Analysis of Ensemble Learning using Simple Perceptrons based on Online Learning Theory

Seiji Miyoshi
Department of Electronic Engineering
Kobe City College of Tech.
Kobe, 651-2194 Japan
E-mail: miyoshi@kobe-kosen.ac.jp

Kazuyuki Hara
Department of Electronics
and Information Engineering
Tokyo Metropolitan College of Tech.
Tokyo, 140-0011 Japan

Masato Okada
RIKEN Brain Science Institute
Wako, Saitama, 351-0198 Japan
Intelligent Cooperation and Control,
PRESTO, JST
Wako, Saitama, 351-0198 Japan

***Abstract*— Ensemble learning of K nonlinear perceptrons, which determine their outputs by sign functions, is discussed within the framework of online learning and statistical mechanics. This paper shows that ensemble generalization error can be calculated by using two order parameters, that is, the similarity between a teacher and a student, and the similarity among students. The differential equations that describe the dynamical behaviors of these parameters are derived analytically in the cases of Hebbian, perceptron and AdaTron learning. These three rules show different characteristics in their affinity for ensemble learning, that is "maintaining variety among students." Results show that AdaTron learning is superior to the other two rules.**

I. Introduction

Ensemble learning has recently attracted the attention of many researchers [1], [2], [3], [4], [5], [6]. Ensemble learning means to combine many rules or learning machines (students in the following) that perform poorly. Theoretical studies analyzing the generalization performance by using statistical mechanics[7], [8] have been performed vigorously[4], [5], [6].

Hara and Okada[4] theoretically analyzed the case in which students are linear perceptrons. Their analysis was performed with statistical mechanics, focusing on the fact that the output of a new perceptron, whose connection weight is equivalent to the mean of those of students, is identical to the mean outputs of students. Krogh and Sollich[5] analyzed ensemble learning of linear perceptrons with noises within the framework of batch learning.

On the other hand, Hebbian learning, perceptron learning and AdaTron learning are well-known as learning rules for a nonlinear perceptron, which decides its output by sign function[9], [10], [11], [12]. Urbanczik[6] analyzed ensemble learning of nonlinear perceptrons that decide their outputs by sign functions for a large K limit within the framework of online learning[13]. Though Urbanczik discussed ensemble learning of nonlinear perceptrons within the framework of online learning, he treated only the case in which the number K of students is large enough. Determining differences among ensemble learnings with Hebbian learning, perceptron learning and AdaTron learning (three typical learning rules), is a very attractive problem, but it is one that has never been analyzed to the best of our knowledge.

Based on the past studies, we discuss ensemble learning of K nonlinear perceptrons, which decide their outputs by sign functions within the framework of online learning and finite K [14], [15]. First, we show that an ensemble generalization error of K students can be calculated by using two order parameters: one is a similarity between a teacher and a student, the other is a similarity among students. Next, we derive differential equations that describe dynamical behaviors of these order parameters in the case of general learning rules. After that, we derive concrete differential equations about three well-known learning rules: Hebbian learning, perceptron learning and AdaTron learning. We calculate the ensemble generalization errors by using results obtained through solving these equations numerically. Two methods are treated to decide an ensemble output. One is the majority vote of students, and the other is an output of a new perceptron whose connection weight equals the means of those of students. As a result, we show that these three learning rules have different properties with respect to an affinity for ensemble learning, and AdaTron learning, which is known to have the best asymptotic property [9], [10], [11], [12], is the best among the three learning rules within the framework of ensemble learning.

II. Model

Each student treated in this paper is a perceptron that decides its output by a sign function. An ensemble of K students is considered. Connection weights of students are $\boldsymbol{J}_1, \boldsymbol{J}_2, \cdots, \boldsymbol{J}_K$. $\boldsymbol{J}_k = (J_{k1}, \cdots, J_{kN}), k = 1, 2, \cdots, K$ and input $\boldsymbol{x} = (x_1, \cdots, x_N)$ are N dimensional vectors. Each component x_i of $\boldsymbol{x}$ is assumed to be an independent random variable that obeys the Gaussian distribution $N(0, 1/N)$. Each component of $\boldsymbol{J}_k^0$, that is the initial value of $\boldsymbol{J}_k$, is assumed to be generated according to the Gaussian distribution $N(0, 1)$ independently. Each student's output is $\mathrm{sgn}(u_1 l_1), \mathrm{sgn}(u_2 l_2), \cdots, \mathrm{sgn}(u_K l_K)$ where

$$\mathrm{sgn}(u_k l_k) = \begin{cases} +1, & u_k l_k \geq 0, \\ -1, & u_k l_k < 0, \end{cases} \tag{1}$$

$$u_k l_k = \boldsymbol{J}_k \cdot \boldsymbol{x}. \tag{2}$$

Here, l_k denotes the length of student $\boldsymbol{J}_k$. This is one of the order parameters treated in this paper and will be described in detail later. In this paper, u_k is called a normalized internal potential of a student.

The teacher is also perceptron that decides its output by a sign function. The teacher's connection weight is $\boldsymbol{B}$. In this paper, $\boldsymbol{B}$ is assumed to be fixed where $\boldsymbol{B} = (B_1, \cdots, B_N)$ is also an N dimensional vector. Each component B_i is assumed to be generated according to the Gaussian distribution $N(0,1)$ independently. The teacher's output is $\mathrm{sgn}(v)$ where

$$v = \boldsymbol{B} \cdot \boldsymbol{x}. \tag{3}$$

Here, v represents an internal potential of the teacher. For simplicity, the connection weight of a student and that of the teacher are simply called student and teacher, respectively.

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

$$|\boldsymbol{x}| = 1, \quad |\boldsymbol{B}| = \sqrt{N}, \quad |\boldsymbol{J}_k^0| = \sqrt{N}, \tag{4}$$

where $|\cdot|$ denotes a vector norm. Generally, a norm of student $|\boldsymbol{J}_k|$ changes as the time step proceeds. Therefore, the ratio l_k of the norm to $\sqrt{N}$ is considered and is called a length of student $\boldsymbol{J}_k$. That is,

$$|\boldsymbol{J}_k| = l_k \sqrt{N}, \tag{5}$$

where l_k is one of the order parameters treated in this paper.

The common input $\boldsymbol{x}$ is presented to the teacher and all students in the same order. Each student compares its output and an output of the teacher for input $\boldsymbol{x}$. Each student's connection weight is corrected for the increasing probability that the student output agrees with that of the teacher. This procedure is called learning, and a method of learning is called learning rule, of which Hebbian learning, perceptron learning and AdaTron learning are well-known examples[9], [10], [11], [12]. Within the framework of online learning, information that can be used for correction other than that regarding a student itself is only input $\boldsymbol{x}$ and an output of the teacher for that input. Therefore, the update can be expressed as follows,

$$\boldsymbol{J}_k^{m+1} = \boldsymbol{J}_k^m + f_k^m \boldsymbol{x}^m, \tag{6}$$

$$f_k^m = f(\mathrm{sgn}(v^m), u_k^m), \tag{7}$$

where m denotes time step, and f is a function determined by learning rule.

III. Ensemble Generalization Error

One purpose of statistical learning theory is to theoretically obtain generalization error. In this paper, two methods are treated to determine an ensemble output. One is the majority vote of K students, which means an ensemble output is decided to be $+1$ if students whose outputs are $+1$ exceed the number of students whose outputs are -1, and -1 in the opposite case.

Another method for deciding an ensemble output is adopting an output of a new perceptron whose connection weight is the mean of the weights of K students. This method is simply called the weight mean in this paper.

We use

$$\epsilon = \Theta\left(-\mathrm{sgn}\left(\boldsymbol{B} \cdot \boldsymbol{x}\right) \mathrm{sgn}\left(\sum_{k=1}^{K} \mathrm{sgn}\left(\boldsymbol{J}_k \cdot \boldsymbol{x}\right)\right)\right), \tag{8}$$

and

$$\epsilon = \Theta\left(-\mathrm{sgn}\left(\boldsymbol{B} \cdot \boldsymbol{x}\right) \mathrm{sgn}\left(\left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{J}_k\right) \cdot \boldsymbol{x}\right)\right), \tag{9}$$

as error ϵ for the majority vote and the weight mean, respectively. Here, $\boldsymbol{\epsilon}$, $\boldsymbol{x}$ and $\boldsymbol{J}_k$ denote $\boldsymbol{\epsilon}^m$, $\boldsymbol{x}^m$ and $\boldsymbol{J}_k^m$, respectively. However, superscripts m, which represent time steps, are omitted for simplicity. Then, $\Theta(\cdot)$ is the step function defined as

$$\Theta(z) = \begin{cases} +1, & z \geq 0, \\ 0, & z < 0. \end{cases} \tag{10}$$

In both cases, $\epsilon = 0$ if an ensemble output agrees with that of the teacher and $\epsilon = 1$ otherwise. Generalization error ϵ_g is defined as the average of error ϵ over the probability distribution $p(\boldsymbol{x})$ of input $\boldsymbol{x}$. The generalization error ϵ_g can be regarded as the probability that an ensemble output disagrees with that of the teacher for a new input $\boldsymbol{x}$. In the case of a majority vote, using Eqs. (2), (3) and (8), we obtain

$$\epsilon = \Theta\left(-\mathrm{sgn}(v) \sum_{k=1}^{K} \mathrm{sgn}\left(u_k\right)\right). \tag{11}$$

In the case of a weight mean, using Eqs. (2), (3) and (9), we obtain

$$\epsilon = \Theta\left(-\mathrm{sgn}\left(v\right) \mathrm{sgn}\left(\sum_{k=1}^{K} u_k\right)\right). \tag{12}$$

That is error ϵ can be described as $\epsilon = \epsilon(\{u_k\}, v)$ by using a normalized internal potential u_k for the student and an internal potential v for the teacher in both cases. Therefore, the generalization error ϵ_g can be also described as

$$\epsilon_g = \int d\boldsymbol{x} p(\boldsymbol{x}) \epsilon = \int \prod_{k=1}^{K} du_k dv p(\{u_k\}, v) \epsilon(\{u_k\}, v), \tag{13}$$

by using the probability distribution $p(\{u_k\}, v)$ of u_k and v. As the thermodynamic limit $N \to \infty$ is also considered in this paper, u_k and v obey the multiple Gaussian distribution based on the central limit theorem. As input $\boldsymbol{x}$ and $\boldsymbol{J}_k$ have no correlation with each other within the framework of online learning, from Eq. (2), the mean and the variance of u_k are 0 and 1, respectively. In the same manner, since an input $\boldsymbol{x}$ and $\boldsymbol{B}$ have no correlation with each other, from Eq. (3), the mean and the variance of v are 0 and 1, respectively.

From these, all diagonal components of the covariance matrix $\boldsymbol{\Sigma}$ of $p(\{u_k\}, v)$ equal unity.

Let us discuss a direction cosine between connection weights as preparation for obtaining non-diagonal components. First, R_k is defined as a direction cosine between a teacher $\boldsymbol{B}$ and a student $\boldsymbol{J}_k$. That is,

$$R_k \equiv \frac{\boldsymbol{B} \cdot \boldsymbol{J}_k}{|\boldsymbol{B}||\boldsymbol{J}_k|} = \frac{1}{l_k N} \sum_{i=1}^{N} B_i J_{ki}. \tag{14}$$

When a teacher $\boldsymbol{B}$ and a student $\boldsymbol{J}_k$ have no correlation, $R_k = 0$, and $R_k = 1$ when the directions of $\boldsymbol{B}$ and $\boldsymbol{J}_k$

agree. Therefore, R_k is called the similarity between teacher and student in the following. Furthermore, R_k is the second order parameter treated in this paper. Next, $q_{kk'}$ is defined as a direction cosine between a student $\boldsymbol{J}_k$ and another student $\boldsymbol{J}_{k'}$. That is,

$$q_{kk'} \equiv \frac{\boldsymbol{J}_k \cdot \boldsymbol{J}_{k'}}{|\boldsymbol{J}_k||\boldsymbol{J}_{k'}|} = \frac{1}{l_k l_{k'} N} \sum_{i=1}^{N} J_{ki} J_{k'i}, \tag{15}$$

where $k \neq k'$. When a student $\boldsymbol{J}_k$ and another student $\boldsymbol{J}_{k'}$ have no correlation, $q_{kk'} = 0$, and $q_{kk'} = 1$ when the directions of $\boldsymbol{J}_k$ and $\boldsymbol{J}_{k'}$ agree. Therefore, $q_{kk'}$ is called the similarity among students in the following, and $q_{kk'}$ is the third order parameter treated in this paper.

Covariance between an internal potential v of a teacher $\boldsymbol{B}$ and a normalized internal potential u_k of a student $\boldsymbol{J}_k$ equals a similarity R_k between a teacher $\boldsymbol{B}$ and a student $\boldsymbol{J}_k$ as follows,

$$\langle v u_k \rangle = \left\langle \frac{1}{l_k} \sum_{i=1}^{N} B_i x_i \sum_{j=1}^{N} J_{kj} x_j \right\rangle = R_k, \tag{16}$$

where $\langle \cdot \rangle$ denotes the average. Covariance between a normalized internal potential u_k of a student $\boldsymbol{J}_k$ and a normalized internal potential $u_{k'}$ of another student $\boldsymbol{J}_{k'}$ equals a similarity $q_{kk'}$ among students as follows,

$$\langle u_k u_{k'} \rangle = \left\langle \frac{1}{l_k l_{k'}} \sum_{i=1}^{N} J_{ki} x_i \sum_{j=1}^{N} J_{k'j} x_j \right\rangle = q_{kk'}. \tag{17}$$

Therefore, Eq. (13) can be rewritten as

$$\epsilon_g = \int \prod_{k=1}^{K} du_k dv p(\{u_k\}, v) \epsilon(\{u_k\}, v), \tag{18}$$

$$p(\{u_k\}, v) = \frac{1}{(2\pi)^{\frac{K+1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \times \exp\left(-\frac{(\{u_k\}, v) \boldsymbol{\Sigma}^{-1} (\{u_k\}, v)^T}{2} \right), \tag{19}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & q_{12} & \cdots & q_{1K} & R_1 \\ q_{21} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & q_{K-1,K} & \vdots \\ q_{K1} & \cdots & q_{K,K-1} & 1 & R_K \\ R_1 & \cdots & \cdots & R_K & 1 \end{pmatrix}. \tag{20}$$

As a result, a generalization error ϵ_g can be calculated if all similarities R_k and $q_{kk'}$ are obtained. Let us thus discuss differential equations that describe dynamical behaviors of these order parameters. In this paper, norms of input, teacher and student are set as Eq. (4); influence of input can be replaced with the average over the distribution of inputs (sample average) in a large N limit. This idea is called self-averaging in statistical mechanics. Differential equations regarding l_k and R_k for general learning rules have been obtained based on self-averaging as follows[9],

$$\frac{dl_k}{dt} = \langle f_k u_k \rangle + \frac{\langle f_k^2 \rangle}{2 l_k}, \tag{21}$$

$$\frac{dR_k}{dt} = \frac{\langle f_k v \rangle - \langle f_k u_k \rangle R_k}{l_k} - \frac{R_k}{2 l_k^2} \langle f_k^2 \rangle, \tag{22}$$

where $\langle \cdot \rangle$ stands for the sample average.

Next, let us derive a differential equation regarding $q_{kk'}$ for the general learning rule. Considering a student $\boldsymbol{J}_k$ and another student $\boldsymbol{J}_{k'}$ and rewriting as $l_k^m \to l_k$, $l_k^{m+1} \to l_k + dl_k$, $q_{kk'}^m \to q_{kk'}$, $q_{kk'}^{m+1} \to q_{kk'} + dq_{kk'}$ and $1/N \to dt$, a differential equation regarding q is obtained as follows,

$$\frac{dq_{kk'}}{dt} = \frac{\langle f_{k'} u_k \rangle - q_{kk'} \langle f_{k'} u_{k'} \rangle}{l_{k'}} + \frac{\langle f_k u_{k'} \rangle - q_{kk'} \langle f_k u_k \rangle}{l_k} + \frac{\langle f_k f_{k'} \rangle}{l_k l_{k'}} - \frac{q_{kk'}}{2} \left(\frac{\langle f_k^2 \rangle}{l_k^2} + \frac{\langle f_{k'}^2 \rangle}{l_{k'}^2} \right), \tag{23}$$

from Eqs. (6), (15), (21) and self-averaging.

IV. Analytical Results

A. Conditions of analytical calculations

Similarities R_k and $q_{kk'}$ increase and approach unity as learning proceeds, leading to R_k and $q_{kk'}$ becoming less irrelevant to each other. For example when $R_k = R_{k'} = 1$, $q_{kk'}$ cannot be $\neq 1$ since a teacher $\boldsymbol{B}$, a student $\boldsymbol{J}_k$ and another student $\boldsymbol{J}_{k'}$ have the same direction. Thus, R_k and $q_{kk'}$ are under a certain restraint relationship each other. When $q_{kk'}$ is relatively smaller when compared with R_k, variety among students is further maintained and the effect of the ensemble can be considered as large. On the contrary, after $q_{kk'}$ becomes unity, a student $\boldsymbol{J}_k$ and another student $\boldsymbol{J}_{k'}$ are the same and there is no merit in combining them.

Let us explain these considerations intuitively by using Figure 1, and let us assume that learning starts from the condition that connection weights of students have no correlation. When learning has proceeded to some degree, K connection weight vectors $\boldsymbol{J}_k$ of K students must distribute at the same distance from connection weight vector $\boldsymbol{B}$ of the teacher, as shown in Figure 1. Especially, Figure 1(a) shows the case in which students are unlike each other — in other words the variety among students is large, that is, q is small. In this case, a mean vector $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{J}_k$ of the connection weights of students can closely approximate the connection weight vector $\boldsymbol{B}$ of the teacher. Thus a combination of students in some sense can approximate the teacher better than each student can do alone. In this case, the effect of ensemble learning is strong. On the contrary, Figure 1(b) shows the case in which students are similar to each other — in other words the variety among students is small, that is, q is large. In this case, the significance of combining three students is small. Therefore, effect of ensemble learning is small when q is large, as in Figure 1(b).

Thus, the relationship between R_k and $q_{kk'}$ is essential to know in ensemble learning. This relationship regarding linear perceptron has already been analyzed quantitatively in very

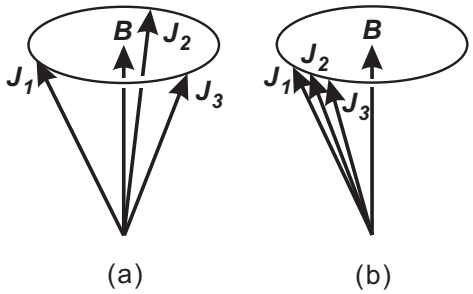

Fig. 1. Effect of ensemble learning ($K = 3$).

clear form[4]. Here, we analytically investigate the relationship between R_k and $q_{kk'}$ with respect to three learning rules of nonlinear perceptrons in the following.

As described above, in this paper each component of initial value $\boldsymbol{J}_k^0$ of student $\boldsymbol{J}_k$ and teacher $\boldsymbol{B}$ is generated independently according to the Gaussian distribution $N(0,1)$, and the thermodynamic limit $N \to \infty$ is considered. Therefore, all $\boldsymbol{J}_k^0$ and $\boldsymbol{B}$ are orthogonal to each other. That is,

$$R_k^0 = 0, \quad q_{kk'}^0 = 0. \tag{24}$$

From Eq. (24) and symmetry of students, we can write

$$\langle f_k u_{k'} \rangle = \langle f_{k'} u_k \rangle, \quad \langle f_k f_{k'} \rangle = \langle f_{k'} f_k \rangle \tag{25}$$

in Eq. (23). From Eq. (24) and symmetry among students, we omit subscript k, k' from order parameters l_k, R_k and $q_{kk'}$ in Eqs. (21)–(23) and write them as l, R and q. In the following sections, we discuss five sample averages $\langle f_k u_k \rangle$, $\langle f_k v \rangle$, $\langle f_k^2 \rangle$, $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ concretely, which are necessary to solve Eqs. (21)–(23) with respect to typical learning rules under the conditions given in Eqs. (24)–(25).

B. Hebbian learning

The update procedure for Hebbian learning is

$$f(\mathrm{sgn}(v), u) = \mathrm{sgn}(v). \tag{26}$$

Using this expression, $\langle f_k u_k \rangle$, $\langle f_k v \rangle$ and $\langle f_k^2 \rangle$ in the case of Hebbian learning can be obtained as follows analytically[9], [16].

$$\langle f_k u_k \rangle = \frac{2R}{\sqrt{2\pi}}, \quad \langle f_k v \rangle = \sqrt{\frac{2}{\pi}}, \quad \langle f_k^2 \rangle = 1. \tag{27}$$

In this section, $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ are derived. Since Eq.(26) is independent of u, we obtain

$$\langle f_k u_{k'} \rangle = \langle f_k u_k \rangle = \frac{2R}{\sqrt{2\pi}}, \quad \langle f_k f_{k'} \rangle = 1. \tag{28}$$

Figure 2 shows a comparison between the analytical results regarding the dynamical behaviors of R and q, which are obtained by solving Eqs.(21)–(25), (27)–(28) numerically and by computer simulation ($N = 10^5$). They closely agree with each other. That is, the derived theory explains the computer simulation quantitatively. Figure 2 shows that q rises more rapidly than R in Hebbian learning; in other words, q is relatively large when compared with R, meaning the variety among students disappears rapidly in Hebbian learning.

C. Perceptron learning

The update procedure for perceptron learning is

$$f(\mathrm{sgn}(v), u) = \Theta(-uv)\,\mathrm{sgn}(v). \tag{29}$$

Using this expression, $\langle f_k u_k \rangle$, $\langle f_k v \rangle$ and $\langle f_k^2 \rangle$ in the case of perceptron learning can be obtained as follows analytically[9], [16].

$$\langle f_k u_k \rangle = \frac{R-1}{\sqrt{2\pi}}, \quad \langle f_k v \rangle = \frac{1-R}{\sqrt{2\pi}}, \tag{30}$$

$$\langle f_k^2 \rangle = 2\int_0^\infty Dv H\left(\frac{Rv}{\sqrt{1-R^2}}\right) = \frac{1}{\pi}\tan^{-1}\frac{\sqrt{1-R^2}}{R}. \tag{31}$$

In this section, $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ are derived. Using Eq. (29), $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ in the case of perceptron learning are obtained as follows analytically:

$$\begin{aligned} \langle f_k u_{k'} \rangle &= \int du_k du_{k'} dv p_3(u_k, u_{k'}, v)\Theta(-u_k v)\mathrm{sgn}(v)u_{k'} \\ &= \frac{R-q}{\sqrt{2\pi}} \end{aligned} \tag{32}$$

$$\begin{aligned} \langle f_k f_{k'} \rangle &= \int du_k u_{k'} dv p_3(u_k, u_{k'}, v)\Theta(-u_k v)\Theta(-u_{k'} v) \\ &= 2\int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty DxH(z) \end{aligned} \tag{33}$$

where

$$z \equiv \frac{-(q-R^2)x + R\sqrt{1-R^2}v}{\sqrt{(1-q)(1+q-2R^2)}} \tag{34}$$

and the definitions of $H(u)$ and Dx are

$$H(u) \equiv \int_u^\infty Dx, \quad Dx \equiv \frac{dx}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right). \tag{35}$$

Figure 3 shows a comparison between the analytical results regarding the dynamical behaviors of R and q, which are obtained by solving Eqs. (21)–(25), (30)–(33) numerically and by computer simulation ($N = 10^5$). They closely agree with each other. That is, the derived theory explains the computer simulation quantitatively. Figure 3 shows that q is smaller than R in the early period of learning ($t < 4.0$), which means perceptron learning maintains the variety among students for a longer time than Hebbian learning.

D. AdaTron learning

The update procedure for AdaTron learning is

$$f(\mathrm{sgn}(v), u) = -u\Theta(-uv). \tag{36}$$

Using this expression, $\langle f_k u_k \rangle$, $\langle f_k v \rangle$ and $\langle f_k^2 \rangle$ in the case of AdaTron learning can be obtained as follows analytically[9], [16]:

$$\langle f_k u_k \rangle = -2\int_0^\infty Duu^2 H\left(\frac{Ru}{\sqrt{1-R^2}}\right) \tag{37}$$

$$= -\frac{1}{\pi}\cot^{-1}\left(\frac{R}{\sqrt{1-R^2}}\right) + \frac{1}{\pi}R\sqrt{1-R^2} \tag{38}$$

$$\langle f_k v \rangle = \frac{\left(1-R^2\right)^{\frac{3}{2}}}{\pi} + R\langle f_k u_k \rangle, \langle f_k^2 \rangle = -\langle f_k u_k \rangle \tag{39}$$

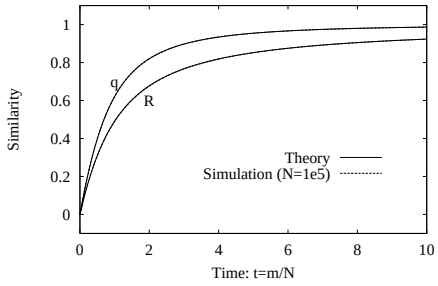

Fig. 2. Dynamical behaviors of R and q in Hebbian learning.

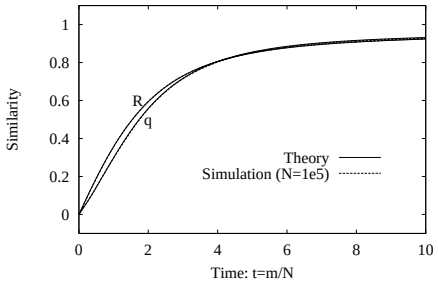

Fig. 3. Dynamical behaviors of R and q in perceptron learning.

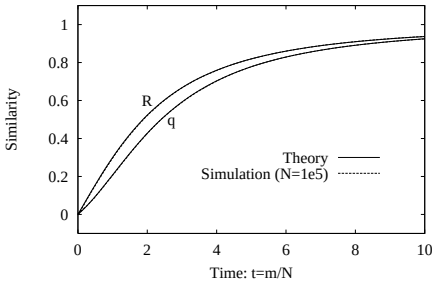

Fig. 4. Dynamical behaviors of R and q in AdaTron learning.

In this section, $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ are derived. Using Eq. (36), $\langle f_k u_{k'} \rangle$ and $\langle f_k f_{k'} \rangle$ in the case of AdaTron learning are obtained as Eqs. (40) and (41) analytically, where the definitions of z, $H(u)$ and Dx are Eqs. (34) and (35), respectively.

Figure 4 shows a comparison between the analytical results regarding the dynamical behaviors of R and q, which are obtained by solving Eqs. (21)–(25), (38)–(41) numerically and by computer simulation $(N = 10^5)$. They closely agree with each other. That is, the derived theory explains the computer simulation quantitatively. Figure 4 shows that q is relatively smaller when compared with R than in the cases of Hebbian learning and perceptron learning. This means AdaTron learning maintains variety among students most out of these three learning rules.

V. Discussion

The results in the previous section showed that AdaTron learning maintains the variety among students best out of the three learning rules. Thus, AdaTron learning is expected to be the best advanced for ensemble learning. To confirm this prediction, we have obtained numerical ensemble generalization errors ϵ_g in the case of $K = 3$ by using R and q for the three learning rules, that is Figures 2–4, and Eqs. (18)–(20). Figures 5–7 show the results. In these figures, MV and WM indicate the majority vote and weight mean, respectively. Numerical integrations of Eq. (18) in theoretical calculations have been executed by using the six-point closed Newton-Cotes formula. In the computer simulation, $N = 10^4$ and ensemble generalization errors have been obtained through tests using 10^5 random inputs at each time step. In each figure, the result of theoretical calculations of $K = 1$ is also shown to clarify the make effect of the ensemble.

These three figures show that the ensemble generalization errors obtained by theoretical calculation explain the computer simulation quantitatively. Though the generalization errors of the three learning rules are all improved by increasing K from 1 to 3, the degree of improvement is small in Hebbian learning and large in AdaTron learning. That is, the effect of the ensemble in AdaTron learning is the largest, as predicted above, due to the relationship between R and q. AdaTron learning originally featured the fastest asymptotic characteristic of the three learning rules[9]. However, it has disadvantage that the learning is slow at the beginning; that is, the generalization error is larger than for the other two learning rules in the period of $t < 6$. This paper shows that AdaTron learning has a good affinity with ensemble learning in regard to "the variety among students" and the disadvantage of the early period can be improved by combining it with ensemble learning.

From the perspective of the difference between the majority vote and the weight mean, Figure 5–7 show that the improvement by weight mean is larger than that by majority vote in all three learning rules. Improvement in the generalization error by averaging connection weights of various students can be understood intuitively because the mean of students is close to that of the teacher in Figure 1(a). The reason why the improvement in the majority vote is smaller than that in the weight mean is considered to be that the variety among students cannot be utilized as effectively by the majority vote as by the weight mean. However, the majority vote can determine an ensemble output only using outputs of students, and is easy to implement. It is, therefore, significant that the effect of an ensemble in the case of the majority vote has been analyzed quantitatively.

Figure 8 shows the results of computer simulations where $N = 10^3$, $K = 1, 3, 11, 31$ until $t = 10^4$ in order to investigate asymptotic behaviors of generalization errors. Asymptotic behavior of generalization error in AdaTron learning in the case of the number K of students at unity is $O(t^{-1})$[9], [12]. Asymptotic order of the generalization error in the case of ensemble learning is considered equal to that of $K = 1$, since properties of $K = 3, 11, 31$ are parallel to those of $K = 1$ in Figure 8. This figure also shows that the effect of ensemble learning on AdaTron learning is maintained asymptotically. Improvement in the generalization error tends to be saturated since the difference between the generalization error at $K = 11$ and that at $K = 31$ on a log scale is very small.

VI. Conclusion

This paper discussed ensemble learning of K nonlinear perceptrons, which determine their outputs by sign functions

$$\langle f_k u_{k'} \rangle = -\int du_k du_{k'} dv p_3(u_k, u_{k'}, v)\Theta(-u_k v)u_k u_{k'} = \frac{1+q}{\pi} R\sqrt{1-R^2} - 2q\int_0^\infty Dv \int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 \tag{40}$$

$$\begin{aligned}\langle f_k f_{k'} \rangle &= \int dv du_k u_k du_{k'} u_{k'} p_3(u_k, u_{k'}, v)\Theta(-u_k v)\Theta(-u_{k'} v) \\ &= \frac{(1-q)^2\left(1+q-2R^2\right)}{2\pi\left(1-R^2\right)^{\frac{3}{2}}}\left(\sqrt{\frac{(1+q)\left(1-R^2\right)}{1-q}} - R\right) + 2(q-R^2)\int_0^\infty Dv\int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx x^2 H(z) \\ &- \frac{2R\left(1+q-R^2\right)}{\sqrt{1-R^2}}\int_0^\infty Dv v\int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx H(z) + 2R^2\int_0^\infty Dv v^2\int_{\frac{Rv}{\sqrt{1-R^2}}}^\infty Dx H(z)\end{aligned} \tag{41}$$

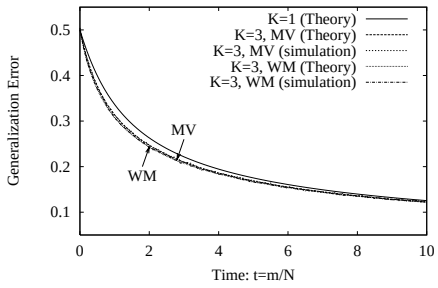

Fig. 5. Dynamical behaviors of ensemble generalization error ϵ_g in Hebbian learning.

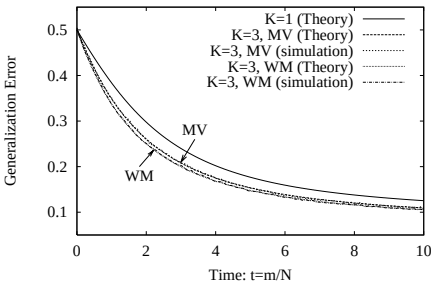

Fig. 6. Dynamical behaviors of ensemble generalization error ϵ_g in perceptron learning.

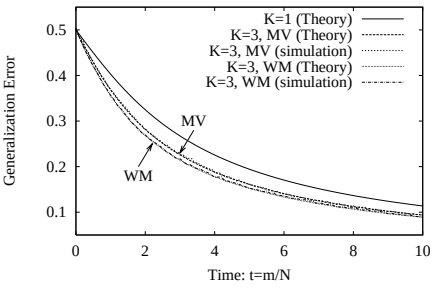

Fig. 7. Dynamical behaviors of ensemble generalization error ϵ_g in AdaTron learning.

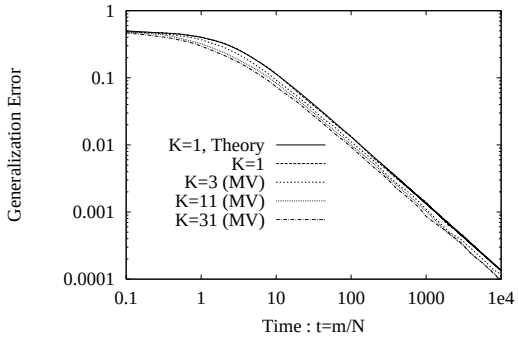

Fig. 8. Asymptotic behavior of generalization error of majority vote in AdaTron learning. Computer simulations, except for the solid line.

within the framework of online learning and statistical mechanics. We have shown that the ensemble generalization error can be calculated by using two order parameters, that is the similarity between the teacher and a student, and the similarity among students. The differential equations that describe the dynamical behaviors of these order parameters have been derived in the case of general learning rules. The concrete forms of these differential equations have been derived analytically in the cases of three well-known rules: Hebbian learning, perceptron learning and AdaTron learning. As a result, these three rules have different characteristics in their affinity for ensemble learning, that is, "maintaining variety among students." The results show that AdaTron learning is superior to the other two rules with respect to that affinity.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, with Grant-in-Aid for Scientific Research 13780313, 14084212, 14580438 and 15500151.

REFERENCES

[1] Y. Freund and R. E. Schapire, Journal of Japanese Society for Artificial Intelligence, **14**(5), 771 (1999) (in Japanese, translation by N. Abe.).
[2] Breiman, L., Machine Learning, **26**(2), 123 (1996).
[3] Y. Freund and R. E. Shapire, Journal of Comp. and Sys. Sci., **55**(1), 119 (1997).
[4] K. Hara and M. Okada, Proc. Fifth Workshop on Information-Based Induction Sciences, 113 (2002) (in Japanese).
[5] A. Krogh and P. Sollich, Phys. Rev. E **55**(1), 811 (1997).
[6] R. Urbanczik, Phys. Rev. E **62**(1), 1448 (2000).
[7] J. A. Hertz, A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
[8] M. Opper and W. Kinzel, in *Physics of Neural Networks III*, edited by E. Domany, J. L. van Hemmen and K. Schulten (Springer, Berlin, 1995).
[9] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001).
[10] J. K. Anlauf and M. Biehl, Europhys. Lett. **10**(7), 687 (1989).
[11] M. Biehl and P. Riegler, Europhys. Lett. **28**(7), 525 (1994).
[12] J. Inoue and H. Nishimori, Phys. Rev. E **55**(4), 4544 (1997).
[13] D. Saad, *On-line Learning in Neural Networks* (Cambridge University Press, Cambridge, 1998).
[14] S. Miyoshi, K. Hara and M. Okada, IEICE Technical Report **103**(228), 13 (2003) (in Japanese).
[15] S. Miyoshi, K. Hara and M. Okada, Proc. Annual Conf. of Japanese Neural Network Society, 104 (2003) (in Japanese).
[16] A. Engel and C. V. Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).