

Symmetric solution of the Bellman optimality equation for repeated harmony game

Hisato Komatsu^{a,b}

^a*Department of Physics, Kindai University, 577-8502, Higashi-Osaka, Osaka, Japan*

^b*Data Science and AI Innovation Research Promotion Center, Shiga
University, 522-8522, Hikone, Shiga, Japan*

Abstract

In social dilemma games, additional rewards or punishments have been studied as means of promoting cooperation. Therefore, it is important to investigate the ideal situation, in which such an additional payoff would change the game. In this study, we investigated the symmetric solution of the Bellman optimality equation for a repeated harmony game. The calculations showed that three types of symmetric solutions exist. One of them corresponds to the trivial All-C strategy, and another to the Win-stay Lose-shift strategy of the prisoners dilemma game. The nontrivial behavior of the strategy corresponding to the last solution is also discussed in detail. In addition, we numerically investigated which strategy the agents actually learn by the reinforcement learning algorithm.

1. Introduction

The methods for promoting cooperation in social dilemma games are important in game theory. One such method is the introduction of additional rewards or punishments by participants [1, 2] or third parties [3, 4]. Along with recent advances in machine learning, multi-agent reinforcement learning methods incorporating such additional rewards have been proposed [5]. Moreover, the learning dynamics of agents in the presence of additional rewards have also been studied [6]. Various studies have been conducted on reinforcement learning applied to repeated games [7, 8, 9]. Usui and Ueda investigated the symmetric solution, i.e., the solution in which the two agents have the same strategy, of the Bellman optimality equation for the repeated prisoner's dilemma [10]. They showed that the All-D strategy creates an equilibrium regardless of the conditions, and the Win-stay Lose-shift (WSLS) [11] and Grim strategies create it when specific conditions are fulfilled. The Bellman optimality equation was also employed by Meylahn and Janssen to investigate the best response to an opponent with a fixed strategy [12, 13].

Email address: `komatsu@phys.kindai.ac.jp` (Hisato Komatsu)

The harmony game is a 2×2 game in which cooperation yields a higher payoff for an agent regardless of the opponent’s strategy. This game is often excluded from theoretical considerations because cooperation is the dominant strategy, at least in one-shot games. However, several previous studies have reported that, even in this game, cooperation can collapse under certain evolutionary [14] or learning dynamics [15], or the spatial structure [16], indicating that the emergence and maintenance of cooperation is not always trivial in practice. Moreover, in repeated games, the folk theorem ensures the existence of strategies that achieve a wide range of average payoffs [17]. Therefore, before considering situations in which additional payoffs are introduced to promote cooperation, we first examine the ideal case in which such additional payoffs completely transform the game into a harmony game. This study investigated the Bellman optimality equation of the harmony game with an approach similar to that in Ref. [10] and lists possible strategies corresponding to symmetric solutions. In addition, we performed numerical experiments using reinforcement learning (Q-learning [20, 21]) to examine which strategies the agents learn under several initial conditions.

The remainder of this paper introduces this model in Sec. 2 and then derives the symmetric solutions of the Bellman optimality equation in Sec. 3. We compare these solutions with the numerical simulations in Sec. 4, and summarize the study in Sec. 5. For reproducibility, details of the simulations in Sec. 4 are provided in Appendix A. We also discuss the optimal strategy when the opponent’s strategy is fixed, in Appendix B.

2. Model

We consider the repeated harmony game, in which two agents (A and B) choose an action, either C (cooperation) or D (defection), at each time step. The reward for each agent, r_A or r_B , depends on their joint action, (a, b) , according to

$$\begin{aligned} r_A(C, C) &= r_B(C, C) = R, \\ r_A(C, D) &= r_B(D, C) = S, \\ r_A(D, C) &= r_B(C, D) = T, \\ r_A(D, D) &= r_B(D, D) = P, \end{aligned} \tag{1}$$

with $R > T > P$ and $R > S > P$. Strictly speaking, the definition of the harmony game varies slightly across references. Our definition is identical to that used in Ref. [16], for example. Throughout this paper, we use the term “reward” instead of “payoff” to align with the terminology of reinforcement learning. In the one-shot game, the unique Nash equilibrium is (C, C) , and C is the dominant strategy of both agents. Moreover, unlike the prisoner’s dilemma, the one-shot game lacks a social dilemma. However, in a repeated game, the folk theorem ensures the existence of Nash equilibrium strategies other than the trivial All- C strategy [17]. Therefore, theoretical and numerical investigations should be carefully conducted to avoid overlooking such nontrivial strategies. For simplicity, we focus on memory-1 strategies, in which agents forget any actions taken

two steps ago or earlier and choose their current action based solely on their previous action. In this case, the observation of the agent was the previous joint action.

3. Bellman optimality equation and its symmetric solution

By ignoring exploration noise such as ϵ in ϵ -greedy method, the Bellman optimality equation is expressed as follows:

$$Q_A^*((a_0, b_0), a_1) = r_A(a_1, \tilde{b}_1) + \gamma \max_{a_2} Q_A^*((a_1, \tilde{b}_1), a_2), \quad (2)$$

$$Q_B^*((a_0, b_0), b_1) = r_B(\tilde{a}_1, b_1) + \gamma \max_{b_2} Q_B^*((\tilde{a}_1, b_1), b_2), \quad (3)$$

$$\text{where } \tilde{a}_1 = \operatorname{argmax}_{a_1} Q_A^*((a_0, b_0), a_1), \tilde{b}_1 = \operatorname{argmax}_{b_1} Q_B^*((a_0, b_0), b_1). \quad (4)$$

To derive Eq. (4), we assume that agents employ the greedy method.

Assuming symmetry of the Q-function between the two agents,

$$Q_A^*((a_0, b_0), a) = Q_B^*((b_0, a_0), a), \quad (5)$$

Eq. (3) can be reduced to Eq. (2). Therefore, we consider Eq. (2) under the following conditions:

$$\tilde{b}_1 = \operatorname{argmax}_{b_1} Q_A^*((b_0, a_0), b_1). \quad (6)$$

In the following discussions, we ignore the case in which the Q-functions for the two actions accidentally coincide: $Q_A^*((a, b), C) = Q_A^*((a, b), D)$, considering that such cases rarely occur.

3.1. Limitation on the symmetric solution

In this section, we prove two lemmas that characterize the allowed pairwise ordering of the Q-function.

Lemma 1. $Q_A^*((D, D), C) > Q_A^*((D, D), D)$

PROOF. If we assume $Q_A^*((D, D), C) < Q_A^*((D, D), D)$, then Bellman equation for $Q_A^*((D, D), a_1)$ is expressed as:

$$Q_A^*((D, D), C) = S + \gamma \max_{a_2} Q_A^*((C, D), a_2), \quad (7)$$

$$Q_A^*((D, D), D) = P + \gamma \max_{a_2} Q_A^*((D, D), a_2) = P + \gamma Q_A^*((D, D), D). \quad (8)$$

Solving Eq. (8), we obtain

$$Q_A^*((D, D), D) = \frac{P}{1 - \gamma}. \quad (9)$$

Considering that the reward at each step has a minimum value, P , the following inequality is derived from the original definition of the Q-function:

$$\begin{aligned} Q_A^*((a_0, b_0), a_1) &= E \left[\sum_{t=0}^{\infty} \gamma^t r_A(a_{t+1}, b_{t+1}) \middle| (a_0, b_0), a_1 \right] \\ &\geq \sum_{t=0}^{\infty} \gamma^t P = \frac{P}{1-\gamma} \end{aligned} \quad (10)$$

Substituting this into Eqs. (7) and (9), we obtain the following inequality:

$$Q_A^*((D, D), C) - Q_A^*((D, D), D) \geq S + \frac{\gamma P}{1-\gamma} - \frac{P}{1-\gamma} = S - P > 0. \quad (11)$$

However, this contradicts the first assumption, $Q_A^*((D, D), C) < Q_A^*((D, D), D)$. Thus, $Q_A^*((D, D), C) > Q_A^*((D, D), D)$. $\square$

Using Lemma 1, the Bellman optimality equation for $Q_A^*((D, D), a)$ can be expressed as

$$Q_A^*((D, D), C) = R + \gamma \max_{a_2} Q_A^*((C, C), a_2), \quad (12)$$

$$Q_A^*((D, D), D) = T + \gamma \max_{a_2} Q_A^*((D, C), a_2), \quad (13)$$

Lemma 2. *Only the following two order-relation patterns are permitted:*

$$Q_A^*((D, C), C) > Q_A^*((D, C), D) \text{ and } Q_A^*((C, D), C) > Q_A^*((C, D), D) \quad (14)$$

$$Q_A^*((D, C), C) < Q_A^*((D, C), D) \text{ and } Q_A^*((C, D), C) < Q_A^*((C, D), D) \quad (15)$$

PROOF. Assuming $Q_A^*((D, C), C) < Q_A^*((D, C), D)$ and $Q_A^*((C, D), C) > Q_A^*((C, D), D)$, we obtain the following equations:

$$Q_A^*((D, C), C) = R + \gamma \max_{a_2} Q_A^*((C, C), a_2) = Q_A^*((D, D), C), \quad (16)$$

$$Q_A^*((D, C), D) = T + \gamma \max_{a_2} Q_A^*((D, C), a_2) = Q_A^*((D, D), D), \quad (17)$$

Here, Eqs. (12) and (13) were used to derive these equations. Considering Eqs. (16) and (17), and Lemma 1, the following inequality holds:

$$Q_A^*((D, C), C) = Q_A^*((D, D), C) > Q_A^*((D, D), D) = Q_A^*((D, C), D). \quad (18)$$

However, this result contradicts the assumption $Q_A^*((D, C), C) < Q_A^*((D, C), D)$. Similarly, $Q_A^*((D, C), C) > Q_A^*((D, C), D)$ and $Q_A^*((C, D), C) < Q_A^*((C, D), D)$ produces the contradiction.

Therefore, Eqs. (14) and (15) are the only permitted order-relation patterns. $\square$

3.2. Explicit solutions

Next, we derive explicit solutions for the four cases allowed by the lemmas in Sec. 3.1. Each solution has the same form as in Ref. [10] because the structure of the Bellman equation does not depend directly on the relationship among R, T, S , and P , but on the assumed pairwise ordering of the Q -function.

In this discussion, we use the following expression:

$$\begin{aligned} x_1 &\equiv R + \gamma \max_{a_2} Q_A^*((C, C), a_2), \\ x_2 &\equiv T + \gamma \max_{a_2} Q_A^*((D, C), a_2), \\ x_3 &\equiv S + \gamma \max_{a_2} Q_A^*((C, D), a_2), \\ x_4 &\equiv P + \gamma \max_{a_2} Q_A^*((D, D), a_2), \end{aligned} \tag{19}$$

for later convenience. This expression is primarily used in Appendix B

3.2.1. The case $Q_A^((C, C), C) > Q_A^*((C, C), D)$, $Q_A^*((C, D), C) > Q_A^*((C, D), D)$, $Q_A^*((D, C), C) > Q_A^*((D, C), D)$, and $Q_A^*((D, D), C) > Q_A^*((D, D), D)$*

This corresponds to an All-C strategy. The solution is given as

$$\begin{aligned} x_1 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R \\ x_2 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma}R \end{aligned} \tag{20}$$

This is always consistent with the condition of the harmony game.

3.2.2. The case $Q_A^((C, C), C) > Q_A^*((C, C), D)$, $Q_A^*((C, D), C) < Q_A^*((C, D), D)$, $Q_A^*((D, C), C) < Q_A^*((D, C), D)$, and $Q_A^*((D, D), C) > Q_A^*((D, D), D)$*

This corresponds to the WSLs strategy in the prisoner's dilemma [11]. In this study, we use the term “WSLS” for convenience, although the notions of “win” and “lose” in the harmony game differ from those in the prisoner's dilemma. The solution is given as

$$\begin{aligned} x_1 &= Q_A^*((C, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R \\ x_2 &= Q_A^*((C, C), D) = Q_A^*((D, D), D) = T + \gamma P + \frac{\gamma^2}{1-\gamma}R, \\ x_3 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = S + \gamma P + \frac{\gamma^2}{1-\gamma}R, \\ x_4 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = P + \frac{\gamma}{1-\gamma}R, \end{aligned} \tag{21}$$

This solution is consistent with the condition of the game when $\gamma > \frac{S-P}{R-P}$.

3.2.3. The case $Q_A^*((C, C), C) < Q_A^*((C, C), D)$, $Q_A^*((C, D), C) > Q_A^*((C, D), D)$,
 $Q_A^*((D, C), C) > Q_A^*((D, C), D)$, and $Q_A^*((D, D), C) > Q_A^*((D, D), D)$

In this case, the solution is given as

$$\begin{aligned} x_1 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma^2}(R + \gamma P), \\ x_2 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma^2}(R + \gamma P), \\ x_3 &= Q_A^*((C, C), C) = S + \frac{\gamma}{1-\gamma^2}(R + \gamma P), \\ x_4 &= Q_A^*((C, C), D) = \frac{1}{1-\gamma^2}(P + \gamma R) \end{aligned} \tag{22}$$

However, this solution contradicts the condition $S > P$.

3.2.4. The case $Q_A^*((C, C), C) < Q_A^*((C, C), D)$, $Q_A^*((C, D), C) < Q_A^*((C, D), D)$,
 $Q_A^*((D, C), C) < Q_A^*((D, C), D)$, and $Q_A^*((D, D), C) > Q_A^*((D, D), D)$

In this case, the solution is given as

$$\begin{aligned} x_1 &= Q_A^*((D, D), C) = \frac{1}{1-\gamma^2}(R + \gamma P), \\ x_2 &= Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma^2}(P + \gamma R), \\ x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, C), C) = S + \frac{\gamma}{1-\gamma^2}(P + \gamma R), \\ x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, C), D) = \frac{1}{1-\gamma^2}(P + \gamma R). \end{aligned} \tag{23}$$

This solution is consistent with the condition of the game when $\gamma > \frac{S-P}{R-S}$.

In this strategy, an agent chooses C if and only if both agents chose D in the previous step. In the repeated prisoner's dilemma, the “handshake mechanisms”, in which agents use particular initial action patterns as the signal for cooperation, have been proposed [18, 19]. Compared to these mechanisms, this strategy seems to require frequent “ D ” handshakes from the opponent at an appropriate time as a condition of cooperation. Hence, in the following, we call this strategy the “Frequent Handshake (FH) strategy”. Note that this strategy is termed the “anti-Grim trigger” in some studies [13], however, the definition of the term “anti-Grim trigger” is confusing because its definition varies across the literature.

The FH strategy may seem odd because it deliberately destroys cooperation even though defection yields a lower reward. Therefore, to illustrate the actual dynamics, we examine several cases in which agent A uses specific strategies while agent B's strategy is fixed as an FH strategy. First, if agent A adopts the FH strategy, the sequence converges to the following limit cycle:

$$(C, C) \rightarrow (D, D) \rightarrow (C, C) \rightarrow \dots \tag{24}$$

In this case, the long-term average of agent A's reward is $\frac{R+P}{2}$. For agent A using the WSLS strategy, the sequence converges to the following limit cycle:

$$(C, C) \rightarrow (C, D) \rightarrow (D, D) \rightarrow (C, C) \rightarrow \dots, \quad (25)$$

and the long-term average of agent A's reward is $\frac{R+S+P}{3}$. If agent A employs the All-C strategy, the sequence converges to a fixed point,

$$(C, D), \quad (26)$$

and the long-term average of agent A's reward is S . Among the three strategies, the FH strategy achieves the highest long-term average when

$$\frac{R+P}{2} > S. \quad (27)$$

The condition that the FH strategy appears, $\gamma > \frac{S-P}{R-S}$, converges to Eq. (27) when $\gamma \rightarrow 1$. In addition, considering the following:

$$S = \min_b \max_a r_A(a, b), \quad (28)$$

Eq. (27) has the same form as the condition under which the folk theorem holds (for γ sufficiently close to 1) [17].

The best response of agent A under the fixed strategy of agent B is investigated in detail in Appendix B.

4. Numerical experiments

In this section, we numerically investigate the strategies adopted by actual reinforcement learning agents. Here, Q-learning [20, 21] was employed as the learning algorithm, and the two agents independently updated their Q-functions Q_A and Q_B . In other words, these functions were updated after the reward was calculated at each step using the following equation:

$$\begin{aligned} Q_A((a_{t-1}, b_{t-1}), a_t) &\leftarrow Q_A((a_{t-1}, b_{t-1}), a_t) \\ &+ \alpha \left[r_A(a_t, b_t) + \gamma \max_{a'} Q_A((a_t, b_t), a') - Q_A((a_{t-1}, b_{t-1}), a_t) \right], \\ Q_B((a_{t-1}, b_{t-1}), a_t) &\leftarrow Q_B((a_{t-1}, b_{t-1}), a_t) \\ &+ \alpha \left[r_B(a_t, b_t) + \gamma \max_{b'} Q_B((a_t, b_t), b') - Q_B((a_{t-1}, b_{t-1}), b_t) \right], \end{aligned} \quad (29)$$

where α denotes the learning rate. The rewards for each joint action were set as $R = 3$, $T = 2$, $S = 1$, and $P = 0$.

One episode is truncated when the time step reaches $T_{\text{ep}} = 100$, but the truncation is not treated as an episode termination when updating the Q-functions according to Eq. (29). In the actual implementation, we realized this by recording the global time step t and resetting the environment when t was a multiple

of $T_{\text{ep}} = 100$; that is, when $t \% T_{\text{ep}} = t \% 100 = 0$. At the beginning of each episode, the observation of agents, (a, b) , is uniformly randomly chosen from (C, C) , (C, D) , (D, C) , and (D, D) . In other words, the previous actions observed by the agents in the initial state are selected uniformly at random. At each step, the agents choose their actions according to the ϵ -greedy method. For each agent, $i = A, B$, the initial value of ϵ_i is set to ϵ_{init} , and ϵ_i is subsequently decreased using the following equation:

$$\epsilon_i \leftarrow \max[\epsilon_{\min}, (1 - \delta_\epsilon)\epsilon_i], \quad (30)$$

We considered two values of ϵ_{init} : $\epsilon_{\text{init}} = 1$ and $\epsilon_{\text{init}} = 0.02$, whereas the other hyperparameters were fixed, as given in Table A.2 of Appendix A. The detailed pseudocode is provided by Algorithm 1 in this appendix.

Several initial Q-functions were also considered: $Q_{\text{init}} = Q_{\text{zero}}$, Q_{WSLS} , and Q_{FH} , where

$$Q_{\text{zero}}((a, b), a') = 0 \text{ for any } a, b, a', \quad (31)$$

$$\begin{aligned} Q_{\text{WSLS}}((C, C), C) &= Q_{\text{WSLS}}((C, D), D) = Q_{\text{WSLS}}((D, C), D) = Q_{\text{WSLS}}((D, D), C) = 1, \\ Q_{\text{WSLS}}((C, C), D) &= Q_{\text{WSLS}}((C, D), C) = Q_{\text{WSLS}}((D, C), C) = Q_{\text{WSLS}}((D, D), D) = 0, \end{aligned} \quad (32)$$

and

$$\begin{aligned} Q_{\text{FH}}((C, C), D) &= Q_{\text{FH}}((C, D), D) = Q_{\text{FH}}((D, C), D) = Q_{\text{FH}}((D, D), C) = 1, \\ Q_{\text{FH}}((C, C), C) &= Q_{\text{FH}}((C, D), C) = Q_{\text{FH}}((D, C), C) = Q_{\text{FH}}((D, D), D) = 0, \end{aligned} \quad (33)$$

with

$$Q_A((a, b), a')|_{t=0} = Q_B((b, a), a')|_{t=0} = Q_{\text{init}}((a, b), a'). \quad (34)$$

As seen later, when $Q_{\text{init}} = Q_{\text{zero}}$, the agents adopt the All-C strategy with high probability. Therefore, we also investigated two other initial Q-functions, Q_{WSLS} and Q_{FH} , under which agents tend to employ WSLS and FH strategies.

When tuning hyperparameters, we conducted 20 or 100 trials using a random seed different from that used when collecting the reported data. Fig. 1 shows the time evolution of the Q-function in each case. In this figure, the solid curves and shaded areas show the median and interquartile intervals (both computed using the `numpy.quantile()`) over 1000 independent trials. The graphs of the cases with $Q_{\text{init}} = Q_{\text{WSLS}}$ or Q_{FH} and $\epsilon_{\text{init}} = 1$ are omitted because their shapes are almost identical to the corresponding graph for $Q_{\text{init}} = Q_{\text{zero}}$ and $\epsilon_{\text{init}} = 1$. Table 1 lists the number of trials (out of 1000) in which both agents selected each strategy. In this table, the learned strategy is identified by the ordering of the Q-functions at the end of the simulation, and the cases in which the two agents learned different strategies are excluded. According to the table, agents almost always learn the All-C strategy when $\epsilon_{\text{init}} = 1$, whereas the learned strategy depends on Q_{init} when $\epsilon_{\text{init}} = 0.02$. Specifically, agents learned the WSLS strategy when $Q_{\text{init}} = Q_{\text{WSLS}}$ and the FH strategy when $Q_{\text{init}} = Q_{\text{FH}}$ in most trials. Under $Q_{\text{init}} = Q_{\text{zero}}$ and $\epsilon_{\text{init}} = 0.02$, the All-C strategy was learned most frequently. However, the WSLS and FH strategies were sometimes

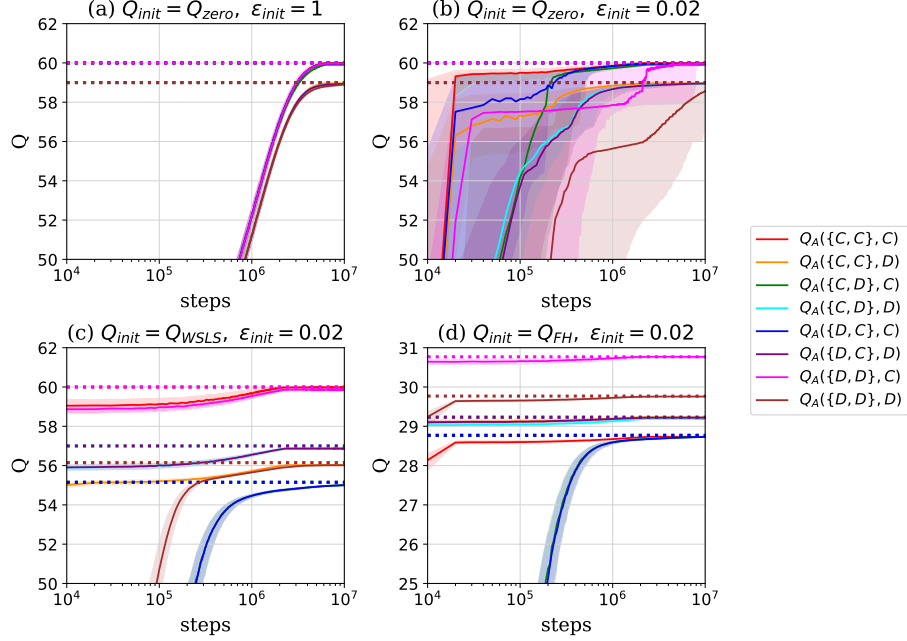


Figure 1: Time evolution of the Q-function (of agent A) with (a) $Q_{\text{init}} = Q_{\text{zero}}$ and $\epsilon_{\text{init}} = 1$, (b) $Q_{\text{init}} = Q_{\text{zero}}$ and $\epsilon_{\text{init}} = 0.02$, (c) $Q_{\text{init}} = Q_{\text{WSLS}}$ and $\epsilon_{\text{init}} = 0.02$, and (d) $Q_{\text{init}} = Q_{\text{FH}}$ and $\epsilon_{\text{init}} = 0.02$. The solid curves and shaded areas of each color represent the median and interquartile intervals of each component $Q_A((a, b), a')$. The dotted lines are the corresponding components of the symmetric solution of the Bellman optimality equation, yielding (a), (b) the All-C strategy (Eq. (20)); (c) the WSLs strategy (Eq. (21)); and (d) the FH strategy (Eq. (23)).

learned. Comparing the results for $Q_{\text{init}} = Q_{\text{zero}}$ with different values of ϵ_{init} , the frequency with which both agents learned the All-C strategy is higher when $\epsilon_{\text{init}} = 1$ than when $\epsilon_{\text{init}} = 0.02$. A possible explanation is that when the opponent behaves randomly, independent of the previous action, the appropriate strategy is always to choose C as in the one-shot game.

To compare the simulation results with the theoretical values discussed in 3.2, we plot the symmetric solutions of the Bellman optimality equation as dotted lines in Fig. 1. These lines correspond to the All-C strategy (Eq. (20)) in graphs (a) and (b), the WSLs strategy (Eq. (21)) in (c), and the FH strategy (Eq. (23)) in (d). Fig. 1 indicates that the medians of the Q-functions converge to a symmetric solution corresponding to the strategies most frequently learned in the simulations. Note that in this simulation, strategies learned with low frequency behaved as outliers, which caused the average over the trials to deviate from the theoretical values. Therefore, we have plotted the median and interquartile intervals in Fig. 1.

initial condition	All-C	WSLS	FH
$Q_{\text{init}} = Q_{\text{zero}}, \epsilon_{\text{init}} = 1$	1000	0	0
$Q_{\text{init}} = Q_{\text{WSLS}}, \epsilon_{\text{init}} = 1$	1000	0	0
$Q_{\text{init}} = Q_{\text{FH}}, \epsilon_{\text{init}} = 1$	1000	0	0
$Q_{\text{init}} = Q_{\text{zero}}, \epsilon_{\text{init}} = 0.02$	823	64	13
$Q_{\text{init}} = Q_{\text{WSLS}}, \epsilon_{\text{init}} = 0.02$	1	999	0
$Q_{\text{init}} = Q_{\text{FH}}, \epsilon_{\text{init}} = 0.02$	0	7	993

Table 1: Number of trials (out of 1000) in which both agents learned each strategy.

5. Summary

In this study, we analyzed the Bellman optimality equation for a repeated harmony game and identified three symmetric solutions: All-C, WSLS, and a third strategy, which we referred to as the Frequent Handshake (FH). In the FH strategy, the agent chooses C (cooperation) only when both it and the opponent have chosen D (defection) in the previous step. If both agents employ this strategy, they repeat (C, C) and (D, D) alternately and punish deviations from this pattern by choosing D .

Numerical experiments using Q-learning showed that agents frequently learn WSLS or FH strategies if the exploration noise is sufficiently weak and the initial Q-function is biased toward these strategies. Our original motivation was to promote cooperation in the social dilemma game by introducing additional rewards, and the harmony game was considered as an ideal case in which these additional rewards completely resolve the dilemma at least in the one-shot game. The fact that such nontrivial strategies exist even in the harmony game makes the design of the additional reward complex. In particular, the FH strategy frequently chooses D and consequently receives less reward than the All-C and WSLS strategies. If frequent defections were observed in more complex environments or under more sophisticated reinforcement learning algorithms, it would be difficult to determine whether they reflected a non-trivial strategy learned by the agents or were merely due to insufficient relaxation or algorithmic issues.

The source code for this work is uploaded to https://github.com/Hisato-Komatsu/Repeated_Harmony.

Acknowledgements

We would like to thank Editage (www.editage.jp) for English language editing.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used DeepL, Google Translate, Microsoft Copilot, and ChatGPT for English language editing, translation,

character	meaning	value
γ	discount factor	0.95
δ_ϵ	decaying rate of ϵ per one step	1×10^{-6}
$\epsilon_{\min}$	minimum value of ϵ	0.002
α	learning rate	0.1
T_{ep}	episode length	100
$T_{\max}$	whole time steps	10^7
R	–	3
T	–	2
S	–	1
P	–	0

Table A.2: Hyperparameters of this study.

research-related queries, and source code review and improvement suggestions. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Appendix A. Details of the numerical experiment

This section describes the pseudocode (Algorithm 1), and lists the hyperparameters (Table A.2) for reproducibility.

Algorithm 1 Algorithm used for the calculation

- 1: Initialize Q_A and Q_B using Eq. (34)
 - 2: $\epsilon_A, \epsilon_B \leftarrow \epsilon_{\text{init}}$
 - 3: Set the initial observation (a_0, b_0) randomly
 - 4: **for** $t \leftarrow 1, \dots, T_{\max}$ **do**
 - 5: Choose agents' action a_t and b_t using ϵ -greedy method with $\epsilon = \epsilon_A$ and $\epsilon = \epsilon_B$, respectively
 - 6: Calculate the rewards r_A and r_B
 - 7: Update Q_A and Q_B using Eq. (29)
 - 8: $\epsilon_A \leftarrow \max[\epsilon_{\min}, (1 - \delta_\epsilon)\epsilon_A]$
 - 9: $\epsilon_B \leftarrow \max[\epsilon_{\min}, (1 - \delta_\epsilon)\epsilon_B]$
 - 10: **if** $t \% T_{\text{ep}} = 0$ **then**
 - 11: truncate the episode and reset the environment by randomly selecting (a_t, b_t)
 - 12: **end if**
 - 13: **end for**
-

Appendix B. Optimal strategy against fixed strategy

To gain insights into the stability of a strategy, it is common practice to consider an agent's learning dynamics while the opponent's strategy is fixed

[12, 13]. Therefore, we briefly discuss agent A's learning when agent B's strategy is fixed. Assuming that agent B follows a deterministic policy π_B , the Bellman optimality equation for agent A is given by Eq. (2) with

$$\tilde{b}_1 = \pi_B(a_0, b_0), \quad (\text{B.1})$$

where $\pi_B(a, b)$ denotes agent B's action when the previous action pair (a, b) is given.

To represent agent B's strategy, we use the following vector:

$$(\pi_B(C, C), \pi_B(C, D), \pi_B(D, C), \pi_B(D, D)). \quad (\text{B.2})$$

For example, the WSLS strategy is expressed as (C, D, D, C) . First, we discuss the case in which agent B takes the strategies that generate the symmetric solutions discussed in Sec. 3.2, WSLS, FH, and All-C.

Appendix B.1. Optimal strategy against the WSLS strategy (The case (C, D, D, C))

When agent B's strategy is fixed as the WSLS strategy, the Bellman optimality equation is expressed as follows:

$$\begin{aligned} Q_A^*((C, C), C) &= Q_A^*((D, D), C) = R + \gamma \max_{a_2} Q_A^*((C, C), a_2), \\ Q_A^*((C, C), D) &= Q_A^*((D, D), D) = T + \gamma \max_{a_2} Q_A^*((D, C), a_2), \\ Q_A^*((C, D), C) &= Q_A^*((D, C), C) = S + \gamma \max_{a_2} Q_A^*((C, D), a_2), \\ Q_A^*((C, D), D) &= Q_A^*((D, C), D) = P + \gamma \max_{a_2} Q_A^*((D, D), a_2). \end{aligned} \quad (\text{B.3})$$

Using Eq. (19), Eq. (B.3) is transformed into

$$\begin{aligned} x_1 &= R + \gamma \max(x_1, x_2), \\ x_2 &= T + \gamma \max(x_3, x_4), \\ x_3 &= S + \gamma \max(x_3, x_4), \\ x_4 &= P + \gamma \max(x_1, x_2). \end{aligned} \quad (\text{B.4})$$

By solving these equations, we obtain the same form as Eq. (21) with $x_1 > x_2$ and $x_3 < x_4$ when $\gamma > \frac{S-P}{R-P}$, and

$$\begin{aligned} x_1 &= Q_A^*((C, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R, \\ x_2 &= Q_A^*((C, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma}S, \\ x_3 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = \frac{1}{1-\gamma}S, \\ x_4 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = P + \frac{\gamma}{1-\gamma}R. \end{aligned} \quad (\text{B.5})$$

with $x_1 > x_2$ and $x_3 > x_4$ when $\gamma < \frac{S-P}{R-P}$. Seeing the ordering of the Q-function values, the former and latter solutions correspond to the WSLS and All-C strategies, respectively.

Appendix B.2. Optimal strategy against the FH strategy (The case (D, D, D, C))

In this case, the Bellman optimality equation is expressed as

$$\begin{aligned}
Q_A^*((D, D), C) &= R + \gamma \max_{a_2} Q_A^*((C, C), a_2), \\
Q_A^*((D, D), D) &= T + \gamma \max_{a_2} Q_A^*((D, C), a_2), \\
Q_A^*((C, C), C) &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = S + \gamma \max_{a_2} Q_A^*((C, D), a_2), \\
Q_A^*((C, C), D) &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = P + \gamma \max_{a_2} Q_A^*((D, D), a_2).
\end{aligned} \tag{B.6}$$

By solving these equations, we obtain the same form as Eq. (23) with $x_1 > x_2$ and $x_3 < x_4$ when $\gamma > \frac{S-P}{R-S}$ and

$$\begin{aligned}
x_1 &= Q_A^*((D, D), C) = R + \frac{\gamma}{1-\gamma} S, \\
x_2 &= Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma} S, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, C), C) = \frac{1}{1-\gamma} S, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, C), D) = P + \gamma R + \frac{\gamma^2}{1-\gamma} S,
\end{aligned} \tag{B.7}$$

with $x_1 > x_2$ and $x_3 > x_4$ when $\gamma < \frac{S-P}{R-S}$. The former and latter solutions correspond to the FH and All-C strategies, respectively.

Appendix B.3. Optimal strategy against the All-C strategy (The case (C, C, C, C))

In this case, by solving the Bellman optimality equation, as in the two cases above, we obtain the same form as Eq. (20), with $x_1 > x_2$. This solution corresponds to the All-C strategy. In other words, when agent B adopts the All-C strategy, the optimal strategy for agent A is also the All-C one.

Note that, for All-C, WSLs, and FH, the condition under which the optimal strategy against each strategy coincides with the respective strategy is the same as that under which the symmetric solution discussed in Sec. 3.2 emerges. This implies that the strategy providing the symmetric solution represents a Nash equilibrium, in which neither of the two agents can unilaterally change their strategy to increase their cumulative reward, when the strategy set is restricted to memory-1 strategies.

Appendix B.4. Optimal strategies against other fixed strategies

We also consider the case in which the strategy of agent B is fixed as one of 13 (deterministic) strategies that do not correspond to symmetric solutions.

Appendix B.4.1. The case (C, C, C, D)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, C), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, C), D) = T + \frac{\gamma}{1-\gamma}R, \\
x_3 &= Q_A^*((D, D), C) = S + \frac{\gamma}{1-\gamma}R, \\
x_4 &= Q_A^*((D, D), D) = P + \gamma S + \frac{\gamma^2}{1-\gamma}R,
\end{aligned} \tag{B.8}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.2. The case (C, C, D, C)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, D), D) = T + \gamma S + \frac{\gamma^2}{1-\gamma}R, \\
x_3 &= Q_A^*((D, C), C) = S + \frac{\gamma}{1-\gamma}R, \\
x_4 &= Q_A^*((D, C), D) = P + \frac{\gamma}{1-\gamma}R,
\end{aligned} \tag{B.9}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.3. The case (C, D, C, C)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma}R, \\
x_3 &= Q_A^*((C, D), C) = \frac{1}{1-\gamma}S, \\
x_4 &= Q_A^*((C, D), D) = P + \frac{\gamma}{1-\gamma}R,
\end{aligned} \tag{B.10}$$

with $x_1 > x_2$ and $x_3 > x_4$, when $\gamma < \frac{S-P}{R-P}$, and

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma}R, \\
x_3 &= Q_A^*((C, D), C) = S + \gamma P + \frac{\gamma^2}{1-\gamma}R, \\
x_4 &= Q_A^*((C, D), D) = P + \frac{\gamma}{1-\gamma}R,
\end{aligned} \tag{B.11}$$

with $x_1 > x_2$ and $x_3 < x_4$, when $\gamma > \frac{S-P}{R-P}$. The former and latter solutions correspond to the All-C and (C, C, D, C) strategies.

Appendix B.4.4. The case (D, C, C, C)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma^2}(R + \gamma S), \\
x_2 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1-\gamma^2}(R + \gamma S), \\
x_3 &= Q_A^*((C, C), C) = \frac{1}{1-\gamma^2}(S + \gamma R), \\
x_4 &= Q_A^*((C, C), D) = P + \frac{\gamma}{1-\gamma^2}(R + \gamma S),
\end{aligned} \tag{B.12}$$

with $x_1 > x_2$ and $x_3 > x_4$, when $\gamma(T - S) < R - T$, and

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = R + \gamma S + \frac{\gamma^2}{1-\gamma}T, \\
x_2 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = \frac{1}{1-\gamma}T, \\
x_3 &= Q_A^*((C, C), C) = S + \frac{\gamma}{1-\gamma}T, \\
x_4 &= Q_A^*((C, C), D) = P + \frac{\gamma}{1-\gamma}T,
\end{aligned} \tag{B.13}$$

with $x_1 < x_2$ and $x_3 > x_4$, when $\gamma(T - S) > R - T$. The former and latter solutions correspond to the All-C and Grim strategies, respectively. Here, following Refs. [10, 12, 13], we use the term ‘‘Grim strategy’’ to refer to the 1-memory strategy that chooses C if and only if the previous joint action is (C, C) , i.e., the (C, D, D, D) strategy. This terminology differs from the conventional Grim Trigger strategy in which a deviation triggers permanent punishment [22, 23].

Appendix B.4.5. The case (C, C, D, D)

In this case, agent B responds with the Tit-for-Tat (TFT) strategy [24]. By solving the Bellman optimality equation, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = T + \gamma S + \frac{\gamma^2}{1-\gamma}R, \\
x_3 &= Q_A^*((D, C), C) = Q_A^*((D, D), C) = S + \frac{\gamma}{1-\gamma}R, \\
x_4 &= Q_A^*((D, C), D) = Q_A^*((D, D), D) = P + \gamma S + \frac{\gamma^2}{1-\gamma}R,
\end{aligned} \tag{B.14}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.6. The case (C, D, C, D)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = Q_A^*((D, C), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = Q_A^*((D, C), D) = T + \frac{\gamma}{1-\gamma}R, \\
x_3 &= Q_A^*((C, D), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}S, \\
x_4 &= Q_A^*((C, D), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma}S,
\end{aligned} \tag{B.15}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.7. The case (D, C, C, D)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = \frac{1}{1-\gamma^2}(R + \gamma S), \\
x_2 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = T + \frac{\gamma}{1-\gamma^2}(R + \gamma S), \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma^2}(S + \gamma R), \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma^2}(S + \gamma R),
\end{aligned} \tag{B.16}$$

with $x_1 > x_2$ and $x_3 > x_4$, when $R - T > \gamma(T - S)$, and

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = R + \gamma S + \frac{\gamma^2}{1 - \gamma} T, \\
x_2 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = \frac{1}{1 - \gamma} T, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((D, D), C) = S + \frac{\gamma}{1 - \gamma} T, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((D, D), D) = P + \gamma S + \frac{\gamma^2}{1 - \gamma} T,
\end{aligned} \tag{B.17}$$

with $x_1 < x_2$ and $x_3 > x_4$, when $R - T < \gamma(T - S)$. The former and latter solutions correspond to the All-C and WSLs strategies.

Appendix B.4.8. The case (D, C, D, C)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = Q_A^*((D, D), C) = \frac{1}{1 - \gamma^2} (R + \gamma S), \\
x_2 &= Q_A^*((C, D), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1 - \gamma^2} (S + \gamma R), \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((D, C), C) = \frac{1}{1 - \gamma^2} (S + \gamma R), \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((D, C), D) = P + \frac{\gamma}{1 - \gamma^2} (R + \gamma S),
\end{aligned} \tag{B.18}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.9. The case (D, D, C, C)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((D, C), C) = Q_A^*((D, D), C) = R + \frac{\gamma}{1 - \gamma} S, \\
x_2 &= Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \gamma R + \frac{\gamma^2}{1 - \gamma} S, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = \frac{1}{1 - \gamma} S, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = P + \gamma R + \frac{\gamma^2}{1 - \gamma} S,
\end{aligned} \tag{B.19}$$

with $x_1 > x_2$ and $x_3 > x_4$, when $T - \gamma S < (1 - \gamma)R$ and $S > \frac{P + \gamma R}{1 + \gamma}$,

$$\begin{aligned}
x_1 &= Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1 - \gamma^2}(R + \gamma P), \\
x_2 &= Q_A^*((D, C), D) = Q_A^*((D, D), D) = T + \frac{\gamma}{1 - \gamma^2}(R + \gamma P), \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = S + \frac{\gamma}{1 - \gamma^2}(P + \gamma R), \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = \frac{1}{1 - \gamma^2}(P + \gamma R),
\end{aligned} \tag{B.20}$$

with $x_1 > x_2$ and $x_3 < x_4$, when $T < \frac{R + \gamma P}{1 + \gamma}$ and $S < \frac{P + \gamma R}{1 + \gamma}$,

$$\begin{aligned}
x_1 &= Q_A^*((D, C), C) = Q_A^*((D, D), C) = R + \frac{\gamma}{1 - \gamma}S, \\
x_2 &= Q_A^*((D, C), D) = Q_A^*((D, D), D) = \frac{1}{1 - \gamma}T, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = \frac{1}{1 - \gamma}S, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = P + \frac{\gamma}{1 - \gamma}T,
\end{aligned} \tag{B.21}$$

with $x_1 < x_2$ and $x_3 > x_4$, when $T - \gamma S > (1 - \gamma)R$ and $S - \gamma T > (1 - \gamma)P$, and

$$\begin{aligned}
x_1 &= Q_A^*((D, C), C) = Q_A^*((D, D), C) = R + \gamma P + \frac{\gamma^2}{1 - \gamma}T, \\
x_2 &= Q_A^*((D, C), D) = Q_A^*((D, D), D) = \frac{1}{1 - \gamma}T, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = S + \gamma P + \frac{\gamma^2}{1 - \gamma}T, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = P + \frac{\gamma}{1 - \gamma}T,
\end{aligned} \tag{B.22}$$

with $x_1 < x_2$ and $x_3 < x_4$, when $T > \frac{R + \gamma P}{1 + \gamma}$ and $S - \gamma T < (1 - \gamma)P$. These solutions correspond to the All-C, (D, C, D, C) , (C, D, C, D) , and All-D strategies.

Appendix B.4.10. The case (C, D, D, D)

In this case, agent B's strategy corresponds to the Grim one. Solving the Bellman optimality equation, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, C), C) = \frac{1}{1-\gamma}R, \\
x_2 &= Q_A^*((C, C), D) = T + \frac{\gamma}{1-\gamma}S, \\
x_3 &= Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}S, \\
x_4 &= Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma}S,
\end{aligned} \tag{B.23}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.11. The case (D, C, D, D)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((C, D), C) = \frac{1}{1-\gamma^2}(R + \gamma S), \\
x_2 &= Q_A^*((C, D), D) = T + \frac{\gamma}{1-\gamma^2}(S + \gamma R), \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma^2}(S + \gamma R), \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma^2}(S + \gamma R),
\end{aligned} \tag{B.24}$$

with $x_1 > x_2$ and $x_3 > x_4$. This solution corresponds to the All-C strategy.

Appendix B.4.12. The case (D, D, C, D)

In this case, we obtain

$$\begin{aligned}
x_1 &= Q_A^*((D, C), C) = R + \frac{\gamma}{1-\gamma}S, \\
x_2 &= Q_A^*((D, C), D) = T + \gamma R + \frac{\gamma^2}{1-\gamma}S, \\
x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}S, \\
x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma}S,
\end{aligned} \tag{B.25}$$

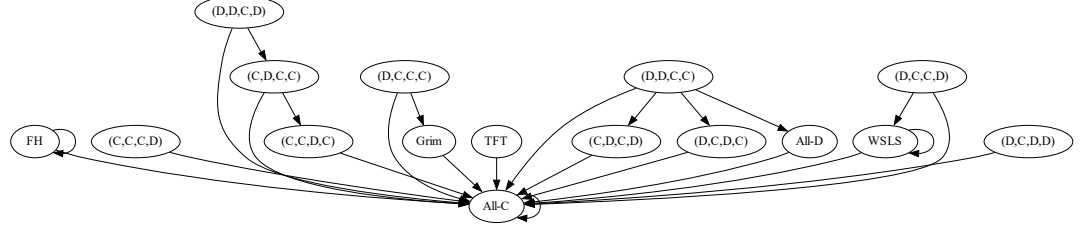


Figure B.2: The best-response network of the Harmony game when the agents' strategies are deterministic.

with $x_1 > x_2$ and $x_3 > x_4$, when $\gamma < \frac{R-T}{R-S}$, and

$$\begin{aligned}
 x_1 &= Q_A^*((D, C), C) = R + \frac{\gamma}{1-\gamma}S, \\
 x_2 &= Q_A^*((D, C), D) = \frac{1}{1-\gamma}T, \\
 x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}S, \\
 x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma}S,
 \end{aligned} \tag{B.26}$$

with $x_1 < x_2$ and $x_3 > x_4$, when $\gamma > \frac{R-T}{R-S}$. The former and latter solutions correspond to the All-C and (C, D, C, C) strategies.

Appendix B.4.13. The case (D, D, D, D) (the All-D strategy)

In this case, we obtain

$$\begin{aligned}
 x_3 &= Q_A^*((C, C), C) = Q_A^*((C, D), C) = Q_A^*((D, C), C) = Q_A^*((D, D), C) = \frac{1}{1-\gamma}S, \\
 x_4 &= Q_A^*((C, C), D) = Q_A^*((C, D), D) = Q_A^*((D, C), D) = Q_A^*((D, D), D) = P + \frac{\gamma}{1-\gamma}S,
 \end{aligned} \tag{B.27}$$

with $x_3 > x_4$. This solution corresponds to the All-C strategy.

In light of the above discussion, we can draw a directed graph in which the edges go from each strategy to its optimal response, as shown in Fig. B.2. This graph is referred to as the best-response network in Ref. [12]. Fig. B.2 shows that the graph contains three absorbing states corresponding to the symmetric solution of Sec. 3.2, that is, the All-C, WSLs, and FH strategies. Therefore, if the agents update their strategies by taking the best response to the present opponent, as in Ref. [12], their strategies converge to one of these three or to the cycle among them.

References

- [1] E. Fehr, S. Gächter, Cooperation and punishment in public goods experiments, *Am. Econ. Rev.* **90**(4) (2000) 980–994. <https://doi.org/10.1257/aer.90.4.980>
- [2] R. Boyd, H. Gintis, S. Bowles, P. J. Richerson, The evolution of altruistic punishment, *Proc. Natl. Acad. Sci. U.S.A.* **100**(6) (2003) 3531–3535. <https://doi.org/10.1073/pnas.0630443100>
- [3] E. Fehr, U. Fischbacher, Third-party punishment and social norms, *Evol. Hum. Behav.* **25**(2) (2004) 63–87. [https://doi.org/10.1016/S1090-5138\(04\)00005-4](https://doi.org/10.1016/S1090-5138(04)00005-4)
- [4] E. W. de Kwaadsteniet, S. A. M. Rijkhoff, E. van Dijk, Equality as a benchmark for third-party punishment and reward: The moderating role of uncertainty in social dilemmas, *Organ. Behav. Hum. Decis. Proc.* **120**(2) (2013) 251–259. <https://doi.org/10.1016/j.obhdp.2012.06.007>
- [5] J. Yang, A. Li, M. Farajtabar, P. Sunehag, E. Hughes, H. Zha, Learning to Incentivize Other Learning Agents, In *Advances in Neural Information Processing Systems* **33** (NeurIPS 2020) (2020)
- [6] C. Zhao, G. Zheng, C. Zhang, J. Zhang, L. Chen, Emergence of cooperation under punishment: A reinforcement learning perspective, *Chaos* **34** (2024) 073123. <https://doi.org/10.1063/5.0215702>
- [7] T. W. Sandholm, R. H. Crites, Multiagent reinforcement learning in the Iterated Prisoner’s Dilemma, *Biosystems* **37** (1996) 147–166. [https://doi.org/10.1016/0303-2647\(95\)01551-5](https://doi.org/10.1016/0303-2647(95)01551-5)
- [8] J. Foerster, R. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, I. Mordatch, Learning with opponent-learning awareness, In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, (2018) 122 - 130. <https://dl.acm.org/doi/abs/10.5555/3237383.3237408>
- [9] Y. Hu, C. Han, H. Li, T. Guo, Modeling opponent learning in multiagent repeated games, *Appl. Intell.* **53**(13) (2023) 17194–17210. <https://doi.org/10.1007/s10489-022-04249-x>
- [10] Y. Usui, M. Ueda, Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner’s dilemma, *Appl. Math. Comput.* **409** (2021) 126370. <https://doi.org/10.1016/j.amc.2021.126370>
- [11] M. Nowak, K. Sigmund A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner’s Dilemma game, *Nature* **364** (1993) 56–58. <https://doi.org/10.1038/364056a0>

- [12] J. M. Meylahn, L. Janssen, Limiting dynamics for Q-learning with memory one in symmetric two-player, two-action Games, *Complexity* **2022**(1) (2022) 4830491. <https://doi.org/10.1155/2022/4830491>
- [13] J. M. Meylahn, Quantifying the likelihood of learning collusive strategy equilibria, *Chaos* **35** (2025) 083106. <https://doi.org/10.1063/5.0281443>
- [14] C. P. Roca, J. A. Cuesta, and A. Sánchez Time Scales in Evolutionary Dynamics, *Phys. Rev. Lett.* **97** (2006) 158701. <https://doi.org/10.1103/PhysRevLett.97.158701>
- [15] M. C. Couto, F. P. Santos, and C. Hilbe, Evolution of noisy learning in games, *Proc. Natl. Acad. Sci. U.S.A.* **123**(20) (2026) e2529959123. <https://doi.org/10.1073/pnas.2529959123>
- [16] J. T. Matamalas, J. Poncela-Casasnovas, S. Gómez, and A. Arenas, Strategic incoherence regulates cooperation in social dilemmas on multiplex networks, *Sci. Rep.* **5** (2015) 9519. <https://doi.org/10.1038/srep09519>
- [17] D. Fudenberg, and E. Maskin The folk theorem in repeated games with discounting or with incomplete information, *Econometrica* **54**(3) (1986) 533-554. <https://doi.org/10.2307/1911307>
- [18] J. Li, P. Hingston, and G. Kendall, Engineering design of strategies for winning iterated prisoner’s dilemma competitions, *IEEE Trans. Comput. Intell. AI Games* **3**(4) (2011) 348-360. <https://doi.org/10.1109/TCIAIG.2011.2166268>
- [19] V. Knight, M. Harper, N. E. Glynatsi, and O. Campbell, Evolution reinforces cooperation with the emergence of self-recognition mechanisms: An empirical study of strategies in the Moran process for the iterated prisoner’s dilemma, *PLoS One*. **13**(10) (2018) e0204981. <https://doi.org/10.1371/journal.pone.0204981>
- [20] C. J. C. H. Watkins and P. Dayan, Q-learning, *Mach. Learn.* **8**(3) (1992) 279-292. <https://doi.org/10.1007/BF00992698>
- [21] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, second edition, MIT press, (2018).
- [22] J. E. Harrington, Jr. The Determination of Price and Output Quotas in a Heterogeneous Cartel, *Int. Econ. Rev.* **32**(4) (1991) 767-792. <https://doi.org/10.2307/2527033>
- [23] F. McGillivray and A. Smith. Trust and cooperation through agent-specific punishments, *Int. Organ.* **54**(4) (2000) 809-824. <https://doi.org/10.1162/002081800551370>
- [24] R. Axelrod, and W. D. Hamilton, The evolution of cooperation, *Science* **211**(4489) (1981) 1390-1396. <https://doi.org/10.1126/science.7466396>