Operations Research and Fuzziology 运筹与模糊学, 2022, 12(2), 312-321
Published Online May 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.122032

Hans汉斯

合作的多智能体强化学习算法

秦前伟，邓喜才

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2022年4月11日；录用日期：2022年5月6日；发布日期：2022年5月11日

摘　要

在多智能体的环境中，智能体的学习行为是一个有价值的研究内容。从系统设计者的角度来看，在同时存在多个智能体的环境中，能够让智能体朝着共同利益的最大化方向调整自己的行为策略，这是值得研究的。本文将提出一种合作的梯度算法(CL-WoLF-IGA)，目的是让智能体朝着使得共同收益最大的策略学习。同时，为了让算法适用于马尔可夫博弈，我们放宽条件，提出CL-WoLF-PHC强化学习算法。该算法在只知道平均共同收益的未知环境中，也能够让使用算法的智能体最终达成能够使共同收益最大化的策略。同时，为了验证算法在实际博弈模型中的表现，我们用经典的博弈模型进行检验CL-WoLF-IGA算法。仿真结果表明，算法具有良好的收敛性。

关键词

多智能体强化学习，博弈论，合作学习

The Reinforcement Learning Algorithm for Cooperative Multi-Agent

Qianwei Qin, Xicai Deng

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Apr. 11th, 2022; accepted: May 6th, 2022; published: May 11th, 2022

Abstract

In the multi-agent environment, the learning behavior of agents is a valuable research content. From the perspective of system designer, it is worth studying that in an environment where multiple agents exist simultaneously, agents can adjust their behavior strategies in the direction of maximizing common interests. In this paper, a cooperative gradient algorithm (CL-Wolf-IGA) is proposed to make the agent learn towards the strategy that maximizes the common benefit. Mean-

文章引用: 秦前伟, 邓喜才. 合作的多智能体强化学习算法[J]. 运筹与模糊学, 2022, 12(2): 312-321.
DOI: 10.12677/orf.2022.122032

while, in order to make the algorithm suitable for Markov games, we relax the conditions and propose CL-Wolf-PHC reinforcement learning algorithm. Even in the unknown environment where only the average common benefit is known, the algorithm can make the agent using the algorithm finally reach the strategy that can maximize the common benefit. At the same time, in order to verify the performance of the algorithm in the actual game model, we use a classical game model to test the CL-Wolf-PHC algorithm. Simulation results show that the algorithm has good convergence.

Keywords

Multi-Agent Reinforcement Learning, Game Theory, Cooperative Learning

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

强化学习又称再励学习、评价学习或增强学习，是一种重要的机器学习方法，用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。随着强化学习的发展，近些年在许多领域都取得了令人瞩目的成果，并且考虑到在现实场景中通常会同时存在多个决策个体，部分研究者逐渐将眼光从单智能体领域延伸到多智能体，即多智能体强化学习(MARL)。

尽管强化学习在许多情况下在单智能体系统上取得了成功，但将强化学习扩展到多智能体系统时，情况将会变得复杂，这是由于其他智能体的存在导致环境不稳定，当一个代理选择一个动作时，该智能体的收益不经受到环境的影响，同时收益也受到其他智能体的影响。此时，对于智能体来说环境将不再是平稳的。比如见表 1 在囚徒困境博弈中，玩家 1 选择行为 C 的情况下，另外一个代理采取行为 C 或者行为 D 会导致不一样的联合收益结果。因此，理性的智能体仅仅追求自己的利益，为了避免被利用，最终将选择纳什均衡策略。

以往，对于多智能体的强化学习算法形成了一定的研究基础，研究人员提出了许多学习算法[1] [2]，来促进智能体在多智能体系统中收敛到纳什均衡策略。然而，在现实生活中，大多数情况下我们希望设计的智能体能够更好的相互合作，而不是仅仅追求个人的利益。因此，为了促进智能体之间的相互合作，本文结合以往的多智能体研究，先提出合作的梯度学习算法 CL-WoLF-IGA。考虑到现实场景中的很多情况下我们并不知道博弈的支付函数和对手的策略，我们放宽条件，提出了更为实用的多智能体强化学习算法 CL-WoLF-PHC。最后，我们将用实验表明在马尔可夫博弈模型中，使用合作算法的智能体将追求共同的最大利益，最终达成最大化共同收益的合作策略。

Table 1. Prisoners' dilemma
表 1. 囚徒困境

玩家 1 的收益		玩家 1 的行为	
玩家 2 的收益		C	D
玩家 1 的行为	C	3，3	0，5
	D	5，0	1，1

本文余下部分如下安排。第二部分是国内外相关的一些研究工作。第三部分将介绍马尔可夫博弈模型和强化学习算法。第四部介绍合作的赢或快速学习算法。第五部分将算法应用于马尔可夫博弈模型中，验证算法的有效性。最后，在第六部分我们将对文章进行总结并指出未来的工作。

2. 相关工作

基于梯度上升多智能体强化学习算法是Singh S.等人于2000年首先提出的无穷小梯度算法(IGA [3])，此算法能够使每个学习者朝着其预期收益的梯度方向更新其策略。IGA 算法的目的是在两人两行为标准型博弈中，使代理收敛到一个特定的纳什均衡。此后，M. Zinkevich 提出了一种称为广义无穷小梯度(GIGA [4])的算法，它将 IGA 算法中只有两行为的情况拓展到了任意多个行为。IGA 算法和 GIGA 算法都可以和 Win or Learn Fast (WoLF)结合以提高智能体在随机博弈中的表现(Wolf-IGA [5], Wolf-GIGA [6])，WoLF-IGA 和 Wolf-GIGA 要求至少指定一个纳什均衡报酬，并且不观察报酬反馈，而是获得报酬的梯度，在文献[5]中还将 WoLF 与 PHC [7]结合形成 WoLF-PHC 算法，WoLF-PHC 算法在实际应用中取得了非常好的效果，并且能够收敛到最优策略。

以上多智能体算法都是促使智能体足够理性，目的是为了使代理的策略收敛到纳什均衡策略，防止被其他智能体利用，而没能都让智能体之间达成共同合作。在文献[8]中 S. Phon-Amnuaisuk 提出一种 CDM-SARSA 算法，在二维空间的多智能协调游戏中表现出来良好的协调性。J.W. Crandall 于 2010 年提出 M-Qubed [9]多智能强化学习算法，M-Qubed 通过在谨慎策略和最佳回应策略之间设置偏差，保证收益不低于最大最小值，同时尽可能协调其他智能体达到更好的结果。Zhang 等人于 2019 年提出 SA-IGA 算法[10]，该算法不但对环境进行学习，也使智能体之间相互的学习，最终收敛到稳定的 Nash 均衡点或者社会最优策略。

大部分的研究都是如何使智能体追求最大的个人收益，如果智能体都是追求个人利益的最大化，最终可能会因对手的利用而选择纳什均衡策略。但在许多情况下，我们往往不需要智能体之间最求个体的最大收益，而是希望彼此合作，向着共同的目标学习，实现总体收益的最大化。

3. 预备知识

在本节中，我们将介绍本文所需的背景。首先，我们概述了相关博弈论的定义。然后简要回顾 Q-learning [11]强化学习算法和赢或快速学习算法。

3.1. 马尔可夫博弈模型

博弈论提供了一个建模智能体交互的框架，以前的研究人员使用该框架来分析 MARL 算法的收敛性。博弈以一种简洁的方式指定了一个代理的报酬依赖于其和其他代理的共同行为。目前常用的博弈模型有矩阵博弈和马尔可夫博弈，分别适用于单状态博弈和多状态博弈。

矩阵博弈通过元组$\langle N, A_1, \cdots, A_N, R_1, \cdots, R_N \rangle$来定义，其中 N 表示博弈中玩家的数量，A_i 是代理人 i 的行动集，且 $R_i : A_1 \times \cdots \times A_N \to \mathbb{R}$。是代理人 i 的报酬，其定义为所有代理人执行的联合行动的函数。如果游戏只有两个代理，那么可以方便地将它们的奖励函数定义为一个支付矩阵，可以如下定义：

$$R_i = \left\{ r_i^{jk} \right\}_{|A_1| \times |A_2|}$$

其中 $i \in \{1,2\}$， $j \in A_j$， $k \in A_k$。矩阵中的每个元素 r_i^{jk} 表示如果代理 i 执行动作 j，而其对手执行动作 k 时代理 i 收到的回报。

马尔可夫博弈是标准型博弈和马尔可夫决策过程(多状态)的推广，用 5 元组$\langle S, N, A_i, T, R_i \rangle$定义，其

中 S 表示状态集，N 表示代理人数，A_i 表示代理 $i \in N$ 的行为集，$T: S \times A \times S \to [0,1]$，状态转移概率函数，表示在状态 s 采取行动 a 后过渡到状态 s' 的概率，$R_i: S \times A \to \mathbb{R}$，表示代理 i 这为在 s 状态下采取行动并过渡到 s' 提供了预期的回报，这里的 $A = A_1 \times \cdots \times A_N$，$A_i$ 表示代理 i 的行为集。

矩阵博弈中代理人 i 的策略用 $\pi_i: A_i \to [0,1]$ 表示，其将行为映射为概率。根据策略 π_i 选择动作 k 的概率为 $\pi_i(k)$，如果执行一个行为的概率为 1，而执行其他操作的概率为 0，则策略是纯策略，否则，该策略是混合的。所有代理的联合策略是单个代理策略的集合，用 $\pi = \langle \pi_1, \cdots, \pi_N \rangle$ 定义，可以简写 $\pi = \langle \pi_i, \pi_{-i} \rangle$，这里 π_{-i} 表示除了代理 i 外其代理的策略。

$$R_i = \begin{bmatrix} r_i^{11} & r_i^{12} \\ r_i^{21} & r_i^{22} \end{bmatrix},\ i \in \{1,2\} \tag{1}$$

在如(1)所示的矩阵博弈中，$R_i(a_i, a_{-i}) = r_i^{a_i a_{-i}}$ 表示代理 i 的收益，在马尔可夫博弈中，$R_i = \sum_{k=0} \gamma^k r_{i,k+1}$ 表示代理 i 在一幕中获得的累积长期收益，其中，$r_{i,k+1}$ 是在 $k+1$ 时收到的即时奖励，γ 是折扣因子。

矩阵博弈和马尔可夫博弈的每个代理的目标都是找到一个策略，使参与者的期望收益最大化。理想情况下，我们希望所有代理都能达到使其个人收益最大化的均衡。然而，当代理都是理性的情况下，即每个代理都是自私的，都不想被对方利用。此时代理将寻找的目标纳什均衡策略，此时将是所有代理的局部收益，没有一个玩家可以通过单方面改变其当前策略来获得更好的预期回报。我们将联合策略的纳什均衡定义如下：

如果对于 $\forall i \in N$，$\forall \pi_i$，有 $V_i(\pi_i^*, \pi_{-i}^*) \ge V_i(\pi_i, \pi_{-i}^*)$，我们称联合策略 $\pi^* = (\pi_i^*, \pi_{-i}^*)$ 为纳什均衡策略。如果构成纳什均衡策略的所有策略都是纯的，则纳什均衡策略是纯的。否则，纳什均衡策略称为混合的。任何博弈至少有一个纳什均衡，但可能没有任何纯均衡。同时，我们用 $\overline{r}$ 表示共同收益的平均值，π_i^{MM} 表示在对手合作的情况下，i 使得平均共同收益最大的策略，其如下定义

$$\pi_i^{MM} = \arg\max_{\pi_i} \max_{a_{-i}} \overline{r_i}(\pi_i, a_{-i}) \tag{2}$$

3.2. Q-Learning 强化学习算法

在强化学习算法中，Q-learning [11]算法因其简单性和鲁棒性，因此被作为单智能体框架中最常用的算法之一，这也是为什么它是最早应用于多智能体环境的 RL 算法之一(Tan, 1993)。Q-learning 是同轨策略(on-policy)下的时序差分(TD [12])控制算法，它的提出是强化学习早期的一个重要突破。Q-learning 学习的过程如下：在某个状态 s 下，智能体选择一个动作 a 执行，然后根据智能体所收到的关于该动作的奖赏值和当前的状态动作值的估计来对动作的结果进行评估。对所有状态下的所有行为进行这样的重复，智能体通过对长期的折扣回报的判断，就可以学习总体上的最优行为。每个智能体的 Q 值函数的迭代公式如下：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \beta\left(r_t + \gamma \max_{a \in A} Q(s_{t+1}, a) - Q(s_t, a_t)\right) \tag{3}$$

其中，参数 $\alpha \in (0,1)$ 为学习率(或学习步长)；$\gamma \in (0,1)$ 为折扣率；$Q(s_t, a_t)$ 是状态动作对的值函数，表示智能体在环境状态 s 下执行动作 a。

强化学习已经是一个在单智能体框架中进行学习的成熟而深刻的理论框架。单个智能体在不确定的环境中运行，必须学会自主行动并实现特定的目标。在这种情况下，已经证明，只要智能体所经历的环境是马尔可夫的[1]，并且智能体可以尝试足够多的操作，强化学习就可以保证收敛到最优策略。尽管这种方法取得了成功，当环境中同时存在多个代理同时学习并可能相互交互时，任务就变得更加复杂，智

能体的益不仅取决与自己的行为，同时也受到其他智能体的行为影响，在以往关于多智能体强化学习的算法中证明了，在两人两行为的博弈中，如果两人都追求个人的最大收益，最终会趋近于纳什均衡，但在许多情况下，纳什均衡解可能是不可取的。我们更希望智能体之间能够相互合作，通过不断的对未知环境进行学习，达成实现共同收益最大的策略。

4. 合作学习的赢或快速学习算法

在实际生活中许多情况下，比如无人驾驶，我们往往不需要智能体仅仅追求自身的最大收益，而是追求有利于社会的最大共同收益，换句话说，我们要求智能体以最大的共同收益为学习目标，而不是以自身的利益为目标。于是，我们结合 WoLF-IGA [5]算法，我们提出合作形式下的梯度学习算法，即合作学习赢或快速学习(CL-WoLF-IGA)。其原理是朝着共同收益的梯度方向学习，当收益大于以往平均策略带来的收益时，策略的学习速率变慢，反之，策略的学习速度变大。对于智能体 i，$i\in\{1,2\}$，其学习模型如下所示

$$V_i(\pi)=\frac{1}{N}\sum_j E_\pi\{R_j\} \tag{4}$$

$$\Delta\pi_i^{(t+1)}\leftarrow \eta l_{i,t}\frac{\partial V_i(\pi^t)}{\partial \pi_i} \tag{5}$$

$$\pi_i^{t+1}\leftarrow \Pi_{[0,1]}\left(\pi_i^{(t)}+\Delta\pi_i^{(t+1)}\right) \tag{6}$$

其中控制学习率变化的参数为 $l_{i,t}$，$\eta(\eta>0)$ 表示步长。$\Pi_{[0,1]}$ 是将输入值映射到有效概率范围 $[0,1]$ 的投影函数，用于防止梯度将策略移出有效概率空间。对于标量 x，有

$$\Pi_{[0,1]}(x)=\arg\min_{z\in[0,1]}|x-z| \tag{7}$$

在(1)所示两人两行为的矩阵博弈中，我们用 (p_1^t,p_2^t) 分别表示两人经过 t 次迭代选择第一个行为的概略，用 $\pi_i=(p_i,1-p_i),i\in\{1,2\}$ 表示第 i 个玩家的策略。则两个人在第 $t+1$ 次的策略为

$$\Delta p_i^{(t+1)}=\eta l_{i,t}\left(\frac{u_i+u_{-i}}{2}p_{-i}^t+\frac{c_i+c_{-i}}{2}\right) \tag{8}$$

$$p_i^{(t+1)}=\Pi_{[0,1]}\left(p_i^{(t)}+\Delta p_i^{(t+1)}\right) \tag{9}$$

其中，

$$l_{i,t}=\begin{cases} l_{\min}, & \text{if } V_i(\pi_1,\pi_2)>V_i(\pi_i^{MM},\pi_2) \\ l_{\max} & \text{其它}\end{cases},\ i\in\{1,2\}\text{且}\ 0<l_{\min}<l_{\max} \tag{10}$$

$$u_i=r_i^{11}+r_i^{22}-r_i^{12}-r_i^{21} \tag{11}$$

$$c_i=r_i^{12}-r_i^{22} \tag{12}$$

CL-WoLF-IGA 算法需要知道每个代理都需要知道其他代理的策略、支付函数以及最大值的策略，这在重复博弈的开始之前通常不知道的，因此，我们结合 PHC [11]，提出策略爬山的合作学习的赢或快速学习(CL-WoLF-PHC)，伪代码如下：

DOI: 10.12677/orf.2022.122032

每个玩家 i 的合作学习算法：

1：设 $\alpha \in (0,1)$，$\delta_l > \delta_w \in (0,1)$.

初始化 $Q(s,a) \leftarrow 0$，$\pi(s,a) \leftarrow \frac{1}{|A_i|}$，$C(s) \leftarrow 0$

2：无限循环：

在状态 s 下根据混合策略 $\pi(s,a)$ 选择动作 a (带有一定的探索)。

观察奖励共同收益的平均值 $\overline{r}_t$ 和下一状态 s'：

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \beta\left(\overline{r}_t + \gamma \max_{a\in A} Q(s_{t+1},a) - Q(s_t,a_t)\right)$$

更新平均策略 $\overline{\pi}$：

$$C(s) \leftarrow C(s) + 1$$

$$\forall a' \in A_i \quad \overline{\pi}(s,a') \leftarrow \overline{\pi}(s,a') + \frac{1}{C(s)}\left(\pi(s,a) - \overline{\pi}(s,a')\right)$$

更新策略：

$$\delta = \begin{cases} \delta_w & \text{if } \sum_{a'} \pi(s,a')Q(s,a') > \sum_{a'} \overline{\pi}(s,a')Q(s,a') \\ \delta_l & \text{其它} \end{cases}$$

$$\delta_{sa} = \min\left(\pi(s,a), \frac{\delta}{|A_i|-1}\right)$$

其中 $\Delta_{sa} = \begin{cases} -\delta_{sa} & \text{if } a \neq \operatorname{argmax}_{a'} Q(s,a) \\ \sum_{a'\neq a} \delta_{sa} & \text{其它} \end{cases}$

$$\pi(s,a) \leftarrow \pi(s,a) + \Delta_{sa}$$

因为 CL-WoLF-IGA 要求的条件过于严苛，而且只适用于矩阵博弈，我们放宽其条件，提出更加实用的 CL-WoLF-PHC。其不仅能实用于矩阵博弈，也能适用于马尔可夫博弈模型，而且此算法仅仅只需要知道博弈过程中所有智能体收益的平均值，就能让使用此算法的智能体收敛到能带来最大共同收益的策略。所以，CL-WoLF-PHC 比 CL-WoLF-IGA 要求的条件更为宽松。

5. 仿真与结果分析

本节内容，我们将在两人两行为的重复博弈下进行数值实验，由于 CL-WoLF-PHC 是 CL-WoLF-IGA 放宽条件下的得到的，所以我们只需要观察使用 CL-WoLF-PHC 的算法是否最终收敛到最大化共同收益的联合策略。图 1~4 分别是使用算法的智能体在囚徒困境、协调博弈、猎鹿问题和性别大战中进行自博弈得到的学习轨迹图。

图 1 是两个智能体在囚徒困境(见表 1)的合作博弈，其中横坐标是重复博弈的次数，纵坐标表示选择第一个行为的概率，我们用蓝色和红色分别表示两人第一个行为的选择概率变化情况。通过实验表明，在囚徒困境博弈中，当博弈重复约 28,000 次时，两智能体都将达成以 1 的概率选择 C，因为(C,C)能够使共同利益最大。

图 2 也是智能体在囚徒困境博弈中的表现，只不过我们将横纵坐标分别表示两个智能体第一个行为的概率，设用 (p_1, p_2) 表示两个人第一个行为的概率，则 $\pi_i = (p_i, 1-p_i), i \in \{1,2\}$ 为第 i 个玩家的策略。同

时，我们随机初始化 16 个策略。其仿真结果表明，无论初始策略如何，随着博弈的重复进行，智能体选择第一个行为的概率越来越大，并最终都以 1 的概率选择第一个行为，即达成联合行为(C,C)。因为在囚徒困境博弈中，(C,C)能使共同收益最大。

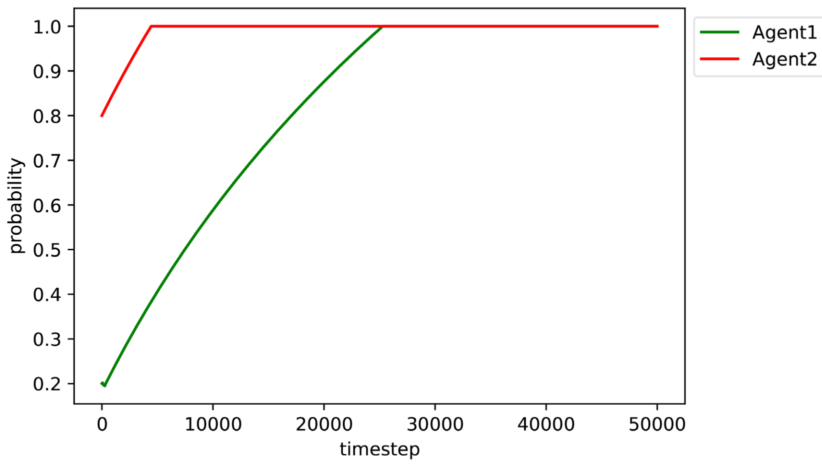

Figure 1. The learning dynamics of two agents in prisoner's dilemma, the horizontal axis represents the number of times the game is played, and the vertical axis represents the probability of the first action of each agent

图 1. 囚徒困境中两个智能体的学习动态，横坐标表示博弈进行的次数，纵坐标表示每个智能体的第一个行为的概率

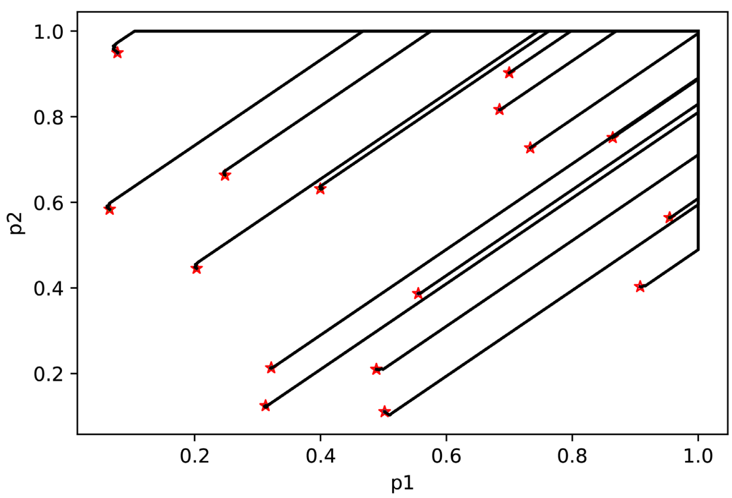

Figure 2. The trajectory of the learning strategies of two agents in the prisoners' dilemma, the horizontal axis is the probability of the first action of the first agent, the vertical axis is the probability of the first action of the second agent, and we randomly select 16 initial policy points

图 2. 两个智能体在囚徒困境中的学习策略轨迹，横坐标是第一个智能体第一个行为的概率，纵坐标第二个智能体第一个行为的概率，并随机初始 16 个策略点

图 3 是两智能体在协调博弈(见表 2)中随机初始 16 个策略的学习轨迹图，由于在协调博弈中联合行为(C,C)和(D,D)都能够使共同收益最大，所以智能体最终达成(C,C)或(D,D)的联合行为，即会收敛到(1,1)或者(0,0)这两点。

Table 2. Coordination game
表 2. 协调博弈

玩家 1 的收益		玩家 1 的行为	
玩家 2 的收益		C	D
玩家 1 的行为	C	2，1	0，0
	D	0，0	1，2

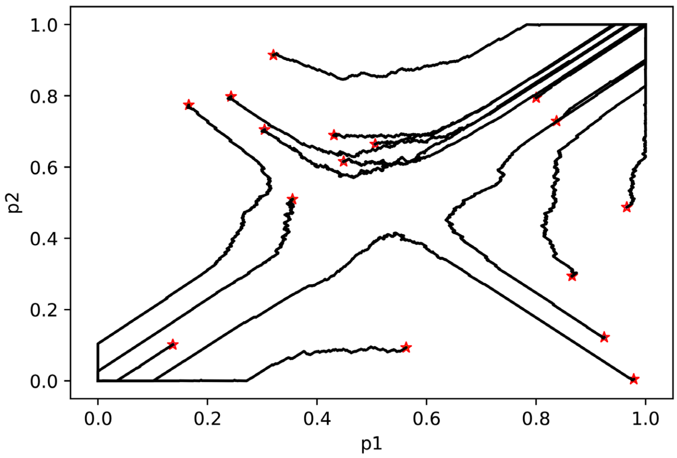

Figure 3. The trajectory of learning strategy of two agents in the coordination game. Horizontal axis and vertical axis respectively represent the probability of the first behavior of the two agents, and we randomly select 16 initial policy points
图 3. 两个智能体在协调博弈中的学习策略轨迹，横纵坐标分别表示两个智能体第一个行为的概率，并随机初始 16 个策略点

图 4 是两智能体在猎鹿问题(见表 3)中随机初始 16 个策略的学习轨迹图，在猎鹿问题中，当只有当两个智能都选者策略 C 时，共同收益才能最大。而实验结果也验证两个智能体通过不断的学习，最后都会以 1 的概率选择第一个行为，即达成联合行为(C,C)。

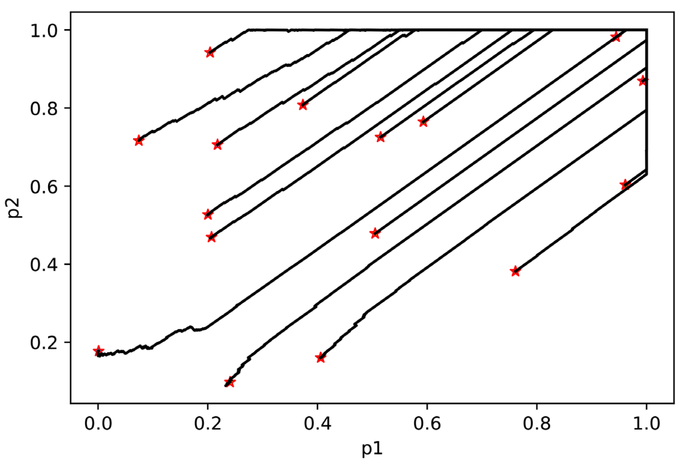

Figure 4. The trajectory of learning strategy of the two agents at randomly initial 16 policy points in Stag hunt. Horizontal axis and vertical axis respectively represent the probability of the first behavior of the two agents
图 4. 两个智能体在猎鹿问题中随机初始 16 个策略点得到的学习策略轨迹，横纵坐标分别表示两个智能体第一个行为的概率

Table 3. Stag hunt
表 3. 猎鹿博弈

玩家 1 的收益		玩家 1 的行为	
玩家 2 的收益		C	D
玩家 1 的行为	C	10，10	0，4
	D	4，0	4，4

如同协调博弈一样，在性别战中(见表 4)，由于联合行为(C,C)和(D,D)都能够使共同收益最大，而图 5 也表明，在只知道平均共同收益的情况下，智能体通过不断的学习，所以最终都会收敛到(1,1)或者(0,0)这两点，即两智能体都将会以 1 的概率选择 C，或者都以 1 的概率选择 D。

Table 4. Battle of the sexes
表 4. 性别战

玩家 1 的收益		玩家 1 的行为	
玩家 2 的收益		C	D
玩家 1 的行为	C	3，2	1，1
	D	0，0	2，3

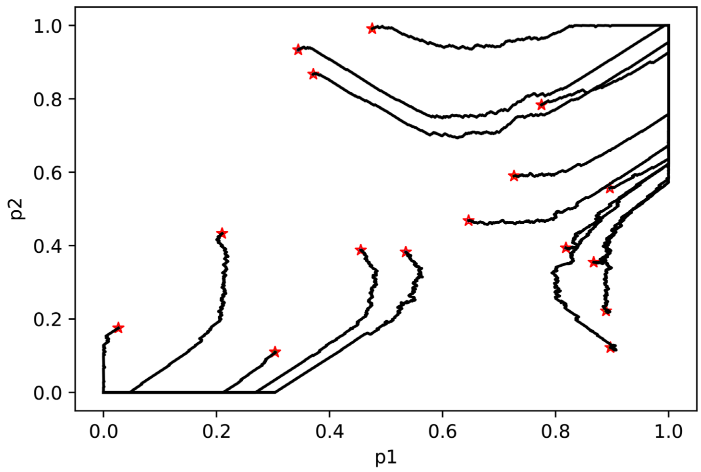

Figure 5. The trajectory of learning strategy of the two agents at randomly initial 16 policy points in Battle of the sexes. Horizontal axis and vertical axis respectively represent the probability of the first behavior of the two agents
图 5. 两个智能体在性别战博弈中随机初始 16 个策略点得到的学习策略轨迹，横纵坐标分别表示两个智能体第一个行为的概率

6. 结束语

本文提出了合作学习的赢或快速学习(CL-WoLF-IGA)梯度算法，通过朝着共同收益的梯度方向更新策略，最终促进智能体达成使共同收益最大化的策略(例如囚徒困境的(C,C))。考虑到算法的实用性，所以我们提出条件更为宽松的 CL-WoLF-PHC 算法，此算法在只知道共同收益的平均值，不需要知道对手的其他信息就能够朝着最大化收益的方向学习。最后，我们在一般和博弈模型中进行仿真实验验证我们的算法，通过在两人两行为的马尔可夫博弈模型中进行实验，并分析使用算法的智能体学习的策略轨迹图，表明了我们的算法具有与理论一致的结果。

参考文献

[1] Littman, M. (1994) Markov Games as a Framework for Multi-Agent Reinforcement Learning. *Machine Learning Proceedings* 1994, New Brunswick, 10-13 July 1994, 1435-1445. https://doi.org/10.1016/B978-1-55860-335-6.50027-1

[2] Hu, J. (1998) Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm. 15*th International Conference on Machine Learning*.

[3] Singh, S. and Kearns, M. (2000) Nash Convergence of Gradient Dynamics in General-Sum Games. *Conference on Uncertainty in Artificial Intelligence*, 541-548.

[4] Zinkevich, M. (2003) Online Convex Programming and Generalized Infinitesimal Gradient Ascent. *Proceedings of the* 20*th International Conference on Machine Learning* (*ICML*-03), 928-936.

[5] Bowling, M. and Veloso, M. (2002) Multiagent Learning Using a Variable Learning Rate. *Artificial Intelligence*, **136**, 215-250. https://doi.org/10.1016/S0004-3702(02)00121-2

[6] Bowling, M. (2005) Convergence and No-Regret in Multiagent Learning. *Advances in Neural Information Processing Systems*, **17**, 209-216.

[7] Abdallah, S. and Lesser, V. (2008) A Multiagent Reinforcement Learning Algorithm with Non-Linear Dynamics. *Journal of Artificial Intelligence Research*, **33**, 521-549. https://doi.org/10.1613/jair.2628

[8] Phon-Amnuaisuk, S. (2009) Learning Cooperative Behaviours in Multiagent Reinforcement Learning. *International Conference on Neural Information Processing*, 570-579. https://doi.org/10.1007/978-3-642-10677-4_65

[9] Crandall, J. and Goodrich, M. (2011) Learning to Compete, Coordinate, and Cooperate in Repeated Games Using Reinforcement Learning. *Machine Learning*, **82**, 281-314. https://doi.org/10.1007/s10994-010-5192-9

[10] Zhang, C. and Hao, J. (2019) SA-IGA: A Multiagent Reinforcement Learning Method towards Socially Optimal Outcomes. *Autonomous Agents and Multi-Agent Systems*, **33**, 403-429. https://doi.org/10.1007/s10458-019-09411-3

[11] Watkins, C. and Dayan, P. (1992) Q-Learning. *Machine Learning*, **8**, 279-292. https://doi.org/10.1023/A:1022676722315

[12] Sutton, R. (1988) Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, **3**, 9-44. https://doi.org/10.1007/BF00115009