Reinforcement learning of coordination in cooperative multi-agent systems

Spiros Kapetanakis and **Daniel Kudenko**
{spiros, kudenko}@cs.york.ac.uk
Department of Computer Science
University of York, Heslington, York
YO10 5DD, U.K.

Abstract

We report on an investigation of reinforcement learning techniques for the learning of coordination in cooperative multi-agent systems. Specifically, we focus on a novel action selection strategy for Q-learning (Watkins 1989). The new technique is applicable to scenarios where mutual observation of actions is not possible.

To date, reinforcement learning approaches for such *independent* agents did not guarantee convergence to the optimal joint action in scenarios with high miscoordination costs. We improve on previous results (Claus & Boutilier 1998) by demonstrating empirically that our extension causes the agents to converge almost always to the optimal joint action even in these difficult cases.

Introduction

Learning to coordinate in cooperative multi-agent systems is a central and widely studied problem, see, for example (Lauer & Riedmiller 2000), (Boutilier 1999), (Claus & Boutilier 1998), (Sen & Sekaran 1998), (Sen, Sekaran, & Hale 1994), (Weiss 1993). In this context, coordination is defined as *the ability of two or more agents to jointly reach a consensus over which actions to perform in an environment*. We investigate the case of *independent* agents that cannot observe one another's actions, which often is a more realistic assumption.

In this investigation, we focus on reinforcement learning, where the agents must learn to coordinate their actions through environmental feedback. To date, reinforcement learning methods for independent agents (Tan 1993), (Sen, Sekaran, & Hale 1994) did not guarantee convergence to the *optimal* joint action in scenarios where miscoordination is associated with high penalties. Even approaches using agents that are able to build predictive models of each other (so-called *joint-action learners*) have failed to show convergence to the optimal joint action in such difficult cases (Claus & Boutilier 1998). We investigate variants of Q-learning (Watkins 1989) in search of improved convergence to the optimal joint action in the case of independent agents. More specifically, we investigate the effect of the estimated value function in the Boltzmann action selection strategy for Q-learning. We introduce a novel estimated value function and evaluate it experimentally on two especially difficult coordination problems that were first introduced by Claus & Boutilier in 1998: the *climbing game* and the *penalty game*. The empirical results show that the convergence probability to the optimal joint action is greatly improved over other approaches, in fact reaching almost 100%.

Our paper is structured as follows: we first introduce the aforementioned common testbed for the study of learning coordination in cooperative multi-agent systems. We then introduce a novel action selection strategy and discuss the experimental results. We finish with an outlook on future work.

Single-stage coordination games

A common testbed for studying the problem of multi-agent coordination is that of repeated cooperative single-stage games (Fudenberg & Levine 1998). In these games, the agents have common interests i.e. they are rewarded based on their joint action and all agents receive the same reward. In each round of the game, every agent chooses an action. These actions are executed simultaneously and the reward that corresponds to the joint action is broadcast to all agents.

A more formal account of this type of problem was given by Claus & Boutilier in 1998. In brief, we assume a group of n agents $\alpha_1, \alpha_2, \ldots, \alpha_n$ each of which have a finite set of *individual actions* A_i which is known as the agent's *action space*. In this game, each agent α_i chooses an individual action from its action space to perform. The action choices make up a *joint action*. Upon execution of their actions all agents receive the reward that corresponds to the joint action. For example, Table 1 describes the reward function for a simple cooperative single-stage game. If agent 1 executes action b and agent 2 executes action a, the reward they receive is 5. Obviously, the optimal joint-action in this simple game is (b, b) as it is associated with the highest reward of 10.

Our goal is to enable the agents to learn optimal coordination from repeated trials. To achieve this goal, one can use either *independent* or *joint-action* learners. The difference between the two types lies in the amount of information they can perceive in the game. Although both types of learners can perceive the reward that is associated with each joint action, the former are unaware of the existence of other agents

		Agent 1	
		a	b
Agent 2	a	3	5
	b	0	10

Table 1: A simple cooperative game reward function.

whereas the latter can also perceive the actions of others. In this way, joint-action learners can maintain a model of the strategy of other agents and choose their actions based on the other participants' perceived strategy. In contrast, independent learners must estimate the value of their individual actions based solely on the rewards that they receive for their actions. In this paper, we focus on individual learners, these being more universally applicable.

In our study, we focus on two particularly difficult coordination problems, the climbing game and the penalty game. These games were introduced by Claus & Boutilier in 1998. This focus is without loss of generality since the climbing game is representative of problems with high miscoordination penalty and a single optimal joint action whereas the penalty game is representative of problems with high miscoordination penalty and multiple optimal joint actions. Both games are played between two agents. The reward functions for the two games are included in Tables 2 and 3:

		Agent 1		
		a	b	c
Agent 2	a	11	-30	0
	b	-30	7	6
	c	0	0	5

Table 2: The climbing game table.

In the climbing game, it is difficult for the agents to converge to the optimal joint action (a, a) because of the negative reward in the case of miscoordination. For example, if agent 1 plays a and agent 2 plays b, then both will receive a negative reward of -30. Incorporating this reward into the learning process can be so detrimental that both agents tend to avoid playing the same action again. In contrast, when choosing action c, miscoordination is not punished so severely. Therefore, in most cases, both agents are easily tempted by action c. The reason is as follows: if agent 1 plays c, then agent 2 can play either b or c to get a positive reward (6 and 5 respectively). Even if agent 2 plays a, the result is not catastrophic since the reward is 0. Similarly, if agent 2 plays c, whatever agent 1 plays, the resulting reward will be at least 0. From this analysis, we can see that the climbing game is a challenging problem for the study of learning coordination. It includes heavy miscoordination penalties and "safe" actions that are likely to tempt the agents away from the optimal joint action.

Another way to make coordination more elusive is by including multiple optimal joint actions. This is precisely what happens in the penalty game of Table 3.

In the penalty game, it is not only important to avoid the miscoordination penalties associated with actions (c, a) and

		Agent 1		
		a	b	c
Agent 2	a	10	0	k
	b	0	2	0
	c	k	0	10

Table 3: The penalty game table.

(a, c). It is equally important to agree on which optimal joint action to choose out of (a, a) and (c, c). If agent 1 plays a expecting agent 2 to also play a so they can receive the maximum reward of 10 but agent 2 plays c (perhaps expecting agent 1 to play c so that, again, they receive the maximum reward of 10) then the resulting penalty can be very detrimental to both agents' learning process. In this game, b is the "safe" action for both agents since playing b is guaranteed to result in a reward of 0 or 2, regardless of what the other agent plays. Similarly with the climbing game, it is clear that the penalty game is a challenging testbed for the study of learning coordination in multi-agent systems.

Reinforcement learning

A popular technique for learning coordination in cooperative single-stage games is one-step Q-learning, a reinforcement learning technique. Since the agents in a single-stage game are stateless, we need a simple reformulation of the general Q-learning algorithm such as the one used by Claus & Boutilier. Each agent maintains a Q value for each of its actions. The value $Q(\text{action})$ provides an estimate of the usefulness of performing this action in the next iteration of the game and these values are updated after each step of the game according to the reward received for the action. We apply Q-learning with the following update function:

$$Q(\text{action}) \leftarrow Q(\text{action}) + \lambda(r - Q(\text{action}))$$

where λ is the learning rate $(0 < \lambda < 1)$ and r is the reward that corresponds to choosing this action.

In a single-agent learning scenario, Q-learning is guaranteed to converge to the optimal action independent of the action selection strategy. In other words, given the assumption of a stationary reward function, single-agent Q-learning will converge to the optimal policy for the problem. However, in a multi-agent setting, the action selection strategy becomes crucial for convergence to *any* joint action. A major challenge in defining a suitable strategy for the selection of actions is to strike a balance between exploring the usefulness of moves that have been attempted only a few times and exploiting those in which the agent's confidence in getting a high reward is relatively strong. This is known as the *exploration/exploitation problem*.

The action selection strategy that we have chosen for our research is the Boltzmann strategy (Kaelbling, Littman, & Moore 1996) which states that agent α_i chooses an action to perform in the next iteration of the game with a probability that is based on its current estimate of the usefulness of that action, denoted by $\mathrm{EV}(\text{action})^1$:

$$P(\text{action}) = \frac{e^{\frac{\mathrm{EV}(\text{action})}{T}}}{\sum_{\text{action}' \in A_i} e^{\frac{\mathrm{EV}(\text{action}')}{T}}}$$

In the case of Q-learning, the agent's estimate of the usefulness of an action may be given by the Q values themselves, an approach that has been usually taken to date.

We have concentrated on a proper choice for the two parameters of the Boltzmann function: the estimated value and the temperature. The importance of the temperature lies in that it provides an element of controlled randomness in the action selection: high values in temperature encourage exploration since variations in Q values become less important. In contrast, low temperature values encourage exploitation. The value of the temperature is typically decreased over time from an initial value as exploitation takes over from exploration until it reaches some designated lower limit. The three important settings for the temperature are the initial value, the rate of decrease and the number of steps until it reaches its lowest limit. The lower limit of the temperature needs to be set to a value that is close enough to 0 to allow the learners to converge by stopping their exploration. Variations in these three parameters can provide significant difference in the performance of the learners. For example, starting with a very high value for the temperature forces the agents to make random moves until the temperature reaches a low enough value to play a part in the learning. This may be beneficial if the agents are gathering statistical information about the environment or the other agents. However, this may also dramatically slow down the learning process.

It has been shown (Singh *et al.* 2000) that convergence to a joint action can be ensured if the temperature function adheres to certain properties. However, we have found that there is more that can be done to ensure not just convergence to *some* joint action but convergence to the *optimal* joint action, even in the case of independent learners. This is not just in terms of the temperature function but, more importantly, in terms of the action selection strategy. More specifically, it turns out that a proper choice for the estimated value function in the Boltzmann strategy can significantly increase the likelihood of convergence to the optimal joint action.

FMQ heuristic

In difficult coordination problems, such as the climbing game and the penalty game, the way to achieve convergence to the optimal joint action is by influencing the learners towards their individual components of the optimal joint action(s). To this effect, there exist two strategies: altering the Q-update function and altering the action selection strategy.

Lauer & Riedmiller (2000) describe an algorithm for multi-agent reinforcement learning which is based on the *optimistic* assumption. In the context of reinforcement learning, this assumption implies that an agent chooses any action it finds suitable expecting the other agent to choose the best match accordingly. More specifically, the optimistic assumption affects the way Q values are updated. Under this assumption, the update rule for playing action α defines that $Q(\alpha)$ is only updated if the new value is greater than the current one.

Incorporating the optimistic assumption into Q-learning solves both the climbing game and penalty game every time. This fact is not surprising since the penalties for miscoordination, which make learning optimal actions difficult, are neglected as their incorporation into the learning tends to lower the Q values of the corresponding actions. Such lowering of Q values is not allowed under the optimistic assumption so that all the Q values eventually converge to the maximum reward corresponding to that action for each agent. However, the optimistic assumption fails to converge to the optimal joint action in cases where the maximum reward is misleading, e.g., in stochastic games (see experiments below). We therefore consider an alternative: the *Frequency Maximum Q Value* (FMQ) heuristic.

Unlike the optimistic assumption, that applies to the Q update function, the FMQ heuristic applies to the action selection strategy, specifically the choice of $\mathrm{EV}(\alpha)$, i.e. the function that computes the estimated value of action α. As mentioned before, the standard approach is to set $\mathrm{EV}(\alpha) = Q(\alpha)$. Instead, we propose the following modification:

$$\mathrm{EV}(\alpha) = Q(\alpha) + c * \mathrm{freq}(\mathrm{maxR}(\alpha)) * \mathrm{maxR}(\alpha)$$

where:

① $\mathrm{maxR}(\alpha)$ denotes the maximum reward encountered *so far* for choosing action α.
② $\mathrm{freq}(\mathrm{maxR}(\alpha))$ is the fraction of times that $\mathrm{maxR}(\alpha)$ has been received as a reward for action α over the times that action α has been executed.
③ c is a weight that controls the importance of the FMQ heuristic in the action selection.

Informally, the FMQ heuristic carries the information of how frequently an action produces its maximum corresponding reward. Note that, for an agent to receive the maximum reward corresponding to one of its actions, the other agent must be playing the game accordingly. For example, in the climbing game, if agent 1 plays action a which is agent 1's component of the optimal joint-action (a, a) but agent 2 doesn't, then they both receive a reward that is less than the maximum. If agent 2 plays c then the two agents receive 0 and, provided they have already encountered the maximum rewards for their actions, both agents' FMQ estimates for their actions are lowered. This is due to the fact that the frequency of occurrence of maximum reward is lowered. Note that setting the FMQ weight c to zero reduces the estimated value function to: $\mathrm{EV}(\alpha) = Q(\alpha)$.

In the case of independent learners, there is nothing other than action choices and rewards that an agent can use to learn coordination. By ensuring that enough exploration is permitted in the beginning of the experiment, the agents have a good chance of visiting the optimal joint action so that the FMQ heuristic can influence them towards their appropriate individual action components. In a sense, the FMQ heuristic

[1]In (Kaelbling, Littman, & Moore 1996), the estimated value is introduced as *expected reward* (ER).

defines a model of the environment that the agent operates in, the other agent being part of that environment.

Experimental results

This section contains our experimental results. We compare the performance of Q-learning using the FMQ heuristic against the baseline experiments i.e. experiments where the Q values are used as the estimated value of an action in the Boltzmann action selection strategy. In both cases, we use only independent learners. The comparison is done by keeping all other parameters of the experiment the same, i.e. using the same temperature function and experiment length. The evaluation of the two approaches is performed on both the climbing game and the penalty game.

Temperature settings

Exponential decay in the value of the temperature is a popular choice in reinforcement learning. This way, the agents perform all their learning until the temperature reaches some lower limit. The experiment then finishes and results are collected. The temperature limit is normally set to zero which may cause complications when calculating the action selection probabilities with the Boltzmann function. To avoid such problems, we have set the temperature limit to 1 in our experiments[2].

In our analysis, we use the following temperature function:

$$T(x) = e^{-sx} * \mathrm{max\_temp} + 1$$

where x is the number of iterations of the game so far, s is the parameter that controls the rate of exponential decay and $\mathrm{max\_temp}$ is the value of the temperature at the beginning of the experiment. For a given length of the experiment $(\mathrm{max\_moves})$ and initial temperature $(\mathrm{max\_temp})$ the appropriate rate of decay (s) is automatically derived. Varying the parameters of the temperature function allows a detailed specification of the temperature. For a given $\mathrm{max\_moves}$, we experimented with a variety of $s, \mathrm{max\_temp}$ combinations and found that they didn't have a significant impact on the learning in the baseline experiments. Their impact is more significant when using the FMQ heuristic. This is because setting $\mathrm{max\_temp}$ at a very high value means that the agent makes random moves in the initial part of the experiment. It then starts making more knowledgeable moves (i.e. moves based on the estimated value of its actions) when the temperature has become low enough to allow variations in the estimated value of an action to have an impact on the probability of selecting that action.

Evaluation on the climbing game

The climbing game has one optimal joint action (a, a) and two heavily penalised actions (a, b) and (b, a). We use the settings $\mathrm{max\_temp} = 500$ and vary $\mathrm{max\_moves}$ from 500 to 2000. The learning rate λ is set to 0.9. Figure 1 depicts the likelihood of convergence to the optimal joint action in the baseline experiments and using the FMQ heuristic with $c = 1$ and $c = 10$. The FMQ heuristic outperforms the baseline experiments for both settings of c. For $c = 10$, the FMQ heuristic converges to the optimal joint action almost always even for short experiments.

[2] This is done without loss of generality.

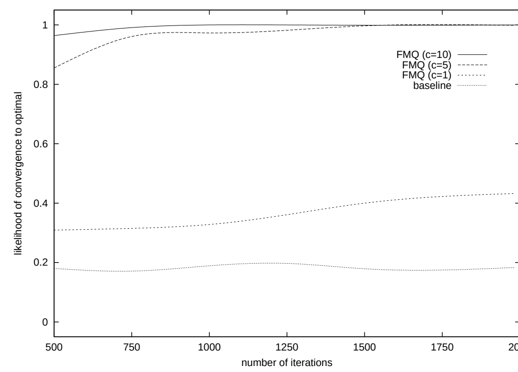

Figure 1: Likelihood of convergence to the optimal joint action in the climbing game (averaged over 1000 trials).

Evaluation on the penalty game

The penalty game is harder to analyse than the climbing game. This is because it has two optimal joint actions (a, a) and (c, c) for all values of $k \leq 0$. The extent to which the optimal joint actions are reached by the agents is affected severely by the size of the penalty. However, the performance of the agents depends not only on the size of the penalty k but also on whether the agents manage to agree on which optimal joint action to choose. Table 2 depicts the performance of the learners for $k = 0$ for the baseline experiments and with the FMQ heuristic for $c = 1$.

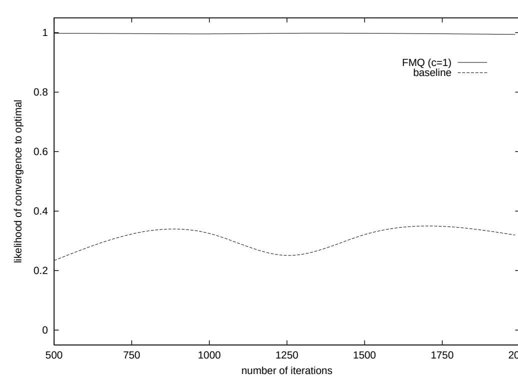

Figure 2: Likelihood of convergence to the optimal joint action in the penalty game $k = 0$ (averaged over 1000 trials).

As shown in Figure 2, the performance of the FMQ heuristic is much better than the baseline experiment. When $k = 0$, the reason for the baseline experiment's failure is not the existence of a miscoordination penalty. Instead, it is the existence of multiple optimal joint actions that causes the agents to converge to the optimal joint action so infrequently. Of course, the penalty game becomes much harder for greater penalty. To analyse the impact of the penalty on the convergence to optimal, Figure 3 depicts the likelihood that convergence to optimal occurs as a function of the penalty. The four plots correspond to the baseline experiments and using Q-learning with the FMQ heuristic for $c = 1$, $c = 5$ and $c = 10$.

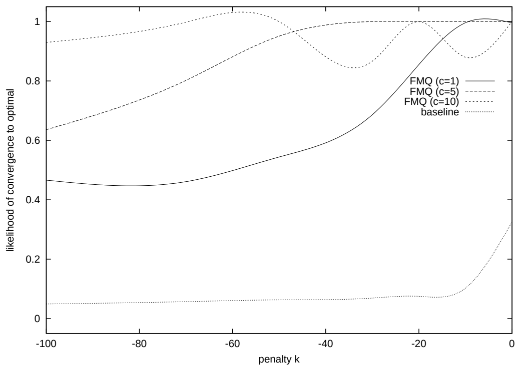

Figure 3: Likelihood of convergence to the optimal joint action as a function of the penalty (averaged over 1000 trials).

From Figure 3, it is obvious that higher values of the FMQ weight c perform better for higher penalty. This is because there is a greater need to influence the learners towards the optimal joint action when the penalty is more severe.

Further experiments

We have described two approaches that perform very well on the climbing game and the penalty game: FMQ and the optimistic assumption. However, the two approaches are different and this difference can be highlighted by looking at alternative versions of the climbing game. In order to compare the FMQ heuristic to the optimistic assumption (Lauer & Riedmiller 2000), we introduce a variant of the climbing game which we term *the partially stochastic climbing game*. This version of the climbing game differs from the original in that one of the joint actions is now associated with a stochastic reward. The reward function for the partially stochastic climbing game is included in Table 4.

		Agent 1		
		a	b	c
Agent 2	a	11	-30	0
	b	-30	14/0	6
	c	0	0	5

Table 4: The partially stochastic climbing game table.

Joint action (b, b) yields a reward of 14 or 0 with probability 50%. The partially stochastic climbing game is functionally equivalent to the original version. This is because, if the two agents consistently choose their b action, they receive the same overall value of 7 over time as in the original game.

Using the optimistic assumption on the partially stochastic climbing game consistently converges to the suboptimal joint action (b, b). This because the frequency of occurrence of a high reward is not taken into consideration at all. In contrast, the FMQ heuristic shows much more promise in convergence to the optimal joint action. It also compares favourably with the baseline experimental results. Tables 5, 6 and 7 contain the results obtained with the baseline experiments, the optimistic assumption and the FMQ heuristic for 1000 experiments respectively. In all cases, the parameters are: $s = 0.006$, $max\_moves = 1000$, $\max\_\mathrm{temp} = 500$ and, in the case of FMQ, $c = 10$.

	a	b	c
a	212	0	3
b	0	12	289
c	0	0	381

Table 5: Baseline experimental results.

	a	b	c
a	0	0	0
b	0	1000	0
c	0	0	0

Table 6: Results with optimistic assumption.

	a	b	c
a	988	0	0
b	0	4	0
c	0	7	1

Table 7: Results with the FMQ heuristic.

The final topic for evaluation of the FMQ heuristic is to analyse the influence of the weight (c) on the learning. Informally, the more difficult the problem, the greater the need for a high FMQ weight. However, setting the FMQ weight at too high a value can be detrimental to the learning. Figure 4 contains a plot of the likelihood of convergence to optimal in the climbing game as a function of the FMQ weight.

From Figure 4, we can see that setting the value of the FMQ weight above 15 lowers the probability that the agents will converge to the optimal joint action. This is because, by setting the FMQ weight too high, the probabilities for action selection are influenced too much towards the action with the highest FMQ value which may not be the optimal joint action early in the experiment. In other words, the agents become too narrow-minded and follow the heuristic blindly since the FMQ part of the estimated value function overwhelms the Q values. This property is also reflected in the experimental results on the penalty game (see Figure 3) where setting the FMQ weight to 10 performs very well in difficult experiments with $-100 < k < -50$ but there is a drop in performance for easier experiments. In contrast, for $c = 1$ the likelihood of convergence to the optimal joint action in easier experiments is significantly higher than in more difficult ones.

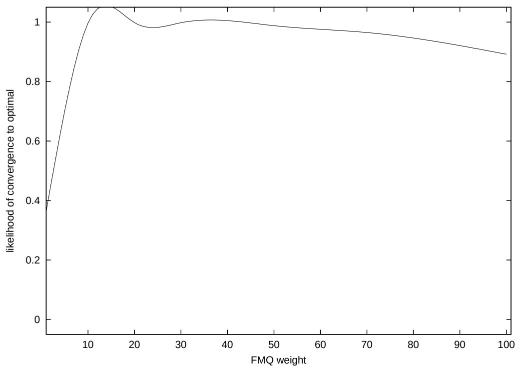

Figure 4: Likelihood of convergence to optimal in the climbing game as a function of the FMQ weight (averaged over 1000 trials).

Limitations

The FMQ heuristic performs equally well in the partially stochastic climbing game and the original deterministic climbing game. In contrast, the optimistic assumption only succeeds in solving the deterministic climbing game. However, we have found a variant of the climbing game in which both heuristics perform poorly: *the fully stochastic climbing game*. This game has the characteristic that *all* joint actions are probabilistically linked with two rewards. The average of the two rewards for each action is the same as the original reward from the deterministic version of the climbing game so the two games are functionally equivalent. For the rest of this discussion, we assume a 50% probability. The reward function for the stochastic climbing game is included in Table 8.

		Agent 1		
		a	b	c
Agent 2	a	10/12	5/-65	8/-8
	b	5/-65	14/0	12/0
	c	5/-5	5/-5	10/0

Table 8: The stochastic climbing game table (50%).

It is obvious why the optimistic assumption fails to solve the fully stochastic climbing game. It is for the same reason that it fails with the partially stochastic climbing game. The maximum reward is associated with joint action (b, b) which is a suboptimal action. The FMQ heuristic, although it performs marginally better than normal Q-learning still doesn't provide any substantial success ratios. However, we are working on an extension that may overcome this limitation.

Outlook

We have presented an investigation of techniques that allows two independent agents that are unable to sense each other's actions to learn coordination in cooperative single-stage games, even in difficult cases with high miscoordination penalties. However, there is still much to be done towards understanding exactly how the action selection strategy can influence the learning of optimal joint actions in this type of repeated games. In the future, we plan to investigate this issue in more detail.

Furthermore, since agents typically have a state component associated with them, we plan to investigate how to incorporate such coordination learning mechanisms in multi-stage games. We intend to further analyse the applicability of various reinforcement learning techniques to agents with a substantially greater action space. Finally, we intend to perform a similar systematic examination of the applicability of such techniques to partially observable environments where the rewards are perceived stochastically.

References

Boutilier, C. 1999. Sequential optimality and coordination in multiagent systems. In *Proceedings of the Sixteenth International Joint Conference on Articial Intelligence (IJCAI-99)*, 478–485.

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

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

Kaelbling, L. P.; Littman, M.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4.

Lauer, M., and Riedmiller, M. 2000. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference in Machine Learning*.

Sen, S., and Sekaran, M. 1998. Individual learning of coordination knowledge. *JETAI* 10(3):333–356.

Sen, S.; Sekaran, M.; and Hale, J. 1994. Learning to coordinate without sharing information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 426–431.

Singh, S.; Jaakkola, T.; Littman, M. L.; and Szpesvari, C. 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning Journal* 38(3):287–308.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, 330–337.

Watkins, C. J. C. H. 1989. *Learning from Delayed Rewards*. Ph.D. Dissertation, Cambridge University, Cambridge, England.

Weiss, G. 1993. Learning to coordinate actions in multi-agent systems. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, volume 1, 311–316. Morgan Kaufmann Publ.