Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Economic Dynamics & Control 32 (2008) 3275–3293
JOURNAL OF Economic Dynamics & Control
www.elsevier.com/locate/jedc

Q-learning agents in a Cournot oligopoly model

Ludo Waltman, Uzay Kaymak*

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Received 24 August 2004; accepted 8 January 2008
Available online 9 February 2008

Abstract

Q-learning is a reinforcement learning model from the field of artificial intelligence. We study the use of Q-learning for modeling the learning behavior of firms in repeated Cournot oligopoly games. Based on computer simulations, we show that Q-learning firms generally learn to collude with each other, although full collusion usually does not emerge. We also present some analytical results. These results provide insight into the underlying mechanism that causes collusive behavior to emerge. Q-learning is one of the few learning models available that can explain the emergence of collusive behavior in settings in which there is no punishment mechanism and no possibility for explicit communication between firms.

JEL classification: C63; D43; D83

Keywords: Collusion; Cournot oligopoly; Q-learning; Reinforcement learning

1. Introduction

In this paper, we model the learning behavior of firms in repeated Cournot oligopoly games using Q-learning. Q-learning is a reinforcement learning model of agent behavior originally developed in the field of artificial intelligence (Watkins, 1989). The model is based on two assumptions. First, for each possible strategy an agent is assumed to remember some value indicating that strategy's performance. This value, referred to as a Q-value, is determined based on the agent's past experience with the strategy. Basically, the

*Corresponding author.
E-mail addresses: lwaltman@few.eur.nl (L. Waltman), kaymak@few.eur.nl (U. Kaymak).

0165-1889/$ - see front matter
doi:10.1016/j.jedc.2008.01.003

Q-value of a strategy is calculated as a weighted average of the payoffs obtained from the strategy in the past, where more recent payoffs are given greater weight. The second assumption of Q-learning states that, based on the Q-values, an agent probabilistically chooses which action to play. A logit model is used to describe the agent's choice behavior. The assumptions made by Q-learning can also be found in other reinforcement learning models. The models of Sarin and Vahid (1999, 2001) and Kirman and Vriend (2001) use ideas similar to Q-values, while the models of, for example, Mookherjee and Sopher (1997) and Camerer and Ho (1999) use a logit model to describe the way in which an agent chooses an action. Q-learning distinguishes itself from other reinforcement learning models in that it combines these two elements in a single model. In the economic literature, the combination of these elements has, to our knowledge, not been studied before.

In this paper, we show that the use of Q-learning for modeling the learning behavior of firms in repeated Cournot oligopoly games generally leads to collusive behavior.[1] This is quite a remarkable result, since most Q-learning firms that we study do not have the ability to remember what happened in previous stage games. The firms therefore cannot use trigger strategies, that is, they cannot threaten to punish each other in case of non-collusive behavior. There is also no possibility for explicit communication between firms. However, despite the absence of punishment and communication mechanisms, collusive behavior prevails among firms. Apart from Q-learning, there are almost no models of the learning behavior of individual economic agents that predict collusive behavior in Cournot games. The only model of which we are aware is the so-called trial-and-error model studied by Huck et al. (2004a). Yet, experimental results (for an overview, see Huck et al., 2004b) indicate that with two firms collusive behavior is quite common in Cournot games. Q-learning is one of the few models that does indeed predict this kind of behavior.

Models of the learning behavior of economic agents are studied both in agent-based computational economics (e.g., Tesfatsion, 2003, 2006) and in game theory (e.g., Fudenberg and Levine, 1998). In agent-based computational economics the methodology of computer simulation is typically adopted, whereas in game theory the analytical methodology is predominant. It seems rather difficult to obtain analytical results for the behavior of multiple Q-learning agents interacting with each other in a strategic setting. In the field of artificial intelligence, it has been proven that under certain conditions a single Q-learning agent operating in a fixed environment learns to behave optimally (Watkins and Dayan, 1992). However, for settings with multiple agents learning simultaneously almost no analytical results are available. Given the difficulty of obtaining analytical results, most of the results that we present in this paper are based on computer simulations. Analytical results are provided only for the special case in which Q-learning firms in a Cournot duopoly game can choose between exactly two production levels, the production level of the Nash equilibrium and some other, lower production level. The analytical results turn out to be useful for obtaining some basic intuition why Q-learning firms may learn to collude with each other.

The remainder of this paper is organized as follows. First, in Sections 2 and 3, we provide an overview of related research and we introduce Q-learning. Then, in Section 4, we discuss the Cournot oligopoly model with which we are concerned throughout the paper. We consider our computer simulations in Sections 5 and 6, in which we discuss the

[1]We refer to all firm behavior that results in a joint profit above the joint profit in the Nash equilibrium as collusive behavior. So, collusive behavior does not always mean that firms make the highest possible joint profit.

simulation setup and present the simulation results. We provide some analytical results in Section 7. Finally, in Section 8, we draw conclusions.

2. Related research

The literature on modeling the learning behavior of economic agents is quite large. Overviews of this literature are provided by Brenner (2006) and Duffy (2006). One can distinguish between individual learning models and social learning models (Vriend, 2000). In individual learning models an agent learns exclusively from its own experience, whereas in social learning models an agent also learns from the experience of other agents. Below, we first discuss the modeling of individual learning behavior, and we then consider the modeling of social learning behavior.

The two most important approaches to modeling individual learning behavior are belief-based learning and reinforcement learning. Examples of belief-based learning models are Cournot adjustment and fictitious play (Fudenberg and Levine, 1998). These two models assume that an agent has the ability both to observe its opponents' action choices and to calculate best responses. In a Cournot oligopoly game, the models predict that firm behavior can converge only to the Nash equilibrium.

Reinforcement learning is based on a very simple idea: the higher the payoffs obtained from a strategy in the past, the more likely the strategy is to be played. Compared with belief-based learning models, reinforcement learning models make few assumptions about both the information available to an agent and the cognitive abilities of an agent. For example, in reinforcement learning an agent needs no information about its opponents' action choices or about the payoffs of the game. An agent is only assumed to have knowledge of the strategies that it can play and, after playing a strategy, of the payoff that it has obtained from that strategy. Reinforcement learning models are studied both in the economic literature and in the artificial intelligence literature (for an overview of the artificial intelligence literature on reinforcement learning, see Kaelbling et al., 1996; Sutton and Barto, 1998). Q-learning is a reinforcement learning model that has been studied extensively by artificial intelligence researchers (e.g., Watkins, 1989; Watkins and Dayan, 1992) but that has received almost no attention from economists. In the economic literature, the reinforcement learning model studied by Roth and Erev (1995) and Erev and Roth (1998) is well-known. Bell (2001) performs a simulation study in which this model is compared with Q-learning. Some other reinforcement learning models have been proposed in the economic literature by Mookherjee and Sopher (1997), Sarin and Vahid (1999, 2001), and Kirman and Vriend (2001). These models are all in some way similar to Q-learning. We discuss their relationship with Q-learning in Section 3.

Some preliminary results on Q-learning behavior in a Cournot oligopoly game are reported by Kimbrough and Lu (2003). In their simulation study, the authors find a small tendency towards collusive behavior. In the present paper, we extend the work of Kimbrough and Lu (2003) by analyzing Q-learning behavior in a Cournot game in more detail and by providing an explanation for the emergence of collusive behavior. Furthermore, in the artificial intelligence literature there are some papers in which Q-learning behavior in iterated prisoner's dilemmas or generalizations thereof is studied (Sandholm and Crites, 1996; Littman and Stone, 2001; Stimpson and Goodrich, 2003; Waltman and Kaymak, 2007). Whether Q-learning agents in an iterated prisoner's

dilemma learn to cooperate with each other turns out to depend on the specific values of the prisoner's dilemma payoffs (Waltman and Kaymak, 2007).

The trial-and-error learning model studied by Huck et al. (2004a) also models individual learning behavior. Like reinforcement learning models, the trial-and-error model makes few assumptions about both the availability of information and the cognitive abilities of an agent. However, the underlying idea of the model is different. In a Cournot oligopoly game, the model assumes that a firm keeps increasing (decreasing) its production level as long as this results in a higher profit. As soon as profit falls, the firm starts decreasing (increasing) its production level. Like Q-learning, the trial-and-error model predicts collusive behavior in Cournot games.

A number of studies have investigated social learning behavior in Cournot oligopoly games. In a well-known study by Vega-Redondo (1997), an evolutionary model of firm behavior is analyzed. Vega-Redondo shows that the model predicts convergence of firm behavior to the Walrasian equilibrium. Alós-Ferrer (2004) and Bergin and Bernhardt (2005) extend the model of Vega-Redondo by providing firms with a memory. When firms have a memory, convergence to any outcome between the Walrasian equilibrium and the Nash equilibrium becomes possible (Alós-Ferrer, 2004) and even collusive behavior may emerge (Bergin and Bernhardt, 2005). Social learning behavior in Cournot games has also been investigated using models based on genetic algorithms (e.g., Arifovic, 1994; Vriend, 2000). Depending on the way in which genetic algorithms are applied, such models predict convergence of firm behavior to either the Walrasian equilibrium or the Nash equilibrium or some outcome in between. We further mention the work of Droste et al. (2002), in which social learning behavior in Cournot games is investigated using a model based on replicator dynamics.

Finally, Dixon (2000) and Oechssler (2002) use models with aspiration levels to investigate learning behavior in Cournot oligopoly games. In their models, a firm changes its production level only if its profit is below some aspiration level. The models predict collusive behavior in Cournot games.

3. Q-learning

In this paper, Q-learning is applied as follows. An agent plays a repeated game. At the beginning of the stage game in period t, the agent's memory is in some state s_t. This state may be determined by, for example, the actions played by the agent and its opponents in the stage game in period $t-1$. Taking into account the state of its memory, the agent chooses to play some action a_t. The choice of an action is made probabilistically based on the so-called Q-values of the agent. Playing action a_t results in some stage game payoff π_t that is obtained by the agent and in a transition of the state of the agent's memory from the old state s_t to some new state s_{t+1}. The agent uses the experience gained during the stage game to update its Q-values, thereby modifying the way in which it chooses actions in stage games in future periods.

For a formal discussion of Q-learning, let $Q_t(s,a)$ denote an agent's Q-value for state $s \in S$ and action $a \in A$ at the beginning of period t. The state space S and the action space A are assumed to be finite. The probability that in period t the agent chooses to play action a is given by

$$\Pr(a) = \frac{\exp(Q_t(s_t,a)/\beta)}{\sum_{a' \in A} \exp(Q_t(s_t,a')/\beta)}, \tag{1}$$

where s_t denotes the state of the agent's memory at the beginning of period t and the parameter $\beta > 0$ denotes the experimentation tendency. The larger the value of β, the higher the probability that the agent chooses to experiment, that is, chooses to play an action that does not have the highest Q-value. As β approaches zero, the probability that the agent chooses to experiment approaches zero too. In the artificial intelligence literature, action choice according to probabilities given by (1) is known as the Boltzmann exploration strategy (e.g., Kaelbling et al., 1996; Sandholm and Crites, 1996). Various other approaches to choosing actions have also been studied in the artificial intelligence literature. We model action choice behavior using (1) because this corresponds to a logit model, which is a quite commonly used model of choice behavior in the economic literature (e.g., McKelvey and Palfrey, 1995; Brock and Hommes, 1997; Mookherjee and Sopher, 1997; Fudenberg and Levine, 1998; Camerer and Ho, 1999; Hofbauer and Sandholm, 2007).[2]

After the agent has played some action a_t in period t, the agent's Q-values are updated according to

$$Q_{t+1}(s,a) = \begin{cases} (1-\alpha)Q_t(s,a) + \alpha\left(\pi_t + \gamma \max_{a' \in A} Q_t(s_{t+1}, a')\right) & \text{if } s = s_t \text{ and } a = a_t, \\ Q_t(s,a) & \text{otherwise,} \end{cases} \tag{2}$$

where π_t denotes the stage game payoff obtained by the agent and the parameters $0 < \alpha \leqslant 1$ and $0 \leqslant \gamma < 1$ denote, respectively, the learning rate and the discount factor. The value of α determines the relative weight that is given to recent experience compared to older experience, while the value of γ indicates the time preference of the agent. The update rule in (2) has the appealing property that when there is only one learning agent (either because there is only one agent or because all other agents use fixed strategies), the update rule allows the agent, under certain conditions, to learn to behave optimally. This property has been proven by Watkins and Dayan (1992).

Unlike Q-learning, most reinforcement learning models studied in the economic literature (e.g., Roth and Erev, 1995; Mookherjee and Sopher, 1997; Erev and Roth, 1998; Sarin and Vahid, 1999, 2001) do not consider the possibility that an agent has a memory for remembering past events. In these models, it is not possible for an agent to learn a strategy in which the choice of an action in the current stage game depends on what happened in previous stage games. In this paper, we consider both agents with a memory and agents without a memory. For an agent without a memory, (1) and (2) simplify to, respectively,

$$\Pr(a) = \frac{\exp(Q_t(a)/\beta)}{\sum_{a' \in A} \exp(Q_t(a')/\beta)}, \tag{3}$$

[2]Note, however, that in each of these papers logit models are used in a somewhat different context. McKelvey and Palfrey, for example, use logit models as the basis of the logit equilibrium concept that they introduce in their paper. Brock and Hommes use logit models to model the way in which agents choose between different predictors based on publicly available information on each predictor's past performance. Hofbauer and Sandholm study settings in which there are a large number of agents and in which each of the agents occasionally changes its action according to, for example, a logit model.

and

$$Q_{t+1}(a) = \begin{cases} (1-\alpha)Q_t(a) + \alpha\pi_t & \text{if } a = a_t, \\ Q_t(a) & \text{otherwise.} \end{cases} \tag{4}$$

Note that an agent without a memory cannot take into account the consequences of the action it plays in the current stage game on the payoffs it obtains in future stage games. For such an agent, the discount factor γ in (2) must therefore equal zero.

Sarin and Vahid (1999) and Kirman and Vriend (2001) propose reinforcement learning models that use the same update rule as Q-learning without a memory. The difference between Q-learning and the model of Sarin and Vahid is that in the latter model there is no experimentation, so that an agent always chooses the action from which it expects to obtain the highest payoff. Sarin and Vahid also propose a variant of their model in which an agent experiments depending on its 'state of mind or mood'. However, they do not seem to study this variant further in other papers (e.g., Sarin and Vahid, 2001). In the model of Kirman and Vriend, agents do experiment, but the probabilities with which the various actions are chosen are not the same as in Q-learning. Q-learning without a memory is also related to the learning model proposed by Mookherjee and Sopher (1997). In a similar way as in Q-learning without a memory, Mookherjee and Sopher use a logit model to describe an agent's choice behavior. Since Mookherjee and Sopher do not specify what kind of update rule to use, Q-learning without a memory can in fact be regarded as a special case of their model.

4. Cournot oligopoly model

We consider a simple Cournot oligopoly model with the following characteristics: the number of firms is fixed, firms produce perfect substitutes, the demand function is linear, firms have identical cost functions, and marginal cost is constant. The inverse demand function is given by

$$p = \max\left(u - v\sum_{i=1}^{n} q_i, 0\right), \tag{5}$$

where n denotes the number of firms, p denotes the market price, q_i denotes firm i's production level, and $u>0$ and $v>0$ denote two parameters. Firm i's total cost equals

$$c_i = wq_i \quad \text{for } i = 1, \ldots, n, \tag{6}$$

where the parameter w denotes a firm's constant marginal cost and satisfies $0 \leqslant w < u$. It follows from (5) and (6) that firm i's profit is given by

$$\pi_i = pq_i - c_i = q_i \max\left(u - w - v\sum_{i'=1}^{n} q_{i'}, -w\right) \quad \text{for } i = 1, \ldots, n. \tag{7}$$

The Nash equilibrium of a Cournot model is obtained if each firm chooses the production level that maximizes its profit given by the production levels of its competitors.

Hence, in the Nash equilibrium $\partial\pi_i/\partial q_i = 0$ for $i = 1, \ldots, n$. In the above Cournot model, this implies that firms' joint production level in the Nash equilibrium is given by

$$q^* = \frac{(u-w)n}{v(n+1)}. \tag{8}$$

Consequently, firms' joint profit in the Nash equilibrium equals

$$\pi^* = \frac{(u-w)^2 n}{v(n+1)^2}. \tag{9}$$

Although in the Nash equilibrium firms individually maximize their profit, they do not maximize their joint profit. Firms maximize their joint profit in the collusive equilibrium, in which they collectively behave as a single monopolist. In the above Cournot model, firms jointly produce a quantity of $(u-w)/2v$ in the collusive equilibrium, which results in a joint profit of $(u-w)^2/4v$. In the collusive equilibrium of a Cournot model, firms produce a smaller quantity than the quantity that maximizes their individual profit. Firms therefore have an incentive to increase their production level. For this reason, a collusive equilibrium is unstable and is not a Nash equilibrium. However, things are different in repeated Cournot games, in which firms may be interested in maximizing their long-term profits. If firms play a Cournot game repeatedly and remember what happened in previous stage games, it may be possible to sustain collusion. Trigger strategies may then be used to support collusive behavior, and collusive behavior may constitute a Nash equilibrium of the repeated game.

In addition to the Nash equilibrium and the collusive equilibrium, another outcome that may be obtained in a Cournot model is the Walrasian equilibrium. As discussed in Section 2, this equilibrium is sometimes encountered in studies on social learning behavior in Cournot models. The Walrasian equilibrium is obtained if firms are not aware of their influence on the market price and therefore behave as price takers. In the above Cournot model, firms' joint production level in the Walrasian equilibrium equals $(u-w)/v$. This production level results in zero profits for all firms.

5. Setup of the computer simulations

In this paper, we focus on the long-run behavior of Q-learning agents when the probability of experimentation approaches zero. In this respect, the approach that we take is similar to the approach that is typically taken to analyze evolutionary game-theoretic learning models (e.g., Vega-Redondo, 1997; Alós-Ferrer, 2004; Bergin and Bernhardt, 2005). We further focus on settings in which the learning behavior of all agents is modeled using Q-learning. An alternative would be to consider settings in which the learning behavior of only one agent is modeled using Q-learning and in which all other agents use fixed strategies. However, such settings are less interesting to study. This is because a single Q-learning agent operating in a fixed environment is, under certain conditions, guaranteed to learn to behave optimally (Watkins and Dayan, 1992). This means that when a Q-learning agent competes with agents that use fixed strategies, the Q-learning agent will simply learn a best response to the strategies of the other agents. The settings on which we focus in this paper, that is, settings with multiple Q-learning agents learning

simultaneously, are more interesting to study, because for such settings analytical results are generally not available. Because of the difficulty of obtaining analytical results, most of the results that we present are based on computer simulations. We now discuss the setup of these simulations.

In the simulations, the Cournot oligopoly model introduced in the previous section was used. The values of the parameters u and v in the inverse demand function were, respectively, 40 and 1. The parameter w, which denotes a firm's constant marginal cost, had a value of 4. Simulations were performed for various values for the number of firms n.

In each simulation run, firms played a repeated Cournot oligopoly game that lasted for one million periods. The learning behavior of firms was modeled using Q-learning. Three types of firms were considered in the simulations: firms without a memory, myopic firms with a memory, and non-myopic firms with a memory. All three types of firms had to choose their production level between 0 and 40. Only integer quantities were allowed. Simulations were performed for various values for the learning rate α. During a simulation run, the experimentation tendency β was gradually decreased over time according to

$$\beta(t) = 1000 \cdot 0.99999^t. \tag{10}$$

In this way, the probability of experimentation was almost one at the beginning of a simulation run and almost zero at the end. In other studies on Q-learning (e.g., Sandholm and Crites, 1996), β is decreased in a similar way. At the beginning of a simulation run, firms' Q-values were initialized to zero. Firms with a memory were able to remember their own production level in the previous period as well as their competitors' joint production level in the previous period. For myopic firms with a memory the discount factor γ had a value of zero, while for non-myopic firms with a memory it had a value of 0.9.

6. Results of the computer simulations

In this section, we present the results of the computer simulations that we performed. We first consider the simulations with firms that did not have a memory, and we then consider the simulations with firms that did have a memory.

Simulations with firms that did not have a memory were performed for various values for both the number of firms n and the learning rate α. For each combination of values for n and α, Table 1 shows firms' joint quantity produced and joint profit. Since we focus on the long-run behavior of firms when the probability of experimentation approaches zero, the quantities and profits in Table 1 were calculated by averaging firms' joint quantity produced and joint profit over the last 100 periods of a simulation run. Moreover, because the outcomes of a simulation run depend on the random numbers that are used, 100 simulation runs with different random numbers were carried out for each combination of values for n and α. Table 1 shows the mean of the outcomes of these 100 simulation runs. The corresponding standard deviation is reported within parentheses. For each value for n, firms' joint quantity produced and joint profit in the Nash equilibrium, calculated using (8) and (9), are also reported in Table 1. Firms' joint quantity produced and joint profit in the collusive equilibrium do not depend on n and are equal to, respectively, 18 and 324 (see Section 4).

Table 1 shows that for all combinations of values for n and α the average outcome that emerged in the simulation runs was somewhere in between the Nash equilibrium and the collusive equilibrium. For each combination of values for n and α, the mean of firms' joint

Table 1
Results of computer simulations with firms that did not have a memory

		Nash	$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 1.00$
$n = 2$	Quantity	24.0	22.8 (1.3)	21.2 (1.4)	20.8 (1.2)	20.8 (1.4)
	Profit	288.0	299.1 (11.6)	312.0 (10.2)	314.7 (6.2)	314.3 (7.0)
$n = 3$	Quantity	27.0	25.1 (1.6)	22.0 (1.8)	21.5 (1.9)	22.1 (1.9)
	Profit	243.0	270.7 (22.9)	304.6 (14.5)	307.8 (14.3)	303.7 (16.6)
$n = 4$	Quantity	28.8	26.3 (1.8)	22.6 (1.9)	22.1 (2.4)	22.9 (2.6)
	Profit	207.4	252.1 (29.8)	299.0 (18.7)	301.4 (19.2)	293.2 (25.8)
$n = 5$	Quantity	30.0	27.6 (1.6)	23.2 (1.8)	22.2 (2.2)	23.3 (2.5)
	Profit	180.0	229.3 (30.0)	294.1 (17.3)	301.1 (19.2)	290.2 (28.7)
$n = 6$	Quantity	30.9	28.3 (1.5)	23.3 (2.2)	22.6 (2.6)	23.1 (3.1)
	Profit	158.7	215.4 (32.2)	290.7 (23.5)	296.3 (27.1)	289.1 (34.8)

quantity produced was significantly lower than firms' joint quantity produced in the Nash equilibrium ($p < 0.0001$), while the mean of firms' joint profit was significantly higher than firms' joint profit in the Nash equilibrium ($p < 0.0001$). So, on average firms learned to collude with each other. Full collusion usually did not emerge, since firms usually did not learn to make the highest possible joint profit. It can further be seen in Table 1 that on average firms' joint quantity produced increased as the number of firms n increased. However, since firms' joint quantity produced in the Nash equilibrium also increases as n increases, a substantial degree of collusion remained even for larger values for n. The same observation can be made if firms' joint profit rather than firms' joint quantity produced is considered. Now consider the effect of the learning rate α, which determined the relative weight that firms gave to recent experience compared to older experience. As can be seen in Table 1, a value of 0.05 for α resulted in a significantly lower degree of collusion than a value of 0.25 or higher. However, even for $\alpha = 0.05$ the degree of collusion was significant. For α equal to 0.25, 0.50, and 1.00, the differences in the degree of collusion were quite small. So, the relative weight that firms gave to recent experience compared to older experience did not have a very large effect on the degree of collusion. A substantial negative effect on the degree of collusion was found only when firms gave a rather low weight to recent experience compared to older experience.

Simulations with firms that had a memory were performed for various values for the number of firms n. In addition, both myopic and non-myopic firms were considered in the simulations. A value of 0.50 was used for the learning rate α. Table 2 shows the results of the simulations. The results in Table 2 were calculated in the same way as the results in Table 1. Comparing the results in the two tables, it can be seen that the results obtained for firms with a memory are quite similar to the results obtained for firms without a memory. Like firms without a memory, firms with a memory on average learned to collude with each other. Full collusion usually did not emerge. The degree of collusion seems to be somewhat lower for firms with a memory, both for myopic and for non-myopic firms, than for firms without a memory, but the difference is not very large. Between myopic and non-myopic firms with a memory, no clear difference can be observed in the degree of collusion.

Table 2
Results of computer simulations with firms that had a memory

		Nash	Myopic ($\gamma = 0.0$)	Non-myopic ($\gamma = 0.9$)
$n = 2$	Quantity	24.0	20.8 (0.9)	19.6 (1.2)
	Profit	288.0	314.2 (7.3)	318.0 (6.8)
$n = 3$	Quantity	27.0	22.9 (1.3)	21.5 (1.7)
	Profit	243.0	297.3 (15.8)	304.8 (18.5)
$n = 4$	Quantity	28.8	24.1 (1.4)	23.8 (1.7)
	Profit	207.4	284.2 (19.8)	277.5 (41.3)
$n = 5$	Quantity	30.0	24.4 (1.4)	23.6 (2.1)
	Profit	180.0	280.1 (20.5)	271.6 (75.8)
$n = 6$	Quantity	30.9	24.7 (1.6)	21.9 (2.0)
	Profit	158.7	274.5 (29.2)	288.5 (51.0)

7. Analytical results

In the previous section, we presented simulation results showing that the use of Q-learning for modeling the learning behavior of firms in a Cournot oligopoly game generally leads to collusive behavior. This turned out to be the case not only for firms with a memory but also for firms without a memory. This is quite remarkable, since firms without a memory cannot use trigger strategies, that is, they cannot threaten to punish each other in case of non-collusive behavior. So, collusive behavior prevails among Q-learning firms without a memory even though there is no punishment mechanism and no possibility for explicit communication between the firms. Interestingly, apart from Q-learning, there are very few learning models that predict collusive behavior in such a setting. In this section, we analyze why Q-learning results in collusive behavior. To do so, we make two simplifying assumptions. First, we assume that there are only two firms in the market. And second, we assume that firms can choose between only two production levels, the production level of the Nash equilibrium and some other, lower production level. Under these two assumptions, a Cournot game reduces to a prisoner's dilemma game. Moreover, the behavior of Q-learning firms becomes analytically tractable, and it becomes clear why Q-learning firms may learn to collude with each other.

Consider the Cournot oligopoly model introduced in Section 4. Let there be two firms in the market, that is, let $n = 2$. It follows from the results in Section 4 that in the Nash equilibrium each firm produces a quantity of $(u - w)/3v$ while in the symmetric collusive equilibrium each firm produces a quantity of $(u - w)/4v$. The following theorem provides sufficient conditions for the emergence of collusive behavior among Q-learning firms without a memory.

Theorem 1. *Consider an infinitely repeated Cournot duopoly game based on the Cournot model introduced in Section* 4 *with the number of firms n equal to* 2. *Let the firms' learning behavior be described by Q-learning, and assume that the firms do not have a memory.*

Assume that the firms can choose between two production levels, denoted by q_C *and* q_N, *that satisfy*

$$(u-w)/4v < q_C < q_N = (u-w)/3v. \tag{11}$$

Let π_{CN}, π_{NN}, π_{CC}, *and* π_{NC} *denote a firm's profit, respectively, if the firm produces* q_C *and its competitor produces* q_N, *if both the firm and its competitor produce* q_N, *if both the firm and its competitor produce* q_C, *and if the firm produces* q_N *and its competitor produces* q_C. *Let the learning rate* α *satisfy*

$$\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}} < \alpha < 1. \tag{12}$$

Let each firm's Q-value of producing q_C *be initialized to a value strictly between* π_{CN} *and* π_{CC}, *and let each firm's Q-value of producing* q_N *be initialized to a value strictly between* π_{NN} *and* π_{NC}. *Then, in the limit as the experimentation tendency* β *approaches zero, the proportion of time in which both firms produce* q_C *equals one.*

The proof of the theorem is provided in Appendix A. The basic intuition of the proof is as follows. Consider first what happens when the firms never experiment, that is, when from the production levels q_C and q_N the firms always choose the one with the higher Q-value. Two special states can be distinguished: a 'collusive state' in which under the assumption of no experimentation both firms keep producing q_C forever and a 'Nash state' in which under the assumption of no experimentation both firms keep producing q_N forever. It can be shown that without experimentation the firms will always end up in one of these states, in which they will then remain forever. Consider now what happens when experimentation is introduced. Let the experimentation tendency β have a value close to zero, so that there is only a very low probability of experimentation. With experimentation, the firms will no longer keep producing the same quantity forever, and two kinds of transitions will start taking place: transitions from the collusive state to the Nash state and transitions the other way around. Since the probability of experimentation is very low, it will usually take a long time before a transition from one state to the other occurs. In which of the two states, the collusive state or the Nash state, the firms will spend most of their time depends on the relative likelihood of the two kinds of transitions. If transitions from the collusive state to the Nash state are more likely than transitions the other way around, the firms will spend most of their time in the Nash state. Conversely, if transitions from the collusive state to the Nash state are less likely than transitions the other way around, the firms will spend most of their time in the collusive state. It turns out that for β close to zero transitions from the collusive state to the Nash state are less likely than transitions the other way around. This can be seen as follows. A transition from the collusive state to the Nash state can be shown to require at least one period in which one of the firm experiments (that is, produces q_N rather than q_C). Assuming that the learning rate α satisfies (12), it can also be shown that for a transition from the Nash state to the collusive state one period in which both firms experiment (that is, produce q_C rather than q_N) will usually be sufficient. The probability with which a firm experiments is given by (1) and depends on the difference between the firm's Q-values of producing q_C and q_N. The larger this difference, the lower the probability of experimentation. It can be shown that in the collusive state the difference between a firm's Q-values will usually be equal to approximately $\pi_{CC}-\pi_{NN}$, while in the Nash state the difference will usually be no larger than $\pi_{NN}-\pi_{CN}$. Because it can also be shown that $\pi_{CC}-\pi_{NN}>2(\pi_{NN}-\pi_{CN})$, the

probability of experimentation will usually be lower in the collusive state than in the Nash state. It even follows from (1) that the probability of one of the firms experimenting in the collusive state will usually be lower than the probability of both firms experimenting simultaneously in the Nash state. For that reason, transitions from the collusive state to the Nash state are less likely than transitions the other way around. As a consequence, the firms will spend most of their time in the collusive state.

The above informal argument provides the basic intuition why Q-learning firms may learn to collude with each other. There turn out to be two opposing forces at work. One force is due to firms trying to optimize their behavior in a given situation. This force is directed towards the Nash equilibrium. The other force is due to the possibility that firms experiment simultaneously and in that way discover the advantages of collusion. This force is directed towards collusive behavior. Somewhat surprisingly, the second force is typically stronger than the first one. This then leads to firms spending most of their time colluding with each other.

It is interesting to note that the argument depends crucially on the specific way in which the probability of experimentation is determined in Q-learning. In a setting with only two actions, the probability of experimentation depends on the difference between the actions' Q-values. The larger this difference, the lower the probability of experimentation. In many other learning models, for example in evolutionary models (e.g., Vega-Redondo, 1997) and in genetic algorithm models (e.g., Vriend, 2000), the probability of experimentation (sometimes referred to as the probability of mutation) has a fixed value and does not depend on past experience. With a fixed probability of experimentation, the informal argument provided above no longer holds. (This is because with a fixed probability of experimentation the probability of one of the firms experimenting in the collusive state is higher than the probability of both firms experimenting simultaneously in the Nash state.) Computer simulations indicate that a fixed probability of experimentation typically leads to firms spending most of their time in the Nash state rather than in the collusive state (see also Waltman and Kaymak, 2007). Apparently, the way in which the probability of experimentation is determined can have a large effect on the learning behavior of economic agents. According to Brenner (2006), psychological research indicates that people take into account past experience when choosing their actions. A learning model like Q-learning, in which the probability of experimentation depends on past experience, therefore seems more in line with psychological findings than learning models with a fixed probability of experimentation.

Finally, we note that the informal argument for the emergence of collusive behavior that we provided above holds not only for two firms but for any number of firms. Although we do not have a formal proof, this suggests that the emergence of collusive behavior is always possible, regardless of the number of firms. This would be a somewhat counterintuitive result, since collusion is generally believed to become much more difficult, if not impossible, when the number of firms increases. However, it would be in line with the simulation results discussed in Section 6, which indicate a substantial degree of collusion for various numbers of firms.

8. Conclusions

We have studied the use of Q-learning for modeling the learning behavior of firms in repeated Cournot oligopoly games. Q-learning, which belongs to the family of

reinforcement learning models, combines two elements that, individually, can also be found in other models of the reinforcement learning type. On the one hand, the way in which the performance of a strategy is measured is similar to the way in which this is done in the models of Sarin and Vahid (1999, 2001) and Kirman and Vriend (2001). On the other hand, the use of a logit model to describe an agent's choice behavior is fairly common and can also be found in the models of, for example, Mookherjee and Sopher (1997) and Camerer and Ho (1999). Q-learning combines both elements in a single model.

Based on computer simulations, we have shown that Q-learning firms generally learn to collude with each other in Cournot oligopoly games, although full collusion usually does not emerge, that is, firms usually do not learn to make the highest possible joint profit. Interestingly, our results hold not only for firms with a memory but also for firms without a memory. The latter firms do not have the ability to remember the quantities produced by their competitors in past periods. Although these firms cannot use trigger strategies to sustain collusion, they still learn to collude with each other. Apart from Q-learning, there are very few learning models that predict collusive behavior among firms without a memory. The analytical results that we have obtained for Cournot duopoly games with two production levels provide some insight into why Q-learning firms may learn to collude with each other. The emergence of collusive behavior seems to depend crucially on the specific way in which the probability of experimentation is determined in Q-learning. More specifically, it seems crucial that in Q-learning the probability of experimentation does not have a fixed value, as is the case in many other learning models, but depends on an agent's past experience.

Whether Q-learning provides a good description of the learning behavior of economic agents is, of course, an empirical question. We have not considered this question in the present paper. However, there is at least some correspondence between Q-learning behavior in Cournot oligopoly games and results from laboratory experiments (for an overview, see Huck et al., 2004b). Experimental results indicate that collusive behavior is quite common in Cournot duopoly games. Unlike most other learning models, Q-learning does indeed predict collusive behavior in these games. However, Q-learning also predicts a substantial degree of collusion in Cournot games with more than two firms. This does not match experimental results. In experimental studies, firm behavior usually turns out to be quite close to the Nash equilibrium when the number of firms is larger than two.

Acknowledgments

We would like to thank Maarten Janssen, Joost van Rosmalen, three anonymous referees, the associate editor, and the editor for their comments. These comments have significantly improved the paper.

Appendix A. Proof of Theorem 1

In this appendix, we provide a proof of Theorem 1. As a shorthand expression, below we sometimes write that the firms produce, for example, (q_C, q_N). With this we mean that one firm, firm 1, produces q_C while the other firm, firm 2, produces q_N.

Proof of Theorem 1. It follows from (7) and (11) that

$$\pi_{CN} = (-3vq_C + 2u - 2w)q_C/3, \tag{A.1}$$

$$\pi_{NN} = (u - w)^2/9v, \tag{A.2}$$

$$\pi_{CC} = (-2vq_C + u - w)q_C, \tag{A.3}$$

$$\pi_{NC} = (u - w)(-3vq_C + 2u - 2w)/9v, \tag{A.4}$$

and that

$$\pi_{CN} < \pi_{NN} < \pi_{CC} < \pi_{NC}. \tag{A.5}$$

For $i = 1, 2$ and $t = 0, 1, \ldots$, let $Q^C_{i,t}$ and $Q^N_{i,t}$ denote, respectively, firm i's Q-value of producing q_C in period t and firm i's Q-value of producing q_N in period t. The theorem assumes that $Q^C_{1,0}, Q^C_{2,0} \in (\pi_{CN}, \pi_{CC})$ and $Q^N_{1,0}, Q^N_{2,0} \in (\pi_{NN}, \pi_{NC})$. It then follows from (4) that $Q^C_{1,t}, Q^C_{2,t} \in (\pi_{CN}, \pi_{CC})$ and $Q^N_{1,t}, Q^N_{2,t} \in (\pi_{NN}, \pi_{NC})$ also holds for $t = 1, 2, \ldots$. Firm i is said to experiment in period t if it produces q_C while $Q^C_{i,t} < Q^N_{i,t}$ or if it produces q_N while $Q^N_{i,t} < Q^C_{i,t}$. For $t = 0, 1, \ldots$, let $X_t \in \{0, 1, 2, 3\}$ denote the state of the learning process in period t. Consider the following three conditions on the firms' Q-values:

$$Q^C_{1,t}, Q^C_{2,t} \leqslant \pi_{NN}, \tag{A.6}$$

$$Q^N_{1,t}, Q^N_{2,t} < \pi_{NN} + \varepsilon < \pi_{CC} - \varepsilon < Q^C_{1,t}, Q^C_{2,t}, \tag{A.7}$$

$$\max(Q^N_{1,t}, \pi_{CC} - \varepsilon) < Q^C_{1,t} \quad \text{and} \quad \max(Q^N_{2,t}, \pi_{CC} - \varepsilon) < Q^C_{2,t}$$
$$\text{and} \quad \max(Q^N_{1,t}, Q^N_{2,t}) \geqslant \pi_{NN} + \varepsilon. \tag{A.8}$$

In these conditions, ε denotes a constant that satisfies

$$0 < \varepsilon < \min((2\pi_{CN} - 3\pi_{NN} + \pi_{CC})/4, (1 - \alpha)\pi_{CN} - \pi_{NN} + \alpha\pi_{CC}), \tag{A.9}$$

where the positivity of the first argument of the min function follows from (11), (A.1)–(A.3) and the positivity of the second argument of the min function follows from (12) and (A.5). Let the state of the learning process be determined by the above three conditions in the following way: $X_t = 1$ if (A.6) is satisfied, $X_t = 2$ if (A.7) is satisfied, $X_t = 3$ if (A.8) is satisfied, and $X_t = 0$ otherwise.

To prove the theorem, three properties of the learning process will be used. Each property is proven separately below.

Property 1. *As the experimentation tendency β approaches zero, the proportion of time in which the learning process is in state* 0 *approaches zero.*

Due to this property, state 0 need not be considered further. Consequently, a transition from state j to state k, where $j, k \in \{1, 2, 3\}$ and $j \neq k$, is said to occur between periods t and t', where $t < t'$, if $X_t = j$, $X_{t+1} = \cdots = X_{t'-1} = 0$, and $X_{t'} = k$.

Property 2. *As the experimentation tendency β approaches zero, the probability that a transition from state* 1 *leads to state* 2 *approaches one.*

Property 3. *Consider the ratio between the average time it takes in state* 1 *before a transition to another state occurs and the average time it takes in state* 2 *before a transition to another state occurs. This ratio approaches zero as the experimentation tendency* β *approaches zero.*

It follows from the last two properties that the ratio between the time in which the learning process is in state 1 and the time in which the learning process is in state 2 approaches zero as β approaches zero. Together with Property 1, this implies that as β approaches zero the proportion of time in which the learning process is in state 0 or 1 approaches zero. If the learning process is not in state 0 or 1, it will be in state 2 or 3. If the learning process is in one of the latter states, the probability that the firms produce (q_C, q_C) approaches one as β approaches zero. It follows that as β approaches zero the proportion of time in which the firms produce (q_C, q_C) approaches one. □

Proof of Property 1. It will first be shown that if the learning process is in state 0 and the firms never experiment, the learning process will leave state 0 within a finite number of periods. Under the assumption that the firms never experiment, the following three observations can be made:

(i) *As long as the learning process is in state* 0, *the firms will not keep producing* (q_N, q_N) *forever*.
This can be seen as follows. Each firm's Q-value of producing q_N will decrease if the firms produce (q_N, q_N). Moreover, as long as the learning process is in state 0, at least one firm, say firm 1, will have a Q-value of producing q_C that is larger than π_{NN} (otherwise the learning process would be in state 1). If the firms produce (q_N, q_N) for a certain finite number of consecutive periods while the learning process is in state 0, firm 1's Q-value of producing q_N will become smaller than its Q-value of producing q_C. Due to the assumption of no experimentation, firm 1 will then produce q_C in the next period.

(ii) *If the learning process is in state* 0 *and the firms produce* (q_C, q_C), *the learning process will leave state* 0 *within a finite number of periods*.
This can be seen as follows. Due to the assumption of no experimentation, in some period t the firms will produce (q_C, q_C) only if $Q^C_{1,t} \geqslant Q^N_{1,t}$ and $Q^C_{2,t} \geqslant Q^N_{2,t}$. If the firms produce (q_C, q_C) in period t, it follows that $Q^C_{1,t+1} > Q^N_{1,t+1}$ and $Q^C_{2,t+1} > Q^N_{2,t+1}$. As a consequence, the firms will produce (q_C, q_C) another time in period $t+1$ and will in fact keep producing it forever. After a finite number of periods, each firm's Q-value of producing q_C will then be larger than $\pi_{CC} - \varepsilon$ and the learning process will have reached state 2 or 3.

(iii) *Regardless of the state of the learning process*, *the firms will produce* (q_C, q_N) *and* (q_N, q_C) *at most a finite number of times*.
To see this, consider one of the two quantity pairs, say (q_C, q_N). Firm 1's Q-value of producing q_C will decrease if the firms produce (q_C, q_N) and increase if the firms produce (q_C, q_C). If firm 1's Q-value of producing q_C decreases a certain finite number of times, denoted by m, and does not increase in between, it will no longer be larger than π_{NN}. It can be seen that the firms will produce (q_C, q_N) at most m times. Two cases have to be distinguished. In the first case, the firms produce (q_C, q_C) before they have produced (q_C, q_N) m times. As shown above, due to the assumption of no

experimentation, the firms will then keep producing (q_C, q_C) forever. In the second case, the firms produce (q_C, q_N) m times and do not produce (q_C, q_C) in between. After the firms have produced (q_C, q_N) m times, firm 1's Q-value of producing q_C will no longer be larger than π_{NN}. Firm 1 will then keep producing q_N forever. This is because firm 1's Q-value of producing q_N is always larger than π_{NN} and because, by assumption, firm 1 will never experiment. In both the first and the second case, the firms produce (q_C, q_N) no more than m times.

So, under the assumption of no experimentation, as long as the learning process is in state 0, the firms will not keep producing (q_N, q_N) forever and they will produce the other three quantity pairs at most a finite number of times. It follows that if the learning process is in state 0 and the firms never experiment, the learning process will leave state 0 within a finite number of periods. The other three states do not have this property. If the learning process is in state 1 and the firms never experiment, the firms will keep producing (q_N, q_N) forever and the learning process will never leave its current state. Similarly, if the learning process is in state 2 or 3 and the firms never experiment, the firms will keep producing (q_C, q_C) forever. Again, the learning process will never leave its current state.

As β approaches zero, the probability that a firm experiments approaches zero. Using the results obtained above, it can be seen that if the probability of experimentation approaches zero and the learning process is in state 0, the probability that within a finite number of periods another state is reached approaches one. Similarly, it can be seen that if the probability of experimentation approaches zero and the learning process is in state 1, 2, or 3, the probability that within a finite number of periods another state is reached approaches zero. It follows that as β approaches zero, the proportion of time in which the learning process is in state 0 approaches zero. □

Proof of Property 2. When the learning process is in state 1, two cases can be distinguished. If $Q^N_{1,t}, Q^N_{2,t} < \pi_{NN} + \varepsilon$, the learning process is said to be in state 1a, otherwise it is said to be in state 1b. The following three observations can now be made:

(i) *If the learning process is in state* 1, *it will leave that state only if the firms produce* (q_C, q_C).
This is because in order to leave state 1, for at least one of the firms the Q-value of producing q_C must increase. Producing (q_C, q_C) is the only way in which a firm's Q-value of producing q_C can increase.

(ii) *If the learning process is in state* 1*a and the firms produce* (q_C, q_C), *the probability that a transition to state* 2 *occurs approaches one as* β *approaches zero*.
This can be seen as follows. Due to (A.9), $\varepsilon < (1-\alpha)\pi_{CN} - \pi_{NN} + \alpha\pi_{CC}$. Therefore, if in some period t learning process is in state 1a and the firms produce (q_C, q_C), it follows that $Q^C_{1,t+1}, Q^C_{2,t+1} > \pi_{NN} + \varepsilon$ and hence that $Q^C_{1,t+1} > Q^N_{1,t+1}$ and $Q^C_{2,t+1} > Q^N_{2,t+1}$. Consequently, the probability that the firms produce (q_C, q_C) another time in period $t+1$ approaches one as β approaches zero. As long as the firms keep producing (q_C, q_C), the probability that they produce it again in the next period approaches one as β approaches zero. After the firms have produced (q_C, q_C) for some finite number of consecutive periods, each firm's Q-value of producing q_C will be larger than $\pi_{CC} - \varepsilon$ and the learning process will have reached state 2.

(iii) *Suppose that in some period t the learning process is in state* 1 *and the firms produce* (q_C, q_C). *The probability that the learning process was in state* 1*a in period* t *then approaches one as* β *approaches zero.*
This can be seen as follows. Let Δt denote the number of consecutive periods in which (q_N, q_N) must be produced so that a firm's Q-value of producing q_N decreases from π_{NC} to a value smaller than $\pi_{NN} + \varepsilon$. Let $t' = t - \Delta t$. For the moment, assume that the learning process has been in state 1 all the time between periods t' and t. This assumption implies that the firms have never produced (q_C, q_C) between periods t' and $t-1$. If the firms have produced (q_N, q_N) all the time between periods t' and $t-1$, the learning process would have been in state 1a in period t. If, on the other hand, the firms have produced either (q_C, q_N) or (q_N, q_C) at least once between periods t' and $t-1$, the learning process could have been in either state 1a or 1b in period t. Given that the learning process was in state 1 in period t', the probability that the firms have produced (q_N, q_N) all the time between periods t' and $t-1$ approaches one as β approaches zero. Furthermore, given the firms' Q-values in period t', the probability that the firms produce (q_C, q_C) in period t is higher if the firms have produced (q_N, q_N) all the time between periods t' and $t-1$ than if they have produced either (q_C, q_N) or (q_N, q_C) at least once between these periods. It now follows that the probability that the learning process was in state 1a in period t approaches one as β approaches zero. This result relies on the assumption that the learning process has been in state 1 all the time between periods t' and t. Since producing (q_C, q_C) in state 1 requires experimentation and since $t - t'$ is a finite number, it can be seen that the probability that this assumption is true approaches one as β approaches zero. Consequently, the result also holds without making the assumption.

As a consequence of the above three observations, the probability that a transition from state 1 leads to state 2 approaches one as β approaches zero. □

Proof of Property 3. First consider state 1. When the learning process is in this state, two cases can be distinguished. If $Q^N_{1,t}, Q^N_{2,t} < \pi_{NN} + \varepsilon$, the learning process is said to be in state 1a, otherwise it is said to be in state 1b. If in some period t the learning process is in state 1a, it will be in state 1b in period $t+1$ only if the firms produce either (q_C, q_N) or (q_N, q_C) in period t. The probability that this happens approaches zero as β approaches zero. If in some period t the learning process is in state 1b, it will reach state 1a if from period t onwards the firms produce (q_N, q_N) for some finite number of consecutive periods. The probability that this happens approaches one as β approaches zero. So, in the limit as β approaches zero, it takes an infinite number of periods to reach state 1b from state 1a, whereas it takes a finite number of periods to reach state 1a from state 1b. Furthermore, if the learning process is in state 1, the number of periods it takes to leave that state approaches infinity as β approaches zero. It now follows that the conditional probability that the learning process is in state 1a given that it is in state 1 approaches one as β approaches zero. If in some period t the learning process is in state 1a, the probability that the firms produce (q_C, q_C) approaches zero as β approaches zero and is of order $\exp(-(Q^N_{1,t} - Q^C_{1,t} + Q^N_{2,t} - Q^C_{2,t})/\beta)$. (To see this, note that the firms choose their production levels independently according to probabilities given by (3).) A lower bound for this order is $\exp(-2(\pi_{NN} - \pi_{CN} + \varepsilon)/\beta)$. Furthermore, if the learning process is in state 1a and the firms produce (q_C, q_C), the probability that a transition from state 1 to

another state occurs approaches one as β approaches zero. (This has been shown in the proof of Property 2.)

Now consider state 2. If in some period t the learning process is in this state, a transition to another state can occur only if the firms produce either (q_C, q_N) or (q_N, q_C). The probability that this happens approaches zero as β approaches zero and is of order $\exp(-\min(Q^C_{1,t} - Q^N_{1,t}, Q^C_{2,t} - Q^N_{2,t})/\beta)$. An upper bound for this order is $\exp(-(\pi_{CC} - \pi_{NN} - 2\varepsilon)/\beta)$.

In summary, if the learning process is in state 1, then in the limit as β approaches zero the rate at which the probability of a state transition approaches zero equals, with probability one, at most $2(\pi_{NN} - \pi_{CN} + \varepsilon)$. Furthermore, if the learning process is in state 2, the rate at which the probability of a state transition approaches zero equals at least $\pi_{CC} - \pi_{NN} - 2\varepsilon$. From $\varepsilon < (2\pi_{CN} - 3\pi_{NN} + \pi_{CC})/4$, which is due to (A.9), it follows that $2(\pi_{NN} - \pi_{CN} + \varepsilon) < \pi_{CC} - \pi_{NN} - 2\varepsilon$. Consequently, the ratio between the average time it takes in state 1 before a state transition occurs and the average time it takes in state 2 before a state transition occurs approaches zero as β approaches zero. □

References

Alós-Ferrer, C., 2004. Cournot versus Walras in dynamic oligopolies with memory. International Journal of Industrial Organization 22, 193–217.

Arifovic, J., 1994. Genetic algorithm learning and the cobweb model. Journal of Economic Dynamics and Control 18, 3–28.

Bell, A.M., 2001. Reinforcement learning rules in a repeated game. Computational Economics 18, 89–111.

Bergin, J., Bernhardt, D., 2005. Cooperation through imitation. Working Paper 1042, Queen's Economics Department.

Brenner, T., 2006. Agent learning representation: advice on modelling economic learning. In: Tesfatsion, L., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 2. Elsevier, Amsterdam, pp. 895–947.

Brock, W.A., Hommes, C.H., 1997. A rational route to randomness. Econometrica 65, 1059–1095.

Camerer, C., Ho, T.-H., 1999. Experience-weighted attraction learning in normal form games. Econometrica 67, 827–873.

Dixon, H.D., 2000. Keeping up with the Joneses: competition and the evolution of collusion. Journal of Economic Behavior and Organization 43, 223–238.

Droste, E., Hommes, C., Tuinstra, J., 2002. Endogenous fluctuations under evolutionary pressure in Cournot competition. Games and Economic Behavior 40, 232–269.

Duffy, J., 2006. Agent-based models and human subject experiments. In: Tesfatsion, L., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 2. Elsevier, Amsterdam, pp. 949–1011.

Erev, I., Roth, A.E., 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88, 848–881.

Fudenberg, D., Levine, D.K., 1998. The Theory of Learning in Games. MIT Press, Cambridge, MA.

Hofbauer, J., Sandholm, W.H., 2007. Evolution in games with randomly disturbed payoffs. Journal of Economic Theory 132, 47–69.

Huck, S., Normann, H.-T., Oechssler, J., 2004a. Through trial and error to collusion. International Economic Review 45, 205–224.

Huck, S., Normann, H.-T., Oechssler, J., 2004b. Two are few and four are many: number effects in experimental oligopolies. Journal of Economic Behavior and Organization 53, 435–446.

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

Kimbrough, S.O., Lu, M., 2003. A note on Q-learning in the Cournot game. In: Proceedings of the Second Workshop on E-Business.

Kirman, A.P., Vriend, N.J., 2001. Evolving market structure: an ACE model of price dispersion and loyalty. Journal of Economic Dynamics and Control 25, 459–502.

Littman, M.L., Stone, P., 2001. Leading best-response strategies in repeated games. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence.

McKelvey, R.D., Palfrey, T.R., 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10, 6–38.

Mookherjee, D., Sopher, B., 1997. Learning and decision costs in experimental constant sum games. Games and Economic Behavior 19, 97–132.

Oechssler, J., 2002. Cooperation as a result of learning with aspiration levels. Journal of Economic Behavior and Organization 49, 405–409.

Roth, A.E., Erev, I., 1995. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. Games and Economic Behavior 8, 164–212.

Sandholm, T.W., Crites, R.H., 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems 37, 147–166.

Sarin, R., Vahid, F., 1999. Payoff assessments without probabilities: a simple dynamic model of choice. Games and Economic Behavior 28, 294–309.

Sarin, R., Vahid, F., 2001. Predicting how people play games: a simple dynamic model of choice. Games and Economic Behavior 34, 104–122.

Stimpson, J.L., Goodrich, M.A., 2003. Learning to cooperate in a social dilemma: a satisficing approach to bargaining. In: Proceedings of the Twentieth International Conference on Machine Learning.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Tesfatsion, L., 2003. Agent-based computational economics. ISU Economics Working Paper 1, Iowa State University.

Tesfatsion, L., 2006. Agent-based computational economics: a constructive approach to economic theory. In: Tesfatsion, L., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 2. Elsevier, Amsterdam, pp. 831–880.

Vega-Redondo, F., 1997. The evolution of Walrasian behavior. Econometrica 65, 375–384.

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, 1–19.

Waltman, L., Kaymak, U., 2007. A theoretical analysis of cooperative behavior in multi-agent Q-learning. In: Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning, pp. 84–91.

Watkins, C.J.C.H., 1989. Learning from delayed rewards. Ph.D. Thesis, University of Cambridge, England, 1989.

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