Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

Dynamical Singularities in Online Learning of Recurrent Neural Networks

Asaki Saito
Future University - Hakodate
Kameda Nakano-cho, Hakodate, Hokkaido 041-8655, Japan
Email: saito@fun.ac.jp

Makoto Taiji
Genomic Sciences Center, RIKEN
Ono-cho, Tsurumi, Yokohama, Kanagawa 230-0046, Japan
Email: taiji@gsc.riken.jp

Takashi Ikegami
Graduate School of Arts and Sciences, University of Tokyo
Komaba, Meguro, Tokyo 153-8902, Japan
Email: ikeg@sacral.c.u-tokyo.ac.jp

Abstract— **We numerically and theoretically demonstrate various singularities, as a dynamical system, of a simple online learning system of a recurrent neural network (RNN) where RNN performs the one-step prediction of a time series generated by a one-dimensional map. More specifically, we show first through numerical simulations that the learning system exhibits singular behaviors ("neutral behaviors") different from ordinary chaos, such as almost zero finite-time Lyapunov exponents, as well as inaccessibility and power-law decay of the distribution of learning times (transient times). Also, we show through linear stability analysis that, as a dynamical system, the learning system is represented by a singular map whose Jacobian matrix has eigenvalue unity in the whole phase space. In particular, we state that the singularity as a dynamical system (shown by the second method) provides a basic reason for the neutral behaviors (shown by the first method) exhibited by the learning system.**

I. Introduction

While learning as well as the learning process has been well theorized on the basis of statistics or statistical mechanics, some studies have also reported that dynamic and complex phenomena are often observed when we actually examine individual learning processes. Indeed, such phenomena not only in biological systems such as the brain but in even much simpler, artificially constructed systems have been numerically revealed by, e.g., Refs. [1], [2], [3]. To understand the highly dynamic phenomena of biological systems such as the brain, it is necessary to analyze the complex dynamics of the learning process itself. The study of nonlinear dynamical systems will provide a promising basis for this purpose. Therefore, in this paper we investigate, not merely numerically but also theoretically, the dynamical characteristics of the online learning process of a simple recurrent neural network.

A recurrent neural network (RNN) is one of the standard artificial neural network architectures, having feedback connections [4]. The presence of the feedback connections makes the RNN a dynamical system with external inputs. Because of this feature, the RNN is more suitable for the present study regarding dynamics than is the other standard architecture, the feedforward network [4]. If we suppose that the RNN is trained using a deterministic online learning algorithm, the resulting total system is also a dynamical system with external inputs, although this feature also makes analyses of the online learning of a RNN extraordinary difficult. In this study, therefore, we focus on one of the simplest online learning systems obtained from a RNN, and clarify the singularities of its learning process as a dynamical system process by using numerical simulations and linear stability analysis.

The outline of the present paper is as follows: Section II provides preliminary materials. We briefly explain RNN and its online learning algorithm which we use in this study. In Sec. III we introduce a simple learning system where the RNN performs the one-step prediction of a time series generated by a one-dimensional map. This learning system itself is represented by a map, and is the exact object of this study. In Sec. IV, using numerical simulations, we show that singular behaviors ("neutral behaviors") different from ordinary chaos are exhibited by the specific learning system where the RNN learns a periodic time series of the logistic map. In Sec. V, using linear stability analysis, we show that, as a dynamical system, the learning system introduced in Sec. III is generally represented by a singular map whose Jacobian matrix has eigenvalue unity in the whole phase space. This fact provides a basic reason for the neutral behaviors observed in Sec. IV. Section VI provides a summary and discussion.

II. Online Learning of RNN

For the RNN, we choose a second-order recurrent neural network [4]. The network is described by the following equation:

$$\begin{aligned} y_i(t+1) &= f(\sum_{j=1}^{m}\sum_{k=1}^{n} w_{ijk}u_j(t)y_k(t) + \sum_{j=1}^{m} w_{ij}u_j(t) \\ &\quad + \sum_{j=1}^{n} w'_{ij}y_j(t) + w_i), \end{aligned} \tag{1}$$

where the activation function is denoted by $f(\cdot)$. Also, the state of the ith unit at discrete times $t = 0, 1, 2, \cdots$ is denoted by $y_i(t)$ $(i = 1, \cdots, n)$, the jth external input at time t by $u_j(t)$ $(j = 1, \cdots, m)$, and the weights by $w_{ijk}, w_{ij}, w'_{ij}, w_i$. ($w_{ijk}$

is the connection weight to the ith unit from the jth input and the kth unit. w_{ij} is the weight from the jth input, whereas w'_{ij} is that from the jth unit. w_i is the bias. The weights w_{ijk}, w_{ij}, w'_{ij}, and w_i are represented as w_* for convenience.) Also, certain of the units are assumed to be visible (i.e., output units) but the others are hidden.

In the case of RNN, learning is the process to make the output trajectory follow a given desired trajectory by improving the weights, with a given input sequence and initial condition. As an online learning algorithm, we choose the real-time recurrent learning (RTRL) algorithm [5], [6]. This algorithm is based on the gradient decent of instantaneous output error $E(t+1) = \frac{1}{2}\sum_{i=1}^{n} \mu_i \left[y_i(t+1) - d_i(t+1)\right]^2$, and its update rule is

$$w_*(t+1) = w_*(t) - \varepsilon \sum_{i=1}^{n} \mu_i \left[y_i(t+1) - d_i(t+1)\right] v_*^i(t+1), \tag{2}$$

where $\varepsilon > 0$ denotes a learning rate parameter, $d_i(t+1)$ denotes a desired response for $y_i(t+1)$, and $v_*^i(t+1)$ denotes $\frac{\partial y_i(t+1)}{\partial w_*}|_{w_*=w_*(t)}$. Output units are specified by $\mu_i = 1$; otherwise $\mu_i = 0$.

By assuming that the weights are constants, the approximate equation for $v_*^i(t)$ is derived from differentiating Eq. (1) by w_*, yielding

$$\begin{aligned} v_*^i(t+1) &= f'(s_i(t)) \left[\sum_{j=1}^{m}\sum_{k=1}^{n} w_{ijk} u_j(t) v_*^k(t) \right. \\ &\qquad \left. + \sum_{j=1}^{n} w'_{ij} v_*^j(t) + \gamma \right], \qquad (3) \\ \gamma &= \begin{cases} \delta_{ia} u_b(t) y_c(t) & \text{if } w_* \equiv w_{abc} \\ \delta_{ia} u_b(t) & \text{if } w_* \equiv w_{ab} \\ \delta_{ia} y_b(t) & \text{if } w_* \equiv w'_{ab} \\ \delta_{ia} & \text{if } w_* \equiv w_a \end{cases} \end{aligned}$$

where $s_i(t)$ denotes the net input to the ith unit at time t, and δ_{ia} denotes the Kronecker delta. We note here that Eq. (3) is a dynamical system with external inputs, with dynamical variables $\{v_*^i\}$. Since the initial state $y_i(0)$ of the network has no dependence on the weights, the initial condition for Eq. (3) is $v_*^i(0) = \frac{\partial y_i(0)}{\partial w_*} = 0$. Thus, the RNN trained using the RTRL algorithm is confirmed to be a dynamical system with external inputs by Eqs. (1)-(3), where the dynamical variables are $\{y_i, w_*, v_*^i\}$ and the external inputs are $\{u_i, d_i\}$.

III. Learning System

A closed dynamical system can be directly constructed by generating the external inputs $\{u_i, d_i\}$ from another dynamical system or another dynamical system with inputs. In this paper, we study such a learning system, in particular the case in which the RNN performs a one-step prediction of a time series generated by a one-dimensional map. Let g and x denote the one-dimensional map and its dynamical variable, respectively. Then, the time series generated by g is given by

$$x(t+1) = g(x(t)) \qquad t = 0, 1, 2, \cdots, \tag{4}$$

and thus the external inputs become $u(t) = x(t)$ and $d(t+1) = x(t+1)$ for our prediction task. If the number of units is only one ($n = 1$), the dynamical system obtained from Eqs. (1)-(4) is given by

$$\begin{aligned} y_1(t+1) &= f(w_{111}(t)x(t)y_1(t) + w_{11}(t)x(t) \\ &\qquad + w'_{11}(t)y_1(t) + w_1(t)) \\ w_{111}(t+1) &= w_{111}(t) - \varepsilon\left[y_1(t+1) - x(t+1)\right] v^1_{111}(t+1) \\ w_{11}(t+1) &= w_{11}(t) - \varepsilon\left[y_1(t+1) - x(t+1)\right] v^1_{11}(t+1) \\ w'_{11}(t+1) &= w'_{11}(t) - \varepsilon\left[y_1(t+1) - x(t+1)\right] v^1_{11}{}'(t+1) \\ w_1(t+1) &= w_1(t) - \varepsilon\left[y_1(t+1) - x(t+1)\right] v^1_1(t+1) \\ v^1_{111}(t+1) &= f'(s_1(t)) \left[w_{111}(t)x(t)v^1_{111}(t) \right. \\ &\qquad \left. + w'_{11}(t)v^1_{111}(t) + x(t)y_1(t)\right] \\ v^1_{11}(t+1) &= f'(s_1(t)) \left[w_{111}(t)x(t)v^1_{11}(t) \right. \\ &\qquad \left. + w'_{11}(t)v^1_{11}(t) + x(t)\right] \\ v^1_{11}{}'(t+1) &= f'(s_1(t)) \left[w_{111}(t)x(t)v^1_{11}{}'(t) \right. \\ &\qquad \left. + w'_{11}(t)v^1_{11}{}'(t) + y_1(t)\right] \\ v^1_1(t+1) &= f'(s_1(t)) \left[w_{111}(t)x(t)v^1_1(t) \right. \\ &\qquad \left. + w'_{11}(t)v^1_1(t) + 1\right] \\ x(t+1) &= g(x(t)). \end{aligned} \tag{5}$$

We treat this 10-dimensional learning system throughout the following study.

IV. Numerical Simulations

In this section, we show various singular behaviors (neutral behaviors) of the learning system [Eq. (5)] by numerical simulations. As a typical example, we present results obtained by the learning system in which RNN learns a periodic time series generated by the logistic map. The logistic map is a one-dimensional map, given by

$$x(t+1) = ax(t)\left[1 - x(t)\right] \qquad t = 0, 1, 2, \cdots,$$

where a is the only parameter [7]. The dynamics of this map has been thoroughly clarified, and we can easily generate not only a chaotic time series but also a periodic time series by setting the parameter a appropriately.

A. Orbital Instability

First, we explore the orbital instability of the learning system using the finite-time Lyapunov exponent. The time-T Lyapunov exponent is the average exponential expansion (or contraction) rate along the trajectory of length T. Actually, the number of Lyapunov exponents is equal to the dimension of the phase space, but we concentrate on the largest one. As shown below, we find two typical classes of dynamical behaviors.

We now show the results obtained from the learning system [Eq. (5)] with the learning rate $\varepsilon = 0.1$ and the activation function $f(x) = 1/(1+e^{-x})$ (these are used in the numerical simulations shown hereafter unless otherwise noted). Figure 1(a) shows $y_1(t)$ and $x(t)$ versus t for $a = 3.3$, where $y_1(0)$, $x(0)$, and $w_*(0)$ [i.e., $w_{111}(0)$, $w_{11}(0)$, $w'_{11}(0)$, and $w_1(0)$] are randomly chosen with uniform distributions in [0, 1] and [-5, 5], respectively. The logistic map at $a = 3.3$ has a stable period-two orbit, and thus the task for RNN is to fit the output $y_1(t)$ to the period-two orbit. In the present example,

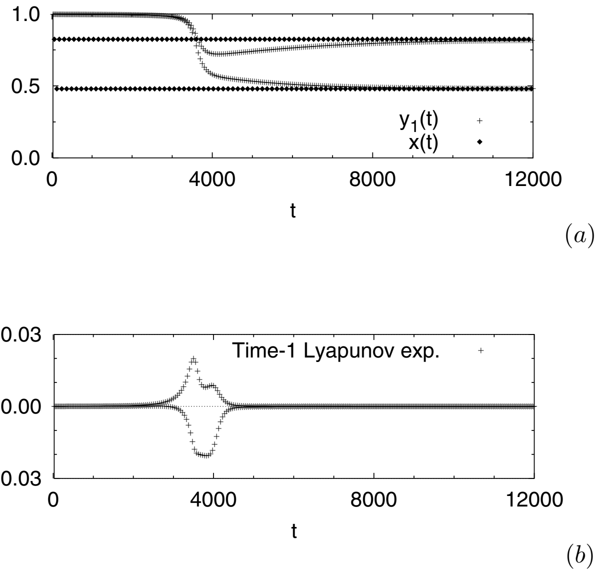


Fig. 1. Time evolution for the period-two learning ($a = 3.3$). (a) $y_1(t)$ and $x(t)$ versus t. (b) The time-1 Lyapunov exponent versus t.

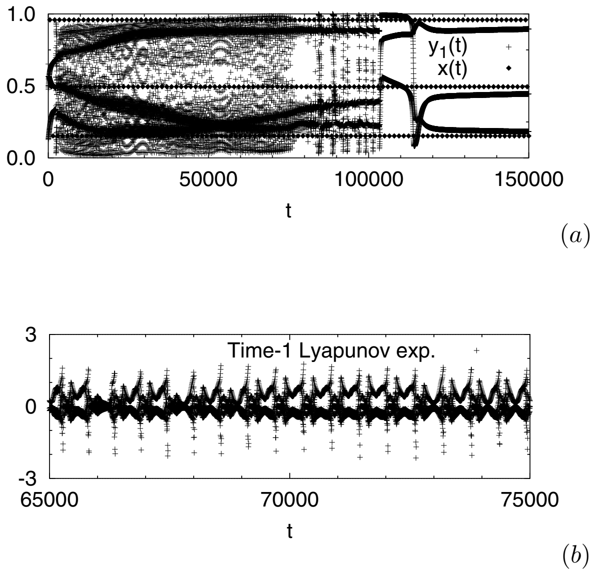


Fig. 2. Time evolution for the period-three learning ($a = 3.835$). (a) $y_1(t)$ and $x(t)$ versus t. (b) The time-1 Lyapunov exponent versus $t \in [65000, 75000]$.

the learning results in success with $y_1(t)$'s smooth approach to $x(t)$. Figure $1(b)$ shows the time-1 Lyapunov exponent versus t, on the same condition as in Fig. $1(a)$. The finite-time Lyapunov exponent is almost 0, but it oscillates around $t = 3500$, where learning progresses substantially. [1] This kind of smooth dynamical behavior is widely observed for other chosen conditions, and forms one of the typical dynamical behaviors of learning systems.

On the other hand, the other typical dynamical behavior is presented in Fig. 2. Figure $2(a)$ shows $y_1(t)$ and $x(t)$ versus t for $a = 3.835$, where the logistic map has a stable period-three orbit. Initial conditions are randomly chosen as before, and the learning also results in success in this example. This example, however, shows complex transient with intermittent behavior; the transient appears to be chaotic, but in many short time intervals it appears to be almost periodic as well. Figure $2(b)$ shows the time-1 Lyapunov exponent versus t, under the same condition as in Fig. $2(a)$. The finite-time Lyapunov exponent oscillates irregularly around 0. In contrast to simple hyperbolic or near-hyperbolic chaos, this result indicates the strong non-hyperbolicity of the dynamical system because of a successively varying number of stable and unstable dimensions under the dynamics. This type of complex dynamical behavior is also widely observed in learning systems.

In either case, the above results show a singularity of the learning system as a dynamical system. Indeed, finite-time Lyapunov exponents ordinarily keep taking either positive or negative values, even if they fluctuate.

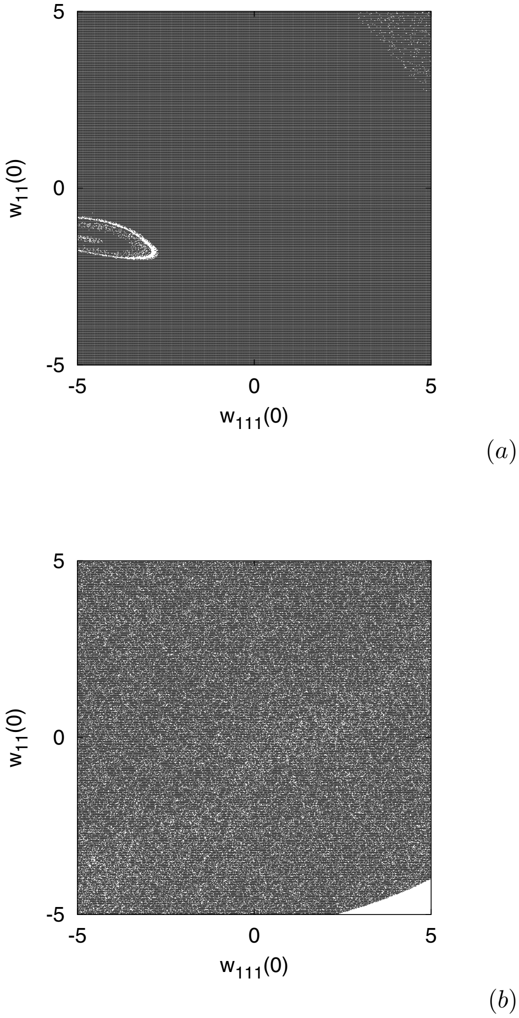
B. Basin Structure and Inaccessibility

In the above two examples, we saw cases in which leaning results in success, but there are also cases in which learning ends in failure, depending on initial conditions. To study the structure of initial conditions (i.e., basin structure), we introduce the following procedure to numerically decide whether a learning process ends in success: If $|y_i(t) - d_i(t)| < \epsilon_{\text{check}}$ for successive n_{check} time steps, then the learning process is regarded as successful, where checking width ϵ_{check} and checking time n_{check} are chosen from sufficiently small positive real numbers and sufficiently large natural numbers, respectively. [2]

Figures $3(a)$ and (b) show a 2-dimensional slice ($w'_{11} = w_1 = -5.0$) through the 4-dimensional initial weight space for the period-two learning ($a = 3.3$) and the period-three learning ($a = 3.835$), respectively. Each initial condition on a 500×500 grid is followed until the time limit of 10^6 time steps, where $w_{111}(0)$ and $w_{11}(0)$ are given by the horizontal and vertical axes, respectively. The other initial values are $y_1(0) = x(0) = 0.3$ and $v^1_{111}(0) = v^1_{11}(0) = {v^1_{11}}'(0) = v^1_1(0) = 0$. Grid points are plotted as black dots for initial conditions from which learning ends in success. Otherwise, points are left blank. As a result, initial weights with success for the period-three learning [Fig. 3(b)] are wholly riddled with white holes, in contrast to those for the period-two learning [Fig. 3(a)]. However, even for the period-two learning [Fig. 3(a)], there is also a region having fine structure where initial weights with success and

[1] The Lyapunov exponent of the logistic map is sufficiently negative at $a = 3.3$, and therefore does not affect the largest finite-time Lyapunov exponent of the total learning system. This remark is also applied to the case of $a = 3.835$ below.

[2] In general, it is impossible to conclude, by the numerical observation of a learning process of finite-length, that the learning ends in failure, because there is a possibility of success if one looks further ahead of that process.



(a)

(b)

Fig. 3. Initial weights with success, plotted as black dots, for (a) the period-two learning, and (b) the period-three learning.

those with failure are complicatedly interwoven. This implies sensitivity to initial conditions.

To rigorously investigate the robustness of the learning process against unavoidable perturbations (noise, measurement errors, etc.), we focus on the basin boundary between two sets of initial weights with different fates (i.e, success or not), and examine the ϵ-dependence of $V(\epsilon)$, i.e., the 4-dimensional volume of the ϵ-neighborhood of the boundary in the 4-dimensional initial weight space. This $V(\epsilon)$ is proportional to the probability of making a mistake in the determination of final fate, if we were to pick an initial weight at random in a bounded region containing the boundary, and if our ability to determine the position of the initial weight had an uncertainty ϵ. Figure 4 shows results of numerical experiments for the period-two and period-three learning, where $V(\epsilon)$ is plotted versus ϵ with a logarithmic scale. In each case, we evaluate $V(\epsilon)$ of the region $-5 \leq w_*(0) \leq 5$. The other settings are the same as the previous basin's case. As a result, $V(\epsilon)$ does not depend on ϵ, regardless of the period-two or period-three learning. This indicates another dynamical singularity of the learning system, i.e., the inaccessibility of the ideal learning process [8], which we now move on to explain.

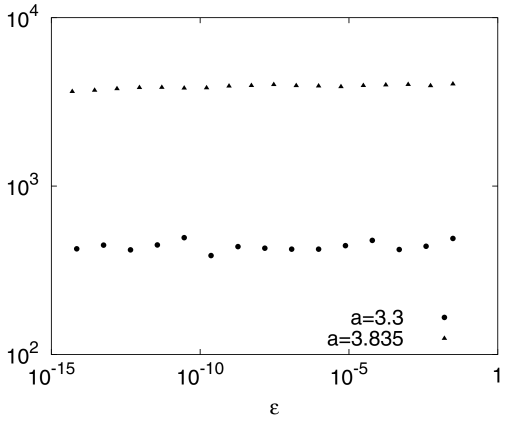


Fig. 4. $V(\epsilon)$ versus ϵ for the period-two learning ($a = 3.3$) and the period-three learning ($a = 3.835$) (log-log plot).

The learning process is determined by dynamical equations. Thus, even if perturbations of amplitude ϵ are added, it is naturally expected that $V(\epsilon)$ can be decreased by decreasing ϵ. In other words, the more one improves accuracy, the better one can follow the true learning process under an ideal condition without perturbations. Indeed, for a fractal boundary in general, $V(\epsilon)$ scales with ϵ as $V(\epsilon) \sim \epsilon^{\phi}$ with $0 < \phi < 1$ [7], and thus $V(\epsilon)$ can be decreased to 0 with a power law. On the other hand, in this special case where $V(\epsilon)$ does not depend on ϵ (i.e., $\phi = 0$), the ideal learning process cannot be approached by decreasing ϵ, as long as there exist perturbations regardless of how small they are. In this sense, the above result shows the inaccessibility of the ideal learning process.

The following should be noted: First, this notion of uncertainty — inaccessibility — is qualitatively different from chaotic unpredictability, which will disappear as accuracy is improved. Second, the above ϕ is called an uncertainty exponent [7], and the box-counting dimension of the boundary, D_0, is given by $D_0 = N - \phi$ where N is the space dimensionality. As mentioned previously, $\phi > 0$ for ordinary fractal sets, e.g., those constructed by transient chaos. On the other hand, fractal sets having $\phi = 0$ are so extraordinary that this class contains the Mandelbrot set and geometric representation of the halting set of a universal Turing machine [9].

The inaccessibility of ideal learning process is widely observed in learning systems, e.g., for other choices of learning rate, periodic time series, the initial value of the network state, the initial value of the logistic map, etc. From the viewpoint of dynamical systems, an extraordinary basin boundary having $\phi = 0$ is based on singular dynamics such as that exemplified

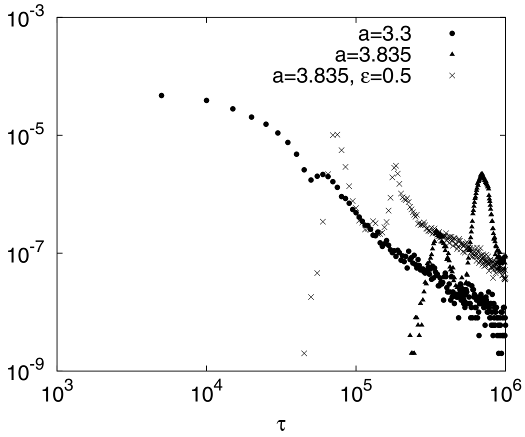


Fig. 5. Distributions of learning times for the period-two learning ($a = 3.3$), the period-three learning ($a = 3.835$), and the period-three learning with a larger learning rate ($a = 3.835$, $\varepsilon = 0.5$) (log-log plot).

in Fig. 2, where the finite-time Lyapunov exponent irregularly fluctuates around 0.

C. Distribution of Learning Times

Now we turn our focus to the construction process of such fractal basins. In particular, we study the distribution of learning times (transient times). Figure 5 shows learning time distribution of the period-two and period-three learning with $a = 3.3$ and $a = 3.835$, respectively. (We later refer to the case with $a = 3.835$ and $\varepsilon = 0.5$, also shown there.) In each case, initial weights are uniformly chosen from the region $-5 \leq w_*(0) \leq 5$. The other settings are the same as before. As a result, in the case of the period-two learning, a fraction of the initial points from which learning ends in success with learning time τ, denoted as $p(\tau)$, is found to decay according to a power law. This "slow" decay is in strong contrast with "fast" decay observed in transient chaos. In general, the distribution of transient times decays exponentially for transient chaos (i.e., the construction process of ordinary fractals) [7]. In the case of the period-three learning, however, we cannot observe any particular decay tendency until the time limit (10^6 time steps) imposed by current computational cost. Nevertheless, we expect power-law decay also in this case, if one looks further ahead over the time limit. Indeed, another period-three learning, denoted by $a = 3.835$ and $\varepsilon = 0.5$ in Fig. 5, shows power-law decay, where the only difference from the period-three learning studied so far is a larger learning rate ($\varepsilon = 0.5$), for the purpose of speed-up.

This singular behavior, the power-law decay of transient time distributions, is typically observed for other learning systems with variously different conditions. From the viewpoint of dynamical systems, singular dynamics such as those having finite-time Lyapunov exponents around 0 underlie the slow decay of these transient time distributions.

V. Linear Stability Analysis

So far we have seen the global properties of the learning system where g in Eq. (5) is the logistic map. In this section, we demonstrate a local property of the general learning system represented by Eq. (5), by using linear stability analysis. We can prove the following fact for the Jacobian matrix of Eq. (5). We give the fact without a proof, but the details are mechanical.

Fact 1: The Jacobian matrix of the learning system [Eq. (5)] has eigenvalue unity at all points in phase space, except for points where the Jacobian matrix is not defined.

As stated in the previous section, we have observed, in various learning systems, singular behaviors (neutral behaviors) different from ordinary chaos, such as that exemplified by zero finite-time Lyapunov exponents. One of the basic reasons for these singularities is that the learning system is represented by such a singular map whose Jacobian matrix has the eigenvalue unity in the whole phase space. Furthermore, the local characteristic stated in Fact 1 can be proven to hold true also for other learning systems with different neuron models [e.g., a first-order model such as $y_1(t+1) = f(w_{11}x(t) + w'_{11}y_1(t) + w_1)$] and different update schemes [e.g., updating $w_*(t)$ based on the gradient of $E(t) = \frac{1}{2}\left[y_1(t) - x(t)\right]^2$, instead of $E(t+1)$]. [3]

VI. Summary and Discussion

In this paper, we have both numerically and theoretically demonstrated various dynamical singularities of a simple learning system in which a RNN learns a time series generated by a one-dimensional map with the RTRL algorithm. In particular, we have numerically shown "neutral behaviors" such as almost zero finite-time Lyapunov exponents, inaccessibility, and power-law decay of the distribution of learning times (transient times), in contrast with ordinary chaos. These characteristics have been widely observed in other learning systems having different network structures (e.g., the number of units $n > 1$), different learning algorithms (e.g., extended Kalman filtering algorithm [4]), and different tasks (e.g., the case in which RNNs learn each other). Furthermore, we have shown that a basic reason for these singularities is that the learning system is represented by a singular map whose Jacobian matrix is proven to have eigenvalue unity in the whole phase space. This characteristic also holds true for certain other learning systems.

On the basis of the above results, we conjecture that unique dynamical singularities such as those reported here will be common in systems capable of adaptation or learning. Indeed, a study similar to the present one can be made for adaptive delayed-feedback control [10]. Further clarification of such universality will be needed to understand the highly dynamic and complex phenomena observed in biological systems such as the brain.

[3] We believe that the same local characteristic will hold true for networks with more than one neuron of the above types, although proving this will require future study.

Acknowledgments

This research is partly supported by a Grant-in-Aid for Young Scientists (B) 18700226 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] J. F. Kolen and J. B. Pollack, "Back propagation is sensitive to initial conditions," *Complex Systems*, vol. 4, pp. 269–280, 1990.

[2] T. Hondou and Y. Sawada, "Analysis of learning processes of chaotic time series by neural networks," *Progress of Theoretical Physics*, vol. 91, no. 2, pp. 397–402, 1994.

[3] H. Nakajima and Y. Ueda, "Riddled basins of the optimal states in learning dynamical systems," *Physica D*, vol. 99, pp. 35–44, December 1996.

[4] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 1999.

[5] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent networks," *Neural Computation*, vol. 1, pp. 270–280, 1989.

[6] K. Doya, "Recurrent networks: Supervised learning," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 1995.

[7] E. Ott, *Chaos in Dynamical Systems*. Cambridge: Cambridge University Press, 1993.

[8] A. Saito, M. Taiji, and T. Ikegami, "Inaccessibility in online learning of recurrent neural networks," *Physical Review Letters*, vol. 93, no. 16, p. 168101, Oct 2004.

[9] A. Saito and K. Kaneko, "Inaccessibility and undecidability in computation, geometry, and dynamical systems," *Physica D*, vol. 155, pp. 1–33, July 2001.

[10] A. Saito and K. Konishi, *in preparation*.