SCIENCE CHINA
Technological Sciences

• Article •

August 2015 Vol.58 No.8: 1392–1400
doi: 10.1007/s11431-015-5848-6

Markov decision evolutionary game theoretic learning for cooperative sensing of unmanned aerial vehicles

SUN ChangHao[1] & DUAN HaiBin[1,2*]

[1] *State Key Laboratory of Virtual Reality Technology and Systems, School of Automation Science and Electronic Engineering, Beihang University, Beijing 100191, China;*

[2] *Bio-inspired Autonomous Flight Systems (BAFS) Research Group, Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electronic Engineering, Beihang University, Beijing 100191, China*

Received January 31, 2015; accepted March 25, 2015; published online June 12, 2015

As one of the major contributions of biology to competitive decision making, evolutionary game theory provides a useful tool for studying the evolution of cooperation. To achieve the optimal solution for unmanned aerial vehicles (UAVs) that are carrying out a sensing task, this paper presents a Markov decision evolutionary game (MDEG) based learning algorithm. Each individual in the algorithm follows a Markov decision strategy to maximize its payoff against the well known Tit-for-Tat strategy. Simulation results demonstrate that the MDEG theory based approach effectively improves the collective payoff of the team. The proposed algorithm can not only obtain the best action sequence but also a sub-optimal Markov policy that is independent of the game duration. Furthermore, the paper also studies the emergence of cooperation in the evolution of self-regarded UAVs. The results show that it is the adaptive ability of the MDEG based approach as well as the perfect balance between revenge and forgiveness of the *Tit-for-Tat* strategy that the emergence of cooperation should be attributed to.

unmanned aerial vehicles (UAVs), iterated prisoner's dilemma (IPD), Markov decision evolutionary game (MDEG), replicator dynamics, cooperation

Citation: Sun C H, Duan H B. Markov decision evolutionary game theoretic learning for cooperative sensing of unmanned aerial vehicles. Sci China Tech Sci, 2015, 58: 1392–1400, doi: 10.1007/s11431-015-5848-6

1 Introduction

In recent years, there has been a tremendous interest in the utilization of unmanned aerial vehicles (UAVs) [1–4] over a distributed environment to cooperatively implement servicing tasks such as surveillance and reconnaissance. With the advances of relatively technologies, an individual UAV can be seen as an intelligent rational agent, which is concerned with its own fitness when making decisions. Generally, the fitness is mainly related to the reward received and the energy consumed. In many applications, the individual fitness of one UAV depends not only on its own action but also on the strategies of the neighbors. Such strategic interactions are typically modeled by game theory, where sharp conflicts between individual interests and collective welfare are often observed in many scenarios known as dilemmas. Thus, it has become a hot topic in the field of cooperative control and learning algorithms [5,6] how to ensure cooperation among self-regarded agents.

As a matter of fact, cooperation plays a vital role in the formulation of complex biological structures, ranging from multi-cellular organisms to human societies [7]. Hence, the evolution of cooperation has been intensively addressed by researchers from natural and social sciences. In particular, as the most stringent scenario of conflicts, the prisoner's dilemma (PD) has long been considered to be a paradig-

*Corresponding author (email: hbduan@buaa.edu.cn)

matic metaphor for studying cooperation. In its original form, the PD is a two-player non-zero-sum game, where each confronts two choices: Cooperation (C) or defection (D) and makes its choice without knowing *a priori* how the other will act. There are four possible outcomes for the conventional version of this game: (1) Mutual cooperation (C, C), (2) mutual defection (D, D), and (3) the situation in which one cooperates while the other defects (C, D) or (D, C). Mutual cooperation offers each a reward R and mutual defection pays each a punishment P. In the case of (3), where one chooses to cooperate and the other prefers to defect, the cooperator gets the sucker's payoff S while the defector gains the temptation T. Under the framework of classical game theory, which involves the study of mathematical models of conflict and cooperation between rational decision-makers [8], defection D yields the dominant strategy. This leads to the dilemma that if both choose to defect in the light of rationalness, then both do worse than if both cooperated [9]. One possible way out for this dilemma is the iterated prisoner's dilemma (IPD), where players meet more than once and play repeatedly, supposing that they remember the results of previous encounters. Much more attention has been given to the paradox of the IPD for the light it may shed on the evolution of altruistic or cooperative behavior since Nowak discovered the astonishingly complex and spatially chaotic patterns in an evolutionary spatial IPD, where cooperation and defection persist indefinitely [10].

Up to now, the evolution of cooperation has been typically approached from a Darwinian evolutionary perspective within the framework of evolutionary game theory [11]. Different from classical game theory, in an evolutionary game players have only bounded rationality and repeatedly engage in strategic evolutions to eventually achieve a refined equilibrium. This progress typically follows the replicator dynamics (RD) [12], which originally models how natural selection affects the frequency of animals (players) in each habitat (game strategy). As its name suggests, evolutionary game approaches have been mostly applied to situations including biological [13], economical [14], and social systems [15], where understanding the conditions for the emergence and persistence of cooperative behavior among selfish individuals is a central problem [16]. Since the evolutionary game theory allows the players to learn from the environment and make individual decisions with little information exchange, it has also been successfully extended to technological problems in many engineering scenarios. Yang and Li [17] solved the vertex cover problem in a distributed networking system by providing an evolutionary snow drift game based optimization framework. The optimal solution belongs to the set of Nash equilibria and is guaranteed by the memory-based best response update rule. The set k-cover problem in the field of wireless sensor networks has also been tackled within the framework of evolutionary game theory, where suboptimal or optimal solutions can be obtained using log-linear learning [18,19] or N-person card game approaches [20,21]. Semasinghe et al. [22] addressed distributed resource allocation in self-organizing small cells by proposing an evolutionary game theoretic approach, where RD is employed to model the strategy adaption. Obando et al. [23] formulated temperature control in buildings as a dynamic resource allocation problem and designed an RD based control law. Other problems that have been studied using evolutionary game theory include joint spectrum sensing and access in cognitive radio networks [24], adaptive filtering networks in wireless sensor networks [25], and cluster analysis in the field of pattern recognition [26].

However, classical evolutionary game theory has great difficulties in dealing with the IPD game involving a long decision process, where players make decisions more than once and a pure strategy corresponds to the action sequence taken during the whole game duration. For an IPD game with size n, there are 2^n pure strategies available for one player, which makes it a computationally expensive task to study the evolution in the IPD by using classical evolutionary game theory and traditional optimization methods [27–29]. By introducing randomness into deterministic evolutionary games, Shapley develops stochastic evolutionary game theory in 1953 [30]. As one typical class of stochastic evolutionary games, Markov decision evolutionary games (MDEGs) [31] prove an efficient approach for designing learning methods by assuming that individuals play and make decisions according to a Markov decision process. So far, the MDEG theory has been applied in energy management in distributed networks [32], analysis of 2×2 spatial games [9] and so on. In the open literature, learning algorithms and relative characteristics of the IPD game have also been studied by using computational evolutionary approaches or mathematical approaches [33–35]. However, to the best knowledge of the authors, the MDEG theory and learning methods for the IPD have been studied independently.

The main contribution of this paper is threefold. Firstly, we formulated the strategic interactions among UAVs in a task of surveillance and reconnaissance as an IPD game. Secondly, we developed an MDEG learning method for one of the UAVs to achieve the optimal solution with respect to the collective fitness. Finally, we analyzed the emergence of cooperative behavior among selfish individuals using numerical simulation results. In the proposed learning algorithm, each individual in the population randomly interacts with the same opponent in an IPD game of length T by following a Markov decision process, i.e., one selects an action in the action space with a probability that depends only on the current state. By introducing a stochastic process and representing one player's strategy by a probability vector, the MDEG based approach simplifies the computational process and improves the performance of the algorithm. Computational experiments show the proposed learning approach not only improves the game payoff of the tagged

UAV but also makes great contributions to the improvement of the collective payoffs of the team.

The structure of this paper is as follows. In Section 2, we give the basic theory of evolutionary games, including the evolutionary stable strategy (ESS) and the replicator dynamics. Section 3 formulates the strategic interactions between two UAVs as an IPD game and presents the architecture of the proposed approach in details. Results and discussions are given in Section 4, which is followed by conclusions in Section 5.

2 Evolutionary game theory

In evolutionary game theory, the games are considered from a perspective different from ordinary game theory, where rational players meet only once in a game and decide the best strategy by taking into account other's behavior [36]. Instead, evolutionary games involve a population of individuals, each programmed to adopt a strategy and interact randomly with other individuals. The probability of a particular individual's survival and reproduction depends on the earned payoff in the game. Furthermore, the evolutionary game is not concerned with the players' choices but focuses on the evolutionary characteristics of the population's behavior, especially on the asymptotic state in the long run.

Consider a large enough population of n individuals (also called players), we assume that each individual has an identical action set $A=\{1,2,\cdots,m\}$. Different from classical game theory, the players here have limited rationality, i.e., they are not smart enough to select the best strategy at once. Instead, at every time instant t, interactions occur between individual i, $i=1, 2,\cdots, n$, and another player randomly drawn from the population. This can be represented by a two-player non-zero-sum bimatrix game. Suppose each adopts a mixed strategy $\boldsymbol{p}$, which corresponds to a probability distribution over the action space A. Denote by $R(\boldsymbol{p},\boldsymbol{q})$ the expected payoff of the individual with mixed strategy $\boldsymbol{p}$ when encountered with a player with strategy $\boldsymbol{q}$. During each interaction, whose payoff matrix is defined as M, the payoff for the tagged individual with strategy $\boldsymbol{p}$ is calculated by $R(\boldsymbol{p},\boldsymbol{q})=\boldsymbol{p}^{\mathrm{T}}M\boldsymbol{q}$, where mixed strategy $\boldsymbol{p}=(p_1, p_2,\cdots,p_{\mathrm{m}})^{\mathrm{T}}$, with p_i representing the probability of choosing the ith pure strategy in the action set A and $\sum_{i=1}^{m} p_i = 1$. Strategy $\boldsymbol{q}$ is similarly defined.

Central in the standard evolutionary game theory are the ESS and the RD, which describe the asymptotic strategy and the evolution dynamics of the game respectively. The ESS, first proposed by Smith and Price in 1973 [11], is a genetically-determined strategy that tends to persist once it is prevalent in a population. It is defined by stability against any mutant strategy $\boldsymbol{q}$, which appears in the population while all the other individuals adopt strategy $\boldsymbol{p}$. A mixed strategy $\boldsymbol{p}$ is said to be evolutionary stable if, whenever a small group of mutants appear, the individuals with strategy p get payoff strictly higher than the mutants. Thus, an ESS in a population is defined as follows.

Definition 1 (*ESS of a population*) [31]. A mixed strategy $\boldsymbol{p}$ is called an ESS of the population if and only if the following condition holds:

$$\begin{aligned}&\exists\bar{\varepsilon}>0,\forall\varepsilon\in(0,\bar{\varepsilon}),\forall\boldsymbol{q}\neq\boldsymbol{p},\\&\varepsilon R(\boldsymbol{p},\boldsymbol{q})+(1-\varepsilon)R(\boldsymbol{p},\boldsymbol{p})>\varepsilon R(\boldsymbol{q},\boldsymbol{q})+(1-\varepsilon)R(\boldsymbol{q},\boldsymbol{p}),\end{aligned}\tag{1}$$

where ε is the fraction of mutant strategy $\boldsymbol{q}$.

If mixed strategy $\boldsymbol{p}$ is a Nash equilibrium, then

$$\begin{aligned}&\forall\boldsymbol{q}\neq\boldsymbol{p},\\&R(\boldsymbol{p},\boldsymbol{p})\geqslant R(\boldsymbol{q},\boldsymbol{p}).\end{aligned}\tag{2}$$

The ESS is a stronger definition compared with the Nash equilibrium, and the relationship between the ESS and Nash equilibrium can be deduced from the following Theorem 1.

Theorem 1 [31]. The condition in eq. (1) holds if and only if the following condition holds:

$$\forall\boldsymbol{q}\neq\boldsymbol{p},R(\boldsymbol{p},\boldsymbol{p})>R(\boldsymbol{q},\boldsymbol{p}),\tag{3}$$

or

$$R(\boldsymbol{p},\boldsymbol{p})=R(\boldsymbol{q},\boldsymbol{p}),R(\boldsymbol{p},\boldsymbol{q})>R(\boldsymbol{q},\boldsymbol{q}).\tag{4}$$

Remark 1. Obviously, when condition (3) holds, mutants gain a lower fitness compared with that of the original population. Hence, the number of the mutants tends to decrease, making strategy $\boldsymbol{p}$ immune to any mutant strategy $\boldsymbol{q}$. On the other hand, when condition (4) is satisfied, there will be more frequent competence between p and q as the population of mutants grows. In these cases, the condition $R(\boldsymbol{p},\boldsymbol{q})>R(\boldsymbol{q},\boldsymbol{q})$ guarantees that the population of mutants narrows down to 0.

Remark 2. An ESS must be a Nash equilibrium, as conditions (3) and (4) indicate the sufficient condition for the Nash equilibrium in eq. (2). However, a Nash equilibrium is not necessarily a ESS, for when $R(\boldsymbol{p},\boldsymbol{p})=R(\boldsymbol{q},\boldsymbol{p})$, eq. (4) cannot be guaranteed to hold.

Replicator dynamics describes the selection process and evolution of strategy distributions in the considered population. Let the state of the population be represented by $\boldsymbol{p}=(p_1, p_2,\cdots,p_{\mathrm{m}})^{\mathrm{T}}$, where p_i is the proportion of the individuals with the ith pure strategy. Note that a state p can also be considered as a mixed strategy, where p_i means the probability of one random player choosing the ith pure strategy. In each iteration generation, the growth rate of the individuals with strategy i is supposed to be in proportion to the difference between the expected payoff earned by an individual with strategy i and the average payoff of the whole population. Hence, the replicator dynamics is represented as

$$\dot{p}_i = p_i(R(e_i, \boldsymbol{p}) - R(\boldsymbol{p}, \boldsymbol{p})), \tag{5}$$

$$e_i = (\underbrace{0,0,\cdots,0}_{i-1},1,\underbrace{0,\cdots,0}_{m-i}), \tag{6}$$

where e_i is represented by an m-dimensional unit vector with the ith entry equal to 1, meaning selecting the ith pure strategy. For example, with $e_1 = (1,\underbrace{0,\cdots,0}_{m-1})$, one individual adopts the first pure strategy, while $e_m = (\underbrace{0,0,\cdots,0}_{m-1},1)$ indicates choosing the last pure strategy.

3 Learning approach in cooperative sensing of UAVs

This section firstly presents the game model of UAV cooperative sensing. To achieve the optimal solution regarding the global performance, we propose a learning algorithm based on the MDEG theory. In the algorithm, each play follows stationary Markov policy, which is updated using the RD equation.

3.1 Markov policy

In the MDEG theory, the decision making process is modeled as a Markov decision process [31]. A Markov decision process model consists of: (1) A set of possible states S, (2) a set of possible actions $A=\{1,2,\cdots,m\}$, and (3) a real valued reward function $R(s,a)$, $s\in S$, $a\in A$, and a set of transition probabilities $P(s'|s,a)$, $s\in S$, $a\in A$, which represents the probability of the state transiting from s to s' when using action a. In a decision process, a player chooses an action $a\in A$ at each time instant t, according to the historical information h_t. Generally, the history information h_t consists of the set of sates $(s_0, s_1,\cdots,s_{t-1})$ previous to time t, the sequence of executed actions $(a_0, a_1,\cdots,a_{t-1})$, and the current state s_t. Denote by $u=(u_1, u_2,\cdots,u_T)$ a general decision policy, where T is the length of the relative process and u_i is the decision function mapping history h_i to the set of distributions over A. Specifically, u_i is given as follows:

$$u_i = (u_i^1, u_i^2, \cdots, u_i^m), \tag{7}$$

where u_i^j is the function mapping history h_i to the probability of choosing the jth action at time i, i.e., the ith interaction, and $m=|A|$ is the number of possible actions. In an IPD game, each player adopts a decision policy and takes actions in accordance with the selected policy and the history.

A process is called a Markov decision process, if the history h_t is reduced to contain only the current state s_t, i.e., one does not rely on the past information when making decisions, but selects actions over the set of possible action A based on its policy u and the current state s only. In this case, the policy u is called a *Markov policy*, which is described as follows:

$$u_i : S \to A, \tag{8}$$

$$u : S^T \to A^T, \tag{9}$$

where $S^T = \underbrace{S\times\ldots\times S}_{T}$ represents the possible state set of the whole decision process, with $(s_1, s_2, ..., s_T)\in S^T$, and $A^T = \underbrace{A\times\ldots\times A}_{T}$ is the set of available action sequences and $(a_1, a_2, ..., a_T)\in A^T$.

A Markov policy u is called a *stationary Markov policy*, if the element u_i is time independent, i.e., one makes decisions according to an identical mapping function u_i, $u_1=u_2=\cdots=u_T$. Similar to the standard game theory, a stationary Markov policy u is defined as a *pure stationary Markov policy*, if u_i maps the current states to a given action. In the same token, a *mixed stationary Markov policy* refers to the ones in which u_i maps the current states to the set of probability distributions over the available action space.

3.2 Game formulation

Consider two UAVs, i.e., UAV 1 and UAV 2, that are sent to carry out a surveillance and reconnaissance task over a specific region R. At each decision moment t, each UAV could choose from actions of cooperation (C) and defection (D). Specifically, each has a basic payoff r_0 if they both choose D. Cs sense the region while Ds do not but overhear the information from Cs' communication with the ground station. Sensing the region alone costs a UAV a certain amount of energy c while cooperative sensing costs each $c/2$. Obtaining the detection information yields a reward of r and overhearing the detection result costs m. As a result, the interaction at each time t can be modeled by a payoff matrix in Table 1. It is easy to check that it is a PD game if and only if $c>r$. Accordingly, the interactions in UAV sensing over a time duration of T can be modeled by an T-IPD.

3.3 MDEG based learning algorithm

Define state s_t of the game at iteration t as the action pair executed by both players at time t–1 as follows:

$$s_t = a_{t-1}^1 a_{t-1}^2, \tag{10}$$

where a_{t-1}^1 and a_{t-1}^2 are the actions taken at time t–1 by

Table 1 Payoff matrix of the sensing game for UAVs

UAV 1\UAV 2	C	D
C	($r-c/2+r_0$, $r-c/2+r_0$)	($r-c+r_0$, $r-m+r_0$)
D	($r-m+r_0$, $r-c+r_0$)	(r_0, r_0)

UAV 1 and 2, respectively. Specifically, for each player in the interaction there are two available actions and four possible states, which are given as follows:

$$A = \{C, D\}, \tag{11}$$

$$S = \{CC, CD, DC, DD\}. \tag{12}$$

A mixed stationary Markov policy in this game is a four-dimension vector $(p_{CC}, p_{CD}, p_{DC}, p_{DD})$, with each element p_s corresponding to the probability of selecting action C in the state of s. Consider a population of n individuals, each of which adopts a pure stationary Markov policy. Such a pure policy assigns each state in S a given action. For example, (0,0,0,0) is the strategy that always defects, whereas strategy P=(1,0,1,1) defects only after receiving a "sucker's payoff" at the state of CD, but cooperates in the other states CC, DC or DD, which is illustrated as in Figure 1. In that case, the whole population can be seen as a mixed stationary Markov policy $P = (p_{CC}, p_{CD}, p_{DC}, p_{DD})$, with p_s being the proportion of the individuals preferring to choose action C when encountered with state s.

Suppose the given strategy for UAV 1 is *Tit-for-Tat*, which is a highly effective strategy for the IPD first introduced by Rapoport in Robert Axelrod's two tournaments held in the 1980s. An agent using this strategy will first cooperate, and then subsequently replicate an opponent's previous action. An illustration for the action sequences taken by both players in a 10-IPD is shown in Figure 2, where UAV 2 with a pure stationary Markov policy (1,1,0,1) plays against the *Tit-for-Tat* strategy. Note that Markov policy (1,1,0,1) takes a random action D at the beginning of the game, while the *Tit-for-Tat* strategy starts with action C. UAV 2 chooses to defect only when it took action C and was betrayed by UAV 1 by action D in the previous encounter.

To ensure cooperation and the optimal solution with

Current States	C,C	C,D	D,C	D,D
Pure policy (1,0,1,1)	p_{CC}=1	p_{CD}=0	p_{DC}=1	p_{DD}=1
Taken actions	C	D	C	C

Figure 1 Actions corresponding to pure policy (1,0,1,1).

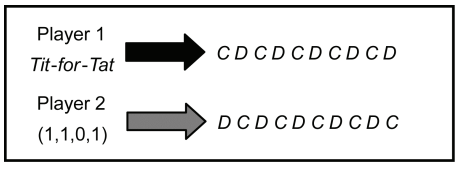

Figure 2 Action sequence taken by UAV 1 with *Tit-for-Tat* and UAV 2 with Markov policy (1,1,0,1).

regards to the collective payoff of both UAVs, we devise a learning algorithm for UAV 2 by using the MDEG theory the RD equation. The algorithm involves a large population of individuals, who randomly interact with UAV 1 in a T-IPD game with a length of T. The population is randomly initialized as a mixed stationary Markov policy $P^1 = (p^1_{CC}, p^1_{CD}, p^1_{DC}, p^1_{DD})$, with $p^1_{CC}, p^1_{CD}, p^1_{DC}$, and p^1_{DD} randomly chosen between 0 and 1. The subscript in the policy P^1 denotes the 1st evolution generation. In each generation, individuals are randomly drawn from the population to play the IPD game with *Tit-for-Tat*. Similar to the replicator dynamics in the standard evolutionary game theory, the mixed stationary Markov policy in the tth generation is updated as follows:

$$\dot{p}^t_s = p^t_s(R(E_s, P_G) - R(P^t, P_G)), s \in S, \tag{13}$$

where $R(E_s, P_G)$ represents the expected payoff of policy E_s against a given policy P_G. $R(P^t, P_G)$ is the average payoff of the whole population. Note that E_s refers to the case where each individual in the population adopts the same pure policy relative to state s while keeping the rest unchanged. In terms of the mixed strategy of the whole population, E_s is expressed specifically as follows:

$$\begin{aligned} E_{CC} &= (1, P^t(2), P^t(3), P^t(4)), \\ E_{CD} &= (P^t(1), 1, P^t(3), P^t(4)), \\ E_{DC} &= (P^t(1), P^t(2), 1, P^t(4)), \\ E_{DD} &= (P^t(1), P^t(2), P^t(3), 1). \end{aligned} \tag{14}$$

The steps involved in the proposed MDEG based learning approach for UAV 2 are given below.

Step 1. Represent the evolutionary polices of UAV 2 by a population of individuals of size n.

Step 2. Initialize the population with a mixed stationary Markov policy $P^t, t = 1$.

Step 3. Each individual in the population is governed by the same Markov decision process according to policy P^t. Individuals are randomly drawn to play the T-IPD game against the given strategy P_G.

Step 4. Calculate the expected payoff of a random individual in the population during a T-IPD game against strategy P_G as follows:

$$R(P^t, P_G) = \frac{1}{n}\sum_{i=1}^{n} r(\Gamma^i_{P^t}, P_G), \tag{15}$$

where $\Gamma^i_{P^t}$ is the action sequence executed by the ith individual in accordance with the current Markov policy P^t, and r is the payoff function. Note that the action sequences selected by individuals differ from one another due to the randomness of the adopted mixed Markov policy.

Step 5. Calculate the average payoff of policy E_s below

$$R(E_s, P_G) = \frac{1}{n}\sum_{i=1}^{n} r(\Gamma_{E_s}^{i}, P_G), \tag{16}$$

where $\Gamma_{E_s}^{i}$ means, by the same token, the ith individual's action sequence following Markov policy E_s, whose specific expressions are shown in eq. (14).

Step 7. Choose and save the best action sequence as $\Gamma_{P^t}^{\text{best}}$.

Step 8. Update the mixed Markov policy P^t according to the replicator dynamics given by eqs. (13) and (14), where the rate of change is proportional to the difference between the expected payoffs of policy E_s and the average payoff earned by the whole population; set $t := t+1$.

Step 9. If the terminal criterion is met, go to Step 10; otherwise, return to Step 3.

Step 10. Output the current Markov policies P^t and the best action sequence $\Gamma_{P^t}^{\text{best}}$.

The flow chart of the proposed approach is shown in Figure 3.

4 Experimental results

This section presents the experimental results of the proposed MDEG based approach for the IPD game. Individuals from the population plays an IPD game of length T=30 against a given strategy *Tit for Tat*. The individuals use the MDEG theory to evolve their strategies for the purpose of maximizing the overall payoff. A population with size n=100 is simulated in a 30-IPD game, and the payoff matrix for a PD game is given as in Table 2, where r=5 and c=6, r_0=m=1. Table 3 shows the parameter setting of the simulations.

Figures 4–6 show the convergence behavior of the population in an MDEG against the *Tit-for-Tat* strategy. The evolution of Markov polices in Figure 4 shows that as the evolution continues, the probabilities P_{CC}, P_{CD}, P_{DC}, and P_{DD} progressively evolve to 1. That is, the mixed stationary

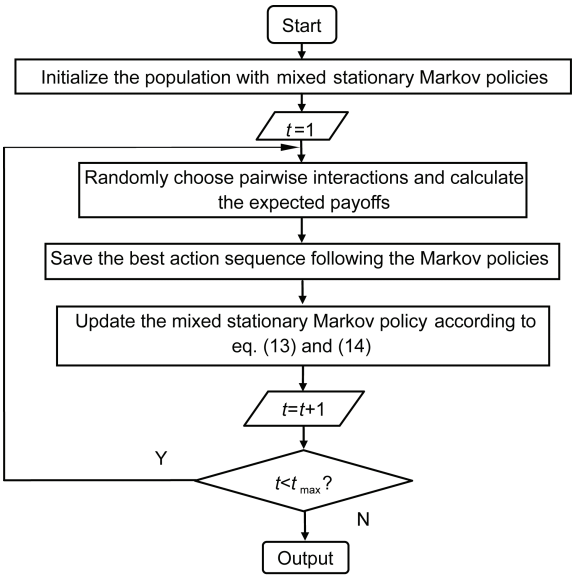

Figure 3 Flowchart of our proposed MDEG based approach.

Table 2 Payoff matrix of the PD game

Player 1\Player 2	*C*	*D*
C	(3,3)	(0,5)
D	(5,0)	(1,1)

Table 3 Parameter setting

Parameter	Meaning	Value
n	population size	100
T	game length	30
Δt	simulation step size	0.1 s
$t_{\max}$	simulation length	30

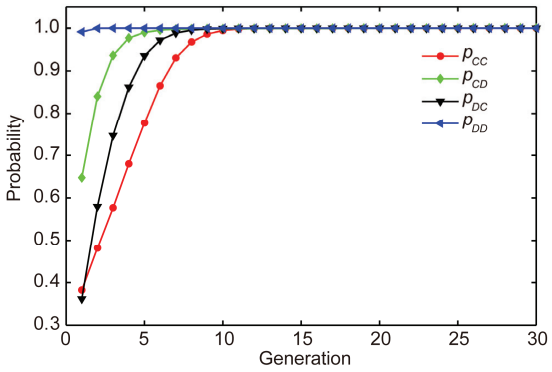

Figure 4 (Color online) Convergence behavior of the Markov policy.

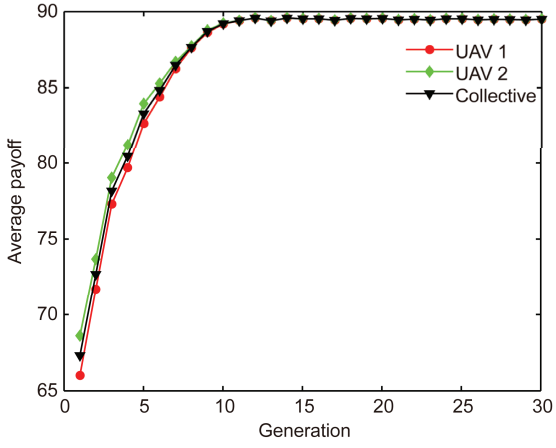

Figure 5 (Color online) Convergence behavior of the average payoff.

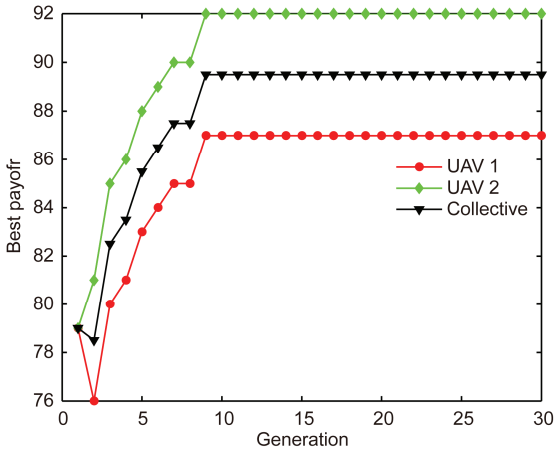

Figure 6 (Color online) Convergence behavior of the best payoff in a large population.

Markov policy represented by the whole population converges to a pure Markov strategy (1,1,1,1). The result indicates that in the assumption of a Markov process, every individual in the population comes to an agreement of seeking cooperation after a long time of evolution [37].

It is interesting to note that the *Tit-for-Tat* strategy is essentially a pure stationary Markov strategy (1,0,1,0), which starts with action *C* and the following actions only depend on the previous action taken by the opponent. *Tit-for-Tat* always chooses to cooperate if encountered with the states of *CC* and *DC*, and insists on defecting when the state is *CD* or *DD*.

As shown in Figure 5, the average payoffs of UAV 1 with the *Tit-for-Tat* strategy, UAV 2 represented by the evolving population, and the collective payoffs of both players all demonstrate an increasing tendency through the evolution of the selfish individuals, whose purposes are to maximize their own payoffs. The detailed evolution process of the Markov policies and payoffs are given in Table 4.

There are mainly two reasons for the emergence of the observed win-win cooperation. In the first place, it is attributed to the *Tit-for-Tat* strategy which starts with cooperation and keeps a perfect balance between revenge and forgiveness in the following interactions [38]. In the second place, the credit goes to the adaptive ability of the proposed MDEG based approach, which dynamically updates its policy according to the difference between the expected payoff of a particular individual and the average payoff of the population.

The proposed learning approach also figures out the best action sequence in *T*-IPD game against the *Tit-for-Tat* strategy, which is presented in Figure 6 and Table 5. Figure 6 shows that when UAV 2 adopts the best action sequence resulting from the current Markov policy of the population, the payoffs of both UAV 1 and UAV 2 increase with the evolution. Table 5 presents the evolution of UAV 2's best action sequence by detailing the action sequences of five sampled generations, i.e., 0, 5, 10, 20 and 30. As the Markov policy evolves with the iteration, the best action sequence converges to such a pattern, where one cooperates in order to earn the "reward" but defects in the last encounter to cheat the opponent out of the "temptation", thus maximizing its overall payoff.

Remark 3. Although the other approaches such as the genetic algorithm (GA) in [33–42] can also achieve the best action sequence against *Tit for Tat*, the proposed learning algorithm indeed has advantages over the GA. The proposed algorithm not only obtains the best action sequence but also a sub-optimal Markov policy that is independent of the game duration. Hence, even the game length extends to T=100, UAV 2 can still make use of the sub-optimal Markov policy $\boldsymbol{p}$=[1,1,1,1] to improve its payoff as well as the global performance. However, the result from the genetic algorithm is only suitable to a fixed game length T=30. In addition, the GA suffers from the need for excessive com-

Table 4 Evolution process of the 30-IPD game

Generation	UAV 2			UAV 1		Average collective payoff
	Markov policy	Average payoff	Best payoff	Average payoff	Best payoff	
0	(0.3433,0.4378,0.2139, 0.8356)	68.61	79	66.01	79	67.31
5	(0.6800,0.9768,0.8619,1.0000)	83.95	88	82.65	83	83.30
10	(0.9861,0.9986,0.9979,1.0000)	89.23	92	89.18	87	89.20
20	(1.0000,0.9996,1.0000,1.0000)	89.55	92	89.55	87	89.55
30	(1.0000,0.9998,1.0000,1.0000)	89.50	92	89.50	87	89.50

Table 5 Evolution of the best action sequence of UAV 2

Generation	Best action sequence of UAV 2
0	*CDCDC CDCCD CDCCC DCDCC CCCDC DDCCC*
5	*DCCCC CDCCC CCCCC DCCCC DCCCC CCCCD*
10	*CCCCC CCCCC CDCCC CCCCC CCCCC CCCCD*
20	*CCCCC CCCCC CCCCC CCCCC CCCCC CCCCD*
30	*CCCCC CCCCC CCCCC CCCCC CCCCC CCCCD*

putational power with the increase of the game length [43].

5 Conclusion

To achieve the optimal solution for UAVs that are carrying out a sensing task, this paper formulated the interactions as an IPD game and presented an MDEG theory based learning approach. Also, the emergence of cooperation among selfish and competitive individuals was also studied. In the proposed learning algorithm, UAV 2 plays the game in accordance with a Markov decision process and adopts a mixed Markov policy. This policy is represented by a large population of selfish and boundedly rational individuals with pure Markov policies. The Markov policy is updated according to the difference between the expected payoffs of a particular policy and the average payoff earned by the whole population. Simulation results with the IPD game showed that by evolving its strategy according to the MDEG theory, UAV 2 not only improves its own game payoff but also makes great contributions to the improvement of the collective payoffs of the team. The obtained results lead us to remark the good performance of MDEG based learning method in terms of accuracy, efficiency and contribution to the explanation of emergence of cooperation in nature and social life.

The cooperation resulting from the Markov decision based evolutionary game is attributed to the following two aspects:

The *Tit-for-Tat* strategy has a perfect balance between revenge and forgiveness.

The proposed MDEG based approach has the adaptive ability to dynamically update its policy according to the difference between the expected payoff of a particular policy and the average payoff of the population.

Our future work will focus on the extension of the MDEG theory to spatial games and its application to other cooperative control problems of UAVs.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61425008, 61333004 and 61273054), Top-Notch Young Talents Program of China, and Aeronautical Foundation of China (Grant No. 20135851042).

1 Duan H B, Li P. Progress in control approaches for hypersonic vehicle. Sci China Tech Sci, 2012, 55: 2965–2970
2 Duan H B, Shao S, Su B W, et al. New development thoughts on the bio-inspired intelligence based control for unmanned combat aerial vehicle. Sci China Tech Sci, 2010, 53: 2025–2031
3 Duan H B, Sun C H. Pendulum-like oscillation controller for micro aerial vehicle with ducted fan based on LQR and PSO. Sci China Tech Sci, 2013, 56: 423–429
4 Zhang X Y, Duan H B, Yu Y X. Receding horizon control for multi-UAVs close formation control based on differential evolution. Sci China Tech Sci, 2010, 53: 223–235
5 Jiang W J, Zhang L M, Wang P. Dynamic scheduling model of computing resource based on MAS cooperation mechanism. Sci China Inf Sci, 2009, 52: 1302–1320
6 Li Y B, Wang H M, Yin Q Y, et al. Fair relay selection in decode-and-forward cooperation based on outage priority. Sci China Inf Sci, 2013, 56: 1–10
7 Chen X, Fu F, Wang L. Influence of different initial distributions on robust cooperation in scale-free networks: A comparative study. Phys Lett A, 2008, 372: 1161–1167
8 Myerson R. Game Theory: Analysis of Conflict. Cambridge, Massachusetts: Harvard University Press, 1991
9 Fort H, Sicardi E. Evolutionary Markovian strategies in 2×2 spatial games. Physica A, 2007, 375: 323–335
10 Nowak M A, May R M. Evolutionary games and spatial chaos. Nature, 1992, 359: 826–829
11 Smith J M, Price G R. The logic of animal conflict. Nature, 1973, 246: 15
12 Taylor P D, Jonker L B. Evolutionary stable strategies and game dynamics. Math Biosci, 1978, 40: 145–156
13 Aristotelous A C, Durrett R. Chemical evolutionary games. Theorl Popul Biol, 2014, 93: 1–13
14 Liu D, Li H, Wang W, et al. Scenario forecast model of long term trends in rural labor transfer based on evolutionary games. J Evol Econ, 2015: 1–22
15 Szolnoki A, Mobilia M, Jiang L L, et al. Cyclic dominance in evolutionary games: a review. J R Soc Interface, 2014, 11: 20140735
16 Du W B, Cao X B, Hu M B, et al. Effects of expectation and noise on evolutionary games. Physica A, 2009, 388: 2215–2220
17 Yang Y, Li X. Towards a snowdrift game optimization to vertex cover of networks. IEEE Trans Cybern, 2013, 43: 948–956
18 Tatarenko T. Proving convergence of log-linear learning in potential games. In: IEEE American Control Conference (ACC), Portland, 2014. 972–977
19 Tatarenko T. Log-linear learning: convergence in discrete and continuous strategy potential games. In: 53rd IEEE Conference on Decision and Control, Los Angeles, 2014. 426–432
20 Wang Q, Yan W, Shen Y. *N*-person card game approach for solving set *k*-cover problem in wireless sensor networks. IEEE Trans Instrum Meas, 2012, 61: 1522–1535
21 Ai X, Srinivasan V, Tham C K. Optimality and complexity of pure Nash equilibria in the coverage game. IEEE J Sel Area Comm, 2008, 26: 1170–1182
22 Semasinghe P, Hossain E, Zhu K. An evolutionary game for distributed resource allocation in self-organizing small cells. IEEE Trans Mobile Comput, 2015, 14: 274–287
23 Obando G, Pantoja A, Quijano N. Building temperature control based on population dynamics. IEEE Trans Control Syst Technol, 2014, 22: 404–412.
24 Jiang C, Chen Y, Gao Y, et al. Joint spectrum sensing and access evolutionary game in cognitive radio networks. IEEE Trans Wireless Commun, 2013, 12: 2470–2483
25 Jiang C, Chen Y, Liu K J R. Distributed adaptive networks: A graphical evolutionary game-theoretic view. IEEE Trans Signal Proc, 2013, 61: 5675–5688
26 Li Q, Chen Z, He Y, et al. A novel clustering algorithm based upon games on evolving network. Expert Syst Appl, 2010, 37: 5621–5629
27 Zhang H Q, Hu X T, Shao X D, et al. IPSO-based hybrid approaches for reliability-redundancy allocation problems. Sci China Tech Sci, 2013, 56: 2854–2864
28 Zuo Y T, Gao Z H, Chen G, et al. Efficient aero-structural design optimization: Coupling based on reverse iteration of structural model. Sci China Tech Sci, 2015, 58: 307–315
29 Gu Y S, Zhang X P, Yang Z C. Robust flutter analysis based on genetic algorithm. Sci China Tech Sci, 2012, 55: 2474–2481
30 Shapley L S. Stochastic games. P Natl Acad Sci USA, 1953, 39: 1095–1100

31 Altman E, Hayel Y. Markov decision evolutionary games. IEEE Trans Autom Control, 2010, 55: 1560–1569
32 Altman E, Hayel Y. A stochastic evolutionary game approach to energy management in a distributed aloha network. In: Proceedings of the 27th Conference on Computer Communications, Phoenix, Arizona, 2008. 2432–2440
33 Azuaje F. A computational evolutionary approach to evolving game strategy and cooperation. IEEE Trans Syst Man Cybern Part B Cybern, 2003, 33: 498–503
34 Putro U S, Kijima K, Takahashi S. Adaptive learning of hypergame situations using a genetic algorithm. IEEE Trans Syst Man Cybern Part A Syst Humans, 2000, 30: 562–572
35 Brede M. Short versus long term benefits and the evolution of cooperation in the prisoner's dilemma game. PLoS ONE, 2013, 8: e56016
36 Duan H B, Sun C H. Swarm intelligence inspired shills and the evolution of cooperation. Sci Rep, 2014, 4, 5210
37 Wang Q, Wang Y Z. Cluster synchronization of a class of multi-agent systems with a bipartite graph topology. Sci China Inf Sci, 2014, 57: 1–11
38 Mu Y F, Guo L. How cooperation arises from rational players? Sci China Inf Sci, 2013, 56: 1–9
39 Zhang Y H, Jiao X H, Li L, et al. A hybrid dynamic programming-rule based algorithm for real-time energy optimization of plug-in hybrid electric bus. Sci China Tech Sci, 2014, 57: 2542–2550
40 Chu F, Wang F L, Wang X G, et al. A model for parameter estimation of multistage centrifugal compressor and compressor performance analysis using genetic algorithm. Sci China Tech Sci, 2012, 55: 3163–3175
41 Gu Y S, Zhang X P, Yang Z C. Robust flutter analysis based on genetic algorithm. Sci China Tech Sci, 2012, 55: 2474–2481
42 Zhao W Z, Wang C Y. Mixed H_2/H_∞ road feel control of EPS based on genetic algorithm. Sci China Tech Sci, 2012, 55: 72–80
43 Luo Q N, Duan H B. An improved artificial physics approach to multiple UAVs/UGVs heterogeneous coordination. Sci China Tech Sci, 2013, 56: 2473–2479