In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems*, pp. 1583-1584, Toronto, Canada, May 2010.

Online Model Learning in Adversarial Markov Decision Processes

(Extended Abstract)

Doran Chakraborty
Department of Computer Science
University of Texas, Austin
chakrado@cs.utexas.edu

Peter Stone
Department of Computer Science
University of Texas, Austin
pstone@cs.utexas.edu

1. ABSTRACT

Consider, for example, the well-known game of Roshambo (Figure 1), or rock-paper-scissors, in which two players select one of three actions simultaneously. One may know that the adversary will base its next action on some bounded sequence of the past joint actions, but may be unaware of its exact strategy. For example, one may notice that every time it selects P, the adversary selects S in the next step; or perhaps whenever it selects R in three of the last four steps, the adversary selects P 90% of the time in the next step. The challenge is that to begin with, neither the adversary function that maps action histories to future actions (may be stochastic), nor even how far back it looks back in the action history (other than an upper bound) may be known. At a high level, this paper is concerned with automatically building such predictive models of an adversary's future actions as a function of past interactions.

Categories and Subject Descriptors

I.2 [**Computing Methods**]: Artificial Intelligence

General Terms

Algorithms, Performance

Keywords

opponent modeling

2. INTRODUCTION

Modeling memory-bounded (a.k.a adaptive [1]) opponents,[2] has received significant attention in the past for two main reasons [1, 5]. First, if we consider opponents whose future behavior depends on the entire history, we lose the ability to (provably) learn anything about them in a single repeated game, since we see a given history only once. The concept of memory-boundedness limits the opponent's ability to condition on history, thereby giving us a chance to learning its policy. Second, there exists an abundance of opponents in the game theory literature which are adaptive. Trigger strategies [4], fictitious play with bounded recall [6], and the polynomial time algorithm that forces the opponent to play the pareto-optimal Nash equilibrium [3] are a few examples of this broad class of opponents.

It has already been shown that play against adaptive opponents can be modeled as an "Adversary Induced Markov Decision Process" (AIM) [1] where the unknown opponent strategy determines the transition and the reward function. In this research we introduce a novel model-based reinforcement learning algorithm designed for an AIM setting. Our algorithm learns the optimal adversary model (given that certain conditions of the underlying AIM hold) and then exploits it by efficiently addressing the following inherent four subproblems:

1. Measuring the goodness (score) of a particular model;
2. Choosing when to exploit the current best model and when to explore alternate models;
3. Selecting the optimal action sequence when exploiting;
4. Planning action sequences when exploring so as to learn better models quickly.

Though our approach addresses each of these subproblems in a unique way, its main strength lies in the solution to the last subproblem of planning exploration. To this end, it uses confidence measures over different models to drive exploration towards states that lead to the most information about promising models. Our approach starts exploiting sub-optimal models early on so as to generate high rewards, but also keeps efficiently exploring in the long-term hunt for the optimal model.

3. PROBLEM DEFINITION

We assume that the opponent, o, acts according to a finite state machine having a memory size of K. K is the memory size of o if the past K-step joint action sequence completely determines the next stochastic action profile of o. The true strategy of o, denoted as π_o, is then a mapping $(A_i \times A_o)^K \mapsto \Delta A_o$. The key insight enabling this research is that the dynamics of playing against a memory-bounded o can be modeled as a Markov Decision Process whose transition probabilities and reward functions are determined by π_o.

Definition An Adversary Induced MDP (AIM) [1] is defined as follows,

[1]Consistent with the literature [5], we call memory-bounded opponents as adaptive opponents
[2]Although we refer to other agents as opponents, we mean any agent (cooperative, adversarial, or neither)

Cite as: Online Model Learning in Adversarial Markov Decision Processes (Extended Abstract), Doran Chakraborty, Peter Stone, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

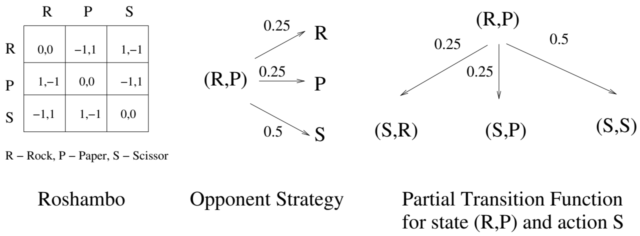

Figure 1: Example of AIM

State : A state is a joint action sequence of size K and the state space is the set of all possible joint action sequences of size K.
Actions : The action space is A_i.
Transition function : The probability of going from state s_1 to state s_2 upon taking action a_i is as follows : 1) the transition probability is the probability of $\pi_o(s_1)$ playing a_o, where (a_i, a_o) is the last pair of joint action in state s_2. 2) For all s_2's which do not end with i playing a_i as the last action, the transition probability is 0.
Reward : The reward obtained on transitioning to state s_2 by taking action a_i in state s_1 is $M_i(a_i, a_o)$, where (a_i, a_o) is the last pair of joint action in state s_2.

The concept of an AIM can be understood intuitively via an example using the game of Roshambo. and illustrated in Figure 1. Assume that o has $K = 1$, meaning that it acts entirely based on the immediate previous joint action. For a history of (R, P), o plays actions R, P and S with probability $0.25, 0.25$ and 0.5 respectively. When i chooses to take action S in state (R, P), the probabilities of transitioning to states $(S, R), (S, P)$ and (S, S) are then $0.25, 0.25$ and 0.5 respectively. For states that have a different action for i, the probability is 0. The reward obtained by i when it transitions to state (S, R) is -1, and so on.

The adversary induces the MDP; hence the name AIM. The optimal policy governing $\mathcal{M}$ is the optimal policy of playing against o. The challenge addressed here is that π_o and K are not known in advance and hence have to be learned in online play. So our goal, is to develop a learning algorithm that will eventually solve for the true K, consequently π_o, and then exploit the opponent optimally.

Finally, it is important to note that there exist opponents in the literature which do not allow convergence to the optimal behavior once a certain set of moves have been played. For example, the *grim-trigger* opponent in the well-known *Prisoner's Dilemma* (*PD*) game, an opponent with memory size 1, plays *cooperate* at first, but then plays *defect* forever once the other agent has played *defect* once. Thus, there is no way of detecting its strategy without defecting, after which it is impossible to recover to the optimal strategy of mutual cooperation. In our analysis, we constrain the class of adaptive opponents to include only those which do not negate the possibility of convergence to optimal exploitation, given any arbitrary initial sequence of exploratory moves.

4. ALGORITHM OVERVIEW

Our objective is to exploit good sub-optimal models early on so as to generate higher rewards and also to act optimally in the limit, against adaptive opponents. From a high-level perspective, our approach is to perform model-based reinforcement learning [7] in an AIM setting. The steps employed by our approach are as follows:

- Maintain a set of models for each possible memory size. A model $\hat{\pi}_k$ is a possible model for π_o, assuming that the opponent has a memory size k. $\hat{\pi}_k$ captures the historical empirical distribution of opponent's play for every possible joint action sequence of size k;
- At each step, the models are updated based on the opponent's behavior from the past step. Based on a score metric, choose the best model ($\hat{\pi}_{best}$);
- Decide to explore or exploit based on how well $\hat{\pi}_{best}$ has predicted the opponent's moves in the past. If $\hat{\pi}_{best}$ is a good predictor of the opponent's moves in the past, then exploit with a high probability assuming $\hat{\pi}_{best}$ to be the true strategy of the opponent, and vice versa. When exploring, take steps that will lead to more information about the promising models (models with higher scores).

We ran experiments against a large set of adaptive opponents in the settings of Roshambo and Kuhn Poker [2]. Our approach successfully bettered a large number of benchmark algorithms.

5. CONCLUSION AND FUTURE WORK

The main contribution of this abstract is highlighting a novel mechanism of model learning and exploitation in modeling adaptive opponents in repeated games. Our approach efficiently explores to learn more about promising models, with an eye towards accruing high rewards on the fly. Fully investigating the empirical success of our approach as a general model based RL algorithm is an important direction for future work.

6. ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (CNS-0615104 and IIS-0917122), ONR (N00014-09-1-0658), DARPA (FA8650-08-C-7812), and the Federal Highway Administration (DTFH61-07-H-00030).

7. REFERENCES

[1] D. Chakraborty and P. Stone. Online multiagent learning against memory bounded adversaries. In *ECML*, pages 211–226, Antwerp,Belgium, 2008.
[2] H. W. Kuhn. A simplified two-person poker. *Contributions to the Theory of Games* ', pages 97–103, 1950.
[3] M. L. Littman and P. Stone. A polynomial-time nash equilibrium algorithm for repeated games. *Decis. Support Syst.*, 39(1):55–66, 2005.
[4] M. J. Osborne and A. Rubinstein. *A Course in Game Theory.* The MIT Press., Massachusetts,USA, 1994.
[5] R. Powers and Y. Shoham. Learning against opponents with bounded memory. In *IJCAI*, pages 817–822, 2005.
[6] A. Sela and D. K. Herreiner. Fictitious play in coordination games. Discussion paper serie b, University of Bonn, Germany, 1997.
[7] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.