ComPlexUs
ORIGINAL RESEARCH PAPER
ComPlexUs 2003;1:112–116; DOI: 22988476098
Received: November 8, 2001
Accepted after revision: December 3, 2002

Episodic Learning: Towards the Emergences of Partial Cooperation

Aviv Bergman[a] Moshe Tennenholtz [a,b]

[a] Center for Computational Genetics and Biological Modeling, Stanford University, Stanford, Calif., USA, and [b] Faculty of Industrial Engineering and Management Technion, Israel Institute of Technology, Haifa, Israel

Key Words
Evolution of Cooperation · Game Theory · Population Dynamics

Abstract
We introduce and apply a new learning mechanism, episodic learning. This learning paradigm is shown to be tightly linked to a natural form of reinforcement learning. Surprisingly, though the defect action in the 'Prisoners' Dilemma' setting is a dominant strategy (and the only evolutionary stable strategy), when using episodic learning partial cooperation is obtained, which is stable against invasion of individuals with defection as their most frequent action. We further show the emergence of partial cooperation ('altruism'), in the context of the 'Centipede' game. Our results are consistent with the experimental data about partial cooperation obtained by cognitive psychologists.

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

©2003 S. Karger AG, Basel
1424–8492/03/0013–0112
$19.50/0
Accessible online at:
www.karger.com/cpu

Aviv Bergman
Center for Computational Genetics and Biological Modeling
Stanford University
Stanford, CA 94305-5020 (USA)
Tel. +1 650 725 7741, Fax +1 650 725 7745, E-Mail Aviv@stanford.edu

Synopsis

Cooperation among multiple independent agents is a widespread phenomenon of importance in sociology and biology. Nevertheless, its mere existence presents a theoretical problem. In the social sciences, theorists generally view social phenomena as emerging from the rational, self-interested actions of individuals; similarly, biologists explain individual and collective phenomena as the result of an evolutionary process in which individuals struggle selfishly to pass their genes down to further generations. In either case, a form of self-interest is the engine of individual action. Hence, the puzzle: how do self-interested individuals learn to cooperate to their mutual advantage, even in situations in which an individual might benefit by cheating?

Much of the research on this question centers on the so-called Prisoner's Dilemma – a simplified scenario of game theory that elegantly brings the essential issues into focus. In the Prisoner's Dilemma, two criminal conspirators face interrogation by the police, and each can either cooperate, by refusing to talk, or cheat (defect), by blaming his partner for the crime. If one defects while the other stays quiet, the police take the word of the defector: he goes free while the other, taken to be the sole perpetrator, receives a severe sentence. If both defect and blame the other, then both go to prison, but for a lesser period, as neither is seen as solely responsible. Finally, if both cooperate and stay quiet, the police can only convict them of a minor charge, and each receives a short 'token' sentence.

In this situation, the prisoners face a tricky dilemma. Each reasons that he ought to defect, regardless of what the other does, as this leads to a shorter sentence in either case. From this perspective, both prisoners should, rationally speaking, choose to defect. But if they do, they will be worse off than if they had both chosen to cooperate. This is the Prisoner's Dilemma.

We present a model of episodic reinforcement learning that results in the emergence of partial cooperation in a population of individuals engaging in a 'one-shot'[1] Prisoners' Dilemma game. We also introduce a local reinforcement learning rule that results in identical behavior. Similar analysis is carried out for other famous games; in particular, we show the emergence of partial 'altruism' in the Centipede game. These results are in accord with observations made by cognitive psychologists.

We define episodic learning as a learning procedure where individuals' strategies are updated only after *sufficient statistics* regarding the different action payoffs are accumulated. Given a set of actions $a = \{1...K\}$, we define an individuals' strategy as its probability distribution over the set of actions. This distribution determines the probability of choosing $i \in a$ as the individual's action. Sufficient statistics are obtained when the number of interactions N required to obtain an estimate of the expected payoff of each action is such that when two individuals with identical strategies differ in their estimation by ϵ where
$\forall\ \epsilon > 0, \exists\ n(\epsilon)$ s.t. if $N > n(\epsilon)$ the difference between the individuals' expected payoffs is smaller than ϵ.

First we analyze the strategies' dynamics in a Prisoners' Dilemma setup using episodic learning.

The Prisoners' Dilemma game can be defined by the following payoff matrix

O	C	D
C	w_{cc}	w_{cd}
D	w_{dc}	w_{dd}

where 'C' and 'D' represent the cooperate and defect actions, and the matrix entries are the payoffs of a player playing the row action against a player playing the column action, where $w_{dc} > w_{cc} > w_{dd} > w_{cd}$. For the purpose of exposition let $w_{dc} = 6$, $w_{cc} = 2$, $w_{dd} = 1$, and $w_{cd} = 0$. Let p_t^j denote the probability of cooperation by individual j at episode t, and p_t denote the overall probability of cooperation in the population. The expected payoff of the cooperation action of agent j, $E_t^j(c)$, when played against a randomly chosen individual is $2p_t$, while its average payoff of the defect action, $E_t^j(d)$, is $1 + 5p_t$. By the end of an episode t, individual j updates its strategy for the next episode according to the following rule:

$$p_{t+1}^j = \frac{E_t^j(c)}{E_t^j(c) + E_t^j(d)}$$

Assuming all individuals posses the same initial strategy, p_0, and all accumulate sufficient statistics, such that $E_t^j(c) = E_t^i(c)$, and $E_t^j(d) = E_t^i(d)\ \forall\ i, j$, the strategies' dynamics will obey the following recursion equation:

$$p_{t+1} = \frac{w_{cc}p_t + w_{cd}(1 - p_t)}{w_{cc}p_t + w_{cd}(1 - p_t) + w_{dc}p_t + w_{dd}(1 - p_t)} \quad (1)$$

From this recursion equation, we can find the equilibrium strategy. For example, using the above values for the payoff matrix, it can be shown that there exists one stable equilibrium point at $\hat{p} = 1/7$. This shows that under the model assumptions, i.e., using episodic reinforcement learning as a mechanism to update an individual's strategy, even though defection is dominant, the system results in partial cooperation as a stable strategy.

Next, to relax the requirement of identical strategies among all individuals, a simulation of a population where each agent is taken to have a randomly chosen initial strategy was studied. Our results show that the population's average strategy converges to a probability of cooperation level as predicted by our analysis, i.e., 1/7. Due to the finite population size, episode duration, and computer roundoff error, suffi-

How can the prisoners – or any pair of agents faced with a similar problem – overcome the dilemma? As one possibility, numerous studies show that agents can learn to cooperate if they play the game on a repeated basis. In the context of this 'Iterated' Prisoner's Dilemma, if each agent has the ability to learn patterns in his partner's actions, then over time a pair can learn strategies for play that achieve a measure of cooperation. But what if agents lack the ability to detect detailed patterns, or play the game with different partners on every occasion? Can agents still learn to cooperate?

In the present paper, Bergman and Tennenholtz explore this issue in the context of the so-called 'one -shot' Prisoner's Dilemma. In this setting, players do not retain a detailed memory concerning the specific outcomes of previous plays. Rather, each of N players follows a strategy for playing that comes down to just one number: a simple probability for cooperating. One can imagine the agents playing this 'one-shot' game on many occasions, each time facing some new agent within a population, and updating their strategies as play proceeds. The authors show that one method for updating these strategies – what they refer to as 'episodic learning' – leads to the emergence of partial cooperation within the population. This process of episodic learning has some highly unrealistic features that make it implausible as a mechanism for cooperation in the real world. However, Bergman and Tennenholtz also demonstrate that a much more natural procedure of 'reinforcement learning' – essentially a method of learning by 'trial and error' – can lead readily to an identical level of cooperation.

The authors depict the basic terms of the one-shot Prisoner's Dilemma in a 'payoff matrix'. In table form, this shows the payoff w_{xy} that a player receives if he follows strategy x while his opponent follows strategy y. Here x and y can be either 'cooperate' c or 'defect' d, and the ordering of the

[1]Individuals may use single or accumulated payoffs from multiple independent trails resulting from playing multiple one-shot games, each with a different individual from a population of players. This scenario is to be distinguished from standard repeated game where memory of previous game results is accounted for (see below).

cient statistics might not be obtained. To overcome this limitation, we modified the dynamics by ensuring that each action is to be played with a non-zero probability, $0 < p \ll 1/7$. Identical results have been obtained when the population size was reduced to 2.

Finally, we have approached the problem of whether the behavior obtained by episodic learning can be obtained by a local reinforcement learning rule. Here we consider two agents engaging in a multiple one-shot Prisoners' Dilemma. Let $Q_t^i(c)$ and $Q_t^i(d)$ denote the so-called Q functions of agent i at time t, i.e. $Q_t^i(c)$ and $Q_t^i(d)$ are measures of success of cooperation and defection by agent i, respectively. Given that at time t a participant i performs action a and gets a payoff w_t, and its current Q value for that action is $Q_t^i(a)$, the Q value of that action at time $t+1$ is given by $Q_{t+1}^i(a) = \delta w_t + (1-\delta)Q_t^i(a)$, where δ determines the learning rate. The Q value of the other action, b, remains unchanged, $Q_{t+1}^i(b) = Q_t^i(b)$. Here the initial Q values are chosen at random from a uniform distribution over the interval [0, 1]. These Q values in turn determine the individual's strategy as follows

$$p_{t+1}^i = \frac{Q_t^i(c)}{Q_t^i(c)+Q_t^i(d)}$$

The actual way that Q values are updated and used for action selection follows the ideas and is in accord with the motivations discussed in cognitive psychology, and in particular in the literature on cognitive game theory [6]. In that literature, as in our study, we are interested in the equilibrium of one-shot games obtained when learning procedures (for updating behavior in the one-shot game) are adopted. This is not to be confused with the equilibrium analysis of infinitely repeated games, in which multi-stage strategies (such as tit for tat) are allowed, and shown to be in equilibrium.

Note that at stationarity

$$\lim_{t\to\infty} Q_t^i(c) = E(c),$$

where $E(c)$ is the expected payoff of an agent playing cooperation, similarly for the defection action. Using the payoff values as defined previously, simulation results show that the strategy the two agents converge to is identical to the strategy obtained by the previous two models, i.e., $p^i = 1/7\ \forall\ i$. To overcome similar limitations as those described in the previous simulation study, we modified the dynamics by ensuring that each action is to be played with a non-zero probability, $0 < p \ll 1/7$.

A similar analysis was carried out for another famous game, the Centipede game. In this game two players, A and B, alternate in their opportunity to quit and stop the game. The payoff to a player making the decision, say A, to quit at time t is higher than the payoff he gets when player B chooses to do so at time $t+1$. However, if the process is not stopped in either of these periods, the payoff to player A is higher than his payoff in both stages. The process ends after an even amount of time steps, T, or when a player chooses to stop. The simplest variant of the Centipede game can be defined as follows. Player A can either take a payoff of 1 (leaving B with a payoff of 0), or continue to period 2, where player B obtains a payoff of 2 leaving player A with 0 payoff[2]. We assume that when two players are matched, the identity of the player to move first is selected randomly (with probably 0.5 to playing each of the roles). The strategy of a player can therefore be described as a specification of whether he will take the payoff of 1 magnitudes $w_{dc} > w_{cc} > w_{dd} > w_{cd}$ defines a game of the Prisoner's Dilemma type. For simplicity, Bergman and Tennenholtz make a specific choice for these values, and begin their study by carrying out a simple calculation to show how 'episodic learning' within a population of N agents can lead to partial cooperation. Suppose the agents play the game in pairs. A pair is chosen at random, they play, and then another pair is chosen, and so on. At any moment, a player's strategy p_i is simply the probability that he will cooperate the next time he is chosen to play. Players update these strategies as play continues, and 'episodic learning' refers to a specific scheme for doing this – a scheme in which all players update their strategies at the same moment, at the end of various 'episodes'. The idea is that as play goes on, each player builds up an empirical estimate of the overall probability of cooperation within the population p. He does this by calculating the fraction of his opponents who have actually cooperated up to that moment. Each player aims to use this estimate to update his strategy p_i, but Bergman and Tennenholtz suppose that no agent actually makes such an update until all have gathered 'sufficient' statistics and have accurate estimates of the community probability p. By choosing the length of these information-gathering episodes to be long enough, the accuracy of these estimates can be made arbitrarily high.

With this special (and somewhat unnatural) scheme for updating, each player's end-of-episode estimate for p will, to a high accuracy, be the same. Each player will then maximize his expected payoff in a community with this level of cooperation by choosing a new strategy, $p' = E(c)/(E(c) + E(d))$, where $E(c)$ and $E(d)$ are the expected payoffs for cooperating and defecting (easily calculated from the payoff matrix). Because of the episodic nature of the updates – which take place only when all players have gathered sufficient statistics – the new strategies of all

[2]In this study we used the two-stage Centipede version of the game, thus standard equilibrium analysis, backwards induction included, are not discussed. Clearly, moving the turn by player A, where player B gets 2 and player A is left with 0 can only be viewed as 'pure altruism' on the part of player A, therefore no sophisticated backwards induction arguments are under consideration in our treatment.

if he is the one to move first, or not. By similar analysis to the one carried out for the Prisoners' Dilemma, we can show that the only stable equilibrium is $p = 3/4$ (i.e. 3/4 of the population will take the payoff of 1 if selected to move first, and the rest will be 'altruistic'). Numerical simulations confirm the predicted analytical result presented here.

Note that the probability of cooperation in the Prisoners' Dilemma game and the probability of 'altruism' in the Centipede game differ in values from those that would maximize the expected overall population fitness. That is to say, the expected global fitness, though higher than the one obtained when individuals choose to employ a dominant strategy, is lower than the maximum attainable expected global fitness.

In this work, we introduced and applied a new learning mechanism, episodic learning. This learning paradigm has been shown to be tightly linked to a natural form of reinforcement learning. Surprisingly, though the *defect action* in the Prisoners' Dilemma setting and the *quit-first* in the Centipede game are dominant strategies (and the only evolutionary stable strategies, ESS), when using episodic learning new results of partial cooperation ('altruism') are obtained, which are stable against invasion of individuals with defection as their most frequent action.

The equilibrium which has been obtained in our analysis is the result of applying a particular learning rule. This equilibrium is different from the one obtained by ESS analysis [1, 2]. It also differs from the one obtained by a classical population dynamics approach [see 3, 4]. The equilibria obtained here, by using a reinforcement learning update rule, are analogous to one obtained by a population of individuals, where individuals adapt their behavioral trait on the basis of population statistics rather than parental inheritance. Hence, these equilibria can be viewed as behavioral equilibria. As we have shown, partial cooperation and 'altruism' are obtained in a *behavioral equilibrium*. These results are in accord with observations made by cognitive psychologists (see Shafir and Tversky [7] for a discussion of partial cooperation in the Prisoners' Dilemma, as well the experimental study of the Centipede game by McKerlvey and Palfrey [5]). The significance of the concept of behavioral equilibrium for the description of emergent social behavior in population and individuals in other situations is to be further explored.

References

1 Axelrod R, Dion E: The further evolution of cooperation. Science 1988;242:1385–1390.
2 Axelrod R, Hamilton WD: The evolution of cooperation. Science 1981;211:1390–1396.
3 Bergman A, Feldman MW: On the evolution of learning: Representation of a stochastic environment. Theor Popul Biol 1995;48:251–276.
4 Eshel I, Feldman MW, Bergman A: Long-term evolution, short-term evolution and population genetic theory. J Theor Biol 1998;191:391–396.
5 McKelvey R, Palfrey R: An experimental study of the centipede game. Econometrica 1992:60:803–836.
6 Roth AE, Erev I: Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. Games Economic Behav 1995;8(suppl):164–212.
7 Shafir E, Tversky A: Thinking through uncertainty: Nonconsequential reasoning and choice. Cogn Psychol 1992;24:449–474.

agents will be identical. This leads to the authors' equation (1), showing how the overall probability of cooperation within the population changes from one episode to the next. With the specific payoff matrix chosen by the authors, the expected payoffs for cooperating or defecting are $E(c) = 2p$ and $E(d) = 1 + 5p$, in which case equation (1) becomes $p' = 2p/(1 + 7p)$.

Notably, this equation indicates that the population can settle into an equilibrium in which the probability of cooperation no longer evolves from one episode to the next. This takes place when $p' = p$, which is the case when $p = 1/7$. Hence, this calculation demonstrates that partial cooperation can indeed emerge in a population of agents playing the one-shot Prisoner's Dilemma, at least if they proceed through a series of 'episodes' as defined here. The strategy of cooperating on one out of seven occasions at random outperforms the strategy of always defecting, even though agents would be 'rationally' tempted to do so.

This analytical result depends on the players making identical and accurate estimates, within each episode, of the community level of cooperation. But the result stands on firmer ground than it might seem. The authors have shown that an identical level of cooperation also emerges in numerical simulations in which players take random initial strategies and in which the population plays a fixed number of games in each episode before the agents update their strategies. In this case, due to random fluctuations, the players will not all have identical accuracy in their estimates of the community probability for cooperation. Nevertheless, these simulations verify the results of the simpler calculation, showing that the errors in the agents' estimates do not spoil the show; episodic learning leads to partial cooperation.

As an aside, Bergman and Tennenholtz note that a subtle technical issue must be addressed in running simulations of this

kind. In any particular episode, even if all agents have a relatively high probability of cooperating, there is still a slim chance that no agent will actually cooperate. In fact, an outcome of this sort is likely in a simulation run through a large number of episodes. Unfortunately, when this happens, all agents will estimate the population-wide level of cooperation as zero, and will adopt this strategy for their ownthemselves, effectively ending the game, as the probabilities will remain at zero forever more. This paradoxical possibility is linked to the existence of a second equilibrium of equation (1) at $p = 0$, representing no cooperation at all. This equilibrium is unstable, however, as any small deviation away from $p = 0$ will tend to lead, after further episodes, to still larger deviations. Because of this instability, this equilibrium has little relevance to the real issue of the emergence of cooperation. To avoid this mathematical artifact in their simulations, the authors demand that the empirical probability of cooperation is greater than zero in each episode (see their constraint that 'each action is to be played with a non-zero probability' that is greater than zero but much less than 1/7). In this way the simulations steer clear of the unimportant unstable equilibrium, but still accurately explore the evolution toward the stable equilibrium.

It is fair to say that the assumptions that go into the procedure of 'episodic learning' are somewhat unnatural. For instance, it seems rather bizarre to demand that all agents wait until they have accumulated good statistics before updating their strategies. Surely agents may differ, for example, in what they judge to be 'sufficient'. To address this issue, Bergman and Tennenholtz move on to explore another mechanism that they refer to as 'reinforcement learning'. The idea is to see if similar results might be achieved with a simple learning mechanism in which the episode is of length one – in which agents do not wait at all, but update their strategies after every encounter. This mechanism of learning is more akin to the procedure of 'traial and error' by which human beings learn to carry out new tasks, trying out various actions, and gradually building up a progressively better picture of which actions work and which do not.

The authors again consider agents playing the one-shot Prisoner's Dilemma on multiple occasions. In this case, however, agents update their playing strategies after each encounter. Each agent maintains a pair of running 'scores' $Q(c)$ and $Q(d)$ to keep track of how well cooperation and defection has paid off in the past, and uses these to decide how to play in the future. Specifically, at any moment, an agent will choose to cooperate with probability $Q(c)/(Q(c) + Q(d))$. After each encounter, the agent then increases the score of the strategy he happened to play in proportion to the payoff he received. If at time t the payoff is w_t, then the score is updated according to the rule

$$Q_{t+1}(a) = \delta w_t + (1 - \delta)Q_t(a)),$$

or, equivalently,

$$Q_{t+1}(a) - Q_t(a) = \delta(w_t - Q_t - Q_t(a)),$$

where δ is an adjustable constant. As the authors suggest, the constant δ can be interpreted as a learning rate, as the agent increases or decreases his score more rapidly as δ grows. In effect, the scores act as running estimates of the average payoff received for the two possible strategies. Hence, if the most recent score is somewhat higher (or lower) than this estimate, the agent will try to make the estimate more accurate by increasing (or decreasing) its value, thereby bringing the new information to bear.

In simulations based on this simple picture, the authors find again that the agents learn to cooperate partially, achieving a long-term probability of cooperation of 1/7. This demonstrates that the same results following from episodic learning can also emerge if agents follow a much simpler strategy based on traial and error updating or 'reinforcement' of strategies in direct proportion to their empirical performance. What's more, this does not appear to be an isolated success. As Bergman and Tennenholtz show, a similar reinforcement-learning approach also achieves partial cooperation among agents faced with another game-theoretic dilemma known as the Centipede Game.

The importance of these results lies in the demonstration that partial cooperation can emerge even without any detailed memory or pattern recognition on the part of the agents. Although sophisticated behavior of this kind can lie behind the emergence of cooperation, it is not necessary. Indeed, even a very rudimentary behavioral strategy of trial and error reinforcement is adequate to the task. Given that this learning mechanism is thought to describe human learning in a variety of situations, this path to the emergence of cooperation may prove to be highly relevant to cooperative phenomena in the real world.

Mark Buchanan