Multiagent Learning

Doran Chakraborty

April 19, 2007

1 Introduction

One of the greatest difficulties about multiagent learning is that the environment is not stationary with respect to the agent. In case of single agent learning problems, the agent has to maximize its expected reward with respect to an environment which is stationary. In case of multiagent scenarios, all the agents learning simultaneously poses a problem of non-stationarity in the environment which other agents have to take into account while computing their optimal behavior in such situations. One of the popular frameworks of addressing the problem of multiagent learning is the framework of stochastic games (SG) introduced by Shapeley [9]. In the following section we would emphasize on the SG framework and some of the well known algorithms that try to address the problem of multiagent learning in the SG Model.

2 Stochastic Games

A stochastic game is a tuple $(n, S, \{A_i\}, T, \{R_i\})$ where n is the number of agents in the system, S is the set of states in the system, A_i is the set of actions that agent i can take, T is the transition function $S \times A \times S \rightarrow [0, 1]$ and R_i is the reward function for agent i. The role of each agent is to maximize its total reward if the interaction is for a limited time period T or maximize the γ discounted reward if the interaction is for an infinite horizon. The solution concept for optimality is Nash Equilibrium [10]. A Nash equilibrium is a collection of strategies for each agent such that each agent 's strategy is the best response strategy with respect to strategies of all other agents in the system. Nash showed that there always exists a Nash equilibrium in single stage games (possibly multiple) though that equilibrium may not be in pure strategies. For example in the game of matching pennies the only Nash equilibrium is when both agents play head or tail with probability 0.5. The universal existence of the Nash Equilibrium is not restricted to only single stage games but also in stochastic games as proved by Shapley [9]. Now we present a brief overview of the popular learning approaches and their restrictions from the literature.

2.1 Minimax-Q

Littman introduced this reinforcement learning algorithm in 1994 that efficiently computes the Nash equilibrium strategy for the agents in a zero-sum SG [8]. The algorithm assumes that the game is of complete (the agents know their payoff and hence their opponents) and perfect (the agents can see each other actions at each time step) information. At each time step the agent gets an observation $< s, a, s', r >$ consisting of a state, a joint-action, the next state and the reward which its uses to update its Q-value pertaining to the joint action a at state s as,

$$Q(s,a) \quad \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma V(s'))$$

where

$$V(s') = max_{\sigma \in PD(A_1)} min_{a_2 \in A_2} \sum_{a_1 \in A_1} \sigma(a_1) \times Q(s, < a_1, a_2 >)$$

where the value of $V(s')$ can be solved in polynomial time using linear programming. Littman showed that the with the usual assumptions on exploration and learning rate made by the single agent Q-learning algorithm, this technique guarantees convergence to the unique Nash equilibrium in the SG game. Note, a zero sum single stage game has only one Nash equilibrium and the corresponding SG version of it also has just one Nash equilibrium which is the single stage Nash equilibrium played at every stage.

2.2 Q-Learning for general sum stochastic games (Nash-Q)

Hu and Wellman introduced this algorithm in 1998 that tries to compute a Nash equilibrium strategy for each agent in general sum games. Assumptions of complete and perfect information is maintained. The agent now has to maintain Q values for every other agent. The entry $Q^k(s,a)$ approximates the average discounted reward for agent k for playing action a in state s and then following the Nash equilibrium strategy from there onwards for the remaining stages. At each time step the agent gets an observation $< s, a, s', r^i >$ consisting of a state, a joint-action, the next state and the reward which its uses to update its Q-value pertaining to the joint action a at state s as,

$$Q^i(s,a) \quad \leftarrow (1-\alpha)Q^i(s,a) + \alpha(r + \gamma V^i(s'))$$

where

$$V^i(s') = Value^i[Q(s')] \tag{1}$$

where $V^i(s')$ operation computes the equilibrium value from state s'. The algorithm guarantees convergence to a Nash equilibrium under restricted settings of the occurrence of a unique globally optimal Nash equilibrium which is also a saddle point (the agent receives a higher payoff if other agents deviate from their policies).

3 Repeated Games

We know shift to a special class of SG's which we call repeated games (RG). RG is a special form of SG with just one state. So the definition of RG is the definition of SG without T. The agents keep playing the same single stage game infinitely or finitely. We now highlight some of the popular approaches and their merits and demerits when trying to solve the problem of computing the optimal strategy profile for RG's.

3.1 Fictitious Play (FP)

Fictitious Play [12] assumes opponents play stationary strategies which is an empirical distribution of the opponent's past actions. A variant of FP is Bounded FP (BFP) where the agent computes a belief of the opponent's past actions through a recent window of opponent actions. When a player has multiple best replies, it chooses each with a strict positive probability. It can be shown in fictitious play, the strategy profile converges to a Nash equilibrium if the game is iterated dominance solvable [6] or cooperative [3]. And in zero sum games, the empirical distribution of actions for the players converge to the unique mixed strategy Nash equilibrium, however the policies of the agents do not [12]. Smooth FP is a variant of FP that can play mixed strategies and thus can converge to mixed Nash strategy equilibriums [6].

3.2 WoLF PHC

Bowling proposed a new criterion for learning in multiagent settings in 2001 [2]. The new criterion stated that the learner should have the properties of convergence in self play and rationality. Convergence in self play means that the learners should converge to a stationary policy against each other in self-play while rationality means that the learners should converge to the best response when playing against stationary opponents. Both the criterion taken together suggest that the learners should converge to a Nash equilibrium in self play. He first introduced a simple Q-learning algorithm that can play mixed strategies [2] and uses hill climbing on the space of mixed policies to generate new mixed policies. The algorithm satisfied the properties of rationality against stationary opponents but failed to converge in self play. He then proposed the WoLF PHC (Win or Lost Fast Policy Hill Climber) algorithm that utilized variable learning rate to ensure convergence in self play. The rationale behind the algorithm is to learn fast when loosing (using a higher learning rate) and play cautiously while

winning so that the opponents can adjust and catch up (using a smaller learning rate). An important property of the algorithm is that each agent just has to know its own payoff and need not need to observe opponent's actions. Bowling further showed empirically that the algorithm converges to Nash equilibrium policy profile in two-action two-player repeated games.

3.3 Incremental Gradient Ascent Learning (IGA)

Singh and colleagues introduced a special class of learners called Incremental gradient ascent learners [13] which converge to Nash equilibrium in self play for a restricted class of two-action two-player games and for the rest converge to a an average payoff that can be sustained by some Nash equilibrium of the repeated game. A two-player two-action general sum game can be defined by the following matrix,

$$G = \begin{pmatrix} r_{11}, c_{11} & r_{12}, c_{12} \\ r_{21}, c_{21} & r_{22}, c_{22} \end{pmatrix}$$

where each entry is the corresponding payoff for the row and column player respectively. Let (α, β) be the mixed strategy played by the players, then the expected payoffs $V_r(\alpha, \beta)$ and $V_c(\alpha, \beta)$ for the row and column players respectively are,

$$\begin{aligned} V_r(\alpha, \beta) &= r_{11}\alpha, \beta + r_{22}(1-\alpha)(1-\beta) + r_{12}\alpha(1-\beta) + r_{21}(1-\alpha)\beta \\ V_c(\alpha, \beta) &= c_{11}\alpha, \beta + c_{22}(1-\alpha)(1-\beta) + c_{12}\alpha(1-\beta) + c_{21}(1-\alpha)\beta \end{aligned}$$

Now taking the partial derivatives of the expected payoffs with the mixed strategy gives rise to the following set of differential equations,

$$\frac{\partial V_r(\alpha, \beta)}{\partial \alpha} = \beta u - (r_{22} - r_{12}) \tag{2}$$

$$\frac{\partial V_c(\alpha, \beta)}{\partial \beta} = \alpha u' - (c_{22} - c_{12}) \tag{3}$$

where $u = (r_{11} - r_{22}) - (r_{12} + r_{21})$ and $u' = (c_{11} - c_{22}) - (c_{12} + c_{21})$. Clearly from equations 3 and 4, it is clear that the $\alpha^* = \frac{(c_{22} - c_{21})}{u'}$ and $\beta^* = \frac{(r_{22} - r_{21})}{u}$, is a Nash pair provided they satisfy legal probability distributions. The concept is to use constrained dynamics and project the values of α and β back inside the unit square to maintain legal probability distributions. The update rule used by IGA is given by,

$$\alpha_{k+1} = \alpha_k + \eta \times \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha_k} \tag{4}$$

$$\beta_{k+1} = \beta_k + \eta \times \frac{\partial V_c(\alpha_k, \beta_k)}{\partial \beta_k} \tag{5}$$

$$\tag{6}$$

where (α_0, β_0) is the arbitrary starting strategy pair and η is a very small learning rate. IGA assumes a full information game and one of the strict assumptions is that each agent at each iteration gets to observe the mixed strategy that its opponent is playing. Singh and colleagues used gradient dynamics in affine dynamical systems [7] to establish the convergence properties of IGA. Bowling and Veloso further extended the convergence properties of IGA using the WoLF principle [1]. They showed that using variable learning rate (which is the heart of the WoLF principle) does in fact guarantee convergence to a unique Nash equilibrium strategy profile in all two by two general sum games. The proof follows the convergence proof presented by Singh and colleagues but differs in one special case [1].

4 Learning in games with more than two players and against unknown opponents

Most of the work on multiagent learning has been based on convergence guarantees in self play and stationary opponents. In the previous section, we previewed a couple of algorithms that attempt to converge to a Nash equilibrium solution is case of self play. But their guarantees applied to a very restricted class of two-player two-action general sum games. Conitzer and Sandholm proposed Awesome in 2003 [4] that provided guarantees in self play for more than two players but their work assumed that the players either played the same algorithm or were stationary. But in reality such assumptions of self-play and stationarity are hard to implement in practice as there are no guarantees about the learning strategies of the opponents. There is a need to come up with learning algorithms that guarantee reasonable performance against any number of arbitrary opponents. Fudenberg and Levine [5] were the first to propose a couple of criterion that should be satisfied by a learning algorithm when playing against unknown opponents.

Safety: *The learning rule must guarantee at least the minimax payoff of the game.*

Consistency: *The learning rule must guarantee that it does at least as well as the best response (in the stage game) to the empirical distribution of play when playing against opponents whose play is governed by independent draws from any fixed probability distribution.*

The conditions coupled together is termed as *universal consistency.* The condition of consistency is often referred as no-regret where the agent does at least as better as the pure strategy best response against the empirical distribution of the opponent's actions. A limitation of this approach is that it makes sense to consider such a criterion in large population games where an agent's actions has no effect on opponent's actions. Vu, Powers and Shoham [14] proposed a new set of criteria for smaller games with fewer opponents,

Targeted Optimality: *Against any member of the target set of opponents, the algorithm achieves with ϵ of the expected value of the best response to the*

actual opponent

Auto Compatibility: *During self-play, the algorithm achieves within ϵ of the payoff of a Nash equilibrium that is not Pareto dominated by any other Nash equilibrium*

Safety: *Against any opponent, the algorithm always receives at least within ϵ of the security value for the game*

In addition, these requirements were required to hold with probability of at least 1 - δ after an initial polynomial period of time. They further extended the criterion against multiple opponents.

Targeted Group Optimality: *When each of the agents in the game is either a self-agent or in the target class, the payoffs of all the self-agents should be at least $V_g - \epsilon$ and within ϵ of an PO outcome, given the actual strategies of agents in the target class.*

Safety: *Against any set of opponents, the agent must achieve at least $V_g - \epsilon$*

where V_g is the maximin payoff of the agent. Then they proposed an algorithm *CORRSTRATEGY(S)* that achieves the convergence guarantees against stationary opponents. They further proposed another algorithm *CORRSTRATEGY(A)* that attempts to achieve the targeted group optimality criterion for adaptive opponents. Note the game is assumed to be of complete information to all agents and each agent can monitor the actions taken by other agents in the system. Shoham and Powers also proposed an algorithm that tries to achieve targeted optimality against adaptive opponents of fixed memory [11] in two player games.

5 Conclusion

In this paper we tried to review the different multiagent learning algorithms that have been proposed over the years. The review is by no means exhaustive but lists essential mile stones in the literature.

References

[1] M. Bowling and M. Veloso. Convergence of gradient dynamics with a variable learning rate. In *Proc. 18th International Conf. on Machine Learning*, pages 27–34. Morgan Kaufmann, San Francisco, CA, 2001.

[2] M. H. Bowling and M. M. Veloso. Rational and convergent learning in stochastic games. In *IJCAI*, pages 1021–1026, 2001.

[3] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI/IAAI*, pages 746–752, 1998.

[4] V. Conitzer and T. Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents, 2003.

[5] F. D. and L. D. Universal consistency and cautious fictitious play. In *Journal of Economic Dynamics and Control*, 1995.

[6] D. Fudenberg and D. K. Levine. In *The Theory of Learning in Games*, 1999.

[7] R. H. In *Differential Equations: Foundations and Applications*, 1987.

[8] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

[9] S. L.S. Stochastic games. In *Classics in game theory*, 1953.

[10] J. F. Nash, Jr. Equilibrium points in n-person games. In *Classics in game theory*, 1997.

[11] R. Powers and Y. Shoham. Learning against opponents with bounded memory. In *IJCAI*, pages 817–822, 2005.

[12] J. Robinson. An iterative method of solving a game. In *Annals of Mathematics*, 1951.

[13] S. Singh, M. Kearns, and Y. Mansour. Nash convergence of gradient dynamics in general-sum games. pages 541–548.

[14] T. Vu, R. Powers, and Y. Shoham. Learning against multiple opponents. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 752–759, New York, NY, USA, 2006. ACM Press.