# 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

**Abstract**

Reinforcement learning can provide a robust and natural means for agents to learn how to coordinate their action choices in multiagent systems. We examine some of the factors that can influence the dynamics of the learning process in such a setting. We first distinguish reinforcement learners that are unaware of (or ignore) the presence of other agents from those that explicitly attempt to learn the value of joint actions and the strategies of their counterparts. We study (a simple form of) Q-learning in cooperative multiagent systems under these two perspectives, focusing on the influence of that game structure and exploration strategies on convergence to (optimal and suboptimal) Nash equilibria. We then propose alternative *optimistic* exploration strategies that increase the likelihood of convergence to an optimal equilibrium.

## 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 [22, 20, 16, 11, 7, 15]. As noted in [16], 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 [21], 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 action selection strategies?

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

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 [24, 5, 6, 9]. We then briefly consider the importance of exploitive exploration strategies and examine, through a series of examples, how game structure and exploration strategies influence the dynamics of the learning process and the convergence to equilibrium. We show that both JALs and ILs will converge to an equilibrium in this specific setting of fully cooperative, repeated games. In fact, even though JALs have much more information at their disposal, they do not perform much differently from ILs in the straightforward application of Q-learning to MASs. We also observe that in games with multiple equilibria, optimality of the "agreed upon" equilibrium is not assured. We then describe several *optimistic* exploration strategies, designed to increase the likelihood of reaching an optimal equilibrium. This provides one way of having JALs exploit the additional information that they possess. We conclude with a discussion of related work and mention several issues 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 the application of RL algorithms to sequential decision problems in which the system is being controlled by multiple agents. However, in the interests of simplicity, our investigations in this paper are focussed on *$n$-player cooperative (or common interest) repeated games*. Sequential optimality will not be of primary interest, though we will discuss this issue in Sections 5 and 6).[1] We can view the prob-



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

lem at hand, then, as a *distributed bandit problem*.

More formally, we assume a collection $\alpha$ 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 distribution over possible rewards; though the rewards are stochastic, for simplicity, we often simply refer to the expected reward $R(a)$. The decision problem is *cooperative* since each agent's reward is drawn from the same distribution, reflecting the utility assessment of all agents. The agents wish to choose actions that maximize (expected) reward.

We adopt some standard game theoretic terminology [13]. 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 $\pi$ 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 $\Pi$ as a joint action. A *reduced profile for agent* $i$ is a strategy profile for all agents but $i$ (denoted $\Pi_{-i}$). Given a profile $\Pi_{-i}$, a strategy $\pi_i$ 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 $\pi'_i$. Finally, we say that the strategy profile $\Pi$ is a *Nash equilibrium* iff $\Pi[i]$ ($i$'s component of $\Pi$) is a best response to $\Pi_{-i}$, 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: we would expect the agents to play $\langle a0, b0\rangle$.

## 2.2 Learning in Coordination Games

Action selection is more difficult if there are multiple optimal joint actions. If, for instance, $x = y > 0$ in the example above, 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* [13] can be addressed in several ways. For instance, communication between agents might be admitted [22] or one could impose conventions or rules that restrict behavior so as to ensure coordination [18]. Here we entertain the suggestion that coordinated action choice might be learned through repeated play of the game with the same agents [5, 6, 9, 11]. (Repeated play with a random selection of similar agents from a large population has also been the object of considerable study [17, 10, 24].)

One especially simple, yet often effective, learning model for achieving coordination is *fictitious play* [3, 5]. 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 $\Pi_{-i}$, 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 simple adaptive strategy will converge to an equilibrium in our simple cooperative games assuming that agents randomize when multiple best responses exist [12], and can be made to converge to an optimal equilibrium if appropriate mechanisms are adopted [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 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, 7], this assumption is often unrealistic. A more general model allows each agent to obtain an *observation* which is related stochastically to the actual joint action selected, where $\Pr_a(o)$ denotes the probability of observation $o$ being obtained by all agents when joint action $a$ is performed. We will not investigate this model further, but mention it here since it subsumes the two special cases we describe below.

## 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. We refer to [8] for a survey of RL techniques.

A simple, well-understood algorithm for single agent learning is *Q-learning* [21]. The formulation of Q-learning for general sequential decision processes is more sophisticated than we need here. 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:

$$Q(a) \leftarrow Q(a) + \lambda(r - Q(a)) \qquad (1)$$

The sample $\langle a, r\rangle$ is the "experience" obtained by the agent: action $a$ was performed resulting in reward $r$. Here $\lambda$ is the learning rate ($0 \leq \lambda \leq 1$), governing to what extent the new sample replaces the current estimate. If $\lambda$ 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 [21].[2]

we will see that interesting issues emerge.

[2] Generally, $Q(a, s)$ is taken to denote the long-term value of per-

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

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

The temperature parameter $T$ can be decreased over time so that the exploitation probability increases (and can be done in such a way that convergence is assured [19]).

The existence of multiple agents, each simultaneously learning, is a potential impediment to the successful employment of Q-learning (or RL generally) in multiagent settings. When agent $i$ is learning the value of its actions in the presence of other agents, it is learning in a nonstationary environment. Thus, the convergence of Q-values is not guaranteed. Naive application of Q-learning to MASs can be successful if we can ensure that each agent's strategy will eventually "settle." This is one of the questions we explore below. Application of Q-learning and other RL methods have met with some success in the past [22, 16, 17, 15].

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. The contrast between ILs and JALs can be illustrated in our example above: if $A$ is an IL, then it will learn Q-values for actions $a0$ and $a1$; if $A$ is a JAL, it will learn Q-values for all four joint actions, $\langle a0, b0\rangle$, etc.

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 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 empirical distributions, possibly with 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

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

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

We note that both JALs and ILs can be viewed as special cases of the partially observable model mentioned above, 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. A preliminary version of this paper [4] studies the methods below within this model.

# 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.

Rather than pursuing this direction, we consider the case where both the ILs and JALs use Boltzmann exploration (other exploitive strategies could be used). 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 op-

forming action $a$ in state $s$, and incorporates consideration of the values of possible states $s'$ to which action $a$ leads. This learning method is, in fact, a basic stochastic approximation technique [14]. We use (perhaps, misuse) the $Q$ notation and terminology to emphasize the connection with action selection.

[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

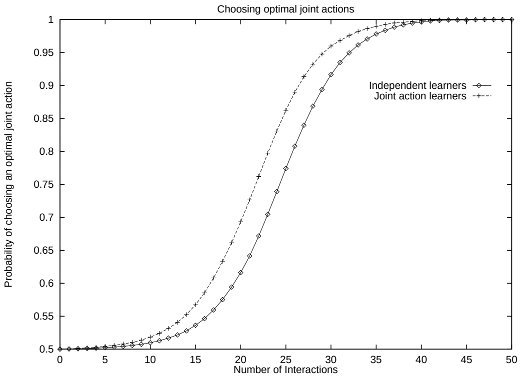


Figure 1: Convergence of coordination for ILs and JALs (averaged over 100 trials).

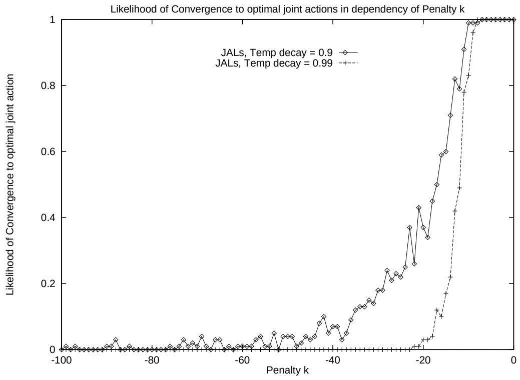


Figure 2: Likelihood of convergence to opt. equilibrium as a function of penalty $k$ (averaged over 100 trials).

timal 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 $0$. 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 4.

The figure also shows convergence for JALs under the same circumstances. 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 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 does not let them exploit this.[4]

of this correlation, especially if frequencies of particular joint actions are ignored.

[4] 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. Note that other exploitive strategies alleviate this problem to a certain degree.

## 4 Convergence and Game Structure

In the simple game considered above, it isn't difficult to see that both independent Q-learners and joint action Q-learners will converge on equilibria, as long as an exploitive exploration strategy with decreasing exploration is used. However, convergence is not always so smooth as illustrated in Figure 1. We know consider the ways in which the game structure can influence the dynamics of the learning process.

Consider the following class of games, with a variable (expected) *penalty* $k \leq 0$.

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

This game (for any penalty) has three deterministic equilibria, of which two ($\langle a0, b0\rangle$, $\langle a2, b2\rangle$) are preferred. If, say, $k = -100$, during initial exploration agent $A$ 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 2 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.[5]

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 following *climbing* game:

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

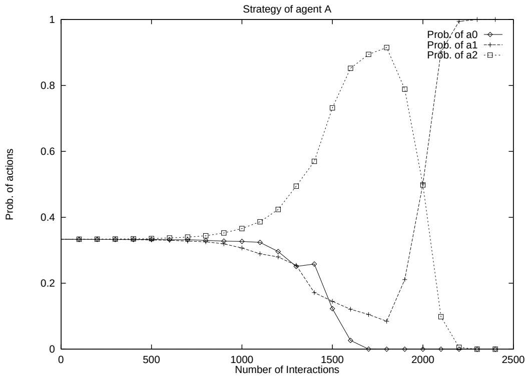


Figure 3: $A$'s strategy in climbing game

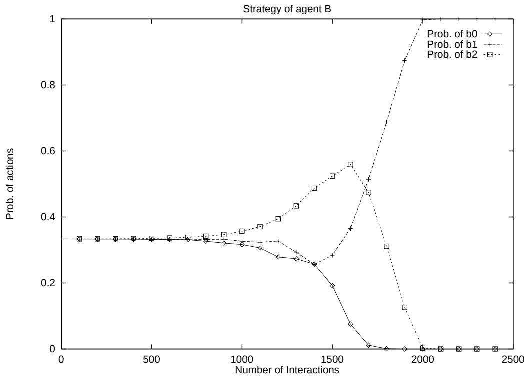


Figure 4: $B$'s strategy in climbing game

| | $a0$ | $a1$ | $a2$ |
|---|---|---|---|
| $b0$ | 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 3, 4 and 5.[6] However, once they "settle" at this point, as long as exploration continues, 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.

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. The conditions that are required in both cases are:

- The learning rate $\lambda$ decreases over time such that $\sum_{\lambda=0}^{t} \lambda = \infty$ and $\sum_{\lambda=0}^{t} \lambda^2 < \infty$.
- Each agent samples each of its actions infinitely often.

[6] Parameter settings for these figures: initial temperature 10000 is decayed at rate $0.995^t$.

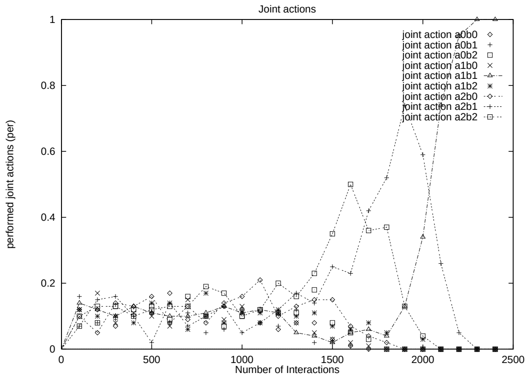


Figure 5: Joint actions in climbing game

- The probability $P_t^i(a)$ of agent $i$ choosing action $a$ is nonzero.
- Each agent's exploration strategy is exploitive. That is, $\lim_{t\to\infty} P_t^i(X_t) = 0$, where $X_t$ is a random variable denoting the event that some nonoptimal action was taken based on $i$'s estimated values at time $t$.

The first two conditions are required of Q-learning, and the third, if implemented appropriately (e.g., with appropriately decayed temperature), will ensure the second. Furthermore, it ensures that agents cannot adopt deterministic exploration strategies and become strictly correlated. Finally, the last condition ensures that agents exploit their knowledge. In the context of ficticious play and its variants, this exploration strategy would be *asymptotically myopic* [5]. This is necessary to ensure that an equilibrium will be reached. Under these conditions we have:

**Theorem 1** *Let $E_t$ be a random variable denoting the probability of a (deterministic) equilibrium strategy profile being played at time t. Then for both ILs and JALs, for any $\delta, \varepsilon > 0$, there is an $T(\delta, \varepsilon)$ such that*

$$\Pr(|E_t - 1| < \varepsilon) > 1 - \delta$$

*for all $t > T(\delta, \varepsilon)$.*

Intuitively (and somewhat informally), the dynamics of the learning process behaves as follows. If the agents are in equilibrium, there is a nonzero probability of moving out of equilibrium; but this generally requires a (rather dense) series of exploratory moves by one or more agents. The probability of this occurring decreases over time, making the likelihood of leaving an equilibrium just obtained vanish over time (both for JALs and ILs). If at some point the agents' estimated Q-values are such that a nonequilibrium is most likely, the likelihood of this state of affairs remaining also vanishes over time. As an example, consider the climbing game above. Once agents begin to play $\langle a2, b2\rangle$ regularly, agent $B$ is still required to explore. After a sufficient sampling of action $b1$—without agent $A$ *simultaneously* exploring and moving away from $a2$—$b1$ will look more attractive than $b2$ and this best reply will be adopted. Decreasing exploration ensures that the odds of simultaneous exploration decrease fast enough to assure that this happens with high probability. Similar reasoning shows that a best reply path will eventually be followed to a point of equilibrium.

This theoretical guarantee of convergence may be of limited practical value for sufficiently complicated games. The key difficulty is that convergence relies on the use of decaying exploration: this is necessary to "approximate" the best-response condition of fictitious play. 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 shift, which can introduce further delays in the process. The longer these delays are, the lower the learning rate $\lambda$ becomes, requiring more experience to overcome the initially biased estimated Q-values.

Finally, the key drawback for JALs (which know the joint Q-values) is the fact that 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 $b2$. Although we don't report on this here, our initial experiments using *windows* or finite histories upon which to base beliefs has shown considerable practical value.[7]

## 5 Biasing Exploration Strategies for Optimality

One thing we notice about the MARL strategies described above is that they do not ensure convergence to an optimal equilibrium. Little can be done about this is the case of ILs.[8] However, JALs have considerably more information at there disposal in the form of joint Q-values. For example, in the penalty game, agents A and B might converge to the suboptimal equilibrium $\langle a1, b1\rangle$; but both agents have learned the game structure and realize their coordinated strategy profile is suboptimal. Once attained, the usual exploration strategies permit escape from this equilibrium only with small, diminishing probability.

Intuitively, we can imagine both agents trying to break out of this equilibrium in an attempt to reach a more desirable point (say, $\langle a2, b2\rangle$). For instance, agent $B$ might sample $b2$ a number of times in order to induce $A$ to switch its strategy to $a2$. In fact, this can be worthwhile if the "penalties" received in the attempt are compensated for by the long run sequence of high rewards obtained once the optimal equilibrium is achieved. Note that this type of action selection runs counter to the requirement that a best response be chosen except for "random" exploration. This type of switch requires that agents intentionally choose (immediately) suboptimal actions.

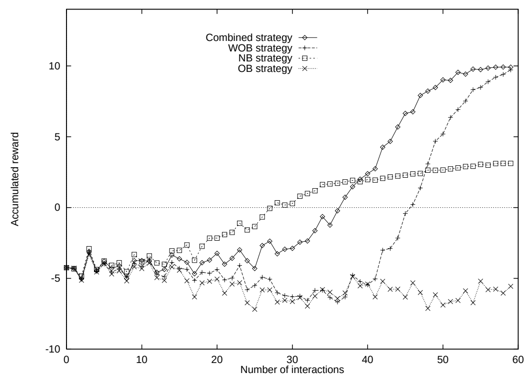


Figure 6: Sliding avg. reward in the penalty game

Ultimately, the decision to attain a long run optimal equilibrium at the expense of a finite sequence of penalties can be cast as a sequential decision problem. For instance, if future rewards are highly discounted, agents may not risk deviating from a suboptimal equilibrium. However, such a decision problem (especially when we move to more complex settings) can be intractable. Instead, we consider augmented exploration strategies that will encourage long run optimality. What we propose below are *myopic heuristics*, based only on the current state, that tend to induce long run optimal behavior. Three such heuristics are:

**Optimistic Boltzmann (OB):** For agent $i$, action $a_i \in A_i$, let *MaxQ*$(a_i) = \max_{\Pi_{-i}} Q(\Pi_{-i}, a_i)$. Choose actions with Boltzmann exploration (another exploitive strategy would suffice) using *MaxQ*$(a_i)$ as the value of $a_i$.

**Weighted OB (WOB):** Explore using Boltzmann using factors *MaxQ*$(a_i) \cdot \Pr_i(\text{optimal match } \Pi_{-i} \text{ for } a_i)$.

**Combined:** Let $C(a_i) = \rho$ *MaxQ*$(a_i) + (1-\rho)$*EV*$(a_i)$, for some $0 \leq \rho \leq 1$. Choose actions using Boltzmann exploration with $C(a_i)$ as value of $a_i$.

OB is optimistic in the sense that an agent assesses each of its actions as though the agents around it will act in order to "match" its choice of an action. WOB is a more realistic version of OB: the assessment of an action is tempered by the likelihood that a matching will be made (according to its current beliefs). Finally the combined strategy is more flexible: it uses a normal exploration strategy but introduces the *MaxQ* factor to bias exploration toward actions that have "potential." The coefficient $\rho$ allows one to tune this bias. The experiment below uses $\rho = 0.5$.

We have performed some preliminary experimentation with these heuristics. Figure 6 illustrates the results of these three heuristics, as well as normal Boltzmann (NB) exploration, for the penalty game ($k = -10$). It shows (sliding) average reward obtained over the last ten interactions for each strategy. Thus it shows not only the convergence behavior, but the penalties incurred in attempting to reach an

[7] Fictitious play based on histories of an appropriately chosen length is shown to converge in [24].

[8] One could imagine that an IL might bias its action selection toward those whose Q-values have high variance, or adhere to a multimodal distribution, perhaps indicative of another agent acting simultaneously; but this seems to run contrary to the "spirit" of ILs.

optimal equilibrium. NB behaves as above, sometimes converging to the optimal (10) and suboptimal (2) equilibrium. Not surprisingly, OB fares poorly: the presence of multiple equilibria make it impossible to do well (although it behaves reasonably well in simpler games). The two agents cannot coordinate because the are not permitted to account for the strategy of the other agent. WOB circumvents the difficulty with OB by using beliefs to ensure coordination; it converges to an optimal equilibrium each time. The Combined strategy also guarantees long run optimality, but it has better performance along the way.

We can draw few formal conclusions at this time; but we think the use of myopic heuristics for exploration deserves considerably more study. Methods like the Combined strategy that allow problem dependent tuning of the exploration strategy seem especially promising. By focusing on particular sequential optimality criteria, intelligent parameter tuning should be possible.

## 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. Convergence guarantees are not especially practical for complex games, but new exploration heuristics may help in this regard.

Several proposals have been put forth that are closely related to ours. Tan [20] and Sen, Sekaran and Hale [16] apply RL to *independent* agents and demonstrate empirical convergence. These results are consistent with ours, but properties of the convergence points (whether they are optimal or even in equilibrium are not considered). Wheeler and Narendra [23] develop a learning automata (LA) model for fully cooperative games. They show that using this model agents will converge to equilibrium if there is a *unique* pure strategy equilibrium; thus the coordination problem that interests us here is not addressed directly. Furthermore, the LA model is different from the Q-learning model we address. However, the connections between the two models deserve further exploration.

A number of important directions remain to be pursued. The most obvious is the generalization of these ideas to general, multistate, sequential problems for which Q-learning is designed (for instance, as addressed in [20, 16]. An interesting issue that emerges when one tries to directly apply fictitious play models to such a setting is estimating the value of actions using the Q-values of future states when the actual future value obtained can hinge on coordination (or lack thereof) at these future states. The application of generalization techniques to deal with large state and action spaces is also of great importance, especially in multiagent domains where the size of joint action spaces can grow exponentially with the number of agents. Finally, we expect these ideas to generalize to other settings (such as zero-sum games) where fictitious play is also known to converge.

**Acknowledgements:** Thanks to Leslie Kaelbling and Michael Littman for their helpful discussions in the early stages of this work and Daniel Koditschek for helpful pointers. This work was supported by NSERC Grant OGP0121843 and IRIS-II Project IC-7.

## References

[1] C. Boutilier. Learning conventions in multiagent stochastic domains using likelihood estimates. *Proc. 12th Intl. Conf. Uncertainty in AI*, pp.106–114, Portland, OR, 1996.

[2] C. Boutilier. Planning, learning and coordination in multiagent decision processes. *Proc. 6th Conf. Theor. Aspects of Rationality and Knowledge*, pp.195–210, Amsterdam, 1996.

[3] G. W. Brown. Iterative solution of games by fictitious play. In T. C. Koopmans, editor, *Activity Analysis of Production and Allocation*. Wiley, New York, 1951.

[4] C. Claus and C. Boutilier. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. *AAAI-97 Work. Multiagent Learning*, pp.13–18, Providence, 1997.

[5] D. Fudenberg and D. M. Kreps. *Lectures on Learning and Equilibrium in Strategic Form Games*. CORE Foundation, Louvain-La-Neuve, Belgium, 1992.

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

[7] J. Hu and M. P. Wellman. Self-fulfilling bias in multiagent learning. *Proc. ICMAS-96*, pp.118–125, Kyoto, 1996.

[8] L. P. Kaelbling, M. L. Littman, A. W. Moore. Reinforcement learning: A survey. *J. Art. Intel. Res.*, 4:237–285, 1996.

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

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

[11] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th Intl. Conf. on Machine Learning*, pp.157–163, New Brunswick, NJ, 1994.

[12] D. Monderer, L. S. Shapley. Fictitious play property for games with identical interests. *J. Econ. Th.*, 68:258–265, 1996.

[13] R. B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, 1991.

[14] H. Robbins and S. Munro. A stochastic approximation method. *Annals Math. Stat.*, 22:400–407, 1951.

[15] T. Sandholm and R. Crites. Learning in the iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995.

[16] S. Sen, M. Sekaran, J. Hale. Learning to coordinate without sharing information. *AAAI-94*, pp.426–431, Seattle, 1994.

[17] Y. Shoham and M. Tennenholtz. Emergent conventions in multi-agent systems: Initial experimental results and observations. *KR-92*, pp.225–231, Cambridge, 1992.

[18] Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. *Proc. AAAI-92*, pp.276–281, San Jose, 1992.

[19] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement learning algorithms. *Machine Learning*, 1998. To appear.

[20] M. Tan. Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents. *Proc. 10th Intl. Conf. on Machine Learning*, pp.330–337, Amherst, MA, 1993.

[21] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[22] G. Weiß. Learning to coordinate actions in multi-agent systems. *Proc. IJCAI-93*, pp.311–316, Chambery, FR, 1993.

[23] R. M. Wheeler and K. S. Narendra. Decentralized learning in Markov chains. *IEEE Trans. Aut. Control*, 31:519–526, 1951.

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