Reinforcement Learning in Extensive Form Games with Incomplete Information: the Bargaining Case Study

Alessandro Lazaric
Politecnico di Milano, DEI
piazza Leonardo da Vinci 32,
I-20133 Milan, Italy
lazaric@elet.polimi.it

Enrique Munoz de Cote
Politecnico di Milano, DEI
piazza Leonardo da Vinci 32,
I-20133 Milan, Italy
munoz@elet.polimi.it

Nicola Gatti
Politecnico di Milano, DEI
piazza Leonardo da Vinci 32,
I-20133 Milan, Italy
ngatti@elet.polimi.it

ABSTRACT

We consider the problem of playing in repeated extensive form games where agents do not have any prior. In this situation classic game theoretical tools are inapplicable and it is common the resort to learning techniques. In this paper, we present a novel learning principle that aims at avoiding oscillations in the agents' strategies induced by the presence of concurrent learners. We apply our algorithm in bargaining, and we experimentally evaluate it showing that using this principle reinforcement learning algorithms can improve their convergence time.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept Learning; I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

General Terms

Algorithms, Design

Keywords

Multi Agent Learning, Game Theoretic Foundations

1. INTRODUCTION

In this paper we focus on the application of the RL approach to extensive form games with incomplete information. It is worth noting that most work in multi-agent RL is centered on learning in repeated strategic form games, where the agents, at each step, simultaneously choose their actions. The application of RL techniques in extensive form games with incomplete information is still largely unexplored.

Our original contribution is a general learning principle (CoLF: Change or Learn Fast) that, by opportunely modifying the learning rate, aims at reducing the non-stationary effects induced by explorative actions performed by the other learning agents. If all the learning agents adopt the CoLF principle exploration rates may be kept higher, thus significantly reducing convergence times.

We experimentally evaluate CoLF in a specific case study (Rubinstein's alternating-offers bargaining [3] with deadlines) comparing it with respect to classic Q-Learning. We show that when a best response learner (based on CoLF or Q-Learning) is coupled with another learner of its same kind, the two learners are able to converge to the equilibrium strategies, in this case subgame perfect. The experimental results show also that the CoLF principle allows to drastically improve the convergence speed with respect to classic Q-Learning.

AAMAS'07, May 14–18, 2007, Honolulu, Hawai'i, USA.

2. THE COLF LEARNING PRINCIPLE

In stochastic games the performance of the learning process is negatively affected by agents' changes in their strategies: every time an agent changes its policy, all the other RL agents perceive a non-stationarity in the environment dynamics. While in general stochastic games Q-Learning with ϵ-greedy exploration strategy is not guaranteed to converge, in a subclass of extensive form games it is possible to prove that it succeeds in approximating the exact values with an error ξ when the exploration factor ϵ is under a given threshold [1]. Although relevant, this theoretical result does not provide any information about the convergence speed of the algorithms and, as we discuss in the experimental section, could be a severe obstacle to the application of learning algorithms to significant problems. In fact, RL algorithms benefit from a long and exhaustive exploration in order to achieve optimal strategies. At the same time, in a stochastic game, high exploration factors may contribute to the non-stationarity perceived by the agents.

CoLF (Change or Learn Fast) principle is inspired by the work of Bowling and Veloso [4], where a variable learning rate is considered. According to the CoLF principle, the learning rate of the algorithm is set as follows: if the achieved outcome is suddenly changing, then agents learn slowly, otherwise agents learn quickly. This principle gives less importance to "unexpected" outcomes (i.e. payoffs that are quite different from those achieved recently in the same state), probably generated by non-stationary causes like exploration activity or normal learning dynamics, while allowing to speed up learning when the agents are playing near-stationary strategies. As a result, a learning agent does not immediately modify her strategy because of the unexpected outcome, so that she can verify whether it is caused either by exploration or by an actual change in the other agents'

978-81-904262-7-5 (RPS)

Algorithm 1 CoLF – Change or Learn Fast

Let $\alpha_{NS} > \alpha_S$, and λ be learning rates, $k > 1$
$S(s, a_i) \leftarrow 0, \forall \mathbf{s} \in \mathcal{S}, a_i \in \mathcal{A}_i$
$Q(s, a_i) \leftarrow \frac{r_{max}}{1-\gamma}, \forall s \in \mathcal{S}, a_i \in \mathcal{A}_i$
for all steps **do**
  choose action a_i^t according to exploration strategy
  execute a_i^t and get the payoff r_i^t
  read the state s^t
  compute $u_i^t = r_i^t + \gamma \arg\max_a Q(s^t, a)$
  $\Delta u_i^t \leftarrow |u_i^t - Q(s^{t-1}, a_i^t)|$
  if $\Delta u_i^t > kS(s^{t-1}, a_i^t)$ **then**
    $\alpha \leftarrow \alpha_{NS}$
  else
    $\alpha \leftarrow \alpha_S$
  end if
  $Q(s^{t-1}, a_i^t) \leftarrow (1-\alpha)Q(s^{t-1}, a_i^t) + \alpha\left(r_i^t + \gamma \arg\max_a Q(s^t, a)\right)$
  $S(s^{t-1}, a_i^t) \leftarrow (1-\lambda)S(s^{t-1}, a_i^t) + \lambda \cdot \Delta u_i^t$
  $t \leftarrow t + 1$
end for

strategies. This principle is very general and may be applied to a variety of different learning algorithms.

Algorithm 1 shows how CoLF principle can be employed in Q-Learning. For each pair $\langle s, a\rangle$, where s is a state and a is an action, besides the Q-value, the algorithm needs to store and update also the S-values. The S-values are exponential averages with weight factor λ of the absolute differences between the current outcome u and the corresponding Q-value. The algorithm exploits two learning rates α_{NS} and α_S, with $\alpha_{NS} > \alpha_S$, and the choice of the learning rate to use to update the Q-value associated to $\langle s^t, a_i^t\rangle$ depends on whether the absolute difference between the current outcome and the Q-value is greater than k times the S-value.

3. BARGAINING MODEL

Alternating-offers bargaining with deadlines is essentially a finite horizon extensive form game, where a buyer $\mathbf{b}$ and a seller $\mathbf{s}$ try to agree on the value of a parameter. A player function $\iota : \mathbb{N} \rightarrow \{\mathbf{b}, \mathbf{s}\}$, such that $\iota(t) \neq \iota(t+1)$, gives the agent that act at time t. The allowed actions of agent $\iota(t)$ at time $t > 0$ are *(1) offer*$(\overline{x})$, where $\overline{x} \in \mathbb{R}$, *(2) exit*, and *(3) accept*; whereas at $t = 0$ agents can make only *(1)* and *(2)*. If at t agent $\iota(t)$ makes *accept* the game stops and the *outcome* is $(\overline{x}, t)$, where $\overline{x}$ is the value offered by agent $\iota(t-1)$ at time $t-1$. If at time t agent $\iota(t)$ makes *exit* the game stops and the outcome is *NoAgreement*. Otherwise the bargaining continues to the next time point.

The gain of an agent i in reaching an agreement (x, t) is given by a utility function $U_i : (\mathbb{R}\times\mathbb{N}) \cup \{NoAgreement\} \rightarrow \mathbb{R}$ that depends on three parameters of agent i: the *reservation price* $RP_i \in \mathbb{R}^+$,the *temporal discount factor* $\delta_i \in (0, 1]$, and the *deadline* $T_i \in \mathbb{N}$, $T_i > 0$. If the outcome is an agreement (x, t), then $U_\mathbf{b}$ and $U_\mathbf{s}$ are respectively:

$$U_\mathbf{b}(x,t) = \begin{cases} (RP_\mathbf{b} - x)\cdot\delta_\mathbf{b}^t & \text{if } t \leq T_\mathbf{b} \\ -1 & \text{otherwise} \end{cases}, U_\mathbf{s}(x,t) = \begin{cases} (x - RP_\mathbf{s})\cdot\delta_\mathbf{s}^t & \text{if } t \leq T_\mathbf{s} \\ -1 & \text{otherwise} \end{cases}$$

if the outcome is *NoAgreement*, then $U_b(NoAgreement) = U_s(NoAgreement) = 0$.

When an agent can make any offer $x \in \mathbb{R}$ and information is complete, alternating-offers has a unique subgame perfect equilibrium (see [2]); we call it SPE1. When offer x is discretized and information is complete, in addition to SPE1, a new equilibrium raises; we call it SPE2. Essentially, SPE2 is very close to SPE1, the difference lays in the offer that agent $\iota(\overline{T}-1)$ would make at $\overline{T}-1$ ($RP_{\iota(\overline{T})}$ in SPE1, $RP_{\iota(\overline{T})} \pm \Delta$ in SPE2 where Δ is the discretization step).

In the translation of alternating-offers to stochastic games, the reward function is non-zero for both the agents only for termination states, where it is the utility of the accepted offer without temporal discount. Exactly, the buyer's reward function is (the seller's one is analogue):

$$\mathcal{R}_\mathbf{b}(x,t) = \begin{cases} (RP_\mathbf{b} - x) & \text{if } t \leq T_\mathbf{b} \\ -1 & \text{otherwise} \end{cases}.$$

This definition is consistent with the individual goal of each learning agent that tries to maximize the sum of its expected reward. In fact, set the value of γ equal to δ, the expected payoff becomes $E[\sum_{j=0}^{t} \gamma^j \mathcal{R}_i(s_j)] = \gamma^t \mathcal{R}_i(s_t)$.

4. EXPERIMENTAL RESULTS

Our experimental activity is directed to verify whether the CoLF learning principle actually allows the reduction of the number of repeated interactions required to converge to an optimal strategy in presence of concurrent learners.

We carry out our experimental activity on the following case study: $\delta_\mathbf{s} = 0.7$, $\delta_\mathbf{b} = 0.8$, $T_\mathbf{s} = T_\mathbf{b} = 9$. Moreover, since in the bargaining problem each player has continuous actions (i.e., the offer $x \in \mathbb{R}$), in order to apply traditional RL algorithms based on a tabular representation of the action-value function, we discretized the offer interval (that we assume, without loss of generality, to be $[0, 1]$) into a finite set of available actions with a discretization step $\Delta = 0.05$.

In order to transfer the theoretical results described in [1] to the bargaining domain we must exclude terminal states that give to an agent the same payoff. Specifically, the two reservation prices must be excluded from the set of available offers (i.e., 0 and 1), and we substitute them with two offers that follows inside the offer interval and are close to the reservation prices (in the experiments we use 0.001 and 0.999). Under these settings the game has a unique SPE, i.e. SPE2, that can be achieved by two Q-learners providing that their exploration probabilities are small enough. Given the values chosen for the bargaining problem, we fixed the exploration probability for Q-Learning to 0.025 (we refer to this version as *Q-Lv1*). For comparison we made experiments with a second version of Q-Learning (*Q-Lv2*) characterized by an optimistic initialization of the Q-values and by a larger exploration probability, so that convergence to the SPE cannot be guaranteed. Finally, we apply the CoLF learning principle to Q-Learning under the same exploration settings of *Q-Lv2* to directly compare their performance.

The experiments can be grouped into two different categories: in self-play and in which the agents are allowed to offer also their reservation prices, thus rising new equilibrium solutions.

4.1 Learning in Self-Play

We compare the learning performance of Q-Learning and CoLF when both the buyer and the seller are learning agents with the same algorithm. In this setting, convergence to the subgame perfect equilibrium means that none of the agents succeeds in exploiting the opponent and, as a consequence, they both learn to play an equilibrium strategy.

The graph in Figure 1 shows the percentage of successful negotiations along the learning process, while the graphs in Figure 2 show the subgame optimality for the seller (at left)

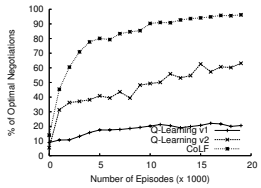

Figure 1: Performance of learning algorithms in self-play in repeated negotiations

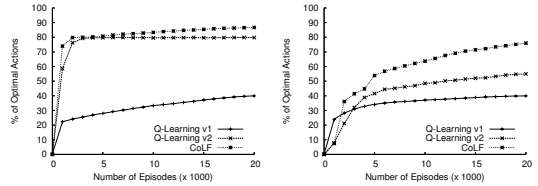

Figure 2: Performance of the seller (left) and the buyer (right) in self-play in random negotiations

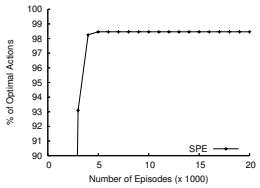

Figure 3: Optimality according to the SPE1 with indifferent actions

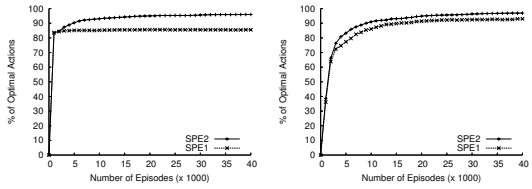

Figure 4: Comparison between optimality according to the SPE1 and SPE2 with indifferent actions for the seller (left) and the buyer (right)

and for the buyer (at right) respectively. The only algorithm that has theoretical convergence guarantees is *QLv1*. Although *QLv2* and CoLF do not have any theoretical guarantee, they show a fast convergence to the SPE solution. In particular, CoLF shows a significant improvement with respect to the other two learning algorithms both in terms of learning speed and optimality of the solution. The experiment with randomly restarted negotiations highlights the intrinsic asymmetry between the seller and the buyer that plays at the deadline and is "forced" to accept profitless offers. This makes more difficult the learning process for the buyer whose performance is quite far from the optimality.

4.2 Learning with Equivalent Strategies

We consider here the situation in which agents can offer opponents' reservation prices, and consequently there are two SPEs. At first we coupled a game theoretical agent that plays the SPE1 strategy with a learning agent using CoLF. In this context the learning agent is expected to learn the optimal strategy. The game theoretical agent starts the negotiation as the seller, while the learner is the buyer and plays at the deadline. The experiment has been carried out with random restart, and it aims at evaluating the percentage of states in which the learning agent has learned the SPE1 strategy. As shown by Figure 3 the optimal strategy learned does not perfectly overlap with the SPE1 strategy. This result is not surprising; in fact, when the buyer plays at the deadline and the other agent has offered buyer's reservation price, independently from the action taken by the buyer its payoff will be 0. While in a game with complete information, by knowing the utility of the other agent, the equilibrium strategy is that the buyer at the deadline should accept her own reservation price, in a game with incomplete information the buyer faces several equivalent strategies.

This phenomenon is even more evident when we consider the negotiation between two learners. Figure 4 displays the percentages of optimality of the learned policies (the seller on the left and the buyer on the right) evaluated with respect to the two SPEs strategies of the game. These graphs show that the two learners in presence of these two SPEs converge to SPE2. This result can be explained as follows: since the agent that plays at the deadline (say the buyer), when receiving an offer equal to her reservation price, will randomize among its actions, the other agent (say the seller) will perceive a low outcome associated to that offer. For this reason the seller will switch to the first available offer below the reservation price of the buyer, and this offer will be unconditionally accepted by the seller.

5. CONCLUSIONS

In this paper we have proposed the CoLF principle applied to Q-Learning in order to limit the non-stationary effects introduced by concurrent learning agents and to improve the convergence speed and the performance after few repeated episodes. The algorithm has been compared to two parameterizations of Q-Learning in the relevant and challenging settings of alternating-offers bargaining games with deadlines. Finally, we analyze the performance of learning algorithms when more than one subgame perfect equilibrium is present.

6. ADDITIONAL AUTHORS

Additional authors: Marcello Restelli,
email: `restelli@elet.polimi.it`.

7. REFERENCES

[1] P. Huang and K. Sycara. Multi-agent learning in extensive games with complete information. In *Proceedings of AAMAS*, pages 701–708, Melbourne, Australia, July 14-18 2003. ACM Press.

[2] S. Napel. *Bilateral Bargaining: Theory and Applications*. Springer-Verlag, Berlin, Germany, 2002.

[3] A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.

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