Learning Pareto-optimal Solutions in 2x2 Conflict Games

Stéphane Airiau and Sandip Sen

Department of Mathematical & Computer Sciences,
The University of Tulsa, USA
{stephane, sandip}@utulsa.edu

Abstract. Multiagent learning literature has investigated iterated two-player games to develop mechanisms that allow agents to learn to converge on Nash Equilibrium strategy profiles. Such equilibrium configurations imply that no player has the motivation to unilaterally change its strategy. Often, in general sum games, a higher payoff can be obtained by both players if one chooses not to respond myopically to the other player. By developing mutual trust, agents can avoid immediate best responses that will lead to a Nash Equilibrium with lesser payoff. In this paper we experiment with agents who select actions based on expected utility calculations that incorporate the observed frequencies of the actions of the opponent(s). We augment these stochastically greedy agents with an interesting action revelation strategy that involves strategic declaration of one's commitment to an action to avoid worst-case, pessimistic moves. We argue that in certain situations, such apparently risky action revelation can indeed produce better payoffs than a non-revealing approach. In particular, it is possible to obtain Pareto-optimal Nash Equilibrium outcomes. We improve on the outcome efficiency of a previous algorithm and present results over the set of structurally distinct two-person two-action conflict games where the players' preferences form a total order over the possible outcomes. We also present results on a large number of randomly generated payoff matrices of varying sizes and compare the payoffs of strategically revealing learners to payoffs at Nash equilibrium.

1 Introduction

The goal of a rational learner, repeatedly playing a stage game against an opponent, is to maximize its expected utility. In a two-player, general-sum game, this means that the players need to systematically explore the joint action space before settling on an efficient action combination[1]. Both agents can make concessions from greedy strategies to improve their individual payoffs in the long run [1]. Reinforcement learning schemes, and in particular, Q-learning [2] have

[1] Though the general motivation behind our work and the proposed algorithms generalize to n-person games, we restrict our discussion in this paper to two-person games.

K. Tuyls et al. (Eds.): LAMAS 2005, LNAI 3898, pp. 86–99, 2006.

been widely used in single-agent learning situations. In the context of two-player games, if one agent plays a stationary strategy, the stochastic game becomes a Markov Decision Process and techniques like Q-learning can be used to learn to play an optimal response against such a static opponent. When two agents learn to play concurrently, however, the stationary environment assumption does not hold any longer, and Q-learning is not guaranteed to converge in self-play. In such cases, researchers have used the goal of convergence to Nash equilibrium in self-play, where each player is playing a best response to the opponent strategy and does not have any incentive to deviate from its strategy. This emphasis on convergence of learning to Nash equilibrium is rooted in the literature in game theory [3] where techniques like fictitious play and its variants lead to Nash equilibrium convergence under certain conditions.

Convergence can be a desirable property in multiagent systems, but converging to just any Nash equilibrium is not necessarily the preferred outcome. A Nash equilibrium of the single shot,i.e., stage game is not guaranteed to be Pareto optimal[2]. For example, the widely studied Prisoner's dilemma (PD in Table 1(b)) game has a single pure strategy Nash equilibrium that is defect-defect, which is

Table 1. Prisoner's dilemma and Battle of Sexes games

(a) Battle of the Sexes

	C	D
C	1,1	3,4
D	4,3	2,2

(b) Prisoners' dilemma

	C	D
C	3,3	1,4
D	4,1	2,2

dominated by the cooperate-cooperate outcome. On the other hand, a strategy that is Pareto Optimal is not necessarily a Nash equilibrium, i.e., there might be incentives for one agent to deviate and obtain higher payoff. For example, each of the agents has the incentive to deviate from the cooperate-cooperate Pareto optima in PD. In the context of learning in games, it is assumed that the players are likely to play the game over and over again. This opens the possibility for such defections to be deterred or curtailed in repeated games by using disincentives. Actually, in the context of repeated games, the Folks Theorems ensure that any payoffs pair that dominates the security value[3] can be sustained by a Nash equilibrium. This means that in the context of the repeated games, Pareto optimal outcome can be the outcome of a Nash equilibrium. In [4], Littman and

[2] A Pareto optimal outcome is one such that there is no other outcome where some agent's utility can be increased without decreasing the utility of some other agent. An outcome X *strongly dominates* another outcome B if all agents receive a higher utility in X compared to Y. An outcome X *weakly dominates* (or simply *dominates*) another outcome B if at least one agent receives a higher utility in X and no agent receives a lesser utility compared to outcome Y. A non-dominated outcome is Pareto optimal.

[3] The security value is the minimax outcome of the game: it is the outcome that a player can guarantee itself even when its opponent tries to minimize its payoff.

Stone present an algorithm that converges to a particular Pareto Optimal Nash equilibrium in the repeated game.

It is evident that the primary goal of a rational agent, learning or otherwise, is to maximize utility. Though we, as system designers, want convergence and corresponding system stability, those considerations are necessarily secondary for a rational agent. The question then is what kind of outcomes are preferable for agents engaged in repeated interactions with an uncertain horizon, i.e., without knowledge of how many future interactions will happen. Several current multi-agent learning approaches [4, 5, 6] assume that convergence to Nash equilibrium in self-play is the desired goal, and we concur since it is required to obtain a stable equilibrium. We additionally claim that any Nash equilibrium that is also Pareto optimal should be preferred over other Pareto optimal outcomes. This is because both the goals of utility maximization and stability can be met in such cases. But we find no rational for preferring convergence to a dominated Nash equilibria. Based on these considerations we now posit the following goal for rational learners in self-play:

> **Learning goal in repeated play:** The goal of learning agents in repeated self-play with an uncertain horizon is to reach a Pareto-optimal Nash equilibria (PONE) of the repeated game.

We are interested in developing mechanisms by which agents can produce PONE outcomes. In this paper, we experiment with two-person, general-sum games where each agent only gets to observe its own payoff and the action played by the opponent, but not the payoff received by the opponent. The knowledge of this payoff would allow the players to compute PONE equilibria and to bargain about the equilibrium. For example the algorithm in [4] assumes the game is played under complete information, and the players compute and execute the strategy to reach a particular equilibrium (the Nash bargaining equilibrium). However, the payoff represents a utility that is private to the player. The player may not want to share this information. Moreover, sharing one's payoff structure requires trust: deceptive information can be used to take advantage of the opponent. The ignorance of the opponent's payoff requires the player to estimate the preference of its opponent by its actions rather than by what could be communicated. By observing the actions played, our goal is to make players discover outcomes that are beneficial for both players and provide incentive to make these outcomes stable. This is challenging since agents cannot realize whether or not the equilibrium reached is Pareto Optimal.

We had previously proposed a modification of the simultaneous-move game playing protocol that allowed an agent to communicate to the opponent its irrevocable commitment to an action [7]. If an agent makes such a commitment, the opponent can choose any action in response, essentially mirroring a sequential play situation. At each iteration of the play, then, agents can choose to play a simultaneous move game or a sequential move game. The motivation behind this augmented protocol is for agents to build trust by committing up front to a "cooperating" move, e.g., a cooperate move in PD. If the opponent myopically chooses an exploitative action, e.g., a defect move in PD, the initiating agent

would be less likely to repeat such cooperation commitments, leading to outcomes that are less desirable to both parties than mutual cooperation. But if the opponent resists the temptation to exploit and responds cooperatively, then such mutually beneficial cooperation can be sustained.

We view the outcome of a Nash equilibrium of the one shot game as an outcome reached by two players that do not want to try to build trust in search of an efficient outcome. Though our ultimate goal is to develop augmented learning algorithms that provably converge to PONE outcomes of the repeated game, in this paper we highlight the advantage of outcomes from our augmented learning schemes over Nash equilibrium outcomes of the single shot, stage game. In the rest of the paper, by Nash equilibrium, we refer to the Nash equilibrium of the stage game, which is a subset of the set of Nash equilibria of the repeated version of the stage game.

We have empirically shown, over a large number of two-player games of varying sizes, that our proposed revelation protocol, that is motivated by considerations of developing trusted behavior, produces higher average utility outcome than Nash equilibrium outcomes of the single-shot games[7]. For a more systematic evaluation of the performance of our proposed protocol, we study, in more detail, all two-player, two-action conflict games to develop more insight about these results and to improve on our previous approach. A *conflict game* is a game where both players do not view the same outcome as most profitable. We are not interested in no-conflict games as the single outcome preferred by both players is easily learned. We use the testbed proposed by Brams in [8] and consisting of all 2x2 structurally distinct conflict games. In these games, each agent rank orders each of the four possible outcomes. On closer inspection of the results from our previous work, we identified enhancement possibilities over our previous approaches. In this paper, we present the updated learners, the corresponding testbed results and the challenges highlighted by those experiments.

2 Related Work

Over the past few years, multiagent learning researchers have adopted convergence to Nash equilibrium of the repeated game as the desired goal for a rational learner [4, 5, 6]. By modeling its opponent, Joint-Action Learners [9] converge to a Nash equilibrium in cooperative domains. By using a variable rate, WoLF [6] is guaranteed to converge to a Nash equilibrium in a two-person, two-actions iterated general-sum game, and converges empirically on a number of single-state, multiple state, zero-sum, general-sum, two-player and multi-player stochastic games. Finally, in any repeated game AWESOME [5] is guaranteed to learn to play optimally against stationary opponents and to converge to a Nash equilibrium in self-play.

Some multiagent learning researchers have investigated other non-Nash equilibrium concepts like *coordination equilibrium* [10] and *correlated equilibrium* [11]. If no communication is allowed during the play of the game, the players choose their strategies independently. When players use mixed strategies, some bad

outcome can occur. The concept of correlated equilibrium [12] permits dependencies between the strategies: for example, before the play, the players can adopt a strategy according to the joint observation of a public random variable. [11] introduces algorithms which empirically converge to a correlated equilibrium in a testbed of Markov game.

Consider the example of a Battle of Sexes game represented in Table 1(a). The game models the dilemma of a couple deciding on the next date: they are interested to go in different places, but both prefer to be be together than alone. In this game, defecting is following one's own interest whereas cooperating is following the other's interest. If both defect, they will be on their own, but enjoy the activity they individually preferred, with a payoff of 2. If they both cooperate, they will also be on their own, and will be worse off, with the lowest payoff of 1, as they are now participating in the activity preferred by their partner. The best (and fair) solution would consists in alternating between (Coordinate, Defect) and (Defect, Coordinate) to obtain an average payoff of 3.5. The Nash equilibrium of the game is to play each action with probability 0.5, which yields an average payoff of 2.5. Only if the players observe a public random variable can they avoid the worst outcomes.

The commitment that one player makes to an action in our revelation protocol can also be understood as a signal that can be used to reach a correlated equilibrium [11]. For example, in the Battle of Sexes game, if a player commits to cooperate, the other player can exploit the situation by playing defect, which is beneficial for both players. When both players try to commit, they obtain 3.5 on average.

3 Game Protocol and Learners

In this paper, we build on the simultaneous revelation protocol [7]. Agents play an nxn bimatrix game. At each iteration of the game, each player first announces whether it wants to commit to an action or not (we will also use reveal an action or not). If both players want to commit at the same time, one is chosen randomly with equal probability. If none decides to commit, then both players simultaneously announce their action. When one player commits an action, the other player plays its best response to this action. Note that for now, the answer to the committed action is myopic, we do not consider yet a strategic answer to the revealed action. Each agent can observe whether the opponent wanted to commit, which agent actually committed, and which action the opponent played. Only the payoff of the opponent remains unknown, since its preferences are considered private.

Let us use as an example matrix #27 of the testbed (Table 2(a)). The only Nash equilibrium of the stage game is when both players play action 0, but this state is dominated by the state where both agents play action 1. If the row player commits to play action 1, the column player plays its best response that is action 1: the row player gets 3, and the column player gets 4, which improves on the payoff of the Nash equilibrium where row gets 2 and column gets 3. The

Table 2. Representative games where proposed strategy enhancement leads to improvement

(a) Game 27

	0	1
0	2, 3	4, 1
1	1, 2	3, 4

(b) Game 29

	0	1
0	3, 2	2, 1
1	4, 3	1, 4

(c) Game 48

	0	1
0	3, 3	2, 1
1	4, 2	1, 4

column player could ensure a payoff of 3 (the payoff of the Nash equilibrium) by revealing action 0, since the row player would play the best response, i.e. action 0. However, by choosing not to commit, the column player let the row player commit: thus the column player obtains its most preferred outcome of 4. If the row player learns to reveal action 1 and the column learns not to reveal in this game matrix, the two learners can converge to a Pareto optimal state that dominates Nash equilibrium.

3.1 Learners

The agents used are expected utility based probabilistic (EUP) learners. An agent estimates the expected utility of each of its action and plays by sampling a probability distribution based on the expected utilities. First, the agent must decide whether to reveal or not. We will use the following notation:

- Q(a,b) is the payoff of the agent when it plays a and the opponent plays b.
- $BR(b)$ denotes the best response to action b.
- p_{OR} is the probability that the opponent wants to reveal.
- $p_{BR}(b|a)$ is the probability that the opponent plays action b when the agent reveals action a.
- $p_R(b)$ is the probability that the opponent reveals b given that it reveals.
- $p_{NR}(b)$ is the probability that the opponent plays action b in simultaneous play, i.e., when no agent reveals.

In [7], the expected utility to reveal an action is

$$EU_r(a) = \sum_{b \in \mathcal{B}} p_{BR}(b|a) Q(a,b)$$

and the expected utility of not revealing is

$$EU_{nr}(a) = \sum_{b \in \mathcal{B}} p_{NR}(b) Q(a,b),$$

where B is the opponent's action set. Back to our example of game #27 (Table 2(a)), the row player quickly learns to reveal action 1, providing it a payoff of 3 and allowing the column player to get its most preferred outcome. However, the expected utility of the column player to reveal action 0 is 3, and the expected utility of not revealing an action should be 4, and not 3 as computed from the above equations used in our previous work. This difference is because

a utility-maximizing opponent will prefer to always reveal in this game. Hence, we need to take into account the possibility of the opponent revealing in the computation of the expected utility. Our augmented expressions for computing the expected utilities to reveal action a is

$$EU_r(a) = \begin{array}{c} (1-p_{OR})\sum_{b\in\mathcal{B}} p_{BR}(b|a)Q(a,b) \\ + \\ \frac{p_{OR}}{2}\sum_{b\in\mathcal{B}} \left(p_R(b)Q(BR(b),b) + p_{BR}(b|a)Q(a,b)\right). \end{array}$$

Two cases can occur. Either the opponent does not want to reveal, in which case the opponent will reply to the agent's revelation, or the opponent also wants to reveal, and with equal probability the opponent and the agent will get to reveal its action. We also have the same cases when computing the expected utility of playing action a, but not revealing. If the opponent reveals, the agent will have to play the best response to the revealed action. If the opponent does not reveal, both agents will announce their actions simultaneously. Hence the expected utility is:

$$EU_{nr}(a) = \begin{array}{c} p_{OR}\sum_{b\in\mathcal{B}} p_R(b)Q(BR(b),b) \\ + \\ (1-p_{OR})\sum_{b\in\mathcal{B}} p_{NR}(b)Q(a,b) \end{array}$$

To choose an action from the expected utilities computed, the agent samples the Boltzmann probability distribution with temperature T and decides to reveal action a with probability :

$$p(\text{reveal } a) = \frac{e^{\frac{EU_r(a)}{T}}}{\sum_{x\in\mathcal{A}}\left(e^{\frac{EU_r(x)}{T}} + e^{\frac{EU_{nr}(x)}{T}}\right)}$$

and it decides not to reveal with probability

$$p(\text{not reveal}) = \frac{\sum_{x\in\mathcal{A}} e^{\frac{EU_{nr}(x)}{T}}}{\sum_{x\in\mathcal{A}}\left(e^{\frac{EU(x)}{T}} + e^{\frac{EU_{nr}(x)}{T}}\right)},$$

where $\mathcal{A}$ is the agent's action set.

If the agent reveals but not the opponent, the agent is done. If the opponent reveals action b, the agent plays its best response: $\text{argmax}_a Q(a,b)$. If no agent has decided to reveal, the agent computes the expected utility to play each action:

$$EU(a) = \sum_{b\in\mathcal{B}} p_{NR}(b)Q(a,b).$$

The agent chooses its action a sampling the corresponding Boltzmann probability distribution

$$p(a) = \frac{e^{\frac{EU(a)}{T}}}{\sum_{b \in \mathcal{B}} e^{\frac{EU(b)}{T}}}.$$

The temperature parameter, T, controls the exploration versus exploitation tradeoff. At the beginning of the game, the temperature is set to a high value, which ensures exploration. At each iteration, the temperature is reduced until the temperature reaches a preset minimum threshold (the threshold is used to prevent exponent overflow computation errors). The use of the Boltzmann probability distribution with a decreasing temperature means that the players converge to play pure strategies. If both agents learn to reveal, however, the equilibrium reached is a restricted mixed strategy (at most two states of the games will be played with equal probability).

4 Experimental Results

In the stage game, the players cannot build any trust required to find a mutually beneficial outcome of the game. The goal of our experiments is to study whether the learners using our augmented revelation protocol and by repeatedly playing a game can improve performance compared to Nash equilibrium payoffs of the stage game. In the following, by Nash equilibrium we refer to the Nash equilibrium of the single shot, stage game.

The testbed, introduced by Brams in [8] consists of all 2x2 conflicting games with ordinal payoff. Each player has a total preference order over the 4 different outcomes. We use the numbers 1, 2, 3 and 4 as the preference of an agent, with 4 being the most preferred. We do not consider games where both agents have the highest preference for the same outcome. Hence games in our testbed contain all possible conflicting situations with ordinal payoffs and two choices per agent. There are 57 structurally different, i.e., no two games are identical by renaming the actions or the players, 2x2 conflict games.

In order to estimate the probabilities presented in the previous section, we used frequency counts over the history of the play. We start with a temperature of 10, and we decrease the temperature with a decay of .5% at each iteration. We are first presenting results on a set of interesting matrices and then provide results on the entire testbed.

4.1 Results on the Testbed

Benefits of the Augmented Protocol. We compared the results over the testbed to evaluate the effectiveness of the augmentation. We found out that in the three games of Table 2, the equilibrium found strictly dominates the equilibrium found with the non-augmented algorithm. The payoffs, averaged over 100 runs are presented in Table 3. In the three games, one player needs to realize that it is better off by letting the opponent reveal its action, which is the purpose of the augmentation. Note that even without the augmentation, the

Table 3. Comparison of the average payoff between the augmented and the non augmented Expected Utility calculations

		Not augmented		Augmented	
	Nash Payoff	average payoff	strategy	average payoff	strategy
Game 27	(2,2)	(2.5, 3.5)	row: reveal 1 col: reveal 0	(3.0, 4.0)	row: reveal 1 col: no rev
Game 29	(2.5, 2.5)	(3.5, 2.5)	row: no rev 0 col: no rev 0	(4.0, 3.0)	row: no rev col: reveal 0
Game 48	(2,3)	(2.5, 3.5)	row: reveal 1 col: reveal 0	(3.0, 4.0)	row: reveal 1 col: no rev
Game 50	(2,4)	(2.3, 3.3)	row: mix col: mix	(2.5, 3.0)	row: reveal 1 col: reveal 0

opportunity of revealing the action brings an advantage since the equilibrium found dominates the Nash equilibrium of the single stage game.

We provide in Figures 1 and 2 the learning curves of the augmented and the non-augmented players, respectively, for game #27 of the testbed (see Table 2(a)). The figures present the dynamics of the expected values of the different actions and the probability distributions for both players when they learn to play. With the augmentation, we see that the row player first learns to play its Nash equilibrium component, before realizing that revealing its action 1 is a better option. The column player first learns to either reveal action 0 or not reveal and then play action 0. But as soon as the column player starts to reveal

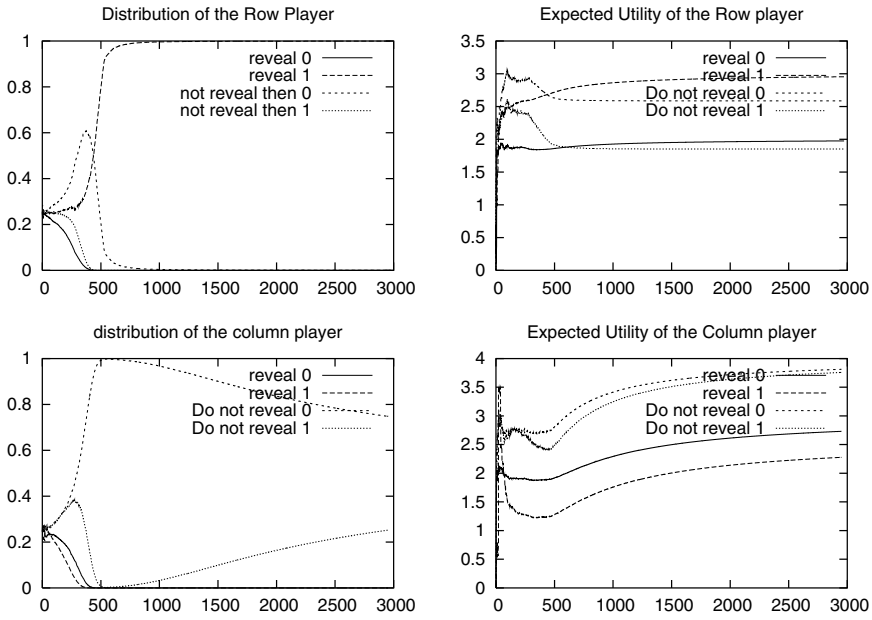


Fig. 1. Learning to play game 27 - augmented

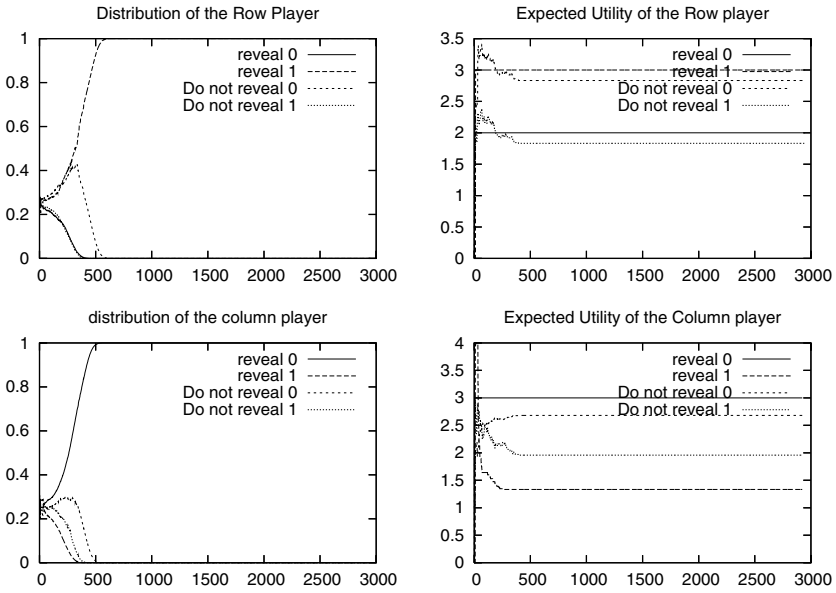


Fig. 2. Learning to play game 27 - not augmented

its action 1, the column player learns not to reveal, which was not possible with earlier expression of the expected utility. These observations confirm that the augmentation can improve the performance of both players.

Comparing protocol outcome with Nash Equilibrium. 51 of the 57 games in the testbed have a unique Nash equilibrium (9 of these games have a mixed strategy equilibrium and 42 have pure strategy equilibrium), the remaining 6 have multiple equilibria (two pure Nash equilibria and and a mixed strategy Nash equilibrium). Of the 42 games that have a unique pure strategy Nash equilibrium, 4 games have a Nash equilibrium that is not Pareto-optimal: the prisoners' dilemma, game #27, #28 and #48 have a unique Nash equilibrium which is dominated.

The Pareto optimal outcome is reached games #27, #28 and #48 with the augmented algorithm. The non-augmented protocol converges to the Pareto equilibrium for game #28, but it failed to do so for games #27 and #48. We noticed that in some games, namely games #41, #42, #44, the players learn not to reveal. Revealing does not help improve utility in these games. Incidentally, these games also have a single mixed strategy Nash equilibrium.

We found that the augmented mechanism fails to produce a Pareto optimal solution in only two games: the Prisoner's dilemma game (Table 4(a)) and game #50 (Table 4(b)) fails to converge because of the opportunity to reveal.

The Prisoner's dilemma game has a single Nash equilibrium where each player plays D. If a player reveals that it is going to cooperate (i.e. play C), the opponent's myopic best response is to play defect (i.e. to play D). With the revelation mechanism, the players learn to play D (by revealing or not). Hence, the players do not benefit from the revelation protocol in the Prisoner's dilemma game.

Table 4. Games for which convergence to a Pareto optimal solution was not achieved

(a) Prisoners' Dilemma

	D	C
D	2, 2	4, 1
C	1, 4	3, 3

(b) Game 50

	0	1
0	2, 4	4, 3
1	1, 1	3, 2

From Table 3, we find that in game #50, the new solution with the augmented protocol does not dominate the old solution. Without the augmentation, there are multiple equilibria. One is when the column player reveals action 0, providing 2 for the row and 4 to the column player. The other is when both players learn to reveal, providing 2.5 for the row player and 3 for the column player. The payoff obtained with the revelation and the payoff of the Nash equilibrium outcome of the stage game do not dominate one another. This game has a single Nash equilibrium which is also a Pareto optima and where each agent plays action 0. By revealing action 0, i.e., its component of the Nash equilibrium, the column player can obtain its most preferred outcome since the best response of the row player is to play action 0. The row player, however, can obtain more than the payoff of the Nash equilibrium by revealing action 1 where the column player's best response is its action 1. The (1,1) outcome, however is not Pareto optimal since it is dominated by the (0,1) outcome. The dynamics of the learning process in this game is shown in Figure 3. Both the players learn to reveal and hence each reveals about 50% of the time, and in each case the other agent plays its best response, i.e., the outcome switches between (0,0) and (1,1). The interesting observation is that the average payoff of the column player is 3, which would

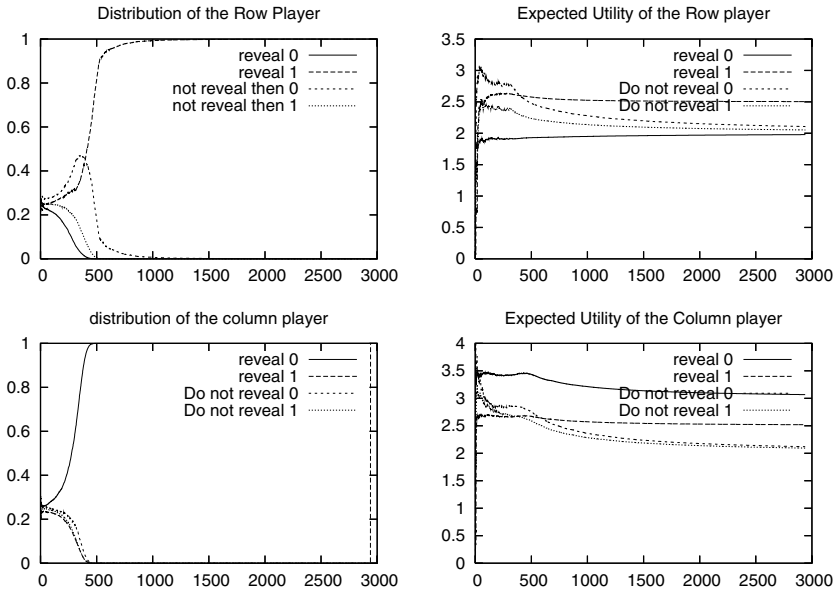


Fig. 3. Learning to play game #50

be its payoff if the column player played 1 instead of a myopic choice of 0 to row player's revealing action 0. Hence, revealing an action does not improve the outcome of this game because of a myopic best response by the opponent.

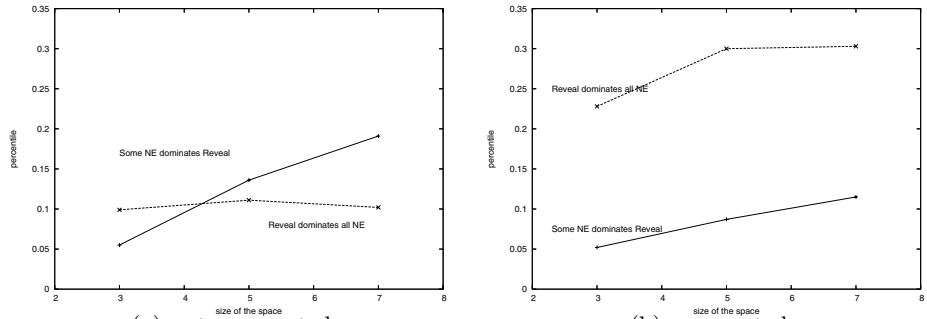
4.2 Results on Randomly Generated Matrices

As shown in the restricted testbed of 2x2 conflicting games with a total preference over the outcomes, the structure of some games can be exploited by the augmented protocol to improve the payoffs of both players. We have not seen cases where both agents would be better off by playing the Nash equilibrium (i.e. we have not encountered cases where revelation worsens the outcome). To evaluate the effectiveness of the protocol on a more general set of matrices, we ran experiments on randomly generated matrices as in [7].

We generated 1000 matrices of size 3x3, 5x5 and 7x7. Each matrix entry is sampled from a uniform distribution in $[0, 1]$. We computed the Nash equilibrium of the stage game of all these games using Gambit [13]. We compare the payoff of the Nash equilibrium with the average payoff over 10 runs of the game played with the revelation protocol. We are interested in two main questions:

- In what proportion of the games does the revelation protocol dominate all the Nash equilibria of the stage game?
- Are there some games where a Nash equilibrium dominates the outcome of the game played with the revelation protocol?

Results from the randomly generated matrices with both the augmented and non-augmented variations are presented in Figure 4. The top curve on each figure represents the percentage of games where all the Nash equilibria (NE) are dominated by the outcome of the revelation protocol. We find that the augmented protocol is able to significantly improve the percentage of Nash dominating outcomes and improves the outcome over Nash equilibria outcomes on 20–30% of the games. The percentage of such games where a Nash Equilibrium is better than the outcome reached by the revelation protocol is represented in the lower curve. We observe that this percentage decreases significantly with the



(a) not augmented (b) augmented

Fig. 4. Results over random generated matrices

augmentation and is now at the 5–10% range. Although these results show that the proposed augmentation is a clear improvement over the previous protocol, there is still scope for improvement as the current protocol does not guarantee PONE outcomes.

5 Conclusion and Future Work

In this paper, we augmented a previous algorithm from [7] with the goal of producing PONE outcomes in repeated single-stage games. We experiment with two-player two-action general-sum conflict games where both agents have the opportunity to commit to an action and allow the other agent to respond to it. Though the revealing one's action can be seen as making a concession to the opponent, it can also be seen as an effective means to force the exploration a subset of the possible outcomes and as a means to promoting trusted behavior that can lead to higher payoffs than defensive, preemptive behavior that eliminates mutually preferred outcomes in an effort to avoid worst-case scenarios. The outcome of a Nash equilibrium of the single shot, stage games can be seen as outcomes reached by myopic players. We empirically show that our augmented protocol can improve agent payoffs compared to Nash equilibrium outcomes of the stage game in a variety of games: the search of a mutually beneficial outcome of the game pays off in many games. The use of the testbed of all structurally distinct 2x2 conflict games [8] also highlights the shortcomings of the current protocol. Agents fails to produce Pareto optimal outcomes in the prisoners' dilemma game and game #50 . The primary reason for this is that a player answers a revelation with a myopic best response.

To find a non-myopic equilibrium, an agent should not be too greedy! We are working on relaxing the requirement of playing a best response when the opponent reveals. We plan to allow an agent to estimate the effects of its various responses to a revelation on subsequent play by the opponent. This task is challenging since the space of strategies, using the play history, used by the opponent to react to one's play is infinite.

Another avenue of future research is to characterize the kind of equilibrium we reach and the conditions under which the algorithm converges to a outcome that dominates all Nash equilibria of the stage game. We plan to actively pursue modifications to the protocol with the goal of reaching PONE outcomes of the repeated game in all or most situations.

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

References

1. Littman, M.L., Stone, P.: Leading best-response strategies in repeated games. In: IJCAI Workshop on Economic Agents, Models and Mechanisms. (2001)
2. Watkins, C.J.C.H., Dayan, P.D.: Q-learning. Machine Learning **3** (1992) 279 – 292

3. Fudenberg, D., Levine, K.: The Theory of Learning in Games. MIT Press, Cambridge, MA (1998)
4. Littman, M.L., Stone, P.: A polynomial-time nash equilibrium algorithm for repeated games. Decision Support Systems **39** (2005) 55–66
5. Conitzer, V., Sandholm, T.: Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In: Proceedings ont the 20th International Conference on Machine Learning. (2003)
6. Bowling, M., Veloso, M.: Multiagent learning using a variable learning rate. Artificial Intelligence **136** (2002) 215–250
7. Sen, S., Airiau, S., Mukherjee, R.: Towards a pareto-optimal solution in general-sum games. In: Proceedings of the Second International Joint Conference On Autonomous Agents and Multiagent Systems. (2003)
8. Brams, S.J.: Theory of Moves. Cambridge University Press, Cambridge: UK (1994)
9. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, Menlo Park, CA, AAAI Press/MIT Press (1998) 746–752
10. Littman, M.L.: Friend-or-foe q-learning in general-sum games. In: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann (2001) 322–328
11. Greenwald, A., Hall, K.: Correlated-q learning. In: Proceedings of the Twentieth International Conference on Machine Learning. (2003) 242–249
12. Aumann, R.: Subjectivity and correlation in randomized strategies. Journal of Mathematical Economics **1** (1974) 67–96
13. McKelvey, R.D., McLennan, A.M., Turocy, T.L.: Gambit: Software tools for game theory version 0.97.0.7. http://econweb.tamu.edu/gambit (2004)