Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2007)

A Theoretical Analysis of Cooperative Behavior in Multi-agent Q-learning

Ludo Waltman and Uzay Kaymak
Erasmus School of Economics, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
Email: {lwaltman,kaymak}@few.eur.nl

***Abstract*— A number of experimental studies have investigated whether cooperative behavior may emerge in multi-agent Q-learning. In some studies cooperative behavior did emerge, in others it did not. This paper provides a theoretical analysis of this issue. The analysis focuses on multi-agent Q-learning in iterated prisoner's dilemmas. It is shown that under certain assumptions cooperative behavior may emerge when multi-agent Q-learning is applied in an iterated prisoner's dilemma. An important consequence of the analysis is that multi-agent Q-learning may result in non-Nash behavior. It is found experimentally that the theoretical results presented in this paper are quite robust to violations of the underlying assumptions.**

I. Introduction

Q-learning [1] is an algorithm for learning how to behave in an unknown environment. The use of Q-learning is most appropriate if the environment is a Markov decision process. For such an environment it has been proven that under certain assumptions Q-learning leads to optimal behavior [2].

In this paper we are concerned with multi-agent Q-learning. In multi-agent Q-learning there are a number of agents that all use Q-learning to choose their actions. The payoff received by an individual agent depends not only on the agent's own behavior but also on the behavior of the other agents. An individual agent in multi-agent Q-learning perceives its environment as non-stationary because the behavior of the other agents changes over time due to learning. Since a non-stationary environment does not have the properties of a Markov decision process, it is quite difficult to analyze multi-agent Q-learning theoretically. In this paper multi-agent Q-learning is analyzed theoretically under certain simplifying assumptions. The analysis is performed using Markov chain theory. It should be noted that attention focuses on the application of the standard Q-learning algorithm in multi-agent settings. Some examples of this approach can be found in [3]–[9]. Variants of the standard Q-learning algorithm specifically designed for multi-agent settings (e.g. [10]–[15]) are not considered.

The theoretical analysis in this paper aims to explain the emergence of cooperative behavior in multi-agent Q-learning. The analysis is closely related to our earlier research described in [8], [9]. In [8], [9] we apply multi-agent Q-learning in a Cournot oligopoly model, which is a well-known model in the field of microeconomics. We show experimentally that agents in a Cournot oligopoly model learn to cooperate with each other by using Q-learning (although they usually do not learn to cooperate to the largest extent possible). Somewhat similar experiments that are described in [16] give the same result. The purpose of this paper is to provide a theoretical analysis that explains results like in [8], [9], [16].

TABLE I

The Payoffs in a One-shot Prisoner's Dilemma Game. The Payoff to the Agent Using the Row (Column) Strategy Is Listed First (Second). The Payoffs Must Satisfy the Inequalities in (1) and (2)

	cooperate	*defect*
cooperate	w, w	u, z
defect	z, u	v, v

Our analysis focuses on multi-agent Q-learning in iterated prisoner's dilemmas. An iterated prisoner's dilemma has similar characteristics as a repeated Cournot oligopoly game but is simpler to analyze. In an iterated prisoner's dilemma there are two agents playing a repeated game. In each period of the repeated game the agents play a one-shot prisoner's dilemma game. The payoffs in a one-shot prisoner's dilemma game are shown in Table I, where u, v, w, and z must satisfy

$$u < v < w < z \tag{1}$$

and

$$2v < z + u < 2w. \tag{2}$$

In a one-shot prisoner's dilemma game *defect* is a dominant strategy for both agents. *(defect, defect)* is therefore a dominant strategy equilibrium. It is also the only Nash equilibrium in the game. The interesting property of the game is that *(defect, defect)* is Pareto dominated by *(cooperate, cooperate)*. In other words, if both agents use the dominant *defect* strategy, they both receive a lower payoff than they would have received if they had both used the dominated *cooperate* strategy.

One of the assumptions we make in our analysis is that agents in an iterated prisoner's dilemma do not have a memory for remembering what happened in the past (of course they do have a memory for remembering estimates of Q values). An agent therefore does not know, for example, which action was executed by its opponent in the previous period of the game. Making this assumption simplifies our analysis considerably. The assumption also has the important implication that there is only one Nash equilibrium in an iterated prisoner's dilemma.

In this equilibrium both agents choose the *defect* action in each period of the game. Other, more cooperative Nash equilibria do not exist because agents do not remember what happened in the past and consequently do not have the ability to punish their opponent in case of defection. Since there are no cooperative Nash equilibria under the assumption that agents do not have a memory, it turns out that by making this assumption we focus our attention on the setting in which cooperative behavior seems most difficult to achieve.

Multi-agent Q-learning in an iterated prisoner's dilemma is studied experimentally in [5]. In one of the experiments in [5] agents without a memory are considered. It turns out that these agents always learn the *defect* strategy, which means that their behavior always converges to the Nash equilibrium. Cooperative behavior between the agents never emerges. A similar result is reported in [4]. The results in [4], [5] may not seem to be very surprising. Using the following reasoning, which can also be found in [17], it may be argued that if multi-agent Q-learning converges it must converge to a (pure strategy) Nash equilibrium. When multi-agent Q-learning has converged, agents no longer perceive their environment as non-stationary, since non-stationarity is caused by learning. Q-learning in a stationary environment is guaranteed to converge to optimal behavior under certain assumptions [2]. In a multi-agent setting optimal behavior means that an agent gives a best response to the strategies of its opponents. Therefore, when multi-agent Q-learning has converged, it must be the case that each agent's strategy has converged to a best response to the opponent strategies and, as a consequence, that a Nash equilibrium has been reached.

This reasoning may seem to explain the non-cooperative behavior reported in [4], [5]. However, it also seems that according to this reasoning the emergence of cooperative behavior (i.e. non-Nash behavior) in the experiments described in [8], [9] should not have been possible. In this paper we investigate in detail the possibility that cooperative behavior emerges in multi-agent Q-learning. We focus our attention on cooperative behavior in iterated prisoner's dilemmas. Using a specific definition of convergence we show that it is possible for multi-agent Q-learning to converge to cooperative behavior. Whether convergence to cooperative behavior takes place turns out to depend on the exploration strategy that is used in the Q-learning algorithm and on the values of the payoffs in the one-shot prisoner's dilemma game. As we already noted above, the theoretical analysis that we provide is based on certain simplifying assumptions. In this paper we also present the results of a number of experiments in which these assumptions were relaxed. The experimental results indicate that the results of our theoretical analysis are quite robust to violations of the underlying assumptions.

Before we present our analysis some other papers in which cooperative behavior in multi-agent Q-learning is studied should be mentioned. In these papers a number of experimental results are reported. In [18] a game similar to a prisoner's dilemma is considered. It is found that cooperative behavior almost never emerges in this game. The agents in [18] do not use the standard Q-learning algorithm but use a variant of this algorithm specifically designed for multi-agent settings. In [3] an experiment is described in which Q-learning agents with a memory learn to cooperate with each other most of the time. However, the game that is studied in [3] does not have all the characteristics of a prisoner's dilemma. In [6] a game is studied that is a generalization of a prisoner's dilemma. It is found that in this game non-cooperative behavior emerges most of the time although Q-learning agents also occasionally learn to cooperate. Finally, in [11] a Q-learning variant for multi-agent settings is presented that allows agents to learn to cooperate in a prisoner's dilemma, at least for certain values of the payoffs.

This paper is organized as follows. We first give an introduction to Q-learning in Section II. We then present a theoretical analysis of cooperative behavior in multi-agent Q-learning in Section III. The robustness of our theoretical results is tested experimentally in Section IV. Finally a discussion is provided in Section V.

II. Q-LEARNING

In this section we give an introduction to Q-learning. The terminology that we use is somewhat adapted to the terminology in the game theoretic literature. Instead of the terms 'reward' and 'policy', which are typically used in the literature on Q-learning, we use the terms 'payoff' and 'strategy'.

Q-learning is an algorithm that belongs to the class of reinforcement learning algorithms [19], [20]. Reinforcement learning is concerned with the problem how an agent can learn to behave optimally from interactions with its environment. The general idea of reinforcement learning is as follows. An agent interacts repeatedly with its environment. During each interaction the agent first observes the state of the environment $s \in S$. The agent then decides to execute an action $a \in A$. This results in a payoff r that is received by the agent and in a transition of the state of the environment from the old state s to the new state s'. Because the state of the environment changes as a result of the action that was executed by the agent, the choice of an action may not only influence the agent's immediate payoff r but also its payoffs in future periods. The environment is usually assumed to be a Markov decision process, which means that the agent's payoff and the new state of the environment only depend (either deterministically or stochastically) on the old state of the environment and on the action that was executed by the agent, i.e. $r = r(s, a)$ and $s' = \delta(s, a)$. In reinforcement learning it is typically assumed that the agent has no prior knowledge of the payoff function $r(s, a)$ and the state transition function $\delta(s, a)$, so the agent has no model of its environment. The goal of the agent is to find an optimal strategy $\pi^* : S \rightarrow A$ for choosing actions. A strategy $\pi(s)$ is optimal if in each state $s \in S$ it selects an action $a \in A$ that maximizes the agent's cumulative payoff, which is the sum of its immediate payoff and its future payoffs. The future payoffs are usually discounted.

Q-learning, introduced in [1], finds an optimal strategy by learning the values of a so-called Q function. The function

$Q(s,a)$ is defined as the expected discounted cumulative payoff that is received by executing action a in state s and following an optimal strategy thereafter. Recursively the Q function can be defined as follows

$$Q(s,a) = \mathrm{E}\left(r(s,a) + \gamma \max_{a' \in A} Q(\delta(s,a), a')\right) \quad (3)$$

where $0 \leq \gamma < 1$ denotes the discount factor. If the values of the Q function are known, an optimal strategy is given by

$$\pi^*(s) = \operatorname*{argmax}_{a \in A} Q(s,a). \quad (4)$$

Q-learning approximates the Q values iteratively. After each interaction of an agent with its environment, the agent's estimated Q values, denoted by $\hat{Q}$, are updated using the update rule

$$\hat{Q}(s,a) \leftarrow (1-\alpha)\hat{Q}(s,a) + \alpha\left(r + \gamma \max_{a' \in A} \hat{Q}(s',a')\right). \quad (5)$$

The parameter $0 \leq \alpha < 1$ is called the learning rate and may be decreased over time. The update rule allows an agent that does not know the functions $r(s,a)$ and $\delta(s,a)$ to learn the values of the Q function and, consequently, to find an optimal strategy for choosing actions. It is proven in [2] that the values $\hat{Q}(s,a)$ estimated using Q-learning converge to the correct values $Q(s,a)$ with probability 1 if the environment is a Markov decision process, the values $\hat{Q}(s,a)$ are stored in a lookup table, all state-action pairs continue to be visited, and the learning rate α is decreased in an appropriate way. In deterministic Markov decision processes convergence of $\hat{Q}$ values can also be proven if a fixed value is used for α [21].

In the special case that the environment is a Markov decision process that has only one state, the action that is executed in the current period cannot influence the payoffs that are received in future periods. This means that future payoffs need not be considered in the Q function. The discount factor γ can therefore be set to 0 in (3) and (5). Denoting the Q function and the payoff function by $Q(a)$ and $r(a)$ respectively, (3) then reduces to

$$Q(a) = \mathrm{E}(r(a)). \quad (6)$$

Therefore, in the case of an environment with only one state, the Q value of an action is simply defined as the expected payoff from that action. In the same way, the update rule in (5) reduces to

$$\hat{Q}(a) \leftarrow (1-\alpha)\hat{Q}(a) + \alpha r. \quad (7)$$

In this paper we refer to Q-learning using this update rule as 'single-state Q-learning'.

In reinforcement learning an agent usually faces a trade-off between exploration and exploitation when choosing an action. On the one hand, an agent may want to explore unknown states and actions to collect new information about its environment. On the other hand, an agent may want to exploit its current knowledge of the environment by executing the action that is expected to maximize the cumulative payoff. In this paper we consider agents that use either the Boltzmann strategy or the ϵ-greedy strategy for choosing between exploration and exploitation.

In the Boltzmann strategy the probability of selecting action a in state s is given by

$$\Pr(a|s) = \frac{\exp\left(\hat{Q}(s,a)/T\right)}{\sum_{a' \in A} \exp\left(\hat{Q}(s,a')/T\right)}. \quad (8)$$

In the case of single-state Q-learning this reduces to

$$\Pr(a) = \frac{\exp\left(\hat{Q}(a)/T\right)}{\sum_{a' \in A} \exp\left(\hat{Q}(a')/T\right)}. \quad (9)$$

Although the Boltzmann strategy favors actions with high $\hat{Q}$ values, all actions have a positive probability of being selected. The temperature parameter $T > 0$ determines how exploration and exploitation are balanced. The probability of exploration may be decreased over time by gradually decreasing T. As T approaches 0, the Boltzmann strategy approaches the greedy strategy of always selecting the action with the highest $\hat{Q}$ value.

In the ϵ-greedy strategy the action (or one of the actions) with the highest $\hat{Q}$ value is selected with probability $1-\epsilon$. With probability ϵ an action is selected randomly using a uniform distribution over all actions. The ϵ-greedy strategy is less sensitive to the exact $\hat{Q}$ values than the Boltzmann strategy. In the Boltzmann strategy the probability of exploration depends on the difference between the highest $\hat{Q}$ value and the other $\hat{Q}$ values. This is not the case in the ϵ-greedy strategy. Also, when exploration takes place, the Boltzmann strategy favors actions with higher $\hat{Q}$ values whereas the ϵ-greedy strategy gives equal probability to all actions.

III. Theoretical Analysis of Cooperative Behavior

In this section we present a theoretical analysis that aims to explain the emergence of cooperative behavior in multi-agent Q-learning. The analysis also provides insight into factors that influence whether cooperative behavior emerges. The focus of the analysis is on multi-agent Q-learning in iterated prisoner's dilemmas.

In the analysis the following assumption is made.

Assumption 1: Agents do not have a memory for remembering what happened in the past and therefore operate in an environment that has only one state.

Because of this assumption agents use single-state Q-learning to choose their actions. As we discussed in Section I, the above assumption implies that there is only one Nash equilibrium in an iterated prisoner's dilemma, namely mutual defection in each period of the game. Other, more cooperative Nash equilibria are ruled out by the assumption. By making Assumption 1 we therefore focus our attention on the setting in which cooperative behavior seems most difficult to achieve.

In addition to Assumption 1 the analysis in this section is also based on the following two assumptions.

Assumption 2: The learning rate α in the Q-learning algorithm has a fixed value of 1.

Assumption 3: Agents almost never explore, i.e. the limit case in which the probability of exploration approaches 0 is considered. If the Boltzmann strategy is used, this is achieved by considering the limit case in which the temperature parameter T approaches 0. If the ϵ-greedy strategy is used, this is achieved by considering the limit case in which ϵ approaches 0.

Assumption 3 states that agents *almost* never explore. It is important to note that the theoretical results presented in this section do not hold if there is no exploration at all. Assumption 2 and 3 match to a limited extent the way in which Q-learning is typically applied in experimental studies. Similarly to Assumption 2 many experimental studies use a fixed value for the learning rate α. Although in these studies α typically has a value that is less than 1, for example 0.2 in [5], [6], [16] and 0.5 in [8], [9], we assume a value of 1 in order to make a theoretical analysis feasible. We also assume that agents almost never explore. This is fairly similar to experimental studies in which the probability of exploration is quite high initially and is decreased over time in such a way that it approaches 0 (e.g. [3], [5], [9]). Although Assumption 2 and 3 are usually not completely satisfied in experimental studies, we do not consider this to be problematic for the purpose of our analysis, which is to explain the emergence of cooperative behavior in multi-agent Q-learning. First of all the analysis shows (see Theorem 1 below) that convergence of multi-agent Q-learning to cooperative behavior (i.e. non-Nash behavior) is not fundamentally impossible. This is an important result on its own. In addition, the results of a number of experiments that we performed indicate that our theoretical results are quite robust to violations of Assumption 2 and 3. We discuss the experiments in Section IV.

The notion of convergence of an agent's strategy is defined as follows in this paper.

Definition 1: Let π denote a pure strategy. In an environment that has only one state, the strategy of an agent is said to converge to strategy π if and only if

$$\lim_{t\to\infty} \Pr(a_t = \pi) = 1 \tag{10}$$

where a_t denotes the action that is executed by the agent in period t.

Note that this definition is restricted to pure strategies and to environments with only one state. This is sufficient for the purpose of this paper. Definition 1 corresponds closely to the way in which empirical convergence is typically established in experimental studies on Q-learning. In a multi-agent setting one may also be interested in the collective behavior of agents. The following definition can then be used.

Definition 2: Let $\pi_1, \ldots, \pi_n$ denote pure strategies. In an environment that has only one state and that is populated by n agents, the strategy profile of the agents is said to converge to strategy profile $(\pi_1, \ldots, \pi_n)$ if and only if for $k = 1, \ldots, n$ the strategy of agent k converges to strategy π_k according to Definition 1.

TABLE II
THE STATES OF THE MARKOV CHAINS THAT ARE USED IN THE PROOFS OF THEOREM 1 AND 2

Symbol	a_1	a_2	$\hat{Q}_1^{\mathrm{c}}$	$\hat{Q}_1^{\mathrm{d}}$	$\hat{Q}_2^{\mathrm{c}}$	$\hat{Q}_2^{\mathrm{d}}$
m_1	*cooperate*	*cooperate*	w	v	w	v
m_2	*cooperate*	*cooperate*	w	v	w	z
m_3	*cooperate*	*cooperate*	w	z	w	v
m_4	*cooperate*	*cooperate*	w	z	w	z
m_5	*cooperate*	*defect*	u	v	w	z
m_6	*cooperate*	*defect*	u	v	u	z
m_7	*cooperate*	*defect*	u	z	w	z
m_8	*cooperate*	*defect*	u	z	u	z
m_9	*defect*	*cooperate*	w	z	u	v
m_{10}	*defect*	*cooperate*	w	z	u	z
m_{11}	*defect*	*cooperate*	u	z	u	v
m_{12}	*defect*	*cooperate*	u	z	u	z
m_{13}	*defect*	*defect*	w	v	w	v
m_{14}	*defect*	*defect*	w	v	u	v
m_{15}	*defect*	*defect*	u	v	w	v
m_{16}	*defect*	*defect*	u	v	u	v

It should be emphasized that the convergence results presented in this section make use of the above two definitions and need not be valid for alternative definitions of convergence.

The following theorem states that cooperative behavior may emerge in an iterated prisoner's dilemma if the assumptions mentioned above are satisfied and agents use the Boltzmann strategy to choose their actions.

Theorem 1: Consider an iterated prisoner's dilemma with an infinite number of periods. Let Assumption 1, 2, and 3 be satisfied. Let both agents use single-state Q-learning combined with the Boltzmann strategy. The strategy profile of the agents then converges to *(cooperate, cooperate)* if and only if the payoffs in the one-shot prisoner's dilemma game satisfy

$$w - v > 2(v - u). \tag{11}$$

It converges to *(defect, defect)* if and only if the payoffs in the one-shot prisoner's dilemma game satisfy

$$w - v < 2(v - u). \tag{12}$$

Proof: For $k = 1, 2$ and $t = 1, 2, \ldots$ we use $a_{kt} \in \{\textit{cooperate}, \textit{defect}\}$ to denote the action that is executed by agent k in period t. We use $\hat{Q}_{kt}^{\mathrm{c}}$ and $\hat{Q}_{kt}^{\mathrm{d}}$ to denote respectively agent k's $\hat{Q}$ value of the *cooperate* action and agent k's $\hat{Q}$ value of the *defect* action immediately after the actions a_{1t} and a_{2t} have been executed and the update rule of Q-learning has been applied. Since in each period the *cooperate* action and the *defect* action both have a positive probability of being executed by an agent, after some time both actions will have been executed at least once by each agent. Let t' denote the first period in which this is the case. Using Assumption 2 it can be seen that for $k = 1, 2$ and $t = t', t' + 1, \ldots$ $\hat{Q}_{kt}^{\mathrm{c}} \in \{u, w\}$ and $\hat{Q}_{kt}^{\mathrm{d}} \in \{v, z\}$. We define $X_t = (a_{1t}, a_{2t}, \hat{Q}_{1t}^{\mathrm{c}}, \hat{Q}_{1t}^{\mathrm{d}}, \hat{Q}_{2t}^{\mathrm{c}}, \hat{Q}_{2t}^{\mathrm{d}})$. Note that X_t is a random variable. We use x_t to denote a particular value that X_t may take. $\{X_t | t = t', t' + 1, \ldots\}$ is a Markov chain because for $t = t', t' + 1, \ldots$ $\Pr(X_{t+1} = x_{t+1} | X_{t'} = x_{t'}, \ldots, X_t = x_t) = \Pr(X_{t+1} = x_{t+1} | X_t = x_t)$. The Markov chain has sixteen states, which we denote by

$m_1, \ldots, m_{16}$. These states are shown in Table II. From each state four state transitions are possible, which correspond to the four action profiles in the one-shot prisoner's dilemma game. Since agents independently choose their actions using the Boltzmann strategy, the probability of a state transition can be calculated by multiplying each agent's probability of choosing an action as given by (9). Transition probabilities are stationary because according to Assumption 3 the temperature parameter T has a fixed value. We use $p_{i\to j}$ to denote the transition probability from state m_i to state m_j, i.e. for $t = t', t'+1, \ldots$ $p_{i\to j} = \Pr(X_{t+1} = m_j | X_t = m_i)$. Obviously, transition probabilities satisfy

$$\sum_{j=1}^{16} p_{i\to j} = 1. \tag{13}$$

Because the Markov chain $\{X_t | t = t', t'+1, \ldots\}$ is irreducible and ergodic, $\lim_{t\to\infty} \Pr(X_t = m_j)$ exists and does not depend on the initial state $X_{t'}$ (e.g. [22]). $P_j = \lim_{t\to\infty} \Pr(X_t = m_j)$ is referred to as a stationary probability of the Markov chain. The stationary probabilities of the Markov chain can be found by solving

$$P_j = \sum_{i=1}^{16} P_i p_{i\to j}, \quad \text{for } j = 1, \ldots, 16 \tag{14}$$

and

$$\sum_{j=1}^{16} P_j = 1. \tag{15}$$

Equation (14) can be written as

$$P_j = \frac{1}{1 - p_{j\to j}} \sum_{\substack{i=1 \\ i\neq j}}^{16} P_i p_{i\to j}. \tag{16}$$

Assumption 3 states that $T \to 0^+$. Using (9) it can be seen that the probability of an agent choosing a particular action approaches either 0 or 1 as $T \to 0^+$. Since the probability of a state transition is calculated by multiplying each agent's probability of choosing a particular action, it follows that for each transition probability $p_{i\to j}$ either $\lim_{T\to 0^+} p_{i\to j} = 0$ or $\lim_{T\to 0^+} p_{i\to j} = 1$. Note especially that $\lim_{T\to 0^+} p_{j\to j} = 0$ for $j = 2, \ldots, 15$. Using (16) we now obtain for $j = 2, \ldots, 15$

$$\lim_{T\to 0^+} P_j = \lim_{T\to 0^+} \sum_{\substack{i=1 \\ i\neq j}}^{16} P_i p_{i\to j}. \tag{17}$$

It follows from (17) that $\lim_{T\to 0^+} P_j = 0$ for $j = 2, \ldots, 15$. This result can be derived in an incremental way, first for $j = 2, 3, 4, 7, 8, 10, 12$, then for $j = 5, 9, 13$, then for $j = 14, 15$, and finally for $j = 6, 11$. Equation (15) subsequently implies that $\lim_{T\to 0^+}(P_1 + P_{16}) = 1$.

Transitions to state m_1 are only possible from the states m_1, m_{13}, m_{14}, m_{15}, and m_{16}. From state m_1 transitions are only possible to the states m_1, m_5, m_9, and m_{13}. Using (13) and (16) we can therefore write

$$P_1 = \sum_{i=13}^{16} \frac{P_i p_{i\to 1}}{p_{1\to 5} + p_{1\to 9} + p_{1\to 13}}. \tag{18}$$

It turns out that for $i = 13, 14, 15$

$$\lim_{T\to 0^+} \frac{P_i p_{i\to 1}}{p_{1\to 5} + p_{1\to 9} + p_{1\to 13}} = 0. \tag{19}$$

The left-hand side of (19) has the form $0/0$. To prove (19) it must be shown that the numerator in the left-hand side approaches 0 faster than the denominator. This can be shown by substituting an upper bound for $\lim_{T\to 0^+} P_i$ into (19). We only work out a proof for $i = 13$. Since $\lim_{T\to 0^+} P_i \leq 1$ for all i, it follows from (17) that

$$\lim_{T\to 0^+} P_2 \leq \lim_{T\to 0^+} (p_{5\to 2} + p_{6\to 2}) \tag{20}$$

$$\lim_{T\to 0^+} P_3 \leq \lim_{T\to 0^+} (p_{9\to 3} + p_{11\to 3}) \tag{21}$$

and

$$\lim_{T\to 0^+} P_4 \leq \lim_{T\to 0^+} (p_{7\to 4} + p_{8\to 4} + p_{10\to 4} + p_{12\to 4}). \tag{22}$$

Using these inequalities and $\lim_{T\to 0^+} P_1 \leq 1$ the following upper bound for $\lim_{T\to 0^+} P_{13}$ results from (17)

$$\begin{aligned} \lim_{T\to 0^+} P_{13} \leq \lim_{T\to 0^+} (& p_{1\to 13} + (p_{5\to 2} + p_{6\to 2}) p_{2\to 13} \\ & + (p_{9\to 3} + p_{11\to 3}) p_{3\to 13} \\ & + (p_{7\to 4} + p_{8\to 4} + p_{10\to 4} + p_{12\to 4}) p_{4\to 13}). \end{aligned} \tag{23}$$

For $i = 13$ the correctness of (19) can be verified by substituting this upper bound into (19) and by taking into account the inequalities in (1). The proof of (19) for $i = 14, 15$ is similar to the proof for $i = 13$.

Using $\lim_{T\to 0^+}(P_1 + P_{16}) = 1$ it follows from (18) and (19) that

$$\lim_{T\to 0^+} P_1 = \lim_{T\to 0^+} \frac{p_{16\to 1}}{p_{16\to 1} + p_{1\to 5} + p_{1\to 9} + p_{1\to 13}}. \tag{24}$$

This expression has the form $0/0$. It can be seen that, depending on the payoffs u, v, and w, either the numerator in (24) approaches 0 faster than the denominator or the numerator and the denominator approach 0 equally fast. This results in

$$\lim_{T\to 0^+} P_1 = \begin{cases} 1, & \text{if } w - v > 2(v - u) \\ \frac{1}{3}, & \text{if } w - v = 2(v - u) \\ 0, & \text{if } w - v < 2(v - u). \end{cases} \tag{25}$$

According to Definition 1 and 2 the strategy profile of agents in an iterated prisoner's dilemma converges to *(cooperate, cooperate)* if and only if $\lim_{t\to\infty} \Pr(a_{kt} = \textit{cooperate}) = 1$ for $k = 1, 2$. Note that

$$\lim_{\substack{T\to 0^+ \\ t\to\infty}} \Pr(a_{1t} = \textit{cooperate}) = \lim_{T\to 0^+} \sum_{j=1}^{8} P_j = \lim_{T\to 0^+} P_1 \tag{26}$$

and similarly that

$$\lim_{\substack{T\to 0^+\\ t\to\infty}} \Pr(a_{2t} = \textit{cooperate}) = \lim_{T\to 0^+} \left(\sum_{j=1}^{4} P_j + \sum_{j=9}^{12} P_j \right) = \lim_{T\to 0^+} P_1. \quad (27)$$

By combining (25), (26), and (27) it turns out that the condition $w - v > 2(v - u)$ is necessary and sufficient for convergence to the *(cooperate, cooperate)* strategy profile. In a similar way it can be shown that the condition $w - v < 2(v - u)$ is necessary and sufficient for convergence to the *(defect, defect)* strategy profile. This completes the proof of Theorem 1. ■

Theorem 1 assumes that agents use the Boltzmann strategy to choose their actions. The following theorem states that cooperative behavior does not emerge in an iterated prisoner's dilemma if instead of the Boltzmann strategy agents use the ϵ-greedy strategy to choose their actions.

Theorem 2: Consider an iterated prisoner's dilemma with an infinite number of periods. Let Assumption 1, 2, and 3 be satisfied. Let both agents use single-state Q-learning combined with the ϵ-greedy strategy. The strategy profile of the agents then converges to *(defect, defect)* for all possible values of the payoffs in the one-shot prisoner's dilemma game.

Proof: Theorem 2 can largely be proven in a similar way as Theorem 1. Using a Markov chain with the states that are shown in Table II and following a similar reasoning as in the proof of Theorem 1, it can be derived that

$$\lim_{\epsilon\to 0} P_1 = \lim_{\epsilon\to 0} \frac{p_{16\to 1}}{p_{16\to 1} + p_{1\to 5} + p_{1\to 9} + p_{1\to 13}}. \quad (28)$$

Note that this expression resembles the expression given by (24) in the proof of Theorem 1. In the ϵ-greedy strategy exploration takes place with probability ϵ. In case of exploration each action has the same probability of being chosen. As we already discussed in the proof of Theorem 1, the probability of a state transition can be calculated by multiplying each agent's probability of choosing a particular action. It can now be seen that the transition probabilities in (28) are given by $p_{16\to 1} = p_{1\to 13} = \epsilon^2/4$ and $p_{1\to 5} = p_{1\to 9} = \epsilon/2 - \epsilon^2/4$. Equation (28) therefore has the form $0/0$ and the numerator in (28) approaches 0 faster than the denominator. This implies that $\lim_{\epsilon\to 0} P_1 = 0$ for all possible values of the payoffs in the one-shot prisoner's dilemma game. It follows from $\lim_{\epsilon\to 0} P_1 = 0$ that the strategy profile of the agents converges to *(defect, defect)*. This completes the proof of Theorem 2. ■

IV. Experimental Results

The theoretical results presented in the previous section are based on certain simplifying assumptions. In this section we test experimentally to what extent these results remain valid if some of the underlying assumptions are relaxed. We focus on Assumption 2 and 3, which state, respectively, that the learning rate α has a fixed value of 1 and that agents almost never explore. Assumption 2 is relaxed by experimenting with fixed values of α that are lower than 1. Assumption 3 is relaxed by experimenting with exploration strategies in which the probability of exploration is quite high initially and is gradually decreased over time in such a way that it approaches 0.

TABLE III

The Number of Experiments in Which Both Agents Cooperated in Period 500,000. The Agents Used the Boltzmann Strategy. The Total Number of Experiments Was 50

u	v	w	z	$\alpha = 0.05$	$\alpha = 0.20$	$\alpha = 0.50$	$\alpha = 1.00$
0	1	9	10	48	50	50	50
0	2	9	10	1	50	49	49
0	2	6	10	0	0	9	10
0	3	6	10	0	0	0	0

TABLE IV

The Number of Experiments in Which Both Agents Cooperated in Period 500,000. The Agents Used the ϵ-greedy Strategy. The Total Number of Experiments Was 50

u	v	w	z	$\alpha = 0.05$	$\alpha = 0.20$	$\alpha = 0.50$	$\alpha = 1.00$
0	1	9	10	48	30	1	0
0	2	9	10	13	2	0	0
0	2	6	10	0	0	0	0
0	3	6	10	0	0	0	0

The setup of the experiments that we performed was as follows. In an experiment an iterated prisoner's dilemma was played by two agents that both chose their actions using single-state Q-learning. A game lasted 500,000 periods. Such a large number of periods turned out to be necessary in order to obtain reliable results. The agents in an experiment used either the Boltzmann strategy or the ϵ-greedy strategy. In both strategies the probability of exploration was gradually decreased over time in such a way that it approached 0. In the Boltzmann strategy this was achieved by decreasing the temperature parameter T as follows

$$T = 10 \cdot 0.999994^t \quad (29)$$

where t denotes the current period in an iterated prisoner's dilemma. In the ϵ-greedy strategy ϵ was decreased according to

$$\epsilon = 0.99999^t \quad (30)$$

In an experiment a fixed value was used for the learning rate α. The following values for α were considered: 0.05, 0.20, 0.50, and 1.00. We also used a number of different values for the payoffs in the one-shot prisoner's dilemma game. For each combination of an exploration strategy, a value of α, and values of the prisoner's dilemma payoffs, we performed 50 experiments. In each experiment different random numbers were used.

We are interested in the number of experiments in which the agents learned to cooperate with each other. For various values of the prisoner's dilemma payoffs and of the learning rate α, the number of experiments in which both agents cooperated in period 500,000 is reported in Table III for the

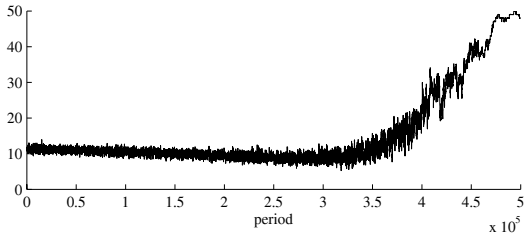

Fig. 1. The number of experiments in which both agents cooperated as a function of the period in the iterated prisoner's dilemma. The following parameter settings were used: $u = 0$, $v = 1$, $w = 9$, $z = 10$, and $\alpha = 0.05$. The agents used the Boltzmann strategy. The graph was smoothed by taking a moving average over 25 periods. The total number of experiments was 50.

Boltzmann strategy and in Table IV for the ϵ-greedy strategy. In experiments in which the agents did not both cooperate in period 500,000 it was almost always found that both agents defected. It should be noted that in some experiments the agents required a lot of time to learn to cooperate with each other. This was the case in, for example, the experiments in which $u = 0$, $v = 1$, $w = 9$, $z = 10$, and $\alpha = 0.05$ and in which the agents used the Boltzmann strategy. In Fig. 1 it is shown how the behavior of the agents changed during these experiments.

We first discuss the results of the experiments in which the learning rate α was equal to 1. These results agree quite well with the theoretical predictions presented in Section III. Of course, the setup of the experiments differed somewhat from the assumptions of the theoretical analysis. The theoretical analysis considered the limit case in which the number of periods in an iterated prisoner's dilemma approaches infinity and the probability of exploration approaches 0. In the experiments, on the other hand, it was only possible to simulate a finite number of periods and a very small probability of exploration. These deviations from the assumptions of the theoretical analysis provide an explanation for small differences between the experimental results and the theoretical predictions. Such a difference was found in the experiments in which $u = 0$, $v = 2$, $w = 9$, and $z = 10$ and in which the agents used the Boltzmann strategy. As can be seen in Table III, in one of these experiments cooperative behavior did not emerge, even though it was predicted by the theoretical analysis.

The results of the experiments in which the learning rate α had a value lower than 1 indicate that the theoretical predictions presented in Section III are quite robust. The results show that Assumption 2, which states that α has a fixed value of 1, can be relaxed without affecting Q-learning's ability to cooperate in an essential way. In the experiments in which the agents used the Boltzmann strategy, cooperative behavior was also observed when α had a fixed value lower than 1 (see Table III). To obtain cooperative behavior it seems that for lower values of α a more restrictive condition on the payoffs in the one-shot prisoner's dilemma game has to be imposed. In a similar way the results of the experiments in which the agents used the ϵ-greedy strategy indicate that the theoretical predictions for this strategy are quite robust. In these experiments cooperative behavior did not emerge for most values of α and most values of the payoffs in the one-shot prisoner's dilemma game (see Table IV). Only when the difference between the payoff of mutual cooperation (w) and the payoff of mutual defection (v) was large and at the same time the value of α was low, cooperative behavior sometimes emerged in the experiments.

V. Discussion

The theoretical analysis presented in Section III shows that under certain simplifying assumptions multi-agent Q-learning in an iterated prisoner's dilemma may converge to cooperative behavior. Under the assumptions of the analysis convergence to cooperative behavior takes place if the Boltzmann strategy is used in the Q-learning algorithm and the payoffs in the one-shot prisoner's dilemma game satisfy the condition in (11). Convergence to cooperative behavior does not take place if the ϵ-greedy strategy is used or if the Boltzmann strategy is used and the condition in (11) is not satisfied. It should be emphasized that these results were derived for the specific definition of convergence provided in Section III. Since in the analysis agents were assumed not to have a memory for remembering what happened in the past, cooperative behavior did not constitute a Nash equilibrium. It therefore follows from the results obtained in Section III that under certain assumptions it is possible for multi-agent Q-learning to converge to a strategy profile that is not a Nash equilibrium.

Some of the assumptions on which the theoretical analysis in this paper is based are rather strong. Moreover, using the proof technique presented in this paper it seems difficult to relax the assumptions. However, the experimental results reported in Section IV indicate that our theoretical results are quite robust to violations of the assumptions. Most importantly, the experimental results show that the assumption of a learning rate α with a fixed value of 1 can be relaxed without affecting Q-learning's ability to cooperate in an essential way. It turns out that as the value of α is decreased cooperation between agents that use the Boltzmann strategy becomes somewhat more difficult whereas cooperation between agents that use the ϵ-greedy strategy becomes somewhat easier.

In addition to explaining the emergence of cooperative behavior in multi-agent Q-learning, the theoretical and experimental results in this paper also draw attention to a more general issue, namely the influence on the behavior of the Q-learning algorithm of such factors as the exploration strategy, the value of the learning rate α, and the payoff values. The results show that in an iterated prisoner's dilemma multi-agent Q-learning may converge to either mutual cooperation or mutual defection, which are two completely opposite outcomes. Which of these outcomes is realized depends on the exploration strategy, the value of α, and the values of the payoffs in the one-shot prisoner's dilemma game. The strong influence of these factors on the behavior of the Q-learning algorithm is likely to be a more general phenomenon that also occurs

in other settings than prisoner's dilemmas. It is important to take this phenomenon into account in experimental studies on (multi-agent) Q-learning. The results of studies that do not consider the influence of factors like those mentioned above may in many cases not be very robust.

Finally it may be interesting to note that Theorem 1 and 2 can also be proven using mathematical techniques from the field of evolutionary game theory. These techniques are discussed in, for example, [23]–[25]. Although the use of mathematical techniques from the field of evolutionary game theory may result in proofs that are intuitively easier to understand than the proofs in this paper, we have chosen not to use these techniques because most readers are probably not familiar with them. However, the techniques may be useful for constructing proofs similar to the ones in this paper. In that way one may, for example, extend the analysis in this paper to other games than prisoner's dilemmas.

REFERENCES

[1] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, England, 1989.

[2] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[3] C. Fang, S. Kimbrough, A. Valluri, Z. Zheng, and S. Pace, "On adaptive emergence of trust behavior in the game of stag hunt," *Group Decision and Negotiation*, vol. 11, pp. 449–467, 2002.

[4] M. Littman and P. Stone, "Leading best-response strategies in repeated games," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

[5] T. Sandholm and R. Crites, "Multiagent reinforcement learning in the iterated prisoner's dilemma," *Biosystems*, vol. 37, pp. 147–166, 1996.

[6] J. Stimpson and M. Goodrich, "Learning to cooperate in a social dilemma: a satisficing approach to bargaining," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[7] N. van Eck and M. van Wezel, "Application of reinforcement learning to the game of Othello," *Computers & Operations Research*, accepted for publication.

[8] L. Waltman and U. Kaymak, "Reinforcement learning in repeated Cournot oligopoly games," in *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and Their Implementation on Smart Adaptive Systems 2004*, 2004, pp. 209–217.

[9] ——, "Q-learning agents in a Cournot oligopoly model," *Journal of Economic Dynamics and Control*, submitted.

[10] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998, pp. 746–752.

[11] J. Crandall and M. Goodrich, "Learning to compete, compromise, and cooperate in repeated general-sum games," in *Proceedings of the Twenty-Second International Conference on Machine Learning*, 2005, pp. 161–168.

[12] A. Greenwald and K. Hall, "Correlated Q-learning," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 242–249.

[13] J. Hu and M. Wellman, "Nash Q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.

[14] M. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proceedings of the Eleventh International Conference on Machine Learning*, 1994, pp. 157–163.

[15] ——, "Friend-or-foe Q-learning in general-sum games," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 322–328.

[16] S. Kimbrough and M. Lu, "A note on Q-learning in the Cournot game," in *Proceedings of the Second Workshop on E-Business*, 2003.

[17] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, pp. 215–250, 2002.

[18] B. Banerjee, R. Mukherjee, and S. Sen, "Learning mutual trust," in *Working Notes of the Autonomous Agents 2000 Workshop on Deception, Fraud and Trust in Agent Societies*, 2000, pp. 9–14.

[19] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[20] R. Sutton and A. Barto, *Reinforcement learning: an introduction*. MIT Press, 1998.

[21] T. Mitchell, *Machine learning*. McGraw-Hill, 1997.

[22] F. Hillier and G. Lieberman, *Introduction to operations research*, 7th ed. McGraw-Hill, 2001.

[23] M. Kandori, G. Mailath, and R. Rob, "Learning, mutation, and long run equilibria in games," *Econometrica*, vol. 61, pp. 29–56, 1993.

[24] M. Kandori and R. Rob, "Evolution of equilibria in the long run: a general theory and applications," *Journal of Economic Theory*, vol. 65, pp. 383–414, 1995.

[25] P. Young, "The evolution of conventions," *Econometrica*, vol. 61, pp. 57–84, 1993.