Convergence of Reinforcement Learning to Nash Equilibrium: a search market

Eric Darmon, Roger Waldeck

Convergence of reinforcement learning to Nash equilibrium: A search-market experiment

Eric Darmon[a,*], Roger Waldeck[b]

[a]*GREDEG-CNRS and Université de Nice Sophia-Antipolis, 250 Avenue Albert Einstein, F-06560 Valbonne, France*
[b]*GET ENST-Bretagne Department LUSSI and ICI, Technopole Brest Iroise CS 83818, F-29238 Brest Cedex 3, France*

Since the introduction of Reinforcement Learning (RL) in Game Theory, a growing literature is concerned with the theoretical convergence of RL-driven outcomes towards Nash equilibrium. In this paper, we apply this issue to a search-theoretic framework (posted-price market) where sellers are confronted with a population of imperfectly informed buyers and take one decision per period (posted prices) with no direct interactions between sellers. We focus on three different scenarios with varying buyers' characteristics. For each of these scenarios, we quantitatively and qualitatively test whether the learned variable (price strategy) converges to the Nash equilibrium. We also study the impact of the temperature parameter (defining the exploitation/exploration trade off) on these results.

Keywords: Reinforcement learning; Nash equilibrium; Search market; Agent-based modeling

1. Introduction

Since the seminal paper of Erev and Roth (see Ref. [1]) introducing Reinforcement Learning (RL) as an efficient modeling tool to approach human actual behavior,

*Corresponding author.
E-mail address: darmon@idefi.cnrs.fr (E. Darmon).

a growing literature tried to compare theoretically the properties of large multi-agent systems driven by RL to that of the Nash outcome. One key issue is to determine in which conditions these two may coincide. For example, Ref. [2] considers this issue in a congestion game analogous to a market entry game. Comparing two specifications of the RL algorithm, the author sketches two situations: in the first one, she considers one isolated RL agent that plays against the $n-1$ other agents endowed with fixed mixed strategies and so do not react on the RL-agent's decisions. In this setting, it is shown that initial conditions play a crucial role on the final strategy played by the RL agent. The case of n RL agents is then examined. The main conclusion is that this second case exhibits more rapidly a stable aggregate behavior and that the relative performance of one RL algorithm depends on the type of considered environment (endogenous evolving versus constant).

This paper provides an illustration of the convergence issue on a simple decentralized market (see Refs. [3,4] for examples of RL applications to decentralized markets). Such markets have been used to analyze price formation when the market is not ruled by an auctionner. Recently, these models have been applied to e-commerce in order to explain the persistant price dispersion despite consumers' lower search costs (see Refs. [5,6]). In this respect, Ref. [7] shows that the link between information and pricing is often misleading in the context of search theory models. Notably, better information may lead to higher price dispersion and a more intensive search by shoppers may lead to higher prices in symmetric mixed strategy equilibrium of the game. We here consider a simple posted-price market of that type: imperfectly informed buyers wish to buy an homogeneous item at the best price on one hand, and sellers try to obtain the maximum profit by setting discrete prices within a bounded range of potential prices on the other hand. There are two types of buyers: (i) uninformed (visit randomly one seller period and shop if the proposed price is less than their reservation price) and (ii) informed buyers (visit k sellers per period and buy at the firm setting the lowest of the k prices if this price is less than their reservation price). The repartition between the two types is governed by an exogenous parameter a. Buyers' behavior is then characterized by the couple (a,k). We can identify two polar cases namely Case I (''competitive setting'') where there is only informed buyers and where k is equal to the number of sellers and Case II (''monopolistic competition'') where there are only uninformed buyers. For Case I, the Nash equilibrium is the Bertrand competitive outcome, while in Case II, the Nash equilibrium is the monopoly outcome. For intermediate cases, Ref. [7] already established the Nash equilibrium in mixed strategies. For these three cases, we test whether the distribution of learned prices converges qualitatively and quantitatively to the Nash mixed equilibrium.

The remainder of this paper is divided as follows: Section 2 presents the simulation model and the implementation of the RL algorithm. Section 3 summarizes the results for the two polar cases (Cases I and II). Section 4 presents the results for one representative intermediate case. Section 5 concludes.

2. The model

We consider a posted-price market where S sellers (indexed by s) and B buyers, respectively, produce and consume an indivisible and homogeneous item. The timeline of a session t ($t = 1, \ldots, T$) is as follows: (i) sellers post prices; (ii) buyers visit sellers and transact; (iii) sellers compute their profits and reward the pricing rule.

On the demand side, B buyers need to purchase one unit of an indivisible good at each period. Their reservation price (v) is identical and will further be equal to 100. We distinguish two types of buyers: (i) informed buyers (in proportion a of the total population of buyers with $a \in [0, 1]$) who systematically visit k sellers per session (fixed sample search strategy) and buy at the lowest proposed price; and (ii) uniformed buyers who visit only one seller per session and buy at that price (since this price not higher than v). We deal with non-repeated purchases, so that the identity of the sampled sellers is randomized at each session (no recall effect).

On the supply side, sellers post price simultaneously and hence independently of the other sellers (no direct imitation). Prices are posted at the beginning of the session (no bargaining). Since sellers know buyers' reservation price v, posted prices cannot be higher than v. Besides, we suppose that sellers incur a constant and identical per unit cost (c) that will be further set without loss of generality to 0. Consequently, possible prices can range from 0 to 100. We will suppose that posted prices (p_t) are discrete within that range with a 1%-step, hence the set of potential prices p_t is the set of integers $\{0, 1, 2, \ldots, 100\}$. We can hence consider each possible price as an independent rule ({Rule #0: Set price 1}, {Rule #1: Set price 1}, etc.) that can be further modeled by a RL. In this setting, rules have no condition part and are then fully determined by a couple {action, fitness}. Let us note $F^i_{t,s}$ the fitness of the pricing rule i of Seller s at period t. RL processes are controlled by two elements: (i) a selection mode that governs which rule should be activated at the current period; and (ii) an updating mode that governs how the agent records its experiences.

- *Selection mode*: The selection mode is usually expressed as a tradeoff between exploration (of new rules) and exploitation (of past ones). Such a tradeoff can be reproduced through a stochastic selection mode using Boltzmann distribution. Each rule is then selected with the following probability:

$$prob\{Select\ Rule\ i\} = \frac{e^{\frac{\tilde{F}^i_{t,s}}{\tau}}}{\sum_j e^{\frac{\tilde{F}^j_{t,s}}{\tau}}} \quad (\forall i, \forall s, \forall t \geqslant 1)\,, \tag{1}$$

  where $\tau > 0$ and where $\tilde{F}_{i,s}$ are the fitnesses normalized between 0 and 1. Parameter τ ('*temperature*') sets the tradeoff between exploration and exploitation: higher τ lead to a more frequent exploration of all possible pricing rules.
- *Updating mode*: Rules are rewarded by the actual profit generated as they are used ($\pi_{t,s} \coloneqq (p_{t,s} - c)(n_{t,s})$ where $n_{t,s}$ is the number of actual transactions of seller s at

period t). The updating mode is the following:

$$F^i_{t+1,s} = F^i_{t,s} + \alpha(\pi_{t,s} - F^i_{t,s}) \text{ with } \pi_{t,s} \coloneqq (p_{t,s} - c)(n_{t,s}) \quad (\forall i, \forall s, \forall t \geqslant 1) . \tag{2}$$

With this formulation, the fitness of a rule is a weighted average of its past payoffs. As parameter α (with $\alpha \in]0, 1]$) increases, sellers' memory decreases i.e. past experiences have a less important weight in his current decisions.

At the first period, sellers have uniform expectations about the potential profit generated by all pricing rules. The initial fitness F_0 is then identical for every seller and every rule. Coefficient F_0 can also be interpreted as sellers' expected belief about market profitability. $B(v - c)$ is the maximum potential profit on this market (case of a single seller selling at the monopoly price v). When parameter δ decreases, sellers' initial beliefs are less and less enthusiastic.

$$F^i_{t=0,s} = F_0 = \delta B(v - c) \quad (\forall i, \forall s, \delta \in [0, 1]) . \tag{3}$$

We implemented a JAVA multi-agent model to simulate the model. The whole material (source code, classes and an executable interface) are available on request to the corresponding author or at the following URL: http://e.darmon.free.fr/fssmarket/. Due to lack of space, we could not report all the related charts and report the reader to [8].

3. Case I- and Case II-results

3.1. Bertrand competitive setting (Case I)

In this setting, all buyers are informed and systematically visit the whole set of sellers ($(a, k) = (1, S)$). Hence, the Nash-theoretical model predicts a degenerated distribution of posted prices, i.e. an equilibrium in pure strategies where the unique posted price is equal to the production cost c. In a discrete setting, one should then expect posted prices to converge to the production cost incremented by one price step, i.e. 1 here.

Fig. 1 shows a representative run of this particular situation.[1] From these figures, we can see that sellers learn to play the competitive outcome despite their initial ignorance of buyers' behavior and of the strategies of other sellers: after 800 time steps, the average posted price converges to the unique competitive price and sellers post this price with a 98% frequency. As expected, the price dispersion of accepted prices is null unlike that of posted prices which decreases after convergence but does not vanish. This can be easily explained: we assumed that the temperature coefficient τ is constant over time and we did not calibrate it arbitrarily to exogenously decrease

[1]It should, however, be noted that while the stationary positions of two different one-shot runs are qualitatively identical, the dynamics leading to those stationary positions can vary from one run to another. As we only deal with stationary outcomes in this paper, we cannot account for this last aspect but some regularities (e.g. periodic movements of prices) can often be reported.

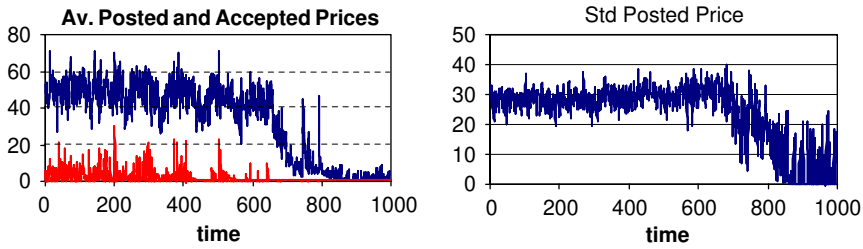

Fig. 1. Representative shot of Case I with $\alpha = 0.8$, $\tau = 0.1$ and $\delta = 1$. Over the last 100 periods, the average posted (resp. accepted) price of the distribution is 2.04 (resp 0.99). The standard deviation of proposed (resp. accepted) prices is 4.32 (resp. 0).

when the system reaches a stationary position. For that reason, sellers never stop exploring alternative strategies. Hence, there is always a non-vanishing fraction of sellers (2% on average) that try other pricing strategies at each period.

The previous result has been obtained with a temperature τ equal to 0.1. To test the impact of the temperature, we ran another simulation with the same parameters but with a higher temperature i.e. $\tau = 0.2$ (see charts in Ref. [8]). Sellers here use alternative price strategies more frequently. The competitive pricing strategy is played only with a 55.85%-frequency (as compared to 98% previously). As expected, higher temperature degrees lead to a more frequent use of alternative rules, i.e. rules that do not have the maximum fitness at period t. However, it should be kept in mind that these price experimentations do not yield better profits: in this setting, buyers have a complete overview of the prices posted on the market, they can hence select only the best price seller. This explains why the gap between the average posted price and the average accepted price increases as the temperature τ increases. From sellers' point of view, deviating from the competitive strategy, leads to a decrease in profits whenever the number of sellers is sufficiently high. In game-theoric terms, pricing strategies other than the competitive outcome ($p_{t,s} = 1$) are then strictly dominated once prices have converged. Using a multi-shot analysis, we can generalize this conclusion by repeating the previous simulation with different τ parameters discretely and randomly drawn in $]0, 0.55[$. For each simulation run, we record the average and standard deviation of the prices posted once the processes have converged (see Fig. 2).

As one can see, sellers play the Nash equilibrium more and more accurately as the temperature decreases. On the contrary, as the temperature increases, average posted prices increase. Considering the magnitude of payoffs normalized between 0 and 1, temperature coefficients higher than 0.5 lead to random choices. When choices are random, the distribution of posted prices converges to a uniform distribution over $[0, v] = [0, 100]$ which mean is hence 50. However, as previously noted, we can notice that temperature variations have a weak impact on the average accepted price as an increase in the temperature leads to an excessive experimentation. Because buyers can perfectly switch to the most competitive seller, the average accepted price remains low. If we consider the highest temperature degrees, we can notice a slight

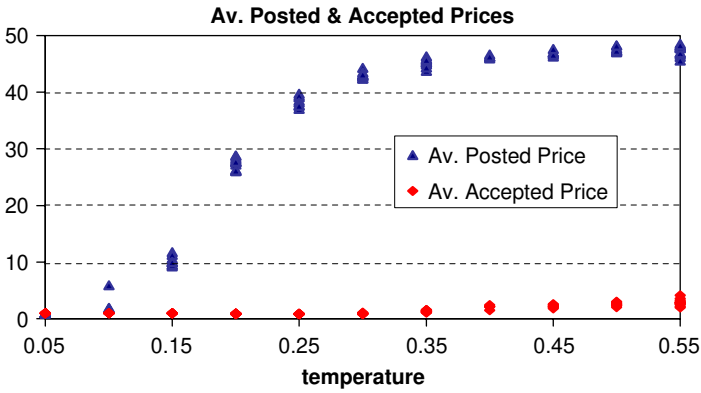

Fig. 2. Average posted and accepted prices with temperature parameters varying from 0.05 to 0.55 ($\alpha = 0.8, \tau = 0.1, \delta = 1$ and 100 iterations). Each point represents the average of the posted (or accepted) prices over the last 100 periods.

increase in accepted prices. This can be easily explained: as choices are random, the probability that one of the S sellers choose the competitive price decreases, thus increasing the average accepted price. In fact, the expected value of the minimum of 20 prices drawn from a uniform distribution on $[0, 100]$ can be shown to be equal to $100/21$.

In short, the temperature plays a more important role by defining the frequency at which the Nash strategy is selected at equilibrium. However, the impact of the temperature coefficient would be less important if this coefficient would be controlled to decrease over time.

3.2. Monopolistic competition (Case 2)

We here consider a case where all buyers are uninformed ($a = 0, \forall k$). In a Nash setting, sellers are informed about this feature. Consequently, buyers are "captive" and sellers can exploit this local monopoly situation by setting the monopoly price v. This leads again to an equilibrium in pure strategies and to a degenerated distribution with no price dispersion (null price variance; average price equal to v). As previously, this result only holds if buyers' characteristics are common and sellers' behaviors are symmetric. Fig. 3 presents a representative run illustrating this situation:

The learned distribution now has one peak located at the monopoly price. From the previous figure, we can deduce that sellers partially learn to adapt to the characteristics of the population of buyers. The average posted price, once the process has converged to a stationary position is here 74.5 against $v = 100$ if sellers would have played the Nash equilibrium. In the same way, the learned standard deviation of posted (accepted) prices is equal to 24.55 against 0 if sellers would have played the Nash equilibrium. In this respect, Cases I and II are not symmetrical: as

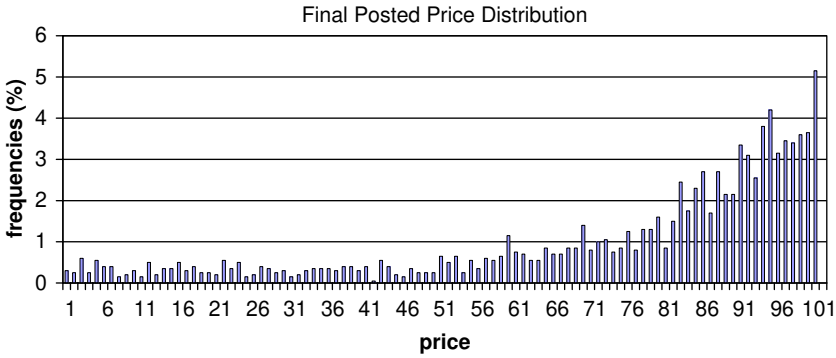

Fig. 3. Case II with $\alpha = 0.8$ and $\delta = 1$. Over the last 100 periods, the average posted (and accepted) price of the distribution is 74.11. The standard deviation of proposed (and accepted) prices is 24.55.

the temperature parameter τ was set at 0.1, sellers played Nash equilibrium with a 98%-frequency in Case I. Here, with the same temperature parameter, sellers play Nash with a 5%-frequency only. This cannot be accounted to a coordination problem among sellers: as $a = 0$, the payoffs of one particular seller is on average independent from the payoffs of other sellers, since the population of buyers is captive to one seller, and since the number of buyers received by one seller at each period (B/S) is on average constant. Cases I and II are asymmetric because the payoff structure (not known by the sellers themselves) is different once a stationary position is achieved. In Case I, a small deviation from the competitive pricing strategy causes a sharp decrease in profits, because buyers can detect any difference in posted prices and thus instantaneously switch to the most competitive seller. This does not hold in Case II: as a seller experiments a strategy close to the monopoly pricing strategy (e.g. $v - 1$), the decrease in profit is smooth and the two pricing rules receives a comparable profit. Considering the stochastic selection mode, those strategies continue being played with a positive probability.

Let us illustrate this by the following simplified numerical example. Consider the two pricing rules $\{p = v\}$ and $\{p = v - 1\}$. On average, any seller receives $B/S = 50$ uninformed buyers at each period. Uninformed buyers conclude a transaction with this seller whatever the posted price (once $p \leqslant v$). If seller sets $p = v$ (resp. $p = v - 1$), it receives $\pi = 20(100) = 2000$ (resp. $\pi = 20(99) = 1980$). Let us assume that this seller just plays these two strategies: considering the stochastic selection mode expressed by Eq. (1) and since $\tau = 0.1$, Strategy $\{p = v\}$ is played with probability 0.512 while Strategy $\{p = v - 1\}$ is played with probability 0.488. As one can see, a slight difference in the two payoffs leads the two strategies to be played with very close probabilities. As we lower the temperature coefficient, these probabilities are distorted and the probability of playing the monopoly price v increases. In Ref. [8], we illustrate this argument for $\tau = 0.01$. As we can note, the monopoly outcome is played with a greater frequency (13.7%), as the temperature is lowered. The whole posted-price distribution shifts left as revealed by its higher average (79.32). Iterating

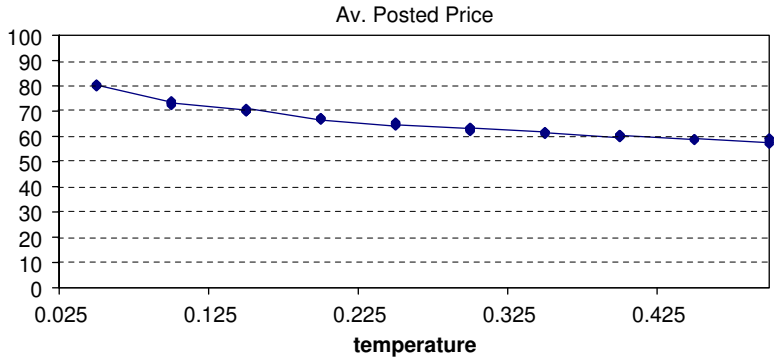

Fig. 4. Average posted price with temperature parameters varying from 0.05 to 0.55 ($\alpha = 0.8$, $\delta = 1$ and 100 iterations). Each point represents the average of the posted prices over the last 100 periods.

the same experiment for a smaller τ coefficient ($\tau = 0.01$), we can see that the learned distribution is slowly converging towards the Nash distribution.

As previously, we used a multi-shot analysis to determine qualitatively the impact of the temperature coefficient on the stationary position (cf. Fig. 4).

Fig. 4 shows that the convergence to the Nash equilibrium is better achieved if the temperature coefficient decreases. However, as indicated by the first representative shots, this convergence is less perfect than that observed in Case I, as a consequence of the payoff structure. Comparing Cases I and II yields to the following conclusion: in this search experiment, the ability of RL sellers to converge to a Nash equilibrium in pure strategies (such as Cases I and II) highly depends on the payoff function once a stationary position has been reached. A non-continuous payoff structure favors here the convergence to the pure equilibrium while a continuous payoff structure makes agents continue playing mixed strategies. The payoff structure would be neutral if sellers would only select the "greedy action" at equilibrium. Again, this could be achieved trough an exogenous decrease in the temperature over time.

4. One illustrative intermediate case

Between the two polar cases presented in the previous sections, many intermediate situations could be considered. In all these intermediate situations, the Nash profit-maximizing strategy can be expressed as a balance between a "high price–low sales" and a "high sales–low price" strategies (see Ref. [7] for a proof). In the first case, sellers target the captive (uninformed) population of buyers while they target the informed fraction of buyers in the second case. As previously, in order to implement such a strategy, sellers would need to know (i) how buyers' types are distributed (i.e. parameters a and k) and (ii) that other sellers will exhibit the same behavior (symmetric equilibrium). Again, we ask to what outcome the final distribution of posted prices converges if both conditions are not filled ex ante.

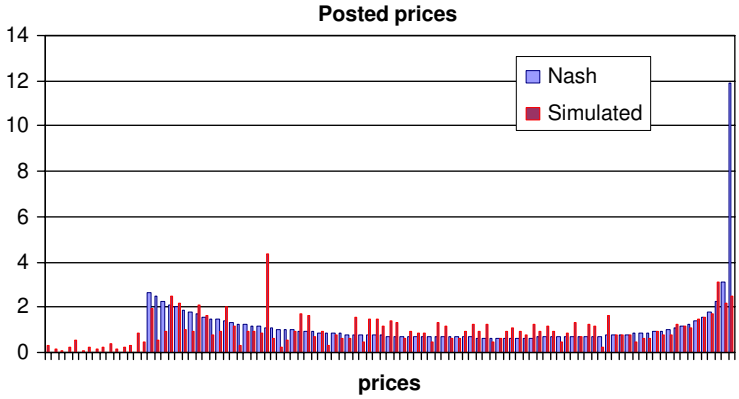

Fig. 5. Learned posted-price distribution (with $a = 0.6$, $k = 4$ and $\tau = 0.05$).

To illustrate this situation, let us consider a particular case[2] where $a = 0.6$ and $k = 4$. Fig. 5 presents the Nash distribution of posted prices (see Ref. [7] for a proof). This distribution can be computed and exhibits the following characteristics: (i) inverted U-shaped bimodal distribution with one peak located at the monopoly price v and a second peak located at the minimum bound of the distribution $p_{\min} = (1-a)v/((k-1)a+1) \simeq 14.3$ strictly superior to the competitive price as $a \neq 1$); (ii) mean (resp. standard deviation) of posted prices is equal to 59.3 (resp. 30.6).

The following chart plots the simulated and the theoretical distribution of posted prices. To obtain the simulated distribution, we recorded the prices stored by all sellers in the last 100 market sessions (i.e. once the process has converged).

We can reject[3] the assumption that the learned distribution converges to the Nash distribution. As previously, we decreased the temperature parameter to determine whether the non-convergence result was caused by an excessive exploration of less-preferred strategies. We ran different runs with the same set of parameters (see Ref. [8]) and noted that despite initial identical parameters, the learned distributions of posted prices still exhibit more variability than what we observed in Cases I and II. However, these differences do weakly impact on the average and standard deviation of posted prices. By plotting the two distributions together, we can identify the origin of the differences between the Nash and the learned distributions: first, sellers continue playing with a positive probability prices inferior to $p_{\min}$. Without knowing buyers characteristics (a,k), sellers fail to coordinate efficiently on the minimum price, and hence play more competitive prices more frequently. Second,

[2]Due to lack of space, we cannot report the results of the simulation model for all (a,k) configurations. The reader can refer to Ref. [9].

[3]We performed a Kolmogorov-Smirnoff adequation test and tested the hypothesis that the learned and the Nash distributions are equal. With a 5% significance level, such tests systematically rejected the equality assumption.

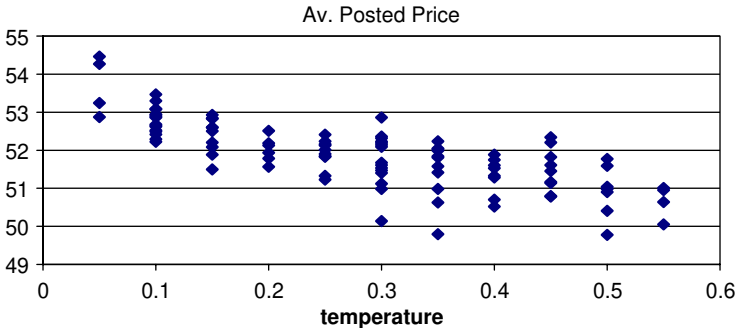

Fig. 6. Average posted price with temperature parameters varying from 0.05 to 0.55 ($\alpha = 0.8, \delta = 1$ and 100 iterations). Each point represents the average of the posted prices over the last 100 periods (with $a = 0.6$, $k = 4$).

the monopoly pricing strategy does not emerge solely but sellers just learn to play "high-price" strategies ($p > 80$) more frequently.

In this intermediate case, increasing or decreasing the temperature does not enable sellers to better approximate the Nash equilibrium. We computed a multi-shot analysis with varying temperature degrees (cf. Fig. 6): Unlike Cases I and II, the temperature has no direct impact on the convergence towards the Nash mixed equilibrium. Again, as τ increases, choices are more and more randomized so that the average posted price tends to 0.5.

5. Conclusions

This aim of this paper was to illustrate the convergence of RL system to a Nash equilibrium within a simple search-market framework. We studied three cases: two polar cases (Nash equilibrium in pure strategy) and one intermediate case (Nash equilibrium in mixed strategies). Inferring from these three cases, we cannot formulate any definite acceptance/reject conclusion regarding the convergence of the RL-based distribution of posted prices towards the Nash one. Considering the competitive outcome case, we showed that sellers learn quite perfectly the Nash distribution whatever the RL coefficients. We showed the role of the temperature (exploration/exploitation parameter) on the final outcome: a decrease in temperature makes choices converge to the "greedy action" (competitive outcome). In Case II (monopolistic competition) it has been shown that sellers learn (although less perfectly) to adapt to buyers' characteristics, by setting prices close to the monopoly price. Comparing Case I with Case II, we noted that sellers endowed with the same exploration/exploitation parameter play the Nash equilibrium in Case II less frequently than what they do in Case I. This illustrates the potential role of the payoff structure in the convergence towards the Nash equilibrium. Finally, we considered one intermediate case where the Nash equilibrium is to play mixed strategies (tradeoff between high sales-low margins and low sales-high margins).

We showed that, whatever the temperature coefficient, sellers were not able to learn such refined pricing strategies. Hence, the system does not converge to the Nash equilibrium. Such a conclusion confirms [10]. In Ref. [10], it has been shown that in normal form games with a unique mixed strategy equilibrium, reinforcement learning, of the type used in [1], does not converge to this type of equilibrium. Moreover, the expected motion of reinforcement learning is given by the evolutionary replicator dynamics and so inherits from the same stability properties. This failure can be analyzed as a coordination problem: to implement such refined strategies, sellers would need (i) to anticipate buyers' behaviors first and then (ii) to anticipate the strategies of other sellers. The first requirement is imperfectly filled: comparing the three situations sketched in this article, we see that sellers learn to adapt to various buyers' behaviors over time despite their initial ignorance. The second requirement is hard to achieve as no communication, neither direct nor indirect, through e.g. imitation is possible.

This finally leads us to a double conclusion: on the one hand, sellers are able to guess an approximate Nash equilibrium in the two polar cases but not in the intermediate one. On the other hand, even if, they are not able to implement a Nash pricing strategy, sellers are able to indirectly infer the characteristics of buyers' behaviors so that the learned distribution of prices self-adjusts to variations in buyers' characteristics. Even if the Nash and the learned distribution of posted prices are different, future work is needed to precisely study the impact of a and k on the characteristics (mean, average) of the price distribution. Another interesting extension would be use alternative learning algorithms (imitation of most successful sellers, refined learning algorithm such as a one-parameter self-tuning EWA, see Ref. [9]).

Acknowledgements

We are grateful to the participants of the First Bonzenfreies Conference on Market Dynamics and Quantitative Economics (Alessandria, September 2004, 9 and 10th) and of the 2nd ELICCIR Workshop (Grenoble, September 2004, 13 and 14th) for their remarks and comments. This paper has partially benefited from the support of the Program "Complex systems in Social Sciences" of the French National Center for Scientific Research (CNRS).

References

[1] I. Erev, A.E. Roth, Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria, Am. Economic Rev. 88 (1998) 848–881.
[2] A.M. Bell, Reinforcement learning rules in a repeated game, Computat. Econ. (2001) 89–111.
[3] A.P. Kirman, N.J. Vriend, Evolving market structure: an ACE model of price dispersion and loyalty, J. Econ. Dynam. Control 25 (2001) 459–502.
[4] T. Brenner, A behavioural learning approach to the dynamics of prices, Computat. Econ. 19 (2002) 67–94.

[5] E. Brynjolfsson, M. Smith, Frictionless commerce? A comparison of internet and conventional retailers, Manage. Sci. 46 (2000) 563–585.
[6] M.D. Smith, J. Bailey, E. Brynjolfsson, Understanding digital markets: review and assessment, in: E. Brinjolffson, B. Kahin (Eds.), Understanding the Digital Economy: Data, Tools and Research, MIT Press, Cambridge, MA, 2000.
[7] R. Waldeck, Search and price competition, J. Econ. Behav. Organ., forthcoming.
[8] E. Darmon, R. Waldeck, Technical appendix of "Convergence of reinforcement learning to Nash equilibrium: a search-market experiment", document available at http://e.darmon.free.fr/in/fssmarket/techphysa.pdf (2005).
[9] E. Darmon, R. Waldeck, Does it matter to play Nash? The case of adaptive sellers, Working Paper downloadable at http://e.darmon.free.fr/in/fssmarket/paperfss.pdf (2004).
[10] M. Posch, Cycling in a stochastic learning algorithm for normal form games, J. Evol. Econ. 7 (1997) 193–207.