Effect of time correlation of input patterns on the convergence of on-line learning

Tsuyoshi Hondou and Mitsuaki Yamamoto
Graduate School of Information Sciences, Tohoku University, Sendai 980-77, Japan

Yasuji Sawada and Yoshihiro Hayakawa
Research Institute of Electrical Communication, Tohoku University, Sendai 980-77, Japan

(Received 17 May 1995)

We studied the effects of time correlation of subsequent patterns on the convergence of on-line learning by a feedforward neural network with the backpropagation algorithm. By using a chaotic time series as sequences of correlated patterns, we found that the unexpected scaling of converging time with the learning parameter emerges when time-correlated patterns accelerate the learning process.

PACS number(s): 87.10.+e, 07.05.Mh, 05.45.+b

It has been reported [1–3] that time correlation of input patterns often largely influences the convergence of on-line learning. As a concrete example, learning of the chaotic map was shown to converge faster when patterns appeared in deterministic order of chaos than when patterns appeared randomly with the same "probability density" with the chaotic time series [1,2]. This showed that on-line learning is sensitive to the order of subsequent patterns. But the influence of the time correlation on the convergence of on-line learning has not been analyzed yet to our knowledge.

If we express the input and output as vectors, supervised learning is a task to acquire the mapping relation $\vec{X}_p$ $(\in \mathbb{R}^n) \mapsto \vec{Y}_p (\in \mathbb{R}^n)$ $(p \in \mathbb{N}$ or $\mathbb{R})$, where the set $\{\vec{X}_p, \vec{Y}_p\}$ is called a "pattern," and p is a pattern index. When the pattern index p is continuous, the number of patterns L is infinite. In gradient descent learning algorithms, the neural network system is updated as follows:

$$\vec{\omega}_{n+1} = \vec{\omega}_n + \delta\vec{\omega}_n ,$$
$$\delta\vec{\omega}_n = -\epsilon \nabla_{\vec{\omega}} E_n |_{\vec{\omega}=\vec{\omega}_n}, \tag{1}$$

where $\vec{\omega}_n$ is a weight vector at discrete time n, E_n is a generalized error, which depends on the learning procedure, and ϵ is a learning parameter.

Among several learning rules, the "backpropagation" algorithm [4], which is a natural extension of the steepest descent method to neural networks, is often used for its ability to realize the desired mapping relation in a network. The algorithm was originally formulated as an on-line learning procedure. The on-line procedure of the backpropagation can be divided into two kinds. The first one is "probabilistic on-line learning" (POL), which uses "local error," E_{p_n}, in Eq. (1): $E_{p_n}(\vec{X}_{p_n}, \vec{\omega}) = [\vec{\sigma}(\vec{\omega}) - \vec{Y}_{p_n}]^2/2$, where a pattern index p_n at discrete time n is drawn with pattern probability P_p satisfying $\Sigma_{p=1}^{L} P_p = 1$, and $\vec{\sigma}$ is an output of the network.

On the other hand, time-correlated input patterns into the network are often used, as in the case of the time series on-line learning. In such cases, the patterns may be presented in the deterministic order of appearance: $p_{n+1} = f(p_n)$, where f is a map that produces the time series of pattern indices. We call this second on-line learning procedure "deterministic on-line learning" (DOL). Although we will mainly analyze, in DOL, the case where the target function and the map that makes the sequence of pattern index coincide, more generally one can use dynamics that is making sequences of patterns different from the target function.

In contrast to the on-line learning, we also discuss "global learning" (GL), which is a modified algorithm of POL. The algorithm uses "global error," $E_{\mathrm{gl}}(\vec{\omega})$, in Eq. (1): $E_{\mathrm{gl}}(\vec{\omega}) = \int E_p(\vec{X}_p, \vec{\omega}) \rho(p) dp (p \in \mathbb{R})$, which is averaged error over patterns, where $\rho(p)$ is a probability density of the pattern with index p. The algorithm often gets easier for analysis, because the error does not depend on the special pattern.

Although on-line learning does not obey the exact gradient descent process of global error as in GL, complete randomness of subsequent patterns in the case of POL makes an analytical approach possible in the context of master equations, which is approximated by the Fokker-Planck equation in the limit of small learning parameters [5–8]. Exactly solvable models are also discussed in the literature [9,10].

Recently Wiegerinck and Heskes [11] showed theoretically that time correlation between subsequent patterns of on-line learning contributes to the diffusion term of a weight vector in the Fokker-Planck equation approximated from the equivalent equation as Eq. (1), and suggested that the result may help us to understand the accelerated on-line learning with time-correlated patterns found in [1,2].

In this paper, we study how the time correlation of subsequent patterns affects the convergence of learning by comparative studies of the two on-line learning procedures; (a) probabilistic on-line learning and (b) deterministic on-line learning. We use the tent map in most cases as a target mapping relation, because the map makes this comparative study easy. But the result is found to be similar for other maps. The tent map [12] is written as

$$x_{n+1} = f(x_n) = r(1 - 2|x_n - 1/2|). \tag{2}$$

We use a sequence of patterns that is produced by the tent map itself in DOL. When $r = 1$, the time series produced by the map has a white Fourier spectrum and a constant invariant density between [0, 1], the same as the uniformly random

number [0, 1], where the deterministic nature of chaotic correlation is expected to appear clearly in comparison with probabilistic randomness.

Let us now consider a conventional feedforward neural network with input and output terminals and $N-2$ hidden layers with M neurons. The output of the ith neuron of the mth layer of the network is as follows:

$$
\begin{aligned}
y_i^2 &= \tanh(\omega_i^1 x - \omega_{i0}^1), \\
y_i^3 &= \tanh\left(\sum_{j=1}^{M} \omega_{ij}^2 y_j^2 - \omega_{i0}^2\right), \\
&\vdots \quad \vdots \\
y_i^{N-1} &= \tanh\left(\sum_{j=1}^{M} \omega_{ij}^{N-2} y_j^{N-2} - \omega_{i0}^{N-2}\right), \\
\sigma &= \sum_{i=1}^{M} \omega_i^{N-1} y_i^{N-1}, \qquad (3)
\end{aligned}
$$

where ω_i^1, ω_{ij}^2, ..., ω_i^{N-1} are the synaptic weights connecting the input terminal to the second layer neurons, second to third layer, etc. and the (N-1)th to the output σ respectively; ω_{i0}^{m-1} is a bias term to the ith neuron of the mth layer. In this paper, we restrict ourselves for simplicity to the case where $N=4$ and $M=3$. The hidden layers (y^2, y^3, ..., y^{N-1}) have full interlayer connections. The local error E is written as

$$E(x_n, \vec{\omega}) = [\sigma(\vec{\omega}) - f(x_n)]^2/2, \qquad (4)$$

where $f(x)$ is the functional relationship of the tent map. Global error is also used in on-line learning to evaluate how learning progresses, because global error does not depend on the special input pattern x_n.

It is known that learning curves decrease suddenly between plateaus for many target functions and models. In the case of this tent map function learning, there usually exists a critical time when the global error $E_{\rm gl}$ decreases sharply, and the map learned by the network shifts abruptly from a constant to a tent [2]. Thus, one can easily define the converging time $t_{\rm cr}$ when the global error crosses the geometrical mean between $E_{\rm gl}$ on the first plateau and that on the second plateau (see Fig. 1). The typical learning curves of the tent map function are shown in Fig. 1. Generically, the three converging times of the tent map learning are found to satisfy the inequality $t_{\rm cr}^c \leq t_{\rm cr}^r \leq t_{\rm cr}^g$, where $t_{\rm cr}^c$, $t_{\rm cr}^r$, and $t_{\rm cr}^g$ are the converging times of DOL, POL, and GL, respectively. Notice that the invariant density $\rho(x)$ of GL and that of POL are always made the same as that of chaotic input (DOL) for comparative purposes. The order of three converging times is consistent with previous reports [1–3]. As one expects from the dynamical equations for weight vectors, the three converging times coincide for $\epsilon \to 0$.

How is the effect of deterministic randomness of subsequent patterns, which follows the chaotic time series, related to that of probabilistic ones? First we concentrate on this problem to discuss the difference of converging time between DOL [(b) in Fig. 1] and POL [(c) in Fig. 1]. Recent studies show that chaotic perturbation has anomalous effects on complex systems such as the Hopfield model [13] and general multistable systems [14], even if the simple statistical quantities (mean, variance, probability density, and Fourier spectrum) of chaos coincide with that of random noise. The effects are known to be related to the unstable fixed points of chaos. Chaotic force has transiently strong time correlation when the input pattern x is in the neighborhood of unstable fixed points; these are $x^*=0$ and 2/3 in the tent map with $r=1$. The nearer the input x injected to one of the unstable fixed points, the longer x stays in the neighborhood. Therefore the network of DOL *sees* biased (or, special) patterns for a while during which the input x stays several times in the vicinity of the unstable fixed point. In this period, the system moves continuously in the direction to reduce the special local error $E(x^*)$ for a while; i.e., the system is largely moved without constraint of global error due to the unstable fixed points of the chaotic map. This phenomenon is easily verified by numerical simulation as in Fig. 2. It should be noticed that the direction of the motion of the weight vector in this period is not necessarily the one that reduces the global error $E_{\rm gl}$. On the other hand, when the input x stays apart from an unstable fixed point, the sequence of input is almost as random as probabilistic; therefore the large change of weight vector in finite time steps is unlikely to occur, and the system is expected to move mostly along a gradient descent path of global error.

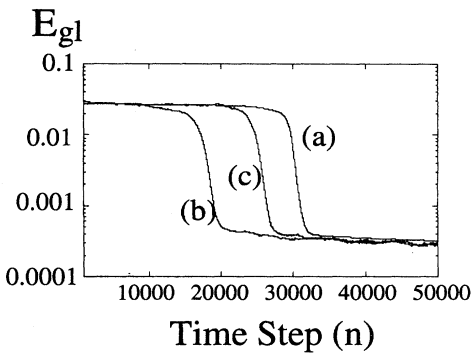

FIG. 1. Typical learning curves of the tent map function by three learning methods: (a) global learning, (b) deterministic on-line learning, and (c) probabilistic on-line learning. Invariant density in (a) and in (c) are the same as in (b). The initial conditions of the weight vectors are the same, $\epsilon=0.05$ and $r=0.95$.

The difference of time correlation of input patterns affects the convergence of learning largely even when all the simple statistical quantities coincide between the tent map chaos and the uniform random, as mentioned above. Therefore, the effect of this chaotic time correlation on the convergence of learning must be clarified. In DOL, the correlation range of input can be varied by changing of iteration number N, as the selection rule of the sequence of patterns x, as $x_{n+1}=f^N(x_n)$ by fixing the target function f, as $x_{n+1}=f(x_n)$. In the strong chaos limit, $N \to +\infty$, the time correlation of the subsequent input x dissappers: the sequence of input pattern is expected to be as random as probabilistic. Figure 3 shows that the time correlation of weak chaotic input ($N=1$) certainly works to accelerate time series learning. Fast decay of the effect of time correlation of subsequent patterns on the acceleration is observed: $t_{\rm cr}^c$ for

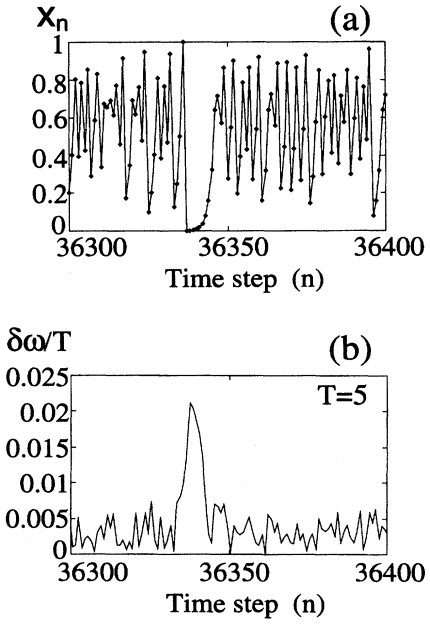

FIG. 2. (a) Typical temporal evolution of input x found in DOL, where $t<t_{\rm cr}^c$. (b) Corresponding time evolution of averaged velocities of a weight vector $\vec{\omega}$ in a finite time interval T, where $\delta\omega\equiv|\delta\vec{\omega}|=|\vec{\omega}_{n+T/2}-\vec{\omega}_{n-T/2}|$. Initial value of $\vec{\omega}$ is drawn from uniformly random number $[-0.05, 0.05]$, $\epsilon=0.05$, and $r=0.9995$.

$N=2$ is nearly equal to that for $N=100$. This is found to be consistent with the exponential decay of deterministic correlation with increasing N [14]. A saturated value of $t_{\rm cr}^c$ is equivalent to the one given by the learning time for random input (POL).

The effect of the time correlation of the input on the learning decreases with decreasing ϵ, and it is completely annihilated in the adiabatic limit, $\epsilon\to 0$, where the change of weight vector per unit time is so small that the evolution of the system is shortly averaged over pattern indices [15]. Therefore the dynamics of POL and DOL (and also GL) should coincide with each other in this limit. Equation (1) indicates that the continuous time T, as used in the Fokker-Planck description [11], should be proportional to ϵ. Therefore the converging time $t_{\rm cr}$ in the discrete model both for POL and for DOL should scale with ϵ^{-1}, and $\epsilon t_{\rm cr}$ should be independent of ϵ in the $\epsilon\to 0$ limit. However, it has not been understood how the finite learning parameter ϵ affects the accelerated learning, that is how ϵt_{cr} should behave with ϵ.

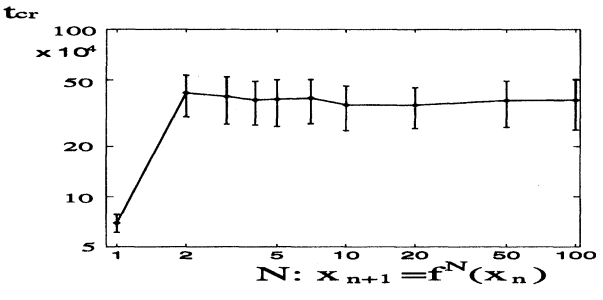

FIG. 3. Converging time $t_{\rm cr}^c$, vs several deterministic time correlation of input patterns x. Lyapunov exponent λ of the sequence of input is $\lambda=N\ln 2$. In the (strong chaos) limit, $N\to\infty$, the system is almost equivalent to POL with uniformly random input $[0,1]$. Ensemble averages over 100 initials are shown. Initial value of $\vec{\omega}$ is drawn from uniformly random number $[-0.05,0.05]$, $\epsilon=0.05$, and $r=0.9995$. It is found that $t_{\rm cr}^c(N\gg 1)\approx t_{\rm cr}^r$ (POL).

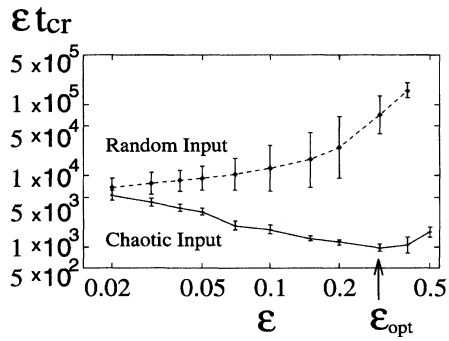

FIG. 4. Dependence of normalized converging time $\epsilon t_{\rm cr}$ on learning parameter ϵ for the tent map learning (solid line for DOL, dotted line for POL). Ensemble averages over 100 initials are shown. Initial value of $\vec{\omega}$ is drawn from uniformly random number $[-0.1, 0.1]$ and $r=0.9995$.

One finds from the result of simulation (Fig. 4) that the two normalized converging times $\epsilon t_{\rm cr}$ approach the same value in the small learning parameter limit ($\epsilon\to 0$). Approach of the normalized converging time to a finite value in the limit shows that there is no local minimum in the learning process. If there are any local minima in the learning process, the normalized converging time must diverge to infinity as $\epsilon\to 0$ [16]. In POL, the normalized converging time increases monotonically with increase of the learning parameter ϵ. However, in DOL, the normalized converging time $\epsilon t_{\rm cr}^c$ decreases first with increase of ϵ, and after some learning parameter $\epsilon_{\rm opt}$ it increases monotonically.

As known in general relaxation methods, finite stepping parameter ϵ is harmful because the possibility of overshooting in phase space increases as ϵ increases. Therefore, the normalized converging time is expected to increase monotonically with increasing learning parameter [10] as the result of overshooting in a learning process without local minima. The simulation showed that this is the case for POL but not necessarily for DOL (Fig. 4). The decrease of $\epsilon t_{\rm cr}^r$ with increase of ϵ was not observed in the simulations (Fig. 4). On the other hand, decrease of $\epsilon t_{\rm cr}^c$ was often found in chaotic patterns, not only in the learning of the tent map but also in that of the logistic map with several parameters.

As one notices, the reduction of converging time with increase of the learning parameter is possible when the system has to escape from local minima to reach the solution of learning [16]. But the present system has no local minima. One might think it strange that the normalized converging time decreases as the learning parameters increase in a pro-

cess without local minima. There should be an alternative that overcomes the harm of overshooting in the region $0<\epsilon<\epsilon_{\rm opt}$ in DOL.

We found that the puzzle may be solved by noticing the fact that in the learning process there are generically plural gradient descent paths to the solution. If chaotic correlation of subsequent patterns works effectively to find a shorter path to the solution by its diffusive motion of weight space, the observed phenomena are understandable. The possibility is strengthened by the fact that the system under chaotic patterns (DOL) should be largely moved away from the exact gradient descent direction of global error due to the unstable fixed points, which would facilitate the system to cross over the potential barrier between the gradient descent paths.

The same order of diffusive motion against gradient descent direction of global error, as found in DOL (see Fig. 1) would be possible, in principle, even in POL, with larger learning parameter ϵ. However, increase of ϵ strengthens the harm of overshooting simultaneously: the harm of overshooting may cancel the merit the diffusive motion in POL. In DOL, the harm of overshooting overcomes the merit of the diffusive motion when ϵ goes over $\epsilon_{\rm opt}$ where the normalized converging time begins to increase.

Finally, we mention an automatic reduction mechanism of the fluctuation of the system, which is characteristic of on-line learning and may weaken the harm of overshooting with a finite learning parameter. As discussed in Refs. [5,6], in "perfectly trainable networks" [17], in which $E_{\rm gl}(\vec{\omega})=0$ is available, the fluctuation in weight vector space (equivalently, the diffusion rate in Fokker-Planck representation [11]) becomes zero when the system reaches the error-free ($E_{\rm gl}=0$) state: the error-free state behaves as a "sink" of probability flow [6]. The reduction of the fluctuation can also occur even if the network is not "perfectly trainable": the system should be stabilized when the residual error is small enough [18].

We showed in this paper that the accelerated on-line learning with chaotic patterns is attributed to the unexpected scaling of the converging time with learning parameter ϵ: the converging time $t_{\rm cr}$ decreases much faster than $t_{\rm cr}\approx\epsilon^{-1}$ with increasing ϵ even without local minima. The results may indicate the beneficial aspects of finite learning parameters of on-line learning with time correlated patterns, because in any case one is forced to use finite learning parameters in realistic learning processes. The studies of the optimal time correlation of general patterns and/or the optimal learning parameter for the network, together with the proof of acceleration mechanism, are under way.

[1] G.J. Mpitsos and R.M. Burton, Neural Networks **5**, 605 (1992).

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

[3] H. Nakajima and Y. Ueda, in *Proceedings of the International Conference on Nonlinear Theory and Its Application* (Institute of Electronics, Information and Communication Engineers, Tokyo, 1993), p. 601.

[4] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel Distributed Processing* (MIT Press, Cambridge, 1986).

[5] G. Radons, H.G. Schuster, and D.Werner, in *Parallel Processing in Neural Systems and Computers*, edited by R. Eckmiller (Elsevier, Amsterdam, 1990), p. 261.

[6] T.M. Heskes and B. Kappen, Phys. Rev. A **44**, 2718 (1991).

[7] L.K. Hansen, R. Pathria, and P. Salamon, J. Phys. A **26**, 63 (1993).

[8] G. Radons, J. Phys. A **26**, 3455 (1993).

[9] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).

[10] D. Saad and S.A. Solla, Phys. Rev. Lett. **74**, 4337 (1995).

[11] W. Wiegerinck and T. Heskes, Europhys. Lett. **28**, 451 (1994).

[12] See, for example, H.G. Schuster, *Deterministic Chaos* (Physik-Verlag, Weinheim, Germany, 1984).

[13] Y. Hayakawa, A. Marumoto, and Y. Sawada, Phys. Rev. E **51**, 2693 (1995).

[14] T. Hondou, J. Phys. Soc. Jpn. **63**, 2014 (1994); T. Hondou and Y. Sawada, Phys. Rev. Lett. **75**, 3269 (1995); **76**, 1005 (1996).

[15] S. Amari, IEEE Trans. **EC-16**, 299 (1967).

[16] H.J. Kushner, SIAM. J. Appl. Math. **44**, 160 (1984).

[17] In feedforward neural networks with no less than three layers, any continuous function can be approximated with any precision. Proof appears in K. Funahashi, Neural Networks **2**, 183 (1989).

[18] An example is found in Fig. 4 of Ref. [2]; T. Hondou, Prog. Theor. Phys. (to be published).