The resilience of cooperation in a dilemma game played by reinforcement learning agents

Koichi Moriyama, Kaori Nakase, Atsuko Mutoh, and Nobuhiro Inuzuka
Department of Computer Science, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
E-mail: moriyama.koichi@nitech.ac.jp

***Abstract*—This work discusses what an (independent) reinforcement learning agent can do in a multiagent environment. In particular, we consider a stateless Q-learning agent in a Prisoner's Dilemma (PD) game. Although it had been shown in the literature that stateless, independent Q-learning agents had been difficult to cooperate with each other in an iterated PD (IPD) game, we gave a condition of PD payoffs and Q-learning parameters that helps the agents cooperate with each other. Based on the condition, we also discussed the ratio of mutual cooperation happening in IPD games. It supposed that mutual cooperation was fragile, *i.e.*, one misfortune defection would have the agents slide down the spiral of mutual defection. However, it is not always correct. Mutual cooperation will reinforce itself and thus it will be robust and resilient. Hence, this work analytically derives how long a series of mutual cooperation continues once it happened while considering the resilience. It gives us further comprehension of the process of reinforcement learning in IPD games.**

I. Introduction

In this paper, we consider an (artificial) learning agent that chooses appropriate actions in a multiagent environment. The most popular learning method for agents is reinforcement learning [1]. In a reinforcement learning process, an agent is given a reward from the environment according to its action, the state of the environment, etc., and learns to take actions that maximize the rewards.

Many reinforcement learning algorithms for a multiagent environment that explicitly consider the other agents have been proposed (*e.g.*, [2]–[9]). Such an algorithm sees the actions of all agents, called joint actions, and changes its action according to the joint actions. However, the number of the joint actions becomes intractable when the number of agents in the environment and/or that of actions each agent can choose increase. Hence, in this work, we discuss what an "independent" reinforcement learning algorithm, which is for a single-agent environment, can do in a multiagent environment.

Game theory [10], which gives a model of such a multiagent environment, analyzes interactions among rational decision-makers who decide behaviors according to payoffs. *Prisoner's Dilemma* (PD), the most famous two-person two-action game in game theory, has been attracting many researchers for decades [11], [12] because it has an interesting property that both players obtain larger payoffs when they "cooperate" although the (individually) rational action is to "defect". We humans often "cooperate" with each other in this game. In an iterative context, called *iterated PD* (IPD), researchers explain these cooperative behaviors as a result of reciprocity, reputations, etc. that will appear in the following turns [11]. For the reciprocity, reputations, etc., we have to remember who did what in the past.

When players who do *not* remember who did what in the past play an IPD, it may result in mutual defection because it is identical to a one-shot PD. Then, what happens when two independent reinforcement learning agents play an IPD? Sandholm and Crites [13] conducted various experiments in which Q-learning [14] agents played an IPD, and investigated whether mutual cooperation occurred or not. They reported that mutual cooperation did not occur when both agents did not take their past actions into account, *i.e.*, when they were independent and "stateless". On the other hand, Wunder et al. [15] showed that stateless Q-learning with an infinitesimal learning rate might escape from mutual defections in an IPD, owing to initial settings.

Those works tell us that stateless, independent Q-learning agents are difficult to cooperate in *certain* IPDs; however, it is not correct in all IPDs. We gave a necessary condition of the payoffs of PD and parameters of Q-learning that achieves mutual cooperation of stateless Q-learning agents [16], [17]. In other words, there exist IPD games in which stateless Q-learning agents can cooperate with each other. Those works show conditions where the Q-value of cooperation exceeds that of defection after one mutual cooperation occurred by "mistakes". Taking cooperation will be reinforced after the mistaken mutual cooperation, and finally, the agent will continue to choose cooperation "intentionally".

Similarly, such a series of intentional mutual cooperation will be ended after misfortune (unilateral) defections. Here we consider how long the intentional mutual cooperation continues. More precisely, when both agents have Q-values that prefer defection to cooperation, how many times will the intentional mutual cooperation appear during the period from the mistaken mutual cooperation to the next intentional defection? We have discussed the expected duration of a series of mutual cooperation in order to derive the expected ratio of mutual cooperation [17]. As a result, the expected ratio is indeed larger than that calculated from the "mistake" probabilities.

In that discussion, we supposed that an action pair except for mutual cooperation always caused at least one agent to take a series of intentional defection afterward. However, according to numerical simulations, such a misfortune defection did not always cause the intentional defection, because mutual cooperation had reinforced itself and thus became robust and resilient. Hence, in this work, we investigate the duration more precisely; we directly derive the expected duration of mutual cooperation from its mistaken start until the intentional defection comes again.

This paper consists of five sections. In Section II, we briefly introduce a PD game and Q-learning, and review the expected ratio of mutual cooperation of Q-learning agents in an IPD game. In Section III, we derive the expected duration of mutual cooperation. In Section IV, the derived duration is verified by numerical experiments. We conclude this paper in Section V.

II. Preparation

This section introduces a PD game and Q-learning, which are used in the following sections. After that, we review our previous work that derived the expected ratio of mutual cooperation in an IPD game played by two stateless Q-learning agents.

A. Prisoner's Dilemma

A Prisoner's Dilemma (PD) game [11], [12] is a two-person two-action game and is often shown by a payoff matrix (Table I). Each player has two actions: C (cooperation) and D (defection). The players choose actions from rows and columns of the matrix, respectively. After choosing its action, each player obtains a payoff ($T, R, P,$ or S) in the matrix. For example, when the row player chooses C and the column player chooses D, they obtain payoffs S and T, respectively.

TABLE I. Prisoner's Dilemma payoffs

Row \ Column	C	D
C	R, R	S, T
D	T, S	P, P

PD has the following relations among the payoffs:

$$T > R > P > S.$$

Under the relations, each player obtains a larger payoff when he/she chooses D regardless of the opponent's action. As a result, both players choose D and obtain P. However, it is more desirable for them to choose C and obtain R that is larger than P.

B. Q-learning

Suppose that an agent chooses an *action* a_t at time t following a probability distribution called a policy π on the available action set $\mathcal{A}$. A reinforcement learning agent updates its policy to obtain the optimal policy π^*. To evaluate a policy, an *action value function* is defined as an expected return, which is a sum of future rewards discounted by $\gamma \in [0, 1)$ when the agent follows the policy. Q-learning [14] has a function Q and updates it to make it approach the optimal action value function Q^* under π^*. Although Q is a function of states and actions in general, in the following sections, we use a stateless version of Q-learning in which Q has only one argument showing actions. Then, the update rule of stateless Q-learning is

$$Q_{t+1}(a) = \begin{cases} Q_t(a_t) + \alpha\, \delta_t & \text{if } a = a_t, \\ Q_t(a) & \text{otherwise,} \end{cases} \tag{1}$$

$$\delta_t \equiv r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(a') - Q_t(a_t),$$

where $r_{t+1} \in \mathbb{R}$ is a reward obtained at t, $\alpha \in (0, 1]$ is a parameter called the learning rate and δ_t is called TD error that approaches 0 when Q approaches Q^*. Q is proved to converge to Q^* with probability one when several conditions hold [14].

If Q^* is known, the agent can choose an optimal action a^* from Q^* by $a^* = \text{argmax}_{a'' \in \mathcal{A}} Q^*(a'')$. However, if the agent always chooses such actions during learning, Q may converge to a local optimum because the convergence conditions are violated. To avoid it, the agent uses a stochastic method like ε-greedy [1] to choose actions. The ε-greedy method chooses an action having the maximum Q with probability $1 - \varepsilon$, or a random action.

C. Expected ratio of mutual cooperation in an IPD game

Suppose that there are two stateless Q-learning agents that play an IPD game. We have derived the expected ratio of mutual cooperation E_{cc} in the IPD game as follows [17].

$$\begin{aligned} E_{cc} &= p_1(1 - p_2) + p_1 p_2 d, \\ d &= \frac{1 - (1 - p_3)}{1 - p_3} = \frac{p_3}{1 - p_3} \end{aligned} \tag{2}$$

where p_1 is the probability that the first mutual cooperation appears by chance when $Q(D) > Q(C)$ in both agents, p_2 is the probability that the first mutual cooperation makes $Q(C) > Q(D)$ in both agents, and p_3 is the probability that the agents take mutual cooperation when $Q(C) > Q(D)$ in both agents. The expected length d of mutual cooperation while $Q(C) > Q(D)$ in both agents is from the expected number of failures before the first success in the geometric distribution with the success probability $1 - p_3$.[1]

Note that the premise of Eq. 2 is restrictive because it only considers the length of mutual cooperation. One misfortune defection will reset everything, *i.e.*, it will have the agents slide down the spiral of mutual defection. However, the numerical verification in that work showed that such a misfortune defection did not always cause the spiral, *i.e.*, mutual cooperation itself had resilience to some extent. Hence, the derived ratio is underestimated.

III. Expected duration of mutual cooperation

We saw that the expected ratio of mutual cooperation in an IPD game played by two stateless Q-learning agents derived in the previous work was underestimated because it ignored the resilience of mutual cooperation. Hence, in order to incorporate the effect of resilience, here we change the definition of mutual cooperation from the actions the agents take to the *intention* the agents have. That is, we say that mutual cooperation ends when $Q(D)$ exceeds $Q(C)$ in *at least one* agent. Then we derive the expected duration of *the* mutual cooperation instead of d of Eq. 2. More precisely, this work analytically derives the expected number of games from the beginning of a series of the mutual cooperation to its end in an IPD game played by two stateless Q-learning agents with the ε-greedy method. The procedure is as follows. Note that the available action set $\mathcal{A}$ is $\{C, D\}$. We derive the expected duration from an agent's view, *i.e.*, we call one agent the *target* and the other the *opponent*,

[1] Here the "success" means the end of mutual cooperation.

and the parameters used are only those of the target. Hereafter, XY stands for the action pair of both players when the target chooses X and the opponent chooses Y.

1) Derive the change of $Q(D)$ caused by the payoff T, the result of unilateral defection of the target (DC).
2) Derive the number of unilateral defection until the end of the mutual cooperation, and the expected number of games necessary for the derived number of unilateral defection. There are two cases due to the cause of the end:
   a) Ended by a series of DCs, and
   b) Ended by one unilateral defection of the opponent (CD) after a series of DCs.
3) Derive the probabilities p_a and p_b corresponding to the above cases, respectively. Finally, derive the expected duration d until the end of the mutual cooperation.

This work supposes that $Q(D)$ becomes the minimum value $P/(1-\gamma)$ before the first CC, and $Q(C)$ becomes the maximum value $R/(1-\gamma)$ before the first DC after a series of CCs. Also, it supposes that the opponent does not choose D except in the above case 2b from the beginning of the mutual cooperation to its end.[2] The learning rate of the target is $\alpha \in (0, 1)$.

A. Change of $Q(D)$

Let t be the start time of the mutual cooperation and t_k be the time from t to the k-th DC ($t_0 = 0$). Then, from Eq. 1, the change of $Q(D)$, or $\Delta_k(D)$, caused by the k-th DC becomes

$$\begin{aligned}\Delta_k(D) &\equiv Q_{t+t_k+1}(D) - Q_{t+t_k}(D)\\ &= \alpha\big(T + \gamma Q_{t+t_k}(C) - Q_{t+t_k}(D)\big)\\ &= \alpha\left(T + \gamma\frac{R}{1-\gamma} - Q_{t+t_k}(D)\right),\end{aligned} \tag{3}$$

because $Q_{t+t_k}(C) > Q_{t+t_k}(D)$. Since $Q(D)$ only changes at $t + t_k + 1$ ($k = 1, 2, 3, ...$),

$$\begin{aligned}&Q_{t+t_k}(D) = Q_{t+t_{k-1}+1}(D)\\ &= (1-\alpha)Q_{t+t_{k-1}}(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\\ &= (1-\alpha)Q_{t+t_{k-2}+1}(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\\ &= (1-\alpha)^2 Q_{t+t_{k-2}}(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\big((1-\alpha)+1\big)\\ &= (1-\alpha)^3 Q_{t+t_{k-3}}(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\\ &\qquad \times \big((1-\alpha)^2 + (1-\alpha) + 1\big)\\ &= \quad \cdots\\ &= (1-\alpha)^{k-1} Q_{t+t_1}(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\sum_{j=1}^{k-1}(1-\alpha)^{j-1}\\ &= (1-\alpha)^{k-1} Q_t(D) + \alpha\left(T + \gamma\frac{R}{1-\gamma}\right)\sum_{j=1}^{k-1}(1-\alpha)^{j-1}.\end{aligned} \tag{4}$$

Finally, by substituting Eq. 4 in Eq. 3 and from the following equation

$$\alpha\sum_{j=1}^{k-1}(1-\alpha)^{j-1} = 1 - (1-\alpha)^{k-1}, \tag{5}$$

we get $\Delta_k(D)$ as follows.

$$\begin{aligned}\Delta_k(D) &= \alpha(1-\alpha)^{k-1}\left(T + \gamma\frac{R}{1-\gamma} - Q_t(D)\right)\\ &= \alpha(1-\alpha)^{k-1}\left(T + \frac{\gamma R - P}{1-\gamma}\right).\end{aligned} \tag{6}$$

Note that $T + (\gamma R - P)/(1-\gamma) > 0$.

B. Duration of the mutual cooperation ended by the target

Based on the above result, let us derive l_a that satisfies

$$Q_{t+t_{l_a}+1}(D) > Q_{t+t_{l_a}+1}(C) = \frac{R}{1-\gamma}.$$

Since

$$Q_{t+t_{l_a}+1}(D) = Q_t(D) + \sum_{k=1}^{l_a}\Delta_k(D),$$

we get

$$\sum_{k=1}^{l_a}\Delta_k(D) > \frac{R-P}{1-\gamma}. \tag{7}$$

From Eqs. 5, 6, and 7, we get

$$\begin{aligned}(1-\alpha)^{l_a} &< 1 - \frac{R-P}{(1-\gamma)T + \gamma R - P}\\ &= \frac{(1-\gamma)(T-R)}{(1-\gamma)T + \gamma R - P}.\end{aligned}$$

Since $\log(1-\alpha) < 0$, we get

$$l_a > \frac{1}{\log(1-\alpha)}\log\frac{(1-\gamma)(T-R)}{(1-\gamma)T + \gamma R - P}. \tag{8}$$

Let the righthand side of Eq. 8 be l_a^{inf}. Then, the minimum integer L_a that satisfies Eq. 8 is $\lceil l_a^{\mathit{inf}} \rceil$.[3]

Next, the expected duration between two DCs when $Q(C) > Q(D)$ in both agents, or n_{dc}, is

$$n_{dc} = \frac{1}{\frac{\varepsilon_s}{2}\left(1 - \frac{\varepsilon_o}{2}\right)} = \frac{4}{\varepsilon_s(2-\varepsilon_o)},$$

where ε_s and ε_o are the parameter of ε-greedy method in the target and the opponent, respectively.

Therefore, the expected duration from the start of the mutual cooperation to its end caused by the target, or d_a, is

$$d_a = n_{dc} \times L_a.$$

[2] In other words, $Q(C) = R/(1-\gamma)$ in the whole duration except for the last game in the "ended by the opponent" case, because CD does not happen.

[3] If l_a^{inf} is an integer, $L_a = l_a^{\mathit{inf}} + 1$.

C. Duration of the mutual cooperation ended by the opponent

Here we consider the case where the mutual cooperation ends due to one CD. It means that $Q(C) > Q(D)$ in both agents before CD, but the difference of them of the target is smaller than the reduction of $Q(C)$ caused by the payoff S.

First of all, let us derive l_b that satisfies

$$\begin{aligned} & Q_{t+t_{l_b}+1}(D) > Q_{t+t_{l_b}+1}(C) \\ & = (1-\alpha)Q_{t+t_{l_b}}(C) + \alpha(S + \gamma Q_{t+t_{l_b}}(C)) \\ & = (1-\alpha+\alpha\gamma)Q_{t+t_{l_b}}(C) + \alpha S. \end{aligned}$$

Similar to the previous subsection, we get

$$l_b > \frac{1}{\log(1-\alpha)} \log \frac{(1-\gamma)\{T-(1-\alpha)R-\alpha S\}}{(1-\gamma)T+\gamma R-P}. \quad (9)$$

Let the righthand side of Eq. 9 be l_b^{inf}. Then, the minimum integer L_b that satisfies Eq. 9 is $\lceil l_b^{\mathit{inf}} \rceil$.[4] Note that $l_a^{\mathit{inf}} > l_b^{\mathit{inf}}$.

Next, the expected duration n_{cd} from the L_b-th DC to the final CD when $Q(C) > Q(D)$ in both agents is

$$n_{cd} = \frac{1}{\left(1-\dfrac{\varepsilon_s}{2}\right)\dfrac{\varepsilon_o}{2}} = \frac{4}{\varepsilon_o(2-\varepsilon_s)}.$$

Hence, the expected duration from the start of the mutual cooperation to its end caused by the opponent, or d_b, is

$$d_b = n_{dc} \times L_b + n_{cd}.$$

D. Expected duration of the mutual cooperation

The above two subsections derive the duration of the mutual cooperation for each case: "ended by the target" and "ended by the opponent". Here we derive two probabilities p_a and p_b corresponding to the former and the latter, respectively, and finally, the expected duration of the mutual cooperation from its start to end. Since the opponent does not choose D except for the last game of the latter case where the target chooses C, there is no mutual defection (DD) in the whole duration.

The latter case happens when the opponent chooses D while the target agent has chosen it not less than L_b but less than L_a times. Thus, the expected duration where the latter case happens, or Δn, is

$$\Delta n = n_{dc} \times (L_a - L_b).$$

Hence, the former, "ended by the target agent", case happens when the opponent consecutively chooses the cooperation Δn times. Then,

$$p_a = \left(1 - \frac{\varepsilon_o}{2}\right)^{\Delta n}.$$

Since there is no mutual defection, the latter, "ended by the opponent", case happens when the former case does not happen. That is,

$$p_b = 1 - p_a = 1 - \left(1 - \frac{\varepsilon_o}{2}\right)^{\Delta n}.$$

Consequently, the expected duration of the mutual cooperation from its start to end, or d, is

$$d = p_a d_a + p_b d_b. \quad (10)$$

IV. Numerical verification

We conducted an experiment 100 times in each of which two stateless Q-learning agents with ε-greedy method played a PD game 1000 times. Each agent updated its Q-function after every game. Here we look at the result and compare it to the derived values in the previous section. Both agents have same parameters: $\alpha = 0.25$, $\gamma = 0$, and $\varepsilon = 0.1$. The payoffs used in the experiment are as follows: $T = 5$, $R = 4.1$, $P = 1$, $S = 0$.[5]

TABLE II. Agents' behaviors in the 100000 games in total

Row \ Column	C	D
C	11091	2894
D	2643	83372

Table II shows how many the action pairs appeared in the experiment. This table shows that CC appeared in about 11% of games. Since the agents used the ε-greedy method for action selection, the probability of CC should be $\varepsilon^2/4 = 0.25\%$ when $Q(D) > Q(C)$ in both agents. The difference is from the fact that one CC from a chance sometimes made $Q(C) > Q(D)$ in both agents and the agents took CC in the probability of $(1-\varepsilon)^2/4 = 90.25\%$ afterward. However, such CC is not continued forever, because a unilateral defection from a chance (in the probability of 9.5%) sometimes made $Q(D) > Q(C)$ in an agent again and they went back mutual defection afterward.

TABLE III. Comparison between derived, experimental, and previous results

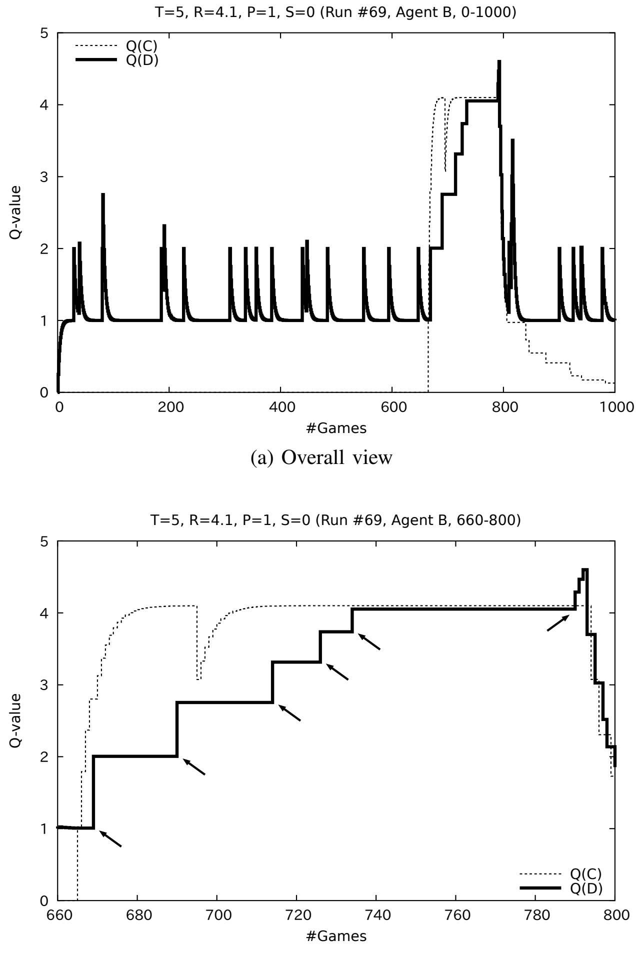
	Derived	Experimental	Previous
L_a	6	4	—
L_b	3	1.992	—
n_{dc}	21.05	—	—
n_{cd}	21.05	—	—
d	85.86	73.13	9.26
d_a	126.32	93.86	—
d_b	84.21	74.66	—
Δn	63.16	—	—
p_a	0.039	0.14	—
p_b	0.961	0.793	—
p_{ab}	—	0.067	—

Table III shows the derived and the experimental results, with the result of the previous work (Section II-C) for comparison. The values of the experimental result in the table are given as follows. First we found sequences where $Q(C) > Q(D)$ in both agents from the whole result. Each sequence was categorized by the reason of its end: "by the target" and "by the opponent".[6] L_a was calculated from the sequences in the "by the target" group, by dividing the total number of

[4] If l_b^{inf} is an integer, $L_b = l_b^{\mathit{inf}} + 1$.

[5] The previous work (Section II-C) [17] requires $R > 4$ in this setting. If not, p_2 becomes 0 that means the first mutual cooperation cannot make $Q(C) > Q(D)$ at all. We also know the payoff values directly influence Q-learning agents' behaviors. See that work [17] for details.

[6] More precisely, if the agents that took the final unilateral defection and had $Q(D) > Q(C)$ after the defection were identical, the sequence was into the "by the target" group. If they were different, it was into the "by the opponent" group. Note that there was another case where both agents had $Q(D) > Q(C)$, which was ignored in the discussion of Section III.



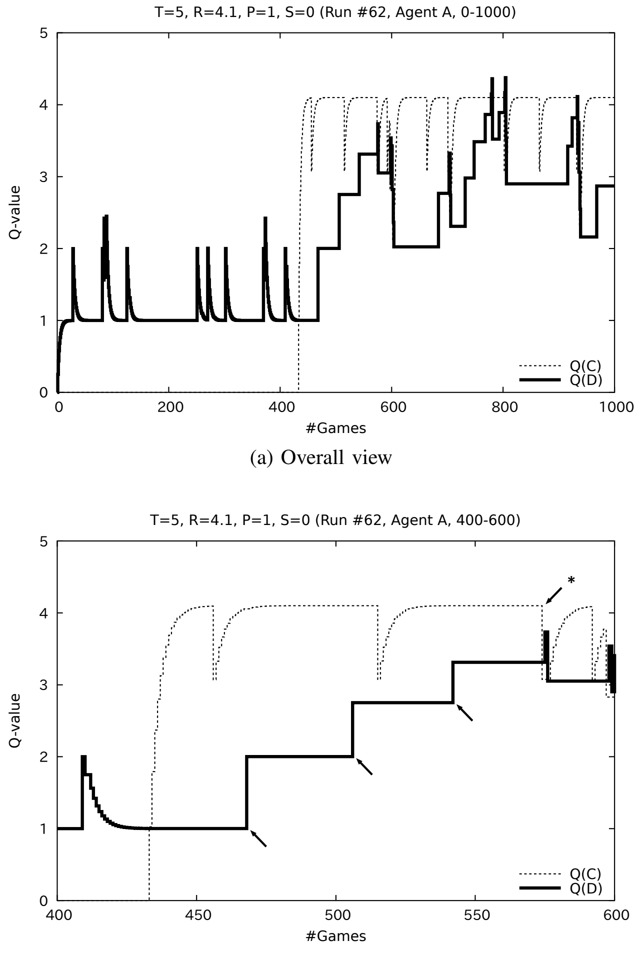
(a) Overall view

(b) Closeup view from the 660th to the 800th game

Fig. 1. Learning curves where the mutual cooperation was ended by the target. The x-axis shows the number of games and the y-axis shows the Q-values the agent had. The arrows in the closeup view indicate when DC happens. We can see that the sixth DC makes $Q(D) > Q(C)$.

unilateral defection of the target by the number of sequences. L_b was calculated similarly from the sequences in the "by the opponent" group. The value d is the average length of all of the sequences. The values d_a and d_b are the average lengths of the sequences in the "by the target" group and in the "by the opponent" group, respectively. The values p_a and p_b are the ratio of the sequences in the "by the target" group and in the "by the opponent" group, respectively, to all of the sequences. The value p_{ab} shows the ratio of another case where the final unilateral defection made $Q(D) > Q(C)$ simultaneously in *both* agents. The value d of the previous work was calculated from Eq. 2, where $p_3 = (1 - \varepsilon/2)^2$.

Figure 1 shows learning curves of one agent in a certain run. It shows the "ended by the target" case. We can see that $Q(C) > Q(D)$, *i.e.*, the agent became to prefer C to D, at the 665th game due to CC at the game. Note that $Q(D)$ was about 1, approximately minimum at that time. After that, the agent took five Ds as indicated by the arrows, and at the sixth D at the 790th game (indicated by the last arrow), the agent became to prefer D again and the mutual cooperation was ended. Note that the number of D is identical to the derived L_a.



(a) Overall view

(b) Closeup view from the 400th to the 600th game

Fig. 2. Learning curves where the mutual cooperation was ended by the opponent. The x-axis shows the number of games and the y-axis shows the Q-values the agent had. The arrows except for that with an asterisk in the closeup view indicate when DC happens. We can see that the asterisked arrow in the closeup view, showing the time of CD, makes $Q(D) > Q(C)$ of the agent.

Figure 2 shows learning curves of one agent in another certain run. It shows the "ended by the opponent" case. We can see that $Q(C) > Q(D)$, *i.e.*, the agent became to prefer C to D, at the 433rd game due to CC at the game. Note that $Q(D)$ was about 1, approximately minimum value at that time. After that, the agent took three Ds as indicated by the arrows. Finally, at the 574th game (indicated by the asterisked arrow), the opponent took D, $Q(C)$ of the agent fell down suddenly, and the agent became to prefer D again. Note that the number of D before the opponent took D is identical to the derived L_b.

The derived numbers of games until the end of the mutual cooperation after it appears (d_a, d_b, and d) are larger than the experimental results. It is due to the premise of the deriving process. In this work, we suppose that $Q(D)$ is minimum when the mutual cooperation starts and $Q(C)$ is maximum at the first defection after the mutual cooperation has started. This premise is satisfied in some cases as shown in the figures, but not always. As we can see in Fig. 2b, $Q(C)$ did not decrease very much immediately because the agent did not choose C for a while after the end of the mutual cooperation. On the other hand, $Q(D)$ decreased quickly due to mutual defection. Thus, $Q(D)$ fell below $Q(C)$ soon and the agent started to choose C again. Furthermore, the agent chose D by chance before $Q(D)$ did not decrease so much, which made $Q(D) > Q(C)$. Hence, it was not necessary to choose D for L_a nor L_b times in order to make $Q(D) > Q(C)$ again. Indeed, the experimental result in Table III shows that L_a and L_b were smaller than the derived values. It shortened the duration until the end of the mutual cooperation. It may also be the reason why p_a in the experiment was larger than the derived value.

On the other hand, the duration d of the previous work is much smaller than the experimental value. As we discussed in Section II-C, it is because the premise of Eq. 2 does not consider the resilience of mutual cooperation played by Q-learning agents at all. Thus, we can say that this work is more appropriate than the previous work.

Note that, in the previous section, we ignored the case where mutual defection appears when $Q(C) > Q(D)$ in both agents. Indeed, the experimental result shows that it happened only five times in the whole experiment.

V. Conclusion

It had been shown in the literature that stateless, independent Q-learning agents had been difficult to cooperate with each other in an iterated Prisoner's Dilemma (IPD) game. However, there is a case where one mutual cooperation occurred by "mistakes" sometimes helps the agents change to prefer cooperation. Mutual cooperation will reinforce itself, and finally, the agent will continue to choose cooperation "intentionally" for a while.

In the previous work, we had discussed the ratio of mutual cooperation in an IPD game played by stateless Q-learning agents, but had not considered the resilience of mutual cooperation at all. Thus, this work derived the expected number of games from such a mistaken mutual cooperation to the end of intentional mutual cooperation, *i.e.*, an intentional defection of one agent, in an IPD game played by two stateless Q-learning agents.

We compared the expected duration of mutual cooperation and that from the experimental result. The derived duration was slightly longer than the experimental results because the premise did not always hold. For example, when the preference of cooperation and that of defection were different from the premise, the number of unilateral defections necessary to end intentional mutual cooperation was less than the derived one. Nevertheless, this work is more appropriate than the previous work that showed the duration much shorter than the experimental result.

As a future work, we will relax the premise and include factors omitted in this work in order to derive the expected duration of mutual cooperation precisely and get further comprehension of the process of reinforcement learning in multiagent systems.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP16K00302, Kayamori Foundation of Informational Science Advancement, and the Hori Sciences & Arts Foundation.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[2] C. Claus and C. Boutilier, "The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems," in *Proc. 15th National Conference on Artificial Intelligence, (AAAI)*, 1998, pp. 746–752.

[3] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, pp. 215–250, 2002.

[4] J. Hu and M. P. Wellman, "Nash Q-Learning for General-Sum Stochastic Games," *Journal of Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.

[5] G. Tesauro, "Extending Q-Learning to General Adaptive Multi-Agent Systems," in *Advances in Neural Information Processing Systems 16*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2004, pp. 871–878.

[6] E. M. de Cote, A. Lazaric, and M. Restelli, "Learning to Cooperate in Multi-Agent Social Dilemmas," in *Proc. 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2006, pp. 783–785.

[7] D. Banerjee and S. Sen, "Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning," *Autonomous Agents and Multi-Agent Systems*, vol. 15, pp. 91–108, 2007.

[8] J. W. Crandall and M. A. Goodrich, "Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning," *Machine Learning*, vol. 82, pp. 281–314, 2011.

[9] J. Hao and H. Leung, "Introducing decision entrustment mechanism into repeated bilateral agent interactions to achieve social optimality," *Autonomous Agents and Multi-Agent Systems*, vol. 29, pp. 658–682, 2015.

[10] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.

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

[12] W. Poundstone, *Prisoner's Dilemma*. New York: Doubleday, 1992.

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

[14] C. J. C. H. Watkins and P. Dayan, "Technical Note: Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[15] M. Wunder, M. Littman, and M. Babes, "Classes of Multiagent Q-learning Dynamics with ϵ-greedy Exploration," in *Proc. 27th International Conference on Machine Learning (ICML)*, 2010, pp. 1167–1174.

[16] K. Moriyama, "Utility based Q-learning to facilitate cooperation in Prisoner's Dilemma games," *Web Intelligence and Agent Systems*, vol. 7, no. 3, pp. 233–242, 2009.

[17] K. Moriyama, S. Kurihara, and M. Numao, "Cooperation-Eliciting Prisoner's Dilemma Payoffs for Reinforcement Learning Agents," in *Proc. 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2014, pp. 1619–1620.