Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Economic Dynamics & Control 27 (2003) 2207–2218
JOURNAL OF Economic Dynamics & Control
www.elsevier.com/locate/econbase

Learning competitive pricing strategies by multi-agent reinforcement learning

Erich Kutschinski[a,*], Thomas Uthmann[b], Daniel Polani[c]

[a] *Centrum voor Wiskunde en Informatica, P.O. Box 94079, Amsterdam, Netherlands*
[b] *Institut für Informatik, Universität Mainz, Germany*
[c] *Institute for Neuro- and Bioinformatics, Medical University Lübeck, Germany*

Abstract

In electronic marketplaces automated and dynamic pricing is becoming increasingly popular. Agents that perform this task can improve themselves by learning from past observations, possibly using reinforcement learning techniques. Co-learning of several adaptive agents against each other may lead to unforeseen results and increasingly dynamic behavior of the market. In this article we shed some light on price developments arising from a simple price adaptation strategy. Furthermore, we examine several adaptive pricing strategies and their learning behavior in a co-learning scenario with different levels of competition. Q-learning manages to learn best-reply strategies well, but is expensive to train.

JEL classification: C63

Keywords: Distributed simulation; Agent-based computational economics; Dynamic pricing; Multi-agent reinforcement learning; Q-learning

1. Introduction

For the last few years the Internet has profoundly affected the retail trade of standardized consumer goods, from books and CDs to intercontinental flights. For a single product consumers enjoy a wide choice of offers and can easily compare prices by using efficient search engines and various on-line price comparison services. But the interaction between the consumer (or the price search agent) and retailer is typically very limited—it mainly consists of obtaining price statements and eventually sending

* Corresponding author.
E-mail addresses: erich.kutschinski@cwi.nl (Erich Kutschinski), uthmann@informatik.uni-mainz.de (Thomas Uthmann), polani@inb.mu-luebeck.de (Daniel Polani).

0165-1889/02/$ - see front matter
PII: S0165-1889(02)00122-7

orders. For the future we envision a much more sophisticated trade on the Internet benefiting both consumers and retailers: personalized agents entering into actual negotiations would be able to act on behalf of consumers or retailers, locating specific products or variants, discussing terms of delivery or special conditions, and performing the transactions automatically, based on their owners' preferences. Market space (Eriksson et al., 1999) is one of the earliest comprehensive designs of such a system. Since then, various other commercial services allowing limited automated interaction have sprung up (for example automated bidding at `ebay.com`).

The currently very simple agents in these environments are in the process of being further refined to adapt their bargaining strategies to user preferences or opponent behavior. This adaptation could be effected directly through the user himself. A more efficient approach though is letting the agents learn to adapt their strategies using past experience and observations. Invariably electronic markets will become more complex and their behavior more difficult to predict, since they will consist of populations of agents adapting to each other. This scenario presents itself as a very interesting field of research.

To examine various models of such markets and their emergent behavior we developed the agent platform DMarks II at the University of Mainz (Kutschinski, 2000). It is an agent framework with decentralized control, and builds on peer-to-peer communication between its entities using the Java-RMI protocol: a design we believe is well suited to model the general structure of the large future electronic markets. The framework allows modelling the agents' strategic behavior from the bottom up, specifying precisely how an agent makes its decisions and how learning from experience is performed, and examine the resulting market quantitatively through simulations.

In this article we want to shed light on price developments in a market with just elementary adaptation rules for both buyers and sellers. Furthermore, we examine the learning capabilities of competitive seller strategies of different complexity. Different types of asynchronous multi-agent reinforcement learning (RL) will be used to determine optimal seller strategies. All experiments are set in a market scenario with an adjustable degree of competition.

The rest of the article is organized as follows: The next section provides some key facts about RL and lists related work. In Section 3 we describe the model of the market, putting more emphasis on the buyers' purchasing strategies. These are left unchanged throughout the experiments in the following sections. Section 4 discusses a market with fixed production and price developments arising from elementary seller pricing strategies. Section 5 introduces more refined pricing based on Q-learning and examines the co-learning of strategies in a market with variable supply. Finally, Section 6 summarizes our results and gives a short outlook.

2. Multi-agent RL

RL is a subfield of machine learning which addresses the task of how to learn to choose optimal actions in order to achieve a goal, given that the effect of these actions towards the goal can be perceived through sensors (Kaelbling et al., 1996).

RL can be applied to robot control tasks as well as to learning to play games, and can approximate successful behavioral strategies through learning from the feedback of its actions. As the problem learning and problem solving phases are not separated, the learner has to balance exploitation of its currently best known strategy to reach its goal and exploration of as yet untested but possibly promising strategies. If the effect of the action towards the goal can only be perceived with temporal delay, or the state of the environment is not fully (or reliably) observable, RL techniques as Q-learning are especially appropriate for the task (Watkins, 1989; Watkins and Dayan, 1992; Sutton and Barto, 1998). As a precondition for RL techniques to perform well on a given task, the learning problem has to have the so-called Markov property: It has to be a static problem in the sense that even though any action chosen by the learner in a particular state may have various effects, the probability distributions over the effects of any action in any state remain fixed during the learning process. Or in other words the environment itself must not change while learning to solve a task in it, and the result of an action may only depend on the current state the learner has reached, and not on its action history.

RL also has been applied to the task of learning in the presence of other agents. This is particularly intriguing since the Markov property obviously does not hold when the other learning agents form part of an agent's environment. Claus and Boutilier (1997) examine a scenario in which two agents have to learn to cooperate to solve a problem, while Littman (1994) deals with a competitive RL setting. Hu and Wellman (1998) formally treat the task of two competitive learners repeatedly playing a fixed matrix game, and prove convergence to one of the game theoretically derived equilibria for their particular setup of full observability of the opponent's strategy.

More recently, Greenwald et al. (1999) and Kephart and Tesauro (2000) addressed the issue of learning to price automatically in electronic markets through application of RL techniques. In Section 5 we will examine a comparable setup.

3. Model of a market

We set up a market scenario in which a single, non-perishable good is being traded against currency in discrete timesteps. In every step sellers produce and store real-valued quantities of the good and fix a price for which it will be sold during the timestep. Buyers inform themselves by querying the sellers' current prices and make a one-shot decision from which seller to buy. They then try to purchase the good according to their demand function from the chosen seller and, if served, consume it and receive a reward. An overview of the sequence of sellers' and buyers' actions is shown in Fig. 1. This model was motivated by the market model of a previous study (Polani and Uthmann, 1999) and by markets examined in Vriend (1996) and Varian (1980). See also Madhavan (2000) and O'Hara (1995).

3.1. Sellers

Sellers keep an account A which is initialized with a small credit c at the start of the simulation. They also keep track of an inventory S of their good, which incurs store

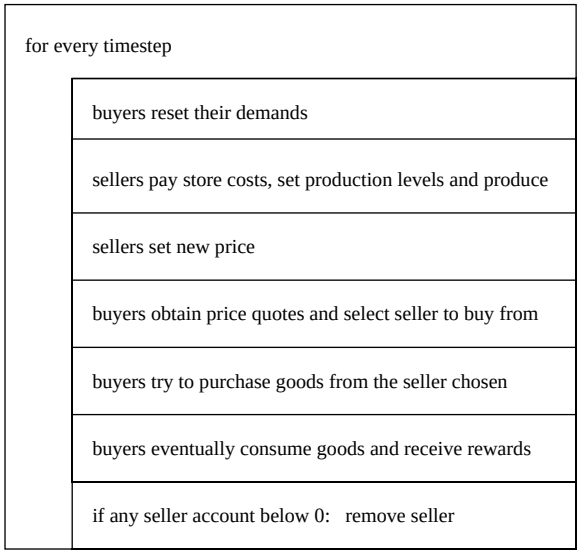

Fig. 1. Sequence of buyer and seller actions per timestep.

costs c_s per unit per timestep. When producing the good to be sold, they are charged production costs c_p per unit. Any profit obtained from selling goods will be placed on the account. If the account ever is depleted, the seller is financially ruined and forced to leave the simulation.

The sellers' behavior in short:

(1) initialize account: $A \leftarrow c$.
(2) pay store costs: $A \leftarrow A - Sp_s$.
(3) set production level to P.
(4) produce: $S \leftarrow S + P, A \leftarrow A - Pp_c$.
(5) set price to p.
(6) serve customer purchases q_i until next timestep: $A \leftarrow A + pq_i$.
(7) if not broke goto (2), else quit.

Setting the production level (3) and fixing a price for the good (5) are the more complex tasks of the sellers. Solution strategies for them will vary throughout our experiments and will be described in more detail in Sections 4 and 5.

3.2. Buyers

Buyers also keep an account A on which they receive a fixed income i every timestep. The account level determines the budget available to satisfy their demands.

Buyers get informed about the current market prices by actively querying the sellers. Assuming that obtaining price quotes may be costly for buyers, we distinguish two

extreme buyer types: The first randomly selects just one of the sellers to query every timestep. The second type will query all sellers in the market and thus is able to fully compare prices. The fraction of fully informed buyers is the only parameter of the buyer population that will vary in our simulations, and we denote it by $\omega \in [0,1]$. Obviously, with increasing ω competition between sellers increases induced by higher price transparency.

A buyer's demand $d(p)$ is determined internally through its inverse demand function $d^{-1}(q)$ which specifies the amount the buyer is willing to pay for a given quantity q of the good, and can therefore be seen as its bid level or marginal rate of substitution (*MRS*). Individual demand is modelled as a stepwise linear function determined by maximum demand $d_{\max}$ and maximum price $p_{\max}$ of a buyer

$$d(p) = \max\left\{0, d_{\max} - p\frac{d_{\max}}{p_{\max}}\right\}, \quad p \geqslant 0. \tag{1}$$

Demand is reduced after consuming goods within a timestep, and reset at the beginning of the next timestep.

After purchasing a quantity q of the good for price p buyers will receive a reward $r(p,q)$ determined as the difference between what they would have been willing to pay for the goods according to their bid level and what they actually paid to the seller

$$r(p,q) = \int_0^q d^{-1}(s)\,\mathrm{d}s - pq. \tag{2}$$

The buyers' behavior in short:

(1) receive income: $A \leftarrow A + i$.
(2) query one or all sellers for prices p_i.
(3) select seller to purchase from: $k \leftarrow \arg\min_i\{p_i\}$.
(4) try to purchase goods: $q \leftarrow \min\{d(p_k), A/p_k\}$.
(5) if successful, consume goods and receive reward: $R \leftarrow R + r(p_k, q)$.
(6) when next timestep goto (1).

After acquiring a number of price quotes for a timestep, buyers thus maximize their expected reward by selecting the seller with the lowest price known and purchasing their full demand for that price. If their account does not allow them to obtain the total quantity, their will purchase what they can afford. If the seller can serve their order, the reward achieved from consumption will be accumulated on the buyer's reward account R, and can be used to measure the efficiency of the buyer's purchasing strategy.

This simple, one-shot behavior of the buyers can be understood within a market in which buyers have different informational preferences (for e.g. one customer type collects all supermarkets' leaflets, while the other just picks up the first one found), and where performing the actual purchase also may be costly (customers only drive to one supermarket to do all their week's shopping; if some goods are sold out, they will have to do without them in the following week). It is a rather basic setup, and provides an environment for the examination of different seller strategies. The described short-term greedy strategy of the buyers is a reduction of the complexity of the market scenario,

and we believe an examination of more complex behavioral strategies for the buyers would be quite rewarding.

We fix general modelling parameters for all our experiments to:

- number of buyers $M = 20$,
- buyer income $i = 0.25$,
- buyer maximum demand $d_{\max} = 1.0$,
- buyer maximum bid level $p_{\max} = 1.0$,
- seller initial credit $c = 100$,
- seller initial price $p = 0.2$.

Some experiments with a range of other parameter values did not lead to significantly different results from those described here. These values also ensure that the income is sufficient to satisfy all of a buyer's demands, and that sellers have a sufficient starting credit to reach more profitable prices unless in very competitive environments. A lower buyer income merely leads to reduced aggregate demand, while a lower credit results in a few sellers dropping out of the market early on, and the rest of the market showing the same behavior as when examined with the reduced number of sellers from the start.

4. Market with fixed production

In this section, we are especially interested in price developments arising from an inventory-based pricing strategy. We fix the production level for each seller to P, thus leaving the price as the only control parameter for the sellers. This can be interpreted as a highly inflexible production chain where sellers must control their profits by adapting their prices. Price adaptation is performed by observing the inventory at every timestep: since store costs per unit per step have to be dealt with, a good pricing strategy will have to keep the inventory close to some value T given by the tradeoff between high store costs and losses due to the unability of serving all buyers' orders. In order to maximize seller profits this value would have to be fine-tuned; but since we are interested in the general behavior of the market for some choice of T, that value remains fixed as well. In general the choice of T does not affect the outcomes of our experiments qualitatively.

Assuming lower prices lead to increased sales, the sellers maximize their long-term profits by raising the price whenever $S < T$, and lowering it when $S > T$. Price adaptation is performed in discrete steps of Δp.

Additional parameters chosen for our experiments with fixed production levels were:

- number of sellers $N = 4$,
- production level $P = 1$,
- inventory threshold $T = 2$,
- price adaptation step $\Delta p = 0.01$.

We found that with increasing price transparency ω in the buyer population competition between the sellers increased, leading to the sellers following each others' prices

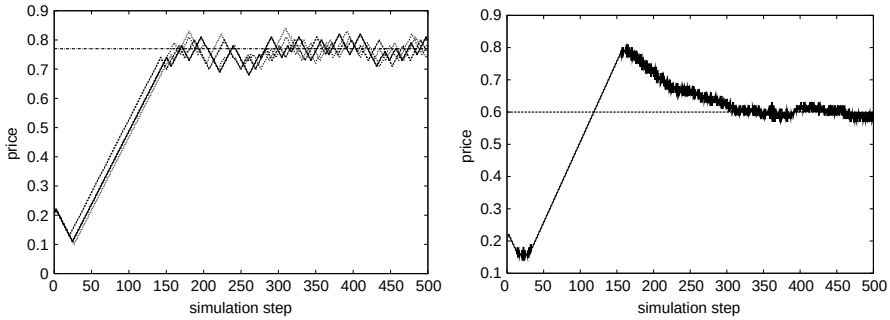

Fig. 2. Price development at $\omega=0.3$ and 1.0, horizontal line indicates average market price between timesteps 300 and 500.

closely. The range of prices chosen by sellers shrunk considerably in more competitive settings: At $\omega = 0.3$ the standard deviation of sellers' prices from the mean market price was $\sigma = 0.019$, while at $\omega = 1.0$ the spread reduced to $\sigma = 0.009$. Eventually, the sellers alternately took turns charging the lowest price of all competitors through implicit coordination (see Fig. 2 left).

In all our experiments the average market price over all sellers for a timestep converged (up to small fluctuations of the sellers as a group) to a value dependent on total supply, total demand, and T, regardless of initial conditions. This value can be computed as follows: Since individual production is fixed, aggregate production can simply be added up over the sellers. Assuming all sellers manage to sell exactly this quantity per timestep, aggregate supply is identical to aggregate production. The market clearing price can now be easily found as the price for which the buyers' aggregated demand equals this quantity. With the given parameter settings one gets

$$q^* = NP = 4.0, \qquad p^* = 1 - \frac{q^*}{M} = 0.8. \tag{3}$$

While competition was not too strong ($\omega \leqslant 0.3$), p^* was quickly and independently reached by all sellers. With growing competition the price reached by the market was significantly lower than the market clearing price. As a reason for this we isolated the concentration of market demand on the single cheapest seller in the most competitive scenarios, and its inability to serve this high demand due to its production limitations. The excess supply of the remaining sellers is taken into the next timestep, and there again presses on the price again. Rather unexpectedly all sellers reach the market clearing price rather quickly, while the price level decays to stabilize at a lower level in the long run (see Fig. 2 right).

This effect occurred in 18 out of 20 independent simulation runs. Further investigation shows this results from all sellers charging exactly identical prices in the early phase sharing demand equally between them. As the market clearing price is reached, this fragile balance breaks up, and sellers enter into the resulting phase of price wars and excess supply, where the market price is driven down again. In the remaining two runs different prices existed in the market from the very start, and the same lower long-run

Table 1
Profits and reward accounts for scenarios with varying level of competition ω

Level of competition ω	Average sellers' A	Average buyers' R
0.0	280.32	55.91
0.3	275.80	56.55
0.5	267.69	57.81
1.0	220.04	69.92

market price was reached without overshooting. This should be the more general case (Table 1).

These experiments show that in a setting with fixed production and an inventory, even a simple price adaptation mechanism just based on inventory levels suffices to reach a consistent market price. Competition induced by price-comparing buyers reduces the variance of the prices charged by individual sellers. Implicit coordination of sellers' pricing strategies appears without any modelling of the opponents, or observing short-term profits. The buyers' one-shot decisions lead to a reduction of the market price below the market clearing price due to excess supply available, and even though buyers frequently cannot satisfy their demand in more competitive scenarios, their average accumulated reward is higher, while the average accumulated seller profits are lower than in less competitive scenarios.

5. Market with variable supply

In the second set of experiments the sellers set their production exactly so that they manage to satisfy all demands. This reflects a scenario with a most flexible production chain. Again sellers maximize their profits by controlling their price only.

5.1. Fixed pricing strategies

In this setting we initially compare $N = 2$ fixed pricing strategies, the derivative follower (DF) and myopically optimal (MO) pricing strategies.

The DF uses the development of profits as compared to the last price change to adapt its future price. It is a variant of an online hillclimbing algorithm that keeps changing its price into the direction of expected profit increase. Doing so it greedily climbs the profit curve. Several DF do reach a common price setting near the monopolistic price, extracting high profits from the buyer population—as long as no other conflicting pricing strategy is introduced into the market (see also Greenwald and Kephart, 1999; Greenwald et al., 1999).

The MO have perfect knowledge of both buyer demand and their competitors' price settings and use this information to select the price yielding the short-term maximum profit, also called myopically optimal best-reply price. In competitive scenarios with discrete fixed price levels this typically leads to undercutting behavior and cyclical

price wars ranging from the monopolistic price down to the competitive equilibrium price, as discussed in Greenwald and Kephart (1999). The MO pricing strategy is, due to its perfect knowledge and short-term rationality, by far superior when competing against the DF strategy. The results achieved in our simulations were comparable to those of the preceding studies already cited. It has to be emphasized that the observed price wars between two or more MO-sellers are a direct result of the discrete price levels used. In a model with continuous price selection no price wars will emerge, but the prices will drop to the Bertrand equilibrium price or competitive price with zero profits for all sellers.

5.2. *RL pricing disregarding the opponent*

To bridge the gap between the reactive DF working from just one previous observation and the myopically rational MO using perfect information, we here introduce a price-profit (PP) adaptation mechanism based on single-state Q-learning. The PP pricing strategy continually uses observed profits to adapt profit expectations for the current price, and chooses new prices through a stochastic selection mechanism based on these expectations. In particular, PP does not use information about the composition of the buyer population or the competitors' price settings (as does MO), but learns from observed profits only (like DF) by building a model of profit expectations for a given price through a robust machine-learning mechanism. Our expectation was that market demand and the competitors' pricing strategies should, in the long run, implicitly be reflected in an average relation between the own price and the observed profit. This would allow PP to adapt to market situations without the need to know about its competitors' prices or pricing strategies. PP developed behavior patterns observed earlier in DF and MO (especially price undercutting) and, as expected, generally outperformed DF in the long run. But lacking an opponent's model and performing the learning rather slowly, it obviously cannot stand against the fully informed MO responding optimally to price changes in any competitive setting.

A more interesting question was how PP learners would adapt their pricing strategies when competing with each other, and whether the individual strategies would converge over time. To examine this we placed several ($N = 2$ and 4) PP sellers in the market, each of them changing its price with a fixed probability every timestep. This way we set up an asynchronous multi-agent reinforcement learning environment. The analysis of these runs showed that the co-learning process has to take place considerably more slowly (i.e. with a smaller learning rate) as compared to learning against fixed pricing strategies to assure convergence of the learning process. The simultaneous learning leads to the development of identical pricing strategies of all PP-sellers with one distinct price maximizing the expected profit (see Fig. 3). This profit maximizing price always lies between the monopolistic and the Bertrand equilibrium price, depending on the amount of competition in the market as determined by the mixed buyer population. Without any competition the PP learners will learn to act as monopolists, while they move closer to the marginal price for $\omega \to 1$.

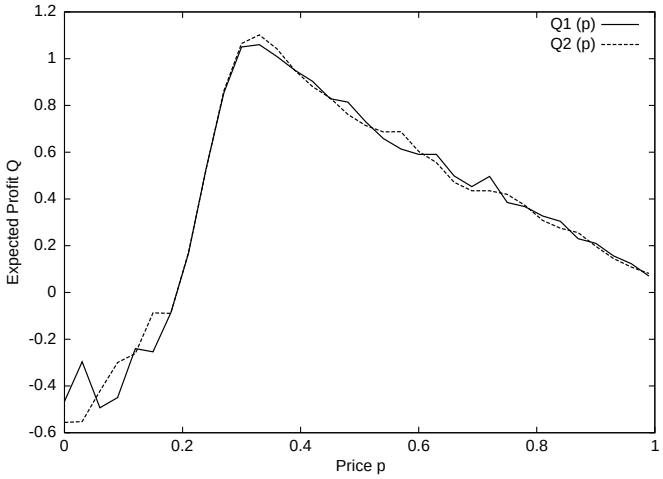

Fig. 3. Learned profit functions $Q_1(p)$ and $Q_2(p)$ of 2 PP sellers at $\omega = 1$.

5.3. RL pricing with a model of the opponent

Finally, we examined a learning price strategy that also takes the competitors' prices into account. As above it is based on Q-learning with a Boltzmann price selection mechanism that trades off exploration against exploitation when choosing future prices. In order to model the influence of competitors' pricing the lowest price on the market determines the state the Q-learning seller is in. This pricing model fully captures the learning task, since only the own price and the lowest of the competitors' prices determine the demand with the given model of the buyer population. We expected the Q-learning sellers to respond more rapidly to the immediate situation on the market than the PP sellers. The identical asynchronous multi-agent reinforcement learning setup as with DF, MO and PP sellers was used.

The simulation runs were performed with no lookahead for the Q-learning algorithm and rewards were immediate. The Q-learning seller managed to learn the profit function and undercutting best-reply pricing strategy against any of the fixed strategies. It developed this optimal pricing policy also against PP and other co-learning Q-adaptive sellers. As could be expected when future rewards are discounted by the factor $\gamma = 0$, the evolved strategy closely resembled the myoptimal pricing strategy (see Fig. 4). This confirms and extends some of the results of Greenwald et al. (1999).

Due to the asynchronous training setup the rate of convergence is considerably slower than in comparable experiments performed by Greenwald et al. (1999). Frequently alternating reinforcements with full knowledge of one single partner's value function are assumed in multi-agent reinforcement learning environments (as in Littman, 1994; Hu and Wellman, 1998). In this work asynchronous learning updates which do not make use of the several partners' value functions led to comparable results, which is an indication that synchronous and asynchronous training processes do not lead to

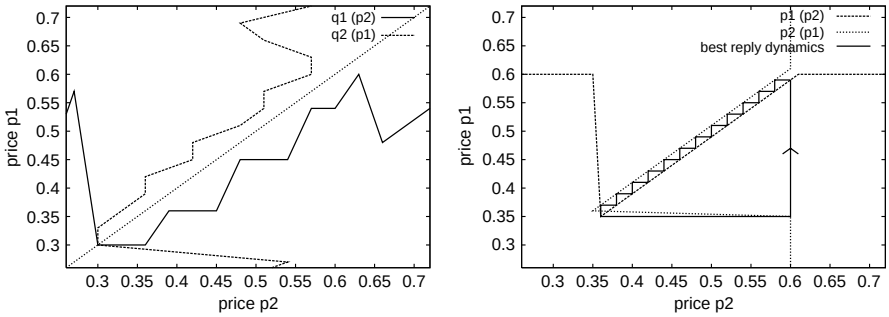

Fig. 4. Best-reply pricing strategies of two competing sellers, plotting the price chosen by a seller as a function of the opponent's price $p_{1,2}$. Learned strategies $q_{1,2}(\cdot)$ on the left; myoptimal strategies $p_{1,2}(\cdot)$ (with price war trajectory) on the right.

qualitatively different results. The number of opponents did not have an impact on the character of the resulting strategies here either. Still, in the asynchronous setup more training updates were needed to remove the additional noise introduced through the stochastic rather than alternate reinforcements. Asynchronous training seems to affect the convergence speed of the algorithms quite badly. Unfortunately simulation runs for discount factors $\gamma > 0$ could not be run with the current implementation due to time constraints. One would expect even slower convergence in this case.

6. Conclusion and outlook

We examined price and pricing strategy development in two different market scenarios, each with varying degrees of competition. Using sellers with a simple price adaptation rule in the scenario with fixed production, convergence to the market clearing price could be observed. Under stronger competition the prices charged by the individual sellers were implicitly coordinated. Excess supply led to a reduction of the market price in the more competitive scenarios.

In the scenario with variable supply the two different pricing strategies using Reinforcement Learning converged to solutions optimal within the scope of their learning model against other fixed and co-learning strategies. While the single-state Q-learners were able to generally distinguish between monopoly and competitive markets, the Q-learners modelling their competitors developed best-reply pricing strategies.

A probably rewarding area of research addresses more complex buyer models and their effect on market price development and the learning of pricing strategies. Furthermore, examining markets of more than one good would allow us to look into dynamics arising from production chains, and form a link between the constant and adaptable supply setting experiments. Here a more simplified setting in which both price and production levels for a single good need to be controlled by the sellers would be a first step.

Acknowledgements

The first author would like to thank the two anonymous referees and K. Somefun for their helpful comments.

References

Claus, C., Boutilier, C., 1997. The dynamics of reinforcement learning in cooperative multiagent systems. In: Sen, S. (Ed.), Collected Papers from the AAAI-97 Workshop on Multiagent Learning. AAAI Press, CA, pp. 746–752.

Eriksson, J., Finne, N., Janson, S., 1999. SICS market space: an agent-based market infrastructure. In: Proceedings of the First International Workshop on Agent-Mediated Electronic Trading, AMET-98, Selected Papers, Lecture Notes in Computer Science, Vol. 1571. Springer, Berlin, pp. 41–53.

Greenwald, A.R., Kephart, J.O., 1999. Shopbots and pricebots. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm.

Greenwald, A.R., Kephart, J.O., Tesauro, G.J., 1999. Strategic pricebot dynamics. In: Proceedings of the ACM Conference on Electronic Commerce (EC-99), Denver.

Hu, J., Wellman, M.P., 1998. Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proceedings of the International Conference on Machine Learning (ICML-98), Madison, Wisconsin.

Kaelbling, L.P., Littman, M.L., Moore, A.W., 1996. Reinforcement learning: a survey. Journal of Artificial Intelligence Research 4, 237–285.

Kephart, J.O., Tesauro, G.J., 2000. Pseudo-convergent Q-learning by competitive pricebots. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML'00), Stanford, CA.

Kutschinski, E., 2000. Simulation eines verteilten Marktnetzwerks mit autonomen Java-Agenten. Master's Thesis, Institut für Informatik, Universität Mainz.

Littman, M.L., 1994. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on Machine Learning (ICML-94), New Brunswick, pp. 157–163.

Madhavan, A., 2000. Market microstructure: a survey. Journal of Financial Markets 3 (3), 205–258.

O'Hara, M., 1995. Market Microstructure Theory. Blackwell, Oxford.

Polani, D., Uthmann, T., 1999. DMarks: Eine verteilte Umgebung fr agentenbasierte Simulationen von Marktszenarien. In: Hohmann, G. (Ed.), Simulationstechnik, 13. Symposium in Weimar, Vol. 3 of Frontiers in Simulation, pp. 391–394.

Sutton, R., Barto, A., 1998. Reinforcement Learning: an Introduction. MIT Press, Cambridge, MA.

Varian, H.R., 1980. A model of sales. American Economic Review, Papers and Proceedings 70 (4), 651–659.

Vriend, N.J., 1996. A model of market-making. Working Paper No. 184, Department of Economics, Universitat Pompeu Fabra, Barcelona.

Watkins, C.C.J.H., 1989. Learning from Delayed Rewards. Ph.D. Thesis, King's College, Cambridge.

Watkins, C.C.J.H., Dayan, P., 1992. Q-learning. Machine Learning 8 (3), 279–292.