Journal of the Physical Society of Japan
Vol. 75, No. 8, August, 2006, 084007

Statistical Mechanics of Time-Domain Ensemble Learning

Seiji MIYOSHI*, Tatsuya UEZU[1] and Masato OKADA[2,3,4]

Department of Electronic Engineering, Kobe City College of Technology, 8-3 Gakuenhigashimachi, Nishi-ku, Kobe 651-2194
[1]*Graduate School of Humanities and Sciences, Nara Women's University, Kitauoyahigashi-machi, Nara 630-8506*
[2]*Division of Transdisciplinary Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561*
[3]*RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198*
[4]*JST PRESTO, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561*

(Received May 9, 2006; accepted June 8, 2006; published August 10, 2006)

Conventional ensemble learning combines students in the space domain. On the other hand, in this paper, we combine students in the time domain and call it time-domain ensemble learning. We analyze the generalization performance of time-domain ensemble learning in the framework of on-line learning using a statistical mechanical method. We use a model in which both the teacher and the student are linear perceptrons with noises. Time-domain ensemble learning is twice as effective as conventional space-domain ensemble learning.

KEYWORDS: ensemble learning, on-line learning, generalization error, statistical mechanics
DOI: 10.1143/JPSJ.75.084007

1. Introduction

Learning is the inference of the underlying rules that dominate data generation using observed data. Observed data are input–output pairs from a teacher and are called examples. Learning can be roughly classified into batch learning and on-line learning.[1)] In batch learning, given examples are used more than once. In this paradigm, a student will give the correct answers after training if that student has adequate degree of freedom. However, a long time and a large memory, in which many examples are stored, are necessary. In contrast, in on-line learning, examples used once are then discarded. In this case, a student cannot give correct answers to all the examples used in training. However, there are merits; for example, a large memory for storing many examples is not necessary, and it is possible to follow a time-variant teacher.

Recently, we[2,3)] have analyzed the generalization performance of some models in a framework of on-line learning using a statistical mechanical method.[3–5)] Ensemble learning means to combine many rules or learning machines (called students in this paper) that perform poorly; it has recently attracted the attention of many researchers.[2,6–9)] The diversity or variety of students is important in ensemble learning. We showed that the three well-known rules, Hebbian learning, perceptron learning, and AdaTron learning have different characteristics in their affinities for ensemble learning, that is "maintaining diversity among students".[3)] In the analyses, the following points were proven subsidiarily.[10–13)] The student vector does not converge in one direction but continues moving in an unlearnable case.[14,15)] Therefore, we also analyzed the generalization performance of a student supervised by a moving teacher that goes around a true teacher.[4)] As a result, it was proven that the generalization error of a student can be smaller than that of a moving teacher, even if the student only uses examples from the moving teacher. In actual human society, a teacher observed by a student does not always present the correct answer. In many cases, the teacher is learning and continues to change. Therefore, the analysis of such a model is interesting for considering the analogies between statistical learning theories and actual human society.

In conventional ensemble learning, the generalization performance is improved by combining students who have diversities. On the other hand, a student does not always converge in one direction but may continue moving in an unlearnable model. Therefore, the generalization performance in such a model must be improved by combining the student at different times, even if there is only one student.[12,13)] Conventional ensemble learning combines students in the space domain. On the other hand, we introduce a method of combining the students in the time domain; we call this time-domain ensemble learning. In this paper, we analyze the generalization performance of the time-domain ensemble learning using a statistical mechanical method. We use a model in which both the teacher and the student are linear perceptrons[2)] with noises. We obtain the order parameters and generalization errors analytically in a framework of on-line learning using a statistical mechanical method.

2. Model

In this paper, we consider a teacher and a student. They are linear perceptrons with the connection weights $\boldsymbol{B}$ and $\boldsymbol{J}^m$, respectively. Here, m denotes the time step. For simplicity, the connection weights of the teacher and student are simply called the teacher and student, respectively. Teacher $\boldsymbol{B} = (B_1, \ldots, B_N)$, student $\boldsymbol{J}^m = (J_1^m, \ldots, J_N^m)$, and input $\boldsymbol{x}^m = (x_1^m, \ldots, x_N^m)$ are N-dimensional vectors. Each component B_i of $\boldsymbol{B}$ is independently drawn from $\mathcal{N}(0,1)$ and fixed, where $\mathcal{N}(0,1)$ denotes a Gaussian distribution with a mean of zero and a variance of unity. Each component J_i^0 of the initial value $\boldsymbol{J}^0$ of $\boldsymbol{J}^m$ is independently drawn from $\mathcal{N}(0,1)$. The direction cosine between $\boldsymbol{J}^m$ and $\boldsymbol{B}$ is R^m and that between $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$ is $q^{m,m'}$. Each component x_i^m of $\boldsymbol{x}^m$ is drawn from $\mathcal{N}(0, 1/N)$ independently. Thus,

*E-mail: miyoshi@kobe-kosen.ac.jp

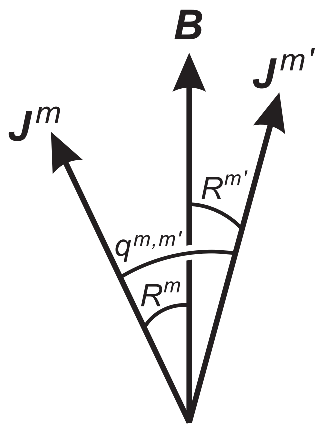


Fig. 1. Teacher $\boldsymbol{B}$ and students $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$. $R^m, R^{m'}$, and $q^{m,m'}$ are direction cosines.

$$\langle B_i \rangle = 0, \quad \langle (B_i)^2 \rangle = 1, \tag{1}$$

$$\langle J_i^0 \rangle = 0, \quad \left\langle \left(J_i^0\right)^2 \right\rangle = 1, \tag{2}$$

$$\langle x_i^m \rangle = 0, \quad \left\langle \left(x_i^m\right)^2 \right\rangle = \frac{1}{N}, \tag{3}$$

$$R^m \equiv \frac{\boldsymbol{B} \cdot \boldsymbol{J}^m}{\|\boldsymbol{B}\| \|\boldsymbol{J}^m\|}, \tag{4}$$

$$q^{m,m'} \equiv \frac{\boldsymbol{J}^m \cdot \boldsymbol{J}^{m'}}{\|\boldsymbol{J}^m\| \|\boldsymbol{J}^{m'}\|}, \tag{5}$$

where $\langle \cdot \rangle$ denotes a mean.

Figure 1 illustrates the relationship among teacher $\boldsymbol{B}$, students $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$ and the direction cosines $R^m, R^{m'}$, and $q^{m,m'}$.

In this paper, the thermodynamic limit $N \to \infty$ is also used. Therefore,

$$\|\boldsymbol{B}\| = \sqrt{N}, \quad \|\boldsymbol{J}^0\| = \sqrt{N}, \quad \|\boldsymbol{x}^m\| = 1. \tag{6}$$

Generally, the norm $\|\boldsymbol{J}^m\|$ of the student changes as the time step proceeds. Therefore, the ratio l^m of the norm to $\sqrt{N}$ is introduced and is called the length of the student. That is, $\|\boldsymbol{J}^m\| = l^m\sqrt{N}$.

Both the teacher and the student are linear perceptrons. Their outputs are $v^m + n_B^m$ and $u^m l^m + n_J^m$, respectively. Here,

$$v^m = \boldsymbol{B} \cdot \boldsymbol{x}^m, \tag{7}$$

$$u^m l^m = \boldsymbol{J}^m \cdot \boldsymbol{x}^m, \tag{8}$$

$$n_B^m \sim \mathcal{N}\left(0, \sigma_B^2\right), \tag{9}$$

$$n_J^m \sim \mathcal{N}\left(0, \sigma_J^2\right), \tag{10}$$

where $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with a mean of zero and variance σ^2. That is, the outputs of the teacher and student include independent Gaussian noises with variances of σ_B^2 and σ_J^2, respectively. Then, v^m and u^m obey Gaussian distributions with a mean of zero and a variance of unity.

Let us define the error ϵ_S^m between the teacher $\boldsymbol{B}$ and the student $\boldsymbol{J}^m$ by the squared error of their outputs:

$$\epsilon_S^m \equiv \frac{1}{2}\left(v^m + n_B^m - u^m l^m - n_J^m\right)^2. \tag{11}$$

Student $\boldsymbol{J}^m$ adopts the gradient method as a learning rule and uses input $\boldsymbol{x}$ and the output of teacher $\boldsymbol{B}$ for updates. That is,

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m - \eta \frac{\partial \epsilon_S^m}{\partial \boldsymbol{J}^m} \tag{12}$$

$$= \boldsymbol{J}^m + \eta\left(v^m + n_B^m - u^m l^m - n_J^m\right)\boldsymbol{x}^m, \tag{13}$$

where η denotes the learning rate of the student and is constant. Generalizing the learning rule, eq. (13) can be expressed as

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + f^m \boldsymbol{x}^m, \tag{14}$$

where f denotes a function that represents the update amount and is determined by the learning rule.

3. Theory

3.1 Generalization error

Ensemble learning means the improvement of performance by combining many students that perform poorly. On the other hand, we use just one student and combine its copies (hereafter called brothers) at different time steps. Conventional ensemble learning combines students in the space domain; on the other hand, we combine students in the time domain. In this paper, K brothers $\boldsymbol{J}^{m_1}, \boldsymbol{J}^{m_2}, \ldots, \boldsymbol{J}^{m_K}$ are combined. Here, $m_1 \le m_2 \le \cdots \le m_K$. We use the squared error ϵ for the new input $\boldsymbol{x}$. Here, it is assumed that the Gaussian noises of eqs. (9) and (10) are independently added to the teacher and each brother of the ensemble, respectively. The weight of each brother $\boldsymbol{J}^{m_k}$ of the ensemble satisfies $C_k > 0$. That is, the error of the ensemble is

$$\epsilon = \frac{1}{2}\left(\boldsymbol{B} \cdot \boldsymbol{x} + n_B - \sum_{k=1}^{K} C_k(\boldsymbol{J}^{m_k} \cdot \boldsymbol{x} + n_k)\right)^2. \tag{15}$$

Here, $n_B \sim \mathcal{N}\left(0, \sigma_B^2\right)$ and $n_k \sim \mathcal{N}\left(0, \sigma_J^2\right)$.

A goal of statistical learning theory is to theoretically obtain generalization errors. Since the generalization error is the mean of errors over the distribution of the new input $\boldsymbol{x}$ and noises $n_B, n_k, k = 1, \ldots, K$, the generalization error ϵ_g of the ensemble is calculated as follows:

$$\epsilon_g = \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}n_B \left(\prod_{k=1}^{K} \mathrm{d}n_k\right) p(\boldsymbol{x}) p(n_B) \left(\prod_{k=1}^{K} p(n_k)\right) \epsilon \tag{16}$$

$$= \int \mathrm{d}v \left(\prod_{k=1}^{K} \mathrm{d}u_k\right) \mathrm{d}n_B \left(\prod_{k=1}^{K} \mathrm{d}n_k\right) p(v, \{u_k\}) p(n_B) \times \left(\prod_{k=1}^{K} p(n_k)\right) \frac{1}{2}\left(v + n_B - \sum_{k=1}^{K} C_k(u_k l^{m_k} + n_k)\right)^2 \tag{17}$$

$$= \frac{1}{2}\left(1 - 2\sum_{k=1}^{K} C_k l^{m_k} R^{m_k} + 2\sum_{k=1}^{K}\sum_{k'>k}^{K} C_k C_{k'} l^{m_k} l^{m_{k'}} q^{m_k, m_{k'}} + \sum_{k=1}^{K} C_k^2 (l^{m_k})^2 + \sigma_B^2 + \sum_{k=1}^{K} C_k^2 \sigma_J^2\right), \tag{18}$$

where $v = \boldsymbol{B} \cdot \boldsymbol{x}$ and $u_k l^{m_k} = \boldsymbol{J}^{m_k} \cdot \boldsymbol{x}$. We performed integration using the following: v and u_k obey $\mathcal{N}(0, 1)$. The covariance between v and u_k is R^{m_k}, and that between u_k and $u_{k'}$ is $q^{m_k, m_{k'}}$. n_B and n_k are independent of other probabilistic variables.

3.2 Differential equations for order parameters and their analytical solutions

In this paper, we examine the thermodynamic limit $N \to \infty$. Therefore, updates for eqs. (13) or (14) must be executed $O(N)$ times for the order parameters l, R, and q to change by $O(1)$. Thus, the continuous times $t_1, \ldots, t_K$, which are the time steps $m_1, \ldots, m_K$ normalized by the dimension N, are introduced as the superscripts that represent the learning process. To simplify the analysis, we introduced the following auxiliary order parameters:

$$r^t \equiv l^t R^t, \tag{19}$$

$$Q^{t,t'} \equiv l^t l^{t'} q^{t,t'}. \tag{20}$$

The simultaneous differential equations in deterministic forms,[16] which describe the dynamical behaviors of order parameters, are obtained on the basis of the self-averaging of thermodynamic limits as follows:

$$\frac{\mathrm{d}l^t}{\mathrm{d}t} = \langle f^t u^t \rangle + \frac{\langle (f^t)^2 \rangle}{2l^t}, \tag{21}$$

$$\frac{\mathrm{d}r^t}{\mathrm{d}t} = \langle f^t v^t \rangle, \tag{22}$$

$$\frac{\mathrm{d}Q^{t,t'}}{\mathrm{d}t'} = l^t \langle f^{t'} \bar{u}^t \rangle, \tag{23}$$

where $t' \geq t$ and $\bar{u}^t = \boldsymbol{x}^{t'} \cdot \boldsymbol{J}^t / l^t \sim \mathcal{N}(0, 1)$.

Since linear perceptrons are used in this paper, the sample averages that appear in the above equations can be easily calculated as follows:

$$\langle f^t u^t \rangle = \eta (r^t / l^t - l^t), \tag{24}$$

$$\langle f^t v^t \rangle = \eta (1 - r^t), \tag{25}$$

$$\langle (f^t)^2 \rangle = \eta^2 (1 + \sigma_B^2 + \sigma_J^2 + (l^t)^2 - 2r^t), \tag{26}$$

$$\langle f^{t'} \bar{u}^t \rangle = \eta \big(r^t - Q^{t,t'}\big)/l^t. \tag{27}$$

Since all components of teacher $\boldsymbol{B}$ and initial student $\boldsymbol{J}^0$ are independently drawn from $\mathcal{N}(0, 1)$ and because the thermodynamic limit $N \to \infty$ is also used, they are orthogonal to each other in the initial state. That is,

$$R^0 = 0. \tag{28}$$

In addition,

$$l^0 = 1 \tag{29}$$

and

$$Q^{t,t} = (l^t)^2 \tag{30}$$

using eqs. (5) and (20). Using these initial conditions, we can analytically solve the simultaneous differential equations (21)–(27) as follows:

$$r^t = 1 - \mathrm{e}^{-\eta t}, \tag{31}$$

$$(l^t)^2 = 1 + \frac{\eta}{2 - \eta}\left(\sigma_B^2 + \sigma_J^2\right) - 2\mathrm{e}^{-\eta t} + \left(2 - \frac{\eta}{2 - \eta}\left(\sigma_B^2 + \sigma_J^2\right)\right)\mathrm{e}^{\eta(\eta - 2)t}, \tag{32}$$

$$Q^{t,t'} = 1 - \mathrm{e}^{-\eta t} + \mathrm{e}^{-\eta t'} + \big((l^t)^2 - 1\big)\mathrm{e}^{-\eta(t' - t)}. \tag{33}$$

Substituting eqs. (31)–(33) into eq. (18), the generalization error ϵ_g can be analytically obtained as a function of time t_k, $k = 1, \ldots, K$ as follows:

$$\begin{aligned}
\epsilon_\mathrm{g} = \frac{1}{2}\Bigg[& 1 - 2\sum_{k=1}^{K} C_k \big(1 - \mathrm{e}^{-\eta t_k}\big) \\
& + 2\sum_{k=1}^{K}\sum_{k'>k}^{K} C_k C_{k'} \big(1 - \mathrm{e}^{-\eta t_k} + \mathrm{e}^{-\eta t_{k'}} \\
& + \big(\bar{\sigma}^2 - 2\mathrm{e}^{-\eta t_k} + \big(2 - \bar{\sigma}^2\big)\mathrm{e}^{\eta(\eta-2)t_k}\big)\mathrm{e}^{-\eta(t_{k'} - t_k)}\big) \\
& + \sum_{k=1}^{K} C_k^2 \big(1 + \bar{\sigma}^2 - 2\mathrm{e}^{-\eta t_k} + \big(2 - \bar{\sigma}^2\big)\mathrm{e}^{\eta(\eta-2)t_k}\big) \\
& + \sigma_B^2 + \sum_{k=1}^{K} C_k^2 \sigma_J^2 \Bigg],
\end{aligned} \tag{34}$$

$$\bar{\sigma}^2 = \frac{\eta}{2 - \eta}\left(\sigma_B^2 + \sigma_J^2\right). \tag{35}$$

4. Results and Discussion

The dynamical behaviors of l^t and R^t are analytically obtained using eqs. (19), (31), and (32). Figures 2 and 3 show some examples of the analytical results and the corresponding simulation results for $N = 2000$. In these figures, the curves represent theoretical results, and the symbols represent simulation results. Figure 2 shows the results of $\sigma_B^2 = \sigma_J^2 = 0.0$ and no noise. Figure 3 shows the results of $\sigma_B^2 = \sigma_J^2 = 0.2$.

Focusing on the signs of the powers of the exponential functions in eq. (32), we can see that l^t diverges if the learning rate satisfies $0 > \eta$ or $\eta > 2$. l^t converges to

$$l^\infty = \sqrt{1 + \bar{\sigma}^2} \tag{36}$$

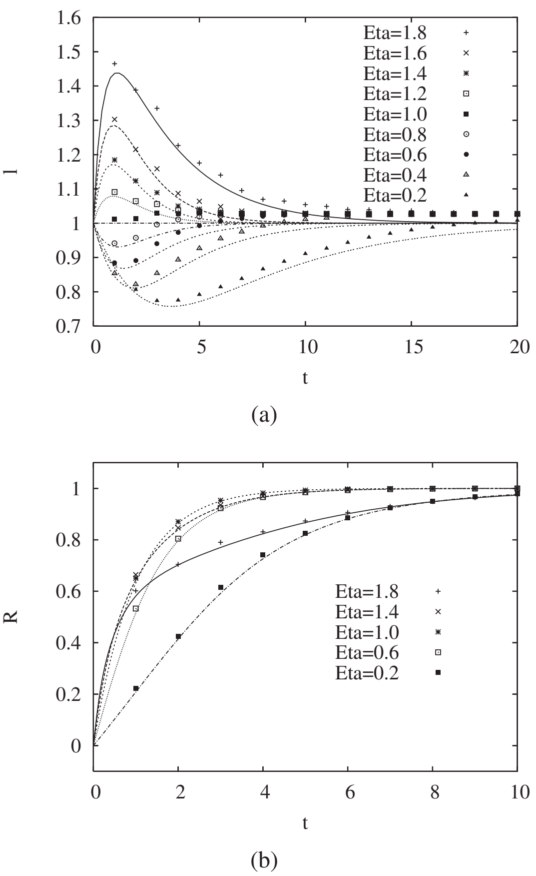


Fig. 2. Dynamical behaviors of (a) l^t and (b) R^t. Theory and computer simulations. $\sigma_B^2 = \sigma_J^2 = 0.0$.

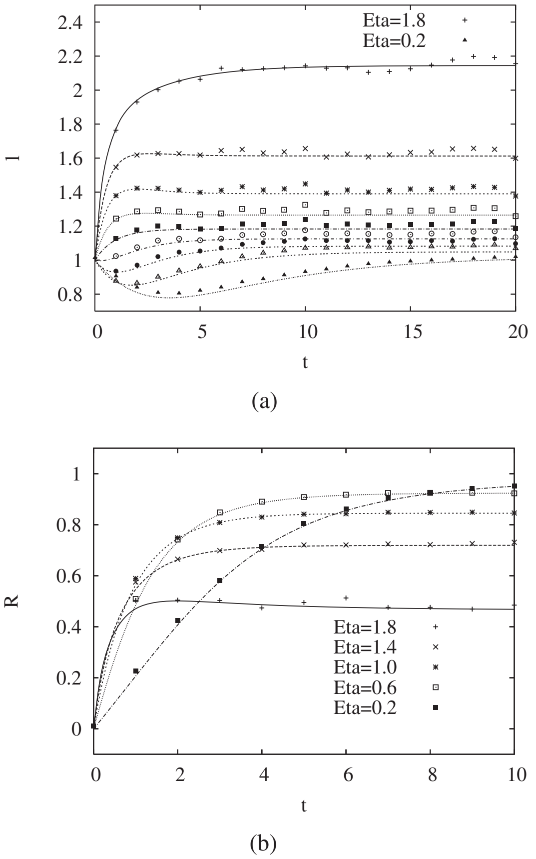


Fig. 3. Dynamical behaviors of (a) l^t and (b) R^t. Theory and computer simulations. $\sigma_B^2 = \sigma_J^2 = 0.2$. $\eta = 1.8$, 1.6, 1.4, 1.2, 1.0, 0.8, 0.6, 0.4, and 0.2 from the top.

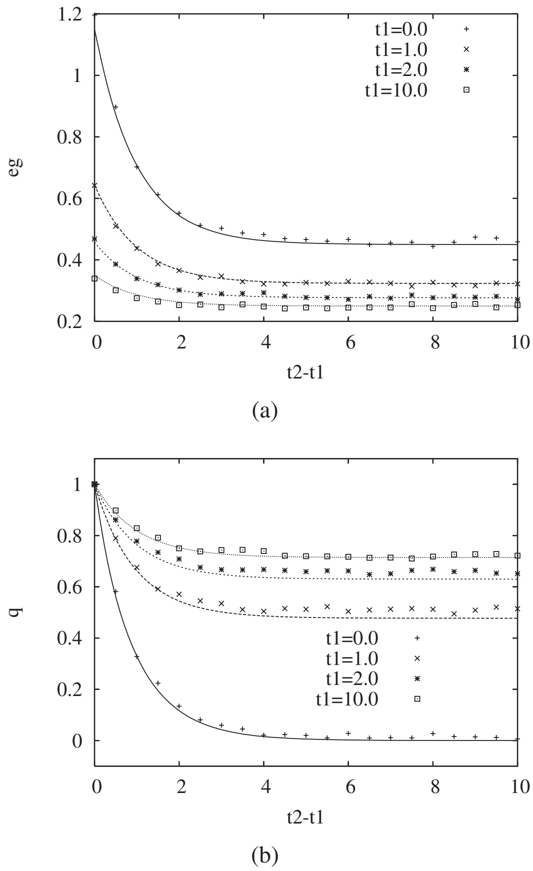


Fig. 4. Relationships between $t_2 - t_1$ and ϵ_g (a), q^{t_1,t_2} (b) in the case of constant leading time t_1. Theory and computer simulation. $\eta = 1.0$, $\sigma_B^2 = \sigma_J^2 = 0.2$.

if $0 < \eta < 2$. Equations (19) and (31) imply that R^t converges to

$$R^\infty = 1/l^\infty. \tag{37}$$

Therefore, we can see that $l^\infty = R^\infty = 1$ in the case of no noise, and $l^\infty > 1$ and $R^\infty < 1$ in the case of noise.

Since eq. (32) implies

$$\left.\frac{\mathrm{d}(l^t)^2}{\mathrm{d}t}\right|_{t=0} \begin{cases} < 0 & \text{when } \eta < \dfrac{2}{2+\sigma_B^2+\sigma_J^2}, \\ = 0 & \text{when } \eta = \dfrac{2}{2+\sigma_B^2+\sigma_J^2}, \\ > 0 & \text{when } \eta > \dfrac{2}{2+\sigma_B^2+\sigma_J^2}, \end{cases} \tag{38}$$

the equation that is a function of t

$$\frac{\mathrm{d}(l^t)^2}{\mathrm{d}t} = 0 \tag{39}$$

has only one solution

$$t = \frac{1}{\eta(1-\eta)} \ln\left(2 - \frac{\eta}{2}\left(2+\sigma_B^2+\sigma_J^2\right)\right), \tag{40}$$

if the learning rate satisfies $0 < \eta < 4/(2+\sigma_B^2+\sigma_J^2)$ and $\eta \neq 1$. Therefore, l^t asymptotically approaches unity after becoming larger than unity if $0 < \eta < 1$, and l^t asymptotically approaches unity after becoming smaller than unity if $1 < \eta < 2$, as shown in Fig. 2(a).

Equations (19), (31), and (32) imply

$$\left.\frac{\mathrm{d}R^t}{\mathrm{d}t}\right|_{t=0} = \eta. \tag{41}$$

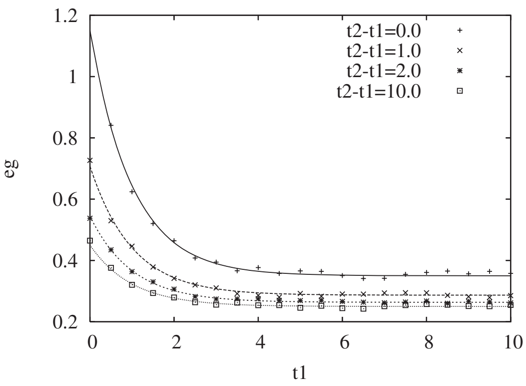
Therefore, the larger η is, the faster R increases, as shown in Figs. 2(b) and 3(b). However, eqs. (19), (31), (32), (36), and (37) imply

$$R^\infty - R^t = \frac{1}{l^\infty} - \frac{r^t}{l^t} \tag{42}$$

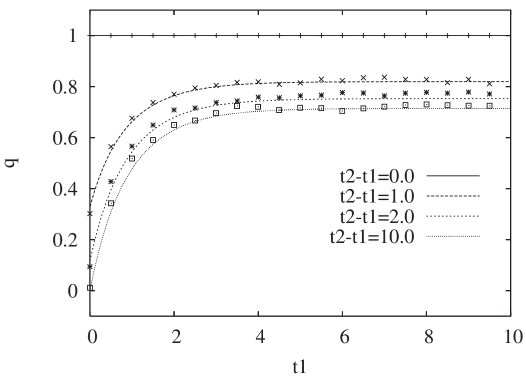
$$\to \left(1+\bar{\sigma}^2\right)\left(\bar{\sigma}^2 \mathrm{e}^{-\eta t} + \frac{2-\bar{\sigma}^2}{2}\mathrm{e}^{\eta(\eta-2)t}\right), \tag{43}$$

when t is large. Since eq. (43) is $O(\mathrm{e}^{-\eta t})$ if $0 < \eta \leq 1$ and $O(\mathrm{e}^{\eta(\eta-2)t}) = O(\mathrm{e}^{((\eta-1)^2-1)t})$ if $1 < \eta < 2$, the convergence of R^t is fastest when the learning rate satisfies $\eta = 1$. This can be confirmed in Figs. 2(b) and 3(b). This phenomenon can be understood by the fact that $\eta = 1$ is a special condition, for which the student uses up the information obtained from input $\boldsymbol{x}$.[5)]

We analytically obtained the dynamical behaviors of the generalization error ϵ_g and the direction cosine q using eqs. (20) and (32)–(35). Figures 4 and 5 show some examples of the analytical results and the corresponding simulation results for $N = 2000$. In these figures, the curves represent theoretical results, and the symbols represent simulation results. ϵ_g was calculated for the simplest case, that is, $K = 2$, $C_1 = C_2 = 1/2$. Other conditions are $\eta = 1.0$ and $\sigma_B^2 = \sigma_J^2 = 0.2$. In the computer simulation, ϵ_g was obtained by averaging the squared errors for 10^4 random inputs at each time step.



(a)



(b)

Fig. 5. Relationships between t_1 and ϵ_g (a), q^{t_1,t_2} (b) in the case of constant time interval $t_2 - t_1$. Theory and computer simulation. $\eta = 1.0$, $\sigma_B^2 = \sigma_J^2 = 0.2$.

Figure 4 shows the relationships between $t_2 - t_1$ and ϵ_g, q^{t_1,t_2} in the case of a constant t_1. When $t_2 - t_1$ increases, ϵ_g increases monotonically, remains constant, or decreases monotonically depending on the value of η. We prove this by the following. Equation (34) implies that

$$\epsilon_{g(K=1)} = \frac{1}{2}\left(\frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right) + \left(2 - \frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)\right)e^{\eta(\eta-2)t}\right). \quad (44)$$

Therefore, $\epsilon_{g(K=1)}$ decreases monotonically, remains constant, or increases monotonically as time t proceeds. The necessary and sufficient conditions for the above three phenomena are

$$\eta < \frac{4}{2+\sigma_B^2+\sigma_J^2}, \quad (45)$$

$$\eta = \frac{4}{2+\sigma_B^2+\sigma_J^2}, \quad (46)$$

$$\eta > \frac{4}{2+\sigma_B^2+\sigma_J^2}, \quad (47)$$

respectively. Since the output of the ensemble is the weighted sum of the outputs of the brothers, the generalization error for $K > 1$ also decreases monotonically, remains constant, or increases monotonically. The necessary and sufficient conditions for these three phenomena are also shown in eqs. (45)–(47). Since the condition of Fig. 4(a) is in agreement with eq. (45), the generalization error decreases monotonically. Equations (20), (32), (33), and (36) imply that in the case of $t_1 = 0$, q^{t_1,t_2} asymptotically approaches zero when $t_2 - t_1 \to \infty$ as shown in Fig. 4(b). This means that after a long time the student is orthogonal with its initial condition.

Since the order parameters and ϵ_g were explicitly obtained as functions of t and t' in eqs. (31)–(34), the relationships between t_1 and ϵ_g, $q^{t,t'}$ in the case of a constant time interval of the brothers or constant $t_{k+1} - t_k$ can be calculated. Figure 5 shows the relationships between t_1 and ϵ_g, q^{t_1,t_2} in the case of constant $t_2 - t_1$. For the same reason as in Fig. 4(a), the generalization error ϵ_g also decreases monotonically in Fig. 5(a). Figure 5(b) shows that q^{t_1,t_2} converges to a value smaller than unity in the case of $t_2 - t_1 \neq 0.0$. This means that the student continues to move after the order parameters l, R, and q reach a steady state.

In Figs. 4 and 5, the generalization error ϵ_g and the direction cosine q^{t_1,t_2} seem to almost reach a steady state by $t_2 - t_1 > 5$ or $t_1 > 5$. The behaviors of ϵ_g when the leading time $t_1 \to \infty$ or the time interval $t_{k+1} - t_k \to \infty$ can be theoretically obtained, since the generalization error and order parameters were analytically obtained as functions of t_k, $k = 1, \ldots, K$, as shown in eq. (34).

Equations (32) and (34) imply that, at first, $(l^t)^2$ diverges unless $0 < \eta < 2$. Therefore, the generalization error diverges unless $0 < \eta < 2$. If $0 < \eta < 2$, the generalization error can be discussed as follows:

When $t_1 \to \infty$, from eqs. (34) and (35) we obtain

$$\begin{aligned}\epsilon_g = \frac{1}{2}\Bigg[& 1 - 2\sum_{k=1}^{K} C_k + 2\sum_{k=1}^{K}\sum_{k'>k}^{K} C_k C_{k'} \\ & \times \left(1 + \frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)e^{-\eta(t_{k'}-t_k)}\right) \\ & + \sum_{k=1}^{K} C_k^2\left(1 + \frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)\right) \\ & + \sigma_B^2 + \sum_{k=1}^{K} C_k^2\sigma_J^2\Bigg]. \end{aligned} \quad (48)$$

In addition, when the time interval $t_{k+1} - t_k \to \infty$, from eq. (48) we obtain

$$\begin{aligned}\epsilon_g = \frac{1}{2}\Bigg[& 1 - 2\sum_{k=1}^{K} C_k + 2\sum_{k=1}^{K}\sum_{k'>k}^{K} C_k C_{k'} \\ & + \sum_{k=1}^{K} C_k^2\left(1 + \frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)\right) + \sigma_B^2 + \sum_{k=1}^{K} C_k^2\sigma_J^2\Bigg]. \end{aligned} \quad (49)$$

Equation (49) shows that the generalization error decreases as the learning rate η decreases regardless of K when $t_1 \to \infty$ and $t_{k+1} - t_k \to \infty$.

In addition, when the weights are uniform or $C_k = C = 1/K$, from eq. (49) we obtain

$$\epsilon_g = \frac{1}{2K}\left(\frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)\right) + \frac{1}{2}\left(\sigma_B^2 + \frac{1}{K}\sigma_J^2\right). \quad (50)$$

Here, considering the special case $K = 1$, we obtain

$$\epsilon_g = \frac{1}{2}\left(\frac{\eta}{2-\eta}\left(\sigma_B^2+\sigma_J^2\right)\right) + \frac{1}{2}\left(\sigma_B^2+\sigma_J^2\right). \quad (51)$$

If $\boldsymbol{B} = \boldsymbol{J}^{t_1}$, the generalization error must equal the residual error

$$\epsilon_g = \frac{1}{2}\left(\sigma_B^2 + \sigma_J^2\right) \tag{52}$$

caused by noise from eq. (15), which is the definition of error. Therefore, the difference between eqs. (51) and (52)

$$\frac{1}{2}\left(\frac{\eta}{2-\eta}\left(\sigma_B^2 + \sigma_J^2\right)\right) \tag{53}$$

is caused by the disagreement between $\boldsymbol{B}$ and $\boldsymbol{J}^{t_1}$.

Next, let us consider another special case, $K = \infty$. If and only if

$$\boldsymbol{B} = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{J}^{t_k}, \tag{54}$$

the generalization error must equal the residual error

$$\epsilon_g = \frac{1}{2}\sigma_B^2 \tag{55}$$

caused by noise from eq. (15), which is the definition of error. Equation (54) is true, since eq. (50) equals eq. (55) when $K = \infty$.

In addition, if $\sigma_B^2 = \sigma_J^2 = \sigma^2$, eq. (50) changes as follows:

$$\epsilon_g = \left(\frac{1}{2K}\frac{2+\eta}{2-\eta} + \frac{1}{2}\right)\sigma^2. \tag{56}$$

The relationship between the learning rate η and the generalization error ϵ_g can be analytically obtained using eq. (56) when both the leading time t_1 and the time interval $t_{k+1} - t_k$ are sufficiently large, and the uniform weight $C_k = C = 1/K$ and $\sigma_B^2 = \sigma_J^2 = 0.5$. Figure 6 shows the analytical results and the corresponding simulation results. In the computer simulation, with $N = 2000$, leading time $t_1 = 10$, and time interval $t_{k+1} - t_k = 10$ ($t_1 = t_{k+1} - t_k = 20$ when $\eta = 0.2$), we obtained ϵ_g by averaging the squared errors for 10^4 random inputs at each time step. Figure 6 confirms the following. The generalization error decreases as the learning rate η decreases. The generalization error decreases and converges to the residual error $\sigma_B^2/2$ as K increases.

In addition, if the learning rate satisfies $\eta = 1$, eq. (56) becomes

$$\epsilon_g = \left(\frac{3}{2K} + \frac{1}{2}\right)\sigma^2. \tag{57}$$

Equation (57) refers to the generalization error ϵ_g of $K = \infty$, which is $1/4$ that of $K = 1$ when the learning rate satisfies $\eta = 1$, uniform weights $C_k = 1/K$, $\sigma_B^2 = \sigma_J^2$, $t_1 \to \infty$, and $t_{k'} - t_k \to \infty$. Since the generalizaion error ϵ_g of conventional space-domain ensemble learning with $K = \infty$, $\eta = 1$, $C_k = 1/K$ and $\sigma_B^2 = \sigma_J^2$ is $1/2$ that of $K = 1$,[2] we can say that time-domain ensemble learning is twice as effective as conventional space-domain ensemble learning. We can explain this difference as follows: In conventional space-domain ensemble learning, the similarities among students become high, since all students use the same examples for learning. On the other hand, in time-domain ensemble learning, the similarities among brothers keep low, since all brothers use almost totally different examples for learning.

5. Conclusion and Future Work

We analyzed the generalization performance of time-domain ensemble learning in the framework of on-line learning using a statistical mechanical method. We used a model in which both the teacher and the student were linear perceptrons with noises. We showed that time-domain ensemble learning is twice as effective as conventional space-domain ensemble learning.

It would be interesting to analyze the time-domain ensemble learning of a model in which the teacher and student are nonlinear perceptrons.[12,13] In that case, it would be difficult to analytically obtain the generalization error and order parameters. However, it is expected that the nonlinear model will show qualitatively different and interesting behaviors. Analysis of the nonlinear model is our future aim.

Acknowledgments

This research was partially supported by Grants-in-Aid for Scientific Research 14084212, 15500151, and 16500093 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

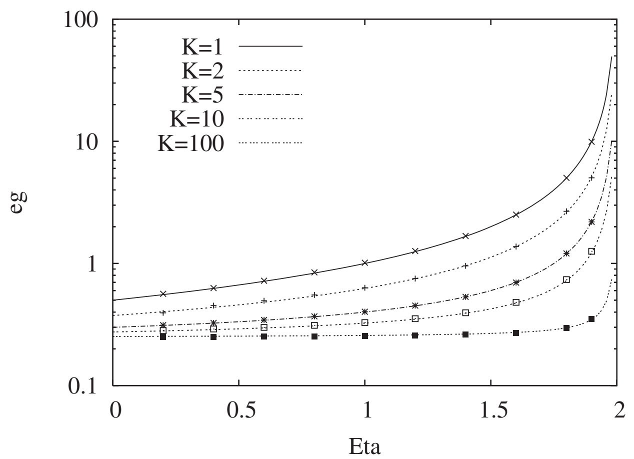


Fig. 6. Relationship between learning rate η and generalization error ϵ_g, when both leading time t_1 and time interval $t_{k+1} - t_k$ are sufficiently large. Theory and computer simulation. $C_k = C = 1/K$ and $\sigma_B^2 = \sigma_J^2 = 0.5$.

1) *On-line Learning in Neural Networks*, ed. D. Saad (Cambridge University Press, Cambridge, 1998).
2) K. Hara and M. Okada: J. Phys. Soc. Jpn. **74** (2005) 2966.
3) S. Miyoshi, K. Hara and M. Okada: Phys. Rev. E **71** (2005) 036116.
4) S. Miyoshi and M. Okada: J. Phys. Soc. Jpn. **75** (2006) 024003.
5) S. Miyoshi and M. Okada: J. Phys. Soc. Jpn. **75** (2006) 044002.
6) Y. Freund and R. E. Schapire: J. Jpn. Soc. Artif. Intell. **14** (1999) 771 [in Japanese, transl. N. Abe].
7) http://www.boosting.org/
8) A. Krogh and P. Sollich: Phys. Rev. E **55** (1997) 811.
9) R. Urbanczik: Phys. Rev. E **62** (2000) 1448.
10) S. Miyoshi, K. Hara and M. Okada: Proc. 7th Workshop Information-Based Induction Sciences, 2004, p. 178 [in Japanese].
11) S. Miyoshi, K. Hara and M. Okada: IEICE Tech. Rep. NC2004-214 (2005) p. 123 [in Japanese].
12) Y. Maeda: Masters Thesis, Graduate School of Humanities and Sciences, Nara Women's University (2002).
13) T. Uezu, Y. Maeda and S. Yamaguchi: in preparation.
14) J. I. Inoue and H. Nishimori: Phys. Rev. E **55** (1997) 4544.
15) J. I. Inoue, H. Nishimori and Y. Kabashima: cond-mat/9708096.
16) H. Nishimori: *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001).