Reaching Pareto Optimality in Prisoner's Dilemma Using Conditional Joint Action Learning

Dipyaman Banerjee and Sandip Sen
Department of Mathematics & Computer Science
The University of Tulsa
{dipu,sandip}@utulsa.edu

Abstract

We consider a repeated Prisoner's Dilemma game where two independent learning agents play against each other. We assume that the players can observe each others' action but are oblivious to the payoff received by the other player. Multiagent learning literature has provided mechanisms that allow agents to converge to Nash Equilibrium. In this paper we define a special class of learner called a conditional joint action learner (CJAL) who attempts to learn the conditional probability of an action taken by the other given its own action and uses it to decide its next course of action. We prove that when played against itself, if the payoff structure of Prisoner's Dilemma game satisfies certain conditions, using a limited exploration technique these agents can actually learn to converge to the Pareto optimal solution that dominates the Nash Equilibrium, while maintaining individual rationality. We analytically derive the conditions for which such a phenomenon can occur and have shown experimental results to support our claim.

Keywords: Multiagent Learning, Game theory, Prisoner's Dilemma.

Introduction

The problem of learning in multi agent system has attracted increasing attention in the recent past (Wellman & Hu 1998; Hu & Wellman 2003; Littman 1994; Claus & Boutilier 1997; Littman 2001; Matsubara, Noda, & Hiraki 1996; Bowling & Veloso 2004; Littman & Stone 2001). As a result a number of learning mechanisms were discovered which were proved to converge to Nash Equilibrium under certain conditions (Hu & Wellman 2003; Littman 1994; Conitzer & Sandholm 2003; Littman & Stone 2005; Bowling & Veloso 2004). However many of these mechanisms assume complete transparency of payoffs for both the players, which may not be always possible in real environments. Moreover, convergence to Nash Equilibrium was assumed to be a desirable criteria for these algorithms, which in many cases may not be Pareto-optimal and may lead to poorer payoff for the players. Under imperfect conditions, where a player can observe the action of all other players but not their payoffs the learning problem is even more difficult as

the agents have less information to compute their optimal action. Though some independent reinforcement learning algorithms have achieved success in the past (Sekaran & Sen 1994; Weiß 1993) the non-stationary nature of the environment precludes the guarantee of convergence for single agent reinforcement learning mechanism.

Claus and Boutilier (Claus & Boutilier 1997) have shown the dynamics of reinforcement learning in a cooperative game. They described two kind of learners: Independent learners and Joint Action learners. An independent learner assumes the world to be stationary and ignores the presence of other players. However, a joint action learner computes the joint probabilities of different actions taken by other players and use them to calculate the expected value of its own actions. Unfortunately, JAL-s do not perform dramatically better than ILs as the Q-values associated with the actions of a JAL learner degenerate to that learned by an IL learner (Claus & Boutilier 1997; Mundhe & Sen 1999). We believe that the primary impediment to JAL's performance improvement is their assumption that actions of different agents are uncorrelated, which is not the case in general. In this paper we present a new learner which understands and tries to use the fact that its own actions affect the action of other agents. Instead of marginal probabilities it uses conditional probabilities of the actions taken by other agents given its own actions, to compute the expected value of its action choices. From now on we will refer to this class of learners as Conditional Joint Action Learner or CJAL.

In self-play, CJAL learners do not to converge to equilibrium every time. On the other hand, they guarantee convergence to a Pareto-optimal outcome under certain restrictions over the payoff structure. We in this paper primarily focus on the game of prisoner's dilemma between two players and derive the conditions for which the players will converge to a Pareto optimal solution. We also describe the effect of exploration strategy on these conditions. We show that under these restrictions a combination of purely explorative and purely exploitative exploration will always eventually lead to Pareto optimality. We have also used an ϵ greedy strategy and derived an upper bound for ϵ, above which agents can never converge to Pareto-optimality. We support our theoretical analysis with experimental results.

The rest of the paper is organized as follows: section 2 de-

scribes the Prisoner's Dilemma game and the CJAL learning algorithm. In section 3 we prove the conditions for reaching Pareto-optimality in prisoner's dilemma for CJAL learners when played against itself and discuss the effect of exploration on the algorithm. In section 4 we provide experimental results and finally in section 5 we conclude the paper and give directions to future work.

CJAL Learning Mechanism

Prisoner's Dilemma

In a 2-player Prisoner's Dilemma (PD) game, two agents play against each other where each agent has a choice of two actions namely, cooperate(C) or defect(D). The bimatrix form of this single stage game is shown below:

	C	D
C	R,R	S,T
D	T,S	P,P

and the following inequalities hold:

$$T > R > P > S$$

and

$$2R > T + S$$

Under these conditions the dominant strategy for a player is to defect and so the defect-defect action combination is a dominant strategy equilibrium and the only Nash Equilibrium. But this is a Pareto suboptimal solution as the cooperate-cooperate action combination dominates this Nash Equlibrium. So the paradox is, even there exists an action-combination which has a better payoff, the players still chose the suboptimal action combination using individual rationality. We claim that under imperfect condition as described above a CJAL learner when played against itself can actually find this cooperate-cooperate solution which maximizes the social welfare and can stick with it given certain payoff structure (still satisfying the inequalities), and suitable exploration techniques.

In this paper we concentrate on two-player games where the players play with one another repeatedly and tries to learn the optimal action choice which maximize their expected utility. We would like to point out that this problem is different from a repeated Prisoner's Dilemma game. Though the players interact repeatedly, they are unaware about the duration for which the game will be played. In other words they ignore the future discounted rewards while computing their expected utility and choose its optimal action only based on the history of interactions they had in the past. This gaming environment is different from a repeated Prisoner's Dilemma problem as dealt by Sandholm et. al (Sandholm & Crites 1995) where agents use the information about duration of the game to compute their expected utility. Also note that, the players have no clue that it is a Prisoner's Dilemma game as they are oblivious to each others' payoffs and are only interested in maximizing individual payoffs.

CJAL Learning

We assume a set S of 2 agents where each agent $i \in S$ has a set of action A_i. The agents repeatedly play a stage game and in every iteration each agent chooses an action $a_i \in A_i$. Let us denote the expected utility of an agent i at time t for an action a_i as $E_t^i(a_i)$. In case of Prisoner's Dilemma $A_i = \{C, D\}$ and is same for both the agents.

We now introduce some notations and definitions to build the framework for CJAL learning. We denote the probability that agent i plays action a_i at iteration t as $Pr_t^i(a_i)$. We also denote the conditional probability that the other agent(j) will play a_j given that i^{th} agent plays a_i at time t as $Pr_t^i(a_j|a_i)$. The joint probability of an action pair (a_i, a_j) at time t is given by $Pr_t(a_i, a_j)$. Each agent maintains a history of interactions at any time t as

$$H_t^i = \bigcup_{\substack{a_i \in A_i \\ a_j \in A_j}} n_t^i(a_i, a_j)$$

where $n_t^i(a_i, a_j)$ denotes the number of times the joint action (a_i, a_j) being played till time t from the beginning. We define

$$n_t^i(a_i) = \sum_{a_j \in A_j} n_t^i(a_i, a_j)$$

Definition 1: *A bimatrix game consists of a pair of Matrices, (M_1, M_2), each of size $|A_1| \times |A_2|$ for a two-agent game, where the payoff of the ith agent for the joint action (a_1, a_2) is given by $M_i(a_1, a_2)$, $\forall (a_1, a_2) \in A_1 \times A_2$, $i = 1, 2$.*

Definition 2: *A CJAL learner is an agent i who at any time instant t chooses an action $a_i \in A_i$ with a probability $f_t(E_t^i(a_i))$ where*

$$\sum_{a_i \in A_i} f_t(E_t^i(a_i)) = 1$$

and

$$E_t^i(a_i) = \sum_{a_j \in A_j} M_i(a_i, a_j) Pr_t^i(a_j|a_i)$$

where a_j is the action taken by the other agent.

Using results from probability theory we can rewrite the expression for expected utility as

$$E_t^i(a_i) = \sum_{a_j \in Aj} M_i(a_i, a_j) \frac{Pr_t(a_i, a_j)}{Pr_t^i(a_i)} \qquad (1)$$

If we define the probability of an event as the fraction of times the event occurred in the past then equation 1 takes the form

$$E_t^i(a_i) = \sum_{a_j \in A_j} M_i(a_i, a_j) * \frac{n_{t-1}^i(a_i, a_j)}{n_{t-1}^i(a_i)} \qquad (2)$$

So, unlike JAL a CJAL learner does not assume that the probability of the other player's taking an action is independent of its own action. A CJAL tries to learn the correlation between its actions and the other agents actions and

uses conditional probability instead of marginal probability to calculate the expected utility of an action. In other words, a CJAL learner splits the marginal probability of an action a_j taken by the other player in conditional probabilities : $Pr_t^i(a_j|a_i) \; \forall a_i \in A_i$ and considers them as the probability distribution associated with the joint action event (a_i, a_j). An intuitive reasoning behind this choice of probability distribution can be obtained by considering each agent's viewpoint. Imagine that each agent views this simultaneous move game as a sequential move game where he is the first one to move. Then in order to calculate the expected utility of its action it must try to find the probability of the other player's action given its own action, which is basically the conditional probability we described above.

We now discuss the learning mechanism used to update the expected utility values. We would like to point out that it would be unreasonable to use a single-agent Q-learning scheme for CJAL to update the expected utility of its individual actions. Because using single agent Q-learning to estimate payoff from a joint action ignores the correlation among actions of the participating agents and hence will be similar to the Q-values learned by an independent learner. Instead we use a joint action Q-learning for CJAL to estimate the expected utilities associated with different joint actions.

So we rewrite the equation 2 as :

$$E_t^i(a_i) = \sum_{a_j \in A_j} Q_t^i(a_i, a_j) * \frac{n_t^i(a_i, a_j)}{n_t^i(a_i)} \qquad (3)$$

where

$$Q_t^i(a_i, a_j) = Q_{t-1}^i(a_i, a_j) + \alpha(M_i(a_i, a_j) - Q_{t-1}^i(a_i, a_j)) \qquad (4)$$

α being the learning rate. Note that, if the reward associated with a particular joint action is deterministic (which is the case for Prisoner's Dilemma game we consider) equation 3 degenerates to equation 2. So from now on in our analysis we will use equation 2 as the equation used to calculate expected utility.

Dynamics of CJAL Learning

Now that we've described the learning mechanism, we try to capture the dynamics of such a mechanism when played against itself. We consider two CJAL learner's to play the Prisoner's Dilemma game against each other. We try to predict analytically the sequence of actions they would take with time.

Exploration Techniques

We use a combination of explorative and exploitative exploration techniques in this paper. We assume that the agents explore each action randomly for some initial time periods N and then uses an ϵ-greedy exploration. Mathematically, $\forall i \in 1, 2$ and

$$\forall a_i \in A_i$$

if $t < N$

$$Pr_t^i(a_i) = \frac{1}{|A_i|}$$

and for $t > N$ let

$$a^* = arg \max_{a_i \in A_i} (E_{t-1}^i(a_i))$$

then,

$$Pr_t^i(a^*) = 1 - \epsilon$$

and,

$$\forall a_i \in A - \{a^*\}$$
$$Pr_t^i(a_i) = \frac{\epsilon}{|A - a^*|}$$

Analysis of CJAL Learning Dynamics

In this setting let us intuitively examine the emergent playing behavior for a two-player Prisoner's Dilemma game if agents take purely greedy actions ($\epsilon = 0$) after the initial N periods. For Prisoner's Dilemma we have $A_i = \{C, D\}$, $i = 1, 2$. Let us also denote $M_i(C, C)$ as R, $M_i(C, D)$ as S, $M_i(D, D)$ as P and $M_i(D, C)$ as T. Initially both the agent assumes that the other agent have an equal probability of playing any action. If N is sufficiently large then we may assume that after N iterations all the conditional probabilities will be close to 1/2. So for both the agents $E_N^i(C) = \frac{R+S}{2}$ and $E_N^i(D) = \frac{T+P}{2}$. Under Prisoner's Dilemma conditions then $E_N^i(D) > E_N^i(C)$. Therefore, both the agents will start playing action D. Now as they play action D , $Pr_t^i(C|D)$ will tend to 0 and $Pr_t^i(D|D)$ will tend to 1. However the $Pr_t^i(C|D)$ and $Pr_t^i(C|C)$ will still remain as $\frac{1}{2}$. So eventually the expected utilities will be $E_i(C) = (S + R)/2$ and $E_i(D) = P$

Now if $\frac{S+R}{2} > P$, then the agents will start playing C. As they both start playing C the $Pr_t^i(C|C)$ will reach 1 and $Pr_t^i(D|C)$ will reach 0. So now $E_i(C) = R$ and $E_i(D) = P$. As $R > P$ they would continue to play C and hence will converge to Pareto optimality. In essence, the agents will learn with time that even though state DC is very lucrative and state CD is equally unattractive, they are almost impossible to reach and will play CC instead, reinforcing each others' trust on cooperation.

Unfortunately, the scenario is not so simple if $\epsilon > 0$. We show below that there exist an ϵ_0 s.t. for $\epsilon > \epsilon_0$, CC can never be achieved. We prove these results below.

Theroem 1: *If the agents randomly explore for a finite time interval N and then adopt an ϵ greedy exploration technique then there exists an ϵ_0 such that for $\epsilon > \epsilon_0$ CJAl can never converge to Pareto-optimality in a game of Prisoner's Dilemma.*

Theorem 2: *If the agents randomly explore for a finite time interval N and then adopts a complete greedy exploration technique ($\epsilon = 0$) then CJAL will always converge to Pareto-optimality if $(R + S) > 2P$ for a Prisoner's Dilemma game.*

Proof: Let us assume that out of N initial interactions each of the four joint actions has been played $\frac{N}{4}$ times (which is a fair assumption if N is sufficiently large). So the agents will play C with probability ϵ and D with probability
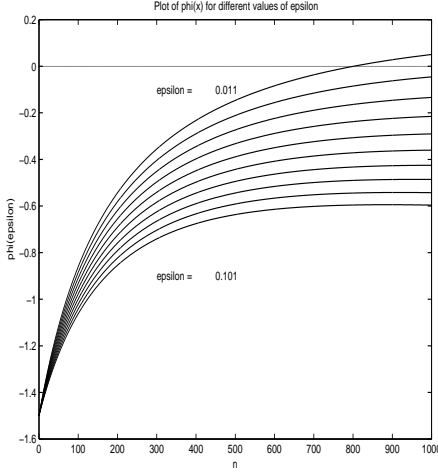
Figure 1: Plot of $\phi(\epsilon)$ with number of iteration

$(1 - \epsilon)$ as $E_N^i(D) > E_N^i(C)$. We observe that expected values for $n_t^i(a_i, a_j)$ at time t is $t * Pr_t^i(a_i, a_j)$. Using the notations given in section 2 at any time instant $t = N + n$ the expected value of the action C and D will then be respectively:

$$E_{N+n}^i(C) = \frac{\frac{N}{4} + n\epsilon^2}{\frac{N}{2} + n\epsilon}R + \frac{\frac{N}{4} + n\epsilon(1 - \epsilon)}{\frac{N}{2} + n\epsilon}S \qquad (5)$$

Similarly,

$$E_i(D) = \frac{\frac{N}{4} + n(1 - \epsilon)^2}{\frac{N}{2} + n(1 - \epsilon)}P + \frac{\frac{N}{4} + n\epsilon(1 - \epsilon)}{\frac{N}{2} + n(1 - \epsilon)}T \qquad (6)$$

Let us now define a function

$$\phi(\epsilon) = E_t^i(C) - E_t^i(D)$$

Now the agents will choose C if $\lim_{\epsilon \to 0}(\phi(\epsilon))$ is positive. which gives us:

$$\frac{n}{N}(R + S - 2P) > \frac{(P + T) - (R + S)}{2} \qquad (7)$$

Now RHS of inequality 7 is a positive constant under the inequalities of Prisoner's Dilemma and LHS is an increasing function in n. So LHS will eventually be larger if,

$$R + S > 2P \qquad (8)$$

which we stated as the condition in Theorem 2.

Now let us observe the nature of $\phi(\epsilon)$ as we increase ϵ, noting that the maximum value of ϵ is 0.5,

$$\lim_{\epsilon \to 0.5}(\phi(\epsilon)) = \frac{(R + S) - (P + D)}{2} \qquad (9)$$

which is always negative for all values of n, R, S, T, P under the conditions of Prisoner's Dilemma.

From equations 7, 9 we conclude there exists some ϵ_0, $0 < \epsilon_0 < 0.5$ s.t for $\epsilon > \epsilon_0$ expected utility for cooperation can never supersede the utility of defect. Hence,

CJAL will never reach Pareto-optimality, which is the claim of Theorem 1.

Now if $\epsilon < \epsilon_0$, Let us assume after n_0 iterations $E_{n_0}^i(C)$ supersedes $E_{n_0}^i(D)$. At this point $(N + n_0)$ an agent will choose D with probability ϵ and C with probability $1 - \epsilon$. So after $N + n_0 + n$ iterations the expected utilities will be:

$$E_{N+n_0+n}^i(C) =$$
$$\frac{\frac{N}{4} + n_0\epsilon^2 + n(1 - \epsilon)^2}{\frac{N}{2} + n_0\epsilon + n(1 - \epsilon)}R$$
$$+ \frac{\frac{N}{4} + n_0\epsilon(1 - \epsilon) + n(\epsilon)(1 - \epsilon)}{\frac{N}{2} + n_0\epsilon + n(1 - \epsilon)}S \qquad (10)$$

$$E_{N+n_0+n}^i(D) =$$
$$\frac{\frac{N}{4} + n_0(1 - \epsilon)^2) + n\epsilon^2}{\frac{N}{2} + n\epsilon + n_0(1 - \epsilon)}P$$
$$+ \frac{\frac{N}{4} + n_0\epsilon(1 - \epsilon) + n(\epsilon)(1 - \epsilon)}{\frac{N}{2} + n\epsilon + n_0(1 - \epsilon)}T \qquad (11)$$

Substituting $\epsilon = 0$ in equations 10, 11, we observe that

$$E_{N+n_0+n}^i(C) = \frac{\frac{N}{4} + n}{\frac{N}{2} + n}R + \frac{\frac{N}{4} + n}{\frac{N}{2} + n}S$$

and

$$E_{N+n_0+n}^i(D) = E_{N+n_0}^i(D)$$

Now under the conditions of Prisoner's Dilemma, $E_{N+n_0+n}^i(C)$ is an increasing function in n. So $E_{N+n_0+n}^i(C)$ will continue to be greater than $E_{N+n_0+n}(D)$, which is the second claim of Theorem 2.

We plot $\phi(\epsilon)$ for $0.011 < \epsilon < 0.101$, varying n from 0 to 1000 shown in figure 1 and for $R = 3$, $S = 0$, $T = 5$, $P = 1$. We observe that $\phi(\epsilon)$ decreases with increasing values of ϵ and is always negative when ϵ is greater than some particular value.

Experimental Results

In our experiments we allow two CJAL learners play a Prisoner's Dilemma game repeatedly. Each agent has two action choices: cooperate or defect. Agents keep count of all the actions played to compute the conditional probabilities and update their beliefs after every iteration. We experiment with different values for R, S, T, P and used two different exploration techniques namely:

1. Choosing actions randomly for first N iterations and then always choose action with highest estimated payoff.

2. Choosing actions randomly for first N iterations and ϵ-greedy exploration thereafter. i.e. explore randomly with probability ϵ, otherwise choose action with highest estimated payoff. We take the value of N as 400.

We use payoff values such that $R + S > 2P$, $R = 3$, $S = 0$, $T = 5$, $P = 1$. We plot the expected utilities of two actions against the number of iterations in Figure 2. We also
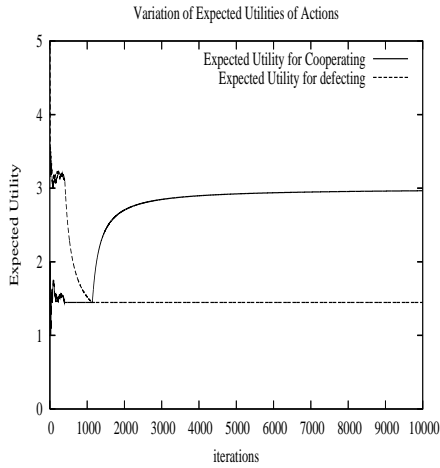
Figure 2: Comparison of Expected Utility when $R+S > 2P$ and $\epsilon = 0$
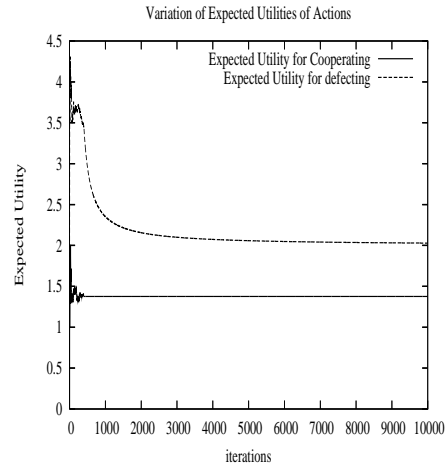

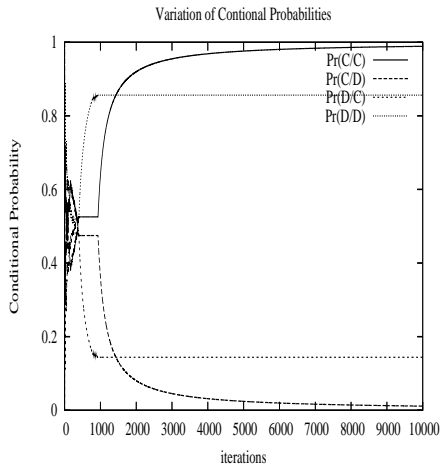Figure 4: Comparison of Expected Utility when $R+S < 2P$ and $\epsilon = 0$

Figure 3: Comparison of Conditional Probability when $R+S < 2P$ and $\epsilon = 0$

Figure 5: Comparison of Expected Utility when $R+S > 2P$ and $\epsilon = 0.1$

compared in Figure 3 the values of four different conditional probabilities mentioned in section 2 and how they vary with time. We observe from Figure 3 that as the players continue to play defect the probabilities of $Pr(D|D)$ increases, but this in turn reduces the expected utility of taking action D where as $Pr(C|C)$ and $Pr(D|C)$ remain unchanged. This phenomena is evident from figure 3 and 2. Around iteration number 1000, expected utility of D falls below that of C and so the agents starts cooperating. As they cooperate $Pr(C|C)$ increases and $Pr(D|C)$ decreases. Consequently, the expected value for cooperating also increases, and hence the agents continue to cooperate.

In the next experiment we continue using the first exploration technique but choose the payoff values such that $R + S < 2P$ ($R = 3$, $S = 0$, $T = 5$, $P = 2$). We plot the expected utilities of two actions against the
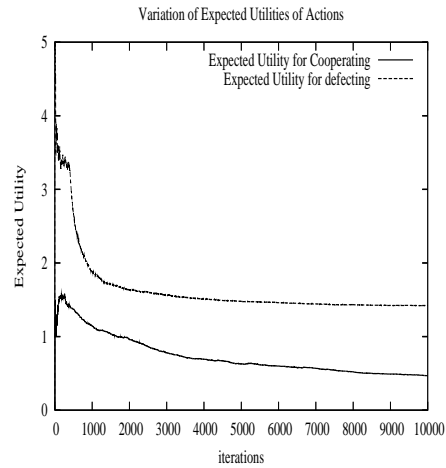
number of iterations. The results are shown in Figure 4. Here we observe due to the condition $R + S < 2P$, though expected utility of defect reduces to the payoff of defect-defect configuration, it still supersedes the expected utility for cooperation. Hence the agents choose to defect and the system converges to the Nash Equilibrium.

In our final experiment we use the second exploration technique taking epsilon value as 0.1 and the same payoff configuration as the first experiment. The results are plotted in figure 5. We observe that though the expected value of defecting reaches below the value of $\frac{R+S}{2}$, due to exploration, $Pr(D|C)$ also increases, which effectively reduces the expected utility of cooperation. In effect, players find it more attractive to play defect, and hence converge to defect-defect option.

Conclusion and Future Work

We described a conditional joint action learning mechanism and analyzed its performance for a 2-player Prisoner's Dilemma Game. Our idea is motivated by the fact that in a multi-agent setting a learner must realize that he is also a part of the environment and his action choices influence the action choices of other agents. We showed both experimentally and analytically that when played against itself under certain restriction on the payoff structure it learns to converge to Pareto-optimality using limited exploration. On the other hand IL or JAL converges to the Nash-equilibrium which is a non-Pareto outcome. We also theoretically derived the conditions for which such a phenomena may occur. In future work, we would like to observe the impact of CJAL for n-person general sum games to deduce the conditions for reaching Pareto-optimality using this learning mechanism. We would also like to observe the performance of CJAL in presence of other strategies such as tit-for-tat, JAL and best response strategies, which does not assume transparency on opponent's payoff.

References

Bowling, M. H., and Veloso, M. M. 2004. Existence of multiagent equilibria with limited agents. *J. Artif. Intell. Res. (JAIR)* 22:353–384.

Claus, C., and Boutilier, C. 1997. The dynamics of reinforcement learning in cooperative multiagent systems. In *Collected papers from AAAI-97 workshop on Multiagent Learning*. AAAI. 13–18.

Conitzer, V., and Sandholm, T. 2003. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In *ICML*, 83–90.

Hu, J., and Wellman, M. P. 2003. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research* 4:1039–1069.

Littman, M. L., and Stone, P. 2001. Implicit negotiation in repeated games. In *Intelligent Agents VIII: AGENT THEORIES, ARCHITECTURE, AND LANGUAGES*, 393–404.

Littman, M. L., and Stone, P. 2005. A polynomial-time nash equilibrium algorithm for repeated games. *Decision Support System* 39:55–66.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 157–163. San Mateo, CA: Morgan Kaufmann.

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 322–328. San Francisco: CA: Morgan Kaufmann.

Matsubara, H.; Noda, I.; and Hiraki, K. 1996. Learning of cooperative actions in multiagent systems: A case study of pass play in soccer. In Sen, S., ed., *Working Notes for the AAAI Symposium on Adaptation, Co-evolution and Learning in Multiagent Systems*, 63–67.

Mundhe, M., and Sen, S. 1999. Evaluating concurrent reinforcement learners. IJCAI-99 Workshop on Agents that Learn About, From and With Other Agents.

Sandholm, T. W., and Crites, R. H. 1995. Multiagent reinforcement learning and iterated prisoner's dilemma. *Biosystems Journal* 37:147–166.

Sekaran, M., and Sen, S. 1994. Learning with friends and foes. In *Sixteenth Annual Conference of the Cognitive Science Society*, 800–805. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Weiß, G. 1993. Learning to coordinate actions in multi-agent systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 311–316.

Wellman, M. P., and Hu, J. 1998. Conjectural equilibrium in multiagent learning. *Machine Learning* 33(2-3):179–200.