Global Behavior of Learning Dynamics in Zero-Sum Games with Memory Asymmetry*

Yuma Fujimoto
CyberAgent
Tokyo, Japan
fujimoto.yuma1991@gmail.com

Kaito Ariu
CyberAgent
Tokyo, Japan
kaito_ariu@cyberagent.co.jp

Kenshi Abe
CyberAgent
Tokyo, Japan
abekenshi1224@gmail.com

ABSTRACT

This study examines the global behavior of dynamics in learning in games between two players, X and Y. We consider the simplest situation for memory asymmetry between two players: X memorizes the other Y's previous action and uses reactive strategies, while Y has no memory. Although this memory complicates their learning dynamics, we characterize the global behavior of such complex dynamics by discovering and analyzing two novel quantities. One is an extended Kullback-Leibler divergence from the Nash equilibrium, a well-known conserved quantity from previous studies. The other is a family of Lyapunov functions of X's reactive strategy. One of the global behaviors we capture is that if X exploits Y, then their strategies converge to the Nash equilibrium. Another is that if Y's strategy is out of equilibrium, then X becomes more exploitative with time. Consequently, we suggest global convergence to the Nash equilibrium from both aspects of theory and experiment. This study provides a novel characterization of the global behavior in learning in games through a couple of indicators.

KEYWORDS

Multi-Agent Learning, Zero-Sum Game, Dynamical Systems, Lyapunov Function

ACM Reference Format:
Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. 2025. Global Behavior of Learning Dynamics in Zero-Sum Games with Memory Asymmetry. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 7 pages.

1 INTRODUCTION

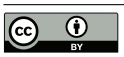
Learning in games targets how multiple agents learn their optimal strategies in the repetition of games [11]. The set of such players' best strategies is defined as Nash equilibrium [24], where every player has no motivation to change his/her strategy. However, this equilibrium is hard to compute in general because one's best strategy depends on the others' strategies. Indeed, the behavior of multi-agent learning is complicated in zero-sum games, where players conflict in their payoffs. Even when players try to learn their optimal strategies there, their strategies often cycle around the Nash equilibrium and fail to converge to the equilibrium.

In order to understand such strange behaviors, which are unique in multi-agent learning, the dynamics of how multiple agents learn their strategies, say, learning dynamics, are frequently studied [5, 9, 33, 34]. The representative dynamics of interest are the replicator dynamics, which is based on the evolutionary dynamics in biology [6, 10, 17, 26, 32]. These dynamics are also known as the multiplicative weight updates (MWU) in its discrete-time version [1, 3]. Furthermore, their connection to other representative learning dynamics, such as gradient ascent [7, 8, 31, 38] and Q-learning [18, 19, 37], should be noted. Such replicator dynamics are known to be characterized by Kullbuck-Leibler (KL) divergence, which is the distance from the Nash equilibrium to the players' present strategies. This KL divergence is conserved during the learning dynamics, and the distance from the Nash equilibrium is invariant [28, 29]. Follow the Regularized Leader (FTRL) is a class of learning algorithms including the replicator dynamics and also has its conserved quantity, which is the summation of divergences for all the players [21, 22]. To summarize, such complex learning dynamics have been discussed based on their conserved quantity.

In this study, we define memory as an agent's ability to change its action choice depending on the outcome of past games. By definition, this memory allows the agent to make more complex and intelligent decisions. When memory is introduced into a normal-form game, the players can achieve a wider range of strategies as the Nash equilibria (known as Folk theorem [12]). Furthermore, memory is also introduced into learning algorithms, such as replicator dynamics [13–16] and Q-learning [4, 20, 23, 35, 36]. Here, since this memory causes feedback from the past, the global dynamics of such learning algorithms become more complex. Indeed, replicator dynamics diverge from the Nash equilibrium under symmetric memory lengths between players [13], while converging under asymmetric memory lengths [14]. Here, KL divergence is no longer useful to capture the global dynamics because it increases or decreases over time. The analysis of the dynamics in with-memory games is limited to the local, linearized stability analysis in the vicinity of Nash equilibrium [14]. To summarize, since memory crucially complicates learning dynamics, the global behavior of the dynamics is still unexplored.

This study provides the first theoretical analysis of the global behavior of learning in with-memory games. We assume games where their memory structure is simplest and asymmetric; One side adopts a reactive strategy that can memorize the other's previous action [2, 15, 16, 25, 27, 30], while the other has no memory. In order to characterize the global behavior of such with-memory games, we extend KL divergence and prove that such extended divergence increases or decreases with time depending on whether the reactive strategy is exploitative or not (see Fig. 1A). We further propose a

*The full version is available at https://arxiv.org/abs/2405.14546



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

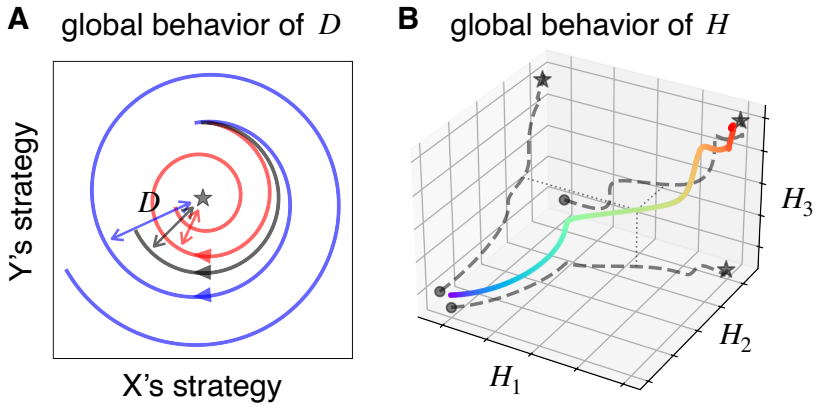


Figure 1: (A) Illustration of the global behavior of the conditional divergence, $D(X, \boldsymbol{y})$. Three trajectories (red, black, and blue) are plotted with the Nash equilibrium (the black star marker). The horizontal and vertical axes show X's strategy (x_1^{st}) and Y's strategy (y_1) in the matching pennies game (formulated in Fig. 2). This divergence decreases (red: $\dot{D} < 0$), cycles (black: $\dot{D} = 0$), or increases (blue: $\dot{D} > 0$) with time. These three lines are plotted for the different initial strategies, i.e., X and $\boldsymbol{y}$. (B) Illustration of the global behavior of the family of Lyapunov functions, $H(X; \boldsymbol{\delta})$. The colored line shows a trajectory (from purple to red) of Lyapunov functions H_1, H_2, and H_3, each of which is $H(X; \boldsymbol{\delta})$ for some specific $\boldsymbol{\delta}$. The gray broken lines are the projections of the black solid line to H_1-H_2, H_2-H_3, and H_3-H_1 planes. All of H_1, H_2, and H_3 monotonically increase with time.

family of Lyapunov functions that characterize the dynamics of the reactive strategy (see Fig. 1B). These Lyapunov functions show that the with-memory side monotonically learns to exploit the no-memory side. As an application of these functions, we suggest the convergence from arbitrary initial strategies to the equilibrium, i.e., global convergence. We prove global convergence in the matching pennies game. We also experimentally confirm that such global convergence is observed in other games equipped with various types of equilibrium.

2 PRELIMINARY

2.1 Settings

First, we formulate two-player normal-form games. We consider two players, denoted as X and Y. X's actions are denoted as $\{a_i\}_{1\le i\le m_{\mathrm{X}}}$, while Y's are $\{b_j\}_{1\le j\le m_{\mathrm{Y}}}$. When X and Y choose a_i and b_j, they obtain the payoffs of $u_{ij} \in \mathbb{R}$ and $v_{ij} \in \mathbb{R}$, respectively. Thus, all their possible payoffs are given by the matrices, $\boldsymbol{U} := (u_{ij})_{ij}$ and $\boldsymbol{V} := (v_{ij})_{ij}$. Here, when $\boldsymbol{V} = -\boldsymbol{U}$ holds, the games are called zero-sum. Although our formulation of learning algorithms can be used for general games, this study focuses on zero-sum games.

We assume that X use reactive strategies, i.e., can change its action choice depending on the other's previous action. This reactive strategy is denoted as $\boldsymbol{X} := (x_{i|j})_{1\le i\le m_{\mathrm{X}}, 1\le j\le m_{\mathrm{Y}}} \in \prod_{1\le j\le m_{\mathrm{Y}}} \Delta^{m_{\mathrm{X}}-1}$, a matrix composed of m_{Y} vectors each of which are an element of a $m_{\mathrm{X}} - 1$-dimensional simplex. Here, $x_{i|j}$ means the probability that X chooses a_i in the condition when Y's previous action is b_j. Thus, $\Sigma_i x_{i|j} = 1$ should be satisfied for all j. On the other hand, Y only can use classical mixed strategies and choose its own action without reference to the previous actions. This mixed strategy is denoted as $\boldsymbol{y} = (y_j)_{1\le j\le m_{\mathrm{Y}}} \in \Delta^{m_{\mathrm{Y}}-1}$, a vector which is an element of a $(m_{\mathrm{Y}} - 1)$-dimensional simplex. Thus, $\Sigma_j y_j = 1$ should be satisfied.

2.2 Stationary State and Expected Payoff

We now discuss the stationary state and expected payoff of repeated games. Since Y determines its action independent of the outcomes of previous rounds, X's stationary strategy, defined as $\boldsymbol{x}^{\mathrm{st}} := (x_i^{\mathrm{st}})_{1\le i\le m_{\mathrm{X}}}$, is given by $x_i^{\mathrm{st}}(\boldsymbol{x}_i, \boldsymbol{y}) = \Sigma_j x_{i|j} y_j$. Here, x_i^{st} means the probability that X chooses a_i in the stationary state. Furthermore, the stationary state is described as $\boldsymbol{P}^{\mathrm{st}} := \boldsymbol{x}^{\mathrm{st}} \otimes \boldsymbol{y}$ with use of $\boldsymbol{x}^{\mathrm{st}}$ and $\boldsymbol{y}$. Last, X's expected payoff is given by $u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}}, \boldsymbol{y}) := \Sigma_i \Sigma_j u_{ij} p_{ij}^{\mathrm{st}} = \Sigma_i \Sigma_j u_{ij} x_i^{\mathrm{st}} y_j$.

2.3 Nash Equilibrium

We now define the Nash equilibrium in the normal-form game. Here, note that this equilibrium is based on games without memories, where X's strategy does not refer to the past games, i.e., $\boldsymbol{x} := (x_i)_i$. By using the expected payoff $u^{\mathrm{st}}(\boldsymbol{x}, \boldsymbol{y})$ for games without memories, the Nash equilibrium $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is formulated as

$$\begin{cases} \boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}} u^{\mathrm{st}}(\boldsymbol{x}, \boldsymbol{y}^*) \\ \boldsymbol{y}^* = \mathrm{argmin}_{\boldsymbol{y}} u^{\mathrm{st}}(\boldsymbol{x}^*, \boldsymbol{y}) \end{cases}. \tag{1}$$

From the definition, $u^{\mathrm{st}}(\boldsymbol{x}, \boldsymbol{y})$ is the linear function for $\boldsymbol{x}$ and $\boldsymbol{y}$, and the Nash equilibrium condition is characterized by the gradient of such expected payoffs as

$$\begin{cases} \partial u^{\mathrm{st}}/\partial x_i = \Sigma_j u_{ij} y_j^* = C & (x_i^* > 0) \\ \partial u^{\mathrm{st}}/\partial x_i = \Sigma_j u_{ij} y_j^* \le C & (x_i^* = 0) \end{cases}, \tag{2}$$

$$\begin{cases} \partial u^{\mathrm{st}}/\partial y_j = \Sigma_i u_{ij} x_i^* = C & (y_j^* > 0) \\ \partial u^{\mathrm{st}}/\partial y_j = \Sigma_i u_{ij} x_i^* \ge C & (y_j^* = 0) \end{cases}. \tag{3}$$

From these conditions, for all i and j such that $x_i^* > 0$ and $y_j^* > 0$ hold, respectively, we obtain

$$\Sigma_i u_{ij} x_i^* = \Sigma_j u_{ij} y_j^* =: u^*. \tag{4}$$

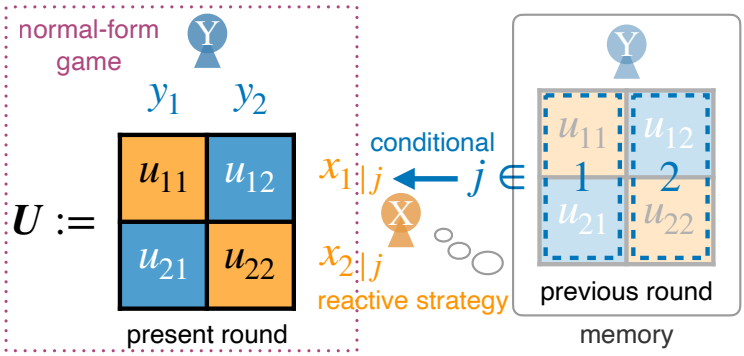


Figure 2: Illustration of games between reactive and zero-memory strategies. The area surrounded by the magenta dotted line shows the normal-form game. In each round, X chooses action $i = 1$ or 2 in the row, following its strategy, i.e., the probability distribution of $\boldsymbol{x} = (x_1, x_2)$. On the other hand, Y chooses action $j = 1$ or 2 in the column, following its strategy, i.e., the probability distribution of $\boldsymbol{y} = (y_1, y_2)$. Depending on their choices i and j, X receives a payoff u_{ij}, given by a matrix form of $U = (u_{ij})_{i,j} = ((u_{11}, u_{12}), (u_{21}, u_{22}))$. Furthermore, in zero-sum games, Y receives $-u_{ij}$. Especially in the matching pennies game, their actions of 1 (2) correspond to the choice of "head" ("tail") of a coin. When their choices match $i = j$, X wins, i.e., $u_{11} = u_{22} = 1$ (the orange blocks). Else when their choices mismatch $i \neq j$, Y wins, i.e., $u_{12} = u_{21} = -1$ (the blue blocks). The area outside of the magenta dotted line shows the difference due to an effect of memory. The gray box shows that X memorizes Y's previous action, represented as $j = 1$ or 2. Thus, X uses a reactive strategy and can choose its action with the conditional probability of $x_{1|j}$ and $x_{2|j}$ for Y's previous action.

Let us interpret this equation. First, $\Sigma_i u_{ij} x_i^* = u^*$ means that when X takes its Nash equilibrium strategy, its own payoff is fixed to u^*, independent of Y's strategy. On the other hand, $\Sigma_j u_{ij} y_j^* = u^*$ similarly means that Y's Nash equilibrium strategy fixes X's payoff to u^*. In other words, either X or Y takes its Nash equilibrium strategy, their payoffs are fixed. This is the special property in zero-sum games.

2.4 Learning Algorithm: Replicator Dynamics

Let us define the replicator dynamics as a representative learning algorithm. X's and Y's replicator dynamics are formulated as

$$\dot{x}_{i|j} = +x_{i|j}\left(\frac{\mathrm{d}u^{\mathrm{st}}}{\mathrm{d}x_{i|j}} - \Sigma_i x_{i|j}\frac{\mathrm{d}u^{\mathrm{st}}}{\mathrm{d}x_{i|j}}\right), \tag{5}$$

$$\dot{y}_j = -y_j\left(\frac{\mathrm{d}u^{\mathrm{st}}}{\mathrm{d}y_j} - \Sigma_j y_j\frac{\mathrm{d}u^{\mathrm{st}}}{\mathrm{d}y_j}\right). \tag{6}$$

Here, following the theorems in [13], X's and Y's replicator dynamics include the gradient for the expected payoff u^{st}. Thus, the update of X's strategy increases its payoff u^{st}, while that of Y's strategy decreases the other's payoff u^{st}. We discuss learning based on the replicator dynamics throughout this study, but we can extend all the following results to another typical learning algorithm, the gradient descent-ascent (see Appendix D for detailed discussion).

3 THEORY ON LEARNING DYNAMICS

This section analyzes the dynamics of Eqs. (5) and (6). First, we compute in detail the gradient terms, which appear to be complex. Next, as a preliminary, we define positive definite matrices for some special vectors. Based on this definition, we introduce two quantities characterizing the dynamics of Eqs. (5) and (6): An extended KL divergence and a family of Lyapunov functions.

3.1 Polynomial Expressions of Learning

First, the gradient terms in Eqs. (5) and (6) are computed as

$$\frac{\mathrm{d}u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}}(\boldsymbol{X},\boldsymbol{y}),\boldsymbol{y})}{\mathrm{d}x_{i|j}} = \frac{\partial x_i^{\mathrm{st}}(\boldsymbol{x}_i,\boldsymbol{y})}{\partial x_{i|j}}\frac{\partial u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}},\boldsymbol{y})}{\partial x_i^{\mathrm{st}}} \tag{7}$$

$$= y_j \Sigma_{j'} u_{ij'} y_{j'}, \tag{8}$$

$$\frac{\mathrm{d}u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}}(\boldsymbol{X},\boldsymbol{y}),\boldsymbol{y})}{\mathrm{d}y_j} = \frac{\partial u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}},\boldsymbol{y})}{\partial y_j} + \Sigma_i \frac{\partial x_i^{\mathrm{st}}(\boldsymbol{x}_i,\boldsymbol{y})}{\partial y_j}\frac{\partial u^{\mathrm{st}}(\boldsymbol{x}^{\mathrm{st}},\boldsymbol{y})}{\partial x_i^{\mathrm{st}}} \tag{9}$$

$$= \Sigma_i u_{ij} x_i^{\mathrm{st}} + \Sigma_i x_{i|j} \Sigma_{j'} u_{ij'} y_{j'}. \tag{10}$$

Here, we remark that Eqs. (5) and (6) are nonlinear functions of $\boldsymbol{X}$ and $\boldsymbol{y}$, which is a feature of learning in with-memory games. Notably, however, these equations are polynomial expressions with $\boldsymbol{X}$ and $\boldsymbol{y}$. Such polynomial expressions cannot be seen in the games of other memory lengths [13, 14] but are special in the games between reactive and no memory strategies.

3.2 Positive Definiteness for Zero-Sum Vectors

Next, let us introduce a definiteness of matrices. Here, however, this definite matrix is for vectors whose elements are summed to 0, named "zero-sum vectors". In mathematics, zero-sum vector $\boldsymbol{\delta} := (\delta_k)_k$ satisfies $\Sigma_k \delta_k = 0$ but $\boldsymbol{\delta} \neq \boldsymbol{0}$.

Definition 1 (Positive definiteness for zero-sum vectors). *A square matrix $\boldsymbol{M}$ is "positive definite for zero-sum vectors" when for all vectors $\boldsymbol{\delta} \neq \boldsymbol{0}$ such that $\Sigma_k \delta_k = 0$, $\boldsymbol{\delta} \cdot (\boldsymbol{M}\boldsymbol{\delta}) < 0$ holds.*

The positive definiteness for zero-sum vectors connects with an ordinary positive definiteness by a simple transformation of a matrix (see Appendix B for details).

3.3 Extended Kullback-Leibler Divergence

The first quantity is an extended version of divergence. Before considering the extension, we introduce the classical version of divergence D_{c}, which is the function of X's mixed strategies ($\boldsymbol{x} := (x_i)_{1\le i\le m_{\mathrm{X}}} \in \Delta^{m_{\mathrm{X}}-1}$) and Y's mixed strategies ($\boldsymbol{y}$) as

$$D_{\mathrm{c}}(\boldsymbol{x},\boldsymbol{y}) := D_{\mathrm{KL}}(\boldsymbol{x}^*\|\boldsymbol{x}) + D_{\mathrm{KL}}(\boldsymbol{y}^*\|\boldsymbol{y}), \tag{11}$$

$$D_{\mathrm{KL}}(\boldsymbol{p}^*\|\boldsymbol{p}) := \boldsymbol{p}^*\cdot\log\boldsymbol{p}^* - \boldsymbol{p}^*\cdot\log\boldsymbol{p}. \tag{12}$$

We now give an intuitive interpretation of this quantity. First, $D_{\mathrm{KL}}(\boldsymbol{p}^*\|\boldsymbol{p})$ is the KL divergence, meaning the distance from the reference point $\boldsymbol{p}^*$ to the target point $\boldsymbol{p}$. Thus, $D_{\mathrm{c}}(\boldsymbol{x},\boldsymbol{y})$ means the total distance from the Nash equilibrium $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ to the current state $(\boldsymbol{x},\boldsymbol{y})$.

Let us extend the classical divergence to the case of this study, where X refers to the previous action of the other and can use reactive strategies $\boldsymbol{X}$. This extended divergence, i.e., $D(\boldsymbol{X},\boldsymbol{y})$, is named the "conditional-sum" divergence, formulated as

$$D(\boldsymbol{X},\boldsymbol{y}) := \Sigma_j D_{\mathrm{KL}}(\boldsymbol{x}^*\|\boldsymbol{x}_j) + D_{\mathrm{KL}}(\boldsymbol{y}^*\|\boldsymbol{y}). \tag{13}$$

We now remark the difference between $D(\boldsymbol{X},\boldsymbol{y})$ and $D_{\mathrm{c}}(\boldsymbol{x},\boldsymbol{y})$. Recall that X's reactive strategy is defined as $(\boldsymbol{x}_j)_{1\le j\le m_{\mathrm{Y}}}$, which shows how to choose its action with the condition that Y chose b_j in the previous round. Hence, $D(\boldsymbol{X},\boldsymbol{y})$ represents the summation of KL divergence from $\boldsymbol{x}^*$ to $\boldsymbol{x}_j$ for all the conditions of j. Here, we also remark that when the reactive strategy does not use memory, i.e., $\boldsymbol{x}_j = \boldsymbol{x}$ for all j, this conditional-sum divergence also captures the behavior of the classical divergence (see Appendix C for details).

This conditional-sum divergence satisfies the following theorem (see Appendix A.1 for its full proof).

Theorem 1 (Monotonic decrease of D for positive definite $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$). *If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is positive definite for zero-sum vector, $D^{\dagger}(\boldsymbol{X};\mathrm{d}\boldsymbol{y}) := \dot{D}(\boldsymbol{X},\boldsymbol{y}) < 0$ for all $\mathrm{d}\boldsymbol{y} := \boldsymbol{y}-\boldsymbol{y}^* \neq \boldsymbol{0}$.*

Proof Sketch. We calculation $\dot{D}(\boldsymbol{X},\boldsymbol{y})$ in practice. In the calculation, the contribution of X's gradient (Eq. (8)) cancels out the contribution of the first term of the gradient of Y (Eq. (10)). (Here, we remark that the same canceling out also occurs in the calculation for the conservation of the classical divergence $D_{\mathrm{c}}(\boldsymbol{x},\boldsymbol{y})$ in games without memory.) However, the contribution of the second term of Eq. (10) is special in games of a reactive strategy. By using the constant payoff condition in the Nash equilibrium (Eqs. (4)), we obtain

$$\dot{D}(\boldsymbol{X},\boldsymbol{y}) = -\mathrm{d}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\mathrm{d}\boldsymbol{y}\ (=: D^{\dagger}(\boldsymbol{X};\mathrm{d}\boldsymbol{y})), \tag{14}$$

which means the difference from Y's equilibrium strategy and is a zero-sum vector. Thus, when $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is positive definite for zero-sum vectors, $\mathrm{d}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\mathrm{d}\boldsymbol{y}$ is always positive, leading to $D^{\dagger}(\boldsymbol{X};\mathrm{d}\boldsymbol{y}) < 0$ for all $\mathrm{d}\boldsymbol{y} \neq \boldsymbol{0}$. □

3.4 Family of Lyapunov Functions

Furthermore, we introduce a Lyapunov function, which characterizes the learning dynamics of X's reactive strategy. Based on an arbitrary zero-sum vector $\boldsymbol{\delta} := (\delta_i)_{1\le i\le m_{\mathrm{X}}}$, this function is defined as

$$H(\boldsymbol{X};\boldsymbol{\delta}) := \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta}. \tag{15}$$

The following theorem holds for this Lyapunov function (see Appendix A.2 for its full proof).

Theorem 2 (Monotonic increase of H). *For all $\boldsymbol{\delta}$ such that $\Sigma_i\delta_i = 0$, $H^{\dagger}(\boldsymbol{y};\boldsymbol{\delta}) := \dot{H}(\boldsymbol{X};\boldsymbol{\delta}) \ge 0$. The equality holds if and only if $d\boldsymbol{y}(=\boldsymbol{y}-\boldsymbol{y}^*) = \boldsymbol{0}$.*

Proof Sketch. By using Eq. (4), we calculate

$$\dot{H}(\boldsymbol{X};\boldsymbol{\delta}) = |\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{U}\mathrm{d}\boldsymbol{y}|^2\ (=: H^{\dagger}(\boldsymbol{y};\boldsymbol{\delta})). \tag{16}$$

This means that $H^{\dagger}(\boldsymbol{y};\boldsymbol{\delta}) \ge 0$ for all $\boldsymbol{\delta}$. If we substitute $\boldsymbol{\delta} = \boldsymbol{x}-\boldsymbol{x}^*$ for some $\boldsymbol{x} \in \Delta^{m_{\mathrm{X}}-1}$, we obtain

$$H^{\dagger}(\boldsymbol{y};\boldsymbol{x}-\boldsymbol{x}^*) = |u^{\mathrm{st}}(\boldsymbol{x},\boldsymbol{y}) - u^*|^2. \tag{17}$$

For any $\boldsymbol{y} \neq \boldsymbol{y}^* \Leftrightarrow \mathrm{d}\boldsymbol{y} \neq \boldsymbol{0}$, there is $\boldsymbol{x}$ such that $u^{\mathrm{st}}(\boldsymbol{x},\boldsymbol{y}) - u^* \neq 0$. Thus, $\mathrm{d}\boldsymbol{y} = \boldsymbol{0}$ is equivalent to $H^{\dagger}(\boldsymbol{y};\boldsymbol{\delta}) = 0$ for all $\boldsymbol{\delta}$. □

Let us interpret the function of $H(\boldsymbol{X};\boldsymbol{\delta})$. First, $H(\boldsymbol{X};\boldsymbol{\delta})$ is the quadratic form of matrix $\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}$ for $\boldsymbol{\delta}$. We now focus on the meaning of $\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}$. The i', i element of $\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}$ is given by $\boldsymbol{u}_{i'}\cdot\log\boldsymbol{x}_i$, in which we denoted vector $\boldsymbol{x}_i := (x_{i|j})_j$. This is the inner product of the payoff under X taking i'-th action ($\boldsymbol{u}_{i'}$) and the logarithmic strategy under X taking i-th action ($\log\boldsymbol{x}_i$). Thus, $\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}$ shows a correspondence matrix between X's strategy and its payoff matrix. Since $H(\boldsymbol{X};\boldsymbol{\delta})$ is the quadratic form of the correspondence matrix, it tends to be large when the diagonal elements of the correspondence matrix are large. The diagonal elements, i.e., $\boldsymbol{u}_i\cdot\log\boldsymbol{x}_i$, indicate how X exploits Y for each i-th action. Thus, $H(\boldsymbol{X};\boldsymbol{\delta})$ means the degree of exploitation of Y by X.

3.5 Global Behavior by Two Quantities

So far, the two theorems (Thms. 1 and 2) explain how the two quantities vary with time. Let us understand the global behavior of the learning dynamics by interpreting the two quantities.

D explains increasing/decreasing of distance: First, recall that $D(\boldsymbol{X},\boldsymbol{y})$ means the distance from the Nash equilibrium. Thm. 1 shows that the distance becomes smaller when $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is positive definite, whereas larger when $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is negative definite. Here, we focus on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ which determines whether the distance becomes smaller or larger. The j, j' element of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is given by $\boldsymbol{x}_j\cdot\boldsymbol{u}_{j'}$, in words, the correspondence between X's strategy for j-th action ($\boldsymbol{x}_j$) and its payoff for j'-th action ($\boldsymbol{u}_{j'}$). The eigenvalues of a matrix are roughly determined by the diagonal elements $\boldsymbol{x}_j\cdot\boldsymbol{u}_j$. This diagonal element is larger when $x_{i|j}$ takes a larger value for larger u_{ij}, meaning that X exploits Y's payoff more. As the simplest example, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is positive definite when X takes $x_{i|j} = 1$ for $i = \hat{i}$ such that $\hat{i} = \arg\max_i u_{ij}$, while $x_{i|j} = 0$ for $i \neq \hat{i}$. To summarize, Thm. 1 captures the global behavior where if X exploits Y, their strategies converge to the Nash equilibrium; otherwise, they diverge.

H explains the monotonic increase of exploitability: Next, recall that $\boldsymbol{U}\log\boldsymbol{X}^{\mathrm{T}}$ in $H(\boldsymbol{X};\boldsymbol{\delta})$ indicates a correspondence matrix between X's strategy and its payoff matrix. Thus, Thm. 2 explains that unless Y takes the equilibrium strategy, this correspondence continues to increase with time. We also pay attention to the change in the degree of correspondence, i.e., $H^{\dagger}(\boldsymbol{y};\boldsymbol{\delta})$. By substituting $\boldsymbol{\delta} = \boldsymbol{x}-\boldsymbol{x}^*$ for some $\boldsymbol{x} \in \Delta^{m_{\mathrm{X}}-1}$, we rewrite it as $H^{\dagger}(\boldsymbol{y};\boldsymbol{x}-\boldsymbol{x}^*) =$

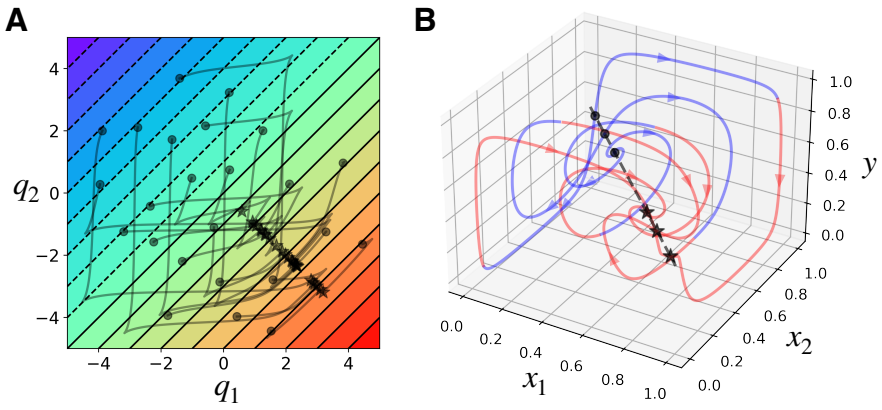


Figure 3: (A) Trajectories of q_1 and q_2. The rainbow contour plot indicates the value of $q_1 - q_2$. All the trajectories monotonically increase $q_1 - q_2$ with time and converge in the area of $q_1 > q_2$ in their final states. (B) Trajectories of the learning dynamics. The black broken line corresponds to the region of Nash equilibria, $x^{\mathrm{st}} = y = (1/2, 1/2)$. Each colored line shows a trajectory of the learning dynamics. First, the circle markers show the initial states. Following the blue lines, the trajectories diverge from the Nash equilibria ($D(X, y)$ increases with time). However, the trajectories stop to diverge and switch to converge to the Nash equilibria ($D(X, y)$ decreases), following the red lines. The star markers are the final states and correspond to one of the Nash equilibria.

$|u^{\mathrm{st}}(\boldsymbol{x}, \boldsymbol{y}) - u^*|^2$. Here, $|u^{\mathrm{st}}(\boldsymbol{x}, \boldsymbol{y}) - u^*|$ indicates the difference in payoff from the equilibrium, i.e., the exploitation by X to Y. Therefore, the correspondence between X's strategy and its payoff matrix becomes larger according to the exploitability.

Remark: We have interpreted both $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ and $\boldsymbol{U} \log \boldsymbol{X}^{\mathrm{T}}$ as the degree of correspondence between X's strategy and its payoff matrix. The interpretation is qualitatively true, but we remark that there are several quantitative differences between $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ and $\boldsymbol{U} \log \boldsymbol{X}^{\mathrm{T}}$, such as the order of multiplication and the existence of logarithm.

4 APPLICATION: GLOBAL CONVERGENCE

By combining the global behaviors obtained from Thms. 1 and 2, we expect that the global convergence to the Nash equilibrium occurs regardless of X's and Y's initial strategies. Thm. 2 shows that as long as Y's strategy is out of equilibrium, the correspondence between X's strategy and its payoff matrix continues to be stronger. Afterward, Thm. 1 shows that if the correspondence is sufficiently strong, Y's strategy is induced to the Nash equilibrium. In the following, our theory and experiment support that such global convergence occurs.

4.1 Example: Matching Pennies

Let us define the matching pennies game (see Fig. 2 for the illustration of its payoff matrix). This game considers the action numbers of $m_{\mathrm{X}} = m_{\mathrm{Y}} = 2$ and the payoff matrix of $\boldsymbol{U} = ((u_{11}, u_{12}), (u_{21}, u_{22})) = ((+1, -1), (-1, +1))$. The Nash equilibrium of this game is only $\boldsymbol{x}^* = \boldsymbol{y}^* = (1/2, 1/2)$. This game is the simplest example of a game equipped with a full-support Nash equilibrium. In addition, it has been known that the replicator dynamics in games without memories cycle around the Nash equilibrium and cannot reach the equilibrium. Nevertheless, by considering the memory asymmetry (Eqs. (5) and (6)), Y's strategy succeeds in the convergence to the equilibrium, as shown in the following corollary (see Appendix A.3 for its full proof). For convenience, we use a special notation available in two-action games; $(x_{1|j}, x_{2|j}) =: (x_j, 1 - x_j)$, $q_j := \log x_j - \log(1 - x_j)$, and $(y_1, y_2) =: (y, 1 - y)$.

Corollary 1 (Global convergence in matching pennies). *In the matching pennies game $\boldsymbol{U} = ((+1, -1), (-1, +1))$, Y's strategy $\boldsymbol{y}$ converges to the equilibrium $\boldsymbol{y}^*$, independent of both the players' initial strategies.*

Proof Sketch. First, note that $q_1 > q_2 \Leftrightarrow x_1 > x_2$. By the direct calculation, we prove $H(\boldsymbol{X}; \boldsymbol{\delta}) \propto q_1 - q_2$. Thm. 2 shows that as long as $\boldsymbol{y} = \boldsymbol{y}^*$, $H(\boldsymbol{X}; \boldsymbol{\delta})$ continues to increase. Thus, after a sufficiently long time, $H(\boldsymbol{X}; \boldsymbol{\delta}) > 0 \Leftrightarrow q_1 > q_2 \Leftrightarrow x_1 > x_2$ continue to hold. We can also prove that $x_1 > x_2$ is equivalent to the positive definiteness of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$. Thm. 1 shows that under positive definite $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$, $\boldsymbol{y}$ asymptotically converges to $\boldsymbol{y}^*$, its equilibrium strategy. □

Our experiments visualize the mechanism of the global convergence based on Thm. 1 and 2. Fig. 3A shows the dynamics of q_1 and q_2. Here, the colors indicate the contour plot for $q_1 - q_2$, showing that $H(\boldsymbol{X}; \boldsymbol{\delta}) \propto q_1 - q_2$ monotonically increases with time and thus Thm. 2 holds. Furthermore, we also see that X's strategy reaches the region of $q_1 > q_2 \Leftrightarrow x_1 > x_2$ after a sufficiently long time passes. In the region, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$ is positive definite, and thus Thm. 1 is applicable after sufficiently long time passes.

Next, Fig. 3B plots the global behavior of the learning dynamics, which is described by the three parameters of x_1, x_2, and y. The gray line shows the region that corresponds to the Nash equilibrium, i.e., $\boldsymbol{x}^{\mathrm{st}}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{y} = (1/2, 1/2)$. Furthermore, the colored lines show example trajectories of the learning dynamics. The blue part of the line shows that $D(\boldsymbol{X}, \boldsymbol{y})$ increases at the beginning of the learning dynamics. This part is in the region of $x_1 < x_2$, following Thm. 1. After that, the red part shows that $D(\boldsymbol{X}, \boldsymbol{y})$ decreases, and

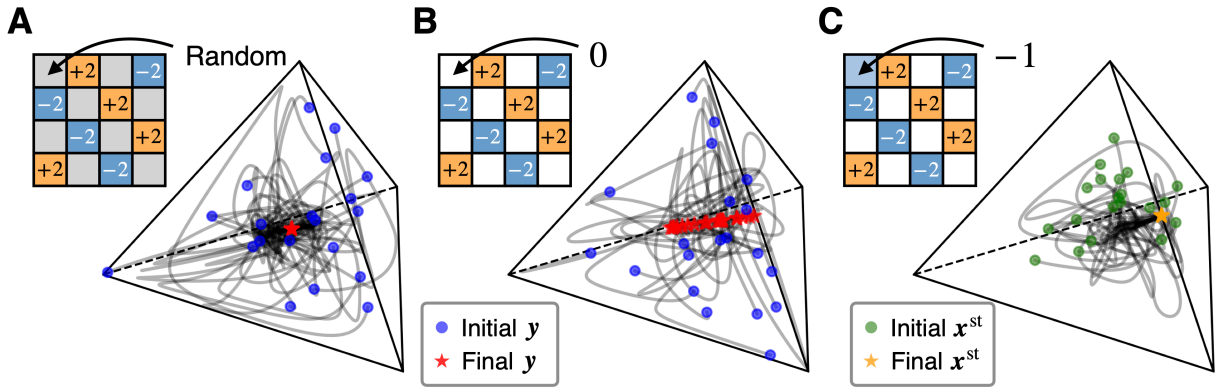


Figure 4: Global convergence in the coupled matching pennies games, where the second, third, fourth, and first actions win the other's first, second, third, and fourth actions, respectively. The winner receives the payoff of 2 (the orange blocks in the matrices for the winning of X), while the loser sends the payoff of 2 (the blue blocks). We now introduce three variants for the other blocks in the payoff matrix. (A) The case of interior equilibrium. We set each of the other blocks by random numbers in $[-1, 1]$ (the gray blocks). Then, Y's strategy converges to the unique Nash equilibrium (the red star marker) independent of its initial state (the blue circle markers). (B) The case of continuous equilibrium. We set each of the other blocks by 0, where the payoff matrix degenerates. Y's strategy converges to one of the Nash equilibria (the line consisting of the red star markers) depending on its initial state. (C) The case of boundary equilibrium. Only the block for the interaction between an action is set to -1, and the others are 0. If so, X's strategy converges to the unique Nash equilibrium (the orange star markers) independent of its initial state (the green circle markers). Instead, Y's strategies do not converge.

the learning dynamics converge to the equilibrium. This part is in the region of $x_1 > x_2$, following Thm. 1.

4.2 Example: Coupled Matching Pennies

We also observe the global convergence in other zero-sum games beyond the matching pennies game. Fig. 4 considers three examples of "coupled" matching pennies games, where a pair of matching pennies games are coupled with some interaction. The game considers the action numbers of $m_X = m_Y = 4$, and some elements of the payoff matrix is fixed as $u_{ij} = +2$ for $j = \sigma(i)$ and $u_{ij} = -2$ for $i = \sigma(j)$. Here, we used the permutation function σ as $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 4$, and $\sigma(4) = 1$. If we consider only X's odd actions and Y's even actions or its reverse, the payoff matrix corresponds to the matching pennies game. Interestingly, there are various types of Nash equilibrium depending on the interaction between these matching pennies games, i.e., the other elements of the payoff matrix, u_{ij} for neither $j = \sigma(i)$ nor $i = \sigma(j)$. Indeed, Fig. 4 shows three cases where Nash equilibrium exists (A) in the interior, (B) continuously, and (C) on the boundary.

We remark that the global behavior obtained from Thm. 1 and 2 are available even though the trajectories of the learning dynamics look complicated (see Fig. 4). Furthermore, Panels A and B show that Y's strategy converges to the equilibrium. The only exception is Panel C, but X's strategy converges to the equilibrium instead of Y's. In the following, we explain in detail this convergence for each panel.

Interior equilibrium: First, Fig. 4A shows the case where the other elements of the payoff matrix are random numbers following the uniform distribution of $[-1, 1]$. In this case, the payoff matrix is linearly independent. In mathematics, there exists no $\boldsymbol{a} = (a_j)_{1 \le j \le m_Y} \in \mathbb{R}^{m_Y}$ such that $\Sigma_j a_j \boldsymbol{u}_j = \boldsymbol{0}$ other than $\boldsymbol{a} = \boldsymbol{0}$. Thus, there is a single Nash equilibrium in the interior of the strategy space. As in the matching pennies game, we observe that Y's strategy always converges to its equilibrium independent of X's and Y's initial strategies.

Continuous equilibrium: Second, Fig. 4B shows the case where all the other elements take 0. In this case, the payoff matrix is not linearly independent in two ways. Indeed, $\Sigma_j a_j \boldsymbol{u}_j = \boldsymbol{0}$ for $\boldsymbol{a} = (0, 1, 0, 1)$ and $(1, 0, 1, 0)$. Thus, Nash equilibria exist continuously as $\boldsymbol{x}^* = r_X(0, 1/2, 0, 1/2) + (1 - r_X)(1/2, 0, 1/2, 0)$ and $\boldsymbol{y}^* = r_Y(0, 1/2, 0, 1/2) + (1 - r_Y)(1/2, 0, 1/2, 0)$ for all $0 \le r_X \le 1$ and $0 \le r_Y \le 1$. Even in such continuous equilibria, we observe that Y's strategy converges to one of the equilibria depending on X's and Y's initial strategies.

Boundary equilibrium: Third, Fig. 4C shows the case where the other elements take 0 in principle except for $u_{11} = -1$. In this case, the payoff matrix is not linearly independent in one way. Indeed, $\Sigma_j a_j \boldsymbol{u}_j = \boldsymbol{0}$ for $\boldsymbol{a} = (0, 1, 0, 1)$. The only Nash equilibrium exists on the boundary of strategy spaces, $\boldsymbol{x}^* = (0, 1/2, 0, 1/2)$, and $\boldsymbol{y}^* = (1/2, 0, 1/2, 0)$. As far as we observe our experiments, Y's strategy fails to converge the equilibrium when the payoff matrix is not linearly independent and is equipped with only the boundary Nash equilibrium. Nevertheless, we observe that X's strategy converges to the equilibrium instead of Y's. We find no case that neither X's nor Y's strategy converges to the equilibrium.

5 CONCLUSION

This study considered the simplest situation of memory asymmetry between two players; only player X memorizes the other's previous

action, while player Y cannot. We formulated their learning dynamics based on the replicator dynamics. Although the existence of memory complicates the dynamics, we captured the global behavior of the learning dynamics by introducing two new quantities. One is the conditional-sum divergence, which is an extension of the previous divergence to the case of reactive strategies. We proved that when X exploits Y, this conditional-sum divergence becomes smaller, meaning that their strategies converge to the Nash equilibrium. The other is a family of Lyapunov functions, meaning X's exploitability to Y. We proved that these Lyapunov functions monotonically increase, meaning that X learns to exploit Y with time. As a valid application of the combination of these two quantities, we suggested the global convergence to the Nash equilibrium. Theoretically, we proved the global convergence in the matching pennies game, the simplest game equipped with an interior Nash equilibrium. Our experiments further support that global convergence occurs in coupled matching pennies games, which can have various types of Nash equilibrium structures, such as interior equilibrium, continuous equilibrium, and boundary equilibrium. It is still a conjecture whether the learning dynamics with memory asymmetry converge to the Nash equilibrium in general zero-sum games. This study provides novel and valid indicators to analyze dynamics in learning in games with memories.

ACKNOWLEDGMENTS

K. Ariu is supported by JSPS KAKENHI Grant No. 23K19986.

REFERENCES

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing* 8, 1 (2012), 121–164.
[2] Seung Ki Baek, Hyeong-Chai Jeong, Christian Hilbe, and Martin A Nowak. 2016. Comparing reactive and memory-one strategies of direct reciprocity. *Scientific reports* 6, 1 (2016), 25676.
[3] James P Bailey and Georgios Piliouras. 2018. Multiplicative weights update in zero-sum games. In *EC*.
[4] Wolfram Barfuss. 2020. Reinforcement learning dynamics in the infinite memory limit. In *AAMAS*.
[5] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research* 53 (2015), 659–697.
[6] Tilman Börgers and Rajiv Sarin. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77, 1 (1997), 1–14.
[7] Michael Bowling. 2004. Convergence and no-regret in multiagent learning. In *NeurIPS*.
[8] Michael Bowling and Manuela Veloso. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136, 2 (2002), 215–250.
[9] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.
[10] Daniel Friedman. 1991. Evolutionary games in economics. *Econometrica: Journal of the Econometric Society* (1991), 637–666.
[11] Drew Fudenberg and David K Levine. 1998. *The theory of learning in games*. Vol. 2. MIT press.
[12] Drew Fudenberg and Eric Maskin. 2009. The folk theorem in repeated games with discounting or with incomplete information. In *A long-run collaboration on long-run games*. World Scientific, 209–230.
[13] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. 2023. Learning in Multi-Memory Games Triggers Complex Dynamics Diverging from Nash Equilibrium. In *IJCAI*.
[14] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. 2024. Memory Asymmetry Creates Heteroclinic Orbits to Nash Equilibrium in Learning in Zero-Sum Games. In *AAAI*.
[15] Yuma Fujimoto and Kunihiko Kaneko. 2019. Emergence of exploitation as symmetry breaking in iterated prisoner's dilemma. *Physical Review Research* 1, 3 (2019), 033077.
[16] Yuma Fujimoto and Kunihiko Kaneko. 2021. Exploitation by asymmetry of information reference in coevolutionary learning in prisoner's dilemma game. *Journal of Physics: Complexity* 2, 4 (2021), 045007.
[17] Josef Hofbauer, Karl Sigmund, et al. 1998. *Evolutionary games and population dynamics*. Cambridge university press.
[18] Aamal Hussain, Francesco Belardinelli, and Georgios Piliouras. 2023. Beyond strict competition: approximate convergence of multi-agent Q-learning dynamics. In *IJCAI*.
[19] Aamal Abbas Hussain, Francesco Belardinelli, and Georgios Piliouras. 2023. Asymptotic Convergence and Performance of Multi-Agent Q-learning Dynamics. In *AAMAS*.
[20] Naoki Masuda and Hisashi Ohtsuki. 2009. A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bulletin of mathematical biology* 71 (2009), 1818–1850.
[21] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. 2018. Cycles in adversarial regularized learning. In *SODA*.
[22] Panayotis Mertikopoulos and William H Sandholm. 2016. Learning in games via reinforcement and regularization. *Mathematics of Operations Research* 41, 4 (2016), 1297–1324.
[23] Janusz M Meylahn, Lars Janssen, et al. 2022. Limiting dynamics for Q-learning with memory one in symmetric two-player, two-action games. *Complexity* 2022 (2022).
[24] John F Nash Jr. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36, 1 (1950), 48–49.
[25] Martin Nowak. 1990. Stochastic strategies in the prisoner's dilemma. *Theoretical population biology* 38, 1 (1990), 93–112.
[26] Martin A Nowak and Karl Sigmund. 2004. Evolutionary dynamics of biological games. *Science* 303, 5659 (2004), 793–799.
[27] Hisashi Ohtsuki. 2004. Reactive strategies in indirect reciprocity. *Journal of Theoretical Biology* 227, 3 (2004), 299–314.
[28] Georgios Piliouras, Carlos Nieto-Granda, Henrik I Christensen, and Jeff S Shamma. 2014. Persistent patterns: Multi-agent learning beyond equilibrium and utility. In *AAMAS*.
[29] Georgios Piliouras and Jeff S Shamma. 2014. Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence. In *SODA*.
[30] Laura Schmid, Christian Hilbe, Krishnendu Chatterjee, and Martin A Nowak. 2022. Direct reciprocity between individuals that use different strategy spaces. *PLoS Computational Biology* 18, 6 (2022), e1010149.
[31] Satinder Singh, Michael J Kearns, and Yishay Mansour. 2000. Nash Convergence of Gradient Dynamics in General-Sum Games.. In *UAI*.
[32] Peter D Taylor and Leo B Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical biosciences* 40, 1-2 (1978), 145–156.
[33] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. 2006. An evolutionary dynamical analysis of multi-agent learning in iterated games. In *AAMAS*.
[34] Karl Tuyls and Gerhard Weiss. 2012. Multiagent learning: Basics, challenges, and prospects. *Ai Magazine* 33, 3 (2012), 41–41.
[35] Masahiko Ueda. 2023. Memory-two strategies forming symmetric mutual reinforcement learning equilibrium in repeated prisoners' dilemma game. *Appl. Math. Comput.* 444 (2023), 127819.
[36] Yuki Usui and Masahiko Ueda. 2021. Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma. *Appl. Math. Comput.* 409 (2021), 126370.
[37] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.
[38] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*.