Correlated-Q Learning

Amy Greenwald
Department of Computer Science
Brown University, Box 1910
Providence, RI 02912
amy@cs.brown.edu

Keith Hall
Department of Computer Science
Brown University, Box 1910
Providence, RI 02912
kh@cs.brown.edu

Abstract

Bowling named two desiderata for multiagent learning algorithms: rationality and convergence. This paper introduces correlated-Q learning, a natural generalization of Nash-Q and FF-Q that satisfies these criteria. Nash-Q satisfies rationality, but in general it does not converge. FF-Q satisfies convergence, but in general it is not rational. Correlated-Q satisfies rationality by construction. This papers demonstrates the empirical convergence of correlated-Q on a standard testbed of general-sum Markov games.

Introduction

This paper introduces *correlated-Q* learning, a multiagent learning algorithm that learns equilibrium policies in Markov games, just as Q-learning learns to optimal policies in Markov decision processes. Correlated-Q learning is named for correlated equilibria (Aumann 1974), which generalize Nash equilibria by allowing for possible dependencies in strategic choices. A Nash equilibrium is vector of *independent* probability distributions over actions, in which all agents optimize with respect to one another's probabilities. A correlated equilibrium is a probability distribution over the *joint* space of actions, in which all agents optimize with respect to one another's probabilities, conditioned on their own.

Recently, there have been several attempts to design a multiagent learning algorithm that converges to equilibrium policies in general-sum Markov games. Hu and Wellman (1998) propose an algorithm called Nash-Q that converges to Nash equilibrium policies in restricted classes of Markov games. Littman's (2001) friend-or-foe-Q (FF-Q) algorithm always converges, but foe-Q (*i.e.*, minimax-Q (Littman 1994)) only learns equilibrium policies in (two-player, constant-sum) Markov games that exhibit adversarial equilibria, and friend-Q only learns equilibrium policies in coordination games with uniquely-valued equilibria. Correlated-Q (CE-Q), which strictly generalizes both Nash-Q and FF-Q, converges empirically to correlated equilibrium policies on a standard testbed of Markov games.

This paper is organized as follows. In the next section, we define Markov games. Next, we define multiagent Q-learning, and show how CE-Q, Nash-Q, and FF-Q are all special cases of this generic algorithm. Next, we compare CE-Q learning with Q-learning and FF-Q in grid games. In the following section, we experiment with the same set of algorithms in a soccer-like game. Overall, we demonstrate that CE-Q learning converges to (correlated) equilibrium policies on a standard testbed of general-sum Markov games.

Markov Games

Stochastic games generalize repeated games and Markov decision processes (MDPs). A ***stochastic game*** is a tuple $\langle I, S, (A_i(s))_{s \in S, 1 \le i \le n}, P, (R_i)_{1 \le i \le n} \rangle$, where I is a set of n players, S is a set of states, $A_i(s)$ is the ith player's set of actions at state s, P is a probability transition function that describes state transitions, conditioned on past states and joint actions, and $R_i(s, \vec{a})$ is the ith player's reward for state $s \in S$ and joint actions $\vec{a} \in A(s) = A_1(s) \times \ldots \times A_n(s)$. Stochastic games for which the probability transitions satisfy the Markov property are called *Markov games*: *i.e.*, for $\vec{a}_t = (a_1, \ldots, a_n)_t$, $P[s_{t+1} | s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0] = P[s_{t+1} | s_t, \vec{a}_t]$.

An MDP is a one-player Markov game. Recall Bellman's equations that characterize the optimal state- and action-values for a single agent and an MDP: for $0 \le \gamma < 1$,

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P[s'|s,a] V^*(s') \qquad (1)$$

$$V^*(s) = \max_{a \in A(s)} Q^*(s,a) \qquad (2)$$

In words, the value $Q^*(s,a)$ is the immediate reward obtained at state s for taking action a plus the expected discounted value of the future rewards obtained by following the optimal policy thereafter. The value function $V^*(s)$ at state s is defined as the value that maximizes $Q^*(s,a)$ over all actions a. The actions that maximize $Q^*(s,a)$ at each state s describe the (deterministic) optimal policy π^*: *i.e.*,

$$\pi^*(s) \in \arg \max_{a \in A(s)} Q^*(s,a) \qquad (3)$$

In Markov games, player i's optimal Q-values are defined over states and action-vectors, rather than state-action pairs:

$$Q_i^*(s,\vec{a}) = R_i(s,\vec{a}) + \gamma \sum_{s'} P[s'|s,\vec{a}]V_i^*(s') \qquad (4)$$

Intuitively, the notion of state-value function also carries over from MDPs to Markov games. But the obvious analogue of Eq. 2, in which all players maximize their respective rewards with respect to one another's actions is not adequate, since (deterministic) actions that satisfy these simultaneous equations need not exist. (As a consequence, Markov games need not exhibit deterministic equilibrium policies: *e.g.*, *Rochambeau*—Rock-paper-scissors.)

Several alternative definitions of the value function have been proposed. Littman (1994) studied two-player, zero-sum Markov games and the minimax value function (von Neumann & Morgenstern 1944). Let $\Sigma_i(s)$ be the probabilistic action space of player i at state s. Now

$$V_1^*(s) = \max_{\sigma_1\in\Sigma_1(s)} \min_{a_2\in A_2(s)} Q_1^*(s,\sigma_1,a_2) = -V_2^*(s) \qquad (5)$$

$Q^*(s,\sigma_1,a_2) = \sum_{a_1\in A_1}\sigma_1(a_1)Q^*(s,a_1,a_2)$. At the opposite extreme, Littman's friend-Q (2001) value function is suited to coordination games—games for which all the players' reward functions are equivalent—with uniquely-valued equilibria:

$$V_i^*(s) = \max_{\vec{a}\in A(s)} Q_i^*(s,\vec{a}) \qquad (6)$$

For the general case of n-player, general-sum games, Hu and Wellman (1998) proposed the following definition of the value function:

$$V_i^*(s) \in \text{NASH}_i(Q_1^*(s),\ldots,Q_n^*(s)) \qquad (7)$$

where $\text{NASH}_i(R_1,\ldots,R_n)$ denotes the ith player's reward according to some Nash equilibrium in the general-sum game determined by reward matrices $r_1,\ldots,r_n$. Note that existence in this case, which is implied by Nash's theorem (1951) relies on probabilistic actions (see Filar and Vrieze (1996)). This definition generalizes the minimax value function, since Nash equilibria and minimax strategies coincide in zero-sum games. But this value function need not be well-defined: in general there are multiple Nash equilibrium policies.

We propose an alternative definition of the value function in Markov games:

$$V_i^*(s) \in \text{CE}_i(Q_1^*(s),\ldots,Q_n^*(s)) \qquad (8)$$

where $\text{CE}_i(R_1,\ldots,R_n)$ denotes the ith player's reward according to some ***correlated equilibrium*** in the general-sum game determined by the rewards $R_1,\ldots,R_n$. Eq. 8 generalizes Eq. 7, since a Nash equilibrium is a correlated equilibrium that can be factored into independent distributions over each individual player's action space. Thus, equilibria that are consistent with Eq. 8 exist, but this value function, too, need not be well-defined.

Correlated Equilibrium

In contrast to Nash equilibria, for which no efficient method of computation is known, correlated equilibria can be computed easily via linear programming. As an example, consider an arbitrary state in a two-player, two-action, general-sum Markov game, where rewards are as in the game of "Chicken":

	L	R
T	6,6	2,7
B	7,2	0,0

The linear program that computes the correlated equilibrium that maximizes the sum of the players' rewards in this game is given by: $\max 12\pi_{TL} + 9\pi_{TR} + 9\pi_{BL}$, subject to the probability constraints $\pi_{TL} + \pi_{TR} + \pi_{BL} + \pi_{BR} = 1$ and $\pi_{TL}, \pi_{TR}, \pi_{BL}, \pi_{BR} \geq 0$, and four rationality constraints that characterize the players disincentive to deviate:

$$\begin{array}{lcl} -1\pi_{TL} + 2\pi_{TR} \geq 0 & \qquad & -1\pi_{TL} + 2\pi_{BL} \geq 0 \\ 1\pi_{BL} - 2\pi_{BR} \geq 0 & & 1\pi_{TR} - 2\pi_{BR} \geq 0 \end{array}$$

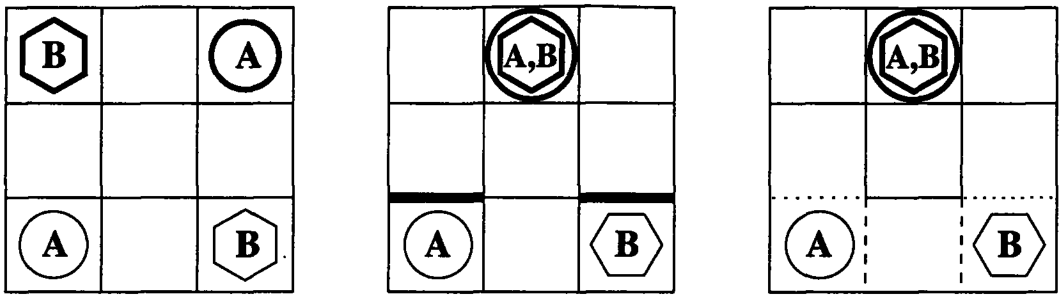
These constraints have a natural interpretation via conditional probabilities. Let $\pi(y|x)$ denote the conditional probability of y given x, and let $\pi(y) = \sum_x \pi(y|x)$ denote the marginal probability of y. Now for players A and B, $\pi_{TL} = \pi(L|T)\pi(T)$ and $\pi_{TR} = \pi(R|T)\pi(T)$. Thus, the first constraint can be restated as $-1\pi(L|T) + 2\pi(R|T) \geq 0$, or equivalently, $6\pi(L|T) + 2\pi(R|T) \geq 7\pi(L|T) + 0\pi(R|T)$. Intuitively, the expected reward to the row player of action T is at least that of action B whenever he in fact plays action T. The other three constraints have analogous interpretations. This linear program has value 10.5, with $\pi_{TL} = 0.5$, $\pi_{TR} = 0.25$, $\pi_{BL} = 0.25$, and $\pi_{BR} = 0$.

Multiagent Q-Learning

```
MULTIQ(MarkovGame, $\gamma$, $\alpha$, $S$, $T$)
  Inputs       discount factor $\gamma$
               learning rate $\alpha$
               decay schedule $S$
               total training time $T$
  Output       action-value functions $Q_i^*$
  Initialize   $s, a_1, \ldots, a_n$ and $Q_1, \ldots, Q_n$

for $t = 1$ to $T$
    1.  simulate actions $a_1, \ldots, a_n$ in state $s$
    2.  observe rewards $r_1, \ldots, r_n$ and next state $s'$
    2.  for $i = 1$ to $n$
        (a)  compute $V_i(s')$
        (b)  update $Q_i(s, a_1, \ldots, a_n)$
               i.  $Q_i(s, a_1, \ldots, a_n) =$
        $(1-\alpha)Q_i(s, a_1, \ldots, a_n) + \alpha[r_i + \gamma V_i(s')]$
    4.  agents choose actions $a_1', \ldots, a_n'$
    5.  $s = s'$, $a_1 = a_1'$, ..., $a_n = a_n'$
    6.  decay $\alpha$ according to $S$
```

Table 1: Multiagent Q-Learning. .



(a) Grid Game 1 (b) Grid Game 2 (c) Grid Game 3

Figure 1: Grid games. Initial States.

In principle, it is straightforward to generalize dynamic programming and reinforcement learning from MDPs to Markov games. A template for multiagent Q-learning is presented in Table 1. In this generic formulation, the means of computing the value function V_i is not made explicit (step 3(a)). Littman's FF-Q algorithm computes V_i according to either Eq. 5 or Eq. 6, as appropriate. Hu's and Wellman's Nash-Q algorithm computes V_i according to Eq. 7. *Correlated-Q* computes V_i in accordance with Eq. 8, using, for example, the utilitarian objective function described above, or perhaps an egalitarian or republican objective function described presently.

In this paper, we report on simulation experiments of a variant of correlated-Q learning that we call ***utilitarian***, since we define the value function in terms of a correlated equilibrium that maximizes the ***sum*** of the players' rewards. Thus, $\text{CE}_i(Q_1^*(s),\ldots,Q_n^*(s)) =$

$$\left\{\sum_{\vec{a}\in A}\sigma^*(\vec{a})Q_i(s,\vec{a}) \mid \sigma^* \text{ satisfies Eq. } 10\right\} \quad (9)$$

where

$$\sigma^* \in \arg\max_{\sigma\in\text{CE}} \sum_{\vec{a}\in A}\sigma(\vec{a})\left(\sum_{i\in I} Q_i(s,\vec{a})\right) \quad (10)$$

In ongoing work, we are also experimenting with correlated equilibria that maximize the ***minimum*** of the players' rewards:

$$\sigma^* \in \arg\max_{\sigma\in\text{CE}} \sum_{\vec{a}\in A}\sigma(\vec{a})\left\{\min_{i\in I} Q_i(s,\vec{a})\right\} \quad (11)$$

and correlated equilibria that maximize the ***maximum*** of the players' rewards:

$$\sigma^* \in \arg\max_{\sigma\in\text{CE}} \sum_{\vec{a}\in A}\sigma(\vec{a})\left\{\max_{i\in I} Q_i(s,\vec{a})\right\} \quad (12)$$

We call these latter two algorithms ***egalitarian*** and ***republican*** correlated-Q learning, respectively.

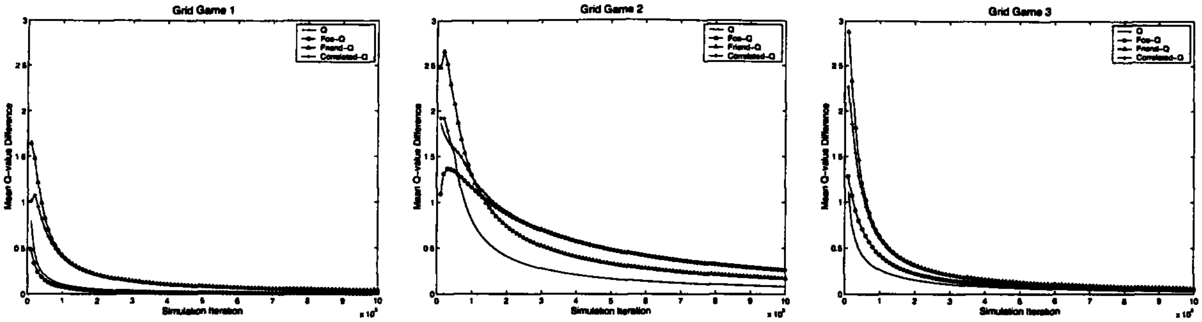
Grid Games

The first set of experimental results we report pertain to grid games (Hu & Wellman 2000). Fig. 1 depicts the initial states of the three games of interest. In grid game 1 (GG1), there are two agents and two goals. The agents' action set includes one step in any of the four compass directions. Actions are executed simultaneously. If both agents attempt to move into the same square, they cannot; instead, they both lose 1 point in GG1 and GG2, and 50 points in GG3. If ever an agent reaches its goal, it scores 100 points and the game is over. Note that it is possible for both agents to score 100 points since moves are executed simultaneously.

Other than the board setup, grid game 2 (GG2) is identical to GG1. In GG2, there are two agents, one goal, and two barriers: if an agent attempts to move through one of the barriers, then with probability 1/2 this move fails. Finally, we introduce grid game 3 (GG3), where like GG2 there is one goal, but in addition there are bonus points to be earned: (i) an agent that successfully enters the center square at the bottom of the grid earns 25 bonus points, or if both agents simultaneously move from their initial positions up the sides, they both earn 20 bonus points. The equilibrium policies of these games are described presently.

In all three of these grid games there exist ***deterministic*** equilibrium policies for both agents. In GG1, there are several pairs of deterministic equilibrium policies in which the agents coordinate their behavior (see Hu and Wellman (2000) for graphical depictions), all of which yield equivalent rewards. In GG2, there are exactly two deterministic equilibrium policies: one agent moves up the center and the other attempts to pass through the barrier, and the same again with the agents' roles reversed. Note that these equilibria are asymmetric: the agent that moves up the center scores 100, but the agent that attempts to move through the barrier scores only 50 on average. The deterministic equilibrium policies of GG2 carry over to GG3.

In addition, all the grid games exhibit ***nondeterministic*** correlated (and Nash) equilibrium policies. For example, in GG2, there exists a continuum of symmetric, nondeterministic, correlated equilibrium poli-



(a) Grid Game 1 (b) Grid Game 2 (c) Grid Game 3

Figure 2: Convergence in the grid games: all algorithms are converging.

Grid Games	GG1		GG2		GG3	
Algorithm	Score	Games	Score	Games	Score	Games
Q	100,100	2500	49,100	3333	100,125	3333
Foe-Q	0,0	0	67,68	3003	120,120	3333
Friend-Q	100, 100	2500	$-\infty, -\infty$	0	$-\infty, -\infty$	0
Correlated-Q	100,100	2500	50,100	3333	117,117	3333

Table 2: Grid Games played repeatedly, allowing 10^4 moves. The number of games played varied with the agents' policies: some move directly to the goal, while others digress.

cies: *i.e.*, for all probabilities p, with probability p one agent moves up the center and the other attempts to pass through the barrier, and with probability $1 - p$ the agents' roles are reversed. More interestingly, in GG3, there exist symmetric, nondeterministic, correlated equilibrium policies in which both agents move up the sides with high probability and each of the deterministic equilibria is played with low probability (assuming utilitarian objectives).

Our experiments reveal that correlated-Q, foe-Q, friend-Q ($\alpha \to 0.001$ and $\gamma = 0.9$.) and on-policy Q-learning (*i.e.*, ϵ-greedy, with $\epsilon \to 0.001$, $\alpha \to 0.001$, and $\gamma = 0.9$) all converge empirically in the three grid games. Littman (2001) proves that FF-Q converges in general-sum Markov games. Fig. 2 shows that in fact ordinary Q-learning (on-policy) and correlated-Q (off-policy) also converge in these games.

The values plotted in Fig. 2 are computed as follows. The error ERR_i^t at time t for agent i is the difference between $Q(s, \vec{a})$ at time t and $Q(s, \vec{a})$ at time $t - 1$: $\text{ERR}_i^t = |Q_i^t(s, \vec{a}) - Q_i^{t-1}(s, \vec{a})|$. The values on the x-axis represent time, and the corresponding y-values are the means of the distributions of the errors ERR_i^t for all $t = 0, \ldots, x$. This mean is converging to zero for all algorithms in all grid games.[1]

Since all the learning algorithms converge, we played the equilibrium policies that the agents learned against one another. Our results are depicted in Table 2. On-policy Q-learning is successful in grid games: it consistently converges to equilibrium policies in which the two agents coordinate their behavior perfectly. In GG1, this leads to symmetric scores, but in GG2 and GG3 their policies and their scores are asymmetric.

Foe-Q learners perform poorly in GG1. Rather than progress toward the goal, they cower in the corners, avoiding collisions, but simultaneously avoiding the goal. In GG2 and GG3, the principle of avoiding collisions leads both foe-Q learners straight up the sides of the grid. These are not equilibrium policies, since both agents have an incentive to deviate to the center, but they do yield reasonable scores in GG2, and Pareto optimal scores in GG3.

Friend-Q learning can perform even worse than foe-Q learning in GG1. This result may appear surprising, since GG1 satisfies the conditions under which friend-Q is guaranteed to converge to equilibrium policies. Indeed, friend-Q learns Q-values that correspond to equilibrium policies, but during play, friends lack the ability to coordinate their policy choices. Whenever one friend chooses a policy that collides with the policy of its so-called friend, both agents obtain infinitely negative scores in one never-ending game. But friend-Q learners do sometimes get lucky and choose coordinated equilibrium policies, as the numbers in Table 2 indicate. Friend-Q's performance is always poor in GG2 and GG3: both friends learn to play the equilibrium policy that uses the center passage, which causes friends to collide repeatedly.

In our implementation of the learning phase of CE-Q, a central coordinator computes an equilibrium policy for each agent. Similarly, in the testing phase, CE-Q

[1] In fact, the actual Q-value differences are converging at all state-action pairs. The mean is simply a useful statistic for summarizing this collective behavior.

play is coordinated. Thus, unlike friend-Q, correlated-Q learners coordinate their play in the grid games. The present comparison is therefore unfair; but in our conclusions, we allude to ongoing research geared towards eliminating CE-Q's dependence on this coordinator.

In GG1, CE-Q learns Q-values that coincide exactly with those of friend-Q (*i.e.*, those Q-values associated with equilibrium policies). CE-Q learns policies much like ordinary Q-learners in GG2. In both GG1 and GG2, utilitarian correlated-Q learning is indifferent between all correlated equilibrium policies, deterministic and nondeterministic, since they all yield equal sums of rewards. In GG3, however, CE-Q learns the particular nondeterministic correlated equilibrium policies that yield symmetric scores, because the sum of rewards at this equilibrium exceeds that of any deterministic equilibrium policies. Consequently, the sum of the scores of CE-Q exceeds that of Q-learning. CE-Q's rewards do not exceed the sum of the foe-Q learners' scores, however; but the foe-Q learners are not rational—they ignore the fact that the reward for using the center passage exceeds that of moving up the sides, given that one's opponent is moving up the side. Utilitarian CE-Q learning converges to rational policies that maximize the sum of the agents' rewards.

Like Nash-Q, correlated-Q learning strictly generalizes friend-Q, since it converges to precisely the same Q-values as friend-Q in games where friend-Q converges. In the next section, we show that correlated-Q learning also appears to generalize foe-Q.

Soccer Game

The grid games are general-sum games for which there exist deterministic equilibria. In this section, we consider soccer (Littman 1994), a zero-sum game for which there do not exist deterministic equilibrium policies.

The soccer field is a grid (see Fig. 3). The circle represents the ball. There are two players, whose possible actions are N, S, E, W, and stick. The agents' actions are executed in random order. If this sequence of actions causes the players to collide, then neither moves. But if the player with the ball moves *second*, then the ball changes possession.[2] If the player with the ball moves into a goal, then he scores +100 if it is in fact his own goal and the other player scores −100, or he scores −100 if it is the other player's goal and the other player scores +100. In either case, the game ends.

In this simple soccer game, there do not exist deterministic equilibrium policies, since at some states there do not exist deterministic equilibria. For example, at the state depicted in Fig. 3 (hereafter, state s), any deterministic policy for player B is subject to indefinite blocking by player A. But if player B employs a nondeterministic policy, then player B can hope to pass player A on his next move.

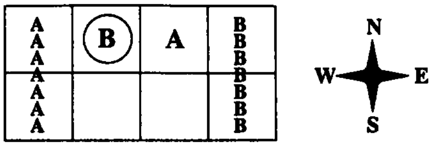


Figure 3: Soccer Game. State s.

We experimented with the same set of algorithms in this soccer game as we did in the grid games. As expected, FF-Q converges at all state-action pairs. Correlated-Q also converges everywhere. In fact, CE-Q learns Q-values (and policies) that coincide exactly with those of foe-Q. But Q-learning does not converge. Intuitively, the rationale for this outcome is clear: Q-learning seeks deterministic optimal policies, but in this game no such policies exist.

Fig. 4 presents an example of a state-action pair at which Q-learning does not converge. In this figure, the values on the x-axis represent time, and the corresponding y-values are the error terms $\text{ERR}_i^t = |Q_i^t(s,\vec{a}) - Q_i^{t-1}(s,\vec{a})|$. The error values shown in Figs. 4(a), (b), and (c) reflect player A's Q-values corresponding to state s, with player A taking action S and player B sticking. These three graphs, corresponding to CE-Q, foe-Q, and friend-Q, respectively, depict converging sequences of error values.

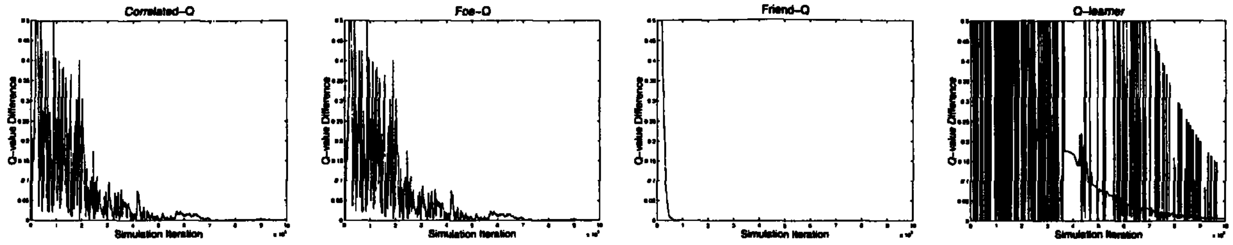
Q-learning agents compute Q-values for each of their own possible actions, ignoring their opponents' actions. The error values shown in Fig. 4(d) reflect player A's Q-values, corresponding to state s and action S. In this figure, although the Q-value differences are decreasing, they are not converging. They are decreasing only because the learning rate $\alpha \to 0.001$. At all times, the amplitude of the oscillations in error values is as great as the envelope of the learning rate.

At state s, CE-Q and foe-Q converge to nondeterministic policies for both players, where each one randomizes between sticking and heading south.

Friend-Q, however, converges to a deterministic policy for player B at state s, namely E. Learning according to friend-Q, player B (fallaciously) anticipates the following sequence of events: player A sticks at state s, and then player A takes action E. Thus, by taking action E, player B passes the ball to player A, with the intent that player A score for him. Player A is indifferent among her actions, since she assumes player B plans to score a goal for her immediately.

In this soccer game, Q-learning does not converge. The friend-Q algorithm converges but its policies are irrational. Correlated-Q learning, however, converges to the same solution as foe-Q learning—the Q-values learned by the two algorithms are identical. Thus, CE-Q learns minimax equilibrium·policies in this two-player, zero-sum game.

[2] In other words, if the player without the ball moves into the player with the ball, attempting to steal the ball, he cannot. But if the player with the ball moves into the player without the ball, the former loses the ball to the latter. This form of the game is due to Littman (1994).



(a) Correlated-Q (b) Foe-Q (c) Friend-Q (d) Q-learning

Figure 4: Convergence in the soccer game. All algorithms—except Q-learning—converge.

Conclusion

Bowling (2001) put forth two desiderata for multiagent learning algorithms: rationality and convergence. Rational agent policies are optimal with respect to one another: *i.e.*, there is no incentive for any agent to deviate from the prescribed behavior. Convergence in this setting refers to Q-values. In this paper, we introduced correlated-Q learning, a natural generalization of existing algorithms that satisfies these criteria. Nash-Q satisfies rationality, but in general it does not converge. FF-Q satisfies convergence, but in general it is not rational. Correlated-Q satisfies rationality by construction, and the experiments described in this paper demonstrate empirical convergence on a standard testbed of general-sum Markov games. It remains to prove convergence of correlated-Q learning in general.

One awkward aspect of correlated-Q learning is that agents must collaborate to jointly learn equilibrium policies. However, there exist adaptive algorithms for individual agents which generate joint play that converges to correlated equilibria (Foster & Vohra 1997). In ongoing work, we are implementing correlated-Q learning without explicit collaboration by replacing the linear programming call with such an adaptive procedure. Similarly, we are implementing an adaptive version of minimax-Q by replacing its linear programming call with an adaptive algorithm that converges to minimax strategies (Freund & Schapire 1996). Preliminary tests suggest that such adaptive algorithms are capable of learning correlated equilibrium policies without relying on any explicit coordination device.

Acknowledgments

The authors are grateful to Michael Littman for inspiring discussions.

References

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1:67–96.

Bowling, M. 2001. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 1021–1026.

Filar, J., and Vrieze, K. 1996. *Competitive Markov Decision Processes.* New York: Springer Verlag.

Foster, D., and Vohra, R. 1997. Regret in the online decision problem. *Games and Economic Behavior* 21:40–55.

Freund, Y., and Schapire, R. 1996. Game theory, online prediction, and boosting. In *Proceedings of the 9th Annual Conference on Computational Learning Theory.* ACM Press. 325–332.

Hu, J., and Wellman, M. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 242–250.

Hu, J., and Wellman, M. 2000. Experimental results of multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 407–414.

Littman, M. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of Eleventh International Conference on Machine Learning*, 157–163.

Littman, M. 2001. Friend or foe Q-learning in general-sum markov games. In *Proceedings of Eighteenth International Conference on Machine Learning*, 322–328.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54:286–295.

von Neumann, J., and Morgenstern, O. 1944. *The Theory of Games and Economic Behavior.* Princeton: Princeton University Press.