Proceedings of the 31st Chinese Control Conference
July 25-27, 2012, Hefei, China

What strategies can induce cooperation between heterogeneous players?

Xin Wang, Yifen Mu, Jing Han
Key Laboratory of Systems and Control, Institute of Systems Science
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
E-mail: hanjing@amss.ac.cn

Abstract:
In many social systems, the emergence and maintenance of cooperation among rational agents is a central topic and a fundamental problem. Consider the case of the Prisoner's Dilemma game which is played between heterogeneous players, i.e. a leader and a follower, cooperation can emerge when both players optimize their payoffs provided that the leader's strategy is restricted in a certain set of k-step-memory strategies. However, it is difficult to analyze the condition for cooperation analytically since the number of leader's strategies increases super-exponentially with the memory length k. So in this paper, a series of computer experiments are used to investigate the condition for cooperation by enumerating the leader's strategies with 2-step-memory while the follower utilizes the Q-learning algorithm to update her strategy. The results illustrate that there are a small amount of strategies for the leader to induce mutual cooperation. Most of them can be featured as "reward of mutual cooperation" and "punishment of cheating kindness". This is the first step to characterize the features of strategies which can induce mutual cooperation for the general memory length k.

Key Words: game theory, Prisoner's Dilemma, cooperation, heterogeneous players, Q-learning.

1 Introduction

Cooperation exists widely in the social and economical systems. In many cases, cooperation is beneficial for both players. So how the system reaches cooperation is a fundamental problem. However, cooperation is unlikely to occur in many cases[5]. Thus, the emergence of cooperation attracts much attention from many scholars and most efforts are based on the Prisoner's Dilemma game.

One notable and influential work is accomplished by Axelrod [6, 7, 8]. By designing the computer tournament, Axelrod proposed a new method to study the cooperation in Prisoner's Dilemma game. In the first tournament, the simplest "Tit for Tat" strategy (which just copies the action of its opponent in the last round) performs surprisingly well. [9] found that the tag mechanism can make it easy for populations to reach cooperation. Meanwhile, by introducing spatial structures on the population, such as lattices and scale-free networks, cooperation can be promoted effectively [10]-[18]. On the other hand, cooperation between two players has also been studied in many ways [19] introduces ε-Nash Equilibrium, [20] discusses the 'good' strategy, ([21, 22]) use the finite automaton to play repeated games. The previous work of the authors [2, 4] present new results of cooperation based on the repeated Prisoner's Dilemma game. By introducing heterogeneous players, i.e. the leader and the follower, cooperation will prevail when the leader takes suitable strategies.

However for the general memory length k, it is not easy to derive the condition for cooperation analytically due to the huge strategy space for the leader with the increase of k. But the computer experiment provides a possible way to characterize the features of strategies which can induce mutual cooperation. In this paper, we will enumerate all the 2-step-memory strategies for the leader and let the follower utilize Q-Learning algorithm to play against it. Then the leader's strategy is evaluated according to different criterions, such as the averaged sum of the payoffs of the players and the averaged relative payoff. Through analysis of the experimental data, it can be found that a small amount of leader's strategies can lead the system to cooperation (more than 90% cooperation level). Meanwhile some features are found for those strategies which can induce mutual cooperation.

The remainder of the paper is organized as follows. The problem is formulated in Section 2 and Section 3 illustrates the simulation design. Section 4 presents the simulation results together with analysis of the data. The conclusion of this paper is drawn in Section 5.

2 Problem

Usually, the Prisoner's Dilemma (PD) game is described by the payoff matrix shown as follows

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} (R,R) & (S,T) \\ (T,S) & (P,P) \end{pmatrix} \end{array}$$

where the parameters satisfy

$$\begin{cases} T > R > P > S \\ R > \frac{T+S}{2} \end{cases} \tag{1}$$

In Axelrod's tournament, the parameters are taken as $T = 5, R = 3, P = 1, S = 0$ (Chapter 3, [8]). Obviously, for the PD game, the 'mutual defection' outcome (D, D) is the unique Nash equilibrium while the 'mutual cooperation' outcome (C, C) is better for both players. This is so-called "Dilemma".

In this paper, the model we study is the same as the one in [2], which consists of two heterogeneous players called "leader" and "follower". It is assumed that the game can be

This work is supported by the National Natural Science Foundation of China (No. 60574068, No. 60821091, and No.60804043) and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KJCX3-SYW-S01).

played infinitely, i.e. $t = 0, 1, 2, \ldots$. At stage t, the leader and the follower take their actions $l(t), f(t) \in \{C, D\}$ simultaneously, where action C and D are denoted as 1 and 0 respectively. Given the memory length k, the system state can be defined as a $2k$-bit number with the form $(l(t-k+1), f(t-k+1), \ldots, l(t), f(t))$. For instance when $k = 2$, there are $2^{2k} = 16$ system states, say $0000, 0001, \ldots, 1111$, denoted as $s_1, s_2, \ldots, s_{16}$. Thus the leader's strategy with k-step memory is considered as the assignment of the value 0 or 1 corresponding to each of 2^{2k} system states. This strategy can be denoted as the vector $A = (a_1, ..., a_{2^{2k}}) \in \{0,1\}^{2^{2k}}$, and all k-step memory strategies constitute a set $\mathcal{A}_k = \{\text{all the strategies with k-step memory}\} \triangleq \{A_1, \ldots, A_{2^{(2^{2k})}}\}$. Naturally the set of the leader's and the follower's strategies, $\mathcal{A}_L$ and $\mathcal{A}_F$, is the subset of $\mathcal{A}_k$.

Given an initial state and the players' strategies, the game can be realized. Both players will get their payoff $p_L(t), p_F(t)$ at stage t according to the payoff matrix. To indicate whether the leader (follower) can get more payoffs than the other, define the relative payoff at stage t for both players as

$$\begin{aligned} w_L(t) &= sgn\{p_L(t) - p_F(t)\} \\ w_F(t) &= -w_L(t) \end{aligned} \tag{2}$$

Over the infinite time horizon, define the overall payoff in the average sense:

$$\begin{aligned} P_L &= \overline{\lim_{T\to\infty}} \frac{1}{T}\sum_{t=1}^{T} p_L(t) \\ P_F &= \overline{\lim_{T\to\infty}} \frac{1}{T}\sum_{t=1}^{T} p_F(t) \end{aligned} \tag{3}$$

The averaged relative payoff W_L, W_F can be defined similarly. So the game played between the leader and the follower can be formulated as a problem of ordered optimization as

$$\begin{cases} F(L) = \underset{F\in\mathcal{A}_k}{argmax}\, P_F(L, F) \\ L^* = \underset{L\in\mathcal{A}_L\subseteq\mathcal{A}_k}{argmax}\, P_L(L, F(L)) \end{cases} \tag{4}$$

The set $\mathcal{A}_{cc} = \{A \in \mathcal{A}_L : P_L = R, W_L = 0\}$ is denoted as the one where the leader using the strategy in $\mathcal{A}_{cc}$ will lead the system to cooperation if the follower optimizes her payoff. However, the number of $\mathcal{A}_k$ increases super-exponentially with k, $|\mathcal{A}_k| = 2^{(2^{2k})} = 2^{(4^k)}$. It causes the problem that for general k, the mathematical analysis is difficult because the sets $\mathcal{A}_k$ and $\mathcal{A}_{cc}$ are discrete so that many popular mathematical tools lose their power. On the other hand, the cardinal number increases too fast, which makes it hard to enumerate the set. The computer experiment provides a possible way to characterize the features of strategies which can induce mutual cooperation. In the next section, we will enumerate the leader's strategy set $\mathcal{A}_k$ for $k = 2$ numerically while the follower utilizes the Q-learning algorithm to update her strategy in order to optimize her payoff.

3 Methods

For $k = 2$, as mentioned above, there are 16 different states, denoted as $s_1, ..., s_{16}$. So there are 2^{16} different strategies for the leader and each one can be represented as a vector $A = (a_1, \ldots, a_{16})$ with an index $I(A) = \sum_j 2^{16-j} a_j$. Therefore, all the leader's strategies can be indexed as $A_0, ...A_I, ..., A_{65535}$. During a repeated PD game, the strategy of the leader L is fixed, while the follower uses Q-learning algorithm to update her strategy for the given L . By using this method, we can find the the best strategy for every given L belonging to $\{A_0, \ldots, A_{65535}\}$.

3.1 Preliminaries

Q-Learning, a type of Reinforcement Learning, was first proposed in [23] as an on-line learning technique. It has been used to study the scenarios under which the information of the environment is not available, so Q-Learning learns the mapping from state to action only through feedback (i.e. the reward or the punishment) it receives. The proof of convergence of the algorithm [24] guarantees that Q-Learning can learn the optimal solution to the given problem under some conditions.

[25] applied Q-Learning to study the repeated PD game where both players utilize Q-Learning to learn the best response to the strategy of their opponent. In this homogeneous scenario, they found that cooperation seldom emerges, which is different from our settings and results.

3.2 Computer experiment Design

For the follower, the update rule of the Q-function is illustrated as follows:

$$\begin{aligned} Q_{t+1}(s(t), f(t)) &= (1-\alpha)Q_t(s(t), f(t)) \\ &\quad +\alpha(r(t) + \omega \max_{b\in\{0,1\}} Q_t(s_{t+1}, b)) \end{aligned} \tag{5}$$

where $s(t)$ and $f(t) \in \{0, 1\}$ are the system state and the action of the follower at stage t respectively, and $r(t) = p_F(t) \in \{R, S, T, P\}$ is the immediate payoff for the follower, which is taken as 3, 0, 5, 1 in the simulation. Additionally, α is the learning rate, taken as 0.1 here; ω is the discount factor, taken as 0.999 here. The Q-function value is initialized as 0.

The action rule of the follower is as follows:

$$Pr(f(t) = 1 | s = s(t)) = \frac{1}{1 + \exp\{\frac{Q_t(s(t),0) - Q_t(s(t),1)}{\tau_t}\}} \tag{6}$$

where $Pr(f(t) = 1|s = s(t))$ is the conditional probability that the follower will cooperate at stage t given the state $s(t)$, the parameter τ_t is a decreasing sequence with t which satisfies $\lim_{t\to\infty} \tau_t = 0$. In our simulation, it is taken as $\tau_t = 5 \cdot 0.999^t$.

Given the leader's strategy, the game can be realized according to the follower's action rule. In our simulation, the initial state is generated by a uniform distribution over the set $\{s_1, \ldots, s_{16}\}$, and then the game is played for 10^5 stages between the leader and the follower. At stage t, the payoffs of the leader and the follower are $p_L(t), p_F(t)$, the sum of their payoffs $psum(t)$ and the leader's relative payoff $w_L(t)$ are obtained. When the game stops at $t = 10^5$, the average payoffs of the leader and the follower avp_L, avp_F, the average sum of the payoffs $avpsum$ and the leader's average relative payoff avw_L can be calculated to analyze the condition of cooperation. In addi-

tion, for each leader's strategy $A_I (I = 0, 1, \ldots, 65535)$, 500 initial states are generated independently for the realization of the repeated game. So 500 groups of $avp_L, avp_F, avpsum, avw_L$ are derived for each A_I, denoted as $avp_L(m), avp_F(m), avpsum(m), avw_L(m), m = 1, 2, \ldots, 500$ respectively. We average them and get $Avp_L = \frac{\sum_m avp_L(m)}{500}, Avp_F = \frac{\sum_m avp_F(m)}{500}, Avpsum = \frac{\sum_m avpsum(m)}{500}, Avw_L = \frac{\sum_m avw_L(m)}{500}$.

4 Results

The curve below in Fig. 1 is typical to demonstrate how the follower's payoff varies during the repeated game. It can be found that in the simulation, the learning process becomes stable after 10^4 times.

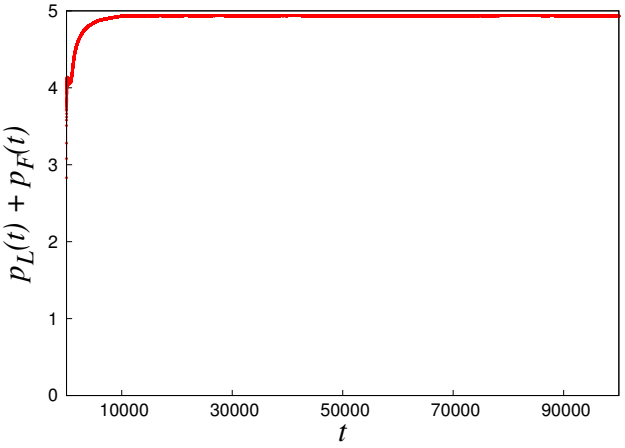

Fig. 1: The payoffs of the follower in the learning process

Note that if and only if at the cooperation state, the sum of the payoffs of players $Avpsum = 2 \times R$, i.e. $2 \times 3 = 6$ in our simulation. At other states, $Avpsum$ is strictly less than 6. Due to the randomness in the Q-learning algorithm, $Avpsum$ is almost always less than 6 even though the system reaches the state of cooperation.

In Fig. 2, different $Avpsum$ values are derived for each leader's strategy $A_I, I = 0, 1, \ldots, 65535$. Fig. 3 further illustrates the number of the leader's strategies under which $Avpsum$ belonging to different intervals. Based on data, there are 713 strategies of the leader which leads to $Avpsum \in (5.6, 5.7]$, denoted as $\#(5.6, 5.7] = 713$. Similarly, we get $\#(5.7, 5.8] = 388, \#(5.8, 5.9] = 708, \#(5.9, 6] = 446$.

Assume during the game, the proportion of the action profile (l, f) at $(1, 1), (1, 0), (0, 1), (0, 0)$ is p_1, p_2, p_3, p_4 respectively, where $p_i \geq 0, \sum_i p_i = 1, i = 1, 2, 3, 4$. The averaged sum of the payoffs of both players is

$$2Rp_1 + (T+S)p_2 + (S+T)p_3 + 2Pp_4 = 6p_1 + 5p_2 + 5p_3 + 2p_4$$

Let it equal to $Avpsum$, we have

$$6p_1 + 5p_2 + 5p_3 + 2(1 - p_1 - p_2 - p_3) = Avpsum$$

then

$$Avpsum - 2 - p_1 = 3(p_1 + p_2 + p_3)$$

where $p_1 + p_2 + p_3 \leq 1$. So we get

$$p_1 \geq Avpsum - 5$$

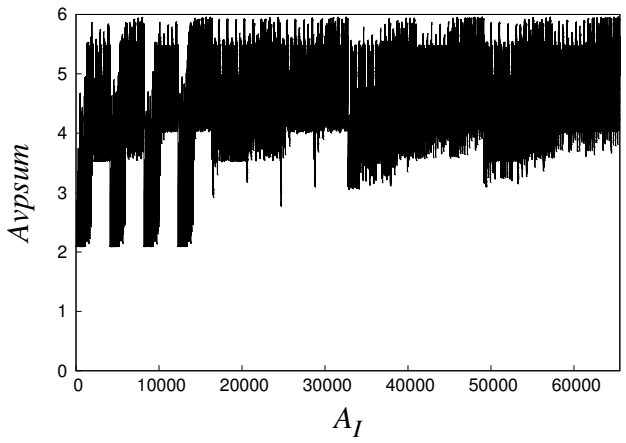

Fig. 2: The $Avpsum$ value for each leader's strategy

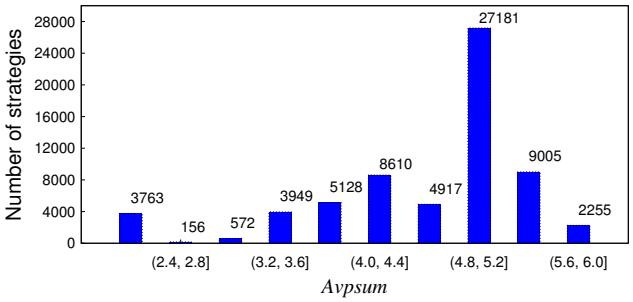

Fig. 3: The distribution of the leader's strategies leading to each payoff intervals

Thus if $Avpsum = 5.9$, then there are more than 90% cooperation level in a whole run. In other words, according to the above data, 446 leader's strategies can lead the system to reach more than 90% cooperation level and $708 + 446 = 1154$ leader's strategies can lead the system to be more than 80% cooperation level.

Furthermore, in those 446 leader's strategies with $Avpsum \geq 5.9$, we need to know what features of strategies can induce mutual cooperation. For simplicity, we average those strategies $((a_1^i, a_2^i, \ldots, a_{16}^i), i = 1, 2, \ldots, 446)$ to get $\frac{\sum_{i=1}^{446} a_k^i}{446}, k = 1, 2, \ldots, 16$. Note that the average strategy may not be a real and feasible one but it can reflect some statistical features of action corresponding to specific system states. In Fig. 4, it can be found that the average action favors cooperation at the state s_4, s_8, s_{12}, s_{16}, which means that if the last action profile is (C, C), the leader should insist on cooperating with the follower in the current move in order to induce mutual cooperation. We call this property *reward of mutual cooperation*. However, at the state s_9, s_{11}, s_{15}, the average action prefers defection and it indicates that if the leader's kindness (i.e. cooperation) is cheated by the follower's following defection, the leader should choose defection to punish the follower. We call this property *punishment of cheating kindness*. At other states, the averaged action has no obvious tendency. These features of strategies are similar to that of "Tit for Tat" strategy, that is, to cooperate when the opponent cooperates in the last move whereas to defect when the opponent's last move is defection. However actually, the player with "Tit for Tat" strategy makes her decision

only based on the opponent's action, regardless of her own action, which is different from both features, "reward of mutual cooperation" and "punishment of cheating kindness" as mentioned above. This difference implies that the purpose of "Tit for Tat" player is not to induce mutual cooperation but to respond to her opponent's action solely. In this case, any unexpected error in action may cause the failure of inducing cooperation when the memory length $k = 2$. In other words, both players are trapped into the "cascade of curse" [4].

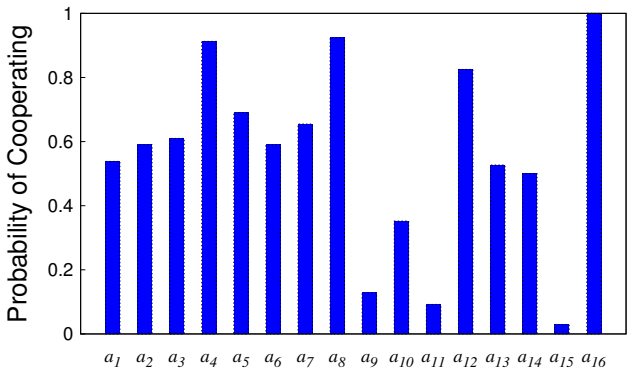

Fig. 4: The averaged action of the leader's strategies inducing $Avpsum \geq 5.9$

This subsection demonstrates that there is a small amount of leader's strategies which can lead the system to cooperation. Specifically, there are almost $\frac{446}{65536} \doteq 0.68\%$ of all 2-step-memory strategies can result to more than 90% cooperation level in the whole run of the repeated PD game. In the meantime, The features of "reward of mutual cooperation" and "punishment of cheating kindness" are found and discussed in the leader's strategy to induce cooperation.

5 Conclusions

This paper makes the first step towards answering the question that what characteristics of strategies can induce mutual cooperation for the general memory length. In the previous work [2], we prove that the system can reach cooperation if one player acts as a leader with some fixed suitable strategy and the other player, acting as a follower, optimizes her averaged payoff in the repeated Prisoner's Dilemma game. However, the number of leader's strategies is huge for the general memory length, which makes it hard to find the appropriate strategies by theoretical analysis.

So we investigate the condition of cooperation in the scenario composed of heterogeneous players, i.e. the leader-follower, by computer experiments. Specifically, all 2-step-memory strategies are considered for the leader while the follower plays against the leader by utilizing the Q-learning algorithm. The results show that there are quite few strategies in the whole set that the leader can choose to induce more than 90% cooperation level. It indicates that it is difficult for the leader to choose suitable strategies to induce mutual cooperation. Meanwhile it can be found that these suitable strategies demonstrate specific properties. Besides, new mathematical tools should be required and used to this problem in our further work.

The idea of characterizing strategies which can induce cooperation in this paper can also be extended to the study of soft control [3]. In [4], the authors investigate a group of shills with the well-designed strategy can lead the system to the state of cooperation. By following the idea here, it can be studied which strategy is optimal for shills to induce cooperation in a group of normal agents which utilize the Q-learning algorithms or other learning methods. In addition, to induce cooperation in the scenario of local interaction, what features of strategies should a shill possess and what is the difference of these strategies in the scenario of local vs. well-mixed interaction? All those problems deserve our further investigation.

REFERENCES

[1] Y. Mu, L. Guo. Optimization and identification in Nonequilibrium dynamica games [C].//*Proceedings of the 48th IEEE Conference on Decision and Control*, Shanghai, China, 2009: 5750-5755.

[2] Y. Mu, L. Guo. How Cooperation Arises From Rational Players? [C].//*Proceedings of the 48th IEEE Conference on Decision and Control*, Atlanta, USA, 2010: 6149-6154.

[3] J. Han, L. Guo, M. Li. Soft control on collective behavior of a group of autonomous agents by a shill agent [J]. *Journal of Systems Science and Complexity*, 2006, 19: 5462.

[4] X. Wang, J. Han, H. W. Han. Special Agents can Promote Cooperation in the Population [J]. *PLoS ONE*, 2011, 6(12): e29182.

[5] G. Hardin. The Tragedy of the Commons [J]. *Science*, 1968, 62: 1243-1248.

[6] R. Axelrod. *The Evolution of Cooperation* [M]. Basic Books, New York, 1984.

[7] R. Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration* [M]. Princeton University Press, Princeton, New Jersey, 1997.

[8] R. Axelrod. The evolution of strategies in the iterated Prisoner's Dilemma, Chapter 3 in: L. Davis, *Genetic algorithms and simulated annealing* [M]. Morgan Kaufman Publishers, Inc.,Los Altos, CA 1987.

[9] R. Riolo. The effects and evolution of tag-mediated selection of partners in populations playing the iterated prisoner's dilemma [C].//Proceedings of 17th International Conference on Genetic Algorithms: 378-385.

[10] M. A. Nowak, R. M. May. Evolutionary games and spatial chaos [J]. *Nature*, 1992, 359: 826-829.

[11] M. A. Nowak, R. M. May. The spatial dilemmas of evolution [J]. *International Journal of Bifurcation Chaos*, 1993, 3: 35-78.

[12] M. A. Nowak, S. Bonhoeffer, R. M. May. Spatial games and the maintenance of cooperation [J]. *Proceedings of National Academy of Sciences, USA*, 1994, 91: 4877-4881.

[13] G. Szabó, C. Töke. Evolutionary prisoner's dilemma game on a square lattice [J]. *P*hysical Review E, 1998, 58: 69-73.

[14] E. Lieberman, C. Hauert, M. Nowak. Evolutionary dynamics on graphs [J]. *Nature*, 2005, 433: 312-316.

[15] H. Ohtsuki, C. Hauert, E. Lieberman, M. A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks [J]. *Nature*, 2006, 441: 502-505.

[16] H. Ohtsuki, M. Nowak. Direct reciprocity on graphs [J]. *Journal of Theoretical Biology*, 2007, 247: 462-470.

[17] G. Szabó, G. Fáth. Evolutionary games on graphs [J]. *Physics Report*, 2007, 446: 97-216.

[18] M. A. Nowak. Five rules for the evolution of cooperation [J]. *Science*, 2006, 314: 1560-1563.

[19] R. Radner. Can bounded rationality resolve the Prisner's Dolemma? [J]. *International Journal of Games Theory*, 1978.

[20] S. Smale. The prisoner's Dilemma and dynamical systems asociated to noncooperative games [J]. *Econometrica*, 1980, 48: 1617-1634.

[21] A. Rubinstein. Finite automata play the repeated Prisoner's Dilemma [J]. *Journal of Economic Theory*, 1986, 39: 83-96.

[22] A. Neyman, D. Okada. Two-person repeated games with finite automata [J]. *International Journal of Games Theory*, 2000, 29: 309-325.

[23] C. J. C. H. Watkins. *Learning from Delayed Rewards* [M]. Ph.D. thesis, Cambridge University.

[24] C. J. C. H. Watkins, P. Dayan. Q-Learning [J]. *Machine Learning*, 1992, 8: 279-292.

[25] T. W. Sandholm, R. H. Crites. Multiagent reinforcement learning in the Iterated Prisoner's Dilemma [J]. *Biosystems*, 1996, 37(1-2): 147-166.