Cooperation-Eliciting Prisoner's Dilemma Payoffs for Reinforcement Learning Agents

(Extended Abstract)

Koichi Moriyama
ISIR, Osaka University,
Osaka, Japan
koichi@ai.sanken.
osaka-u.ac.jp

Satoshi Kurihara
Graduate School of
Information Systems,
The University of
Electro-Communications,
Tokyo, Japan

Masayuki Numao
ISIR, Osaka University,
Osaka, Japan

ABSTRACT

This work considers a stateless Q-learning agent in iterated Prisoner's Dilemma (PD). We have already given a condition of PD payoffs and Q-learning parameters that helps stateless Q-learning agents cooperate with each other [2]. That condition, however, has a restrictive premise. This work relaxes the premise and shows a new payoff condition for mutual cooperation. After that, we derive the payoff relations that will elicit mutual cooperation from the new condition.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*intelligent agents, multiagent systems*

Keywords

reinforcement learning; game theory

1. INTRODUCTION

In this paper, we consider a learning agent that chooses appropriate actions in a multiagent environment. In particular, we will discuss what an "independent" reinforcement learning algorithm can do in a multiagent environment.

Prisoner's Dilemma (PD) [1] has the property that both players obtain larger payoffs when they "cooperate" although the (individually) rational action is to "defect", and we humans often "cooperate" with each other. What happens when two independent stateless reinforcement learners play *iterated PD* (IPD)? According to Sandholm and Crites [3], mutual cooperation did not occur.

However, it is not the case in all IPDs. We have already given a condition of PD payoffs and Q-learning parameters that helps stateless Q-learning agents cooperate with each other [2]. That work shows a condition that the Q-value of "cooperation" overcomes that of "defection" after one mutual cooperation occurred by mistakes.

Nevertheless, that condition assumes that the Q-value of "defection" is minimum at the time the mutual cooperation occurs. In this paper, we relax the minimum Q-value premise and show a new condition of PD payoffs where stateless Q-learning agents can cooperate with each other. After that, we derive the payoff relations that will elicit mutual cooperation from the new condition.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. COOPERATION-ELICITING PAYOFFS

In this section, we introduce a new condition relaxing the minimum Q-value premise by considering the "Temptation" payoff of PD. After that, we derive the cooperation-eliciting payoff relations. Let XY be the action pair when the target agent chooses X and the opponent chooses Y, while C and D show "cooperation" and "defection", respectively. Also, let the learning rate α and the discount factor γ be constants in $(0, 1)$, and let $Q(C)$ and $Q(D)$ be the Q-values of C and D, respectively. The payoffs are shown as T, R, P and S when the action pair is DC, CC, DD, and CD, respectively. Note that $T > R > P > S$ in PD.

2.1 Effect of the "Temptation" payoff

Figure 1 shows the movement of Q-values as a schematic view. In this figure, DC happens at time τ, DD continues from $\tau + 1$ to $\tau + l$ ($\equiv t$), and CC happens at $t + 1$. $Q(D)$ becomes highest when T is given by DC. After that, it decreases by P given by DD. If DD continues infinitely, $Q(D)$ returns to the limit as supposed in the previous work. However, if $Q(D)$ does not return to the limit with a paucity of DDs like in the figure, the previous condition is not valid because its premise is unsatisfied.

The following is the new condition that makes $Q(C) \geq Q(D)$ by one mutual cooperation in the case of Figure 1.

THEOREM 1. *Suppose that $Q_{\tau-1}(D) = P/(1-\gamma)$, $Q_{\tau-1}(C) = S + \gamma P/(1-\gamma)$, and, after DC happens at time τ, DD continues from $\tau + 1$ to $\tau + l$ ($\equiv t$). Then, when CC happens at $t + 1$, the condition s.t. $Q_{t+1}(C) \geq Q_{t+1}(D)$ is*

$$\alpha R \geq P - (1-\alpha)S + \alpha(1-\alpha\gamma)\zeta_l(T-P)$$

where $\zeta_l = (1 - \alpha + \alpha\gamma)^l$.

2.2 Worst case analyses

We know that $Q(D)$ has the maximum value $T/(1-\gamma)$ when continuing DC infinitely, whereas $Q(C)$ is already minimum in Theorem 1. Then, we can get the following corollary as the worst case of Theorem 1.

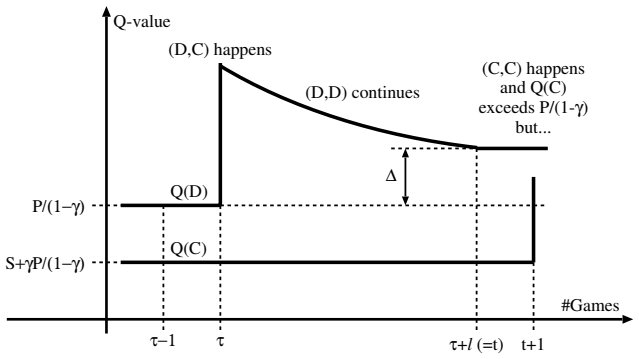

Figure 1: Movement of Q-values

COROLLARY 1. *Suppose that $Q_\tau(D) = T/(1-\gamma)$, $Q_\tau(C) = S + \gamma P/(1-\gamma)$, and DD continues from $\tau+1$ to $\tau+l$ ($\equiv t$). When CC happens at $t+1$, the condition s.t. $Q_{t+1}(C) \geq Q_{t+1}(D)$ is*

$$\alpha R \geq P - (1-\alpha)S + \frac{1-\alpha\gamma}{1-\gamma}\zeta_l(T-P)$$

where $\zeta_l = (1-\alpha+\alpha\gamma)^l$.

Suppose that both players use ε-greedy action selection method. Let ε_s be the target agent's random action probability and ε_o the opponent's one. The probabilities that affect the number of mutual cooperations are (i) the probability that the first CC appears by chance when $Q(C) < Q(D)$ in both agents (p_1), (ii) the probability that the first CC makes $Q(C) \geq Q(D)$ in both agents (p_2), and (iii) the probability that the agents take CC when $Q(C) > Q(D)$ in both agents (p_3). We know that $p_1 = \varepsilon_s/2 \times \varepsilon_o/2$ and $p_3 = (1-\varepsilon_s/2)(1-\varepsilon_o/2)$ from the definition of ε-greedy action selection method. Let us consider p_2 in the following.

From Corollary 1, we can get the condition of l as follows:

$$l \geq \frac{1}{\log(1-\alpha+\alpha\gamma)} \log \frac{(1-\gamma)(\alpha R - P + (1-\alpha)S)}{(1-\alpha\gamma)(T-P)} .$$

Note that it is valid only when $\alpha R > P - (1-\alpha)S$. Let $l_{\min}$ be the right-hand side of this formula. Since l follows a geometric distribution, the probability that $Q(C) \geq Q(D)$ after one CC can be derived from the cumulative distribution function where the interval is not shorter than $m \equiv \max\{\lceil l_{\min} \rceil, 0\}$, *i.e.*, $P(l \geq m) = P(l > m-1) = 1 - P(l \leq m-1) = (1-\varepsilon_s/2)^m(1-\varepsilon_o/2)^m$. Note that it is slightly restrictive because it divides the interval when CD occurs which does not affect $Q(D)$. Let m_s be the interval m of the target agent and m_o that of the opponent; then, $p_2 = (1-\varepsilon_s/2)^{\max\{m_s,m_o\}}(1-\varepsilon_o/2)^{\max\{m_s,m_o\}}$.

Next, let us consider the number of mutual cooperation. There are two directions after the first CC appeared by chance: (i) CC continues for a while because $Q(C)$ becomes larger than $Q(D)$ in both agents by the first CC, or (ii) CD, DC, or DD happens in the next iteration because $Q(C)$ cannot overcome $Q(D)$ in at least one agent by the first CC.

Once $Q(C) \geq Q(D)$ in both agents, CC will continue $p_3/(1-p_3)$ times because the number follows a geometric distribution. Then, n_{cc}, the expected number of mutual cooperation per iteration when $Q(C) < Q(D)$, becomes

$$n_{cc} = p_1(1-p_2) + \frac{p_1 p_2 p_3}{1-p_3} .$$

Table 1: Total numbers and probabilities of mutual cooperations in the experiment

R	$\#CC$	Prob. CC
10	10255	0.0114
50	94843	0.1054
90	608490	0.6761
99	660637	0.7340

Finally, let us investigate how n_{cc} changes when the payoffs change. When both agents have same ε and m, the partial derivative of n_{cc} with respect to m is

$$\frac{\partial n_{cc}}{\partial m} = \left(-p_1 + \frac{p_1 p_3}{1-p_3}\right)\left\{2\left(1-\frac{\varepsilon}{2}\right)^{2m} \cdot \log\left(1-\frac{\varepsilon}{2}\right)\right\} .$$

It is smaller than 0 when $p_3 > 1/2$, *i.e.*, $\varepsilon < 2-\sqrt{2} \simeq 0.5858$. It means that more mutual cooperations will occur as m becomes smaller. Since $m \equiv \max\{\lceil l_{\min} \rceil, 0\}$, let us see the partial derivative of $l_{\min}$ with respect to each payoff: $\partial l_{\min}/\partial T > 0$, $\partial l_{\min}/\partial R < 0$, $\partial l_{\min}/\partial P > 0$, and $\partial l_{\min}/\partial S < 0$. It means that T and P should be small while R and S be large. That is, PD with $T \simeq R \gg P \simeq S$ gives a relatively large n_{cc}, which means, in such PD games, stateless Q-learning agents with ε-greedy action selection method are more likely to cooperate with each other.

3. EXPERIMENT

Let us verify the cooperation-eliciting payoff relations $T \simeq R \gg P \simeq S$ shown in Section 2.2. We used the payoffs $T = 100$, $P = 1$, $S = 0$, and R was set to 10, 50, 90, and 99, respectively. The learning rate $\alpha = 0.25$, the discount factor $\gamma = 0$, and the random action probabilities $\varepsilon = 0.1$. Q was initialized by random real values between $S\,(=0)$ and R.

Table 1 shows the total numbers and probabilities of mutual cooperations for each R. They are of 1000 runs each of which contains 1000 games but first 100 games in each run are excluded. When $R = 99$, it is very surprising that the probability of mutual cooperation was over 0.7. This R also satisfies the PD relations and the IPD rule $T + S < 2R$. From this result, we should know that the probability of mutual cooperation in IPD by stateless Q-learning agents highly depends on the payoff values themselves.

4. CONCLUSION

This work extended the condition in our previous work [2] to handle the "Temptation" payoff. In particular, we considered the case that $Q(D)$ does not decrease enough with a paucity of mutual defections. In the worst case analyses, we derived n_{cc}, the expected number of mutual cooperation per iteration, and the payoff relations that will elicit mutual cooperation while they satisfy PD relations.

5. REFERENCES

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

[2] K. Moriyama. Utility based Q-learning to facilitate cooperation in Prisoner's Dilemma games. *Web Intelligence and Agent Systems*, 7(3):233–242, 2009.

[3] T. W. Sandholm and R. H. Crites. Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *BioSystems*, 37:147–166, 1996.