Learning to Commit in Repeated Games

Stéphane Airiau
Mathematical & Computer Sciences Department
University of Tulsa, USA
stephane@utulsa.edu

Sandip Sen
Mathematical & Computer Sciences Department
University of Tulsa, USA
sandip@utulsa.edu

ABSTRACT

Learning to converge to an efficient, i.e., Pareto-optimal Nash equilibrium of the repeated game is an open problem in multiagent learning. Our goal is to facilitate the learning of efficient outcomes in repeated plays of incomplete information games when only opponent's actions but not its payoffs are observable. We use a two-stage protocol that allows a player to unilaterally commit to an action, allowing the other player to choose an action knowing the action chosen by the committed player. The motivation behind commitment is to promote trust between the players and prevent them from mutually harmful choices made to preclude worst-case outcomes. Our agents learn whether commitment is beneficial or not. Interestingly, the decision to commit can be thought of as expanding the action space and our proposed protocol can be incorporated by any learning strategies used for playing repeated games. We show the improvement of the outcome efficiency of standard learning algorithms when using our proposed commitment protocol. We propose convergence to Pareto optimal Nash equilibrium of repeated games as desirable learning outcomes. The performance evaluation in this paper uses a similarly motivated metric that measures the percentage of Nash equilibria for repeated games that dominate the observed outcome.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms

Keywords

Repeated Game, Learning, Commitment

1. INTRODUCTION

A rational agent, playing an iterated games, tries to maximize 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 action combination. Both agents can make concessions from greedy strategies to improve their individual payoffs in the long run [12]. Reinforcement learning schemes, and in particular, Q-learning [17] have been widely used in single-agent learning situations. In the context of multi-player games, if one agent plays a stationary strategy, the stochastic game becomes an MDP 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 assumption does not hold any longer. In addition, it is no longer clear what an optimal strategy is. Researchers have focused on convergence to Nash equilibrium (NE) in self-play, where each player is playing a best response to the opponent strategy and does not have any incentive to deviate from his strategy.

Convergence is a desirable property in multiagent systems, but converging just to any NE is not the preferred outcome since NE is not guaranteed to be Pareto optimal (an outcome is Pareto optimal if no agent can improve its payoff without decreasing its opponent's payoff). For example, the widely studied Prisoner's Dilemma game (PD in Table 1(b)) has a single pure strategy NE that is defect-defect, which is dominated by the cooperate-cooperate outcome. A Pareto Optimal outcome may not be appealing to players if that outcome is also not a NE, i.e., there might be incentives for one agent to deviate and obtain higher payoff. For example, each agent has the incentive to deviate from the cooperate-cooperate Pareto optima in PD. In repeated games, folk theorems[9] ensure that, when players are "patient enough", any payoff dominating a reservation payoff can be sustained by a NE. Hence, in repeated games, there are Pareto Optimal outcomes that are also NE outcomes.

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.

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

We are interested in developing mechanisms by which agents can produce PONE outcomes. [13] provides a solution under complete knowledge. This assumption is unrealistic in most cases: opponent valuation is in general intrinsic and private. Moreover, payoff communication opens the door for deceptive behavior. Hence, we believe that not observing the opponent payoff is a more realistic assumption. We are interested in two-person, general-sum games

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.

where each agent only gets to observe its own payoff and the action played by the opponent, but opponent's payoff is unknown. Under these conditions, it may be difficult to guarantee convergence to a PONE. In order to compare the performance of different algorithms that are trying to converge to a PONE, we introduce a new metric: given an outcome of the game, the metric relates to the relative number of outcomes dominating the current outcome.

2. RELATED WORK

Researchers have focused on convergence to NE in self-play. This emphasis on convergence of learning to Nash equilibrium is rooted in the literature in game theory [8] where techniques like fictitious play and its variants lead to NE convergence under certain conditions. More recently, multiagent learning researchers have also adopted convergence to NE as the desired goal for a rational learner [6, 13]. By modeling the opponent, Joint-Action Learners [5] converge to a NE in cooperative domains. By using a variable rate, WoLF [3] is guaranteed to converge to a Nash equilibrium in a two-person, two-actions repeated 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 [6] is guaranteed to learn to play optimally against stationary opponent and to converge to NE in self-play.

In [15], Powers and Shoham propose new criteria for learners in a MAS: converging to near best response against any stationary players, converging to a PONE in self play, and close to minimax payoff against any other players. They propose an algorithm that meets these criteria. It requires, however, knowledge of the opponent's payoff. This is not the case in [7] where Crandall and Goodrich have a similar goal to our work. They propose an algorithm that guarantees an outcome that is not lower than the minimax outcome (this outcome can be sustained by a NE). Moreover, they propose that a learner should learn to accept compromises that increase their average payoff (Compromise/cooperate property). Although they cannot guarantee this property, they present empirical results showing convergence to Pareto efficient outcome in many games (e.g. PD, Stag Hunt, Chicken).

We had previously proposed a modification on the simultaneous-move game playing protocol that allowed an agent to communicate to the opponent its irrevocable commitment to an action [1, 16]. 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. Our use of commitment is different from the use of commitment in [11] where players cannot observe the actions of other players, and they commit to play the same action for a sequence of time slots. In [1], we compared the outcome obtain by an had hoc learner with the outcome of a NE of the stage game (or one shot game). In this paper, we show that the commitment protocol can be used with arbitrary multiagent reinforcement learning algorithms and that it facilitates convergence to near-efficient Nash equilibria of the repeated games and not just to efficient Nash equilibria of single-stage games. In addition, under the assumption that players are greedy, we provide a correspondence between learning in the traditional protocol and learning in the commit protocol. Finally, we propose a metric based on the folk theorem that relates to the relative number of outcomes dominating the reservation outcome and the current outcome.

3. EQUILIBRIUM IN REPEATED GAMES

To motivate our metric and the importance of considering equilibrium of the repeated game in multiagent learning, we review the equilibrium concepts in the context of repeated games. We are interested in repeated games where the agents play a normal form game (called the stage game), infinitely and try to optimize the average payoff received. **Notation:** In the following, we consider an $n \times n$ two-player game that can be represented by two matrices. r and $\mathcal{R}$ (respectively c and $\mathcal{C}$) denote the row player and its payoff matrix (resp the column player and its payoff matrix), and p_r (resp p_c) is the mixed strategy of the row (resp column) player. We will use $\mathcal{R}(a_r, a_c)$ (resp $\mathcal{R}(p_r, p_c)$) to denote the payoff received by the row player when r plays a_r and c plays a_c (resp the expected utility of the row player when it uses the strategy p_r and its opponent plays strategy p_c).

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

(a) Battle of the Sexes

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

(b) Prisoners' dilemma

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

3.1 Outcome candidates for equilibrium

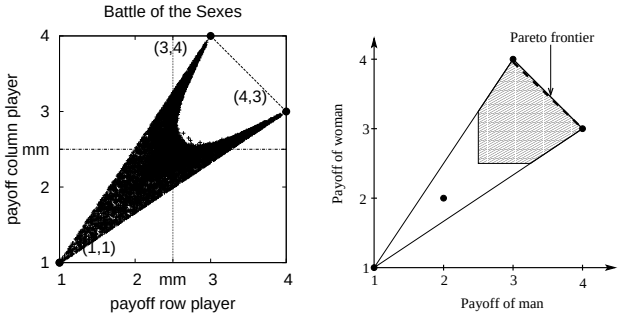
For any (infinite) history of play, we can compute the proportion of each pair of payoffs obtained by the players. The average payoff obtained is a convex combination of the pairs of payoffs of the game: $\mathcal{V}(H) = \{(\mathcal{R}(i,j), \mathcal{C}(j,i)) | (i,j) \in [1..n]^2\}$. Hence, all possible payoffs of the repeated game can be represented by the convex hull $\mathcal{H}$ with vertices in $\mathcal{V}(H)$.

If no communication is allowed during the play of the game, the players choose their strategies independently. Note that all the points of the convex hull cannot be produced by independent mixed strategy. The concept of correlated equilibrium [2] 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. [10] introduces algorithms that empirically converge to a correlated equilibrium in a testbed of Markov games. 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 in going to different places, but both prefer to be together than being alone. The best (and fair) solution would consists in alternating between (Coordinate, Defect) and (Defect, Coordinate) to obtain an average payoff of 3.5. However, playing independent uniform strategy leads to an average payoff of 2.5. To avoid bad outcomes, players can use the observation of a public random variable to coordinate their actions. The convex hull containing all possible payoff of the repeated game is a triangle represented in Figure 1. The shaded area inside the triangle in Figure 1(a) is the payoff pairs that can be obtained by players using independent mixed strategies. In this game, a large portion of high payoffs for the row and the column player cannot be reached using independent mixed strategies.

Each player can guarantee a minimum payoff by playing its *maxmin* strategy. The payoff of the minimax equilibrium is defined by:

$$v_r = \min_{p_c} \left\{ \max_{p_r} \mathcal{R}(p_r, p_c) \right\} \text{ for the row player}$$

$$v_c = \min_{p_r} \left\{ \max_{p_c} \mathcal{C}(p_c, p_r) \right\} \text{ for the column player}$$



(a) Possible payoff for independent mixed strategy

(b) $\mathcal{F}$ (outcome that can be sustained by a NE in repeated game) for Battle of sexes ($v_r = v_c = \frac{5}{2}$)

Figure 1: Payoff of Battle of sexes

Hence, not all points in $\mathcal{H}$ are candidates to be equilibrium outcomes: only outcomes that dominates the minimax value are of interest. The region of feasible payoff $\mathcal{F}$ is defined by

$$\mathcal{F} = \{(x, y) \in \mathcal{H} | x \geq v_r, y \geq v_c\}.$$

In the example of the Battle of Sexes, the minimax value for both player is $\frac{5}{2}$ and occur when both players play the mixed strategy $(\frac{3}{4}, \frac{1}{4})$. The feasible payoffs $\mathcal{F}$ are represented by the shaded area in Figure 1(b). Note that most of these points are not reachable when both players play independent mixed strategy.

3.2 Pareto and Nash Equilibrium

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 Y if all agents receive a higher utility in X compared to Y. An outcome X *weakly dominates* (or simply *dominates*) another outcome Y 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. In the 2-dimensional representation, if there are no points "above" and "on the right" of a point x, x is Pareto optimal. The Pareto optimal outcomes are located at the edges $\mathcal{P}$ of the convex hull $\mathcal{H}$. The set of Pareto optimal points of the battle of sexes is represented in Figure 1(a).

Strategies are in *Nash equilibrium* when they are mutual best responses. Assuming that the opponent plays its component of the Nash equilibrium, the player cannot do better than playing its own component. For any single shot game, there exists at least a mixed strategy NE. For the battle of sexes, there are two pure NE which are (Cooperate, Defect) and (Defect, Cooperate) and a mixed strategy NE which is the minimax equilibrium $(\frac{3}{4}, \frac{1}{4})$. For infinitely repeated games, the situation is very different since there is an infinite number of Nash equilibria. A set of "folk theorems" ensure that if players are sufficiently patient, for each feasible payoff $v \in \mathcal{F}$, there is a NE of the repeated game with payoff v. The idea behind the theorem is exploited by Littman and Stone in [13] where they introduce an algorithm to converge to a Pareto-optimal NE. In their approach, if a player deviates from the equilibrium with outcome $v \in \mathcal{F}$, it will be punished by playing the minimax strategy long enough. The punishment is designed so as to make it irrational to deviate from the chosen equilibrium.

Hence, in the repeated game, any point in $\mathcal{F}$ is an outcome of a NE. Points in $\mathcal{F}$ and not Pareto optimal are by definition dominated, which make them poor candidate for good equilibrium points. The NE that are also Pareto optimal, hence points in $\mathcal{S}_{PONE} = \mathcal{F} \cap \mathcal{P} \neq \emptyset$ are preferable. A bargaining argument found in [14] highlights a best candidate, and their algorithm converge to one particular NE on the Pareto frontier.

3.3 Metric for a two-player game

To compare two equilibrium outcomes, we can use the concept of dominance. However, when there is no dominance between the outcomes, additional metrics are needed. Investigating the sum of the payoff of the player (a measure of the social welfare), or the product of the payoff (a measure of fairness) provides insight to the equilibrium properties of the learning algorithms. Another approach is to consider the number of equilibria that dominate the current equilibrium: the fewer outcomes that dominate the current outcome, the closer this outcome is to a Pareto Optimum. The folk theorems [9] ensure that when an outcome dominates the minimax outcome, it can be sustained by a NE of the repeated game. For an outcome x, let $d(x)$ denotes the area containing all points that dominates x in the payoff space. If $d(x) = 0$ and x dominates the minimax outcome, then x is a PONE.

Definition 1. Performance metric of an equilibrium outcome x: $\delta(x) = \frac{d(x)}{d(x_{mm})}$ where $x_{mm} = (v_r, v_c)$ is the minimax outcome.

$\delta(x)$ represents the proportion of NE outcomes of the repeated games that dominates x. The smaller $\delta(x)$, the better the outcome x is with respect to convergence to a PONE. When one outcome x dominates an outcome y, $\delta(x) < \delta(y)$. The opposite is not true: when there is no dominance between x and y, $\delta(x)$ may be less, equal, or greater than $\delta(y)$.

4. COMMITMENT

We now present our proposed commitment protocol that can be added onto any stage game playing algorithm. The motivation behind the protocol is for agents to improve payoffs by building trust via up-front commitment to "cooperating" moves that can be mutually beneficial, 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 use an augmented game playing protocol where the players are allowed to announce the action they are going to play. The first effect of this modification of the simultaneous play protocol is to increase the space of possible payoff since players can play some correlated equilibrium[1]. For example, in the battle of the sexes game, it is possible to reach the fair equilibrium where both players gets a reward of 3.5. Commitment to an action can also reduce some uncertainty and can help players to reach better outcome. In the remaining of this paper, we show that myopic exploitation of a commitment can improve the outcome of the game, but non-myopic solutions are needed to reach a PONE.

[1]we have not proved that all possible outcome are possible, and it is not clear whether all correlated equilibrium can be reached with these assumptions

4.1 Protocol

We build on the simultaneous revelation protocol [1, 16]. Agents repeatedly play an $n \times n$ bimatrix game. At each iteration of the game, each player first announces whether or not it wants to commit to an action. If both players want to commit at the same time, one is chosen randomly. If no player decides to commit, then both players simultaneously announce their action, as in the traditional simultaneous play protocol. When one player commits to an action, the other can choose any action given its opponent's action. Each agent can observe which agent actually revealed, and which action the opponent played. In this paper we consider two-player games where agents play best response action to opponent's committed action. We believe that this protocol can be easily extended to a n-player game with $n \geq 2$ when only one player commits to an action.

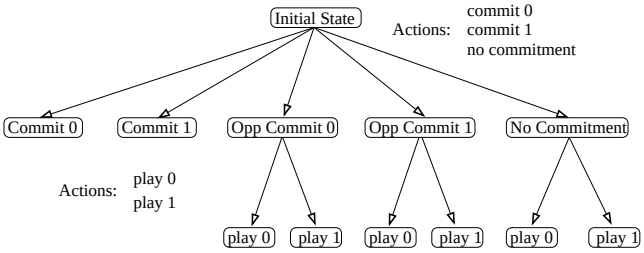


Figure 2: Game tree for a two-action game.

Such games can be represented by game trees, e.g., Figure 2 presents the tree for a two-action game. In the initial state, the agents have $n+1$ actions: it can plan to commit to any of the n actions of the game, or decide not to commit. The transition from the root of the tree depends on the decision of the opponent. The *commit* states are reached when the player commits and the opponent does not, or when both players are willing to commit, but the player wins the toss. From the *commit* state, no further decision is needed, and the payoff received depends on the play of the opponent. When the player decides not to commit, the transition can lead to any one of the n states where the opponent commits or to the state where no players is willing to commit. In both cases, the player has n actions available. From the *opp commit* states, the transition depends only on the current players' decision. From the state where there is no commitment, the transition also depends on the opponent decision. Any multiagent learning algorithm can be used to estimate the utility of different actions, including the commitment actions, from repeated play against an opponent.

Definition 2. A **pure strategy** has one of the following form:
• "does not want to commit, when other player does not announce, play action k" that we denote by ($\neg$, k).
• "want to commit to action k" that we denote by (k, -).

4.2 Examples

The following examples illustrate the possible effects of playing with the possibility to announce. We used matrices from the testbed introduced by Brams in [4]. These examples shows that in some cases, playing with announce is beneficial, and in other cases, different equilibrium can be reached. In the first two examples we show that one or both player announcing its play can be benefited. In the last two examples, we show that announcing may improve on the outcome of the NE of the stage game, but the equilibrium may not be a PONE.

For game 27 presented in Table 2(a), there is a unique Nash equilibrium in the single shot game where both players play action 0 with an outcome of $(2, 3)$. Note that the NE is dominated by the pure strategy s^* where both agents play action 1 with outcome $(3, 4)$. Assume that the column player plays the strategy $(\neg, 0)$. The row player can get 3 by playing $(\neg, 0)$ or $(0, -)$. More interestingly, if the column player plays a best response to the committed action, the row player can obtain 4 by committing to action 1 (i.e. playing $(1, -)$). Note that this solution is also beneficial for the row player that gets 3. For this game, the myopic exploitation results in a beneficial outcome for both players.

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

(a) Game 27

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

(b) Game 50

Table 2: Payoffs Matrices of Two games in Brams' Testbed

The second example is the battle of the sexes of Table 1(a). The strategies where both players play $(0, -)$ is in NE. Half of the time, a player receives 3, and half of the time, it receives 4. This game exploits the possibility to play a coordinated equilibrium.

For Game 50 represented in Table 2(b), there is a single pure NE of the stage game where both players play action 0. Note that this solution is a PONE. The row player gets only its third preferred outcome, when the column player gets its most preferred one. When the players are allowed to commit and are myopic (they will exploit the commitment of the opponent by playing a best response), the row player can get a payoff of 3 by playing $(1, -)$, since the myopic column player will respond by playing action 1. This situation does not benefit the column player that, with the same argument, can get 4 by playing $(0, -)$. By playing myopically, the agents will be in a correlated equilibrium where they gets (2,4) and (3,2) with equal probability. Note that this equilibrium is not a PONE, since the correlated equilibrium is strongly dominated by the strategies $(\neg, 0)$ for the row and $(\neg, 1)$ for the column. In this case, committing to an action improves the outcome of the row player, but decreases the payoff of the column.

Finally, we consider the Prisoners' dilemma game in Table 1(b). In this game, if the agents are myopic, a commitment does not provide any advantage: if a player commits to play cooperate, the opponent greedily exploits the situation by playing defect. The correlated equilibrium where both agents reveals cooperate provides better results than the NE, but it is still dominated by the $(\neg, C)$ $(\neg, C)$ payoff. If a players commits to play defect, the best response is also to play defect. Hence, if the players are limited to play a best response when an agent reveals its action, they can improve on the NE, but the equilibrium reached is not a PONE. These last two games illustrate that non-myopic exploitation of a commitment is needed to improve the payoff of both players.

5. ESTIMATING PAYOFFS

Learning in repeated games can be viewed as a reinforcement learning task where, at each repetition t of the stage game, the player chooses a course of actions and gets a reward r_t for it. Players discount the future utilities using a discount factor $\gamma \in [0, 1]$ and try to maximize the sum $u = \sum_{t=0}^{\infty} \gamma^t r_t$. A simple, model-free online technique for reinforcement learning is Q-learning [17]. The update rule for Q-learning when a learner played action a in

state $\mathcal{S}$, and observe the reward r and the new state $\mathcal{S}'$ is

$$Q(a,\mathcal{S}) \leftarrow \alpha Q(a,\mathcal{S}) + (1-\alpha)\left(r + \gamma \max_{b \in Action(,\mathcal{S}')} Q(b,\mathcal{S}')\right).$$

The parameter α is the learning rate that controls the importance of the new information compared to past information.

Q-learning can learn payoff in a Markov Decision Process (MDP). When both players are learning, the Markovian assumption is violated. Because Q-learning updates has been used in multiagent learning, we use this method to estimate the payoff online.

In learning the game tree of Figure 2, the reward provided to the players are the payoffs of the stage game. For the terminal states which are successors of the states *Opp Commit* i, the utility of these states can be learned. For any other state, the payoff depends on the policy of the opponent. Assume that both players play a static strategy and that the Q-values have converged. If the players are no longer exploring, greedy exploitation of these values results in playing a best response, since the player will try to optimize their expected values. In particular, when players use a greedy exploitation, they will play a best response to a commitment. Exploration schemes such as the ϵ-greedy, the use of Boltzmann probability distribution, or the use of probability distribution learned by WoLF will learn to play a best response to a commitment. In the following, we will use this hypothesis to reason about the play of the game. Yet we recognize that to avoid myopic behavior, a learner should not use a greedy exploitation.

If we assume that the opponent plays a best response to a committed action, given an $n \times n$ game G_c played in the protocol with commitment, it is possible to build a game G_{eq} played with the traditional simultaneous game protocol: each player can directly play one of the $2n$ pure strategies available. An example when $n = 2$ is provided in Table 3 and can be extended to any n. Note that if we relaxed the assumption of playing a best response to a committed action, the payoff of all cells where at least one agent commit would depend on the action of the opponent. In this case, a learner could still use this table to learn its expected payoff.

	0	1
0	$a_{0,0}$, $b_{0,0}$	$a_{0,1}$, $b_{0,1}$
1	$a_{1,0}$, $b_{1,0}$	$a_{1,1}$, $b_{1,1}$

equivalent to

	$(\neg, 0)$	$(\neg, 1)$	$(0, -)$	$(1, -)$
$(\neg, 0)$	$a_{0,0}$, $b_{0,0}$	$a_{0,1}$, $b_{0,1}$	BR(-,0)	BR(-,1)
$(\neg, 1)$	$a_{0,0}$, $b_{0,0}$	$a_{0,1}$, $b_{0,1}$	BR(-,0)	BR(-,1)
$(0, -)$	BR(0,-)	BR(0,-)	BR(0,0)	BR(0,1)
$(1, -)$	BR(1,-)	BR(1,-)	BR(1,0)	BR(1,1)

where:

- BR(i,-) is the pair of payoff where row commits to i and column plays the best response to i
- BR(-,j) is the pair of payoff where column commits to j and row plays the best response to j
- BR(i,j) is the average pair of payoff of BR(i,-) and BR(-,j)

Table 3: Equivalence of games in the traditional protocol and the commit protocol when agents are greedy

Compared to NE outcome of a traditional protocol, the NE outcome with the commit protocol may differ. We hypothesize that, under rational play, the outcome of a game played with the commit protocol is not strictly dominated by the outcome of the game played with the traditional protocol. Assume that players are in a NE of the stage game and are provided the opportunity to commit. A player i commits only when it is beneficial, hence getting a higher payoff. If the other player j is improving due to the commitment, both players improve their respective payoffs. Else, j's payoffs is worse. In this case, j may improve by committing, which might decrease i's payoff. If on average both players' payoffs decrease, the players will ultimately learn not to reveal. When i commits and j cannot improve its payoff by committing, e.g. committing to any action yields a lesser payoff, the players reached a different equilibrium (i improves and j is worse off but there is no dominance). In any case, players should only benefit from the commit protocol.

6. RESULTS

We compared the use of the protocol with commitment with the traditional protocol of simultaneous play on various set of matrices. We first experiment with the testbed proposed by Brams in [4] which represents all the conflicted 2x2 games with ordinal payoff. We then compared the results on a set of random matrices.

Any traditional algorithm for game playing can be used to learn the game tree of Figure 2. For reason of simplicity, we use the assumption that the players learn best response when an opponent commits, and we used the equivalent matrix presented in Table 3. We chose to use WoLF-PHC[2] (Win or Learn Fast - policy hill climbing) [3] as the learning algorithm. The algorithm learns mixed strategy and is guaranteed to converge to a NE in a 2-person, 2-actions repeated game. The outcome of a play the traditional protocol (resp. the commit protocol) are referred as to WoLF (resp. WoLF(commit)).

6.1 Testbed of 2x2 conflicted games

We first use a neutral but extensive testbed of games introduced by Brams in [4]: the testbed is composed of all possible conflicting situations that can occur in a two-action two-player game with a total preference order over the four outcomes of the game. This testbed represents a wide variety of situations, including often-studied games like PD, the chicken game, battle of the sexes, etc. We use the numbers 1, 2, 3, 4, as the preference of an agent for a state in the 2x2 matrix, with 4 being the most preferred. Though these numbers correspond to ordinal payoff, we treat them as cardinal payoffs. There is no game where agents can simultaneously obtain their most preferred outcome, which implies that each game represents a conflicting situation. There are 57 structurally different 2x2 conflict games (no two games are identical by renaming the actions or the players). Learners typically have access to only their own payoff matrices but can observe opponent actions. Lack of knowledge of opponent payoff is a more realistic assumption in an open environment, but puts the learners at a disadvantage compared to the static players.

Among the game of the testbed, 51 games have a unique NE (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 NE). Of the 42 games that have a unique pure strategy NE, 4 games have a pure NE that is not Pareto-optimal (the prisoners' dilemma, game 27, 28 and 48 have a unique NE which is dominated), and 2 games which have a single mixed strategy NE are dominated by a pure strategy.

[2]WoLF-PHC settings: $\alpha(t) = \frac{1}{10+\frac{t}{100}}$, $\delta_W = \frac{1}{10+t}$, $\delta_L = 4\delta_W$. The games were played over 10,000 iterations, and results were averaged over 40 runs.

In five games, the outcome of WoLF(commit) strictly dominates the outcome of WoLF. Three of them are games where the NE is dominated (games 27, 28, 29 and 48). The remaining two games are the games where the NE is a mixed strategy NE dominated by a pure strategy. In 9 other games, the equilibrium reached is different than the NE of the stage game, but there is no dominance. We found that the augmented mechanism fails to produce a Pareto optimal solution in only two games: the prisoner's dilemma game (Table 1(b)) and game 50 (Table 2(b)).

6.2 Results on randomly generated matrices

As shown in the previous experiments, the structure of some games can be exploited by the commit protocol to improve the payoff of both players. To evaluate the effectiveness of the protocol on a more general set of matrices, we ran experiments on randomly generated matrices as in [16]. We generated 1000 matrices of sizes 3x3, 5x5 and 7x7. Each matrix entry is sampled from a uniform distribution in $[0, 1]$. We compare the outcome of WoLF(commit) and WoLF.

In Figure 3, we plot different areas: the average area containing all the outcome of NE (i.e. dominating the minimax outcome), the area that dominates the outcome of the traditional and the commitment protocol. We first observe that the minimax outcome is dominated by more outcomes for larger games, i.e. the space of NE is larger. When we compare with the area that dominates the outcome found by WoLF we find that the outcome with the protocol with commitment is dominated by less outcomes, and the difference increases with the game size. In Figure 4, we plot our δ metric that provides the percentage of sustainable NE of the repeated game that dominates the outcome of the algorithm. The plot indicates that the outcome obtained with protocol with commitment is dominated by at most 10% of the possible NE, when the outcome of the traditional simultaneous game is dominated by 3 times more NE. This suggests that the commitment protocol produces more efficient equilibrium than the traditional simultaneous game protocol.

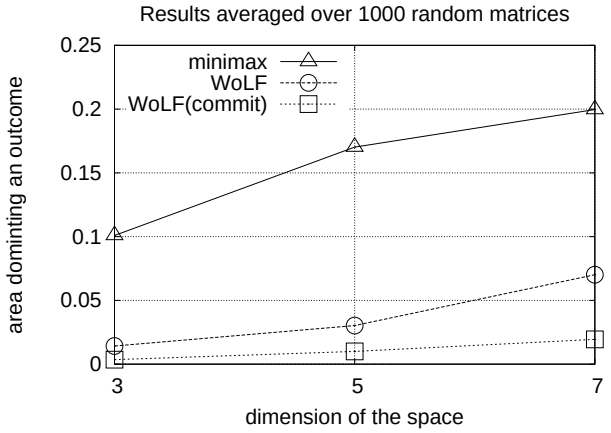


Figure 3: Results over randomly generated matrices: area of the points that dominates the minimax outcome, WoLF and WoLF(commit)

7. CONCLUSION AND FUTURE WORK

In this paper, we built on a previous algorithm from [1, 16] with the goal of producing PONE outcomes in repeated single-stage games. We propose a metric that can be used to measure the quality of an outcome: it represents the relative number of Nash equilibria

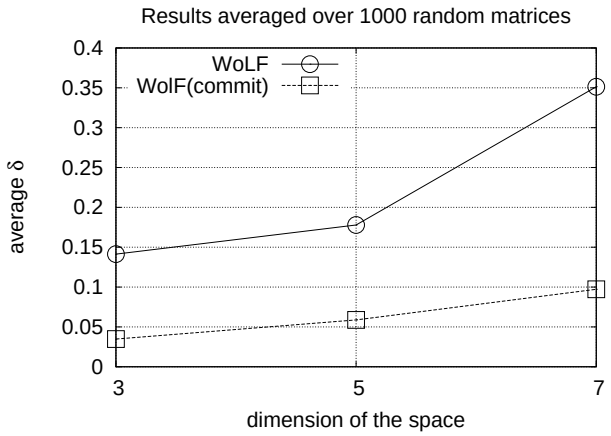


Figure 4: Results over randomly generated matrices: metric δ for WoLF and WoLF(commit)

of the repeated game that dominate the outcome reached. Under the assumption that the opponent payoff matrix is unknown, it might be difficult to ensure convergence to a PONE. Our proposed metric is helpful in comparing the relative efficiency of different outcomes.

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. The opportunity of revealing its action should not be seen as making a concession to the opponent, but rather as a means to explore the possibility of mutually beneficial outcomes. Any learning algorithm can be augmented to incorporate the commit protocol, which improves the payoffs in most cases: we empirically show that our protocol improve the payoffs obtained by WoLF-PHC in a variety of games. The experiments also show shortcomings of the current commitment protocol in that it fails to reach PONE outcomes: the primary reason for this is that a player responds to a commitment with a myopic best response.

We assume that a player does not know the payoff matrix of the opponent, which makes it difficult to estimate whether the equilibrium reached is acceptable for both players. In particular, there are situations where not playing a best response to a committed action can be beneficial for both players. To find a non-myopic equilibrium, an agent should not be too greedy! Currently, the agents are learning only their own payoff, and learn to play a best response to a committed action. We are working on learning action-utility estimates that incorporates an estimate of the preference of the opponent in the game tree presented in Figure 2. We expect that the agents will be able to more consistently discover states beneficial for both learners, and thereby converge to PONE outcomes.

Acknowledgment: US National Science Foundation award IIS-0209208 partially supported this work.

8. REFERENCES

[1] S. Airiau and S. Sen. Learning pareto optimal solutions in 2x2 conflict games. In *Lecture Note in Artificial Intelligence (LNAI 3898) AAMAS-05 Workshop on Learning and Adaptation in MAS (LAMAS), 2005*, 2005.

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

[3] M. Bowling and M. Veloso. Multiagent learning using a

variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[4] S. J. Brams. *Theory of Moves*. Cambridge University Press, Cambridge: UK, 1994.

[5] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, Menlo Park, CA, 1998. AAAI Press/MIT Press.

[6] V. Conitzer and T. Sandholm. 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.

[7] J. W. Crandall and M. A. Goodrich. Learning to compete, compromise, and cooperate in repeated general-sum games. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 161–168, New York, NY, USA, 2005. ACM Press.

[8] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

[9] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

[10] A. Greenwald and K. Hall. Correlated-q learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 242–249, 2003.

[11] S. Kapetanakis, D. Kudenko, and M. Strens. Learning of coordination in cooperative multi-agent systems using commitment sequences. *Artificial Intelligence and the Simulation of Behavior*, 1(5), 2004.

[12] M. L. Littman and P. Stone. Leading best-response strategies in repeated games. In *IJCAI Workshop on Economic Agents, Models and Mechanisms*, 2001.

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

[14] J. F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, April 1950.

[15] R. Powers and Y. Shoham. New criteria and a new algorithm for learning in multi-agent systems. In *Proceedings of NIPS*, 2005.

[16] S. Sen, S. Airiau, and R. Mukherjee. Towards a pareto-optimal solution in general-sum games. In *Proceedings of the Second International Joint Conference On Autonomous Agents and Multiagent Systems*, pages 153–160, 2003.

[17] C. J. C. H. Watkins and P. D. Dayan. Q-learning. *Machine Learning*, 3:279 – 292, 1992.