The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems

Caroline Claus and Craig Boutilier
Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada V6T 1Z4
{cclaus,cebly}@cs.ubc.ca

1 Introduction

The application of learning to the problem of coordination in multiagent systems (MASs) has become increasingly popular in AI and game theory. The use of reinforcement learning (RL), in particular, has attracted recent attention [11, 9, 8, 6]. As noted in [9], using RL as a means of achieving coordinated behavior is attractive because of its generality and robustness.

Standard techniques for RL, for example, Q-learning [10], have been applied directly to MASs with some success. However, a general understanding of the conditions under which RL can be usefully applied, and exactly what form RL might take in MASs, are problems that have not yet been tackled in depth. We might ask the following questions:

- Are there differences between agents that learn as if there are no other agents (i.e., use single agent RL algorithms) and agents that attempt to learn both the values of specific *joint* actions and the strategies employed by other agents?
- Are RL algorithms guaranteed to converge in multiagent settings? If so, do they converge to (optimal) equilibria?
- How are rates of convergence and limit points influenced by the system structure and learning dynamics?

In this paper, we begin to address some of these questions in a very specific context, namely, repeated games in which agents have common interests (i.e., cooperative MASs). We focus our attention on Q-learning, due to its relative simplicity (certainly not for its general efficacy), and consider some of the factors that may influence the dynamics of multiagent Q-learning, and provide partial answers to these questions.

We first distinguish and compare two forms of multiagent RL (MARL). *Independent learners* (ILs) apply Q-learning in the classic sense, ignoring the existence of other agents. *Joint action learners* (JALs), in contrast, learn the value of their own actions in conjunction with those of other agents via integration of RL with equilibrium (or coordination) learning methods [12, 5, 4, 7]. We also examine the influence of partial observability on JALs, and how game structure and exploration strategies influence the dynamics of the learning process and the convergence to equilibrium. We conclude by mentioning several problems that promise to make the integration of RL with coordination learning an exciting area of research for the foreseeable future.

2 Preliminary Concepts and Notation

2.1 Single Stage Games

Our interest is in *n-player cooperative repeated games*.[1] We assume a collection α of n (heterogeneous) agents, each agent $i \in \alpha$ having available to it a finite set of *individual actions* A_i. Agents repeatedly play a *stage game* in which they each independently select an individual action to perform. The chosen actions at any point constitute a *joint action*, the set of which is denoted $\mathcal{A} = \times_{i \in \alpha} A_i$. With each $a \in \mathcal{A}$ is associated a (possibly stochastic) reward $R(a)$; the decision problem is *cooperative* since there is a single reward function R reflecting the utility assessment of all agents. The agents wish to choose actions that maximize reward.

A *randomized strategy* for agent i is a distribution $\pi \in \Delta(A_i)$ (where $\Delta(A_i)$ is the set of distributions over the agent's action set A_i). Intuitively, $\pi(a^i)$ denotes the probability of agent i selecting the individual action a^i. A strategy π is *deterministic* if $\pi(a^i) = 1$ for some $a^i \in A_i$. A *strategy profile* is a collection $\Pi = \{\pi_i : i \in \alpha\}$ of strategies for each agent i. The expected value of acting according to a fixed profile can easily be determined. If each $\pi_i \in \Pi$ is deterministic, we can think of Π as a joint action. A *reduced profile for agent i* is a strategy profile for all agents but i (denoted Π_{-i}). Given a profile Π_{-i}, a strategy π_i is a *best response* for agent i if the expected value of the strategy profile $\Pi_{-i} \cup \{\pi_i\}$ is maximal for agent i; that is, agent i could not do better using any other strategy π'_i. Finally, we say that the strategy profile Π is a *Nash equilibrium* iff $\Pi[i]$ (i's component of Π) is a best response to Π_{-i}, for every agent i. Note that in cooperative games, deterministic equilibria are easy to find. An equilibrium (or joint action) is *optimal* if no other has greater value.

As an example, consider the simple two-agent stage game:

	$a0$	$a1$
$b0$	x	0
$b1$	0	y

Agents A and B each have two actions at their disposal, $a0, a1$ and $b0, b1$, respectively. If $x > y > 0$, $\langle a0, b0\rangle$ and $\langle a1, b1\rangle$ are both equilibria; but only the first is optimal.

[1] Most of our conclusions hold *mutatis mutandis* for sequential, *multiagent Markov decision processes* [2] with multiple states.

2.2 Learning in Coordination Games

Action selection is more difficult if there are multiple optimal joint actions. If, for instance, $x = y > 0$, neither agent has a reason to prefer one or the other of its actions. If they choose them randomly, or in some way reflecting personal biases, then they risk choosing a suboptimal, or *uncoordinated* joint action. The general problem of *equilibrium selection* can be addressed in several ways. We consider the notion that an equilibrium (i.e., coordinated action choice in our cooperative setting) might be learned through repeated play of the game [4, 5, 7, 8].

One especially simple, yet often effective, model for achieving coordination is *fictitious play* [3, 4]. Each agent i keeps a count $C^j_{a^j}$, for each $j \in \alpha$ and $a^j \in A_j$, of the number of times agent j has used action a^j in the past. When the game is encountered, i treats the relative frequencies of each of j's moves as indicative of j's current (randomized) strategy. That is, for each agent j, i assumes j plays action $a^j \in A_j$ with probability $\Pr^i_{a^j} = C^j_{a^j} / (\sum_{b^j \in A_j} C^j_{b^j})$. This set of strategies forms a reduced profile Π_{-i}, for which agent i adopts a best response. After the play, i updates its counts appropriately, given the actions used by the other agents. We think of these counts as reflecting the beliefs an agent has regarding the play of the other agents (initial counts can also be weighted to reflect priors).

This very simple adaptive strategy will converge to an equilibrium in our simple cooperative games, and can be made to converge to an optimal equilibrium [12, 1]; that is, the probability of coordinated equilibrium after *k* interactions can be made arbitrarily high by increasing *k* sufficiently. It is also not hard to see that once the agents reach an equilibrium, they will remain there—each best response simply reinforces the beliefs of the other agents that the coordinated equilibrium remains in force.

We note that most game theoretic models assume that each agent can observe the actions executed by its counterparts with certainty. As pointed out and addressed in [1, 6], this assumption is often unrealistic. We will be interested below in the more general case where each agent obtains an *observation* which is related stochastically to the actual joint action selected. Formally, we assume an observation set O, and an observation model $\rho : \mathcal{A} \rightarrow \Delta(O)$. Intuitively, $\rho(a)(o)$, which we write in the less cumbersome fashion $\Pr_a(o)$, denotes the probability of observation o being obtained by all agents when joint action a is performed. Each agent is aware of this function.[2]

2.3 Reinforcement Learning

Action selection is more difficult still if agents are unaware of the rewards associated with various joint actions. In such a case, *reinforcement learning* can be used by the agents to estimate, based on past experience, the expected reward associated with individual or joint actions.

A simple, well-understood algorithm for single agent learning is *Q-learning* [10]. In our stateless setting, we assume a *Q-value*, $Q(a)$, that provides an estimate of the value of performing (individual or joint) action a. An agent updates its estimate $Q(a)$ based on sample $\langle a, r \rangle$ as follows:

$$Q(a) + \alpha(r + -Q(a)) \qquad (1)$$

The sample $\langle a, r \rangle$ is the "experience" obtained by the agent: action a was performed resulting in reward r. Here α is the learning rate ($0 \leq \alpha \leq 1$), governing to what extent the new sample replaces the current estimate. If α is decreased "slowly" during learning and all actions are sampled infinitely, Q-learning will converge to true Q-values for all actions in the single agent setting [10].

Convergence of Q-learning does not depend on the *exploration strategy* used. An agent can try its actions at any time—there is no requirement to perform actions that are currently estimated to be best. Of course, if we hope to enhance overall performance during learning, it makes sense (at least intuitively) to bias selection toward better actions. We can distinguish two forms of exploration. In *nonexploitive exploration*, an agent randomly chooses its actions with uniform probability. There is no attempt to use what was learned to improve performance—the aim is simply to learn Q-values. In *exploitive exploration* an agent chooses its best estimated action with probability p_x, and chooses some other action with probability $1 - p_x$. Often the exploitation probability p_x is increased slowly over time. We call a nonoptimal action choice an *exploration step* and $1 - p_x$ the exploration probability. Nonoptimal action selection can be uniform during exploration, or can be biased by the magnitudes of Q-values. A popular biased strategy is *Boltzmann exploration*: action a is chosen with probability

$$\frac{e^{Q(a)/T}}{\sum_{a'} e^{Q(a')/T}} \qquad (2)$$

The temperature parameter T can be decreased over time so that the exploitation probability increases.

There are two distinct ways in which Q-learning could be applied to a multiagent system. We say a MARL algorithm is an *independent learner* (IL) algorithm if the agents learn Q-values for their individual actions based on Equation (1). In other words, they perform their actions, obtain a reward and update their Q-values without regard to the actions performed by other agents. Experiences for agent i take the form $\langle a^i, r \rangle$ where a^i is the action performed by i and r is a reward. If an agent is unaware of the existence of other agents, cannot identify their actions, or has no reason to believe that other agents are acting strategically, then this is an appropriate method of learning. Of course, even if these conditions do not hold, an agent may choose to ignore information about the other agents' actions.

A *joint action learner* (JAL) is an agent that learns Q-values for joint actions as opposed to individual actions. The experiences for such an agent are of the form $\langle a, r \rangle$ where a is a joint action. This implies that each agent can observe the actions of other agents. We can generalize the picture slightly by allowing experiences of the form $\langle a^i, o, r \rangle$ where a^i is the action performed by i, and o is its (joint action) observation.

For JALs, exploration strategies require some care. In the example above, if A currently has Q-values for all four joint actions, the expected value of performing $a0$ or $a1$ depends

[2] Generalizations of this model could also be used.

crucially on the strategy adopted by B. To determine the relative values of their *individual* actions, each agent in a JAL algorithm maintains beliefs about the strategies of other agents. Here we will use simple empirical distributions, possibly with biased initial weights as in fictitious play. Agent A, for instance, assumes that each other agent B will choose actions in accordance with A's current beliefs about B (i.e., A's empirical distribution over B's action choices). In general, agent i assesses the expected value of its individual action a^i to be

$$EV(a^i) = \sum_{a^{-i} \in A_{-i}} Q(a^{-i}i \cup \{a^i\}) \prod_{j \neq i} \{\mathrm{Pr}^i_{a^{-i}[j]}\}$$

Agent i can use these values just as it would Q-values in implementing an exploration strategy.[3]

Maintaining a belief distribution by means of fictitious play is problematic if agents have imprecise observational capabilities. Following [1], we use a simple Bayesian updating rule for beliefs:

$$Pr(a[j] = a^j | a[i] = a^i, o) = \frac{Pr(o|a[j] = a^j, a[i] = a^i) Pr(a[j] = a^j)}{Pr(o|a[i] = a^i)}$$

Agent i then updates its distribution over j's probabilities using this "stochastic observation;" in particular, the count $C^j_{a_k}$ is incremented by $Pr(a^j_k|o)$ (intuitively, by a "fractional" outcome).[4]

3 Comparing Independent and Joint-Action Learners

We first compare the relative performance of independent and joint-action learners on a simple coordination game of the form described above:

	$a0$	$a1$
$b0$	10	0
$b1$	0	10

The first thing to note is that ILs using nonexploitive exploration will not deem either of their choices (on average) to be better than the other. For instance, A's Q-values for both action $a0$ and $a1$ will converge to 5, since whenever, say, $a0$ is executed, there is a 0.5 probability of $b0$ and $b1$ being executed. Of course, at any point, due to the stochastic nature of the strategies and the decay in learning rate, we would expect that the learned Q-values will not be identical; thus the agents, once they converge, might each have a reason to prefer one action to the other. Unfortunately, these biases need not be coordinated.

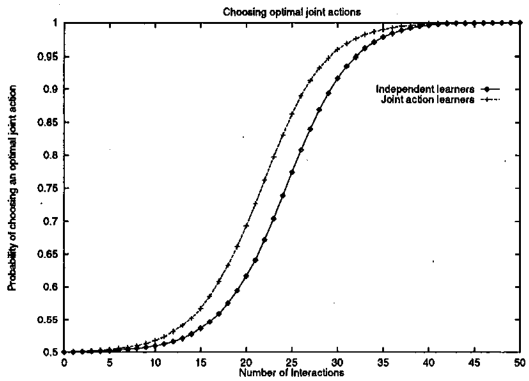

Figure 1: Convergence of coordination for ILs and JALs (averaged over 100 trials).

Rather than pursuing this direction, we consider the case where both the ILs and JALs use Boltzmann exploration. Exploitation of the known values allows the agents to "coordinate" in their choices for the same reasons that equilibrium learning methods work when agents know the reward structure. Figure 1 shows the probability of two ILs and JALs selecting an optimal joint action as a function of the number of interactions they have. The temperature parameter is $T = 16$ initially and decayed by a factor of 0.9^t at the $t + 1$st interaction. We see that ILs coordinate quite quickly. There is no preference for either equilibrium point: each of the two equilibria was attained in about half of the trials. We do not show convergence of Q-values, but note that the Q-values for the actions of the equilibria attained (e.g., $\langle a0, b0 \rangle$) tended to 10 while the other actions tended to 0. We note that probability of optimal action selection does not increase smoothly within individual trials; the averaged probabilities reflect the likelihood of having reached an equilibrium by time t, as well as exploration probabilities. We also point out that much faster convergence can be had for different parameter settings (e.g., decaying temperature T more rapidly). We defer general remarks on convergence to Section 5.

The figure also shows convergence for JALs under the same circumstances (full observability is assumed). JALs do perform somewhat better after a fixed number of interactions, as shown in the graph. While the JALs have more information at their disposal, convergence is not enhanced dramatically. In retrospect, this should not be too surprising. While JALs are able to distinguish Q-values of different joint actions, their ability to use this information is circumscribed by the strictures of the action selection mechanism. An agent maintains beliefs about the strategy being played by the other agents and "exploits" actions according to expected value based on these beliefs. In other words, the value of individual actions "plugged in" to the exploration strategy is more or less the same as the Q-values learned by ILs—the only distinction is that JALs *compute* them using explicit belief distributions and joint Q-values instead of updating them directly. Thus, even though the agents may be fairly sure of the relative Q-values of joint actions, Boltzmann exploration

[3] The expression for $EV(a^i)$ makes the justifiable assumption that the other agents are selecting their actions independently. Less reasonable is the assumption that these choices are uncorrelated, or even correlated with i's choices. Such correlations can often emerge due to the dynamics of belief updating without agents being aware of this correlation, especially if frequencies of particular joint actions are ignored.

[4] This essentially corresponds to using the empirical counts as Dirichlet parameters, and treating the i's beliefs as a Dirichlet distribution over j's set of mixed strategies.

does not let them exploit this.[5]

4 The Effects of Partial Action Observability

When JALs are in a situation where partial observability holds sway, the updating of Q-values should be undertaken with more care. Given an experience $\langle a^i, o, r\rangle$, where a^i is agent i's action and o is the resultant observation, there can be a number of joint actions a consistent with a^i and o, according to the observation model, that give rise to this experience. Intuitively, i should update the Q-values for *each* of such a, something that conflicts with the usual Q-learning model. To deal with this, we propose the use of "fractional" updates of Q-values: the Q-value for joint action a will be updated by reward r; but the magnitude of this update will be "tempered" by the probability $\Pr(a|o, a^i)$ that a was taken given o, a^i.[6] Specifically, agent i updates joint Q-values for all actions a where $a[i] = a^i$ using:

$$Q(a) = Q(a) + \alpha \Pr(a|o, a^i)(r - Q(a)) \tag{3}$$

where $\Pr(a|o, a^i)$ is computed in the obvious way by i using its beliefs and Bayes rule.

We first note that if agent i is learning in a stationary environment (i.e., all other agents play fixed, mixed strategies), then this update rule will converge; that is, the limiting Q-values are well-defined, under the same conditions required of Q-learning. Informally:

Proposition 1 *Convergence of Q-values using update rule (3) is assured in stationary environments.*

We note that convergence of Q-values is enhanced when the "fractional" nature of the updates is accounted for when decaying the learning parameter α—we typically maintain a separate α_a for each joint action a and decay α_a using the total experience with a (i.e., weighting the decay rate using $\Pr(a|o, a^i)$). We also note that convergence cannot generally be to the true Q-values of joint actions unless the observation model is perfect. In our running example, for instance, if observations do not permit agent A to distinguish $b0$ from $b1$ perfectly, then the Q-value for, say, $\langle a0, b0\rangle$ must lie somewhere between 0 and 10.

Figure 2 shows how convergence to equilibrium is influenced by the accuracy of the observation model in the game above. The model associates a unique observation o_a with each of the four joint actions a. The observation probability refers to $\Pr(o_a|a)$, and ranges from 0.25 (fully unobservable) to 1 (fully observable). If $o \neq o_a$ then $\Pr(o|a) = (1 - \Pr(o_a|a))/3$. Rate of convergence refers to the expected number of interactions until convergence (i.e., the probability of selecting an optimal joint action reaches 0.99). The more accurate the observation model, the more quickly the agents converge. Figure 3 shows the convergence of joint Q-values for observation probability 0.5.

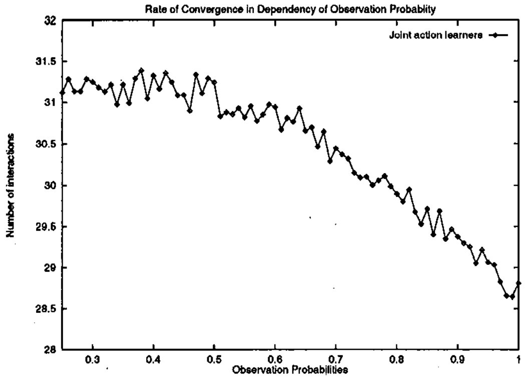

Figure 2: Time to convergence (opt. joint action with prob. 0.99) as a function of observation prob. (averaged over 1000 trials; $\alpha = 0.2, T = 16, T$ decayed at rate 0.9).

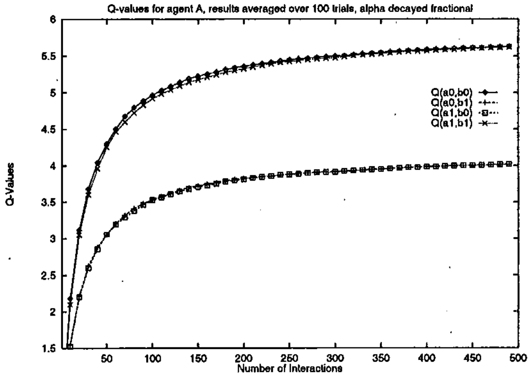

Figure 3: Convergence of Q-values with observation probability 0.5 (averaged over 1000 trials).

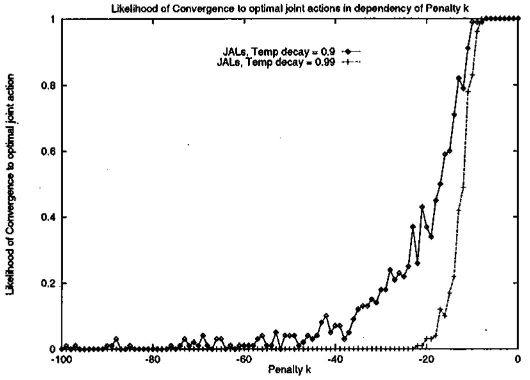

Figure 4: Likelihood of convergence to opt. equilibrium as a function of penalty k (averaged over 100 trials).

[5] The key reason for the difference in ILs and JALs is the larger difference in Q-values for JALs, which bias Boltzmann exploration slightly more toward the estimated optimal action.

[6] An alternative way of viewing this: a single $Q(a)$ is chosen by the agent, with probability $\Pr(a|o, a^i)$, for update with reward r.

5 Convergence and Game Structure

We will argue below that Q-learning, in both the IL and JAL cases, will converge to an equilibrium under appropriate conditions. Of course, in general, convergence to an optimal equilibrium cannot be assured. Before making this case, we consider the ways in which the game structure can influence the dynamics of the learning process.

	$a0$	$a1$	$a2$
$b0$	10	0	k
$b1$	0	2	0
$b2$	k	0	10

When $k < 0$, this game has three deterministic equilibria, of which two are preferred. If $k = -100$, agent A, during initial exploration, will find its first and third actions to be unattractive because of B's random exploration. If A is an IL, the average rewards (and hence Q-values) for $a0, a2$ will be quite low; and if A is a JAL, its beliefs about B's strategy will afford these actions low expected value. Similar remarks apply to B, and the self-confirming nature of equilibria virtually assure convergence to $\langle a1, b1 \rangle$. However, the closer k is to 0, the lower the likelihood the agents will find their first and third actions unattractive—the stochastic nature of exploration means that, occasionally, these actions will have high estimated utility and convergence to one of the optimal equilibria will occur. Figure 4 shows how the probability of convergence to one of the optimal equilibria is influenced by the magnitude of the "penalty" k. Not surprisingly, different equilibria can be attained with different likelihoods.[7]

Thus far, our examples show agents proceeding on a direct route to equilibria (albeit at various rates, and with destinations "chosen" stochastically). Unfortunately, convergence is not so straightforward in general. Consider the game:

	$a0$	$a1$	$a2$
$b0$	11	-30	0
$b1$	-30	7	6
$b2$	0	0	5

Initially, the two learners are almost certainly going to begin to play the nonequilibrium strategy profile $\langle a2, b2 \rangle$. This is seen clearly in Figures 5 and 6. However, once they "settle" at this point, as long as exploration continues (here Boltzmann exploration is used), agent B will soon find $b1$ to be more attractive—so long as A continues to primarily choose $a2$. Once the nonequilibrium point $\langle a2, b1 \rangle$ is attained, agent A tracks B's move and begins to perform action $a1$. Once this equilibrium is reached, the agents remain there. Figures 5 and 6 clearly show the agents settling into nonequilibria at various points and "climbing" toward an equilibrium over time along a best reply path. The convergence (and resting points) for the joint Q-values are shown in Figure 7.[8]

This phenomenon will obtain in general, allowing one to conclude that the multiagent Q-learning schemes we have proposed will converge to equilibria almost surely—but only if certain conditions are met in the case of ILs. First, exploration must continue at any point with nonnegative probabil-

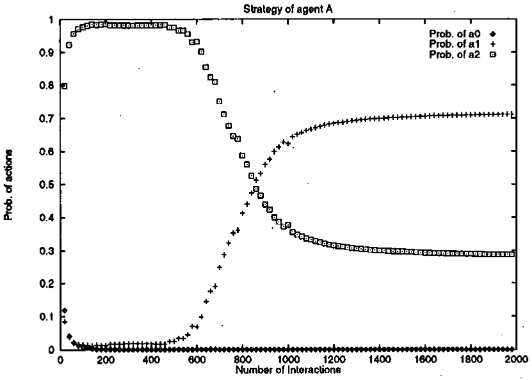

Figure 5: Dynamics of A's strategy.

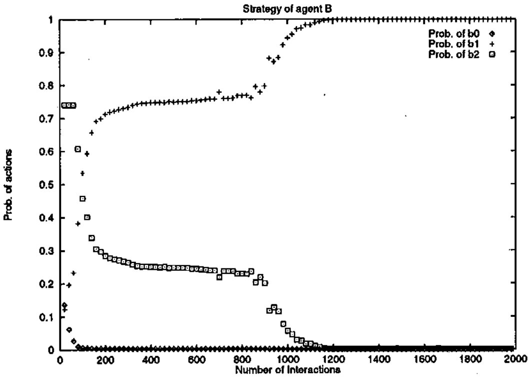

Figure 6: Dynamics of B's strategy.

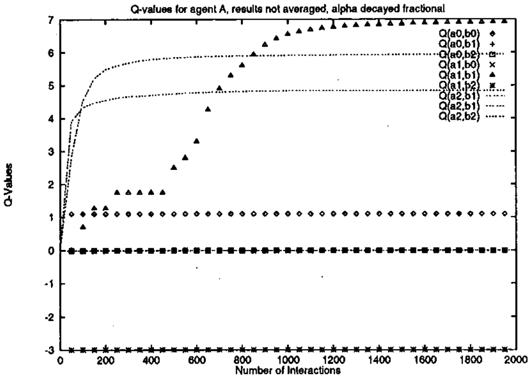

Figure 7: Dynamics of joint Q-values ($\alpha = 0.1$).

[7] These results are shown for JALs; but the general pattern holds true for ILs as well.

[8] These results are based on Boltzmann exploration with a temperature setting of 1 and *no decay*—this is to induce considerable exploration, otherwise convergence is quite slow (see below). This explains why agent A converges to $\Pr(a1) = 0.7$.

ity. This is required by Q-learning in any case, but is necessary to ensure that a nonequilibrium is not prematurely accepted. Second, when agents are exploring, if the probability of both moving away from a nonequilibrium point simultaneously is sufficiently high, then we cannot expect agent A to notice that it has a new best reply. So, for instance, if we use uniform exploration over time, we cannot induce B to learn a good enough Q-value for its "break out move" to actually move away (e.g., every time $b1$ is tried, there is a high enough probability that A does $a0$ that $b1$ never looks better than $b2$).

We note that Boltzmann exploration with a decaying temperature parameter ensures infinite exploration. Furthermore, as the temperature decreases over time, so too does the probability of simultaneous exploration. Thus, at some finite time t (assuming bounded rewards), the probability of B executing $b1$ while A performs $a2$ will become high enough to render $b1$ more attractive after some finite number of experiences with $b1$. At this point, (with high probability) B will perform $b1$, allowing A to respond in a similar fashion.

These arguments can be put together to show that, with proper exploration and proper use of best responses (i.e., that are at least *asymptotically myopic* [4]), we will eventually converge to an equilibrium. Informally:

Proposition 2 *IL and JAL algorithms will converge to an equilibrium if sufficient, but decreasing, exploration is undertaken.*[9]

We remark that this theoretical guarantee of convergence may not be of much practical value for sufficiently complicated games. The key difficulty is that convergence relies on the use of decaying exploration: this is necessary to ensure that the agents' estimated values are based (ultimately) on a stationary environment. This gradual decay, however, makes the time required to shift from the current entrenched strategy profile to a better profile rather long. If the agents initially settle on a profile that is a large distance (in terms of a best reply path) from an equilibrium, each shift required can take longer to occur because of the decay in exploration. Furthermore, as pointed out above, the probability of concurrent exploration may have to be sufficiently small to ensure that the expected value of a shift along the best reply path is greater than no such shift, which can introduce further delays in the process. We note that the longer these delays are, the lower the learning rate α becomes, requiring even more experience to overcome the initially biased estimated Q-values. Finally, the largest drawback lies in the fact that—even when JALs have perfect joint Q-values—beliefs based on a lot of experience require a considerable amount of contrary experience to be overcome. For example, once B has made the shift from $b2$ to $b1$ above, a significant amount of time is needed for A to switch from $a2$ to $a1$: it has to observe B performing $b1$ enough to overcome the rather large degree of belief it had that B would continue doing $b2$.

6 Concluding Remarks

We have seen described two basic ways in which Q-learning can be applied in multiagent cooperative settings, and examined the impact of various features on the success of the interaction between equilibrium selection learning techniques with RL techniques. We have demonstrated that the integration requires some care, and that Q-learning is not nearly as robust as in single-agent settings.

These considerations point toward a number of interesting avenues of research directed toward improving the performance of MARL. We remark on some of the directions we are currently pursuing. The first is the enhancement of practical convergence by the use of "windowing" methods for estimating beliefs in JALs. By using only the last k experiences to determine empirical belief distributions, the abatement of convergence caused by strongly held prior beliefs can be alleviated. We note that a memory-based technique of this sort might be applied to ILs if reward experiences are stored and old experiences discounted in determining Q-values. The second avenue we are investigating are possible techniques for encouraging agents to converge to optimal equilibria. This is one area where JALs have a distinct advantage over ILs: even if they have converged to an equilibrium, they can tell—since they have access to joint Q-values—if a better equilibrium exists. Coordination learning techniques (see, e.g., [1]) might then be applied, as could other exploration techniques that attempt to induce a shift from one equilibrium to another. Finally, we hope to explore the details associated with general, multistate sequential decision problems and investigate the application of generalization techniques in domains with large state spaces.

Acknowledgements: Thanks to Leslie Kaelbling and Michael Littman for their helpful discussions in the early stages of this work.

References

[1] C. Boutilier. Learning conventions in multiagent stochastic domains using likelihood estimates. *UAI-96*, pp.106–114, Portland.

[2] C. Boutilier. Planning, learning and coordination in multiagent decision processes. *TARK-96*, pp.195–210, Amsterdam.

[3] G. W. Brown. Iterative solution of games by fictitious play. T. C. Koopmans (ed.), *Activity Analysis of Production and Allocation*. Wiley, 1951.

[4] D. Fudenberg and D. M. Kreps. *Learning and Equilibrium in Strategic Form Games*. CORE Foundation, Louvain, 1992.

[5] D. Fudenberg and D. K. Levine. Steady state learning and Nash equilibrium. *Econometrica*, 61(3):547–573, 1993.

[6] J. Hu and M. P. Wellman. Self-fulfilling bias in multiagent learning. *ICMAS-96*, pp.118–125, Kyoto.

[7] E. Kalai and E. Lehrer. Rational learning leads to Nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.

[8] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. *ML-94*, pp.157–163, New Brunswick.

[9] S. Sen, M. Sekaran, and J. Hale. Learning to coordinate without sharing information. *AAAI-94*, pp.426–431, Seattle.

[10] C. J. C. H. Watkins and P. Dayan. Q-learning. *Mach. Learn.*, 8:279–292, 1992.

[11] G. Weiß. Learning to coordinate actions in multi-agent systems. *IJCAI-93*, pp.311–316, Chambery.

[12] H. Peyton Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993.

[9] We have not yet constructed a formal proof, but see no impediments to this conjecture. The convergence conjecture also relies on the absence of cycles in best reply paths of cooperative games.