Review and reproduction of *Learning dynamics in social dilemmas* Macy, M.W., Flache, A., 2002

Julien Baudru[1], Damien Decleire[1], Hamza El Miri[1] and Anthony Zhou[1]

[1]Université Libre de Bruxelles, Brussels, Belgium

Abstract

The Nash equilibrium is incapable of accurately predicting the outcome of repeated mixed-motive games, nor can it describe how a population of players moves from one equilibrium to the next. These limitations have prompted efforts to explore alternatives to analytical game theory; in this paper, we adopt a learning-theoretic approach based on the work of Michael W. Macy[1] and Andreas Flache[2]. More precisely, computational experiments with adaptive agents identify a fundamental solution to social dilemmas, stochastic collusion, by a random walk from a self-limiting noncooperative equilibrium into a self-reinforcing cooperative one. However, we demonstrate that this method is only feasible within a small range of aspiration levels. Agents are dissatisfied with mutual cooperation above an upper threshold and tend towards a deficient equilibrium under a lower one. Additionally, aspirations that adjust with experience do not produce viable results.

Introduction

Generally, games are defined as social dilemmas when a player (agent) receives a higher payoff by defecting than by cooperating with other players. Thus, in the following pages we will look at how to train agents in order to make them work together, to do so we will study different combinations of parameters as well as the effects that these produce on the learning of agents.

In the case where players play only one game of these games their best choice is in the Nash Equilibrum of the payoff matrix, however in these pages we will deal with social dilemmas with two players in the repeated game setting. Thus, in the case of repeated games, the search for the Nash Equilibrum does not lead to the best reward on average. To do so, the two agents will use the experience they acquired during the previous games to choose the best action and thus maximize their profits, therefore the agents learn from their mistakes and their success, they evolve over time. These mistakes and successes are defined here by rewards or punishments.

Many solutions have been proposed to allow agents to learn as they progress in the game. One of the first solutions we can think of would be to never again reproduce an action that led the agent to a punishment. However, this naive solution does not allow much experimentation and is therefore of little interest. Another more interesting solution often used to allow agents to learn is Q-Learning. The latter, introduced by Watkins and Dayan (1992), allows agents to choose the best action according to the state they are in. The solution on which we will focus was proposed by Bush and Mosteller (1953), the BM (Bush-Mosteller) model. This model was modified by the authors of the article we are reproducing, Macy and Flache (2002), who introduced the concepts of aspiration and habituation.

In the following pages we will use this learning model on three types of social dilemmas: Prisoner's Dilemma, Chicken and Stag Hunt game. For each of these games, we will compare the different results obtained by varying the values of aspiration and habituation, we will also analyze for which values the agents reach (or not) the complete cooperation called the **SRE** (Self-Reinforcing Equilibrium).

Method

As players are only matched into pairs for those games, we can use a matrix to compute the payoff of the players. At every round, the payoff is given by the following matrix:

		Player 2 C	Player 2 D
Player 1	C	(R, R)	(S, T)
	D	(T, S)	(P, P)

Table 1: The payoff matrix for every round

Each player can choose to either cooperate (C) or either defect (D). If both of them choose to cooperate they will

[1]Department of Sociology, Cornell University, Ithaca, NY 14853

[2]Interuniversity Center for Social Science Theory and Methodology, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands

both receive *R* (reward). Oppositely, if they both choose to defect, they will both receive *P* (punishment). Moreover, if one choose to cooperate and the other choose to defect, the cooperative one will receive *S* (sucker) and the defective one will receive *T* (temptation).
Considering all that, we want to maximize the reward of a player in a repetitive game. To do so, we will focus on the learning process of our players as it affects their decision-making when choosing the best action.

Learning

As previously mentioned, we will use the BM model for the learning process of our players. It is a stochastic model that generates positive or negative stimuli depending on the decision-making of our players. This stimulus allows us to simulate the satisfaction or the dissatisfaction of an action taken by the players. And this feeling of satisfaction depends on how far the reward received is from its current aspiration level. The aspiration is what the player expect from the game, the furthest away, the more satisfaction or disappointment is felt. The player can get used to that feeling of satisfaction/dissatisfaction when similar rewards are received and change its strategy by updating its probability of choosing an action.

The stimulus of an action s_a is calculated as follows:

$$s_a = \frac{\pi_a - A}{sup[|T - A|, |R - A|, |P - A|, |S - A|]}, a \in \{C, D\} \tag{1}$$

where π_a is the payoff of the action a, A is the aspiration level, and T, R, P, S are the values of the game as expressed in Table 1. The denominator is the supremum (upper bound) of the difference between the set of different payoff and the aspiration.

The aspiration is updated every time as follow:

$$A_{t+1} = (1 - h)A_t + h\pi_t \tag{2}$$

where h indicates the habituation to stimulus of the player, i.e. the degree to which the aspiration level tends toward the payoffs of the last iteration, and π_t is the payoff received at time t. When $h = 1$, the aspiration floats immediately to the payoff of the last iteration, this means unless receiving exactly the same payoff as before, the player will feel (dis)satisfaction, but this can be very volatile. We will only look at when $h = 0$ and $h = 0.2$ to see what happens without and with habituation. Other values of h could be tested but since it will only increase the speed of the change in aspiration, there shouldn't be any concrete change in the way the player's aspiration change.
With all that, the model can then update the probabilities of an action a as follows:

$$p_{a,t+1} = \begin{cases} p_{a,t} + (1 - p_{a,t}) \cdot l \cdot s_{a,t}, & \text{if } s_{a,t} \geq 0 \\ p_{a,t} + p_{a,t} \cdot l \cdot s_{a,t}, & \text{otherwise} \end{cases}, a \in \{C, D\} \tag{3}$$

where $p_{a,t}$ is the probability of picking the action a at instant t, l is the learning rate with $0 < l < 1$, $s_{a,t}$ is the stimulus. At $t = 0$, $p_a = 0.5 \, \forall a \in \{C, D\}$. The probability of picking the other action will also be updated so that the sum of the two probabilities is always equal to 1.

The equation (3) will favor the repetition of rewarded actions and will try to avoid the punished actions by decreasing the current probability of this action. This is done by increasing/decreasing the current probability by a factor that is the function of the current probability, the non-negative learning-rate, and the positive/negative stimulus received.

Results

Effect of the aspiration

Aspiration is a key parameter for training agents. If the aspiration value is too low, choosing mutual or unilateral defection will be considered as a good outcome even if these outcomes are socially negative. Aspiration will also be a source of motivation for agents. With sufficient aspiration, agents will tend not to get bogged down in a strategy that seems suitable without exploring alternatives.

In some cases, such as in the prisoner's dilemma, the agent will tend to be satisfied with the temptation to cheat. This is why a sufficiently large aspiration value is necessary to allow agents to try different strategies even if the temptation to cheat is high. When the reward obtained is greater than the aspiration, the probability that the behavior will be repeated increases and the probability of seeking alternatives decreases. The reverse is also true if the reward obtained is lower than the aspiration.

Fixed aspirations

In this section, we have tested different values for the aspiration in order to illustrate what we explained in the previous point. The figure 1 represents the cooperation rate between the two agents for each of the games with an aspiration set to 2, this value being the median of the list π.

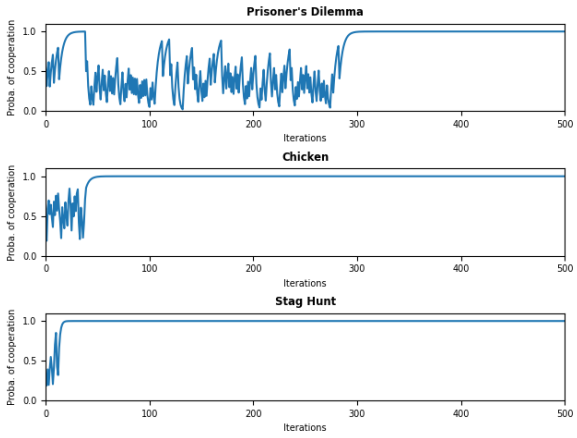

Figure 1: Change in p_c over 500 iterations with initially high aspirations [$\pi = (4, 3, 1, 0), A_0^0 = 2, h = 0, l = 0.5, p_{c,0} = 0.5$]

For this particular value of A_0^0, we notice that each of the three games ends in a state of total cooperation between the agents. This result is particularly interesting for the games where the value of temptation (T) is higher than the value of R. Indeed, it means that the agents manage to overcome the desire to defect from each other, which shows that they have effectively learned to cooperate. Moreover, we note that among the three games, agents tend to reach their stable states of joint cooperation faster in the Stag Hunt game than in the Chicken game, the Prisonner Dilemma game being the one for which agents take the longest to reach the **SRE**.

The figure 2 represents the cooperation rate between the two agents for each of the games with an aspiration set to 0.5, this value being the smaller than all the value in the list π.

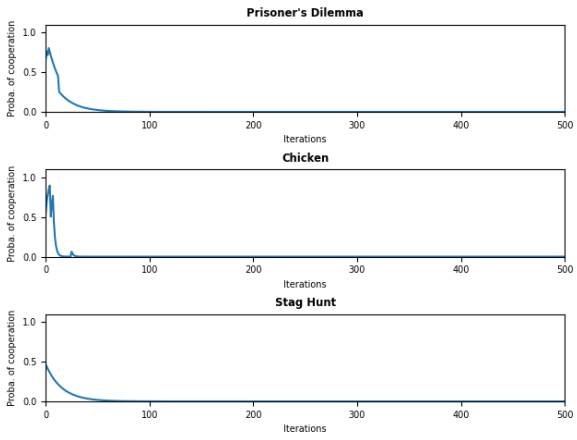

Figure 2: Change in p_c over 500 iterations with initially low aspirations [$\pi = (4, 3, 1, 0), A_0^0 = 0.5, h = 0, l = 0.5, p_{c,0} = 0.5$]

We notice that for a low value of A_0^0, for all three games the agents will tend to end their learning process in a stable situation of mutual defection. This result is due to the fact that agents receive positive stimuli for all values of π being greater than 0.5, i.e. T, R and S for the Chicken game and T, R and P for the Prisoner's Dilemma game and the Stag Hunt game. In this case, the agents are satisfied with the reward obtained even when it is equal to 0 (in the case of the double defection, i.e. P), so as nothing pushes them to seek for a higher reward, once they have reached a mutual defection state they will stay there until the end of their learning.

The figure 3 illustrates the cooperation rate between the two agents for each of the games with an aspiration set to 3.5, this value being the larger than all the value in the list π except 4.

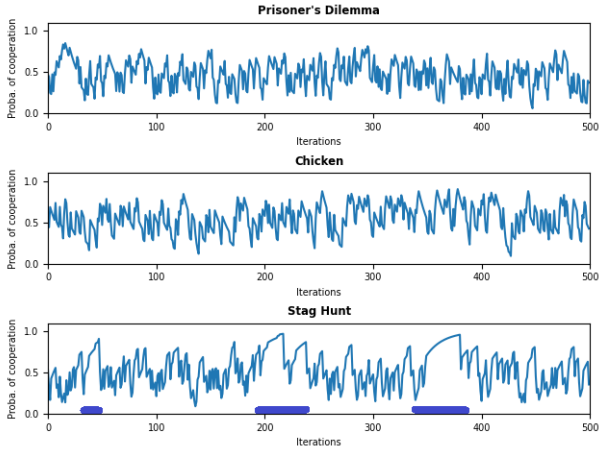

Figure 3: Change in p_c over 500 iterations with initially low aspirations [$\pi = (4, 3, 1, 0), A_0^0 = 3.5, h = 0, l = 0.5, p_{c,0} = 0.5$]

In the case where we fix a high value for A_0^0, we notice that for the three games, the agents will neither reach the **SRE** (i.e. full cooperation) nor reach the full defection, they will rather oscillate between the two states in an almost random way. This behavior seems to be explained by the fact that only one reward value satisfies both agents for each of the games, R for the Stag Hunt game and T for the Chicken game and the Prisoner's Dilemma game. What is surprising here for the Stag Hung game is that despite the fact that it is R the only satisfying reward for the agents, and that this reward leads to mutual collaboration, the agents do not converge to the **SRE** as shown in the original article by Macy and Flache (2002). However, we note that on several occasions, for the intervals delimited by the purple stroke, the agents seem to be progressing towards this state of full cooperation.

Floating aspirations

Based on the previous results, it is safe to assume that there must exist an interval of values for the aspiration that favors mutual cooperation. Assigning a value outside of this range destabilizes the equilibrium and pulls the agents into a deficient one.

Rather than testing every possible aspiration value, one might want to let the agents discover the optimal balance point during the learning process. Surprisingly, this approach does not produce the expected results. This behavior can be explained by how agents adapt to a recurrent stimulus. Agents become dissatisfied with mutual cooperation and numb to social costs in addition to an increased sensitivity to changes in the stimulus. In other words, agents who have become habituated to rewards in a **SRE** will react more aversively to a punishment and vice-versa. As a result, habituation not only reduces the self-reinforcing effect of mutual cooperation but also amplifies the effects of defection.

Fig. 4 illustrates the destabilizing effects of habituation in comparison to Figure [FIGURE 1 NEEDED] with identical conditions except for an increase in habituation ($h = 0.2$). We observe that players can achieve mutual cooperation but cannot maintain it. As they become habituated to the rewards, they become increasingly sensitive to the cost of **S** due to the amplifying effects of a high aspiration. As a result, whenever a chance defection occurs, the agents become considerably less willing to cooperate and are drawn into the **SCE** resetting their habituation to the **SRE** in the process.

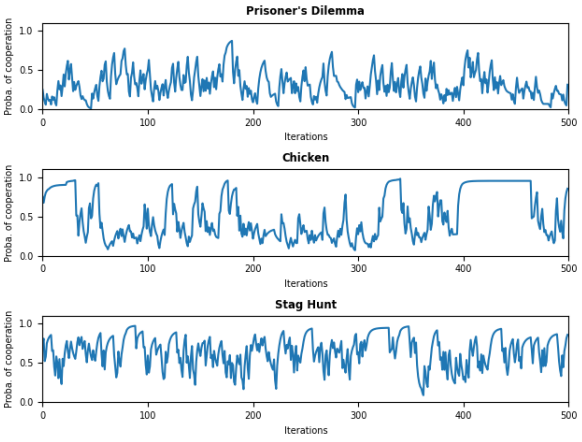

Figure 4: Change in p_c over 500 iterations with initially high aspirations [$\pi = (4, 3, 1, 0), A_0^0 = 2, h = 0.2, l = 0.5, p_{c,0} = 0.5$]

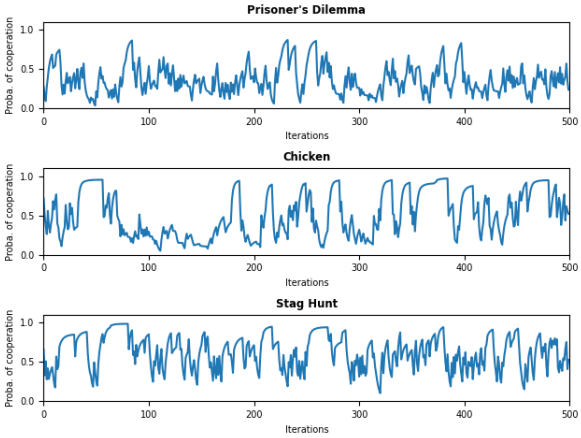

Figure 5: Change in p_c over 500 iterations with initially low aspirations [$\pi = (4, 3, 1, 0), A_0^0 = 0.5, h = 0.2, l = 0.5, p_{c,0} = 0.5$]

Effect of the habituation

The role played by the habituation parameter in the model described in these pages is to prevent agents from falling into vicious circles but also, paradoxically, to prevent them from continuing in virtuous circles. This phenomenon occurs because habituation to a reward will decrease the stimulus associated with it, so if a reward is received by an agent a large number of times, its importance will gradually decrease.

Discussion

Conclusion

References

Bush, R. R. and Mosteller, F. (1953). A Stochastic Model with Applications to Learning. *The Annals of Mathematical Statistics*, 24(4):559 – 585.

Macy, M. W. and Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7229–7236.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. In *Machine Learning*, pages 279–292.