Inaccessibility in Online Learning of Recurrent Neural Networks

Asaki Saito,[1] Makoto Taiji,[2,3] and Takashi Ikegami[4]

[1]*Future University-Hakodate, 116-2 Kameda Nakano-cho, Hakodate, Hokkaido 041-8655, Japan*
[2]*Genomic Sciences Center, RIKEN, Ono-cho, Tsurumi, Yokohama, Kanagawa 230-0046, Japan*
[3]*Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato, Tokyo 106-8569, Japan*
[4]*Graduate School of Arts and Sciences, University of Tokyo, Komaba, Meguro, Tokyo 153-8902, Japan*

(Received 18 December 2003; revised manuscript received 7 May 2004; published 13 October 2004)

We apply nonlinear dynamical system techniques to recurrent neural networks. In particular, we numerically analyze the dynamical system characteristics of the online learning process. By introducing the notion of inaccessibility, we show that the learning process is well characterized by strong nonhyperbolicity and inaccessibility, which is a greater uncertainty than chaotic unpredictability. These results are clearly contrasted with a gradient descent dynamics, or ordinary chaos.

DOI: 10.1103/PhysRevLett.93.168101 PACS numbers: 87.18.Sn, 05.45.Pq, 07.05.Mh, 89.75.Hc

Learning and adaptation are essential features of brains, living organisms, social systems, and so forth. Attempts have also been made to equip artificial systems with such a learning ability. Although these theoretical investigations have achieved significant success (see, e.g., Ref. [1]), most attention has been paid to rather static learning aspects such as efficiency, convergence to the optimal solution, etc. To understand the highly dynamic phenomena of the realistic learning systems as brains, it is necessary to analyze the complex dynamics of the learning process itself, including the cases such that the efficiency of that process is low or the process does not eventually converge, as is often the case with real-life learning systems. The methods used in the study of nonlinear dynamical systems can be effectively utilized for this purpose. In this work we study the dynamical characteristics in the simplest class of learning systems. That is, we study the learning process of a recurrent neural network trained with an online learning algorithm.

A recurrent neural network (RNN) is one of the main artificial neural network architectures that has feedback connections [1]. This feature makes the RNN a dynamical system with external inputs, where the dynamical variables are the states of the units. A RNN deterministically transforms an input time series into an output time series. Suppose a RNN is trained using a deterministic online learning algorithm. In online learning, connection weights are sequentially updated, depending on the previous weights and other "information" extracted from given training data (e.g., an error covariance matrix in a Kalman filtering). This stored information is also updated sequentially. These two updates are determined by specifying an online learning algorithm. Thus, by regarding both the weights and the information as new dynamical variables, a RNN trained by an online learning algorithm is also a dynamical system with inputs. In this case, however, teacher signals are also the inputs in addition to the original input signals in the RNN.

To clarify dynamical features of the learning process, we construct a closed dynamical system by taking the inputs from another dynamical system. We then study the dynamics of the total learning system. In this Letter, we examine orbital instability and basin structure (inaccessibility), all in cases where a RNN learns a periodic time series generated by the logistic map. Because such a case is one of the simplest, it is appropriate to start with. Our study employs numerical methods because there are no theoretical methods established for analyzing the dynamics of the online learning of a RNN. As shown later, even in this simple case, the dynamics is extraordinarily complex and difficult to analyze.

We choose as our RNN a second-order RNN [1]:

$$y_i(t+1) = f\Big(\sum_{j=1}^{m}\sum_{k=1}^{n} w_{ijk}u_j(t)y_k(t) + \sum_{j=1}^{m} w_{ij}u_j(t) + \sum_{j=1}^{n} w'_{ij}y_j(t) + w_i\Big). \quad (1)$$

The nonlinear function f is given by $f(x) = 1/(1+e^{-x})$. The state of the ith unit at time t is denoted by $y_i(t)$ ($i = 1, \ldots, n$), the jth external input at time t by $u_j(t)$ ($j = 1, \ldots, m$), and the weights by w_{ijk}, w_{ij}, w'_{ij}, w_i. (w_{ijk} is the 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 output units.

In the case of RNNs, learning is the process of making the output trajectory follow a given desired trajectory by improving the weights. For an online learning algorithm, we choose that of real-time recurrent learning (RTRL) [2]. RTRL is based on the gradient descent of current output error. The update rule of RTRL is

$$w_*(t) = w_*(t-1) - \varepsilon\sum_{i=1}^{n} \mu_i[y_i(t) - d_i(t)]v^i_*(t), \quad (2)$$

where $\varepsilon > 0$ denotes a learning rate parameter, $d_i(t)$ denotes a desired response for $y_i(t)$, and $v^i_*(t)$ denotes $\frac{\partial y_i(t)}{\partial w_*}|_{w_*=w_*(t-1)}$. Output units are specified by $\mu_i = 1$; otherwise $\mu_i = 0$. By assuming that the weights are constant in time, the approximate equation for $v^i_*(t)$ is derived from differentiating Eq. (1) by w_*, yielding

$$v^i_*(t+1) = f'(s_i(t))\Big[\sum_{j=1}^{m}\sum_{k=1}^{n} w_{ijk}u_j(t)v^k_*(t) + \sum_{j=1}^{n} w'_{ij}v^j_*(t) + \gamma\Big], \tag{3}$$

$$\gamma = \begin{cases} \delta_{ia}u_b(t)y_c(t) & \text{if } w_* \equiv w_{abc} \\ \delta_{ia}u_b(t) & \text{if } w_* \equiv w_{ab} \\ \delta_{ia}y_b(t) & \text{if } w_* \equiv w'_{ab} \\ \delta_{ia} & \text{if } w_* \equiv w_a \end{cases}$$

where $s_i(t)$ is the net input to the ith unit at time t, and δ_{ia} is the Kronecker delta. As can be seen from Eqs. (1)–(3), a RNN trained by the RTRL is a dynamical system with inputs, where the dynamical variables are $\{y_i(t), w_*(t), v^i_*(t)\}$ and the inputs are $\{u_i(t), d_i(t)\}$ [3].

It is straightforward to construct a closed dynamical system by generating the above inputs $\{u_i(t), d_i(t)\}$ from another dynamical system. In this work, we use the well-known logistic map [4], given by

$$x(t+1) = ax(t)[1 - x(t)] \qquad t = 0, 1, 2, \ldots, \tag{4}$$

where a is the only parameter. In our study, the RNN performs a one-step prediction of a periodic time series generated by the logistic map [i.e., $u(t) = x(t)$ and $d(t) = x(t+1)$]. If there is only one unit ($n = 1$), the obtained closed dynamical system given by Eqs. (1)–(4) is ten dimensional, where the dynamical variables are $\{x, y_1, w_{111}, w_{11}, w'_{11}, w_1, v^1_{111}, v^1_{11}, v'^1_{11}, v^1_1\}$. We use this ten-dimensional system throughout the following studies.

First we explore the orbital instability of the learning system, by using the finite-time Lyapunov exponent [5]. The time-t Lyapunov exponent is the average exponential expansion rate along the trajectory of length t. In this study, we focus on the largest exponent. As shown below, we find two typical classes of dynamical behaviors.

In the following, all the numerical results are obtained from the ten-dimensional learning system with the learning rate $\varepsilon = 0.1$. Figure 1(a) shows $y_1(t)$ and $x(t)$ versus t for $a = 0.5$, where $y_1(0)$ and $x(0)$ are randomly chosen from $[0, 1]$, whereas $w_{111}(0)$, $w_{11}(0)$, $w'_{11}(0)$, and $w_1(0)$ are from $[-5, 5]$. The logistic map at $a = 0.5$ has a stable fixed point at $x = 0$, and thus the task for a RNN is to fit the output $y_1(t)$ to 0. In this example, the learning results in success with $y_1(t)$'s smooth approach to 0. Figure 1(b) shows the time-1 Lyapunov exponent versus t, on the same condition as in Fig. 1(a). (We later explain the estimate also shown there.) The finite-time Lyapunov exponent is almost 0, but it oscillates smoothly around $t = 700$, where learning progresses substantially [6]. This kind of smooth dynamical behavior is widely observed for other choices of conditions, and forms one of the typical dynamical behaviors in the 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 results in success in this example as well. However, this example shows complex transient with intermittent behaviors with respect to time. The transient as a whole appears to be chaotic. However, many short time intervals are embedded in the transient as well, where the trajectory appears to be almost periodic. In this case, the finite-time Lyapunov exponent oscillates around 0 irregularly with intermittent bursts [Fig. 2(b)]. This result indicates strong nonhyperbolicity of the dynamical system, because the number of stable and unstable dimensions is successively varied under the operation of the dynamics.

So far, we have evaluated the finite-time Lyapunov exponent of the whole ten-dimensional system. The smooth behavior as shown in Fig. 1, however, can be well explained by assuming that the learning system

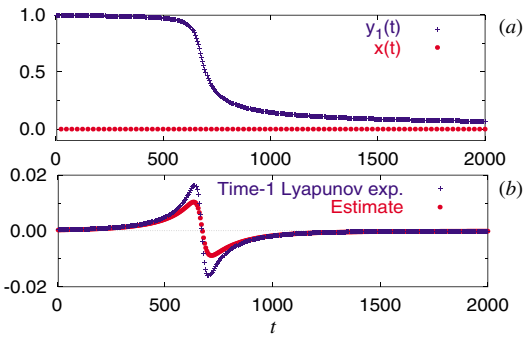

FIG. 1 (color online). Time evolution for the fixed point learning ($a = 0.5$). (a) $y_1(t)$ and $x(t)$ versus t. (b) The time-1 Lyapunov exponent and its estimate versus t.

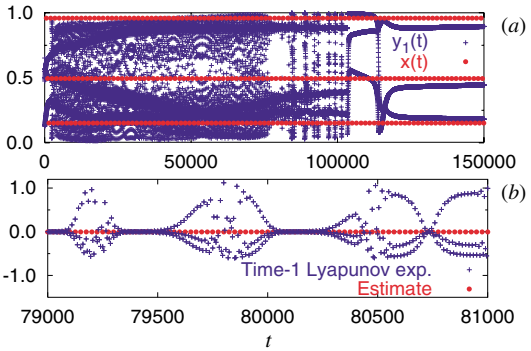

FIG. 2 (color online). Time evolution for the period three learning ($a = 3.835$). (a) $y_1(t)$ and $x(t)$ versus t. (b) The time-1 Lyapunov exponent and its estimate versus $t \in (79\,000, 81\,000)$.

obeys a simple gradient descent dynamics only within weight space (four-dimensional space, in this case, by neglecting the other variables). Here, by the term "gradient descent dynamics," we consider the following:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \varepsilon \nabla E_t|_{\boldsymbol{w}=\boldsymbol{w}(t)} \tag{5}$$

where E_t is a cost function of variables $\boldsymbol{w}$, on which gradient descent is performed with respect to $\boldsymbol{w}$ at time t. If we consider two nearby orbits, $\boldsymbol{w}(t)$ and $\boldsymbol{w}'(t)$, then the evolution of the displacement $\Delta\boldsymbol{w}(t) = \boldsymbol{w}'(t) - \boldsymbol{w}(t)$ is approximated by $\Delta\boldsymbol{w}(t+1) = (\mathbf{I} - \varepsilon \mathbf{H}_t|_{\boldsymbol{w}=\boldsymbol{w}(t)})\Delta\boldsymbol{w}(t)$ where $\mathbf{I}$ is the identity matrix and $\mathbf{H}_t$ is the Hessian matrix of E_t. Thus, the largest time-1 Lyapunov exponent of (5) is given by $\log(1 - \varepsilon\lambda)$, where λ is the smallest eigenvalue of $\mathbf{H}_t|_{\boldsymbol{w}=\boldsymbol{w}(t)}$.

As for the original problem of the ten-dimensional dynamical system, the cost function at time t is $\frac{1}{2}[y_1(t+1) - x(t+1)]^2$, on which gradient descent is performed with respect to the weights (i.e., w_{111}, w_{11}, w'_{11}, and w_1). It is easy to evaluate the smallest eigenvalue $\tilde{\lambda}$ of the Hessian matrix of this cost function. Thus, if the dynamics of the ten-dimensional map is effectively a gradient descent dynamics only within four-dimensional weight space, then the actual finite-time Lyapunov exponent will well coincide with its estimate $\log(1 - \varepsilon\tilde{\lambda})$.

Figures 1(b) and 2(b) also show the estimates of the finite-time Lyapunov exponents based on the above assumption. In Fig. 1(b), the actual finite-time Lyapunov exponent well coincides with its estimate. Thus, this learning system can be regarded as simply performing a gradient descent within the weight space. We expect that the landscape of the cost function may have one local minimum, and that its downward slope may consist of an interval with negative $\tilde{\lambda}$ (second derivative) first, followed by an interval with a positive one near the bottom, that corresponds to the positive and negative finite-time Lyapunov exponents, respectively. In this case, even though stretching is achieved while the exponent is positive, folding and mixing do not occur. On the other hand, in Fig. 2(b), the actual exponent behavior is qualitatively different from its estimate. Instead of showing strong nonhyperbolicity, the estimate is almost always 0 or less (we confirmed that the estimate is slightly less than 0). Thus, this system does not perform a simple gradient descent. Indeed, this strong nonhyperbolicity is attributable to the complex motion of $\boldsymbol{v}^i_*(t)$.

In the two examples above, learning results in success, but these two types of success (gradient descent and no gradient descent) should be clearly distinguished. The qualitative difference in the ways to success can be well identified by measuring a deviation of a finite-time Lyapunov exponent from its estimate. From the viewpoint of dynamical systems, the latter dynamical behavior of the learning systems is remarkable, exhibiting strong nonhyperbolicity in contrast with simple hyperbolic chaos or near-hyperbolic chaos [7].

Generally, there are many cases where learning ends in failure, depending on an initial condition. In the following, we study the structure of initial conditions, i.e., basin structure of the above dynamics, and show how inaccessibility is present in the learning process.

Figure 3(a) shows a two-dimensional slice ($w_{111} = w_1 = -5.0$) through the four-dimensional initial weight space for the period three learning ($a = 3.835$) where $y_1(0) = x(0) = 0.3$. Each initial condition on a 500×500 grid is followed until a certain time limit T (10^6 time steps), where $w_{11}(0)$ and $w'_{11}(0)$ are given by the horizontal and vertical axes, respectively. Grid points are plotted as black dots, for initial conditions from which learning ends in success. Points are left blank for initial conditions that do not end in success until T. It shows fine structure on which initial weights resulting in success and those in failure are complicatedly interwoven. This implies sensitivity to initial conditions [10].

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 we examine the ϵ dependence of $V(\epsilon)$, the four-dimensional volume of the ϵ neighborhood of the boundary. This $V(\epsilon)$ is proportional to the probability of error in determining the final fate, if both of the following conditions are satisfied: We pick an initial weight at random in a bounded region containing the boundary, and our ability to determine the position of the initial weight has an uncertainty ϵ. Figure 3(b) shows the numerical results for $a = 3.835$ and $y_1(0) = x(0) = 0.3$, where $V(\epsilon)$ is plotted with ϵ with a logarithmic scale [we evaluate $V(\epsilon)$ of the region $-5 \le w_*(0) \le 5$]. The results show that $V(\epsilon)$ does not depend on ϵ.

The learning process is determined by dynamical equations. Thus, in normal cases, $V(\epsilon)$ can be decreased by decreasing ϵ, even if perturbations of amplitude ϵ are

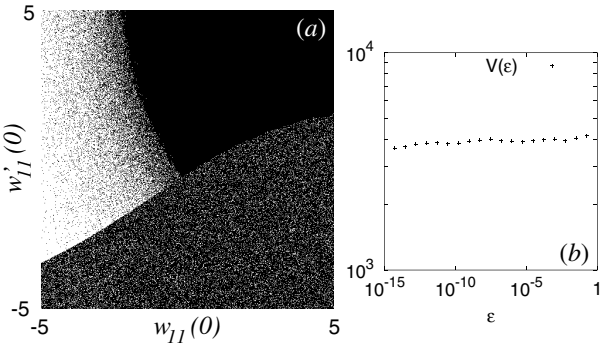

FIG. 3. (a) Initial weights with success (plotted as black dots), and (b) $V(\epsilon)$ versus ϵ, for the period three learning ($a = 3.835$, $T = 10^6$).

added. In other words, the more one improves accuracy, the better one can follow an ideal learning process (one without perturbations). Indeed, for a fractal boundary in general, $V(\epsilon)$ scales with ϵ as $V(\epsilon) \sim \epsilon^{\phi}$ with $0 < \phi < 1$ [4], 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., $\phi = 0$), an ideal learning process cannot be approached by decreasing ϵ, as long as perturbations exist (no matter how small). We call this situation inaccessibility of an ideal learning process.

The following should be noted. First, this newly introduced notion of uncertainty—inaccessibility—is qualitatively different from chaotic unpredictability, which disappears as accuracy improves. Second, the above ϕ is called the uncertainty exponent [4], and the box-counting dimension of the boundary is given by $N - \phi$, where N is the space dimensionality. As mentioned already, $\phi > 0$ for ordinary fractals, e.g., those constructed by transient chaos. On the other hand, several known fractals with $\phi = 0$ are so extraordinary that this class contains the Mandelbrot set and geometric representation of the halting set of a universal Turing machine [13].

Inaccessibility of an ideal learning process is widely observed in many learning systems, for example, with other choices of learning rate, periodic time series (e.g., period two), the initial value of the network state, or the initial value of the logistic map. However, there are also cases where $\phi > 0$, i.e., an ideal learning process is accessible [e.g., for $a = 0.5$ and $y_1(0) = x(0) = 0.3$]. From the viewpoint of dynamical systems, the extraordinary basin boundary with $\phi = 0$ is based on dynamics such as that exemplified in Fig. 2, where the finite-time Lyapunov exponent fluctuates around 0.

In this Letter, we have shown the dynamical system characteristics of learning systems where a RNN learns a periodic time series generated by the logistic map with the RTRL algorithm. Especially, we have shown cases exhibiting strong nonhyperbolicity and inaccessibility, in contrast with cases of gradient descent dynamics or with cases of ordinary chaotic dynamics. In the cases of strong nonhyperbolicity and inaccessibility, furthermore, we have found a power law decay of the distribution of learning times (transient times), although we do not show it in this Letter. As far as we know, inaccessibility is always accompanied by this power law decay, together with strong nonhyperbolicity [13]. Here we should emphasize again that it is very difficult to analyze the dynamical properties, especially inaccessibility (basin structure), of high-dimensional learning systems. Thus, to clarify the dynamical features of learning systems, we have to start with learning systems that are as low-dimensional as possible, such as our simple model. Because dynamics turns out extraordinarily singular even in low-dimensional learning systems, we expect it is naturally so in high-dimensional systems. Indeed, the characteristics reported here have been widely observed in other learning systems with different network structures (e.g., the number of units $n > 1$), different learning algorithms (e.g., extended Kalman filtering), and different tasks (e.g., the case where RNNs learn each other).

Finally, inaccessibility has been shown to give a characterization of undecidability in computation theory [13]. Based on the present results of inaccessibility, we expect that learnability is also intrinsically limited as universal computability. Clarifying potential limitations of learning will not just deepen our understanding of learning. It will also be crucial in the understanding of brains, living organisms, etc., along with clarifying the significance of the type of learning, such as exemplified in Fig. 2, which fails to follow an expected learning scheme but results in success (by somehow utilizing instability, for example).

We thank S. Amari, K. Kaneko, N. Murata, T. Hondou, M. Tatsuno, and I. Frank for their suggestions. This work is partly supported by the Special Postdoctoral Researchers Program of RIKEN.

[1] S. Haykin, *Neural Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1999).

[2] R. J. Williams and D. Zipser, Neural Comput. **1**, 270 (1989); K. Doya, in *The Handbook of Brain Theory and Neural Networks*, edited by M. Arbib (MIT Press, Cambridge, MA, 1995).

[3] The initial condition for the partial dynamical system (3) is $v_*^i(0) = \frac{\partial y_i(0)}{\partial w_*} = 0$. In this example, $\{v_*^i\}$ corresponds to the information utilized in online learning.

[4] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, Cambridge, England, 1993).

[5] T. Sauer *et al.*, Phys. Rev. Lett. **79**, 59 (1997).

[6] The Lyapunov exponent of the logistic map is sufficiently negative at $a = 0.5$, and thus does not affect the largest finite-time Lyapunov exponent of the total learning system. The same applies to the case of $a = 3.835$.

[7] This dynamical behavior is quite different from that of adaptive control reported in [8,9], where gradient descent dynamics was basically studied.

[8] B. A. Huberman and E. Lumer, IEEE Trans. Circuits Syst. **37**, 547 (1990).

[9] S. Sinha *et al.*, Physica (Amsterdam) **43D**, 118 (1990).

[10] Similar sensitivity was reported in [11], for the learning of feedforward networks. This research, however, did not treat total dynamics, because such networks can represent only a static mapping. Also, Hondou and Sawada studied parameter dynamics for feedforward network learning [12].

[11] J. F. Kolen and J. B. Pollack, Complex Syst. **4**, 269 (1990).

[12] T. Hondou and Y. Sawada, Prog. Theor. Phys. **91**, 397 (1994).

[13] A. Saito and K. Kaneko, Physica (Amsterdam) **155D**, 1 (2001); Prog. Theor. Phys. **99**, 885 (1998).