An Adaptive Strategy via Reinforcement Learning for the Prisoner's Dilemma Game

Lei Xue, Changyin Sun, *Member, IEEE*, Donald Wunsch, *Fellow, IEEE*, Yingjiang Zhou, and Fang Yu

***Abstract*—The iterated prisoner's dilemma (IPD) is an ideal model for analyzing interactions between agents in complex networks. It has attracted wide interest in the development of novel strategies since the success of tit-for-tat in Axelrod's tournament. This paper studies a new adaptive strategy of IPD in different complex networks, where agents can learn and adapt their strategies through reinforcement learning method. A temporal difference learning method is applied for designing the adaptive strategy to optimize the decision making process of the agents. Previous studies indicated that mutual cooperation is hard to emerge in the IPD. Therefore, three examples which based on square lattice network and scale-free network are provided to show two features of the adaptive strategy. First, the mutual cooperation can be achieved by the group with adaptive agents under scale-free network, and once evolution has converged mutual cooperation, it is unlikely to shift. Secondly, the adaptive strategy can earn a better payoff compared with other strategies in the square network. The analytical properties are discussed for verifying evolutionary stability of the adaptive strategy.**

***Index Terms*—Complex network, prisoner's dilemma, reinforcement learning, temporal differences learning.**

I. Introduction

GAME theory has become the natural mathematical method to discuss strategic and social interactions, particularly in a competitive environment [1]. The prisoner's dilemma, which exists in many areas, serves as a useful tool for studying human behavior in various social settings and has contributed insights to engineering science, economics, game theory, the analysis of social network structures, and psychology [2]. The iterated prisoner's dilemma (IPD) is a widely used model for analyzing the individual behavior of an agent within a given system. In the IPD, mutual cooperation could provide the highest total income, although selfish individual reasoning often leads to other choices. There are many examples of the Prisoner's dilemma in real life, when people have to choose between being selfish or altruistic. Therefore, the famous computer tournaments for IPD were held by Robert Axelrod [3]. He invited game theorists to submit strategies for playing IPD. The highest payoff was earned by the strategy named "tit-for-tat", it is a strategy which cooperates in the first round and repeats what the opponent has done in the previous move. Some impressive results were collected in [3], the relevant message for people facing a prisoner's dilemma can be summarized as follows:

1) don't be envious;
2) don't be the first to defect;
3) reciprocate both cooperation and defection;
4) don't be too clever.

For one-shot prisoner's dilemma, there is no doubt that betrayal will earn the best payoff for the agent. However, it is seldom that people just face between being selfish or altruistic only once. Thereafter, scholars turned their attention to seek the mutual cooperation during IPD. The spatial evolutionary game demonstrated that local interactions within a spatial structure can maintain cooperative behavior [4]. Reference [4] dealt with the relative merits of various strategies when players who recognized each other meet repeatedly. This spatial version of the prisoner's dilemma can generate chaotically changing spatial patterns. Reference [5] introduced a measure for the cluster shape and demonstrated that the macroscopic patterns can be used to determine the characteristics of the underlying microscopic interactions. Reference [6] studied the competition and strategy selections between a class of generalized strategies.

With the development of evolutionary computation, many scholars have made significant contributions to the research of IPD [7]. Different agents within IPD games may utilize different strategies. Scholars have introduced numerous technologies which can be used to identify or modify IPD strategies. Some strategies are fixed and can be implemented using the finite state machine or Markov decision process [8]. Other strategies are adaptive based on different representation schemes [9], [10]. In order to determine which kind of

Manuscript received March 9, 2016; accepted October 31, 2016. This work was supported by the National Natural Science Foundation (NNSF) of China (61603196, 61503079, 61520106009, 61533008), the Natural Science Foundation of Jiangsu Province of China (BK20150851), China Postdoctoral Science Foundation (2015M581842), Jiangsu Postdoctoral Science Foundation (1601259C), Nanjing University of Posts and Telecommunications Science Foundation (NUPTSF) (NY215011), Priority Academic Program Development of Jiangsu Higher Education Institutions, the open fund of Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education (MCCSE2015B02), and the Research Innovation Program for College Graduates of Jiangsu Province (CXLX1309). Recommended by Associate Editor Qinglai Wei. (*Corresponding author: Changyin Sun.*)

Citation: L. Xue, C. Y. Sun, D. Wunsch, Y. J. Zhou, and F. Yu, "An adaptive strategy via reinforcement learning for the prisoner's dilemma game," *IEEE/CAA J. of Autom. Sinica*, vol. 5, no. 1, pp. 301−310, Jan. 2018.

L. Xue and C. Y. Sun are with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China (e-mail: rayxue@seu.edu.cn; cysun@seu.edu.cn).

D. Wunsch is with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA (e-mail: dwunsch@mst.edu).

Y. J. Zhou is with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: zhouyj@njupt.edu.cn).

F. Yu is with the Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 210096, China (e-mail: fangyu1985@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2017.7510466

strategy has a better performance under specific conditions, some studies have investigated using fingerprint and agent-case embedding to analyze them [11]−[13]. The existing research results also indicated that evolutionary strategies can compete well against some of the fixed strategies [14], [15]. Inspired by these outstanding results, we propose an adaptive strategy based on temporal difference learning method which can consider both the achievement of mutual cooperation and better performance.

The reinforcement learning paradigm can be applied to solve many practical problems [16], [17]. References [18]−[20] applied reinforcement learning method to solve nonlinear system problems. Moreover, learning to predict involves using past experience with an incompletely known system to predict its future behavior, and this type of learning is significant for the IPD. Reinforcement learning is an effective way to teach the agent how to make a decision based on the previous experiences. Reference [20] brought together cooperative control, reinforcement learning, and game theory to present a multi-agent formulation for the online solution of the team games. In [21], the scholars explored interactions in a coevolving population of model-based adaptive agents and fixed non-adaptive agents, and identified that the reinforcement learning method can serve as a useful tool for developing an adaptive strategy. Hingston and Kendall also incorporated reputation as the mechanism for evolving into the existing co-evolutionary learning model for IPD games, where the mechanism for evolving cooperative behaviors is reputation [22]. Reference [23] designed a team of agents which can accomplish consensus over a common value according to cooperative game theory approach. Reference [24], [25] investigated the evolution of cooperation in the prisoner's dilemma when individuals change their strategies subject to performance evaluation of their neighbors over variable time horizons. The main contribution of this paper is shown as follows:

1) A temporal difference (TD) learning method is applied to design an adaptive strategy. The feature of adaptive strategy is that it can balance the short-term rational decision for self-interest against the long-term decision for overall interest. The evolutionary stability of the adaptive strategy is studied.

2) Three kinds of tournaments based on different complex networks are provided. During the tournaments in square lattice network which contains different strategies, the adaptive strategy earns a better payoff. As to the scale-free network constituted by adaptive agent, all the agents will cooperate with each other for long-term reward.

Therefore, the simulation results verify that the adaptive strategy is willing to choose cooperation without losing competitiveness.

This paper is organized as follows. In Section II, IPD is introduced. In Section III, TD(λ) method for prisoner's dilemma is presented. In Section IV, three tournaments are given to verify the feasibility of adaptive strategy. Section V states the conclusions of our study.

II. Iterated Prisoner's Dilemma

Life is filled with paradoxes and dilemmas. A very lifelike paradox is called "prisoner's dilemma", discovered by Melvin Dresher and Merrill Flood [12]. The prisoner's dilemma is a canonical example of two non-zero-sum game. Each agent has two options in each round. One is to cooperate (C), and the other is to defect (D). Based on its choice, the agent will receive a payoff governed by a payoff matrix, as shown in Table I. where R is the payoff when both agents choose cooperation. When only one of the agents defects, it will receive a payoff T, and the opponent will receive a payoff S. If both of the agents decide to defect, each will receive a payoff P. The basic rule of the payoff matrix is $T > R > P > S$ and $2R > T + S$.

The standard IPD is played repeatedly between two players, each with its own strategy. They may have different strategies which can be represented by lookup tables, finite-state machines, and neural networks. The IPD based on the strategies which are represented by finite state machines can be analyzed as a Markov process. This allows an average score to be determined for any pair of strategies using standard techniques in stochastic processes [12]. Some typical types of the fixed strategies are described in Table II [26].

TABLE I
IPD Payoff Matrix

Agent 1\ Agent 2	Cooperate	Defect
Cooperate	$R \setminus R$	$S \setminus T$
Defect	$T \setminus S$	$P \setminus P$

TABLE II
Some Fixed IPD Strategies

Name	Behavior
Always defect	Always plays D.
Always cooperate	Always plays C.
Tit-for-tat	Plays C initially and then repeats the other player's last action.
GRIM	Always defects if the opponent ever defects.
Pavlov	Plays C initially and then cooperates thereafter if its action and its opponent's action matched during the previous round.
Tit-for-2tit	Chooses D only if its opponent has chosen D for its last two moves.
Ripoff	Alternates between C and D until its opponent chooses D for the first time. On the round after this defection, it cooperates and then plays tit-for-tat thereafter.
Psycho	Chooses D initially and then plays the opposite of its opponent's last action.
Random	Simply flips a fair coin to decide how to play. It cannot be represented by the finite state machine.

Definition 1: Denoted by $\mathbf{a}$ as an n-tuple of mixed actions and if $a = (a_1, a_2, \ldots, a_n)$, then the payoff of $\mathbf{a}$ can be written as $U = (U_1, U_2, \ldots, U_n)$. For convenience we introduce the substitution notation $\mathbf{a} = (\boldsymbol{a}_{-i}; a_i)$, where $\boldsymbol{a}_{-i} = (a_1, a_2, \ldots, \boldsymbol{a}_{i-1}, a_{i+1}, \ldots, a_n)$.

Definition 2 (Nash Equilibrium) [3]: A joint action set $(\boldsymbol{a}_{-i}; a_i)$ is a pure Nash Equilibrium point if, for the ith player

$$U_i(\boldsymbol{a}_{-i}; a_i) = \max_{a_i \in a} U_i(\boldsymbol{a}_{-i}; a_i^*). \tag{1}$$

During the IPD, a Nash equilibrium (NE) can be achieved by maximizing the payoffs of agents. If the agents are selfish and want to maximize their own payoffs without considering the interests of the others, the mutual defection can be obtained as a NE. Therefore, one of the two following situations will occur:

1) If the opponent chooses C, the agent may have payoff R or T, $T > R$, then the agent will choose D.

2) If the opponent chooses D, the agent may have payoff S or P, $P > S$, then the agent will choose D.

Therefore, mutual defection leads to the unique NE, which represents a better short-term payoff. However, Axelrod showed that although mutual defection yields a better short-term reward [3], mutual cooperation is a better solution in the long run. Furthermore, when the IPD presents more than two choices, the evolution of defection may be a result of strategies effectively having more opportunities to exploit others when more choices exist [11]. Based on the fixed strategies shown in Table II, each of them has their own personality and can make decisions based on the opponent's move. However, the fixed strategies are of passive type which means they choose the actions based on history without considering the actions of next step. The main contribution of this paper is designing a competitive adaptive strategy which can predict the actions and achieve mutual cooperation without losing competitiveness. In the next section, the reinforcement learning method is introduced to design the adaptive strategy.

III. TD(λ) Method for Prisoner's Dilemma

During the IPD, the known information includes the decisions and payoffs of the previous steps. Therefore, the goal of our study is solving a multi-step prediction problem regarding how to teach the agent predict its own future scores based on the available options. Learning to predict involves using past experience of the unknown system to predict its future behavior. One advantage of prediction learning is that the training examples can be taken directly from the temporal sequence of ordinary sensory input; no special supervisor or teacher is required. TD learning is a prediction method which combines Monte Carlo and dynamic programming ideas. It learns by sampling the environment according to some policy and then approximates its current estimate based on previously learned estimates.

A. Adaptive Design Strategy by TD(λ) Method

In this paper, the adaptive agent should have some features as follows. An adaptive agent should consider the long-term reward based on the situation. In other words, an adaptive agent should learn cooperative coevolution. For instance, if the agent identifies that the opponent will choose "cooperate", the adaptive agent should choose cooperate. If the opponent choose "defect", for protecting its payoff, the adaptive agent should choose "defect". Learn to cooperate with others is a significance character of human beings. Therefore, for an adaptive agent, learning to cooperate is vital.

As a prediction method, when observation is possible, TD learning can be adjusted to better match the observation. TD methods also are more incremental, easier to compute, and tend to make more efficient use of their experience.

Based on the model mentioned in [1], the TD(λ) learner can be expressed as Fig. 1.

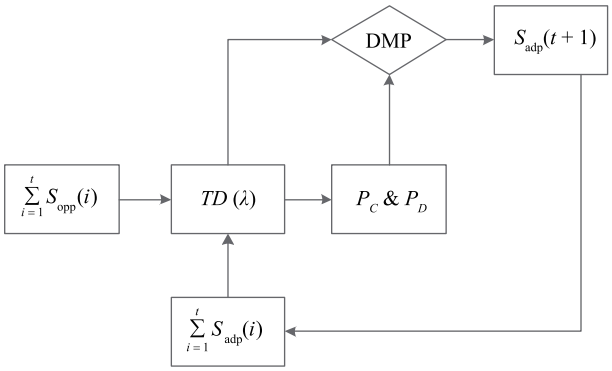


Fig. 1. The decision making process for the adaptive agent.

Epochs $t = 1, 2, \ldots, T$. The scores of the adaptive agents and their opponents are $S_{\text{adp}}(t)$ and $S_{\text{opp}}(t)$, while the forecasting cooperative and defective earnings of the adaptive agent at time t are $S_{\text{cp}}(t)$ and $S_{\text{dp}}(t)$, respectively. The equation for calculating the forecasting earnings of adaptive agent is shown as follows:

$$T_{\text{cp}}(t+1) = \sum_{i=1}^{t} P_C^{t-i}(S_{\text{adp}}(i) - S_{\text{opp}}(i)) \tag{2}$$

$$T_{\text{dp}}(t+1) = \sum_{i=1}^{t} P_D^{t-i}(S_{\text{adp}}(i) - S_{\text{opp}}(i)) \tag{3}$$

$$S_{\text{cp}}(t+1) = T_{\text{cp}}(t+1) + 2P_C(t)R \tag{4}$$

$$S_{\text{dp}}(t+1) = T_{\text{dp}}(t+1) + 2P_D(t)P \tag{5}$$

where $P_C \in (0, 1)$ and $P_D \in (0, 1)$ represent the possibility of cooperation and defection, respectively, for the adaptive agent. R, T, S and P are the payoffs of the IPD, as shown in Table I.

Therefore, the TD(λ) learner can be described as follows:

1) Initialization: The state set and action set of the ith agent are $Z = \{C, D\}$, where C is cooperate, and D is defect. The payoff matrix of the agents is shown as Table I.

2) Calculating $S_{\text{cp}}(t+1)$ and $S_{\text{dp}}(t+1)$ by (2)–(5).

3) The agent makes decision based on the decision making process (DMP).

4) Back to 2) until the iteration stops.

The DMP is shown as follows:

1) If $S_{\text{cp}}(t+1) > S_{\text{dp}}(t+1)$, the adaptive agent will cooperate with the opponent. The possibility of cooperating will increase to $P_C(t+1) = P_C(t) + F(P_C(t+1))$; the

possibility of defecting will decrease to $P_D(t+1) = P_D(t) - F(P_D(t+1))$.

2) If $S_{\rm cp}(t+1) < S_{\rm dp}(t+1)$, the adaptive agent will defect with the opponent. The possibility of defecting will increase to $P_D(t+1) = P_D(t) + F(P_D(t+1))$; the possibility of cooperating will decrease to $P_C(t+1) = P_C(t) - F(P_C(t+1))$.

3) If $S_{\rm cp}(t+1) = S_{\rm dp}(t+1)$, the adaptive agent will cooperate with the opponent. However, the possibility of defecting cannot be reduced. The possibility of cooperating will increase to $P_C(t+1) = P_C(t) + F(P_C(t+1))$. Therefore, the adaptive agent can be encouraged to choose cooperation for the long-term team reward. The function $F(\varepsilon)$ is modified Fermi function shown as follows:

$$F(\varepsilon(t+1)) = \frac{1}{1+e^{[\varepsilon(t)-\varepsilon(t-1)]/k}}. \tag{6}$$

The DMP of adaptive agent clearly indicates that the decisions rely not only on the previous steps, but also on the learning method of the adaptive agent. P_D and P_C decide the tendency of the adaptive agent. Moreover, due to $P_C, P_D \in [0,1]$, P_D^{t-i} and P_C^{t-i} will decrease. In other words, the effect of the last step is greater than that of the other steps, and effect of the first decision within the IPD becomes increasingly weaker as the number of iterations increases. For investigating the convergence in the mentioned manner, the analytical properties are discussed in following subsection.

B. The Analytical Properties of Adaptive Strategy

Different IPD strategies use different history lengths of memory to determine their choices. In a finite length IPD which has L rounds, the largest history length which a strategy can access is L. As to the strategy mentioned in this section, the memory size is as long as the length of finite IPD game. A nontrivial question is whether the adaptive strategy can be the counter strategy (CS) against other strategy? It has been proven that every finite history length is possible to occur in an infinite length IPD which can be expressed as following theorem.

Definition 3 (Counter Strategy) [26]: A strategy S is a counter strategy (CS) against another strategy S_1, if for any strategy S'

$$U(S, S_1) \geq U(S', S_1). \tag{7}$$

Lemma 1 [26]: For any strategy that uses a limited history length, there always exist some strategies with longer memory against which the strategy cannot be a counter strategy.

Theorem 1: The adaptive strategy mentioned in Section II has a higher probability of being a CS against fixed strategy in an infinite length IPD.

Definition 4 [26]: Any strategy that uses a limited history cannot be an evolutionarily stable strategy (ESS) in an infinite length or indefinite length IPD. ESS is a strategy such that, if all the members of a population adopt it, then no mutant strategy can invade the population under the influence of natural selection.

Therefore, the condition for a strategy S to be ESS is that for any S'

$$U(S, S) \geq U(S', S).$$

There is a relationship between CS and ESS. A strategy is ESS if it is the only CS against all IPD strategies [26]. As to the adaptive strategy, the evolutionary stability is discussed as follows.

Theorem 2: As to IPD consisted of fixed strategies and adaptive strategy, the adaptive strategy mentioned in Section II is ESS.

The adaptive strategy designed in this paper is a kind of strategy with the memory size as the length of the IPD game. During the IPD game, the adaptive strategy has the longest memory compared with other strategies. As to the designed game, the adaptive strategy has a higher possibility to be CS of the other fixed strategies. Therefore, the adaptive strategy will earn a better payoff when played against other fixed strategies. During next section, some simulations are given to illustrate the effectiveness of the mentioned manner.

IV. Simulations

In order to measure the effectiveness of agents using the adaptive strategy, three IPD tournaments are simulated in different complex networks. The square lattice network and scale-free network are introduced to be the environments of the simulations. Each of the game theoretical tournaments can be represented as tuple $G(N, A, F)$, where N is the number of the agents, $A = \{C, D\}$ is the action set of the agents, C and D represents the cooperate and defect, and F is the payoff function for each action. The practical payoff matrix of the Prisoner's dilemma is shown in Table III.

TABLE III
IPD Payoff Matrix

Agent 1\ Agent 2	Cooperate	Defect
Cooperate	3 \ 3	0 \ 4
Defect	4 \ 0	1 \ 1

The simulations based on square lattice network and scale-free networks are given in the following sections.

A. The Simulations Based on the Square Lattice Network

Based on the characteristic of the 100×100 square lattice network, there are $N = 1, 2, \ldots, i, \ldots, 10\,000$ agents on the network. In this section, two tournaments are provided to illustrate the effectiveness of the adaptive strategy. During first tournament, the designed adaptive strategy will play against the fixed strategies mentioned in Table II. During the second tournament, two kinds of strategies based on Q-learning [27] and self-adaptive method [28] are given to play against the designed strategy. The structure of the partial square lattice network is described as Fig. 2. As to each agent, it has eight neighbors. The coordinates of agents represent the relative position between them.

During each iteration, the ith agent plays against its neighbors according to the rules of IPD. Based on the feature of the lattice network, the adaptive agents can play against agents with all kinds of strategies. The cooperative rate p and average fitness value are introduced to illustrate the effectiveness of mentioned manner.

$$p_t = \frac{N_c}{N} \tag{8}$$

where p_t represents cooperative rate of the tth iteration; N_c represents the number of cooperating agent; N is the number of all the agents.

$$f_{\text{ave}}(Stra) = \frac{\sum_{t=1}^{T} S_t}{T} \tag{9}$$

where $f_{\text{ave}}(Stra)$ means the average fitness value of strategy $Stra$; S_t is the payoff of tth round-robin game; T is the number of total iterations.

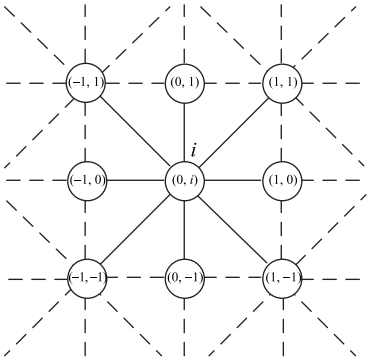


Fig. 2. The structure of the square lattice network.

1) The Tournaments Between Designed Strategy and Fixed Strategies: In this section, the tournament between designed strategy based on the TD learning method and fixed strategies is provided to verify whether the designed strategy can earn a better payoff than the other fixed strategies. The steps of the tournament are represented as follows:

Step 1: Initializing the number of iteration as 100. Randomly generating the positions of the agents with different strategies. The numbers of agents with different strategies are equal. The initial states of the agents are cooperation and defection. The total initial cooperative rate is $p_1 = 0.3$. The cooperators will be marked as red squares. The defectors will be marked as blue squares.

Step 2: During each round of the round-robin games, ith agent will play against its neighbors by the prescribed sequence which is shown as Fig. 3, and update its states according to the characteristics.

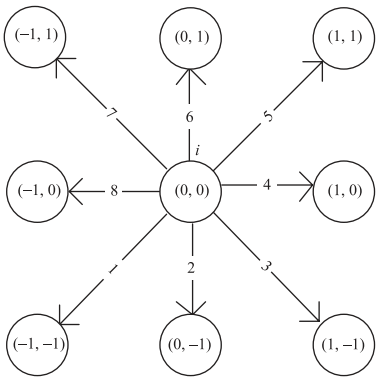


Fig. 3. The order of each round-robin game.

Step 3: Calculating the cooperative rates and average fitness values for verifying effectiveness of the mentioned manners.

During the first example, the average fitness values of the different strategies in IPD tournament are given in the Table IV. The statistical results illustrate that adaptive strategy earns a better payoff compared with other strategies. In the tournament, defective actions became extinct after 9 generations, and cooperative actions occupy the majority of the population.

TABLE IV
RESULTS OF FIRST ROUND-ROBIN IPD

Strategy	Score
Always defect	2.367
Ripoff	2.347
Always cooperate	2.289
Psycho	2.462
Tit-for-tat	2.774
Adaptive strategy	3.135
GRIM	2.763
Pavlov	2.334
Tit-for-2tit	2.597

The simulation results of the first tournament are shown in Fig. 4. Therefore, a conclusion can be obtained that mutual cooperation becomes popular among the agents and spreads fast. However, another problem is that how does the initial cooperative rate p_1 influence the result. Therefore, the simulation results based on the different initial conditions are shown as Fig. 5. Fig. 5 mentions that the three different initial values which are $p_1 = 0.3$, $p_1 = 0.5$ and $p_1 = 0.8$ lead to different balance points. However, the equilibrium points are all above 75 percent which is significantly larger than the percentage of agent with always cooperate. Therefore, mutual cooperation can be achieved between majority of the agents.

2) The Tournaments Between Designed Strategy and Other Evolutionary Learning Methods: In this section, two evolutionary learning methods are introduced into the tournaments. One is a Q-learning strategy [27], and the other one is a self-adaptive win-stay-shift reference selection strategy [28]. The environment of this tournament is also the square lattice network. The steps of the tournament are represented as follows:

Step 1: Initializing the number of iteration as 100. Randomly generating the positions for the agents with different strategies on the square lattice network. Three kinds of evolutionary strategies are evenly distributed on the network. The initial states of the agents are cooperate and defect. The initial cooperative rate is $p_1 = 0.3$. The cooperators will be marked as green squares. The defectors will be marked as blue squares.

Step 2: During each round of the round-robin, ith agent will play against its neighbors by the prescribed sequence which is shown as Fig. 3, and update its states according to their own features.

Step 3: Calculating the cooperative rates and average fitness values of the three different evolutionary strategies.

The simulation results of the second tournament are shown in Fig. 6. The simulation results illustrate that the cooperation spreads fast among the agents. The network becomes stable within 5 iterations, and the cooperative rate of 100th iteration is 0.904. The results verify that the mutual cooperation can be achieved between the agents with evolutionary strategies.

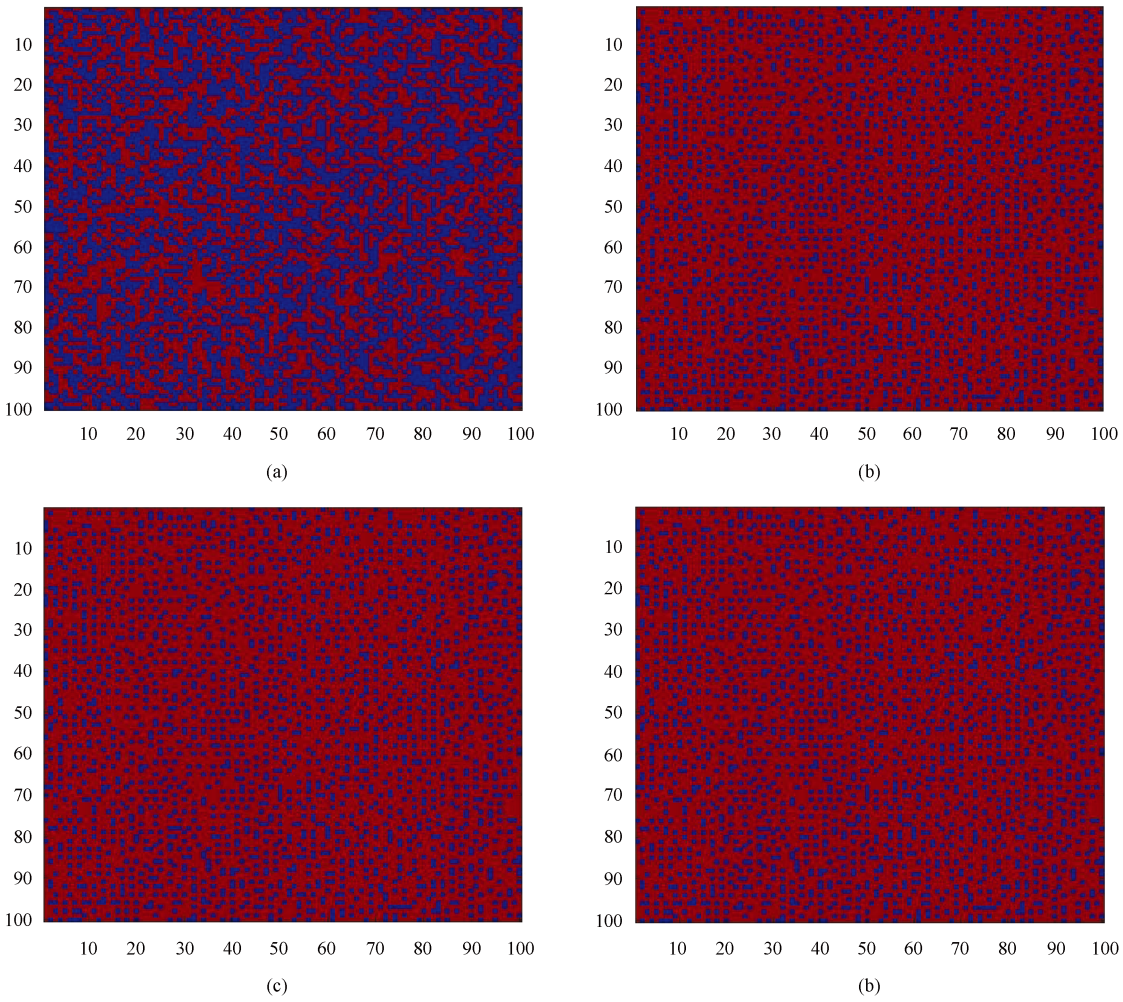


Fig. 4. The simulation results of round-robin game. (a) The distributions of the 1st round-robin game. (b) The distributions of the 10th round-robin game. (c) The distributions of the 50th round-robin game. (b) The distributions of the 100th round-robin game.

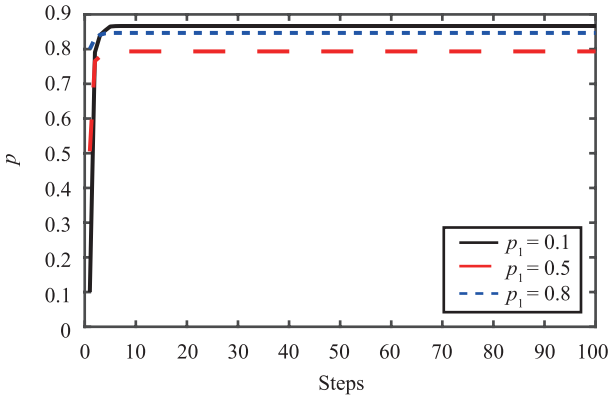


Fig. 5. The cooperative rates based on different initial values.

The average fitness values of the different strategies are shown in Table V. The results show that although the TD Learning Strategy earns a better score, the differences between the scores are minor. The reason of this phenomenon is that the cooperation spreads fast and maintained within 5 rounds.

To sum up, the simulations based on the square lattice network show that the payoff of adaptive strategy is higher than other strategies during the tournament. Most of the agents choose cooperation with others. However, not all the agents choose cooperating with the others. We give another example to verify whether the mutual cooperation can be achieved among the agents with the designed adaptive strategy.

TABLE V
RESULTS OF SECOND ROUND-ROBIN IPD

Strategy	Score
Q-learning strategy	2.83
Self-adaptive strategy	3.15
TD learning strategy	3.26

B. The Simulations Based on the Scale-free Networks

For the scale-free networks, the relationships between the agents are not homogenous. The agents prefer to make connections with the agents with advantage of fitness which are named hubs. Therefore, they may have different degrees. This kind of network is widely used in the real world, such as Internet, social networks, and so on. References [29], [30] studied the game behaviors between the two agents on the scale-free network. These studies drew a conclusion that the scale-free network can promote the mutual cooperation by

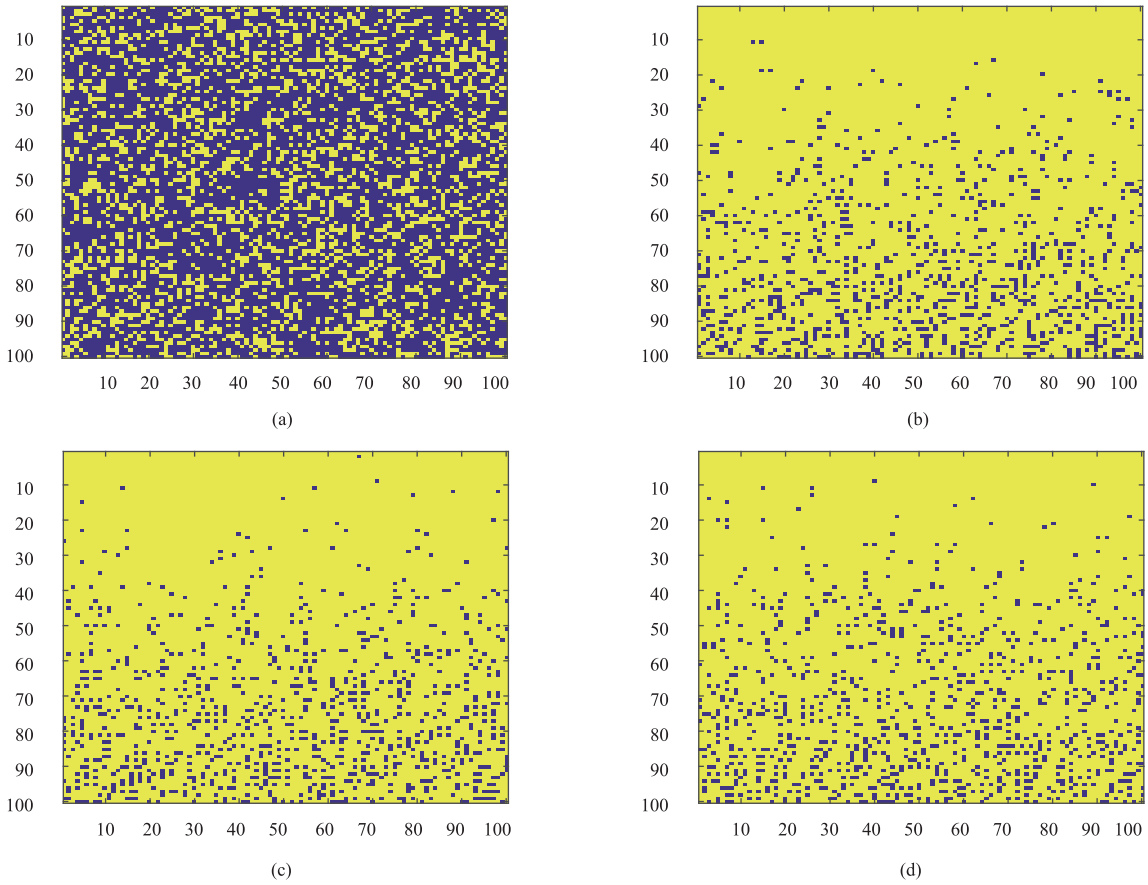


Fig. 6. The simulation results of round-robin game. (a) The distributions of the 1 st round-robin game. (b) The distributions of the 5 th round-robin game. (c) The distributions of the 10 th round-robin game. (d) The distributions of the 100 th round-robin game.

the most successive agents. Reference [29] found that the scale-free networks were extremely vulnerable to attacks, i.e., to the selection and removal of a few nodes that play the most important role in assuring the network's connectivity. The conditions of the hubs are significant for the scale-free network.

The purpose of this experiment is to figure out whether the adaptive agents can achieve mutual cooperation by their own strategies to increase the fault tolerance of the scale-free network. The typical graph of scale-free network is shown as Fig. 7 [7].

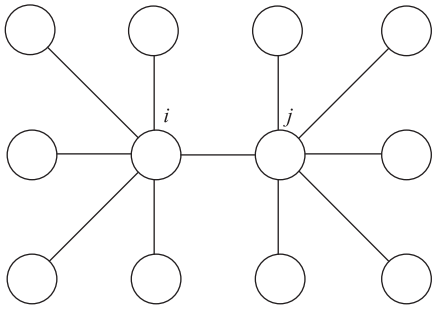


Fig. 7. The typical structure of scale-free network.

All the agents in this tournament are adaptive agents. During this tournament, the IPD runs for 100 rounds per trial with different initial values of the cooperative rate. The initial values of cooperative rate are 0.1, 0.5 and 0.8, respectively. Based on this situation, the initial states of some hub agents may be defection. Therefore, it is very difficult for the agents to achieve mutual cooperation by the characteristic of the scale-free network. The main propose of this tournament is inspecting whether the mentioned manner can optimize decision making process for helping the agents to obtain mutual cooperation. The steps of the tournament are shown as follows.

Step 1: Initializing the number of iterations as 100. Initializing the states of the agents in the scale-free network. The initial values of the cooperative rates are 0.1, 0.5 and 0.8, respectively.

Step 2: During each round of the tournament, ith agent will play against its neighbors, and update its states according to the characteristics. Since all the agents are adaptive agents, the agents will try to achieve mutual cooperation according to optimizing the decision making process.

Step 3: Calculating the cooperative rates for verifying effectiveness of the mentioned manners.

As Fig. 8 shows, the cooperative rates will convergence to 1. Therefore, the simulation results of this tournament show that no matter what the initial value is, all the agents will choose cooperation. The experimental results indicate that although the scale-free network encourages the agent to copy the strategies of the hubs which may lead the agents to

defect with each other, the adaptive agents can obtain mutual cooperation. They can make their own decisions based on their strategies to increase the fault tolerance of the scale-free network. The experimental results indicate the effectiveness of the adaptive strategies.

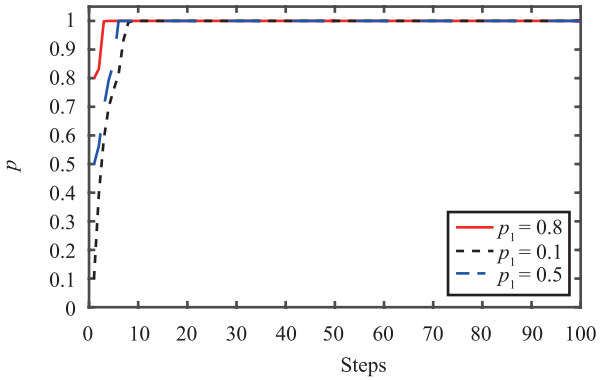


Fig. 8. The cooperative rates based on different initial values.

V. Conclusion

A reinforcement learning method was introduced to design an adaptive strategy for the IPD. The agent with adaptive strategy can make decisions under a consideration of the long-term reward. In order to verify the effectiveness of this method, three kinds of tournaments under two different environments were discussed in Section IV. The simulation results illustrated that the adaptive agents were able to cooperate with their opponents without losing competitiveness. They could achieve mutual cooperation, which is not only meaningful for the long-term reward of the team, but also the fault tolerance of the scale-free network. In our future research, we will investigate the essential relationship between the IPD and multi-agent systems in different complex networks, as well as the application of game theory for analyzing dynamics of multi-agent system.

Appendix A
Proof of Theorem 1

As for the infinite length IPD, suppose S_L is a fixed strategy with memory-L. The CS against S_L must play a periodic sequence with period less than L. Let $q_i (i = 1, \ldots, L), \sum_{i=1}^{L} q_i = 1$ denotes the probability that CS plays a sequence whose period is equivalent to i.

Suppose $Q_L(A|S_L)$ is the probability of adaptive strategy being a CS against a fixed strategy S_L with memory-L. As to the strategy with zero memory S_0, there is $Q_L(A|S_0) = q/2$. If S_0 plays a periodic sequence whose period is equivalent to two, the $Q_L(A|S_0) = q/4$. Furthermore, S_0 plays a periodic sequence whose period is equivalent to L, $Q_L(A|S_0) = q_L/2^L$.

Thus, the highest value of $Q_L(A|S_0)$ is represented as

$$Q_L(A|S_0) = \max\left(\frac{1}{2}q_1, \frac{1}{4}q_2, \ldots, \frac{1}{2^L}q_L\right).$$

The strategy with memory-1 can shift its play sequence between a determined sequence and a period-two sequence. Thereforethe maximums of $Q_L(A|S_1)$ is

$$Q_L(A|S_1) = \max\left(\frac{1}{2}q_1 + \frac{1}{4}q_2, \frac{1}{4}q_2 + \frac{1}{4}q_2, \ldots, \frac{1}{2^L}q_L + \frac{1}{4}q_2\right).$$

There are $Q_L(A|S_L) = q_1/2 + q_2/4 + \cdots + q_L/2^L$
For the adaptive strategies, the memory length is as long as the IPD. Therefore, during the infinite length IPD, the adaptive strategy has the length K which is larger than the L. Therefore, the highest value of $Q_L(A|S_L)$ is $Q_L(A|S_L) = q_1/2 + q_2/4 + \cdots + q_L/2^L$

Thus,

$$\begin{cases} Q_L(A|S_L) > Q_L(A|S_{\mathrm{L}-1}) > \cdots > Q_L(A|S_1) > Q_L(A|S_0) \\ Q_L(A|S_K) = Q_L(A|S_{\mathrm{L}}), \qquad K > L. \end{cases}$$

Due to

$$Q(S_L) = \sum_{0}^{\infty} q_i Q_i(A|S_L)$$
$$Q(S_K) = \sum_{0}^{\infty} q_i Q_i(A|S_K)$$

based on the former information

$$\begin{cases} Q_i(A|S_L) < Q_i(A|S_{L+1}), \qquad i > L \\ Q_i(A|S_L) = Q_i(A|S_{L+1}), \qquad i \leq L. \end{cases}$$

There must be $Q(A) = Q(S_K) > Q(S_{\mathrm{L}})$ for any limited number L in an infinite length IPD. Therefore, the adaptive strategies have a higher probability of being a CS against fixed strategy in certain conditions. The concepts of CS can be used to verify the evolutionary stability of the adaptive strategy.

Appendix B
Proof of Theorem 2

Based on the Theorem 1, for any fixed strategy S',

$$U(S, S) \geq U(S', S)$$

and

$$U(S, S') \geq U(S', S').$$

According to [26], the condition for a strategy S to be ESS is that for any S'

$$U(S, S) \geq U(S', S) \tag{10}$$
$$U(S, S') > U(S', S'). \tag{11}$$

Comparing with the inequalities with (10) and (11), the adaptive strategy is ESS, when played against other fixed strategy. Therefore, the adaptive strategy is ESS under certain condition.

REFERENCES

[1] J. Seiffertt, S. Mulder, R. Dua, and D. C. Wunsch, "Neural networks and Markov models for the iterated prisoner's dilemma," in *Proc. Int. Joint Conf. Neural Networks*, Atlanta, GA, USA, 2009, pp. 2860−2866.

[2] H. Y. Quek, K. C. Tan, C. K. Goh, and H. A. Abbass, "Evolution and incremental learning in the iterated prisoner's dilemma," *IEEE Trans. Evol. Comput.*, vol. 13, no. 2, pp. 303−320, Apr. 2009.

[3] R. Axelrod, *The Evolution of Cooperation*. New York, USA: Basic, 1984.

[4] M. A. Nowak, R. M. May, "Evolutionary games and spatial chaos," *Nature*, vol. 359, no. 6398, pp. 826−829, Oct. 1992.

[5] F. Fu, M. A. Nowak, and C. Hauert, "Invasion and expansion of cooperators in lattice populations: Prisoner's dilemma vs. snowdrift games," *J. Theor. Biol.*, vol. 266, no. 3, pp. 358−366, Oct. 2010.

[6] J. Liu, Y. Li, C. Xu, and P. M. Hui, "Evolutionary behavior of generalized zero-determinant strategies in iterated prisoner's dilemma," *Phys. A Stat. Mech. Appl.*, vol. 430, pp. 81−92, Jul. 2015.

[7] G. Szabó, G. Fath, "Evolutionary games on graphs," *Phys. Rep.*, vol. 446, no. 4−6, pp. 97−216, Jul. 2007.

[8] D. C. Wunsch and S. Mulder, "Evolutionary algorithms, Markov decision processes, adaptive critic designs, and clustering: Commonalities, hybridization and performance," in *Proc. Int. Conf. Intelligent Sensing and Information Processing*, Chennai, India, 2004, pp. 477−482.

[9] H. Ishibuchi and N. Namikawa, "Evolution of iterated prisoner′s dilemma game strategies in structured demes under random pairing in game playing," *IEEE Trans. Evol. Comput.*, vol. 9, no. 6, pp. 552−561, Dec. 2005.

[10] H. Ishibuchi, H. Ohyanagi, and Y. Nojima, "Evolution of strategies with different representation schemes in a spatial iterated prisoner's dilemma game," *IEEE Trans. Comput. Intell. AI Games*, vol. 3, no. 1, pp. 67−82, Mar. 2011.

[11] D. Ashlock and E. Y. Kim, "Fingerprinting: Visualization and Automatic analysis of prisoner's dilemma strategies," *IEEE Trans. Evol. Comput.*, vol. 12, no. 5, pp. 647−659, Oct. 2008.

[12] D. Ashlock, E. Y. Kim, and W. Ashlock, "Fingerprint analysis of the noisy prisoner's dilemma using a finite-state representation," *IEEE Trans. Comput. Intell. AI Games*, vol. 1, no. 2, pp. 154−167, Jun. 2009.

[13] D. Ashlock and C. Lee, "Agent-case embeddings for the analysis of evolved systems," *IEEE Trans. Evol. Comput.*, vol. 17, no. 2, pp. 227−240, Apr. 2013.

[14] J. S. Wu, Y. Q. Hou, L. C. Jiao, and H. J. Li, "Community structure inhibits cooperation in the spatial prisoner's dilemma," *Phys. A Stat. Mech. Appl.*, vol. 412, pp. 169−179, Oct. 2014.

[15] Y. Z. Cui and X. Y. Wang, "Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks," *Phys. A Stat. Mech. Appl.*, vol. 407, pp. 7−14, Aug. 2014.

[16] S. P. Nageshrao, G. A. D. Lopes, D. Jeltsema, and R. Babuška, "Port-hamiltonian systems in adaptive and learning control: A survey," *IEEE Trans. Autom. Control*, vol. 61, no. 5, pp. 1223−1238, May 2016.

[17] C. M. Liu, X. Xu, and D. W. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 45, no. 3, pp. 385−398, Mar. 2015.

[18] Y. J. Liu, Y. Gao, S. C. Tong, and Y. M. Li, "Fuzzy approximation-based adaptive backstepping optimal control for a class of nonlinear discrete-time systems with dead-zone," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 1, pp. 16−28, Feb. 2016.

[19] Y. Gao and Y. J. Liu, "Adaptive fuzzy optimal control using direct heuristic dynamic programming for chaotic discrete-time system," *J. Vibrat. Control*, vol. 22, no. 2, pp. 595−603, 2016.

[20] Y. J. Liu, L. Tang, S. C. Tong, C. L. P. Chen, and D. J. Li, "Reinforcement learning design-based adaptive tracking control with less learning parameters for nonlinear discrete-time MIMO systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 165−176, Jan. 2015.

[21] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598−1611, Aug. 2012.

[22] P. Hingston and G. Kendall, "Learning versus evolution in iterated prisoner's dilemma," in *Proc. Congr. Evolutionary Computation*, Portland, OR, USA, 2004, pp. 364−372.

[23] S. Y. Chong and X. Yao, "Multiple choices and reputation in multiagent interactions," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 689−711, Dec. 2007.

[24] E. Semsar-Kazerooni and K. Khorasani, "Multi-agent team cooperation: A game theory approach," *Automatica*, vol. 45, no. 10, pp. 2205−2213, Oct. 2009.

[25] D. Ashlock, J. A. Brown, and P. Hingston, "Multiple opponent optimization of prisoner's dilemma playing agents," *IEEE Trans. Comput. Intell. AI Games*, vol. 7, no. 1, pp. 53−65, Mar. 2015.

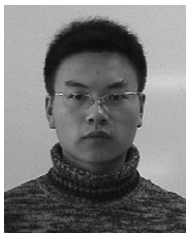
[26] J. W. Li and G. Kendall, "The effect of memory size on the evolutionary stability of strategies in iterated prisoner's dilemma," *IEEE Trans. Evol. Comput.*, vol. 18, no. 6, pp. 819−826, Dec. 2014.

[27] K. Moriyama, "Learning-rate adjusting Q-learning for prisoner's dilemma games," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology*, Sydney, NSW, Australia, 2008, pp. 322−325.

[28] X. Y. Deng, Z. P. Zhang, Y. Deng, Q. Liu, and S. H. Chang, "Self-adaptive win-stay-lose-shift reference selection mechanism promotes cooperation on a square lattice," *Appl. Math. Comput.*, vol. 284, pp. 322−331, Jul. 2016.

[29] F. C. Santos and J. M. Pacheco, "Scale-free networks provide a unifying framework for the emergence of cooperation," *Phys. Rev. Lett.*, vol. 95, no. 9, pp. Article ID 098104, Aug. 2005.

[30] F. C. Santos and J. M. Pacheco, "A new route to the evolution of cooperation," *J. Evol. Biol.*, vol. 19, no. 3, pp. 726−733, May 2006.



Lei Xue received the B.S. and M.S. degrees from Heilongjiang University and Southeast University, China, in 2009 and in 2012, respectively. He is currently a Ph.D candidate in School of Automation, Southeast University. His research interests include game theory, multi-agent system, and optimization control.

Changyin Sun (M'17) is a Professor in School of Automation, Southeast University, Nanjing, China. He received the Bachelor Degree in College of Mathematics, Sichuan University, China, and the M.S. and Ph.D. degrees in electrical engineering from the Southeast University, Nanjing, China, respectively, in 2001 and 2004. He is the Associate Editor of *IEEE Transactions on Neural Networks and Learning Systems*, *Neural Processing Letters*, *IEEE/CAA Journal of Automatica Sinica*. His research interests include intelligent control, flight control, pattern recognition, optimal theory, etc.

Donald Wunsch (F'05) is the Mary K. Finley Missouri Distinguished Professor at Missouri University of Science and Technology. He received his B.S. and M.S. degrees from University of New Mexico and University of Washington, USA, respectively. He received his Ph.D. degree from University of Washington (Seattle), USA. His key research contributions are in: clustering/unsupervised learning; adaptive resonance and reinforcement learning architectures, hardware and applications; neurofuzzy regression; traveling salesman problem heuristics; robotic swarms; and bioinformatics. He is an IEEE Fellow and previous INNS President, INNS Fellow and Senior Fellow 2007−2013, NSF CAREER Award winner, and winner of the 2015 INNS Gabor Award. He served as IJCNN General Chair, and on several boards, including the St. Patrick's School Board, IEEE Neural Networks Council, International Neural Networks Society, and the University of Missouri Bioinformatics Consortium, Chaired the Missouri University of Science and Technology Information Technology and Computing Committee as well as the Student Design and Experiential Learning Center Board.

Yingjiang Zhou received the M. S. degree from Hohai University in 2010 and the Ph.D. degree from the Southeast University, China, in 2014. He is currently working at College of Automation, Nanjing University of Posts and Telecommunications. His research interests include consensus of multi-agent, formation control of UAV and nonlinear system control.

Fang Yu received the B.S. and M.S. degrees from Huzhou Teachers College University and Henan University of Science and Technology, China, in 2006 and in 2009, respectively. She received her Ph.D. degree from Kobe University, Japan, in 2013. She is currently a researcher at Shanghai Maritime University. Her research interests include game theory, supply chain management, optimization control.