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 repated 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 in iterated stage games, should maximize long-term 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. Reinforcement learning schemes, and in particular, Q-learning [6], 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, and Q-learning is not guaranteed to converge in self play. In such cases, 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) 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[3] 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. We believe that 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. [4] 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 where each agent only gets to observe its own payoff and the action played by the opponent, but opponent's payoff is unknown. Under

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

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 states dominating the current outcome.

2. 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 build on the simultaneous revelation protocol [5]. 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, agents play best response action to opponent's committed action.

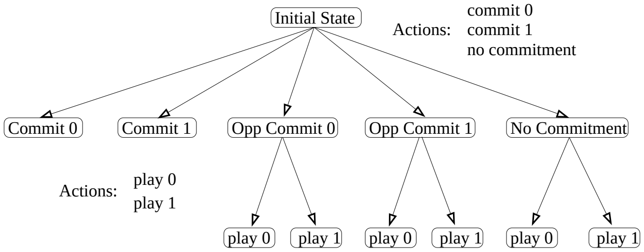

Figure 1: Game tree for a two-action game.

Such games can be represented by game trees, e.g., Figure 1 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.

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.

3. RESULTS

We compared the use of the commit protocol with the traditional protocol of simultaneous play on various set of matrices. Any traditional algorithm for game playing can be used in combination with the commit protocol. We chose to use WoLF-PHC[1] (Win or Learn Fast - policy hill climbing) [1] 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.

3.1 Metric

To compare the 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 [3] 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} 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 $\delta(y)$.

3.2 Testbed of 2x2 conflicted games

We first use a neutral but extensive testbed of games introduced by Brams in [2]: 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. There is no game where agents can simultaneously obtain their most preferred outcome, which implies that each game represents a

[1] 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.

conflicting situation. There are 57 structurally different 2x2 conflict games (no two games are identical by renaming the actions or the players). In five games, the equilibrium reached by WoLF(commit) strictly dominates the equilibrium reached with WoLF. Three of them are games where the NE is dominated. The remaining two are the games where the NE is a mixed strategy NE dominated by a pure strategy. In nine other games, the outcome obtained by WoLF(commit) is different than the outcome of the NE of the stage game played with the traditional protocol, but there is no dominance (one player gains and the other looses). We found that the use of the commit protocol fails to produce a Pareto optimal solution in only two games, one of which is the prisoner's dilemma game.

3.3 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 [5]. 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 the top plot of Figure 2, 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 the bottom plot of Figure 2, we plot our δ metric that provides the percentage of sustainable NE of the repeated game that dominates the outcome of the algorithm. The chart 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.

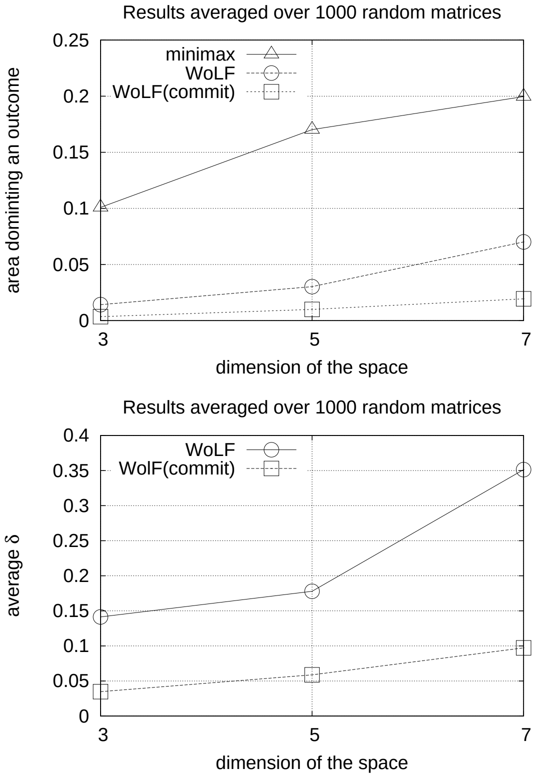

Figure 2: Results over randomly generated matrices.

4. CONCLUSION AND FUTURE WORK

In this paper, we built on a previous algorithm from [5] 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 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 1. 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.

5. REFERENCES

[1] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

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

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

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

[5] 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.

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