International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 1633–1641
DOI: https://doi.org/10.2991/ijcis.d.210520.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

The Value Function with Regret Minimization Algorithm for Solving the Nash Equilibrium of Multi-Agent Stochastic Game

Luping Liu, Wensheng Jia*

College of Mathematics and Statistics, Guizhou University, Guizhou, Guiyang, 550025, China

ARTICLE INFO

Article History
Received 02 Feb 2021
Accepted 13 May 2021

Keywords
Regret minimization
Multi-agent
Stochastic game
Nash equilibrium
Spatial prisoner's dilemma

ABSTRACT

In this paper, we study the value function with regret minimization algorithm for solving the Nash equilibrium of multi-agent stochastic game (MASG). To begin with, the idea of regret minimization is introduced to the value function, and the value function with regret minimization algorithm is designed. Furthermore, we analyze the effect of discount factor to the expected payoff. Finally, the single-agent stochastic game and spatial prisoner's dilemma (SDP) are investigated in order to support the theoretical results. The simulation results show that when the temptation parameter is small, the cooperation strategy is dominant; when the temptation parameter is large, the defection strategy is dominant. Therefore, we improve the level of cooperation between agents by setting appropriate temptation parameters.

1. INTRODUCTION

Fudenberg and Levine [1] put forward the game learning theory, the goals of the bounded rationality agents are to maximum their long-term payoff and regret minimization by constantly adjusting their strategies from their known information. Recently, many scholars focused on the study Nash equilibrium (NE) of multi-agent stochastic game (MASG). Yang and Wang [2] researched the multi-agent reinforcement learning (MARL) from the perspective of game theory. Rubinstein [3] investigated that bounded rational participants continuously modified their cognition in repeated games in which compared current strategies with previous ones to make optimal strategy choices. Asienkiewicz and Balbus [4] presented the existence analysis of NE for random games under certain conditions. Watkins [5,6] first proposed the Q-learning method, and proved the convergence of Q-learning. Littman [7] proposed a minimax Q-learning for two-person zero-sum stochastic game. Shoham *et al.* [8] discussed *Nash-Q* learning in general-sum stochastic game. Therefore, Q-learning and various improved learning algorithms play an important role in the implementation of NE for the MASG.

Reinforcement learning [9] solved the MASG by interacting with complex environment and learning from experiences. Bowling and Velson [10,11] proposed a classic method to evaluate MARL algorithm. The MARL has recently been extensively used in wireless sensor networks [12], event-triggered consensus system [13], traffic signal controllers [14], numerical algorithm [15], comparative analysis [16], integrodifferential algebraic [17], computational algorithm [18], and other fields. But the majority MARL algorithms either lack a rigorous convergence guarantee [19], potentially converge only under strong assumptions such as the existence of an unique NE [20,21], or provably non-convergent in all cases [22]. Zinkevich [23] identified the nonconvergent behavior of the value-function method in general-sum stochastic game. Minagawa [24] considered a sufficient condition for the uniqueness of NE in strategic-form game. However, Hansen *et al.* [25] proposed the concept of no regret to measure convergence, which came up with a new criteria to evaluate convergence in zero-sum self-plays [26,27]. Regret minimization has been used in a variety of games in recent years [28]. Inspired by research works mentioned above, we mainly studies the value function with regret minimization algorithm for solving the NE of MASG. The central idea of regret minimization is that the agent obtained a payoff after the agent has taken an action in the learning process, agents can retrospect the history of actions and payoff taken so far, and the agent regret not having taken another action, namely, the best action in hindsight. The agents' goal is to minimize the cumulative regret, written as the sum $\sum_{t=1}^{T}(V^*(s,a) - V^t(s,a))$ of the difference between the values of V at the action a at time t and the true optimum of V^* of the action a. Different from [29] in which considered the expected average time payoff and limited space for states/actions, this paper considered the expected sum of discount payoff in an unlimited time range, which

*Corresponding author. Email: wsjia@gzu.edu.cn

means regret minimization can be regard as discounted expected payoff optimization criterion. In this paper, the idea of regret minimization is introduced to the value function, and the value function with regret minimization algorithm is designed. Furthermore, we analyze the effect of discount factor to the expected payoff. Finally, the single-agent stochastic game (SASG) and spatial prisoner's dilemma (SDP) are investigated in order to support the theoretical results. The simulation results show that when the temptation parameter is small, the Cooperation strategy is dominant; when the temptation parameter is large, the defection strategy is dominant. Therefore, we improve the level of cooperation between agents by setting appropriate temptation parameters.

The remainder of this paper is structured as follows: in Section 2, we introduce the model of MASG, and analyze the discount factor to the influence of discounted expected payoff. In Section 3, the idea of regret minimization is introduced to the value function, and the value function with regret minimization algorithm is designed. In Section 4, a simple stochastic game and the SDP game are investigated in order to support the theoretical results. Finally, we present some brief summaries.

2. PROBLEM DESCRIPTION AND PREREQUISITES

2.1. The Model of Multi-Agent Stochastic Game

A framework of MASG is given as follows [7]:

Let $N = \{1, \dots, n\}$ denote the set of all agents, a MASG is a tuple $< N, \mathcal{S}, \mathcal{A}_i, \mathcal{D}, R_i, \gamma >$, where,

- N is the number of agents;
- $\mathcal{S}$ is the set of states;
- $\{\mathcal{A}_i\}_{i\in N}$ is the set of action for the i-th agent, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the joint action set of all agents;
- $\mathcal{D}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ $(\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \sum_{s'\in\mathcal{S}} \mathcal{D}(s'|s, a) = 1)$ is a state transition probability function, and s' represents the possible state at the next moment;
- $R_i : S \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, $i \in N$ is the payoff function of the i-th agent, giving the expected payoff received by the agent under joint actions in each state;
- $\gamma \in (0, 1)$ denotes the discount factor. When $\gamma \to 0$, the agent is regarded as myopic, which means that the agent is only worried about immediate payoff. When $\gamma \to 1$, the agent is known as farsighted, which means that the agent more interested about future payoff.

In infinite-horizon process [9], the agents' discounted payoff from time step t to horizon is,

$$R_i^t = r_i^{t+1} + \gamma r_i^{t+2} + \gamma^2 r_i^{t+3} + \cdots = \sum_{l=1}^{\infty} \gamma^{l-1} r_i^{t+l}. \tag{1}$$

The model of MASG as shown in Figure 1.

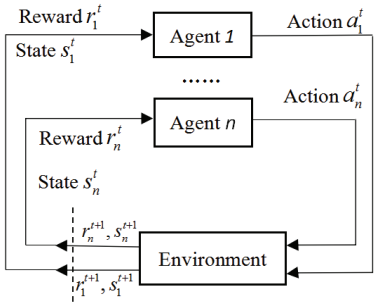


Figure 1 The interaction of multi-agent and environment.

$\pi_i : \mathcal{S} \to \mathcal{A}_i$ denotes the strategy of agent i. Let $\pi = (\pi_1, \dots, \pi_n)$ be all agents' joint strategy, the value function V_i^π defines the long-term cumulative payoff of agent i in any state s at time t, taking action a under the joint strategy π as follows:

$$V_i^t(s, a) = \sum_{a\in\mathcal{A}} \pi_i(s, a) \cdot \left(R_i^t(s'|s, a) + \gamma \sum_{s'\in S} \mathcal{D}(s'|s, a) V_i^\pi(s') \right) , \forall s \in S, t \in \{1, 2, \dots, T\}, \tag{2}$$

where T denotes terminate time, i.e. horizon. Equation (2) is referred to as Bellman updated equation of V^π for agent i, and records the payoff value by obtaining on the Markov chain $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_t, \mathcal{S}_{t+1}, \dots$ with the state s as the initial state. The item $R_i^t(s'|s, a^t) + \gamma V_i^\pi(s')$ denotes staring from the state s, taking action a_t at time t, the agent i's payoff value obtained by 1-step transition to s' and plus the discounted expected payoff collected from the state s'. To solve the MASG, Equation (2) has rewritten as an iterative formula of the dynamic programming equations as follows:

$$V^{k+1}(s, a) = R(s'|s, a) + \gamma \sum_{s'\in S} \mathcal{D}(s'|s, a) V^k(s', a). \tag{3}$$

The optimal value function of agent i is defined by

$$V_i^*(s, a) = \max_{a\in\mathcal{A}} \left(R_i(s'|s, a) + \gamma \sum_{s'\in S} \mathcal{D}(s'|s, a) V_i^*(s', a) \right), \tag{4}$$

as rational agents, they attempt to find the best response policy in favor of all their states.

Definition 2.1. (Nash equilibrium NE of the MASG) Let $< N, \mathcal{S}, \mathcal{A}_i, \mathcal{D}, R_i, \gamma >$ be the MASG, if a policy $\pi^* = (\pi_1^*, \dots, \pi_i^*, \dots, \pi_n^*)^{\mathrm{T}}$ is a NE, $\pi_{-i} \triangleq (\pi_1, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_n)^{\mathrm{T}}$, then $\forall a_i \in \mathcal{A}_i (i \in N)$, $\forall s \in \mathcal{S}$, the following inequality holds:

$$V_i(s, \pi_i^*, \pi_{-i}^*) \geq V_i(s, \pi_i, \pi_{-i}^*), \forall \pi_i \in \Pi_i,$$

where Π_i is the strategy space of agent i, $V_i(s, \pi_i^*, \pi_{-i}^*)$ denotes the discount accumulation payoff. π^* is the NE of the MASG such that each individual strategy π_i^* is a best response to others. The NE of the MASG describes each agent maximize own discounted expected payoff, and no agent can obtain higher benefit by unilaterally changing its strategy as long as all other agents keep their strategies invariant.

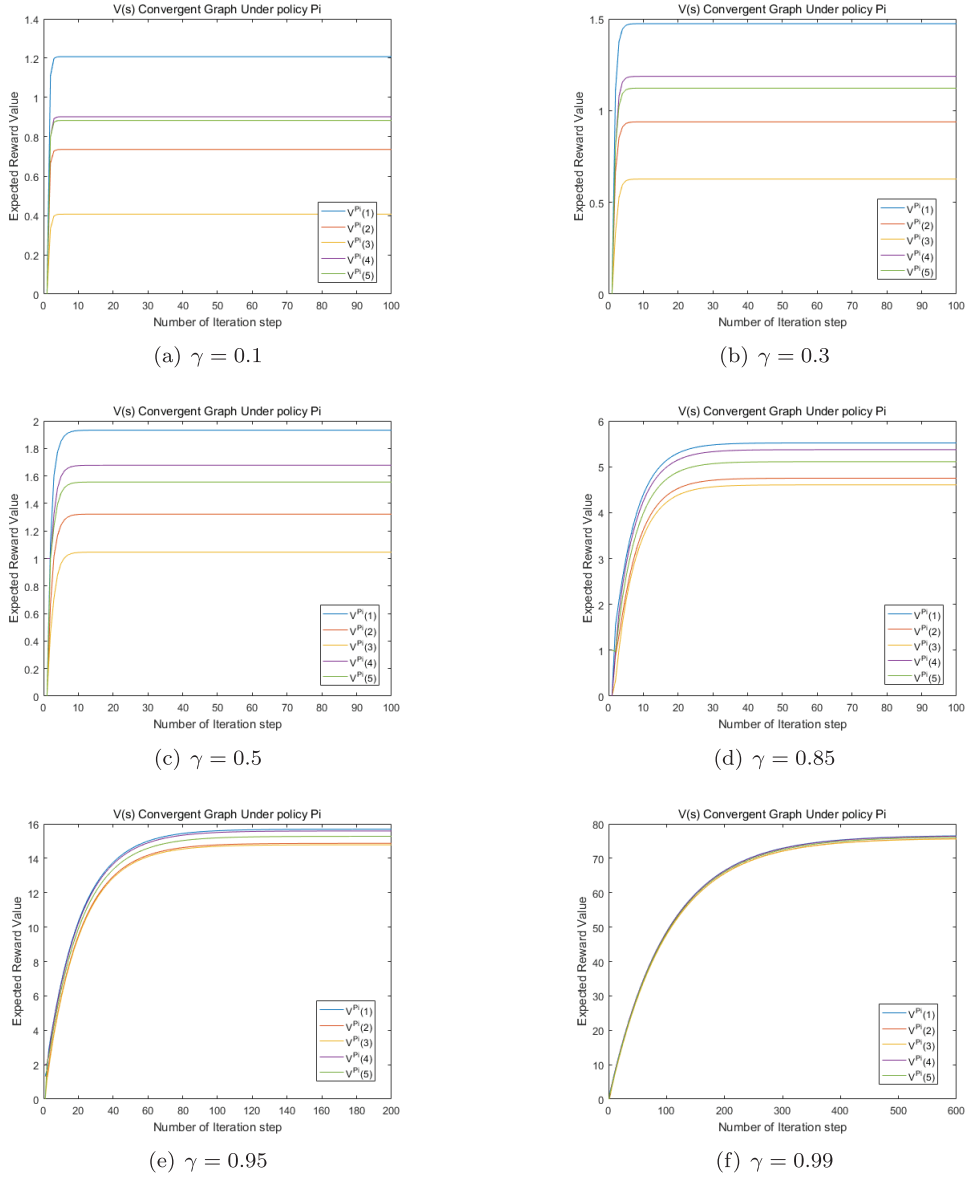
2.2. The Analysis of Discount Factor

For a Markov decision process (MDP) system, we consider the discount factor to the influence of the expected payoff in MASG. Now we make a simple experiment about the MASG $< N, \mathcal{S}, \mathcal{A}_i, \mathcal{D}, R, \gamma >$, where different discount factor γ, the transition probability function $\mathcal{D}$ and payoff function R are as follows:

$$\mathcal{D} = \begin{pmatrix} 2/3 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1/3 & 2/3 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \\ 0 & 1/4 & 1/2 & 1/4 & 0 \end{pmatrix}, R = \begin{pmatrix} 2 & 1 & 0 & 0 & 3 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 \\ 2 & 0 & 0 & 2 & 1 \\ 0 & 2 & 0 & 2 & 1 \end{pmatrix}.$$

Meanwhile, $V^0 = [0, 0, 0, 0, 0]^{\mathrm{T}}$ denotes initial value function vector, where T represents the transposed operator, the value function's convergence property under different discount factor is shown Figure 2.

According to Figure 2(a–f), we can observe that starting from V_0, the value function V finally converges with the number of iteration step, and the optimal value function is unique. Through this experiment, we know that the value function is sensitive to the value of discount factor. When $\gamma \to 0$, the agent is myopic, the expected payoff is small. When $\gamma \to 1$, the agent is farsighted, the expected payoff is large. Consequently, it is easy to know that myopic agents only care about immediate benefits, and hyperopic agents are more likely to obtain higher benefits in the future. Another point that needs to be



(a) $\gamma = 0.1$ (b) $\gamma = 0.3$

(c) $\gamma = 0.5$ (d) $\gamma = 0.85$

(e) $\gamma = 0.95$ (f) $\gamma = 0.99$

Figure 2 The diagram of the convergence of the value function under different discount factor.

explained is that the expected payoff value will not converge when $\gamma \geq 1$.

3. THE VALUE FUNCTION WITH REGRET MINIMIZATION ALGORITHM

3.1. The Cumulative Regret Minimization

Assuming that the finite-horizon, a policy π is obliged to approach to the optimal strategy at any iteration. ϑ_t^k is equivalent to the difference between V^*, (4) and the value function of V^π, (2) in time step.

$$\vartheta_t^k = V^*(s,a) - V^t(s,a), \tag{5}$$

where ϑ_t^k denotes regret degree under adopting strategy π in state s. Our goal is to minimize the agents' accumulative regret,

$$\Theta_{\mathrm{T}}^K = \sum_{t<\mathrm{T}} \vartheta_t^{\mathrm{K}}, \tag{6}$$

where K denotes the terminal time step.

In Bubeck [30], the formula (6) is the normalized object. The loss ϑ_t^k is defined as follows [31],

$$\vartheta_{\mathrm{t}}^K = \min_{k<\mathrm{K}} \vartheta_t^{\mathrm{k}}. \tag{7}$$

We define the cumulative regret minimization of all agents in the MASG as follows:

$$\Theta_{\mathrm{T}}^K = \sum_{t<\mathrm{T}} \sum_{k<\mathrm{K}} \vartheta_t^{\mathrm{k}}, \tag{8}$$

where Θ_{T}^K denotes an upper bound of the agent, which means minimization gap between the strategic value and the optimal strategy value.

3.2. Arithmetic Flow of the Value Function with Regret Minimization Algorithm

The implementation steps of the value function with regret minimization algorithm are as follows:

step 1: Initialization parameters. Some strategies are randomly generated, and set the discounted factor, the value of cumulative regret degree is 0.0001.

step 2: Each agent's the value function $V^t(s,a)$ is calculated by Equation (2), and the optimal strategy $V^*(s,a)$ is calculated by formula (4).

step 3: The cumulative regret degree $\Theta_{\mathrm{T}}^K = \sum_{t<\mathrm{T}} \sum_{k<\mathrm{K}} \vartheta_t^{\mathrm{k}}$ is calculated by Equation (8).

step 4: Stopping condition of iterations: does the cumulative regret degree satisfy $\Theta_{\mathrm{T}}^K < 0.0001$ for all agents? If yes, we output the optimal strategy π^*; otherwise, we return **step 1**.

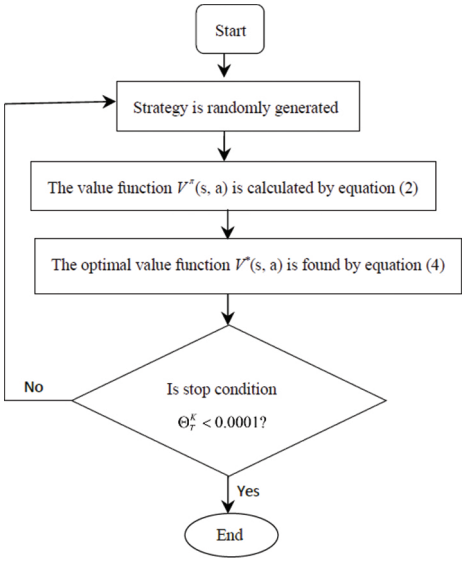


Figure 3 Flow chart of value function under regret minimization algorithm.

Once the optimal strategy is obtained by satisfying the cumulative regret, the value of the discounted expected payoff is defined. The value function under the regret minimization algorithm see Figure 3.

4. THE NEOF A MASG

4.1. A Simple Stochastic Game

The aim of the agent is to maximize their long term discounted expected payoff making respond to others agents. We give an example of a SASG as follows:

Example 1.

Let $N = 1$, the agent's state is $\mathcal{S} = \{s_1, s_2, s_3\}$. In state s_1 and s_2, we select an action from the agent's action sets $\mathcal{A}(s_1) = \mathcal{A}(s_2) = \{a_1, a_2\}$; in state s_3, we choose action from $\mathcal{A}(s_3) = \{a_2\}$. If we select action a_1 in state s_1, then the payoff is $R(s_1, a_1) = 2$, and move state s_2 with probability 1. If we choose action a_2 in state s_1, then the payoff is $R(s_1, a_2) = 3$, and remain in state s_1 with probability 1. In state s_2, if we select a_1, then we receive $R(s_2, a_1) = 5$, and move state s_1 with probability 1, whereas the payoff choosing action a_2 devotes $R(s_2, a_2) = 10$ and we shift to state s_3 with probability 0.5 and reserve in state s_2 with probability 0.5. If we can only select a_2 in state s_3, which means $R(s_3, a_2) = 0$ and we remain in state s_3 with probability 1. Assume that the agent has enough farsighted and the discount factor is $\gamma = 0.9$ by the analysis of the discount factor in Section 2.2.

The above description can be shown in Table 1. We suppose that the horizon is T and the final payoff is $r_T(s), \forall s \in \mathcal{S}$.

Assume that the decision will be made at $t = 0, 1, 2$, i.e. $T = 3$. Moreover, $R_3(s_1) = R_3(s_2) = R_3(s_3) = 0$. In state s_3, $\pi^{t,*}(s_3) = a_2$

Table 1 The SASG was described in Example 1.

Agent(N)	1					
State($\mathcal{S}$)	s_1		s_2		s_3	
Action($\mathcal{A}$)	a_1	a_2	a_1	a_2	–	a_2
Payoff(R)	2	3	5	10	–	0
Transition probability($\mathcal{D}$)	(0.0, 1.0, 0.0)	(1.0, 0.0, 0.0)	(1.0, 0.0, 0.0)	(0.0, 0.5, 0.5)	–	(0.0, 0.0, 1.0)

and $V^{t,*}(s_3) = 0, \forall t$. In terminal time, the agent's payoff $R^T(s_1) = 0$ and $V^3(s_1) = V^3(s_2) = 0$.

By the backward induction method, at time $t = 2$ in state s_1, we have

$$V^2(s_1, a_1) = 2 + V^3(s_2) = 2,$$
$$V^2(s_1, a_2) = 3 + V^3(s_1) = 3.$$

So $V^{2,*}(s_1) = 3$ and the optimal strategy $\pi^*(s_1, 2) = a_2$. In state s_2, we have

$$V^2(s_2, a_1) = 5 + V^3(s_1) = 5,$$
$$V^2(s_2, a_2) = 10 + 0.9V^3(s_2) + 0.9V^{3,*}(s_3) = 10.$$

So $V^{2,*}(s_2) = 10$ and the optimal strategy $\pi^*(s_2, 2) = a_2$.

At time $t = 1$ in state s_1, we have

$$V^1(s_1, a_1) = 2 + V^{2,*}(s_2) = 12,$$
$$V^1(s_1, a_2) = 3 + V^{2,*}(s_1) = 6,$$

So $V^{1,*}(s_1) = 12$ and the optimal strategy $\pi^*(s_1, 1) = a_1$. In state s_2, we have

$$V^1(s_2, a_1) = 5 + V^{2,*}(s_1) = 8,$$
$$V^1(s_2, a_2) = 10 + 0.9V^{2,*}(s_2) + 0.9V^{2,*}(s_3) = 19.$$

So $V^{1,*}(s_2) = 19$ and the optimal strategy $\pi^*(s_2, 2) = a_2$.

At time $t = 0$ in state s_1, we have

$$V^0(s_1, a_1) = 2 + V^{1,*}(s_2) = 21,$$
$$V^0(s_1, a_2) = 3 + V^{1,*}(s_1) = 15.$$

So $V^{0,*}(s_1) = 21$ and the optimal strategy $\pi^*(s_1, 0) = a_1$. In state s_2, we have

$$V^0(s_2, a_1) = 5 + V^{1,*}(s_1) = 17,$$
$$V^0(s_2, a_2) = 10 + 0.9V^{1,*}(s_2) + 0.9V^{1,*}(s_3) = 27.1.$$

So $V^{0,*}(s_2) = 27.1$ and the optimal strategy $\pi^*(s_2, 0) = a_2$.

Therefore, the optimal value function and the optimal strategy in any state as follows,

$$V^* = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{c} \begin{array}{ccc} t=0 & t=1 & t=2 \end{array} \\ \begin{pmatrix} 21 & 12 & 3 \\ 27.1 & 19 & 10 \\ 0 & 0 & 0 \end{pmatrix} \end{array}, \quad \pi^* = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{c} \begin{array}{ccc} t=0 & t=1 & t=2 \end{array} \\ \begin{pmatrix} a_1 & a_1 & a_2 \\ a_2 & a_2 & a_2 \\ a_2 & a_2 & a_2 \end{pmatrix} \end{array},$$

where the payoff is $V^*(s_1) = 17$ in state s_1, $V^*(s_2) = 27.1$ in state s_2 or $V^*(s_3) = 0$ in state s_3. In decision horizon, the NE of the SASG is (a_1, a_1, a_2) in state s_1, the NE of the SASG is (a_2, a_2, a_2) in state s_2, the NE of the SASG is (a_2, a_2, a_2) in state s_3. Therefore, the agent's learning behavior be able to convergence to the NE of the SASG, and the agent is no-regret under fixed discount factor.

4.2. The Spatial Prisoners' Dilemma

The SPD [32] can be regarded as a two-agent two-action stochastic game $< N, \mathcal{S}, \mathcal{A}_i, \mathcal{D}, R_i, \gamma >$, where $N = 2$, agents' state set $\mathcal{S}$ corresponds to different temptation factors, agents' action set is $\mathcal{A}_i = \{C, D\}$ $(i = 1, 2)$. When all agents fixed strategy, we can obtain one of the four possible payoff: R(Payoff), S(Sucker), T(Temptation), and P(Penalty). In the multi-agent setting, if all agents select Cooperation (C), then they receive R(Payoff); if all agents choose Defection (D), then they obtain P(Penalty); if some agents select Cooperation (C) and some Defection (D), cooperators obtain Sucker (S) and defectors gain Temptation (T). The four payoff value of SPD satisfy the inequalities: $T > R > P > S$ and $2R > T + S$.

Example 2.

Let $N = \{1, 2\}$ be the set of two agents, agents' state sets are $\mathcal{S} = \{s_1, s_2\}$, $b(b > 1)$ denotes the temptation parameter. In state s_1 and s_2, we can choose an action from $\mathcal{A}(s_1) = \mathcal{A}(s_2) = \{C, D\}$. In state s_1, the immediate payoff of agent 1 is $R(s_1, C, C) = 1$, $R(s_1, C, D) = 0$, $R(s_1, D, C) = b$, and $R(s_1, D, D) = 0$, the immediate payoff of agent 2 is $R(s_1, C, C) = 1$, $R(s_1, C, D) = 0$, $R(s_1, D, C) = b$, and $R(s_1, D, D) = 0$. If the agent choose the action pair $\{C, D\}$, $\{D, C\}$, $\{D, D\}$ in state s_1, then the agent will move to state s_2 with probability 1; if the agent select the action pair $\{C, C\}$ in state s_1, then the agent remain in state s_1 with probability 1. In state s_2, the immediate payoff of agent 1 is $R(s_2, C, C) = 1$, $R(s_2, C, D) = 0$, $R(s_2, D, C) = b$, and $R(s_2, D, D) = 0$, the immediate payoff of agent 2 is $R(s_2, C, C) = 1$, $R(s_2, C, D) = 0$, $R(s_2, D, C) = b$, and $R(s_2, D, D) = 0$. Once the the agent reach state s_2, the agent remain in state s_2 with probability 1. In some cases, a finite-horizon problem for the SPD must be improved by identifying the horizon T and the terminal payoff $R_T(s), \forall s \in \mathcal{S}$. The above description is represented by Table 2.

The game starts in state s_2, the NE of the agent is (D, D). The MASG game is the prisoners' dilemma, and no agent can obtain higher benefit by unilaterally changing its strategy as long as all other agents keep their strategies invariant. The discount factor can be analyzed in state s_2, we obtain,

$$V_i^*(s_2) = \frac{1}{1-\gamma}.$$

Table 2 The description of the MASG, where state s_1 (left) and state s_2 (right).

		Agent 2 C	Agent 2 D
Agent 1	C	1,1 / (1,0)	0,b / (0,1)
	D	b,0 / (0,1)	0,0 / (0,1)

State s_1

		Agent 2 C	Agent 2 D
Agent 1	C	1,1 / (0,1)	0,b / (0,1)
	D	b,0 / (0,1)	0,0 / (0,1)

State s_2

In state s_1, the MASG game be expressed as follows:

		Agent 1 C	Agent 1 D
Agent 2	C	$1+\gamma V_1^*(s_1)$, $1+\gamma V_2^*(s_1)$	$\gamma V_1^*(s_2)$, $b+\gamma V_2^*(s_2)$
	D	$b+\gamma V_1^*(s_2)$, $\gamma V_2^*(s_2)$	$\gamma V_1^*(s_2)$, $\gamma V_2^*(s_2)$

or

		Agent 1 C	Agent 1 D
Agent 2	C	$1+\gamma V_1^*(s_1)$, $1+\gamma V_2^*(s_1)$	$\frac{\gamma}{1-\gamma}$, $b+\frac{\gamma}{1-\gamma}$
	D	$b+\frac{\gamma}{1-\gamma}$, $\frac{\gamma}{1-\gamma}$	$\frac{1}{1-\gamma}$, $\frac{1}{1-\gamma}$

where $b(b>1)$ represents the temptation factor by using Defection strategy for agents.

Evidently, (D,C) and (C,D) aren't an equilibrium of the MASG game by virtue of agents are motivated to change their strategies, and (D,D) is the NE for all values of γ. the pair of actions (C,C) is the NE, if we have,

$$1+\gamma V_i(s_1) \geq b+\frac{\gamma}{1-\gamma}$$
$$\Rightarrow V_i(s_1) \geq \frac{(2-b)\gamma+b-1}{\gamma(1-\gamma)}, (i=1,2).$$

Assume that both of agents select Cooperation in state s_1, then

$$V_i(s_1) = 1+\gamma\pi_i(s_1)$$
$$\Rightarrow V_i(s_1) = \frac{1}{1-\gamma}.$$

Meanwhile,

$$\frac{1}{1-\gamma} \geq \frac{(2-b)\gamma+b-1}{\gamma(1-\gamma)}$$
$$\Leftrightarrow (\gamma-1)[(1-b)\gamma-(1-b)] \geq 0,$$
$$\Leftrightarrow \gamma \geq 1.$$

On the one hand, (D,D) is a NE of the MASG, but (C,C) isn't a NE due to $\gamma \geq 1$. On the other hand, $\gamma \geq 1$ is out of the range of discount factor, so the SDP not converge to the NE. $\gamma \geq 1$ shows that the agent does not converge to the strategy (C,C) [32,33] in the classic prisoner's dilemma game. Thus, the payoff value of SDP is independent of the discount factor. and we consider the influence of different temptation factors in order to raise the level of Cooperation all agents.

Therefore, we simulate the SDP on a 300×300 grid with an even 50-50 split between cooperators and defectors randomly distributed on the grid and the simulation is for 300 generations. Assume that the temptation parameter set as $b=1.1$, $b=1.3$, $b=1.5$, $b=1.7$, $b=1.8$, $b=1.9$, and the temptation parameter is independent of the discount factor γ. In the graph of the final state the cooperators are expressed by *yellow*, the defectors are indicated by *blue*, cooperators to defectors (C to D) are represented by *red*, defectors to cooperators (D to C) are *green*.

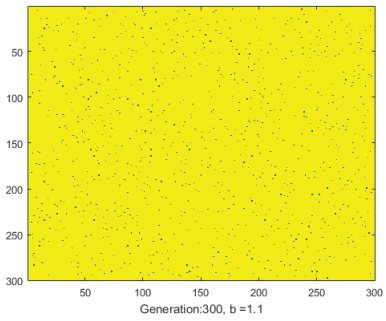
From Figure 4(a–f) we can see the number of cooperators suddenly dropped owing to agents being isolated and surrounded by defectors, hence a few cooperators that survive will create small clusters and then rise in numbers, so the cooperators quickly become dominant over the defectors after a few generations when the temptation factor b is lower.

Up until $b=1.7$ the cooperators are dominant but the defectors are outnumbering the cooperators when the temptation becomes continually higher. This means that there exist some transition point between $b=1.7$ and $b=1.9$ when the defectors overtake the cooperators. Now we can investigate how the parameter b influences the model by looking at the density of cooperators each round and over time, see how the model changes behavior for different values of b, especially in the region $1.7<b<1.9$.

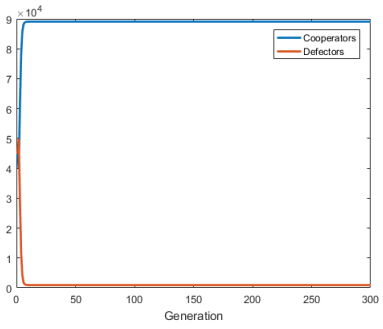
According to Figure 5(a–d), when $b \geq 1.8$ the numbers of the defectors become further rising, then the defectors are dominant. The game selects different temptation factor corresponds to different states, the adoption strategy of the agent is closely related to the temptation factor. When $1<b<1.8$, the agent's cooperation strategy is dominant, the level of cooperation is higher. When $b>1.8$, the agent's defection strategy is dominant, and the level of defection is higher with b is bigger. Obviously, we can improve the level of cooperation between agents by setting appropriate temptation parameter $1<b<1.8$ and $b \to 1$.

5. CONCLUSION

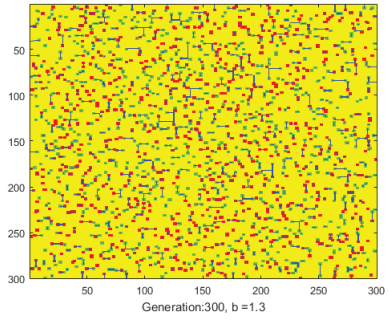
In this paper, we give a new attempt to solve NE of MASG by using the value function with regret minimization algorithm. We consider the expected payoff as an optimization criterion between agents. To begin with, the idea of regret minimization is introduced to the



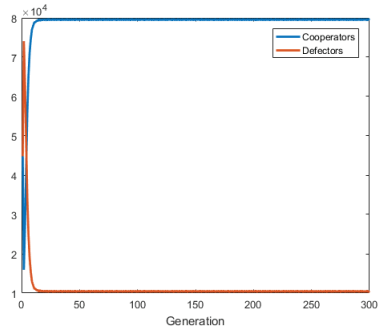
(a) Final state for $b = 1.1$



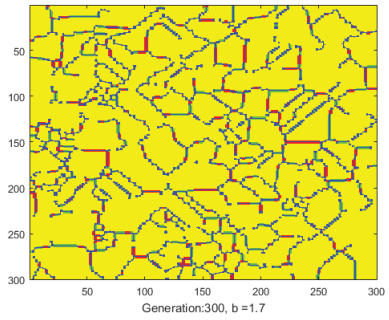
(b) The agent changes over time for $b = 1.1$



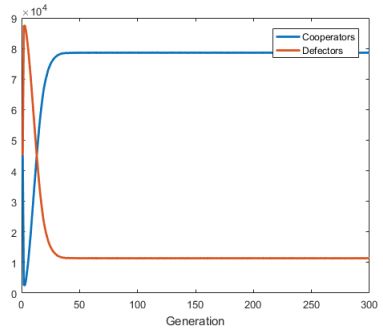
(c) Final state for $b = 1.3$



(d) The agent changes over time for $b = 1.3$



(e) Final state for $b = 1.7$



(f) The agent changes over time for $b = 1.7$

Figure 4 The agent strategy change with different $b = 1.1$, $b = 1.3$, $b = 1.7$ (left); the numbers of cooperators and defectors (right).

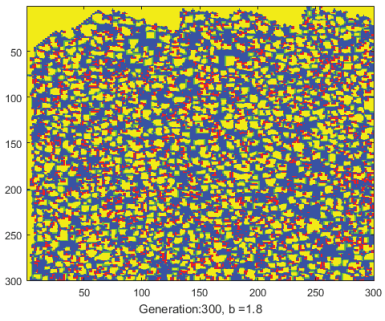
value function, and the value function with regret minimization algorithm is designed. Furthermore, we analyze the effect of discount factor to the discounted expected payoff. Finally, the simulation results show that when the temptation parameter is small, the cooperation strategy is dominant; when the temptation parameter is large, the defection strategy is dominant, we improve the level of cooperation between agents by setting appropriate temptation parameters $1 < b < 1.8$ and $b \to 1$. Hence, the value function with regret minimization algorithm is an effective way to solve the NE of the stochastic game. We are also interested in further research to explore whether the value function with regret minimization algorithm can be used to solve more complexity stochastic game for the large-scale action set or continuous action space.

CONFLICTS OF INTEREST

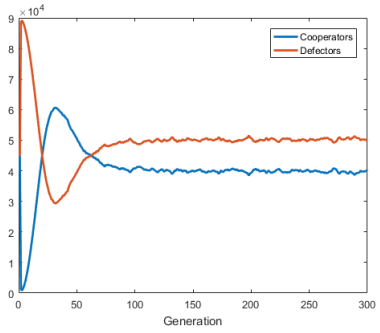
The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

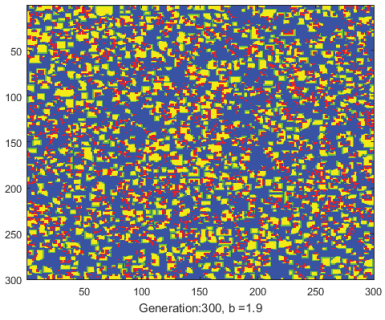
All authors read and approved the final manuscript



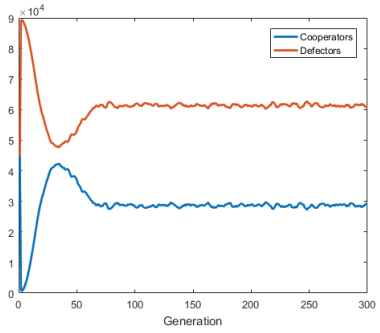
(a) Final state for $b = 1.8$



(b) The agent changes over time for $b = 1.8$



(c) Final state for $b = 1.9$



(d) The agent changes over time for $b = 1.9$

Figure 5 The agent strategy change with different b (left); the numbers of cooperators and defectors (right).

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. [12061020], [71961003]), the Science and Technology Foundation of Guizhou Province (Grant No.20201Y284, 20205016, 2021088), the Foundation of Guizhou University (Grant No. [201405], [201811]). The authors acknowledge these supports.

REFERENCES

[1] D. Fudenberg, D.K. Levine, The Theory of Learning in Games, Cambridge University Press, 1988.
[2] Y.D. Yang, J. Wang, An overview of multi-agent reinforcement learning from game theoretical perspective, Multiagent Syst. arXiv: 2011.00583, (2021).
[3] A. Rubinstein, Modeling Bounded Rationality, MIT Press, Cambridge, MA, USA, 1997.
[4] H. Asienkiewicz, L. Balbus, Existence of Nash equilibria in stochastic games of resource extraction with risk-sensitive players, Top. 11 (2019), 502–518.
[5] C.J.C.H. Watkins, Learning from Delayed Rewards, PhD Thesis, University of Cambridge, Cambridge, England, 1989.
[6] C.J.C.H. Watkins, Q-learning, Mach. Learn. 8 (1989), 279–292.
[7] M.L. Littman, Markov games as a framework for multi-agent reinforcement learning, in 11th International Conference on Machine Learning, New Brunswick, NJ, USA, 1994, pp. 157–163.
[8] Y. Shoham, R. Powers, T. Grenager, Multi-agent Reinforcement Learning: a Critical Survey, Technical Report, Stanford University, 2003. Web Manuscript, https://www.cc.gatech.edu/classes/AY2008/cs7641_spring/handouts MALearning_ACriticalSurvey_2003_0516.pdf
[9] R. Sutton, A. Barto, Reinforcement Learning: an Introduction, MIT Press, Cambridge, MA, USA, 1998.
[10] M. Bowling, M. Veloso, Rational and convergent learning in stochastic games, in International Joint Conference on Artificial Intelligence, 2001, vol. 17, pp. 1021–1026. Web Manuscript, http://www.cs.cmu.edu/~mmv/papers/01ijcai-mike.pdf
[11] M. Bowling, M. Veloso, Multiagent learning using a variable learning rate, Artif. Intell. 136 (2002), 215–250.
[12] A.P. Renold, S. Chandrakala, MRL-SCSO: multi-agent reinforcement learning-based self-configuration and self-optimization protocol for unattended wireless sensor networks, Wireless Personal Commun. 96 (2017), 5061–5079.
[13] Y.L. Cui, M.R. Fei, D.J. Du, *et al.*, Event-triggered consensus of multi-agent systems with data transmission delays and random packet dropouts, Control Theory Appl. 32 (2015), 1208–1218.
[14] S. Tantawy, B. Abdulhai, H. Abdelgawad, Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto, IEEE Trans. Intell. Transp. Syst. 14 (2013), 1140–1150.
[15] O.A. Arqub, A.H. Zaer, Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm, Inf. Sci. 279 (2014), 396–415.
[16] O.A. Arqub, Numerical algorithm for the solutions of Fractional order systems of Dirichlet function types with comparative analysis, Fundamenta Informaticae. 166 (2019), 111–137.

[17] O.A. Arqub, R. Hasan, The RKHS method for numerical treatment for integrodifferential algebraic systems of temporal two-point BVPs, Neural Comput. Appl. 30 (2018), 2595–2606.
[18] O.A. Arqub, Computational algorithm for solving singular Fredholm time-fractional partial integrodifferential equations with error estimates, J. Appl. Math. Comput. 59 (2019), 227–243.
[19] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: a selective overview of theories and algorithms, Mach. Learn, arXiv: 1911.10635, (2019).
[20] J. Hu, M.P. Wellman, Nash Q-learning for general-sum stochastic games, J. Mach. Learn. Res. 4 (2003), 1039–1069.
[21] M.L. Littman, Value-function reinforcement learning in Markov games, Cogn. Syst. Res. 2 (2001), 55–66.
[22] E. Mazumdar, L.J. Ratliff, S. Sastry, *et al.*, Policy gradient in linear quadratic dynamic games has no convergence guarantees, in Smooth Games Optimization and Machine Learning Workshop, Bridging Game, arXiv: 1907.03712, 2019.
[23] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 2003, pp. 928–936. Web Manuscript, http://www.stanford.edu/class/cs369/files/Zinkevich-GradDescent-ICML03.pdf
[24] J. Minagawa, On the uniqueness of Nash equilibrium in strategic-form games, J. Dyn. Games. 7 (2020), 97–104.
[25] N. Hansen, S.D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evol. Comput. 11 (2003), 1–18.
[26] M. Bowling, Convergence and no-regret in multiagent learning, in Advances in Neural Information Processing Systems, 2005, pp. 209–216.
[27] M. Zinkevich, M. Johanson, M. Bowling, *et al.*, Regret minimization in games with incomplete information, in Advances in Neural Information Processing Systems, 2008, pp. 729–1736.
[28] Y. Zhang, T. Chen, S. Chang, Existence of solution to n-person non-cooperative games and minimax regret equilibria with set payoffs, Appl. Anal. (2020), 1–16.
[29] K. Zhang, Z. Yang, H. Liu, *et al.*, Fully decentralized multi-agent reinforcement learning with networked agents, Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, arXiv: 1802.08757, 2018, pp. 5872–5881.
[30] S. Bubeck, N. Cesabianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Found. Trends Mach. Learn. 5 (2012), 1–122.
[31] S. Bubeck, R. Munos, G. Stoltz, Pure exploration in multi-armed Bandits problems, in: R. Gavaldà, G. Lugosi, T. Zeugmann, S. Zilles (Eds.), Algorithmic Learning Theory, Springer, Berlin, Heidelberg, Germany, 2009, pp. 23–37.
[32] R. Axelrod, W.D. Hamilton, The evolution of cooperation, Science. 211 (1981), 1390–1396.
[33] R. Axelrod, Effective choice in the prisoner's dilemma, J. Conflict Resolut. 24 (1980), 3–25.