Emergence of exploitation as symmetry breaking in iterated prisoner's dilemma

Yuma Fujimoto[1,*] and Kunihiko Kaneko[1,2,†]

[1]*Department of Basic Science, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan*

[2]*Research Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo, 3-8-1 Komaba, Tokyo 153-8902, Japan*

(Received 24 June 2019; published 5 November 2019)

In society, mutual cooperation, defection, and exploitative relationships are common. Whereas cooperation and defection are studied extensively in the literature on game theory, exploitative relationships between players, in which one receives a larger benefit than the other while the game itself is symmetric, are little explored. In a recent seminal study, Press and Dyson demonstrated that if only one player can learn about the other, asymmetric exploitation is achieved in the prisoner's dilemma game. In their study, however, asymmetry is assumed in decision making between persons; the exploiting player one-sidedly determines and fixes the strategy and the exploited player follows it. It is unknown whether such exploitation emerges and is stably established even when both players learn about each other symmetrically and try to optimize their payoffs. Here, we first formulate a dynamical system that describes the change in a player's probabilistic strategy with reinforcement learning to obtain greater payoffs, based on the recognition of the other player. By applying this formulation to the standard prisoner's dilemma game, we numerically and analytically demonstrate that an exploitative relationship can be achieved despite symmetric strategy dynamics and symmetric rule of games. This exploitative relationship is stabilized by both the players: The exploiting player demands the other's unfair cooperation. Even though the exploited player, who receives a lower payoff than the exploiting player, has optimized the own strategy, the player accepts the other's defection to some degree. Whether the final equilibrium state is mutual cooperation, defection, or exploitation crucially depends on the initial conditions. Response to decrease the cooperation probability against a defector leads to oscillations in the probabilities of cooperation between the players and thus a complicated basin structure to the final equilibrium. In particular, any slight difference between both players' initial strategies can be amplified and fixed as a large difference in the probabilities of cooperation, leading to fixation of exploitation. In other words, symmetry breaking between the exploiting and exploited players results. Considering the generality of the result, this study provides another perspective on the origin of exploitation in society.

DOI: 10.1103/PhysRevResearch.1.033077

I. INTRODUCTION

Equality is not easily achieved in society; instead, inequality among individuals is common. Exploitative behavior, in which one individual receives a greater benefit at the expense of others receiving lower benefits, is frequently observed. Of course, such exploitation can originate from *a priori* differences in individual capacities or environmental conditions. However, such exploitation is also developed and sustained historically. Even when inherent individual capacities or environmental conditions are not different, and even when individuals are able to choose other actions to escape exploitation and optimize their benefits, exploitation somehow remains.

In this study, we consider how such exploitation emerges and is sustained. Of course addressing this question completely is too difficult, as the answer may involve economics, sociology, history, and so forth. Instead, we simplify the problem by adopting a game theoretic framework and investigate whether exploitative behavior can emerge *a posteriori* as a result of dynamics in individuals' cognitive structures. By using a symmetric game in which both players have an identical payoff matrix, the exploitation is defined as a state in which one player chooses an action to accept a lower score than the other, even though the former player can potentially recover the symmetry and receive the same payoff as the other. With the change in strategies of the players through learning, we check whether "symmetry breaking" can occur when individuals have symmetric capacities and environmental conditions.

For this analysis, we adopt the celebrated prisoner's dilemma game. In this game, both players can independently choose cooperation or defection. Regardless of the other player's choice, defection is more beneficial than cooperation, but the payoff when both players defect is lower than that when both players cooperate. In the prisoner's dilemma game, the emergence and sustainability of cooperation, even though defection is any individual player's best choice, has been extensively investigated [1,2].

*yfujimoto@complex.c.u-tokyo.ac.jp

†kaneko@complex.c.u-tokyo.ac.jp

Cooperation can indeed emerge in repeated games in which each player chooses his or her own action (cooperation or defection) depending on the other's previous actions. In other words, a cooperative relationship emerges with the potential for punishment. Players cooperate conditionally with cooperators and defect against defectors [e.g., by a tit-for-tat (TFT) strategy]. In evolutionary games, cooperation is known to stably emerge from the introduction of a "space structure" [3], "hierarchical structure" [4], "stochastic transition of rule" [5], and so forth, in which a certain punishment mechanism against defection is commonly adopted.

In contrast to the intensive and extensive studies on cooperative relationships, however, studies on exploitative relationships (i.e., asymmetric cooperation between two players) are very much limited. In this game, an exploitative relationship is represented by unequal cooperation probabilities between the players, as a defector can get higher benefit at the expense of a cooperator. A recent study proposes zero-determinant strategies [6], classified as one-memory strategies, in which one player stochastically determines whether to cooperate or defect depending on the condition on the previous actions of both players. If a player one-sidedly adopts and fixes the zero-determinant strategy while the other accordingly optimizes his or her own strategy, the former player can exploit the latter. Here, however, the study focuses only on one-way learning. Hence, the two players have different abilities in the beginning. Thus, whether reciprocal optimization between two symmetric players can generate an exploitative relationship remains unresolved. Indeed, in the evolution of strategies with zero-determinant ones, the exploitative relationship does not last forever; rather, cooperation or generosity is promoted [7–11].

To consider the possibility of exploitation, we take a learning process into account here. Indeed, the coupled replicator model was introduced in game theory for reciprocal changes in strategies by learning [12–14]. Such models use a deterministic reinforcement learning process in which every player has a probability distribution that provides a probabilistic strategy for taking actions. Each player updates his or her strategy based on the experience of repeated games. In other words, a probability to take a successful action is reinforced. If the other player's strategy is fixed, a player deterministically increases his or her own payoff throughout the repeated game. There are several models for the reinforcement learning based on the experience of game, like fictitious play [15,16] and Q learning [17]. As far as we know, however, such previous models have no memories in the player's strategy, and thus the exploitation does not emerge. Here, we seek the possibility of whether tiny differences in the initial strategies between the players are amplified through learning, and symmetry breaking in the strategies and accordingly in the scores between the players results, so that one player keeps on getting a higher score than the other.

In the next section, we extend the coupled replicator model to take the reference to the other's previous action so that the player conditionally determines his or her action depending on it. In fact, such a strategy with the reference of other's action is justified by an ability to make a model on the other [19–21]. We formulate the learning dynamics of strategies accordingly and apply it into the prisoner's dilemma. In Sec. III, we demonstrate that an exploitative relationship emerges by this extended coupled replicator. Then, we analyze how the exploitation is stabilized, even though both the players optimize their own strategies. We uncover the condition in which exploitation emerges in the symmetric prisoner's dilemma game between the players under symmetric learning dynamics. In Sec. IV, we then investigate the dynamics of the learning process in depth to demonstrate that a slight difference in initial strategies between the players is amplified, and this symmetry breaking results in a large payoff difference, i.e., exploitation.

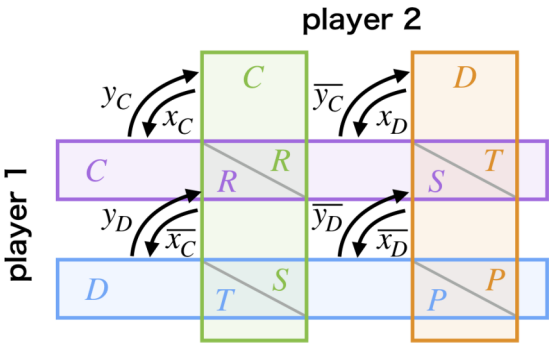

FIG. 1. Schematic diagram for the prisoner's dilemma game and the strategy. Player 1 (horizontal) and 2 (vertical) independently choose their own actions from C and D. The resultant payoffs T, R, P, and S are displayed. The black arrows indicate the stochastic transition from the previous action to the next one, by using the probabilistic strategy in the text.

II. MODEL

We study the well-known prisoner's dilemma (PD) game (see Fig. 1 for the payoff matrix), in which each of two players, referred to as players 1 and 2, chooses to cooperate (C) or defect (D). Thus, a game involves one of four possible actions, CC, CD, DC, and DD, where the left (right) index shows player 1's (2's) choice. For actions CC, CD, DC, and DD, player 1's score is given by R, S, T, and P, respectively. In the PD game, defection is more beneficial regardless of the other player's action, meaning that both $T > R$ and $P > S$ hold. In addition, mutual cooperation (CC) is more beneficial than the mutual defection (DD), meaning that $R > P$ holds. A repeated game requires the additional condition that $2R > T + S$. In other words, sequential cooperation (i.e., always choosing CC) is more beneficial than reciprocal defection and cooperation (i.e., repeatedly alternating between CD and DC).

The payoff (Fig. 1) between the players is different only for the case with asymmetric actions, CD and DC. If DC is more frequently achieved than CD, player 1 gains T more frequently than S. Thus, in such a state, player 2 is exploited by player 1. Indeed, in the original notation in the prisoner's dilemma game, T stands for the "temptation" to exploit the other and S stands for the exploited "sucker" payoff.

We next define a class of strategy (see Fig. 1), in which one player stochastically determines whether to choose C or D based on the other player's action in the previous round. Player 1's strategy is given by two variables that represent the probabilities of cooperation in the next round, x_C and x_D, when

player 2 was previously a cooperator or defector, respectively. Conversely, $\overline{x_C} := 1 - x_C$ ($\overline{x_D} := 1 - x_D$) indicates the probability that player 1's present action is D when the other's previous action is C (D). Throughout this study, we use the definition $\overline{X} := 1 - X$. Similarly, player 2's strategy is given by y_C and y_D. These strategies include several well-known strategies, All-D ($x_C = x_D = 0$), All-C ($x_C = x_D = 1$), and TFT ($x_C = 1, x_D = 0$), as extreme cases.

A. Repeated game for fixed strategies

Before considering the dynamics of each player's strategy, we consider each player's resulting action and payoff when the strategies (i.e., x_C, x_D, y_C, and y_D) are fixed. We assume that (CC, CD, DC, DD) is played with probability $\mathbf{p} := (p_{CC}, p_{CD}, p_{DC}, p_{DD})^{\mathrm{T}}$ in the previous period. Then, the probabilities of the occurrence of (CC, CD, DC, DD) in the next round are obtained by operating the 4×4 Markov matrix $\mathbf{M}$, which is given by

$$\mathbf{M} := \begin{pmatrix} x_C\, y_C & x_D\, y_C & x_C\, y_D & x_D\, y_D \\ x_C\, \overline{y_C} & x_D\, \overline{y_C} & x_C\, \overline{y_D} & x_D\, \overline{y_D} \\ \overline{x_C}\, y_C & \overline{x_D}\, y_C & \overline{x_C}\, y_D & \overline{x_D}\, y_D \\ \overline{x_C}\, \overline{y_C} & \overline{x_D}\, \overline{y_C} & \overline{x_C}\, \overline{y_D} & \overline{x_D}\, \overline{y_D} \end{pmatrix}. \tag{1}$$

For a given fixed (x_C, x_D, y_C, y_D), the probability is updated as $\mathbf{p}' = \mathbf{M}\mathbf{p}$. Thus, after a sufficient number of iterated games, the probabilities converge to an equilibrium, $\mathbf{p}_{\mathrm{e}}$. Here, this equilibrium state is uniquely defined at least when $0 < x_C, x_D, y_C, y_D < 1$ is satisfied by the full connectivity of $\mathbf{M}$. The equilibrium state, $\mathbf{p}_{\mathrm{e}}$, is represented as the eigenvector of the above matrix corresponding to the 1-eigenvalue, which is written with only two variables, x_{e} and y_{e}, as

$$\mathbf{p}_{\mathrm{e}} = (x_{\mathrm{e}} y_{\mathrm{e}}, x_{\mathrm{e}} \overline{y_{\mathrm{e}}}, \overline{x_{\mathrm{e}}} y_{\mathrm{e}}, \overline{x_{\mathrm{e}}}\, \overline{y_{\mathrm{e}}})^{\mathrm{T}} \tag{2}$$

(see the Supplemental Material [22] for the derivation). Here, note that each player unconditionally cooperates with probabilities x_{e} and y_{e} in the equilibrium state, which are given by

$$x_{\mathrm{e}}(x_C, x_D, y_C, y_D) = \frac{x_D + (x_C - x_D) y_D}{1 - (x_C - x_D)(y_C - y_D)},$$
$$y_{\mathrm{e}}(x_C, x_D, y_C, y_D) = \frac{y_D + (y_C - y_D) x_D}{1 - (x_C - x_D)(y_C - y_D)}. \tag{3}$$

At the equilibrium state, the payoff of player 1 (2), denoted by u_{e} (v_{e}), is given by

$$u_{\mathrm{e}}(x_C, x_D, y_C, y_D) = \mathbf{p}_{\mathrm{e}} \cdot (R, S, T, P)^{\mathrm{T}},$$
$$v_{\mathrm{e}}(x_C, x_D, y_C, y_D) = \mathbf{p}_{\mathrm{e}} \cdot (R, T, S, P)^{\mathrm{T}}. \tag{4}$$

We emphasize that the equilibrium state for a repeated game is denoted by the subscript e, but it is unrelated to the equilibrium of learning dynamics discussed in the following subsection.

B. Learning dynamics of strategies

Next, we consider the dynamic changes in strategies created by a reinforcement learning process. During a repeated game, every player takes actions following his or her own strategy and updates the probability of each action to gain a higher payoff. Here, we assume that the strategy updates occur much more slowly than the repetition of games does. Under this assumption, every player can accurately evaluate the benefit gained by a single action and update his or her own strategy to increase his or her payoff under an assumption that the other player's strategy is fixed. Then, depending on the frequency of actions adopted, and according payoffs, the probability to choose each action is updated.

First, we compute the amount of benefits gained by player 1's cooperative action in a single game, which is denoted by u_C: When the player 1 cooperates, the present state is given by $\mathbf{p} = \mathbf{p}_{1C} := (y_{\mathrm{e}}, \overline{y_{\mathrm{e}}}, 0, 0)^{\mathrm{T}}$ in equilibrium, because CC (CD) occurs with probability y_{e} ($\overline{y_{\mathrm{e}}}$) and neither DC nor DD occurs. This cooperation of the player 1 at the present round continues to influence on the benefit for the future rounds, which, however, decays toward zero as $\mathbf{M}^t(\mathbf{p}_{1C} - \mathbf{p}_{\mathrm{e}})$ for larger $t \to \infty$. Then, the total payoff brought by 1's single cooperation is given by

$$u_C := \left\{ \sum_{t=0}^{\infty} \mathbf{M}^t (\mathbf{p}_{1C} - \mathbf{p}_{\mathrm{e}}) + \mathbf{p}_{\mathrm{e}} \right\} (R, S, T, P)^{\mathrm{T}}$$
$$= \sum_{t=0}^{\infty} \mathbf{M}^t (\mathbf{p}_{1C} - \mathbf{p}_{\mathrm{e}}) (R, S, T, P)^{\mathrm{T}} + u_{\mathrm{e}}. \tag{5}$$

In the same way, we obtain the probability of defection of the player 1 $\mathbf{p}_{1D}$ and the resulting payoff u_D as

$$\mathbf{p}_{1D} := (0, 0, y_{\mathrm{e}}, \overline{y_{\mathrm{e}}})^{\mathrm{T}},$$
$$u_D := \left\{ \sum_{t=0}^{\infty} \mathbf{M}^t (\mathbf{p}_{1D} - \mathbf{p}_{\mathrm{e}}) + \mathbf{p}_{\mathrm{e}} \right\} (R, S, T, P)^{\mathrm{T}}. \tag{6}$$

Second, we consider the update of x_C by player 1 based on the above payoffs u_C and u_D. The advantage of cooperation relative to the average is given by $u_C - (x_C u_C + \overline{x_C} u_D)$. Then, x_C increases proportionally. Note that since player 2's previous action and player 1's present action need to be C and C, respectively, the probability of using strategy x_C is given by $y_{\mathrm{e}} x_C$. Then, we obtain the evolution of x_C over time as

$$\dot{x_C} = y_{\mathrm{e}} x_C \{u_C - (x_C u_C + \overline{x_C} u_D)\}$$
$$= x_C \overline{x_C} y_{\mathrm{e}} (u_C - u_D). \tag{7}$$

Here, $(u_C - u_D)$ is given by

$$u_C - u_D = \frac{(y_C - y_D)\{x_{\mathrm{e}}(R - S) + \overline{x_{\mathrm{e}}}(T - P)\} - \{y_{\mathrm{e}}(T - R) + \overline{y_{\mathrm{e}}}(P - S)\}}{1 - (x_C - x_D)(y_C - y_D)} \tag{8}$$

(see the Supplemental Material [22] for a detailed calculation). The dynamics of x_D are similarly obtained as

$$\begin{aligned} \dot{x_D} &= \overline{y_e} x_D \{u_C - (x_D u_C + \overline{x_D} u_D)\} \\ &= x_D \overline{x_D}\, \overline{y_e} (u_C - u_D). \end{aligned} \tag{9}$$

In the same way, the dynamics of player 2's strategy are given by

$$\begin{aligned} \dot{y_C} &= y_C \overline{y_C} x_e (v_C - v_D), \\ \dot{y_D} &= y_D \overline{y_D}\, \overline{x_e} (v_C - v_D), \end{aligned} \tag{10}$$

$$v_C - v_D = \frac{(x_C - x_D)\{y_e(R-S) + \overline{y_e}(T-P)\} - \{x_e(T-R) + \overline{x_e}(P-S)\}}{1 - (x_C - x_D)(y_C - y_D)}. \tag{11}$$

Note that x_e and y_e are also time dependent, because x_e and y_e are given as functions of time-dependent variables (x_C, x_D, y_C, y_D).

The above learning dynamics can be divided into three terms. For example, we focus on the dynamics of x_C, given by Eq. (7). The first term, $x_C \overline{x_C}$, represents frequency-dependent selection. When x_C is close to 0 or 1, evolution proceeds slowly over time because the nondominant strategy rarely appears. Thus, the evolution to this strategy takes a long time under a biased population distribution. The second term, y_e, represents the dependence of the evolutionary speed of x_C upon its frequency of use, because the other player cooperates with the probability y_e in the previous action. The third term, $u_C - u_D$, represents that the change rate of the strategy is proportional to the difference in resultant payoffs by C and D, due to the reinforcement learning.

The learning dynamics extend the previous "coupled replicator model" [12–14] to include memory of the other's previous action. Indeed, in the coupled replicator model, reinforcement learning of conditional strategies is not adopted. The first term, the effect of frequency-dependent selection, is common to this model and previous models. However, the second term, i.e., the effect of conditional time evolution, is not found in the previous studies [12–14]. A term that corresponds to our third term, i.e., the effect of the payoff gap, exists therein, but the computation of the payoff differs. Specifically, in the previous studies, only the payoff in the present period is considered because the deviation from the equilibrium state is completely relaxed by a single game, and no conditional strategies are used. In contrast, in the present model, we need to consider the whole process by which a deviation from the equilibrium state affects future periods over the long term, as is shown in Eqs. (5) and (6).

C. Intuitive interpretation of the model

The above equilibrium state [Eq. (2)] and learning dynamics [Eqs. (7), (9), and (10)] seem complicated at first glance. However, we can intuitively interpret them by employing the concept of the response function [23].

First, we introduce the response function. We consider the situation in which player 2 cooperates with probability y independent of player 1's previous actions. Player 1, with strategies given by x_C and x_D, also becomes an unconditional cooperator with probability $f_x(y) = y(x_C - x_D) + x_D$ (see the Supplemental Material [22] for a detailed calculation). Indeed, against player $y = 1$ (i.e., a pure cooperator), $f_x(1) = x_C$ holds, whereas, against a pure defector, $f_x(0) = x_D$ holds. Since f_x is player 1's probability of cooperating given player 2's probability of cooperating, we call it the "response function," following the previous studies [23].

Second, the equilibrium probabilities of cooperation, x_e and y_e in Eq. (2), are interpreted as the crossing point of both the response functions, as shown in Fig. 2. In other words,

$$\begin{aligned} x_e &= f_x(y_e), \\ y_e &= f_y(x_e) \end{aligned} \tag{12}$$

hold. Indeed, Eq. (12) is equivalent to Eq. (3).

Third, the above learning dynamics [Eqs. (7), (9), and (10)] can be easily written by using the response function (see the Supplemental Material [22] for a detailed calculation). Here, we only focus on Eq. (7) as an example. The second term, y_e, corresponds to the contribution to a change in the crossing point against a change in x_C. Thus, we obtain

$$y_e \propto \frac{\partial x_e}{\partial x_C}. \tag{13}$$

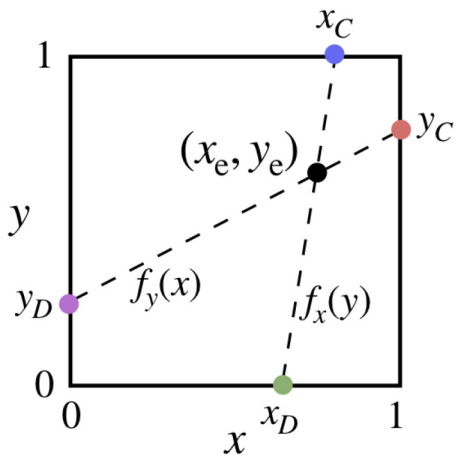

FIG. 2. Interpretation of the equilibrium state for repeated games. The horizontal (vertical) axis indicates the unconditional probability of player 1 (2) to cooperate. Blue, green, red, and magenta dots indicate the strategies x_C, x_D, y_C, and y_D, respectively. Accordingly, response function $f_x(y)$ [$f_y(x)$] is given by connecting x_C (y_C) with x_D (y_D). The crossing point of response functions (black dot) agrees with (x_e, y_e), which is each player's probability to cooperate in the equilibrium of repeated game.

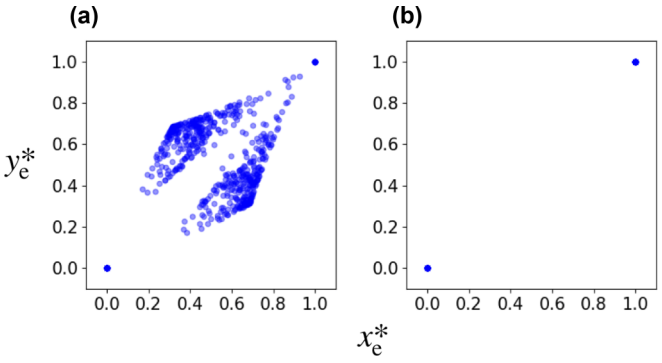

FIG. 3. (a) Final state of learning dynamics in the case of $(T, R, P, S) = (5, 3, 1, 0)$. Many sets of fixed (x_e^*, y_e^*) satisfy $0 \leqslant x_e^*, y_e^* \leqslant 1$ with asymmetry between them. (b) Final state of learning dynamics in the case of $(T, R, P, S) = (5, 4.5, 1, 0)$. Only two sets, $(x_e^*, y_e^*) = (0, 0)$ and (1,1), are reachable. For both the cases, initial states are uniformly chosen as $(x_C^o, x_D^o, y_C^o, y_D^o) = ((2i - 1)/2N, (2j - 1)/2N, (2k - 1)/2N, (2l - 1)/2N)$ with $(i, j, k, l) = 1, \ldots, N$ and $N = 10$. Accordingly, x_e^o and y_e^o can take values in [0,1]. See the Supplemental Material [22] for the animation of each dynamics.

In addition, the third term, $u_C - u_D$, corresponds to the gradient of a player's payoff on the other player's response function. In other words, we obtain

$$u_C - u_D \propto \left.\frac{\partial u(x_e, f_y(x_e))}{\partial x_e}\right|_{y=y_e}. \tag{14}$$

From Eqs. (13) and (14), with canceling the extra components, we can rewrite Eq. (7) as

$$\dot{x_C} = x_C\overline{x_C}\frac{\partial u_e}{\partial x_C}. \tag{15}$$

The same equation holds for the dynamics of x_D, y_C, and y_D. The learning dynamics is interpreted by associating the frequency-dependent selection term, $x_C\overline{x_C}$, and the adaptive learning term, $\partial u_e/\partial x_C$.

III. ANALYSIS OF LEARNING EQUILIBRIUM

Now, we actually simulate the above learning dynamics. Figure 3 shows the final states of (x_e^*, y_e^*) given various initial states $(x_C^o, x_D^o, y_C^o, y_D^o)$. Below, the superscript o $(*)$ denotes the initial (final) value of learning dynamics. Here, instead of directly plotting four-dimensional players' strategies $(x_C^*, x_D^*, y_C^*, y_D^*)$, we plot only their two-dimensional projection to (x_e^*, y_e^*), which is the crossing point generated by their response functions.

From the figure, we see that in the case of $T - R - P + S \leqslant 0$, only (1) pure DD $(x_e^* = y_e^* = 0)$ and (2) pure CC $(x_e^* = y_e^* = 0)$ strategies can be achieved. In the case of $T - R - P + S > 0$, however, (3) the intermediate states $0 < x_e^*, y_e^* < 1$, which include the case of $x_e^* \neq y_e^*$, can also be achieved. We now analyze these fixed points mathematically.

A. Analysis of each fixed point

(1) The pure DD fixed point is given by $y_e^* = x_D^* = x_e^* = y_D^* = 0$, which satisfies $\dot{x_C} = \dot{x_D} = \dot{y_C} = \dot{y_D} = 0$. Here, $x_D^* = y_D^* = 0$ is clearly satisfied from $x_e^* = y_e^* = 0$. Instead, x_C^* and y_C^* are arbitrary. Then, the linear stability analysis shows that the fixed point is stable if $u_C^* - u_D^* \leqslant 0$ and $v_C^* - v_D^* \leqslant 0$ are additionally satisfied. These conditions are equivalent to $x_C^*, y_C^* \leqslant (P - S)/(T - P)$. Thus, the pure DD fixed-point attractor exists on a two-dimensional plane with continuous values of x_C and y_C.

(2) The pure CC fixed point is given by $x_C^* = y_e^* = y_C^* = x_e^* = 1$, which satisfies $\dot{x_C} = \dot{x_D} = \dot{y_C} = \dot{y_D} = 0$. Here, $x_C^* = y_C^* = 1$ is clearly satisfied from $x_e^* = y_e^* = 1$. Instead, x_D^* and y_D^* are arbitrary. Then, the fixed point is linearly stable if $u_C^* - u_D^* \geqslant 0$ and $v_C^* - v_D^* \geqslant 0$. These conditions are equivalent to $x_D^*, y_D^* \leqslant 1 - (T - R)/(R - S)$. Thus, the pure CC fixed-point attractor also exists on a two-dimensional plane in which x_D^* and y_D^* continuously change. Note that $x_C^* - x_D^* \geqslant (T - R)/(R - S)$ and $y_C^* - y_D^* \geqslant (T - R)/(R - S)$ hold, implying that both players sufficient punish the other's defection.

The pure DD and CC states are both well known as Nash equilibrium and as Pareto optimal, respectively. Because the dominance of these states has been extensively studied, their achievements here are not surprising. In these pure states, no exploitation appears, and both players' actions and payoffs are symmetric. Other states on the boundary of actions (such as $x_e = 1, y_e = 0$) cannot be stable fixed points (see the Supplemental Material [22] for details). The only other fixed points are given by the next case.

(3) When both x_e^* and y_e^* are neither 0 or 1, $u_C^* - u_D^* = v_C^* - v_D^* = 0$ should hold to satisfy the fixed-point condition. Then, $\dot{x_C} = \dot{x_D} = \dot{y_C} = \dot{y_D} = 0$ is satisfied. In such cases, the condition of a fixed point for learning dynamics is

$$\begin{aligned} & u_C^* - u_D^* = 0 \\ \Leftrightarrow\ & y_C^* - y_D^* = \frac{y_e^*(T - R) + \overline{y_e^*}(P - S)}{x_e^*(R - S) + \overline{x_e^*}(T - P)}, \\ & v_C^* - v_D^* = 0 \\ \Leftrightarrow\ & x_C^* - x_D^* = \frac{x_e^*(T - R) + \overline{x_e^*}(P - S)}{y_e^*(R - S) + \overline{y_e^*}(T - P)}. \end{aligned} \tag{16}$$

From Eqs. (16), the set of $(x_C^*, x_D^*, y_C^*, y_D^*)$ achieving (x_e^*, y_e^*) is uniquely given by

$$\begin{aligned} x_C^* &= x_e^* + \overline{y_e^*}\frac{x_e^*(T - R) + \overline{x_e^*}(P - S)}{y_e^*(R - S) + \overline{y_e^*}(T - P)}, \\ x_D^* &= x_e^* - y_e^*\frac{x_e^*(T - R) + \overline{x_e^*}(P - S)}{y_e^*(R - S) + \overline{y_e^*}(T - P)}, \\ y_C^* &= y_e^* + \overline{x_e^*}\frac{y_e^*(T - R) + \overline{y_e^*}(P - S)}{x_e^*(R - S) + \overline{x_e^*}(T - P)}, \\ y_D^* &= y_e^* - x_e^*\frac{y_e^*(T - R) + \overline{y_e^*}(P - S)}{x_e^*(R - S) + \overline{x_e^*}(T - P)}. \end{aligned} \tag{17}$$

Note that as long as the two conditions $u_C^* - u_D^* = v_C^* - v_D^* = 0$ are satisfied within the region $0 \leqslant x_C^*, x_D^*, y_C^*, y_D^* \leqslant 1$, the fixed point condition for $(x_C^*, x_D^*, y_C^*, y_D^*)$ is satisfied. Thus, the fixed points for learning dynamics exist again on a two $(= 4 - 2)$–dimensional space. Then, all such fixed points are represented just as two variables (x_e^*, y_e^*). According to Eq. (17), there is a one-to-one correspondence between the

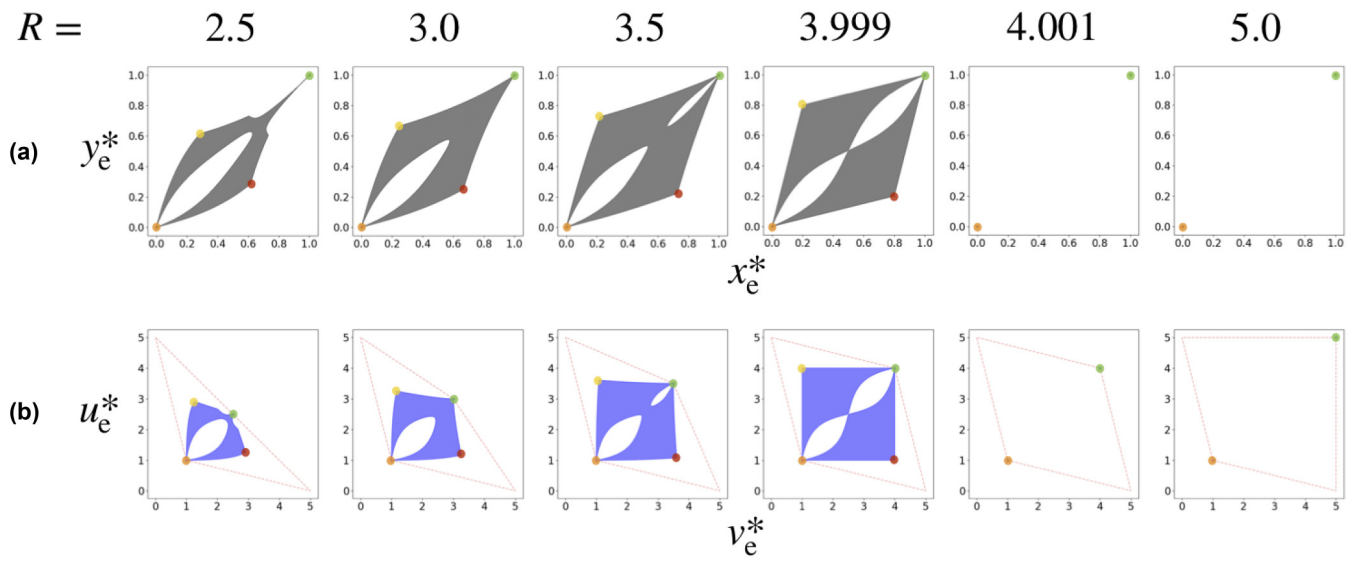

FIG. 4. (a) The region of stable fixed points in which both eigenvalues are negative. (b) Both players payoffs mapped from fixed points. $(T, P, S) = (5, 1, 0)$ are fixed in all figures, and R is 2.5, 3.0, 3.5, 3.999, 4.001 and 5.0 from left to right. In panel (a), the horizontal (vertical) axis indicates x_e^* (y_e^*). In the case of $(T + S)/2 \leqslant R < T - P + S$, in other words, $2.5 \leqslant R < 4.0$, there are two-dimensional fixed points on (x_e^*, y_e^*) with player asymmetry. However, all of the fixed points become unstable in the case of $T - P + S < R \leqslant T$, in other words, $4.0 < R \leqslant 5.0$. In panel (b), the horizontal (vertical) axis indicates u_e^* (v_e^*). Red broken line indicates the possible set of both the payoffs. Orange (green) dot indicates the state of pure DD (CC) in all figures. Yellow (red) dot indicates the most exploitative state from 1 to 2 (from 2 to 1).

four-dimensional strategies of both players $(x_C^*, x_D^*, y_C^*, y_D^*)$ and (x_e^*, y_e^*). Accordingly, we use the plot (x_e^*, y_e^*) in Fig. 3, instead of the four-dimensional space for the fixed points, and will be adapted later.

Although such two-dimensional fixed points exist for all sets of T, R, P, S, not all of them are always reachable from the initial conditions. We further study the stability of the fixed point by performing linear stability analysis around it. Here, we recall that there are only two constraints on the four-dimensional dynamics. Thus, two of four eigenvalues always are zero, and the stability is neutral in two-dimensional space.

Now, we examine the stability by the other two eigenvalues, as seen in Fig. 4(a). The figure shows that in the case of $T - R - P + S \leqslant 0$, none of these novel fixed points has linear stability. Thus, only the symmetric states, pure DD and CC, are achieved by learning dynamics.

In contrast, in the case of $T - R - P + S > 0$, the two-dimensional part of the fixed points satisfies linear stability. For almost all of these points, $x_e^* \neq y_e^*$ holds. Because $x_e^* \neq y_e^*$ is equivalent to the payoff inequality $(u_e^* \neq v_e^*)$, we refer to such states as exploitative relationships in which one player receives more benefit than the other. Such stable two-dimensional exploitation also appears even if the update speeds of the strategies are changed (see the Supplemental Material [22] for the detailed results).

B. Characterization of the exploitative relationship

We now characterize the exploitative state by comparing the payoffs for $T - R - P + S > 0$. Figure 4(b) shows both players' payoffs at the stable fixed points. As the representative state, we especially focus on the most exploitative relationship from 1 to 2 (see yellow dot in Fig. 4), where both $y_e^* - x_e^*$ and $u_e^* - v_e^*$ are maximal. The state is given by

$$
\begin{aligned}
x_e^* &= \frac{P - S}{R + P - 2S}, \quad y_e^* = \frac{T - P}{2T - R - P}, \\
u_e^* &= R + \frac{(T - R)(R - S)(T - R - P + S)}{(R + P - 2S)(2T - R - P)}, \\
v_e^* &= P + \frac{(T - P)(P - S)(T - R - P + S)}{(R + P - 2S)(2T - R - P)}.
\end{aligned} \tag{18}
$$

If $T - R - P + S > 0$, both $u_e^* > R$ and $v_e^* > P$ hold. These equations show that the exploitative relationship is favorable for different reasons between the exploiting and exploited players. First, $u_e^* > R$ means that the exploiting player 1 obtains a higher payoff than that in the case with pure CC. The player 1 could be motivated to defect more, which, however, would increase the probability of player 2's defect and the chance of DD. Hence, player 1 will not increase the probability of defect further. On the other hand, player 2 receives a higher payoff than under the pure DD, because $v_e^* > P$. Then player 2's motivation to defect more is suppressed to avoid DD, so that the player 2 accepts exploitation over mutual defection.

Next, we see how the above exploitative relationship is established. By substituting Eq. (18) for Eq. (17), we get both the players' strategies as

$$
\begin{aligned}
x_C^* &= \frac{(P - S)(2T - R - P)}{(R + P - 2S)(T - P)}, \\
x_D^* &= 0, \\
y_C^* &= 1, \\
y_D^* &= 1 - \frac{(R + P - 2S)(T - R)}{(R - S)(2T - R - P)}.
\end{aligned} \tag{19}
$$

Here, recall that TFT strategy always cooperates (defects) to the other's cooperation (defection). In comparison to TFT, the exploited player 2's strategy is generous ($y_C^* = 1$ but $y_D^* > 0$) and can be termed as generous TFT (we use this term by extending its original definition in Ref. [18]). On the other hand, the exploiting player 1's strategy is narrower-minded than TFT ($x_D^* = 0$ but $x_C^* < 1$), say, narrow-minded TFT. It should be noted that the exploitation is stabilized by both the players. The generous and narrow-minded strategies are stabilized by each other. The exploiting (exploited) player with narrow-minded TFT (generous TFT) demands (accepts) the other's cooperation (defection) and thus promotes the other to be generous (narrow-minded). This result is unexpected from previous studies in which both players' TFT lead to the cooperative relationship. In addition, it should be noted here that this exploitative relationship is completely different from that by one-way optimization in Press and Dyson's study, because the present exploitation is achieved as a result of both players' optimization.

The condition $T - R - P + S > 0$ can be intuitively interpreted from the perspectives of both the exploiting and exploited players. From the perspective of exploiting player, the condition written as $T - R > P - S$ implies that a player's change of action from C to D is more beneficial when the other is C than D. In other words, the exploiting player is more motivated to defect than the exploited player is. In contrast, from the perspective of the exploited player, the condition written as $R - T < S - P$ means that a player's change of action from D to C is more beneficial when the other is D than C. In other words, the exploited player is more motivated to cooperate than the exploiting player is. Thus, the exploitative relationship is stabilized by both the players; the exploiting (exploited) one's motivation to defect (cooperate) is more. This condition $T - R - P + S > 0$ is known as "submodular PD" in economics [24], so that we use this term for this condition. In addition, the same condition is also observed in a biological study [25]. However, why and how such a condition leads to the exploitation is first noted here.

IV. TRANSIENT DYNAMICS TO THE LEARNING EQUILIBRIUM

In Sec. III, we analyzed the fixed points and the linear stability in their neighborhoods. However, this analysis is limited to only a small partition (i.e., the neighborhood of a two-dimensional space at most) of the whole four-dimensional phase space given by (x_C, x_D, y_C, y_D). We now study the transient dynamics to reach the learning equilibrium from arbitrary initial conditions of the two players $(x_C^o, x_D^o, y_C^o, y_D^o)$.

A. Characterization of transient dynamics

Although the attractors consist of the pure DD, CC, and various degrees of exploitative state with two dimensionality, the transient dynamics are categorized into the following several cases.

Case (1): Direct convergence to a cooperative relationship. As easily guessed, a large x_C and a small x_D encourage the other player to cooperate by punishing the other's defection. Thus, as Fig. 5(a) shows, when both players have sufficiently strong punishments, they evolve toward a cooperative relationship and converge to pure CC.

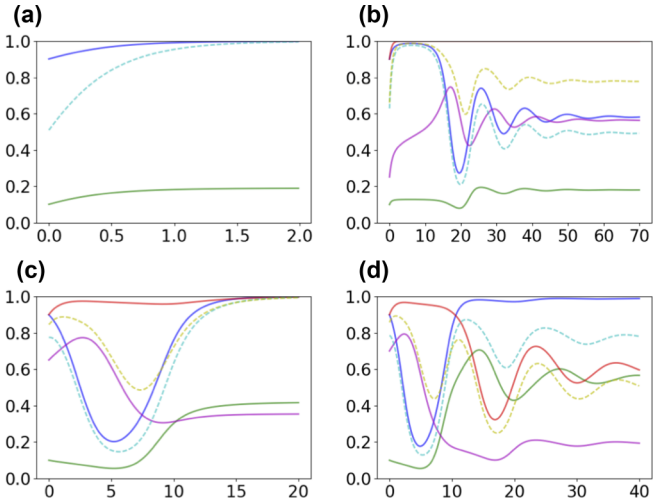

FIG. 5. Trajectories of strategies during the learning dynamics with a payoff matrix of $(T, R, P, S) = (5, 3.25, 1, 0)$. In all figures, $(x_C^o, x_D^o, y_C^o) = (0.9, 0.1, 0.9)$ are fixed, with (a) $y_D^o = 0.1$, (b) $y_D^o = 0.25$, (c) $y_D^o = 0.65$, and (d) $y_D^o = 0.7$. Thus, player 1's strategy is fixed close to TFT, but player 2's strategy departs from TFT, ranging from (a) (closest) to (d) (farthest). Blue, green, red, and magenta solid lines indicate x_C, x_D, y_C, and y_D, respectively. Yellow and cyan broken lines indicate x_e and y_e, respectively. Note that only player 1's trajectory is plotted in panel (a) because each of y_C, y_D, y_e is equal to x_C, x_D, x_e. (a) Trajectories of case (1). The player's probabilities of cooperation increase throughout the dynamics and converge to the pure CC. (b) Trajectories of case (2). Both players first move toward the pure CC. At time 10, however, player 1 takes advantage of player 2's generous strategy (i.e., too much unconditional cooperation) and increases his or her probability of defection. Against player 1's behavior, player 2 does not increase punishment to maintain the previous high probability of cooperation, which further increases player 1's defection probability. The finite degree of exploitation from player 1 to player 2 is thus fixed. (c) Trajectories of case (3). The initial asymmetry is larger than that in panel (b). Around time 5, the same exploitation as in panel (b) emerges. From time 5 to time 10, however, player 2 increases his or her punishment of player 1 decreasing y_D. From time 10, both players punish each other and finally reach the pure CC. (d) Trajectories of case (4). Until time 5, the exploitation of player 2 by player 1 emerges, and from time 5 to time 10, player 2 increases his or her punishment of player 1. From time 10 to time 20, however, player 2's excessive, one-sided punishment demands player 1's unconditional cooperation, which results in the reverse exploitative relationship from that in case (b).

Here, we emphasize that the extreme limit of the punishment strategy is given by $x_C = 1$ and $x_D = 0$, which is the TFT strategy. In general, strategy (x_D', x_C') is closer to TFT than strategy (x_C, x_D) is when both of $x_C' \geqslant x_C$ and $x_D' \leqslant x_D$ are satisfied. When only one of the inequalities holds, however, the strategy that is closer to TFT is not defined.

Case (2): Exploitative relationship as a failure to reach cooperation. Figure 5(b) shows an example of trajectory that reaches an asymmetric relationship in which one player exploits the other. Initially, one player is closer to TFT than the other is. Both players pursue a cooperative relationship

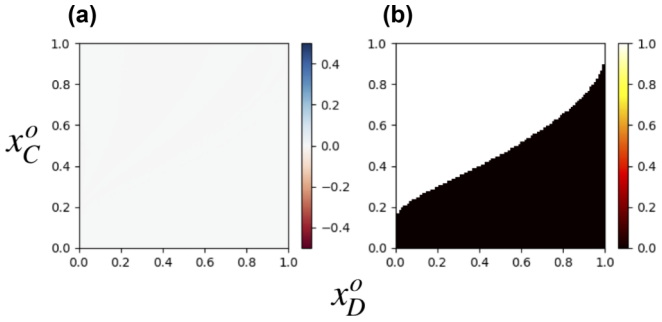

FIG. 6. The degree of (a) exploitation ($y_e^* - x_e^*$) and (b) cooperation ($x_e^* + y_e^*$)/2 in the final state of learning dynamics is plotted by a color, against the initial condition (x_D^o, x_C^o) for $(T, R, P, S) = (5, 4.5, 1, 0)$. The horizontal (vertical) axis indicates x_D^o (x_C^o), and the player 2's strategy is fixed at $y_C^o = 0.8, y_D^o = 0.2$. In this case, only pure CC and DD strategies are stable fixed points for learning dynamics. The basin to the pure CC (DD) strategy is plotted by white (black) points in the right figure.

by punishing each other [as in case (1)] in the beginning, but the latter player becomes too cooperative to punish the other. Thus, the former player switches to defection, and the latter player's strategy conversely increases the probability of cooperation regardless of the former player's defection. Thus, an exploitative relationship is achieved.

Case (3): Cooperative relationship recovered from exploitation. As seen in Fig. 5(c), the initial difference in the strategies is larger than that in case (2). The player closer to TFT initially starts to exploit the other [as in case (2)]. This exploitation, however, is too strong to become stable, and the latter player increases punishment, leading to the cooperative relationship found in case (1).

Case (4): Reversed exploitative relationship. An exploitative relationship is constructed between asymmetric players as in case (2), but now the relationship is reversed. Instead, the player who is initially farther from TFT exploits the closer player, as seen in Fig. 5(d). The degree of punishment oscillates over time, and the player who is more cooperative switches. If the difference in initial strategies increases further, the oscillation lasts longer, and which player exploits the other follows a complicated switching pattern. Finally, a reverse exploitative relationship is achieved.

B. Basin structure for exploitative state

In the above, we have shown transient trajectories reaching final cooperative or exploitative states. Now, we study the dependence of the final state after learning on the initial conditions.

First, if PD is not submodular, only the pure DD and CC strategies are stable. In these cases, if the initial state $(x_C^o, x_D^o, y_C^o, y_D^o)$ reaches the pure CC, any initial condition closer to TFT (i.e., with either $x_C^{o'} > x_C^o$ or $x_D^{o'} < x_D^o$) also reaches the pure CC, as shown in Fig. 6. Thus, the basin structure, how each initial state reaches a final state, is simple.

On the other hand, when PD is submodular, the basin structure is complicated, as seen in Fig. 7, in which pure CC and DD strategies and various degrees of exploitative relationships are achieved. Slight differences in initial states lead to changes in the final state, especially near the boundary of the basin to the pure DD.

From Fig. 7(a), we observe successive changes of cases (1)–(4) as well as further oscillation of punishments, plotted against the change in the difference between both players' initial strategies. In addition, note that the payoff (and action) at the basin boundary between the pure CC and exploitation [i.e., case (1) and (2)] is discontinuous. The collapse of cooperation results in a finite (indeed, rather large) degree of asymmetry in the payoffs. This discontinuous transition is due to positive feedback loop between generous and narrow-minded strategies. Once the cooperative relationship is collapsed, the generous player becomes more generous (i.e., decreases the probability of defect) to maintain cooperation.

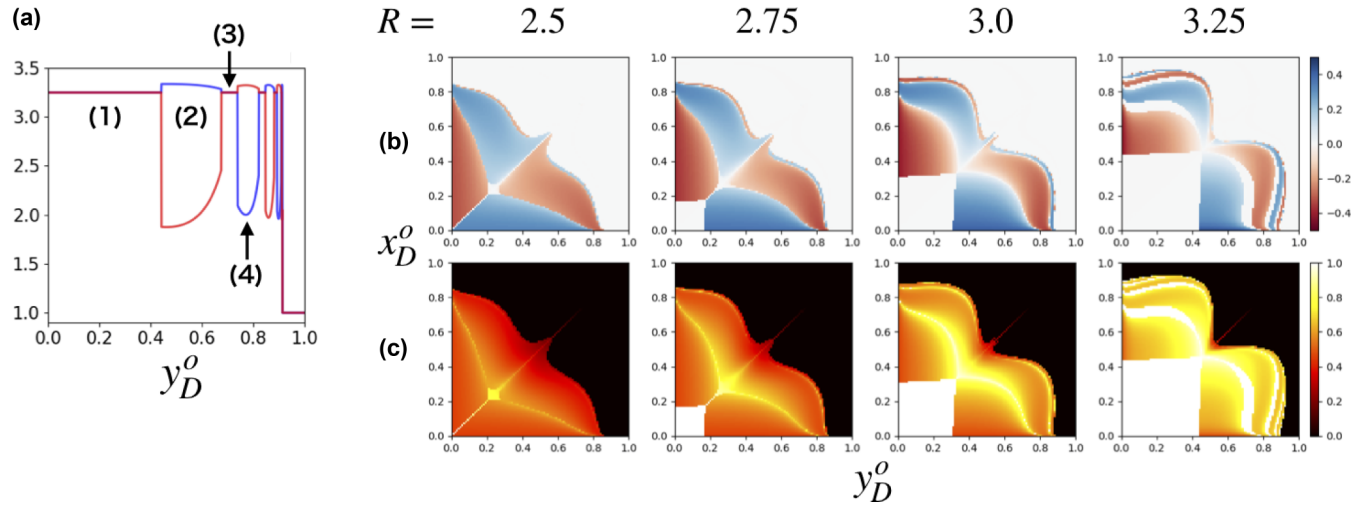

FIG. 7. Panel (a) shows both players' payoffs u_e^* (blue) and v_e^* (red) when $x_D^o = 0.1$ in $R = 3.25$. (#) indicates the trajectory case classified in Sec. IV A. Degrees of (b) exploitation of player 2 by player 1 ($:= y_e^* - x_e^*$) and (c) cooperation ($:= (x_e^* + y_e^*)/2$) are plotted against the initial values of (x_D^o, y_D^o). The horizontal (vertical) axis commonly indicates y_D^o (x_D^o), and both x_C^o and y_C^o are fixed to 0.999. In both figures, $(T, P, S) = (5, 1, 0)$ is fixed, and R equals 2.5, 2.75, 3.0, and 3.25 from left to right.

On the other hand, the narrow-minded player becomes more narrow-minded (i.e., increases the probability of defect) against the generous other.

Furthermore, Figs. 7(b) and 7(c) show how the basin structure changes depending on the payoff matrix. When the benefit of reciprocal cooperation is minimal for the PD [$R = (T + S)/2$], the region of cooperation is almost nonexistent even from the initial conditions with $x_D^o \sim y_D^o$. However small the difference between both players' initial strategies is, a finite amount of exploitation is finally achieved (in some cases with the reversal of the initial difference). In other words, symmetry breaking between both the strategies occurs as a result of amplification of any tiny difference in the initial strategies. As R is increased, however, cooperative relationship (pure CC) is finally established if the difference in the initial strategies is small [as seen in the enhancement of the cooperation region near the diagonal line in Fig. 7(b)]. A certain amount of difference between the initial strategies is necessary to reach the exploitative relationship, and the spontaneous symmetry breaking does not occur strictly.

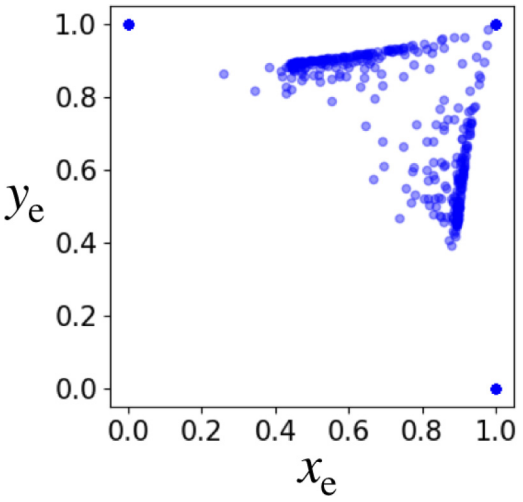

FIG. 8. The equilibrium states for the learning dynamics in the case of $(T, R, P, S) = (5, 4, 0, 1)$.

V. SUMMARY AND DISCUSSION

In this study, we formulated learning dynamics in which two players mutually update their probabilistic conditional strategies through a repeated game. This learning process is decomposed into frequency-dependent selection (i.e., the term $x_C\overline{x_C}$) and adaptive learning (i.e., the term $\partial u_e/\partial x_C$).

We analyzed the fixed-point attractors of a dynamical system of strategies. Interestingly, in addition to pure DD and CC strategies, two-dimensional neutral fixed points with an exploitative relationship can be stably reached if PD is submodular. Even though the two players have the same learning dynamics and intend to optimize their payoffs, an asymmetric relationship can be achieved under certain conditions. Accordingly, when we observed exploitative relationship, it is difficult to reason why one side is exploited by the other.

Our finding is that the exploitative relationship is stabilized by both the exploiting and exploited players. The exploiting player receives a higher payoff than the other player does and often receives a higher payoff than that under the pure CC. The exploited player receives a lower payoff than the other player does but secures at least the minimax payoff, which is obtained under the pure DD. In addition, this exploitative relationship is structured by asymmetric punishments against the other player's defection. Both players punish each other, but the exploiting (exploited) player defects (cooperates) more than the other does.

We then analyze the transient dynamics for reaching the exploitative state. For submodular PD, the feedback of punishment leads to the temporal oscillation of final state from cooperation to exploitation, cooperation, exploitation by the other player, and so forth, depending on how close the initial strategies are to TFT. The basin structure is complicated, and slight differences in the initial strategies can lead to the drastic changes in the final state.

Complicated strategies with memories over many previous actions are sometimes studied by using multiagent learning models, such as the coupled neural networks. As a result of reciprocal learning, an emergence of exploitative relationship [26] and the endogenous acquisition of punishment [27] are observed at some stage in the iterated PD. However, whether the state is stable or transient is not explored. To analyze such state is rather difficult, because the dynamics are nondeterministic and extremely high-dimensional, as is generally seen in machine learning studies. In contrast, our model is deterministic and low dimensional, so that the stationary exploitative state is clearly analyzed, which will also provide a basis to study the behavior of complicated multiagent systems.

There is a recent report of exploitation by symmetry breaking in a game between players on two two-dimensional lattices [28]. Each player's strategy is updated by Monte Carlo simulation within each lattice, whereas the payoff is given by the game between the same positions of two lattices. Thus, the exploitation, the difference between the fraction of cooperators in the two lattices, can be allowed as each player does not have any motivation to increase the average payoff lattice. Further, the population dynamics are nondeterministic and high dimensional, which make the analysis for exploitation harder. Note that the condition for the payoff matrix to achieve exploitation is restricted and totally different from ours. In our deterministic game, there exist only two players who are motivated to increase their own payoff over the other's. Under this natural, simple setup, we have demonstrated the emergence of exploitation and uncovered the analytic condition for it.

We can straightforwardly adopt our learning model to games with general payoff matrix. Then, equilibria with different degrees of asymmetry between the two players coexist as in the PD game. For instance, the equilibria for the snowdrift (chicken) game in which $T > R > S > P$ holds are shown in Fig. 8, as in Fig. 3 for the PD. Note, however, that in this snowdrift game, there already exists asymmetric Nash equilibrium as pure CD and DC. Thus, the emergence of various asymmetric equilibria by our cognitive learning dynamics is not so striking.

Note that the PD game is the classic paradigm for the study of cooperation and defection. Thus, the results of this study have general implications for the issues of cooperation, exploitation, and defection. Here, it is interesting

to note that the emergence of exploitation depends on the payoff matrix (T, R, P, S). We have shown that submodular PD [which includes the standard case adopted in most previous studies, i.e., the matrix (5,3,1,0)] generally justifies exploitation from both the exploiting and exploited player's perspectives.

It is often thought that exploitative relationships result from differences in players' abilities or environmental conditions. Whether and how players with the same learning abilities evolve toward the "symmetry breaking" associated with exploitation remains unknown. We have shown that exploitation can emerge even between players with same learning rule and the same payoff based on differences in their initial strategies. Furthermore, the complicated basin structure that we observe implies that slight difference in the initial strategies can lead to an unexpected exploitation relationship with regard to which player exploits the other. This result provides a novel perspective on the origins of exploitation and complex societal relationships.

ACKNOWLEDGMENTS

The authors would like to thank E. Akiyama, T. Sekiguchi, and H. Ohtsuki for useful discussions. This research was partially supported by JSPS KAKENHI Grants No. JP18J13333 and No. JP17H06386, and Hitachi, the University of Tokyo Laboratory.

[1] R. Axelrod and W. D. Hamilton, The evolution of cooperation, Science **211**, 1390 (1981).
[2] R. Axelrod and D. Dion, The further evolution of cooperation, Science **242**, 1385 (1988).
[3] M. A. Nowak and R. M. May, Evolutionary games and spatial chaos, Nature (London) **359**, 826 (1992).
[4] A. Traulsen and M. A. Nowak, Evolution of cooperation by multilevel selection, Proc. Natl. Acad. Sci. USA **103**, 10952 (2006).
[5] C. Hilbe, S. Simsa, K. Chatterjee, and M. A. Nowak, Evolution of cooperation in stochastic games, Nature (London) **559**, 246 (2018).
[6] W. H. Press and F. J. Dyson, Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent, Proc. Natl. Acad. Sci. USA **109**, 10409 (2012).
[7] C. Hilbe, M. A. Nowak, and K. Sigmund, Evolution of extortion in iterated prisoner's dilemma games, Proc. Natl. Acad. Sci. USA **110**, 6913 (2013).
[8] A. J. Stewart and J. B. Plotkin, From extortion to generosity, evolution in the iterated prisoner's dilemma, Proc. Natl. Acad. Sci. USA **110**, 15348 (2013).
[9] Z. X. Wu and Z. Rong, Boosting cooperation by involving extortion in spatial prisoner's dilemma games, Phys. Rev. E **90**, 062102 (2014).
[10] A. Szolnoki and M. Perc, Defection and extortion as unexpected catalysts of unconditional cooperation in structured populations, Sci. Rep. **4**, 5496 (2014).
[11] Z. Rong, Z. X. Wu, D. Hao, M. Z. Chen, and T. Zhou, Diversity of timescale promotes the maintenance of extortioners in a spatial prisoner's dilemma game, New J. Phys. **17**, 033032 (2015).
[12] T. Börgers and R. Sarin, Learning through reinforcement and replicator dynamics, J. Econ. Theory **77**, 1 (1997).
[13] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, UK, 1998).
[14] Y. Sato, E. Akiyama, and J. D. Farmer, Chaos in learning a simple two-person game, Proc. Natl. Acad. Sci. USA **99**, 4748 (2002).
[15] G. W. Brown, Iterative solution of games by fictitious play, in *Activity Analysis of Production and Allocation*, edited by T. C. Koopmans (Wiley, New York, 1951), Vol. 13, pp. 374–376.
[16] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, MA, 1998).
[17] S. Dridi and E. Akcay, Learning to cooperate: The evolution of social rewards in repeated interactions, Am. Naturalist **191**, 58 (2018).
[18] M. A. Nowak and K. Sigmund, Tit for tat in heterogeneous populations, Nature (London) **355**, 250 (1992).
[19] D. Premack and G. Woodruff, Does the chimpanzee have a theory of mind? Behavioral Brain Sci. **1**, 515 (1978).
[20] R. Saxe and S. Baron-Cohen, *Theory of Mind: A Special Issue of Social Neuroscience* (Psychology Press, London, 2007).
[21] R. W. Lurz, *Mindreading Animals: The Debate over What Animals Know about Other Minds* (MIT Press, Cambridge, MA, 2011).
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.1.033077 for the detailed calculation and extra data.
[23] Y. Fujimoto and K. Kaneko, Functional dynamic by intention recognition in iterated games, New J. Phys. **21**, 023025 (2019).
[24] S. Takahashi, Community enforcement when players observe partners' past play, J. Econ. Theory **145**, 42 (2010).
[25] M. Nowak, Stochastic strategies in the prisoner's dilemma, Theor. Pop. Biol. **38**, 93 (1990).
[26] T. W. Sandholm and R. H. Crites, Multiagent reinforcement learning in the iterated prisoner's dilemma, Biosystems **37**, 147 (1996).
[27] M. Taiji and T. Ikegami, Dynamics of internal models in game players, Physica D (Amsterdam, Neth.) **134**, 253 (1999).
[28] Q. Jin, L. Wang, C. Y. Xia, and Z. Wang, Spontaneous symmetry breaking in interdependent networked game, Sci. Rep. **4**, 4095 (2014).