Auton Agent Multi-Agent Syst (2007) 15:91–108
DOI 10.1007/s10458-007-0020-8

# Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning

**Dipyaman Banerjee · Sandip Sen**

Published online: 30 April 2007

**Abstract** We consider the learning problem faced by two self-interested agents repeatedly playing a general-sum stage game. We assume that the players can observe each other's actions but not the payoffs received by the other player. The concept of Nash Equilibrium in repeated games provides an individually rational solution for playing such games and can be achieved by playing the Nash Equilibrium strategy for the single-shot game in every iteration. Such a strategy, however can sometimes lead to a Pareto-Dominated outcome for games like Prisoner's Dilemma. So we prefer learning strategies that converge to a Pareto-Optimal outcome that also produces a Nash Equilibrium payoff for repeated two-player, n-action general-sum games. The Folk Theorem enable us to identify such outcomes. In this paper, we introduce the Conditional Joint Action Learner (CJAL) which learns the conditional probability of an action taken by the opponent given its own actions and uses it to decide its next course of action. We empirically show that under self-play and if the payoff structure of the Prisoner's Dilemma game satisfies certain conditions, a CJAL learner, using a random exploration strategy followed by a completely greedy exploitation technique, will learn to converge to a Pareto-Optimal solution. We also show that such learning will generate Pareto-Optimal payoffs in a large majority of other two-player general sum games. We compare the performance of CJAL with that of existing algorithms such as WOLF-PHC and JAL on all structurally distinct two-player conflict games with ordinal payoffs.

**Keywords** Multiagent learning · Game theory · Prisoner's dilemma

D. Banerjee (✉) · S. Sen
Department of Computer Science, University of Tulsa, Tulsa OK, USA
e-mail: dipyaman@gmail.com

S. Sen
e-mail: sandip@utulsa.edu

## 1 Introduction

It has been argued that a learning algorithm playing a repeated game should at least converge in self-play and ensure the safety or minimax value for both the players [1]. To achieve this, researchers have tried to develop algorithms that converge to the Nash Equilibrium of the single-stage game that is being played repeatedly. We refer to such an equilibrium as Nash Equilibrium for Single-Stage Game (NESSG). NESSG is a strategy profile such that no player has any incentive to unilaterally deviate from its own strategy. Also an NESSG for the single-stage game is a Nash Equilibrium strategy profile for the repeated game and it always guarantees safety value for all the players. Due to these reasons, a number of learning mechanisms have been developed which have been proved to converge to NESSG under certain conditions [1;2;6;12;16]. However, there are two problems with choosing NESSG as a solution. First, for a given game, there can be multiple NESSGs and to reach a mutual agreement among the players about choosing one of them can be a non-trivial task. Second, for games such as the Prisoner's Dilemma, the only NESSG outcome is Pareto-Dominated and playing that repeatedly will result in a Pareto-Dominated outcome for the iterated version of the game as well[1]. Hence, instead of trying to converge to an NESSG, we set our learning goal to converge to a solution that is Pareto-Optimal and produces a Nash Equilibrium payoff for the repeated game. We refer to these outcomes as Pareto-Optimal Outcomes sustained by Nash Equilibrium (POSNE). We show in Sect. 3 how the Folk Theorem [9] can help us identify such POSNE outcomes of a repeated game.

Most of the previous learning mechanisms developed for game playing assume complete transparency of the payoffs [6;10;12], i.e., players can observe the payoffs received by all the players after they make their action choices. This may not be possible in a large number of real environments. A more realistic assumption would be to allow the players to observe the actions of all other players but not their payoffs. A game in which players cannot observe opponent's payoffs is known as an *incomplete information game*. In this paper we consider only incomplete information games. In such games, a player cannot compute the opponent's best response to its own strategy, as that will depend on the opponent's payoff structure. This in turn prevents the players from directly computing the Nash Equilibria of the game where every player's strategy is a best response to other players' strategy. We believe, however, that it is still possible to reach mutually preferred outcomes against a class of agents by observing the opponent's action choices over time.

Claus and Boutilier have shown the dynamics of reinforcement learning in a cooperative game [5]. They described two kind of learners: Independent Learners and Joint Action Learners. An Independent Learner (IL) assumes the world to be stationary and ignores the presence of other players. A Joint Action Learner (JAL), however, acknowledges the impact of the other player and computes the joint probabilities of different actions taken by all the players and uses them to calculate the expected utility of its own actions. Unfortunately, JALs do not necessarily perform significantly better than ILs as the $Q$-values associated with the actions of a JAL learner can degenerate to that learned by an IL learner [5;20]. We believe that the primary impediment to JALs performance improvement is their assumption that actions of different agents

[1] An outcome is Pareto-Dominated by another outcome if at least one player prefers the latter and no player prefers the former. An outcome that is not Pareto-dominated is Pareto-Optimal.

are uncorrelated, which is not true in general. In this paper, we present a new learner which understands and uses the fact that its own actions influence the action choices of other agents. Instead of marginal probabilities, it uses conditional probabilities of the actions taken by the opponent given its own actions, to compute the expected utility of its actions. We refer to this class of learners as Conditional Joint Action Learners (CJALs).

In self-play, CJALs do not converge to Nash Equilibrium every time. On the other hand, they converge to a Pareto-Optimal outcome under certain restrictions over the payoff structure. In this paper, we primarily focus on repeated play of the game of Prisoner's Dilemma between two players and derive the conditions for which two CJAL players will converge to a Pareto-Optimal outcome. We also describe the effect of different exploration strategies on these conditions. We show that under these restrictions a combination of purely explorative and purely exploitative exploration will lead to Pareto-Optimal outcomes. We support our claim with experimental results.

The rest of the paper is organized as follows: we discuss relevant game theory concepts and related work in this domain in Sect. 2. In this section, we also show how to identify the POSNE outcomes using the Folk Theorem [9]. Section 3 describes the Prisoner's Dilemma game and the CJAL learning algorithm. In Section 4, we derive the conditions for converging, in self-play, to the mutually cooperative outcome in the Prisoners Dilemma game for CJAL learners. In Section 5 we provide experimental results and finally, in Sect.6, we conclude the paper and suggest directions for future work.

## 2 Background

In this section we review relevant concepts from game theory and discuss the related work from this field. We also show how the Folk Theorem can be used to identify POSNE outcomes.

### 2.1 Relevant game theory concepts

**Stage Game:** A stage game can be defined by a tuple $(S, A_1 \ldots A_{|S|}, U_1 \ldots U_{|S|})$, where $S$ is the set of players. A player $i \in S$ can choose an action $a_i$ from its action set $A_i$. Choosing an action deterministically corresponds to a *pure strategy*. A pure strategy profile $q$ is a set of pure strategies chosen by all the players. We denote the set of all possible pure strategy profiles as $Q$, which is the Cartesian product $\times_{i \in S} A_i$. $Q$ also corresponds to the set of pure strategy outcomes. A *strategy* for a player $i$ is defined as a probability distribution $\pi_i$ over $A_i$ according to which $i$ chooses its actions. Note that a pure strategy is a strategy where one action is chosen with probability 1. A strategy profile $p$ is defined as the set of strategies used by all players: $\{\pi_1, \ldots, \pi_{|S|}\}$. We define $X$ as the set of all possible probability distributions over $Q$. The set of all possible strategy profiles is a subset of $X$ that can be factored into independent probability distributions for the players. The payoff function for a player $i$ is defined as a function $U_i : X \to \mathbb{R}$ that decides the payoff a player would receive for a particular strategy profile. In general, for a strategy profile $\{\pi_1, \ldots, \pi_{|S|}\}$ the payoff to agent $i$ will be denoted as $U_i(\pi_1, \ldots, \pi_{|S|})$.

**Nash Equilibrium:** A strategy profile is said to be in Nash Equilibrium if no player has an incentive to unilaterally change its strategy when every player plays its part of that strategy profile. Formally, for a game played by $n$ players if $\pi_i$ denotes the strategy for player $i$, then the strategy profile $\pi_i \dots \pi_n$ is said to be in Nash Equilibrium iff

$$\forall i, \quad U_i(\pi_1, \dots, \pi_i, \dots \pi_n) \geq U_i(\pi_1, \dots, \hat{\pi}_i, \dots, \pi_n)$$

where $\hat{\pi}_i$ denotes any strategy for player $i$ other than $\pi_i$. There exists at least one Nash Equilibrium for any stage game. At any Nash Equilibrium, every player's strategy is a best response to the others' strategies.

**Minimax Strategy and Safety Value:** A minimax strategy for a player is a strategy which ensures the largest payoff it can receive regardless of the strategy adopted by the other player. For a player $i$, we denote the set of all possible probability distributions over $A_i$ as $\sigma_i$. In a two-player game, the minimax value for player $i$ playing against a player $j$ is then given by

$$U_i^{\text{minimax}} = \max_{\pi_i \in \sigma_i} \min_{\pi_j \in \sigma_j} U_i(\pi_i, \pi_j).$$

This value is also known as the *safety value* of a player and can be computed using linear programming techniques without the knowledge of opponent's payoffs.

**Pareto-Optimality:** An outcome or a payoff assignment is said to be Pareto-Optimal in a game if there exists no other possible payoff assignment where at least one player is better off and the others are at least as well off. A probability distribution $x^* \in X$ is said to produce a Pareto-Optimal (PO) outcome iff there exists no $x \in X$ s.t.

$$\forall i \in S, \quad U_i(x) \geq U_i(x^*)$$

and $\exists j \in S$ s.t.

$$U_j(x) > U_j(x^*).$$

In that case the payoff vector $U^* = \{U_1(x^*), \dots, U_{|S|}(x^*)\}$ is called a Pareto-Optimal outcome.

**Pareto-Dominance:** A payoff assignment corresponding to some $x_1 \in X$ is said to Pareto-Dominate another payoff assignment corresponding to some $x_2 \in X$ iff

$$\forall i \in S, \quad U_i(x_1) \geq U_i(x_2)$$

and $\exists j \in S$ s.t

$$U_j(x_1) > U_j(x_2).$$

Note that a payoff assignment is Pareto-Optimal if it is not Pareto-Dominated by any other outcome.

**Finitely and Infinitely Repeated games:** When the same stage game is played repeatedly, it is called a repeated game. A repeated game can be finite or infinite. However, an infinitely repeated game does not mean that it is played an infinite number of times. Infinitely repeated games can eventually terminate but the players are uncertain about the exact point of time when the game will terminate. In a finitely repeated game, however, the players know *a priori* exactly how many times the game will be played. It can be proved using backward induction that playing the

Nash Equilibrium strategy of the stage game is the only Nash Equilibrium for a finitely repeated version of the game. For infinitely repeated games, however, the set of Nash Equilibria can be much larger. In this paper, we consider only infinitely repeated games.

**Average Payoff Criterion:** The total payoff received by the players for an infinitely repeated game can be infinite and hence incomparable. To alleviate this problem, game theorists use averaging mechanisms such as the *$\delta$-discounted average payoff criterion* or the *limit of average payoff criterion*. In this paper, we only consider the latter criterion to model and compare payoff values received by the players. If $U_i^j$ is the payoff received by player $i$ at the $j^{th}$ iteration, then according to this criterion one sequence of payoffs $U_i = (U_i^1, U_i^2, \ldots)$ is better for player $i$ than another sequence $\hat{U}_i = (\hat{U}_i^1, \hat{U}_i^2, \ldots)$ iff,

$$\lim_{M\to\infty} \sum_{k=1}^{M} \frac{U_i^k}{M} \geq \lim_{M\to\infty} \sum_{k=1}^{M} \frac{\hat{U}_i^k}{M}.$$

We do not discuss the merits and demerits of such a payoff criterion but point out that any payoff that can be achieved as an expected payoff corresponding to some strategy-profile of the stage game can also be achieved as an average payoff of the repeated version of that game [19].

## 2.2 Multiagent learning algorithms

A number of multiagent learning algorithms have been developed by both game theorists and multiagent system researchers in the past few years for playing repeated games. It has been argued that a learning algorithm should at least converge under self-play and guarantee the safety value for all the players. The second objective is easily justified as the players can simply switch to their minimax strategy to guarantee the safety value instead of receiving a lower payoff. The first constraint is comparatively more strict and is harder to achieve. Some researchers also include rationality as a desired criterion which requires playing a best response against an opponent using a stationary strategy [1;22]. A player following a stationary strategy always chooses its action from a fixed probability distribution. We now discuss some multiagent learning techniques from current literature.

**Best Response Learners:** In any iteration of a repeated game, a Best Response learner plays that action which is a best response to the action chosen by the opponent in the last iteration. Unfortunately, BR learners perform poorly against an opponent who changes its strategy frequently enough. For example, in the game of matching pennies shown in Table 1, a BR learner would always receive a payoff of −1 against a player who chooses its two actions alternately. Also, under self-play, BR learners are not guaranteed to converge and can produce payoffs less than the safety value.

**Table 1** Payoff matrix for Matching Pennies

| | H | T |
|---|---|---|
| H | 1,−1 | −1,1 |
| T | −1,1 | 1,−1 |

**Fictitious Play Learners:** An improvement over the Best Response learner is the *fictitious play* (FP) learner [4;9]. Instead of playing best response to the last action, FP would play the best response to the average of the strategies played in the past by the opponent. FP performs better than the Best Response learner but fails to overcome the problems of Best Response learning in general. Both BR and FP players choose their strategy without knowledge of opponents' payoff structure.

**Infinitesimal Gradient Ascent Learner:** Singh, et al. [26] introduced and analyzed the concept of an Infinitesimal Gradient Ascent Learner (IGA) which changes its strategy in the direction of positive payoff gradient, but with an infinitesimal step size. They have shown that under self-play IGA will either converge to a Nash Equilibrium or will generate the expected payoff of a Nash Equilibrium. Bowling and Veloso introduced the WOLF-PHC algorithm which provably converges to a Nash Equilibrium for any two-action, two-player general sum game [1]. This is one of the strongest known convergence property for a multi-agent learning algorithm. In Sect. 5 we empirically compare the performances of WOLF-PHC and CJAL.

**No-regret Learners:** Researchers have also developed algorithms with the goal of minimizing *regret* for not playing an action at a particular iteration. There are two types of regrets: Internal regret and External regret. It has been proven that no-external regret for all the players is a necessary and sufficient condition for ensuring convergence to the Nash Equilibrium. On the other hand, no-internal regret ensures convergence to correlated equilibrium. A number of algorithms such as FPL [13], Weighted Majority [15], and Regret Matching [25] have been developed that attempts to minimize the external regret. Jafari and Greenwald [11], on the other hand, have developed an algorithm which is targeted to minimize internal regrets.

Apart from these classes of learners, researchers have also used single-agent reinforcement learning techniques such as Q-Learning and extended it for multi-agent domains with some success [7;10;12;16–18;23;31]. These algorithms have been shown to be effective for particular types of games. Most of these algorithms, however, assume complete information games and compute their strategies using the opponent's payoff structure.

Other multiagent systems literature that make the incomplete information assumption used in this paper include work that requires no communication between agents, e.g., use of aspiration levels [27], and approaches that use some form of inter-agent communication external to the learning algorithms to facilitate concurrent learning, e.g., use of commitment sequences to enforce stationary environments [14], action revelation [24], trigger transition between different phases of learning [29], use of expert's advice [8], etc. Though Vidal and Durfee explicitly models the influence of a player's action on others their calculations require more information than we assume is available [30]. For a more complete list of multiagent reinforcement learning research, refer to recent surveys of the literature[21;28].

### 2.3 Identifying POSNE outcomes

We now describe how to identify the POSNE outcomes for any infinitely repeated 2-player game using the Folk Theorem [9]. Let us consider the game of Prisoner's Dilemma. In the two-player Prisoner's Dilemma (PD) game, each agent has a choice of two actions: cooperate ($C$) or defect ($D$).

**Table 2** The Prisoner's Dilemma Game

| | C | D |
|---|---|---|
| C | R,R | S,T |
| D | T,S | P,P |

Table 2 provides the payoff vectors for each of the pure strategy profiles in $Q$ for the PD game. The following inequalities hold for the PD payoff of matrix:

$$T > R > P > S \tag{1}$$

and

$$2R > T + S. \tag{2}$$

Using the values $R = 3$, $S = 0$, $T = 4$, and $P = 2$, we can plot the payoffs in Fig. 1. The points A, B, C and D represent these payoff vectors in a 2-dimensional plane, where the $x$ and $y$ axes represent the row and column players payoffs, respectively. According to the average payoff criterion, any payoff corresponding to points A, B, C and D can be achieved in the repeated version of the game by simply repeating the action-pair that produces that payoff for the stage game. Let $P$ denote this set of points. The convex hull for the points in $P$ is the smallest convex polygon $C$ such that $\forall p \in P$, $p$ is either on the boundary or inside $C$. The convex hull for the Prisoner's Dilemma game is shown as the quadrilateral ABCD in Fig. 1. The region bounded by the convex hull for a set of points $P$ contains all the points that can be generated using a convex combination of the points in $P$. In Fig. 1, the quadrilateral ABCD includes all the payoff vectors that are feasible as expected payoffs for some probability distributions over the pure strategy profiles in $Q$ and can also be realized as an average payoff of the corresponding repeated game.

The *minimax point* is the point corresponding to the minimax payoffs for both the players. Such a point will always be either inside or on the boundary of the convex hull. In Fig. 1 the point A(1,1) is the minimax point. The Folk Theorem tells us that

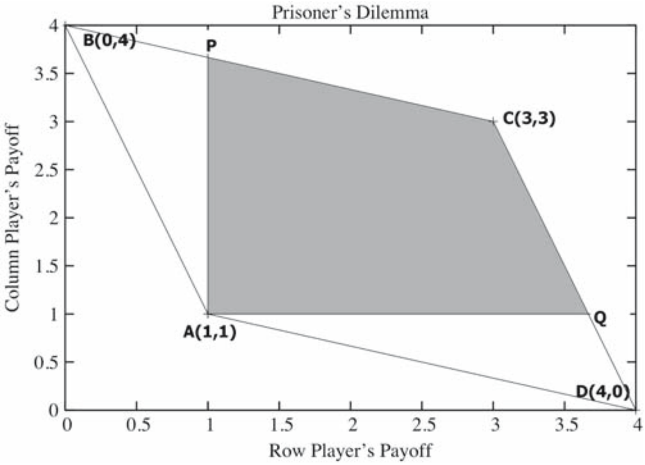

**Fig. 1** Payoff space for the prisoner's dilemma game

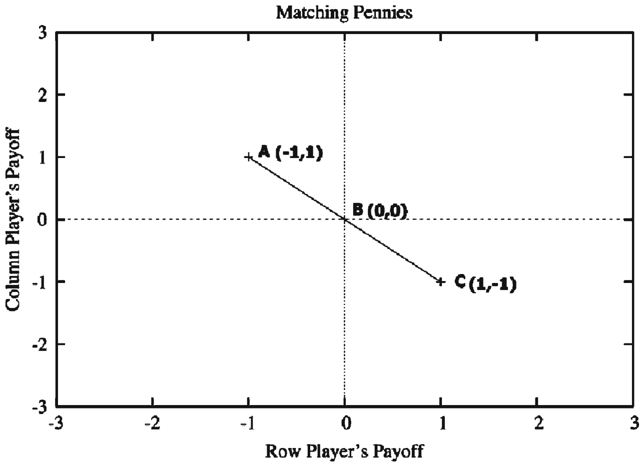

**Fig. 2** Payoff space for matching pennies

any point inside the convex hull of feasible payoffs and which Pareto-Dominates the minimax point corresponds to some Nash Equilibrium payoff for the repeated game according to the average payoff criterion [9]. We observe that a point in the payoff space is Pareto-Dominated by all other points which are to the right and above it (including the points that are either on the vertical or on the horizontal line passing through it). Hence, in Fig. 1, any point in the shaded region will coincide with the average payoff that can be obtained by playing some Nash Equilibrium strategy-profile of the repeated game. We refer to this shaded region as Area Dominating Minimax (ADM). If such an ADM exists, then we can define POSNE outcomes as the points in the ADM that are Pareto-Optimal. In Fig. 1, any point on the boundary PCQ is a POSNE outcome. We call PCQ the Pareto-Frontier for this game. In general, we define the Pareto-Frontier as the set of Pareto-Optimal points in the ADM. Note that if there exists an ADM for a game, an outcome for the game is POSNE if and only if it resides on the Pareto-Frontier. In the particular case when no point dominates the minimax point, the ADM does not exist. In this case, the minimax-point will be both a Nash Equilibrium and a POSNE outcome. For example, in Fig. 2, which shows the payoffs for the game of Matching Pennies (see Table1), no point Pareto-Dominates the minimax point B. Hence, B is both a POSNE outcome and a Nash Equilibrium for this game.

## 2.4 POSNE outcomes in the prisoner's dilemma game

From Table 2 and inequalities 1 and 2 we can observe that the dominant strategy for a player in the Prisoner's Dilemma game is to defect and the defect-defect action combination is a dominant strategy equilibrium and the only Nash equilibrium, NESSG, for this stage game. Though this solution guarantees minimax payoff for both the players, it is Pareto-Dominated by the outcome (C,C) and hence is not a POSNE outcome. In fact, all pure strategy outcomes, except the Nash Equilibrium outcome are Pareto-Optimal. Note that, (C,C) not only maximizes social welfare (sum of payoffs), but is the only POSNE outcome from the set of pure strategy combinations and resides on the Pareto-Frontier PCQ (see Fig. 1). The outcomes (D,C) and (C,D) are not POSNE as they reside outside the ADM represented by the shaded region APCQ. So the paradox is that even though there exists an action-combination which Pareto-Dominates the NESSG outcome, the players still converge to the Pareto-Dominated Nash Equilibrium outcome using individual rationality. We claim that a CJAL learner, under self-play and without the knowledge of opponent's payoff, can find this cooperate-cooperate solution which maximizes the social welfare and can converge to it given certain payoff structures (still satisfying the inequalities 1 and 2) and suitable exploration techniques. We provide an intuitive analysis and show empirical results to justify our claim.

## 3 Conditional joint action learner

We now present the details of the CJAL learning mechanism. We assume a set $S$ of 2 agents where each agent $i \in S$ has a set of actions $A_i$ and repeatedly plays a stage game. In every iteration each agent chooses an action $a_i \in A_i$. We denote the expected utility of an agent $i$ at iteration $t$ for an action $a_i$ as $E^i_t(a_i)$. In case of the Prisoner's Dilemma game, $A_i = \{C, D\}$ 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 $Pr^i_t(a_i)$ and the conditional probability that the other agent, $j$, will play $a_j$ given that the $i$th agent plays $a_i$ at iteration $t$ as $Pr^i_t(a_j|a_i)$. The joint probability of an action pair $(a_i, a_j)$ at iteration $t$ is given by $Pr_t(a_i, a_j)$. Each agent maintains a history of interactions at any iteration $t$ as

$$H^i_t = \bigcup_{\substack{a_i \in A_i \\ a_j \in A_j}} n^i_t(a_i, a_j)$$

where $n^i_t(a_i, a_j)$ denotes the number of times the joint action $(a_i, a_j)$ has been played till iteration $t$ from the beginning. We define the number of times agent $i$ has played action $a_i$ until iteration $t$ as

$$n^i_t(a_i) = \sum_{a_j \in A_j} n^i_t(a_i, a_j).$$

**Definition 1** A CJAL learner is an agent i who at any iteration $t$ chooses an action $a_i \in A_i$ with a probability proportional to $E^i_t(a_i)$ where

$$E^i_t(a_i) = \sum_{a_j \in A_j} U_i(a_i, a_j) Pr^i_t(a_j|a_i)$$

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

$$E^i_t(a_i) = \sum_{a_j \in Aj} U_i(a_i, a_j) \frac{Pr_t(a_i, a_j)}{Pr^i_t(a_i)}. \tag{3}$$

If we approximate the probability of an event as the fraction of times the event occurred in the past then Eq. 3 takes the form

$$E_t^i(a_i) = \sum_{a_j \in A_j} U_i(a_i, a_j) * \frac{n_{t-1}^i(a_i, a_j)}{n_{t-1}^i(a_i)}. \tag{4}$$

Unlike JAL [5], 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 player learns the correlation between its actions and the other agent's action and uses the conditional probabilities instead of marginal probabilities to calculate the expected utility of an action. Therefore, a CJAL learner splits the marginal probability of an action $a_j$ taken by the other player into 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 it is the first one to move. To calculate the expected utility of its action, it must then try to find the probability of the other player's action given its own action, which is basically the conditional probability we described above.

### 3.1 Learning utility estimates

We now discuss the learning mechanism used to update the expected utility values when utilities for outcomes are either not known *a priori* or can be non-deterministic. We note that it would be unreasonable to use a single-agent Q-learning scheme for CJAL to update the expected utility of its individual actions. Instead, we use a joint action Q-learning for CJAL to learn the expected utilities associated with every action combination (joint actions) and weight them by their probability of occurrence. So we rewrite the equation 4 as:

$$E_t^i(a_i) = \sum_{a_j \in A_j} Q_t^i(a_i, a_j) * \frac{n_{t-1}^i(a_i, a_j)}{n_{t-1}^i(a_i)}, \tag{5}$$

where $Q_t^i(a_i, a_j)$, the estimated payoff from joint action $(a_i, a_j)$ is updated after the $(t-1)$th interaction as

$$Q_t^i(a_i, a_j) \leftarrow Q_{t-1}^i(a_i, a_j) + \alpha(U_i(a_i, a_j) - Q_{t-1}^i(a_i, a_j)), \tag{6}$$

where $0 < \alpha \leq 1$ is the learning rate. Note that if the reward associated with a particular joint action is deterministic (which is the case for the Prisoner's Dilemma game we consider) Eq. 5 simplifies to Eq. 4. Henceforth, we will use Eq. 4 to calculate expected utility.

### 3.2 Exploration techniques

We use two distinct exploration phases in CJAL. We assume that the agents explore actions randomly for $N$ initial interactions and thereafter use an $\epsilon$-greedy exploration. In the $\epsilon$-greedy exploration phase, an agent chooses the action that produces maximum expected utility with a probability $1-\epsilon$ and explores other actions randomly with probability $\epsilon$. Therefore, the probability that at iteration $t$, agent $i$ will choose action $a_i$, $Pr_t^i(a_i)$, $\forall i \in 1, 2$ and $\forall a_i \in A_i$, is given by

$$Pr_t^i(a_i) = \begin{cases} \frac{1}{|A_i|}, \text{ if } t < N \\ 1-\epsilon, \text{ if } t > N \text{ and } a = a^* \\ \frac{\epsilon}{|A_i|-1}, \text{ if } t > N \text{ and } a \neq a^* \end{cases}$$

where $a^*$, the highest expected utility action, is defined as

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

## 4 Dynamics of CJAL learning

We now intuitively analyze the dynamics of the CJAL learning mechanism in self-play. We consider two CJAL learners playing the Prisoner's Dilemma game and analytically predict the sequence of actions they would take with time.

For the Prisoner's Dilemma game, we have $A_i = \{C, D\}$, $i = 1, 2$. We denote $U_i(C, C) = R$, $U_i(C, D) = S$, $U_i(D, D) = P$ and $U_i(D, C) = T$. In the exploration phase, as both agents choose their actions from a uniform distribution, the expected number of occurrences of each outcome will be $N/4$ after $N$ iterations. Also the expected number of times an agent would play each of its two actions is $N/2$. So, if $N$ is sufficiently large, the expected values of the conditional probabilities will be:

$$Pr_N^i(D|C) = Pr_N^i(C|C) = Pr_N^i(D|D) = Pr_N^i(C|D) = 1/2$$

for both the players. Therefore, it can be expected that the expected utility of the two actions after $N$ iterations will converge to

$$E_N^i(C) = U_i(C, C)Pr_N^i(C|C) + U_i(C, D)Pr_N^i(D|C) = \frac{R+S}{2} \tag{7}$$

and

$$E_N^i(D) = U_i(D, C)Pr_N^i(C|D) + U_i(D, D)Pr_N^i(D|D) = \frac{T+P}{2}. \tag{8}$$

Based on the constraints on the payoffs in the Prisoner's Dilemma game (see inequalities 1 and 2) we then have $E_N^i(D) > E_N^i(C)$, i.e., the defect action is preferable to the cooperate action. Therefore, if both agents choose their actions greedily ($\epsilon = 0$), they will play action $D$. They will continue playing action $D\,t$ interactions after $N$ as long as $E_{N+t}^i(D) > E_{N+t}^i(C)$. Now as they continue playing action $D$, $Pr_t^i(C|D)$ will tend to 0 and $Pr_t^i(D|D)$ will tend to 1. However, $Pr_t^i(D|C)$ and $Pr_t^i(C|C)$ will still remain at $\frac{1}{2}$. Eventually the expected utilities will be $E_i(C) = (S + R)/2$ and $E_i(D) = P$. Now if at some iteration $t$ after $N$, $\frac{S+R}{2} > E_{N+t}(D) \geq P$, then both players will switch to $C$ and receive a payoff $R$. As $R > S$, $E_{N+t}(C)$ will monotonically increase and will always be greater than $E_{N+t}(D)$. So they would continue playing $C$ in subsequent iterations and hence will converge to the $(C, C)$ outcome. We make this claim based on the assumption that after $N$ iterations each outcome will occur approximately $N/4$ times. Using the *weak law of large numbers*, we can say that it will indeed be true if $N \to \infty$. However, if $R + S < 2P$ then $E_{N+t}(C)$ will never supersede $E_{N+t}(D)$ as $E_{N+t}(D)$ can never be less than $U_i(D, D)$ and the learners will converge to the $(D, D)$ outcome.

The scenario, unfortunately, is not so simple if $\epsilon > 0$. We will experimentally show that for $\epsilon$ greater than some threshold $\epsilon_0$, convergence to $(C, C)$ may not be achieved. From the above discussion we can however make the following conjecture:

**Conjecture 1** *In the Prisoner's Dilemma game, if* $(R + S) > 2P$ *and if the agents randomly explore for a finite number of iterations N and then adopt a greedy exploration technique* ($\epsilon = 0$) *then the probability of CJAL players converging to the* $(C, C)$ *outcome tends to 1 as N approaches infinity.*

## 5 Experimental results for CJAL

We experiment with two CJAL learners repeatedly playing the Prisoner's Dilemma game. Agents keep a count of all the actions played to compute the conditional probabilities and update their beliefs after every iteration. We vary the $R, S, T$ and $P$ values and use two different exploration techniques:

1. Choosing actions randomly for the first $N$ iterations and then always choose actions with highest estimated payoff.
2. Choosing actions randomly for first $N$ iterations and $\epsilon$-greedy exploration thereafter, i.e, explore randomly with probability $\epsilon > 0$, otherwise choose action with the highest estimated payoff.

### 5.1 CJAL in self-play in prisoner's dilemma

We used $N$ as 400 in all the experiments. In the first experiment 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 Fig. 3. We also compare in Fig. 4 the values of four different conditional probabilities mentioned in Sect. 3. We observe from Fig. 4 that as the players continue to play defect after the first $N$ interactions, the probability $Pr(D|D)$ increases, but this reduces the expected utility of taking action D whereas $Pr(C|C)$ and $Pr(D|C)$ remain unchanged. This phenomena is evident from Figs. 4 and 3. A little after iteration number 1,000, 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 $\epsilon = 0$ 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 Fig. 5. We observe that as $R + S < 2P$, though the expected utility of defect reduces to $P$, the payoff of the defect-defect outcome, it still supersedes the expected utility for cooperation $R + S/2$. Hence the agents continue to defect and the system converges to the NESSG of the PD game.

In our next experiment we use $\epsilon = 0.1$ and the same payoff configuration as in the first experiment. The results are plotted in Fig. 6. We observe that though the expected value of defecting reaches below the value of $\frac{R+S}{2}$, due to exploration, $Pr(D|C)$ is also more than the $\epsilon = 0$ case, which effectively reduces the expected utility of cooperation. As a result, players find it more attractive to play defect, and hence converge to the defect-defect outcome.

### 5.2 Comparison with other learners in general-sum games

In the next set of experiments, we compare the performance of CJAL with WOLF-PHC and JAL for arbitrary 2-player general sum games. We used all possible

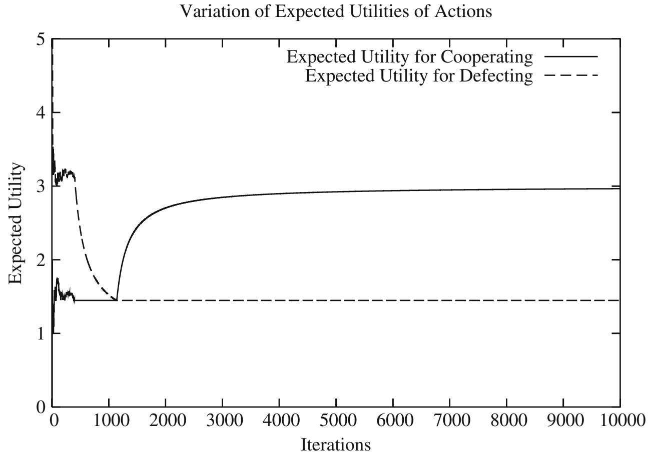

**Fig. 3** Comparison of Expected Utility when $R + S > 2P$ and $\epsilon = 0$

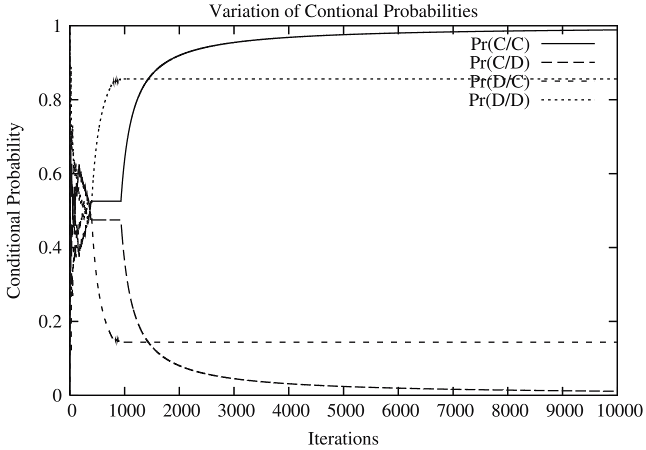

**Fig. 4** Conditional probabilities when $R + S > 2P$ and $\epsilon = 0$

structurally distinct two-player, two-action conflict games as a testbed for CJAL. In each game, each player rank orders the four possible outcomes from 1 to 4. We use the rank of an outcome as the payoff to that player for any outcome. In a conflict game, there exists no outcome that is most preferred by both the players. Steven Brams lists 57 such game matrices with ordinal payoffs [3]. We used these game matrices as a testbed to empirically verify convergence behavior of CJAL.

We played two CJAL players against each other in all of these 57 games repeatedly and observed their convergence behavior. In these experiments we used the first exploration technique i.e., $\epsilon = 0$. To evaluate the performance of CJAL we used the following criteria:

**Average Social Welfare:** Sum of the payoffs obtained by the two players in their converged state, averaged over 57 games.

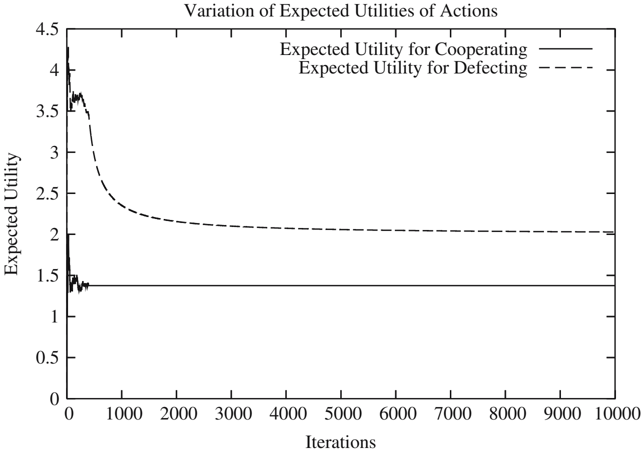

**Fig. 5** Comparison of Expected Utility when $R + S < 2P$ and $\epsilon = 0$

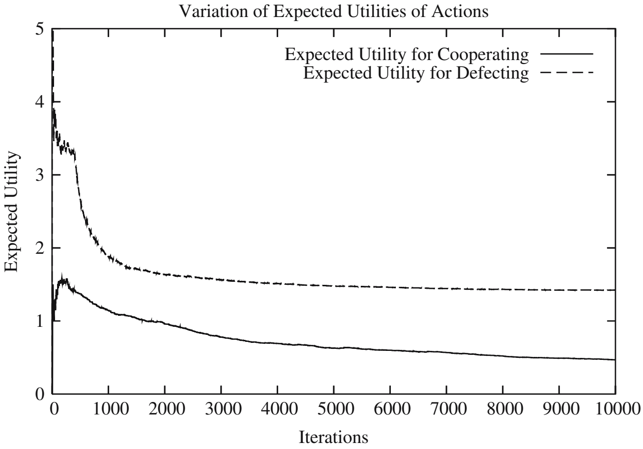

**Fig. 6** Comparison of Expected Utility when $R + S > 2P$ and $\epsilon = 0.1$

**Average Product of Payoffs:**Product of the payoffs obtained by two players in their converged state, averaged over 57 games.

**Success Rate:**Percentage of games, out of the 57 games, in which the players converge to a POSNE outcome.

We compared our results with that of WOLF-PHC and JAL, in self-play, on these 57 games using these evaluation criteria. We used 10,000 interactions for each run and averaged the results over 20 runs for each game. We only took the average of the payoffs obtained in the last 1000 iterations to approximate the final converged value. The results are then averaged over all the 57 games and are shown in Table 3. The first two columns represent the average social welfare and product of the payoffs, respectively. The third column represents the proportion of games in which the algorithms have converged to a POSNE outcome. We also compared our results with the average Nash Equilibrium payoffs for all these stage games. We observe that CJAL outperforms

**Table 3** Comparison of JAL, WOLF-PHC and CJAL on $2 \times 2$ Conflict Games

| | Social welfare | Product of payoffs | Success rate |
|---|---|---|---|
| JAL | 6.1 | 9.13 | 0.81 |
| CJAL | 6.14 | 9.25 | 0.86 |
| WOLF-PHC | 6.03 | 9.01 | 0.75 |
| Nash | 6.05 | 9.04 | 0.75 |

**Table 4** Payoff matrix for chicken game

| | C | D |
|---|---|---|
| C | 3,3 | 2,4 |
| D | 4,2 | 1,1 |

**Table 5** Payoff matrix for battle of sexes

| | F | O |
|---|---|---|
| F | 4,3 | 0,0 |
| O | 0,0 | 3,4 |

**Table 6** Social welfare generated by JAL, WOLF-PHC and CJAL on standard games

| | PD1 | PD2 | Chicken | BS | MP |
|---|---|---|---|---|---|
| JAL | 2 | 4 | 6 | 7 | 0 |
| CJAL | 6 | 4 | 6 | 7 | 0 |
| WOLF-PHC | 2 | 4 | 6 | 7 | 0 |
| Nash | 2 | 4 | 6 | 7 | 0 |

WOLF-PHC and JAL on all these metrics. For example, CJAL converges to a POSNE outcome in 86% of the games, whereas WOLF-PHC and JAL converges to a POSNE outcome in 75% and 81% cases, respectively. Also note that, in 75% of the games the single-stage Nash Equilibrium, NESSG, solutions are also POSNE. This is the success rate of WOLF-PHC which is proved to converge to the NESSG.

We also used some standard games such as Prisoner's Dilemma, Chicken games (Table 4), Battle of Sexes (BS) (Table 5) and Matching Pennies (MP) (Table 1) to analyze CJAL performance in more detail. Although most of these games are already included in these 57 conflict games, we explicitly show the Social Welfare generated for these games in Table 6. For this experiment, we used two versions of the Prisoner's Dilemma game: PD1 (Table 7) and PD2 (Table 8). For PD1, where $R + S > 2P$, CJAL converges to the cooperate-cooperate solution producing a Social Welfare of 6, whereas both JAL and WOLF-PHC converges to the NESSG and produces a Social Welfare of only 2. For PD2, where $R+S < 2P$, CJAL, as well as JAL and WOLF-PHC, converges to the NESSG and produces a Social Welfare of 4. For games like Chicken, Battle of Sexes and Matching Pennies, all the algorithms produce the same Social Welfare. The Social Welfare for these games also coincides with the Social Welfare produced by the NESSG.

**Table 7** Payoff matrix for The Prisoner's Dilemma Game, $R + S > 2P$ (PD1)

| | C | D |
|---|---|---|
| C | 3,3 | 0,4 |
| D | 4,0 | 1,1 |

**Table 8** Payoff matrix for The Prisoner's Dilemma game, $R + S < 2P$ (PD2)

| | C | D |
|---|---|---|
| C | 3,3 | 0,4 |
| D | 4,0 | 2,2 |

## 6 Conclusion and future work

We described a Conditional Joint Action Learning (CJAL) scheme and analyzed its performance for the two-player Prisoner's Dilemma game and all possible structurally distinct $2 \times 2$ conflict games with ordinal payoffs. The motivation behind CJAL was to provide the agents with the opportunity to converge to a mutually beneficial outcome against like-minded agents in incomplete information games. We propose that the goal of concurrent multiagent learning should be to reach a Pareto-Optimal outcome with Nash Equilibrium payoffs of the repeated game. We developed CJAL as a learning algorithm that can actually find such an outcome if one exists and converges to that against similar players. Our design principle underlying CJAL was motivated by the fact that in a multiagent setting a learner must consider that its action choices influence the action choice of other learners. We analyzed the convergence of CJAL in self-play in the Prisoner's Dilemma game under different payoff constraints. We showed that under certain restrictions on the payoff structures, CJAL can learn to converge to the POSNE outcome in self-play. On the other hand IL or JAL always converges to the stage game Nash Equilibrium which is a Pareto-Dominated outcome.

We also experimentally demonstrated the convergence of CJAL using limited exploration in self-play to POSNE outcomes on a representative testbed. Though CJAL was not explicitly designed to optimize measures like social welfare, fairness (measured by the product of player payoffs) and success in converging to POSNE outcomes, it outperforms JAL and WOLF-PHC on these metrics.

The results presented in this paper are for two-player, two-action games. We would like to evaluate CJAL's performance in more general $n$-player $m$-action settings.

The goal of our research is to find a generic multi-agent learning strategy that will always produce a POSNE outcome in self-play in incomplete information settings. Though CJAL is a significant step toward this goal, we would like to improve our algorithm so that it can always guarantee such a convergence. We would also like to understand and observe CJAL's behavior in a heterogeneous population where players use different learning strategies.

**Acknowledgements** This work has been supported in part by an NSF award IIS-0209208.

## References

1. Bowling, M. H., & Veloso, M. M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence, 136*(2), 215–250.

2. Bowling, M. H., & Veloso, M. M. (2004) Existence of multiagent equilibria with limited agents. *Journal of Artificial Intelligence Res. (JAIR), 22*, 353–384.
3. Brams, S. J. (1994) *Theory of moves*. Cambridge, UK: Cambridge University Press.
4. Brown, G. W. (1951). Iterative solution of games by fictiious play. In *activity analysis of production and allocation*. New York: Wiley.
5. Claus, C., & Boutilier, C. (1997). The dynamics of reinforcement learning in cooperative multiagent systems. In *Collected papers from AAAI-97 workshop on Multiagent Learning*, (pp. 13–18). AAAI.
6. Conitzer, V., &Sandholm, T. (2003). Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In *ICML*, (pp. 83–90).
7. Crandall, J. W., & Goodrich, M. A. (2005). Learning to compete, compromise, and cooperate in repeated general-sum games. In *Proceedings of the nineteenth international conference on machine learning*, pp. 161–168.
8. de Farias, D. P., & Megiddo, N. (2003). How to combine expert (and novice) advice when actions impact the environment? In *NIPS*.
9. Fudenberg, D., & Levinem K. (1998). *The theory of learning in games*. Cambridge, MA: MIT Press.
10. Greenwald, A. R., & Hall, K. (2003). Correlated q-learning. In *ICML*, pp. 242–249.
11. Greenwald, A. R., & Jafari, A. (2003). A general class of no-regret learning algorithms and game-theoretic equilibria. In *COLT*, pp. 2–12.
12. Hu J., & Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, *4*, 1039–1069.
13. Kalai, A., & Vempala, S. (2002). Geometric algorithms for online optimization. Technical Report MIT-LCS-TR-861, MIT Laboratory for Computer Science.
14. Kapetanakis, S., Kudenko, D., & Strens, M. (2004). Learning of coordination in cooperative multiagent systems using commitment sequences. *Artificial Intelligence and the Simulation of Behavior, 1*(5).
15. Littlestone, N., & Warmuth, M. K. (1989). The weighted majority algorithm. In *IEEE symposium on foundations of computer science*, pp. 256–261.
16. Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, (pp. 157–163). San Mateo, CA: Morgan Kaufmann.
17. Littman, N. L. (2001). Friend-or-foe q-learning in general-sum games. In *Proceedings of the eighteenth international conference on machine learning*, (pp. 322–328) San Francisco, CA: Morgan Kaufmann.
18. Littman, M. L., & Stone, P. (2001). Implicit negotiation in repeated games. In *Intelligent agents VIII: Agent theories, architecture, and languages*, pp. 393–404.
19. Littman, M. L., & Stone, P. (2005). A polynomial-time nash equilibrium algorithm for repeated games. *Decision Support System, 39*, 55–66.
20. Mundhe, M., & Sen, S. (1999). Evaluating concurrent reinforcement learners. IJCAI-99 workshop on agents that learn about, from and with other agents.
21. Panait, L., & Luke, S. (2005). Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems, 11*(3), 387–434.
22. Sandholm, T. W., & Crites, R. H. (1995). Multiagent reinforcement learning and iterated prisoner's dilemma. *Biosystems Journal, 37*, 147–166.
23. Sekaran, M., & Sen, S. (1994). Learning with friends and foes. In *Sixteenth annual conference of the cognitive science society*, (pp. 800–805). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
24. Sen, S., Mukherjee, R., & Airiau, S. (2003). Towards a pareto-optimal solution in general-sum games. In *Proceedings of the second intenational joint conference on autonomous agents and multiagent systems* (pp. 153–160). New York, NY: ACM Press.
25. Mas-Colell, A., & Hart, S. (2001). A general class of adaptive strategies. *Journal of Economic Theory, 98*(1), 26–54.
26. Singh, S. P., Kearns, M. J., & Mansour, Y. (2000) Nash convergence of gradient dynamics in general-sum games. In *UAI*, pp. 541–548.
27. Stimpson, J. L., Goodrich, M. A., & Walters, L. C. (2001) Satisficing and learning cooperation in the prisoner's dilemma. In *Proceedings of the seventeenth international joint conference on artificial intelligence*, pp. 535–540.
28. Tuyls, K., & Nowé, A. (2006). Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review, 20*(1), 63–90.

29. Verbeeck, K., Nowé, A., Lenaerts, T., & Parentm, J. (2002). Learning to reach the pareto optimal nash equilibrium as a team. In *LNAI 2557: Proceedings of the fifteenth Australian joint conference on artificial intelligence*, Vol. (pp. 407–418). Springer-Verlag.
30. Vidal, J. M., & Durfee, E. H. (2003). Predicting the expected behavior of agents that learn about agents: the CLRI framework. *Autonomous Agents and Multi-Agent Systems, 6*(1), 77–107.
31. Weiß, G. Learning to coordinate actions in multi-agent systems. In *Proceedings of the international joint conference on artificial intelligence*, pp. 311–316, August 1993.