Cooperation In Stochastic Games Through Communication

Raghav Aras
Loria \ INRIA-Lorraine
B.P. 239, Campus Scientifique
Vandœuvre-lès-Nancy, France
Cedex 54506, France
aras@loria.fr

Alain Dutech
Loria \ INRIA-Lorraine
B.P. 239, Campus Scientifique
Vandœuvre-lès-Nancy, France
Cedex 54506, France
dutech@loria.fr

François Charpillet
Loria \ INRIA-Lorraine
B.P. 239, Campus Scientifique
Vandœuvre-lès-Nancy, France
Cedex 54506, France
charpillet@loria.fr

1. INTRODUCTION

The application of reinforcement learning principles to the search of equilibrium policies in stochastic games (SGs) has met with some success ([3], [4], [2]). The key insight of this approach is that each agent can learn his own β-discounted reward equilibrium policy by keeping track of Q-values of all the agents including himself, and considering the Q-value matrix for each state as his payoff matrix. Each agent sees what actions other agents take, and what payoffs they receive. There is some evidence that in practice, agents that do not observe the actions and payoffs of other agents (hereby denoted as imperfectly observing agents), can still learn adversarial equilibrium (AE) policies in general-sum SGs ([1]) using naive Q-learning. Considering the Prisoners' Dilemma stage game (Table 1) as an abstraction of a SG, this implies that, even by ignoring other agents' play, agents still learn to play DD, which is the adversarial equilibrium joint action. The payoff received in DD can be thought of as each agent's *security level*.

It is of interest to inquire if imperfectly observing agents can improve upon what they learn solipsistically in SGs without global optima. In other words, whether they can get higher payoffs than their security level. We observe that by setting ϵ to any positive real, CC can be made arbitrarily better for both the agents than DD. The distinguishing quality of CC is that it affords the highest payoffs to *both* the agents than any other joint-action. We define joint-actions such as CC, with some abuse, as a best compromise equilibrium or BCE. It is an (unstable) equilibrium, since both agents have an incentive to deviate from it profitably, and both are aware of this fact. Since CC is not a stable equilibrium, the crux of learning is to make agents force one another into playing it rather than DD. Toward this end, we present a reinforcement learning algorithm that incorporates a signaling faculty (endowed to the agents) and an additional payoff interpretion rule. The objective of the learning agent is to learn a BCE policy instead of an AE policy under the β-discounted reward criterion.

Table 1: Prisoners' Dilemma game

	C	D
C	$1 + \epsilon$, $1 + \epsilon$	1 - 2ϵ, $1 + 2$ ϵ
D	$1 + 2$ ϵ, 1 - 2 ϵ	1 - ϵ, 1 - ϵ

2. FRAMEWORK

A two-agent general-sum stochastic game is represented by a set of payoff matrices $\{U_1, \ldots, U_z\}$, corresponding to a set of states $\{s_1, \ldots, s_z\}$. Each payoff matrix is of size $A^1 \times A^2$, where A^k is agent k's action set. Each entry of U_s consists of: $r^k(i, j)$, the payoff agent k will get when the two agents play actions i, j in s, and $p(s, i, j)$, the state transition probability vector, giving the probability of moving to every state, from 1 to z, on taking actions i, j in s. A policy is a mapping $s \rightarrow i$. The β-discounted reward for agent 1 when the game is in state s is,

$$v^1_\beta(s, \pi, \sigma) = \sum_{n=0}^{\infty} \beta^n r^1_n \qquad (1)$$

where π is agent 1's policy and σ is agent 2's and r^1_n is agent 1's payoff in the state the game finds itself at step n. $\beta \in [0, 1)$. An analogous definition can be given for agent 2. A pair of policies for the two agents, (π, σ), is a Nash equilibrium for the stochastic game, if $\forall s$, $v^1_\beta(s, \pi, \sigma) \geq v^1_\beta(s, x, \sigma)$, and $v^2_\beta(s, \pi, \sigma) \geq v^2_\beta(s, \pi, y)$, where x and y are other policies. In an adversarial equilibrium (AE), not only may an agent not deviate profitably from its equilibrium policy, but additionally the deviation will profit the other agent if he does not deviate. Given an AE (π, σ), we define the pair $(\hat{\pi}, \hat{\sigma})$ as a best compromise equilibrium (BCE) if $\forall$ s:

$$(\hat{\pi}, \hat{\sigma}) = \arg \max_{(x,y) \neq (\pi, \sigma)} \min_{v^k} (v^1_\beta(s, x, y), v^2_\beta(s, x, y)) \qquad (2)$$

As learning conditions, first of all, we assume that agents are imperfectly observing. We assume that the SG has a unique AE and a BCE. Naturally, we assume that it does not have a global optimal. Given the imperfect observability, we require that agents know at least their adversarial policy payoffs as well as their payoffs when others deviate from it; the latter we assume are individually maximum for the agents. Finally, we assume that all payoffs are non-negative reals. The learning task is a non-trivial one since payoffs in the BCE might be lower for an agent than in some other joint-action.

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

3. THE ALGORITHM

The agents are given a certain number of rounds L to learn by playing the SG, and another number of rounds P, to execute fixed policies of their choosing. Each round consists of a fixed number of agent joint-actions; in fact it approximates an "episode". The objective manifests in the form of maximizing average payoff in the P rounds. At the heart of our algorithm lies a signaling faculty that we endow to the agents. Each agent has the option of executing an action with or without sending a message to the other agent. Thus each action has a boolean variable, M, to indicate if the action is with or without a message. Similarly, each state has a boolean variable, N to indicate if a message was received in the state or not. The action and state space of each agent thus doubles. Each agent maintains Q-values for all states and all individual actions, which are initialized to 0. During the L learning rounds, agents update Q-values using naive Q-learning. However, before updating a Q-value, each learning agent k assigns a new value to received payoff r^k using the following rule, which we label BCE Q: (R^k_{ae} is his AE payoff, while R^k_{max}, his payoff when the other agent deviates from the AE):

> If the agent neither sends nor receives a message, his payoff is unchanged if it is less than R^k_{max}, otherwise it is set to 0. If the agent either only sends a message without receiving one or receives one, without sending one, his payoff is $(r^k - R^k_{max})$ if r^k is higher than R^k_{ae}, otherwise r^k is set to -R^k_{max}. If the agent receives as well as sends a message, his payoff is $(R^k_{max} - r^k)$, if it is less than R^k_{max} but higher than R^k_{ae}, otherwise it is set to 0.

The principle of this algorithm is as follows: agents can avoid the AE by negating severely the effect of deviating from it. Then, of the remaining joint-actions, there remains the problem of assymetrical payoffs in which one agent gets less than what he would in the BCE. Communication can potentially give him a much higher payoff, but it also risks giving him a much lower payoff if the other agent does not wish to communicate. Only if the payoffs are symmetrical would both agents want to communicate. In the P rounds, agents execute their optimal policies just as in naive Q-learning.

4. EXPERIMENTAL RESULTS

We conducted experiments on a stochastic game with 5 states, numbered 0 to 4. In each state, each agent can take two actions, 0 or 1. Each of four joint-actions from state 0, leads a specific state (and to no other) from 1 to 4 with probability 1. Any joint action from any other state leads to state 0. Payoffs are (0, 0) for any joint-action in state 0, and they are according to Table 2 in other states (agent 1 is row). Each "action" in this table represents a 2-step agent policy. The exponents indicate state number. For example, the action pair $\{00, 10\}$ implies that starting in state 0, agent 1 took two successive 0 actions, while agent 2 took a 1 followed by a 0. The AE of the SG is thus the policy $\{11, 11\}$ when starting in state 0, while the BCE is $\{00, 00\}$. AE gives a 2-step payoff of (5, 5) while BCE gives (7.5, 7.5). We compared a pair of BCE Q learners with a pair of naive Q and Nash Q learners on this SG. Results of the agents' performance in the P rounds is shown in Table 3. L and P equaled 50,000 each, and there were a 100 runs. Round length was 3.

Table 2: Payoff matrix for the SG

	00	01	10	11
00	$(15,15)^1$	$(0,50)^1$	$(2,2)^2$	$(1,400)^2$
01	$(50,0)^1$	$(1,1)^1$	$(1,1)^2$	$(2,2)^2$
10	$(2,2)^3$	$(1,1)^3$	$(1,1)^4$	$(2,2)^4$
11	$(400,1)^3$	$(2,2)^3$	$(2,2)^4$	$(10,10)^4$

Table 3: Performance during P rounds

	Average Payoff	Variance
(Q, Q)	3.90, 3.90	0.02, 0.02
(Nash Q, Nash Q)	5.23, 6.22	4.64, 4.82
(BCE Q, BCE Q)	7.51, 7.34	0.21, 0.34

5. CONCLUSIONS

In game-theoretic terms, agents can improve upon their equilibrium payoffs by signing a contract of playing chosen joint-actions, if the payoff matrix is known. A contract has a correlational property; it requires the participation of all the agents for it to be self-enforceable as an equilibrium is, and centralized mediation. The idea of contracts extends to that of a correlated equilibrium, and this idea has been utilized in the Correlated Q-learning algorithm [2] for SGs with perfectly observing agents and a central mediator. Learning non-equilibrium policies that are self-enforceable, when agents are imperfectly observing and there is no central mediator, is thus a challenging problem. In this paper we have attempted to address this issue by introducing the notion of payoff evaluation or interpretation and BCE. To relate BCE to typical reinforcement learning problems such as grid-games, considering the example of Chicken, a BCE is the joint-policy where both agents eschew the corridor and go for the barrier. The BCE Q rule for payoff interpretation demands a certain knowledge of the SG which normally agents are expected to learn, and that the SG have certain restrictive properties; this limits the applicability of the algorithm. This should not be surprising for such an algorithm considering that self-enforcement of non-equilibria in SGs can be shown to be impossible in the general case for imperfectly observing agents. Our research is more veered toward distributed reinforcement learning of goal-oriented tasks, and hence the rules proposed are meant to exploit the structure of the SGs that model these tasks.

6. REFERENCES

[1] M. Bowling. Convergence problems of general-sum multiagent reinforcement learning. *Seventeenth International Conference on Machine Learning*, pages 89–94, 2000.

[2] A. Greenwald and K. Hall. Correlated-Q learning. *Twentieth International Conference on Machine Learning*, pages 242–249, 2003.

[3] J. Hu and M. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, pages 1039–1069, 2003.

[4] M. Littman. Friend-or-foe Q-learning in general-sum games. *Eighteenth International Conference on Machine Learning*, pages 322–328, 2001.