A Bayesian Approach for Learning and Tracking Switching, Non-Stationary Opponents

(Extended Abstract)

Pablo Hernandez-Leal
Instituto Nacional de Astrofísica, Óptica y Electrónica
Puebla, México
pablohl@ccc.inaoep.mx

Benjamin Rosman
Council for Scientific and Industrial Research, and the University of the Witwatersrand, South Africa
brosman@csir.co.za

Matthew E. Taylor
Washington State University, Pullman, Washington, USA
taylorm@eecs.wsu.edu

L. Enrique Sucar
Instituto Nacional de Astrofísica, Óptica y Electrónica
Puebla, México
esucar@inaoep.mx

Enrique Munoz de Cote
Instituto Nacional de Astrofísica, Óptica y Electrónica
Puebla, México
jemc@inaoep.mx

ABSTRACT

In many situations, agents are required to use a set of strategies (behaviors) and switch among them during the course of an interaction. This work focuses on the problem of recognizing the strategy used by an agent within a small number of interactions. We propose using a Bayesian framework to address this problem. In this paper we extend Bayesian Policy Reuse to adversarial settings where opponents switch from one stationary strategy to another. Our extension enables online learning of new models when the learning agent detects that the current policies are not performing optimally. Experiments presented in repeated games show that our approach yields better performance than state-of-the-art approaches in terms of average rewards.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

Keywords

Policy reuse; non-stationary opponents; repeated games

1. INTRODUCTION

A core problem in multiagent systems and human-computer interaction is being able to identify the behaviors of other (target) agents. When such a problem is tackled correctly, it allows an agent to derive an optimal policy against such behavior [1]. Consider, for example, a poker player whose behavior has been identified by the opponent player. Then, the opponent can act optimally against such player, anticipating every move. Another example are service robots. They need to interact with persons that have not one but a set of different behaviors (strategies) depending, for example, on factors such as personality or physical capabilities. Moreover, it is desirable that robots learn new tasks and learn how to optimize those.

The target opponent agents that we focus on use a fix (stationary) strategy for an unknown number of interactions, switching to another fixed strategy and repeating this switching to different strategies. There has been some literature addressing this problem [2, 3, 5], however, no work has focused on the problem of reusing previously seen strategies to boost learning speeds.

The Bayesian policy reuse (BPR) framework [7] has been proposed to determine *quickly* the best policy to select when faced with an unlabeled (but previously seen) task. However, standard BPR assumes knowledge of all possible tasks and optimal policies from the start. Our contribution, BPR+, is an extension to BPR for adversarial settings against non-stationary opponents. In particular, we relax the assumption of knowing all opponent strategies *a priori* by providing an online learning approach for incorporating models of new strategies when current policies perform suboptimally.

2. BPR+

Our proposal BPR+ handles non-stationary opponents and learns new models in an online manner. BPR was presented in a single agent environment facing different tasks (represented by MDPs). BPR+ extends to a multiagent setting. Now the tasks correspond to opponent strategies and the policies correspond to optimal policies against those stationary strategies. To cope with these new type of environment, some considerations need to be taken.

(i) An exploration for switch detection is added: drift exploration. Against non-stationary opponents, exploration cannot be terminated. Strategy switches can be difficult to

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

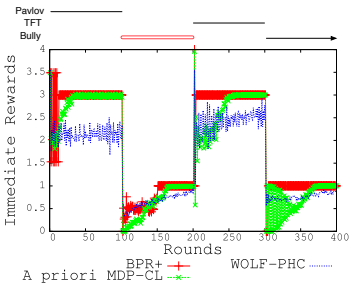

Figure 1: BPR+ and other approaches against a non-stationary opponent (thick line on top of the figure means a unknown strategy to BPR+).

Table 1: Average rewards with std. dev. (±) of the learning agents (average of 100 trials). The opponent changes strategies randomly. Agents can have initial information about the opponent strategies (*Known* column) or start without any information (*Unknown* column), ∗,† represent statistical significance with *a priori* MDP-CL and WOLF-PHC respectively.

	Opponent strategies	
Learning agent	*Known*	*Unknown*
Omniscient	2.53 ± 0.00	2.32 ± 0.00
BPR+	**2.24 ± 0.30** ∗†	**2.17 ± 0.27** ∗†
A priori MDP-CL	2.18 ± 0.31	2.11 ± 0.24
WOLF-PHC	1.71 ± 0.27	1.77 ± 0.19

detect, particularly if the strategies in question are very similar. Such similarities can produce a "shadowing" effect [4] in the agent's perception — an agent's optimal policy $\pi^\star$ will produce an ergodic set of states against some opponent strategy, but if the opponent's switching strategy induces a similar MDP where the policy $\pi^\star$ produces the same ergodic set, the agent will not detect something has changed (unless some exploration occurs).

(ii) A method for detecting when an unknown opponent strategy appears is needed. This scenario is identified through receiving a sequence of rewards which are unlikely given the known opponent models. Two parameters are needed, ρ measures how different the probabilities need to be (compared to the known opponent strategies) and n controls the number of rounds needed before consider learning a new opponent strategy.

(iii) A method for learning the new opponent strategy (and a new optimal policy). The opponent behavior is modeled trough an MDP, $\mathcal{M}_{new}$ [1]. To learn its parameters an exploratory phase is needed. We assume the opponent will not change of strategies during the learning phase (a number of rounds) after which an MDP representing the opponent strategy is obtained. Since the rewards are deterministic, solving the MDP is done through value iteration, obtaining an optimal policy $\pi^\star_{new}$.

(iv) An algorithm to update the known models to add the the recently learned model, $\mathcal{M}_{new}$, is needed.

3. EXPERIMENTS

Our approach is evaluated in the context of repeated games, in particular using the iterated prisoner's dilemma. Well-known strategies in this domain are Tit-for-Tat (TFT), Pavlov and Bully [6].

We show the behavior of BPR+ against a switching opponent which uses known and unknown strategies. BPR+ only has information about TFT and Pavlov strategies. However, the opponent will also use Bully. Figure 1 depicts rewards for BPR+, *a priori* MDP-CL [5], and WOLF-PHC [2]. For the known opponents (TFT and Pavlov) BPR+ takes about 4 rounds to reach the optimal reward. From round 100 the opponent uses Bully which is unknown to BPR+. Then, BPR+ detects this as a new opponent strategy and starts a learning phase which finishes approximately at round 160. Now, BPR+ computes and uses an optimal policy against Bully and updates its models to cope with future switches (round 300).

We evaluated BPR+ against an opponent that switches randomly among the three mentioned strategies. Table 1, under column *Known*, shows the average rewards with standard deviations for the three learning approaches and the Omniscient player against an opponent that switches strategies randomly every 200 rounds. Both BPR+ and *a priori* MDP-CL know the strategies the opponent may use but not the order or the switching times (the opponent should use all strategies at least once). Lastly, we tested the approaches without any prior information, this is shown under column *Unknown*. Results show that BPR+ obtained statistically significant score improvements over the other approaches.

4. CONCLUSIONS

We propose a Bayesian approach to cope with non-stationary opponents (that change from one stationary strategy to another) in repeated games. Our approach, BPR+, exploits information of how the policies behave against different strategies in order to asses the best policy faster than any algorithm in literature. We also provided BPR+ with an online learning algorithm for adding models to its library.

Acknowledgments

This research has taken place in part at the Intelligent Robot Learning (IRL) Lab, Washington State University. IRL research is supported in part by grants AFRL FA8750-14-1-0069, AFRL FA8750-14-1-0070, NSF IIS-1149917, NSF IIS-1319412, USDA 2014-67021-22174, and a Google Research Award.

REFERENCES

[1] B. Banerjee and J. Peng. Efficient learning of multi-step best response. In *Proceedings of the 4th AAMAS*, pages 60–66, Utretch, Netherlands, 2005. ACM.

[2] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[3] B. C. Da Silva, E. W. Basso, A. L. Bazzan, and P. M. Engel. Dealing with non-stationary environments using context detection. In *Proceedings of the 23rd ICML*, pages 217–224, Pittsburgh, Pennsylvania, 2006.

[4] N. Fulda and D. Ventura. Predicting and Preventing Coordination Problems in Cooperative Q-learning Systems. In *Proceedings of the 20th IJCAI*, pages 780–785, Hyderabad, India, 2007.

[5] P. Hernandez-Leal, E. Munoz de Cote, and L. E. Sucar. Using a priori information for fast learning against non-stationary opponents. In *Advances in Artificial Intelligence – IBERAMIA*, pages 536–547, Santiago de Chile, 2014.

[6] M. L. Littman and P. Stone. Implicit Negotiation in Repeated Games. *ATAL '01: Revised Papers from the 8th International Workshop on Intelligent Agents VIII*, 2001.

[7] B. Rosman, M. Hawasly, and S. Ramamoorthy. Bayesian Policy Reuse. *Machine Learning*, (in press) 2016.