Tractable Infinite Order Markov Analysis for Iterated Games with Learners

Shohei Hidaka, Takuma Torii, and Akira Masumi
School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan 923–1293
Email: {shhidaka, tak.torii, a-masumi}@jaist.ac.jp

***Abstract*—The theory of games involving players who adaptively learn from their past experiences is not yet well understood. We analyze games in which players make on each turn a probabilistic choice of actions determined by a k^{th}-order Markov process which signifies how they learn from their past k actions for a fixed number k. As the number of states in such Markov processes grows exponentially with k, the analysis of games involving learners with long memories has been viewed as computationally intractable. This study develops a technique which enables feasible analysis of these long-memory Markov process. We further show that, for two players involved in an iterated prisoners' dilemma, the probability of mutual defection increases with the size of their memories. This result is consistent with the classical prisoners' dilemma with two rational players.**

I. Games with adaptive learners

Conventional game theory treats each player as a rational decision maker who is making decisions based on sufficient information regarding the game being played. The dynamics in such a game are generally characterized by its Nash equilibria — states in which none of the agents may profit by changing their actions [1].

Real social problems, however, are often vastly more complex than such formulations [2], [3], [4], [5]. In reality, each agent has limited computational resources and limited information about the game it is playing. Under such constraints, learning – inductive inference of future behaviour from a limited experience of past behaviour – plays a crucial role in finding locally optimal actions. One key question regarding games in which agents have limited information but employ learning techniques regards their long-term equilibria [3].

A class of iterative games with *reinforcement learning* [6] has been investigated in both theoretical [7], [8] and empirical studies [2], [3], [4], [5]. In this class of games, the only information available to each agent is a number of its own actions and the rewards to these actions. The probabilities for the players' next actions are computed according to the weighted averages of rewards for the possible actions.

As each player's decision is probabilistic, a game of this class in which players have access to their history going back k turns can be formulated as a k^{th}-order finite Markov process. The equilibria of such a game are then characterized by its stationary distributions. Althogh this formulation is mathematically simple, not many past studies have taken this approach to the analysis of iterative games with reinforcement learning. The reason for this is that the number of states in an iterated game between learners with k-step memory grows exponentially with k. To the best of our knowledge, only few special cases, such as those in which $k = 1$ [9] and some specific games in which k is large [10], have been analyzed.

The present study shows a new technique which approximately computes a marginal stationary distribution of a k^{th}-order Markov process. This technique is applied to an iterated game of prisoners' dilemma with learning players. Our analysis demonstrates the utility of the proposed computational technique as a numerical tool for the analysis of dynamic games with learners.

In Section II, we describe the iterated version of prisoners' dilemma and reinforcement learning. In Section III, we introduce a Markov process formalism for the iterated prisoners' dilemma. In Section IV, we present the main result of the paper about the marginal stationary distribution of an infinite order Markov process. In Section V, we demonstrate an application of the new computational technique for an iterated prisoners' dilemma.

II. Iterated prisoners' dilemma

The prisoners' dilemma is a classical game which has long been used as a minimal model demonstrating difficulty of mutual cooperation. In prisoners' dilemma, each of two players chooses an action from either Cooperation (C) or Defection (D), and each player is given a certain payoff depending on the actions of both players. Each player benefits fairly when both choose Cooperation (CC). However, one player can gain even more by choosing Defection on the condition that the other chooses Cooperation. This incentivizes Defection for both players, and the game results in mutual Defection (DD) with individual payoffs lower than those in the mutual Cooperation case. This is the only Nash equilibrium in the prisoners' dilemma. The "dilemma" is that two rational players cannot escape from the mutual Defection with unhappy payoffs although there is a more beneficial option in mutual Cooperation.

The basic game has been extended to games with multiple agents, iterated steps, stochastic strategies, situations affected by noise, and certain topologies for agent interactions [11], [12], [13]. In the iterative variant of models, each agent can adaptively choose its action on the basis of a series of past actions and payoffs. One of the simplest cases is completely analyzed based on the finite Markov formalism [9], but more general cases remain open for further research.

978-1-4799-5955-6/14/$31.00 ©2014 IEEE

In a recent study [14], we analyzed the iterated prisoners' dilemma (IPD) with two probabilistic learners. This analysis suggests that IPD with two learners can result in mutual cooperation, but this is limited only to learners with short memories due to the previously mentioned computational difficulty involved in analyzing iterated games involving players with long memories. The present study develops a new computational tool which extends this analysis to learners with long memories.

A. IPD with reinforcement learners

We briefly outline the IPD with learners who learn using reinforcement learning.

Definition 1 (Iterated prisoners' dilemma)**.** *We label the players* 1 *and* 2*, and we label the two moves available to each player at each step as* 0 *(for Cooperation) and* 1 *(for Defection). For each integer* t *and for each* $i = 1, 2$*, we denote by* $x_{t,i}$ *the choice of action that player* i *made on turn* t*. Define a function* $f : \{0,1\}^2 \to \{1,2,3,4\}$ *by*

$$f(a_1, a_2) := 2a_1 + a_2 + 1. \tag{1}$$

The function f *encodes each of the four possible outcomes on a single turn of the game as the integers* $1, 2, 3,$ *and* 4*. We define* $X_t := f(x_{t,1}, x_{t,2})$*. We write* $X_t^{t+k-1} := (X_t, X_{t+1}, \ldots, X_{t+k-1})$ *for a sequence of* k *states from the step* t*.*

We assume the existence of a payoff map $r = (r_1, r_2) : \{1,2,3,4\} \to \mathbb{R}^2$ *such that* $r_i(X_t)$ *is the payoff to the* i^{th} *agent resulting from the actions taken by both agents at turn* t*. Note that, under our assumption, the payoff scheme remains constant over time. We further assume that the payoffs are symmetric so that, for any* $a, b \in \{0,1\}$*,*

$$r_1(f(a,b)) = r_2(f(b,a)) =: R_{a,b}.$$

An iterated prisoners' dilemma is a game which satisfies these assumptions as well as the inequalities

$$R_{01} < R_{11} < R_{00} < R_{10}, \text{ and}$$

$$R_{01} + R_{10} < 2R_{00}.$$

Definition 2 (IPD with reinforcement learners)**.** *In the reinforcement learning model, we begin with an IPD as defined above. However, each agent chooses an action based on a function of the rewards it received for its past* k *actions, with* k *specified at the outset. Specifically, for agent* i*, this function is*

$$\phi_{i,x}^{\alpha_i}\left(X_{t-k}^{t-1}\right) = \sum_{s=1}^{k} \alpha_i^s \delta_{x,x_{t-s,i}} r_i(X_{t-s}),$$

where $\alpha_i \in [0,1]$ *is a memory-retention parameter, and* $\delta_{x,y}$ *is the Kronecker delta, which takes the value 1 when* x *and* y *agree and is 0 otherwise.*

Using this weighted rewards function along with a sensitivity parameter $\beta_i \geq 0$*, the probability that the* i^{th} *agent chooses action* x *at step* t *is*

$$P\left(x \mid X_{t-k}^{t-1}\right) = \frac{\exp\left(\beta_i \phi_{i,x}^{\alpha_i}\left(X_{t-k}^{t-1}\right)\right)}{\sum_{x=0}^{1} \exp\left(\beta_i \phi_{i,x}^{\alpha_i}\left(X_{t-k}^{t-1}\right)\right)} \tag{2}$$

We assume that the agents choose their actions independently at each turn so that, for any $(a_1, a_2) \in \{0,1\}^2$*,*

$$P\left(X_t = f(a_1, a_2) \mid X_{t-k}^{t-1}\right) = \prod_{i=1}^{2} P\left(a_i \mid X_{t-k}^{t-1}\right). \tag{3}$$

III. Markov process formulation

Equation (3) allows us to calculate the conditional probabilities $P(X_{t-k+1}^t \mid X_{t-k}^{t-1})$. In this manner, we construct a Markov chain with states consisting of all possible length k move sequences of the players in an IPD with reinforcement learning. This is the k^{th}-order Markov process corresponding to the variables X_t. We encode the states in this k^{th}-order Markov process with the integers $1, 2, \ldots, 4^k$ and describe its transition matrix with respect to this encoding as follows:

Definition 3 (Transition matrix)**.** *For any integer* t*, the indexing map*

$$h_k\left(X_{t-k+1}^t\right) := 1 + \sum_{j=0}^{k-1} (X_{t-k+1+j} - 1)4^j. \tag{4}$$

assigns a unique integer $1 \leq i \leq 4^k$ *to each of the states* X_{t-k+1}^t*.*

For $1 \leq j \leq 4^k$*, the set*

$$\mathcal{H}_j := \left\{h_k(X_1^k) : j = h_k\left(X_0^{k-1}\right)\right\}.$$

consists of the indices of those states in the k^{th}*-order Markov process which can be reached from the state indexed by* j*. Denote by* $M_n(\mathbb{R})$ *the set of* $n \times n$ *matrices with real entries. The transition matrix* $Q \in M_{4^k}(\mathbb{R})$ *for the* k^{th}*-order Markov process is defined by*

$$Q_{i,j} := P\left(h_k^{-1}(i) \mid h_k^{-1}(j)\right). \tag{5}$$

Observe that, unless $i \in \mathcal{H}_j$*,* $Q_{i,j} = 0$*.*

For $t \in \mathbb{Z}$, write

$$p_t = \left(P(h_k(X_t^{t+k-1}) = 1), \ldots, P(h_k(X_t^{t+k-1}) = N^k)\right)^T,$$

and suppose we start with some initial probability vector p_0. Then, p_t is obtained by $p_t = Qp_{t-1}$ for $t > 0$. Applying this infinitely many steps, we obtain the *stationary probability distribution*

$$p_\infty = \lim_{t \to \infty} Q^t p_0, \tag{6}$$

if the limit exists. The Perron-Frobenius theorem [15] concerns the existence of this limit (6). In this study, unless otherwise specified, we simply assume the existence of a stationary distribution.

A. The exchangable case ($\alpha_i = 1$)

In general, the computation of the stationary vector for $Q \in M_{4^k}(\mathbb{R})$ has complexity $O(4^{k+1})$ [10]. If $\alpha_i = 1$ for each $i = 1, 2$, however, we can compute a stationary distribution more efficiently. Exploiting the exchangability of the actions in a state sequence X_t^{t+k-1}, the size of the state space of this special case is reduced to $(k+1)k(k-1)$, which is much smaller than the size of the original state space, 4^k. In [10], we showed that a stationary vector for this special case could be computed efficiently for even relatively large values of k.

B. Block-diagonalization of transition matrix

For further extension, we present a mathematical property of the k^{th}-order transition matrix Q defined in (5). The following theorem [10] shows that a k^{th} order transition matrix can be rewritten as a product of a permutation matrix and block-diagonal matrix. We set our notation before stating the theorem.

Let us write the number of states N. We have $N = 4$ in he IPD introduced in the previous section. Let us denote the identity matrix by $E_N \in \mathbb{R}^{N\times N}$ and the unit vector by

$$e_{N,i} := (0, \ldots, 0, \overset{i}{\overset{\vee}{1}}, 0, \ldots, 0)^T \in \mathbb{R}^N,$$

zero vector by $\mathbf{0} = (0, 0, \ldots, 0)^T$, and

$$E_{N,i} := (\mathbf{0}, \ldots, \mathbf{0}, \overset{i}{\overset{\vee}{e_{N,i}}}, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{R}^{N\times N}.$$

We define a special permutation matrix called *commutation matrix* by [16]:

$$C_{M,N} := \sum_{i=1}^{N} E_M \otimes E_{N,i} \tag{7}$$

where $\otimes$ denotes Kronecker product. For an $M \times N$ matrix X, we write

$$\text{vec}\,(X) := (X_{1,1}, \ldots, X_{M,1}, \ldots, X_{1,N}, \ldots, X_{M,N})^T,$$

and we call $\text{vec}\,(X)$ the vectorization of X. The commutation matrix satisfies

$$\text{vec}\,(X) = C_{M,N}\text{vec}\left(X^T\right).$$

Theorem 1. *Let $\lambda \in \mathbb{R}$ and $\theta \in \mathbb{R}^{N^k}$ be the eigenvalue and its corresponding eigenvector of the k^{th}-order transition matrix Q with states N defined in (5). Then, it has the following decomposition:*

$$Q = C_{N,N^{k-1}} \sum_{i=1}^{N^{k-1}} E_{N^{k-1},i} \otimes Q_i$$

where the block diagonal matrix $Q_m \in M_N(\mathbb{R})$ has its (i, j) element

$$\{Q_m\}_{i,j} = Q_{a,b}$$

where

$$a = f_k(i, N(m-1)+j),\ b = N(m-1)+j.$$

With this theorem, we obtain the following result.

Corollary 1. *For an arbitrary vector $x \in \mathbb{R}^{N^k}$ and transition matrix $Q \in \mathbb{R}^{N^k\times N^k}$ with its block diagonal matrices $Q_i \in \mathbb{R}^{N\times N}, i = 1, \ldots, N^{k-1}$,*

$$Qx = \text{vec}\left((Q_1x_1, Q_2x_2, \ldots, Q_{N^{k-1}}x_{N^{k-1}})^T\right) \tag{8}$$

where $x_i \in \mathbb{R}^N (i = 1, \ldots, N^{k-1})$ satisfies

$$x = \text{vec}\left((x_1, x_2, \ldots, x_{N^{k-1}})\right).$$

IV. m-SHIFT STATIONARY DISTRIBUTION

As the number, 4^k, of states of the k^{th}-order Markov process defined in Section III grows exponentially, we cannot in practice compute the stationary distributions of such processes for large values of k. This is a major obstacle when in the analysis of games like IPD with reinforcement learning introduced in Section I.

In analyzing a $(k+m)^{\text{th}}$ order Markov process, it is sometimes sufficient to calculate its *marginal* stationary distribution conditional upon the states in the corresponding k^{th}-order process. This is a probability over series of k observations, rather than its full stationary distribution over series of $(k+m)$ observations.

We define the k^{th}-order marginal stationary distribution to be the probability distribution over the final k observations in states of the $(k+m)^{th}$-order Markov chain obtained by taking the expectation of the corresponding state probabilities over all possible initial states in the k^{th}-order Markov chain. The probability vector corresponding to this distribution is Write a stationary distribution

$$\theta_{k+m} = \lim_{t\to\infty}\left(P(h_{k+m}(X_{t+1}^{t+k+m}) = i)\right)_{i=0}^{N^{k+m}-1}.$$

Write $\text{mod}_k(x) := x \mod k$ and $\lfloor x \rfloor$ for the greatest integer $y \le x$. For each non-negative integer j, write

$$g_{N,k,m}(i) := 1 + \text{mod}_{N^k}\left(\lfloor\frac{i-1}{N^m}\rfloor\right).$$

The k^{th}-order marginal distribution of the $(k+m)^{\text{th}}$-order Markov process is

$$\theta_{k,m} := \lim_{t\to\infty}\left(P\Big(g_{N,k,m}\left(h_{k+m}(X_{t+1}^{t+k+m})\right) = i\Big)\right)_{i=1}^{N^k}.$$

For instance, analysis of the long-term dynamics of IPD involving reinforcement learners with long memories requires the first-order marginal stationary probabilities of the four possible outcomes on the final observations of states in the associated $(k+m)^{\text{th}}$-order Markov chain for large values of m. This first-order marginal stationary distribution shows how frequently each pair of moves (CC, CD, DC, DD) is to occur when the game involves two learners with memories of length k.

Suppose we wish to obtain a marginal distribution $\theta_{k,m}$ in the limit $m \to \infty$ for a small k. We consider a inductive construction in which, for $k < m$, a probability vector $x_k \in \mathbb{R}^{N^k}$ with k past steps is mapped to x_{k+1} at step $m-k$. Denote such a transition matrix from a probability vector $x_k \in \mathbb{R}^{N^k}$ to $x_{k+1} \in \mathbb{R}^{N^{k+1}}$ by $U_{N,k}$, and $x_{k+1} := U^{N,k}x_k$. Multiply an infinite series of transition matrix $U_{N,k}, U_{N,k+1}, \ldots$ with an arbitrary probability vector x_k, we obtained a stationary vector x_∞, which converges to θ_∞ uniquely under the regular condition. Then, consider the diagonalization matrix $Z_{N,k}$ with which $\theta_{k,1} = Z_{N,k}\theta_{k+1}$. The desired marginal distribution is obtained by applying these matrices to an arbitrary probability vector x_k

$$\theta_{k,\infty} = \lim_{m\to\infty} Z_{N,k} \ldots Z_{N,k+m-1}U_{N,k+m-1} \ldots U_{N,k}x_k.$$

Although an artibtrary x_k converges in the limit, it is obvious that $x_k = \theta_{k,\infty}$ converges at the fastest rate for a finite m. This motivates to define the m-shift stationary distribution

$$\omega_{k,m} := F_{k,m}\omega_{k,m}$$

of its transition matrix

$$F_{k,m} := Z_{N,k} \dots Z_{N,k+m-1} U_{N,k+m-1} \dots U_{N,k}.$$

Observe that the 1-shift transition matrix is exactly the k^{th}-order transition matrix, $F_{k,1} = Q^{(k)}$. We expect this m-shift stationary distribution converges the desired marginal distribution at the best rate. In the following, we give a specific form of $Z_{N,k}$ and $U_{N,k}$, and describe the mathematical properties of m-shift stationary distribution.

Definition 4 (Matrix form)**.** *Define the marginalization matrix*

$$Z_{N,k} := \mathbf{1}_N^T \otimes E_{N^k}.$$

For an arbitrary transition matrix $Q^{(k)} \in M_{N^k}(\mathbb{R})$ which is block diagonal with $Q_i^{(k)} \in M_N(\mathbb{R}), i = 1, \dots, N^{k-1}$, write $Q_i^{(k)} = \left(Q_{i,1}^{(k)}, Q_{i,2}^{(k)}, \dots, Q_{i,N}^{(k)}\right)$. With these matrices, we define the shift matrix

$$U_{N,k} := \sum_{j=1}^{N} \sum_{i=1}^{N^{k-1}} E_{N^{k-1},i} \otimes E_{N,j} \otimes Q_{i,j}^{(k)}.$$

For a k^{th}-order Markov process, the corresponding *m-shift transition matrix* $F_{k,m} \in M_{N^k}(\mathbb{R})$ is defined as follows:

Define

$$\begin{aligned}\mathcal{H}_{i,j,m} \quad := \quad & \Big\{ l : i = h_k\left((g(l, m+k-1), \dots, g(l,m))\right), \\ & j = h_k\left((g(l, k-1), \dots, g(l,0))\right) \Big\}.\end{aligned}$$

$$(F_{k,m})_{i,j} := \sum_{l \in \mathcal{H}_{i,j,m}} P\left(h_{k+m}^{-1}(l) \mid h_k^{-1}(j)\right). \tag{9}$$

Write

$$\mathcal{H}_{i,j} = \{l : 1 \le l \le N^{k+m}, g_{N,k,m}(l) = i, g_{N,k,0}(l) = j\}.$$

Proposition 1 (Elements of m-shift transition matrix)**.** *The m-shift transition matrix corresponding to the original k^{th}-order Markov chain, $F_{k,m} \in M_{N^k}(\mathbb{R})$ is defined by*

$$(F_{k,m})_{i,j} := \sum_{l \in \mathcal{H}_{i,j}} P\left(g_{N,k,m}(l) = i \mid g_{N,k,0}(l) = j\right). \tag{10}$$

An m-shift transition matrix can easily be expressed in two forms — as a product (as per Definition 4), or recursively (as per Proposition 2). These two forms are equivalent, but differ in terms of computational complexity. In general, the recursive form requires less space for computation but the factored form is more efficient in time. Theorem 1 is used to make explicit the factored and recursive forms of the m-shift transition matrix.

Proposition 2 (A recursive expression of m-shift transition matrix)**.** *For an arbitrary transition matrix $Q^{(l)} \in M_{N^l}(\mathbb{R})(l = k, k+1, \dots, k+m-1)$ with block diagonal consisting of matrices $Q_i^{(l)} \in M_N(\mathbb{R})$ where $i = 1, \dots, N^{l-1}$, write*

$$Q_i^{(l)} = \left(Q_{i,1}^{(l)}, Q_{i,2}^{(l)}, \dots, Q_{i,N}^{(l)}\right).$$

For $1 \le i \le N^{k+m-2}$, define $\overline{F}_i^{(k+m-1)} := Q_i^{(k+m-1)}$. For $k \le l < k+m-1$ and $1 \le i \le N^{l-1}$, define

$$\overline{F}_i^{(l)} := \left(\overline{F}_{N(i-1)+1}^{(l+1)} Q_{i,1}^{(l)}, \dots, \overline{F}_{N(i-1)+N}^{(l+1)} Q_{i,N}^{(l)}\right).$$

Then we have the recursive form of the m-shift transition matrix

$$F_{k,m} = \overline{F}_1^{(k)}.$$

The m-shift transition matrices can be used to estimate the k^{th}-order marginal stationary distribution:

Proposition 3. *Let $\theta_{k,\infty} \in \mathbb{R}^{N^k}$ be a marginal stationary distribution in the limit $m \to \infty$. Then,*

$$\theta_{k,\infty} = \lim_{m \to \infty} F_{k,m}\theta_{k,\infty}.$$

This can be rephrased using the recursive expression for the m-shift transition matrix as stating that, for every $1 \le i \le N^{m-k}$,

$$\limsup_{m \to \infty} \left\| \overline{F}_i^{(m)} - Q_i^{(m)} \right\| = 0. \tag{11}$$

Moreover, the error term of (11) is a non-increasing function of m.

Proposition 3 motivates to approximate the limiting marginal distribution $\theta_{k,\infty}$ with m-shift stationary distribution $\omega_{k,m}$ with a finite m via the following corollary.

Corollary 2. *By Proposition 3, there is a series of distributions $\omega_{k,1}, \omega_{k,2}, \dots$ corresponding to $Q^{(k)}, Q^{(k+1)}, \dots$ for which*

$$\limsup_{m \to \infty} \|\omega_{k,m} - \theta_{k,m}\| = 0.$$

V. Numerical Experiments

We present an analysis of the iterated prisoners' dilemma with rewards $R_{00} = 1, R_{01} = -2, R_{10} = 2$, $R_{11} = 0$ and where the players are reinforcement learners with identical sensitivity parameters $\beta_1 = \beta_2 = \beta = 1/2$, and identical memory retention parameters $\alpha_1 = \alpha_2 = \alpha \in [0, 1]$.

Computing the m-shift transition matrix and its stationary vector, we obtained the marginal stationary probabilities of mutual cooperation ($P(CC)$), mutual defection ($P(DD)$), and one-side defection ($P(CD) = P(DC)$, with equality due to the symmetry between two players). On our best computational resource, we set $m = 12$. The calculation of these stationary probabilities takes on the order $4^{m+1} \approx 10^{7.82}$ steps, and is pushing the boundaries of what we can reasonably compute. However, we suspect that a more sophisticated algorithm, using the recursive form of the m-step transition matrix shown in Proposition 2, can mitigate this computational problem to some extent.

Figure 1 shows the marginal stationary probabilities estimated by the m-shift stationary probabilities as functions of the memory retention parameter α. The multiple lines of the same color show the marginal probabilities for different values of the shift m for a fixed value of α. The arrows indicate

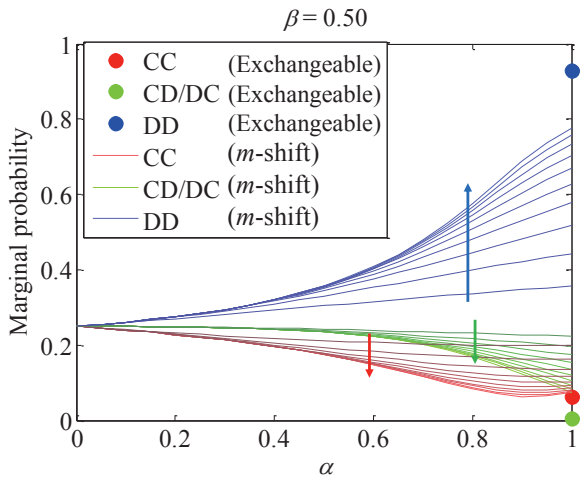

Fig. 1. The m-shift stationary distributions (curves) and marginal distributions for $\alpha = 1$ and $k = 1000$ (circles) as functions of the memory retention parameter α.

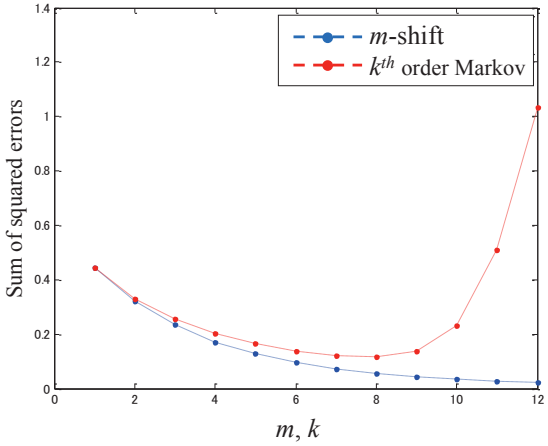

Fig. 2. The sum of squared errors of the m-shift stationary distributions (blue) and the k^{th} order Markov process as functions of m or k, respectively, by taking the corresponding marginal distribution for $\alpha = 1$ and $k = 1000$ as normative values.

the directions in which these groups of lines changes from $m = 1, 2, \ldots, 12$.

Observe that the probability for mutual defection $P(DD)$ increases with the memory retention parameter α. This result is qualitatively consistent with the outcome of the classical prisoners' dilemma with two rational players. Our analysis illustrates the counterpart of the classical Nash equilibrium in the iterated version of the game with probabilistic reasoners capable of remembering all the previous outcomes. Interpreting the memory retention parameter α as the degree of rationality of the agents, this indicates that, as players become more rational, they are more likely to mutually defect.

For the exchangable case $\alpha = 1$, the special computational procedure described in Section III-A can be used even for large values of k. We used this procedure to perform an analysis of the exchangable case with $k = 1000$. In our analysis the estimated marginal distributions appear to converge. The results of this analysis are indicated by the filled circles in Figure 1.

Treating the estimates obtained using the special property of $\alpha = 1$ as a the true stationary distributions, we analyzed the sum of squared errors (SSE) in the estimated m-shift stationary distributions with $\alpha = 1$. The blue line in Figure 2 shows these SSEs as a function of m. As expected from Proposition 3, the SSE is a decreasing function of m. This error analysis validates the statement of Corollary 2, that the m-shift stationary distribution approaches a marginal stationary distribution in the limit $m \to \infty$.

In theory, as $k \to \infty$, the difference between the k^{th}-order stationary distributions and the corresponding m-shifted marginal stationary distributions could vanish in the limit. Such convergence, however, is not obvious. The red line in Figure 2 shows the sum of squared errors of the marginal distributions calculated by the k^{th} order Markov process. The errors of the k^{th} order Markov process shows slower congergence (higher errors at each $k = m$) than the corresponding m-shift stationary distribution. More importantly, it shows the error is a non-monotonic function of k. This suggests that, regarding a marginal stationary distribution, a higher-order Markov process analysis does not always give a better approximation to stationary distributions in the limit $k \to \infty$ than a lower-order one.

The non-monotonic error in estimation using k^{th}-order Markov processes also suggests that a certain range of finite k may show a special pattern in its maginal stationary distribution. In fact, our recent work [14] shows a consistent result with this result: the IPD of a relatively high-order Markov process $k \approx 10$ tends to exhibit mutual cooperation in its long-term dynamics, as compared to mutual defection dynamics when k is either smaller or very, very large. Combined with the analysis of the present study, this suggests that a certain range of memory lengths k enables mutual cooperation, but learning with very large memory lengths results in mutual defection, as in the Nash equilibrium of classical prisoners' dilemma.

VI. Concluding Remarks

When analyzing a game with probabilistic learners, higher order Markov processes are unavoidable. Often, however, it is information about the marginal stationary distribution which one truly desires. The numerical evaluation of the previous section confirms the computational advantage of using m-shift stationary distributions over the stationary distributions of the corresponding k^{th}-order Markov chains. Our technique is not limited to the current specific case, and is applicable in general to the analysis of any higher order Markov process.

Acknowledgment

The authors would like to thank Dr. Neeraj Kashyap for his helpful clarification of this paper. This study was supported by the NeuroCreative Lab, Grant-in-Aid for Scientific Research B No. 23300099, and Grant-in-Aid for Exploratory Research No. 25560297.

REFERENCES

[1] J. F. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[2] C. F. Camerer and T. Hua Ho, "Experience-weighted attraction learning in normal form games," *Econometrica*, vol. 67, no. 4, pp. 827–874, 1999.

[3] C. F. Camerer, "Behavioural studies of strategic thinking in games," *Trends in Cognitive Sciences*, vol. 7, no. 5, pp. 225–231, 2003.

[4] C. Camerer, *Behavioral game theory*. New Age International, 2010.

[5] A. E. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, vol. 8, no. 1, pp. 164–212, 1995.

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

[7] T. Borgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *Journal of Economic Theory*, vol. 77, no. 1, pp. 1–14, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002205319792319X

[8] Y. Sato, E. Akiyama, and J. D. Farmer, "Chaos in learning a simple two-person game," *Proceedings of the National Academy of Sciences*, vol. 99, no. 7, pp. 4748–4751, 2002.

[9] M. Nowak, "Stochastic strategies in the prisoner's dilemma," *Theoretical Population Biology*, vol. 38, no. 1, pp. 93–112, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/004058099090005G

[10] S. Hidaka, T. Torii, and A. Masumi, "Which types of learning make a simple game complex?" under review.

[11] R. Axelrod, "The evolution of cooperation," 1984.

[12] T. Galla, "Intrinsic noise in game dynamical learning," *Physical Review Letters*, vol. 103, no. 19, p. 198702, 2009.

[13] T. W. Sandholm and R. H. Crites, "Multiagent reinforcement learning in the iterated prisoner's dilemma," *Biosystems*, vol. 37, no. 1-2, pp. 147–166, 1996.

[14] T. Torii, S. Hidaka, and A. Masumi, "Emergence of cooperation in the iterated prisoner ' s dilemma between reinforcement learners," in *Proceedings of The Twenty Eighth Annual Conference of the Japanese Society for Articial Intelligence*, 2014, 4H1-3.

[15] E. Seneta, *Non-negative Matrices and Markov Chains*. Springer, 2006.

[16] J. R. Magnus and H. Neudecker, *Matrix differential calculus*. New York: Cambridge Univ Press, 1988.