IPSJ Transactions on Mathematical Modeling and Its Applications Vol. 4 No. 1 72–81 (Jan. 2011)

Regular Paper

Statistical Mechanics of On-line Node-perturbation Learning

Kazuyuki Hara,[†1] Kentaro Katahira,[†2,†3,†4]
Kazuo Okanoya[†3,†2] and Masato Okada[†4,†3,†2]

Node-perturbation learning (NP-learning) is a kind of statistical gradient descent algorithm that estimates the gradient of an objective function through application of a small perturbation to the outputs of the network. It can be applied to problems where the objective function is not explicitly formulated, including reinforcement learning. In this paper, we show that node-perturbation learning can be formulated as on-line learning in a linear perceptron with noise, and we can derive the differential equations of order parameters and the generalization error in the same way as for the analysis of learning in a linear perceptron through statistical mechanical methods. From analytical results, we show that cross-talk noise, which originates in the error of the other outputs, increases the generalization error as the output number increases.

1. Introduction

Learning in neural networks [1] can be formulated as optimization of an objective function that quantifies the system's performance. This is achieved by following the gradient of the objective function with respect to the tunable parameters of the system. This optimization is computed directly by calculating the gradient explicitly and updating the parameters by a small step in the direction of the locally greatest improvement. However, computing a direct gradient to follow can be problematic. For instance, reinforcement learning has no explicit form of objective function, so we cannot calculate the gradient.

A stochastic gradient-following method to estimate gradient information is proposed for problems where a true gradient is not directly given. Node-perturbation learning (NP-learning) [2),3)] is one of the stochastic learning algorithms. NP-learning estimates the gradient of an objective function by using the change in the objective function in response to a small perturbation. NP-learning can be formulated as reinforcement learning with a scalar reward, and all the weight vectors are updated by using the scalar reward while the gradient method uses the reward vector. Hence, as well as being useful as a neural network learning algorithm, NP-learning can be formulated as reinforcement learning [2),4)] or it can be used in a brain model [5),6)].

On-line learning in a perceptron has been analyzed by many researchers using statistical mechanics methods and the exact behavior of the system has been depicted [7)–12)]. The network used in NP-learning has several outputs, each of which is a learning agent. The agent is treated as a simple linear perceptron [3)]. The objective function of NP-learning is the total error of the agents. The error is distributed according to the quantity of noise added to each output. When the output number is one, NP-learning is the same as on-line learning in a perceptron using a gradient descent algorithm (linear perceptron learning). When the output number is more than two, NP-learning can formulated as on-line learning in a linear perceptron with noise (noisy linear perceptron learning) [13),14)]. As we will show, NP-learning is a learning method similar to noisy linear perceptron learning, but it has not yet been analyzed using statistical mechanics methods.

In this paper, we analyze NP-learning following the analysis of noisy linear perceptron learning through a statistical mechanical method, and derive order parameter equations which depict NP-learning behavior. We then derive the generalization error using order parameters. From our results, we show that the cross-talk noise, which originates in the error of the other outputs, affects the learning performance in NP-learning and our analysis provides a new perspective regarding the analysis of NP-learning.

2. Model

In this section, we formulate the teacher and student networks, and an NP-learning algorithm employing a teacher-student formulation. We assume the teacher and student networks receive N-dimensional input $\boldsymbol{x}^{(m)} = (x_1^{(m)}, \ldots, x_N^{(m)})$ at the m-th learning iteration as shown in **Fig. 1**. Here, we

†1 College of Industrial Technology, Nihon University
†2 Japan Science Technology Agency, ERATO Okanoya Emotional Information Project
†3 Brain Science Institute, RIKEN
†4 Graduate School of Frontier Science, The University of Tokyo

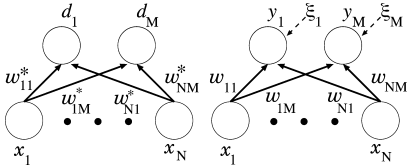

Fig. 1 Network structure of teacher (left) and student (right) networks, both with the same network structure. ξ_i represents a noise added to i-th student network output.

assume the existence of a teacher network vector $\boldsymbol{w}_k^*$ that produces desired output, so the teacher output d_k is a target of the student output y_k. The learning iteration m is ignored in the figure.

The teacher networks shown in Fig. 1 have N inputs and M outputs, and are identical to M linear perceptrons. Each student network has the same architecture as the teacher. $\boldsymbol{w}^*$ denotes the weight matrix of teacher networks with $M \times N$ elements. $\boldsymbol{w}^{(m)}$ denotes the weight matrix of student networks with $M \times N$ elements at the m-th learning iteration, the same as the teacher networks. We also assume that the elements $x_i^{(m)}$ of independently drawn input $\boldsymbol{x}^{(m)}$ are uncorrelated random variables with zero mean and variance $1/N$; that is, the i-th element of input is drawn from identical Gaussian distribution $P(x_i)$. In this paper, the thermodynamic limit of $N \to \infty$ is assumed. In the thermodynamic limit, the law of large numbers and the central limit theorem can apply. We can then depict the system behavior by using a small number of parameters. Statistics of the inputs at the thermodynamic limit are as follows.

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

where $\langle \cdot \rangle$ denotes the average over possible inputs and $|| \cdot ||$ denotes the norm of a vector.

M linear perceptrons are used as the teacher networks and are not subject to learning. Thus, the weight vectors $\{\boldsymbol{w}_k^*\}$, $k = 1, \ldots, M$ are fixed in the learning process. The output of the k-th teacher $d_k^{(m)}$ for N-dimensional input $\boldsymbol{x}^{(m)}$ at the m-th learning iteration is

$$d_k^{(m)} = \sum_{i=1}^{N} w_{ik}^* x_i^{(m)} = \boldsymbol{w}_k^* \cdot \boldsymbol{x}^{(m)}, \tag{2}$$

where teacher weight vectors $\{\boldsymbol{w}_k^*\}$, $\boldsymbol{w}_k^* = (w_{1k}^*, \ldots, w_{Nk}^*)$ are N-dimensional vectors, and each element w_{ik}^*, $i = 1, \ldots, N$ of teacher weight vectors $\boldsymbol{w}_k^*$ is drawn from a probability distribution of zero mean and unit variance. Assuming the thermodynamic limit of $N \to \infty$, statistics of the k-th teacher weight vector are

$$\langle w_{ik}^* \rangle = 0, \quad \left\langle (w_{ik}^*)^2 \right\rangle = 1, \quad ||\boldsymbol{w}_k^*|| = \sqrt{N}. \tag{3}$$

The distribution of the k-th output of the teacher networks follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit.

M linear perceptrons are used as student networks, and each student network has the same architecture as the teacher network. For the sake of analysis, we assume that each element of $w_{ik}^{(0)}$, which is the initial value of the k-th student vector $\boldsymbol{w}_k^{(m)}$, is drawn from a probability distribution of zero mean and unit variance. The norm of the k-th initial student weight vector $||\boldsymbol{w}_{ik}^{(0)}||$ is $\sqrt{N}$ in the thermodynamic limit of $N \to \infty$. Statistics of the k-th student weight vector are

$$\left\langle w_{ik}^{(0)} \right\rangle = 0, \quad \left\langle (w_{ik}^{(0)})^2 \right\rangle = 1. \tag{4}$$

The k-th student output $y_k^{(m)}$ for the N-dimensional input $\boldsymbol{x}^{(m)}$ at the m-th learning iteration is

$$y_k^{(m)} = \sum_{i=1}^{N} w_{ik}^{(m)} x_i^{(m)} = \boldsymbol{w}_k^{(m)} \cdot \boldsymbol{x}^{(m)}. \tag{5}$$

Generally, the norm of student weight vector $||\boldsymbol{w}_k^{(m)}||$, $\boldsymbol{w}_k^{(m)} = (w_{1k}^{(m)}, \ldots, w_{Nk}^{(m)})$ changes as the time step proceeds. Therefore, the ratio $l_k^{(m)}$ of the norm to $\sqrt{N}$ is considered and is called the length of student weight vector $\boldsymbol{w}_k^{(m)}$. The norm at the m-th iteration is $l_k^{(m)}\sqrt{N}$, and the order of $l_k^{(m)}$ is $O(1)$.

$$||\boldsymbol{w}_k^{(m)}|| = l_k^{(m)}\sqrt{N}. \tag{6}$$

The distribution of the k-th output of student $P(y_k)$ follows a Gaussian distribution of zero mean and l_k^2 variance in the thermodynamic limit of $N \to \infty$.

Next, we formulate the learning algorithm of the NP-learning. We follow Werfel's formulations of the learning [3]. For the possible inputs $\{\boldsymbol{x}\}$, we want to train the student network to produce desired outputs $\boldsymbol{y} = \boldsymbol{d}$. Here, $\boldsymbol{d} = (d_1, \ldots, d_M)$ and $\boldsymbol{y} = (y_1, \ldots, y_M)$. We employ the squared error as an error function. The squared error is defined by

$$E = \frac{1}{2}||\boldsymbol{d} - \boldsymbol{y}||^2 = \frac{1}{2}\sum_{k=1}^{M}(d_k - y_k)^2. \tag{7}$$

We train the student network within the framework of on-line learning. That is, input $\boldsymbol{x}^m$ given in each learning step is used to update a student weight vector according to the learning algorithm and is not used in the future learning. NP-learning is a stochastic learning, and a random noise vector $\boldsymbol{\xi}$ is added directly to the student network outputs $\boldsymbol{y}$ to perturb the error function E to estimate the gradient (see Fig. 1). Perturbed squared error E_{NP} is defined by

$$E_{NP} = \frac{1}{2}||\boldsymbol{d} - (\boldsymbol{y} + \boldsymbol{\xi})||^2 = \frac{1}{2}\sum_{k=1}^{M}(d_k - (y_k + \xi_k))^2. \tag{8}$$

If addition of the noise vector $\boldsymbol{\xi}$ lowers the error, the student weight vectors are adjusted in the direction of the noise vector. Here, each element ξ_k of the noise vector $\boldsymbol{\xi}$ is drawn from Gaussian distribution of zero mean and σ^2 variance. Werfel, et al. proposed the learning equation [3] as follows:

$$\boldsymbol{w}_k^{(m+1)} = \boldsymbol{w}_k^{(m)} - \frac{\eta}{\sigma^2}(E_{NP}^{(m)} - E^{(m)})\xi_k^{(m)}\boldsymbol{x}^{(m)}. \tag{9}$$

Here, $E_{NP}^{(m)} - E^{(m)}$ is the difference between the error with and without noise. Note that the difference $E_{NP}^{(m)} - E^{(m)}$ is assigned for each output $y_k^{(m)}$ from the independence of noise $\xi_k^{(m)}$ for each output. In this paper, Eq. (9) is used as the learning equation. η denotes the learning rate.

Next, we will show the relation between NP-learning and on-line learning in a perceptron using the gradient descent algorithm. On-line learning in a perceptron using the gradient descent algorithm is referred to as linear perceptron learning for simplicity. From Eq. (7), E with $M = 1$ is identical to the output error of the linear perceptron (student). Keeping this in mind, we get the next equation by substituting Eqs. (7) and (8) into Eq. (9).

$$\begin{aligned}\boldsymbol{w}_1^{m+1} &= \boldsymbol{w}_1^{(m)} - \frac{\eta}{\sigma^2}(E_{NP}^{(m)} - E^{(m)})\xi_1^{(m)}\boldsymbol{x}^{(m)} \\ &= \boldsymbol{w}_1^{(m)} + \frac{\eta}{2\sigma^2}\left(2(\xi_1^{(m)})^2(d_1^{(m)} - y_1^{(m)}) + (\xi_1^{(m)})^3\right)\boldsymbol{x}^{(m)}\end{aligned} \tag{10}$$

Here, if we assume $|\xi_1^{(m)}| \ll 1$, the term $(\xi_1^{(m)})^3$ is negligible and Eq. (10) becomes equivalent to the learning equation of linear perceptron learning. Thus, NP-learning with $M = 1$ can be considered linear perceptron learning.

On the other hand, from Eq. (7), E with $M \geq 2$ becomes the sum of all the students' output errors, and the error for each student is not given. Considering this, we get the next equation by substituting Eqs. (7) and (8) into Eq. (9):

$$\begin{aligned}\boldsymbol{w}_k^{(m+1)} &= \boldsymbol{w}_k^{(m)} - \frac{\eta}{\sigma^2}(E_{NP}^{(m)} - E^{(m)})\xi_k^{(m)}\boldsymbol{x}^{(m)} \\ &= \boldsymbol{w}_k^{(m)} + \frac{\eta}{2\sigma^2}\left[\sum_{k'=1}^{M} 2\xi_{k'}^{(m)}(d_{k'}^{(m)} - y_{k'}^{(m)}) - (\xi_{k'}^{(m)})^2\right]\xi_k^{(m)}\boldsymbol{x}^{(m)} \\ &= \boldsymbol{w}_k^{(m)} + \frac{\eta}{2\sigma^2}\Bigg[2(\xi_k^{(m)})^2(d_k^{(m)} - y_k^{(m)}) - (\xi_k^{(m)})^3 \\ &\quad + \sum_{k'\neq k}^{M} 2\xi_k^{(m)}\xi_{k'}^{(m)}(d_{k'}^{(m)} - y_{k'}^{(m)}) - \xi_k^{(m)}(\xi_{k'}^{(m)})^2\Bigg]\boldsymbol{x}^{(m)}.\end{aligned} \tag{11}$$

If we assumed $|\xi_k^{(m)}| \ll 1$, the terms of $(\xi_k^{(m)})^3$ and $\xi_k^{(m)}(\xi_{k'}^{(m)})^2$ become negligible. Then, from the viewpoint of signal-to-noise (S/N) analysis, the first term within the square brackets of Eq. (11) can be considered the signal because it relates to the k-th output which we observe. The third term within the brackets of the equation is the sum of all outputs except for the k-th output, and is a random variable not correlated to the k-th output error. The third term is thus considered the noise added to the output. Therefore, NP-learning with $M \geq 2$ is considered linear perceptron learning with noise [13]. Consequently, we analyze NP-learning through the same method as for linear perceptron learning with noise. (For an analysis of linear perceptron learning with noise, see the Appendix.)

3. Theory

In this paper, we consider the thermodynamic limit of $N \to \infty$ to analyze the

dynamics of the generalization error of the present system through statistical mechanics. In the following paragraphs, the iteration number m is omitted to simplify the notation of equations.

As pointed out, we will discuss learning based on on-line learning. In on-line learning, the input $\boldsymbol{x}$ is not used after the learning and the weight vector $\boldsymbol{w}_k$ is statistically independent of a new learning input. The squared error is then defined using the outputs of the teachers and those of students as given in Eqs. (2) and (5), respectively. The generalization error ε_g is given by the squared error E averaged over the possible input $\boldsymbol{x}$ drawn from a Gaussian distribution $P(\boldsymbol{x})$ of zero mean and $1/N$ variance.

$$\begin{aligned}\varepsilon_g &= \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) E \\ &= \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) \frac{1}{2} ||\boldsymbol{d} - \boldsymbol{y}||^2 \\ &= \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) \frac{1}{2} \left(\sum_{k=1}^{M} \left(\sum_{i=1}^{N} w_{ik}^* x_i - \sum_{i=1}^{N} w_{ik} x_i \right) \right)^2 . \end{aligned} \tag{12}$$

This calculation is the N-dimensional Gaussian integral with $\boldsymbol{x}$. We employ coordinate transformation from $\boldsymbol{x}$ to $\{d_k\}$ and $\{y_k\}$, $k = 1, \ldots, M$. Note that the distribution of the output of the student $P(y_k)$ follows a Gaussian distribution of zero mean and l_k^2 variance in the thermodynamic limit of $N \to \infty$. For the same reason, the output distribution for teacher $P(d_k)$ follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit. At the limit of $N \to \infty$, the distribution $P(d_k, y_k)$ of the k-th teacher output d_k and the k-th student output y_k is [11)]

$$P(d_k, y_k) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp\left[-\frac{(d_k \ y_k)\Sigma_k^{-1} \ (d_k \ y_k)^T}{2} \right], \tag{13}$$

$$\Sigma_k = \begin{pmatrix} 1 & r_k \\ r_k & l_k^2 \end{pmatrix}. \tag{14}$$

Here, T denotes the transpose of a vector and $r_k = R_k l_k$. R_k is the overlap between the teacher weight vector $\boldsymbol{w}_k^*$ and the student weight vector $\boldsymbol{w}_k$. Overlap R_k is defined as

$$R_k = \frac{\boldsymbol{w}_k^* \cdot \boldsymbol{w}_k}{||\boldsymbol{w}_k^*|| ||\boldsymbol{w}_k||} = \frac{\boldsymbol{w}_k^* \cdot \boldsymbol{w}_k}{N l_k}. \tag{15}$$

Hence, by using this coordinate transformation, the generalization error in Eq. (12) can be rewritten as

$$\begin{aligned}\varepsilon_g(t) &= \int \prod_{k=1}^{M} \mathrm{d}d_k \mathrm{d}y_k P(\boldsymbol{d}, \boldsymbol{y}) \frac{1}{2} \sum_{k=1}^{M} (d_k - y_k)^2 \\ &= \frac{1}{2} \sum_{k=1}^{M} (1 - 2r_k(t) + l_k^2(t)) \end{aligned} \tag{16}$$

Here, $t = m/N$. Consequently, we calculate the dynamics of the generalization error by substituting the time step value of $l_k(t)$ and $r_k(t)$ into Eq. (16).

Next, we derive the differential equations of order parameters l_k and r_k by following analysis of linear perceptron learning with noise [13),14)]. (For an analysis of linear perceptron learning with noise, see the Appendix.) For the sake of convenience, we write the overlap as r_k. From Eqs. (38) and (41) in the appendix, the differential equations of two order parameters of the k-th student l_k^2 and the overlap r_k are given by the equations

$$\frac{\mathrm{d}l_k^2}{\mathrm{d}t} = 2\eta \left\langle f_k y_k \right\rangle + \eta^2 \left\langle f_k^2 \right\rangle, \tag{17}$$

$$\frac{\mathrm{d}r_k}{\mathrm{d}t} = \eta \left\langle f_k d_k \right\rangle, \tag{18}$$

where $\langle \cdot \rangle$ denotes the average over possible inputs and perturbation noises, and

$$f_k = \frac{1}{2\sigma^2} \left[2\xi_k^2 (d_k - y_k) - \xi_k^3 + \sum_{k' \neq k}^{M} \left(2\xi_k \xi_{k'} (d_{k'} - y_{k'}) - \xi_k \xi_{k'}^2 \right) \right]. \tag{19}$$

The averages in Eqs. (17) and (18) are calculated as

$$\begin{aligned}\langle f_k y_k \rangle &= \frac{1}{2\sigma^2} \left\langle 2\xi_k^2 \ (d_k - y_k) y_k - \xi^3 y_k + \sum_{k' \neq k}^{M} 2\xi_k \xi_{k'} (d_{k'} - y_{k'}) y_k - \xi_k \xi_{k'}^2 y_k \right\rangle \\ &= r_k - l_k^2, \end{aligned} \tag{20}$$

$$\left\langle f_k^2 \right\rangle = \frac{1}{4\sigma^4} \left\langle \left(2\xi_k^2 (d_k - y_k) - \xi_k^3 + \sum_{k' \neq k}^{M} \left(2\xi_k \xi_{k'} (d_{k'} - y_{k'}) - \xi_k \xi_{k'}^2 \right) \right)^2 \right\rangle$$

$$= 3(1 - 2r_k + l_k^2) + \sum_{k' \neq k}^{M} (1 - 2r_{k'} + l_{k'}^2) + \frac{1}{4}(M+2)(M+4)\sigma^2, \quad (21)$$

$$\langle f_k d_k \rangle = \frac{1}{2\sigma^2} \left\langle 2\xi_k^2 (d_k - y_k) d_k - (\xi_k)^3 d_k + \sum_{k' \neq k}^{M} 2\xi_k \xi_{k'} (d_{k'} - y_{k'}) d_k - \xi_k \xi_{k'}^2 d_k \right\rangle$$
$$= 1 - r_k. \quad (22)$$

Here, we have used $\langle x_k^2 \rangle = 1/N$, $\langle d_k^2 \rangle = 1$, $\langle y_k^2 \rangle = l_k^2$, $\langle \xi_k^2 \rangle = \sigma^2$, $\langle \xi_k^4 \rangle = 3\sigma^4$, $\langle \xi_k^6 \rangle = 15\sigma^6$ and $\langle \xi_k \rangle = \langle \xi_k^3 \rangle = \langle \xi_k^5 \rangle = 0$ since these variables obey zero mean Gaussian distributions. The cross-talk noise, which originates in the error of the other outputs, appears in Eq. (21) from the average of the second-order cross-talk noise $\langle \xi_k^2 \xi_{k'}^2 \rangle$, while the average of the first-order, the cross-talk noise $\langle \xi_k \xi_{k'} \rangle$, is eliminated from Eqs. (20) and (22). By substituting Eqs. (20)–(22) into Eqs. (17) and (18), we get the following differential equations.

$$\frac{\mathrm{d}l_k^2}{\mathrm{d}t} = 2\eta(r_k - l_k^2) + \eta^2 \Bigg[3(1 - 2r_k + l_k^2) + \sum_{k' \neq k}^{M} (1 - 2r_{k'} + l_{k'}^2) + \frac{1}{4}(M+2)(M+4)\sigma^2 \Bigg], \quad (23)$$

$$\frac{\mathrm{d}r_k}{\mathrm{d}t} = \eta(1 - r_k), \quad (24)$$

Here, $k' \neq k$. From the above results, we found that the effect of cross-talk noise appears in l_k^2 but not in r_k.

4. Results

In this section, we discuss the dynamics of the order parameters and their asymptotic properties, and then derive the analytical solution of generalization error. Finally, we discuss the validity of analytical results by comparison with simulation results.

For the sake of simplicity, the initial weight vectors of the teachers and students are homogeneously correlated, so we assume $l_k^{(0)} = l^{(0)}$ and $r_k^{(0)} = r^{(0)}$. From the symmetry of the evolution equation for updating the weight vector,

$$l_k^{(t)} = l^{(t)}, r_k^{(t)} = r^{(t)}, \quad (25)$$

are obtained. Substituting Eq. (25) into Eqs. (23) and (24), we get

$$\frac{\mathrm{d}l^2}{\mathrm{d}t} = 2\eta(r - l^2) + \eta^2 \left[(M+2)(1 - 2r + l^2) + \frac{1}{4}(M+2)(M+4)\sigma^2 \right], \quad (26)$$

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \eta(1 - r). \quad (27)$$

Here, Eqs. (26) and (27) form closed differential equations. Equation (27) can be solved analytically.

$$r^{(t)} = 1 - (1 - r^{(0)})e^{-\eta t}. \quad (28)$$

By substituting Eq. (28) into Eq. (26), $l^{(t)}$ is also solved analytically:

$$(l^{(t)})^2 = \left(1 + \frac{(M+2)(M+4)\eta\sigma^2}{4(2 - (M+2)\eta)} \right) - 2\left(1 - r^{(0)}\right) e^{-\eta t} + \left(1 - 2r^{(0)} + (l^{(0)})^2 - \frac{(M+2)(M+4)\eta\sigma^2}{4(2 - (M+2)\eta)} \right) e^{-\eta(2-(M+2)\eta)t}, \quad (29)$$

where $(l^{(0)})^2$ is the initial value of $(l^{(t)})^2$. From Eq. (29), cross-talk noise appears on $(l^{(t)})^2$ and it makes $(l^{(t)})^2$ larger as output number M increases.

Next, we derive an analytical solution of the generalization error. By substituting Eq. (25) into Eq. (16), we get

$$\varepsilon_g^{(t)} = \int \prod_{k=1}^{M} \mathrm{d}d_k \mathrm{d}y_k P(\boldsymbol{d}, \boldsymbol{y}) \frac{1}{2} \sum_{k=1}^{M} (d_k - y_k)^2 = \frac{M}{2}(1 - 2r^{(t)} + (l^{(t)})^2). \quad (30)$$

By substituting Eqs. (29) and (28) into Eq. (30), we can rewrite the generalization error of NP-learning:

$$\varepsilon_g^{(t)} = \frac{M}{2} \Bigg[\frac{\eta\sigma^2(M+2)(M+4)}{4(2 - (M+2)\eta)} + \left(1 - 2r^{(0)} + (l^{(0)})^2 - \frac{\eta\sigma^2(M+2)(M+4)}{4(2 - (M+2)\eta)} \right) e^{-\eta(2-(M+2)\eta)t} \Bigg]$$

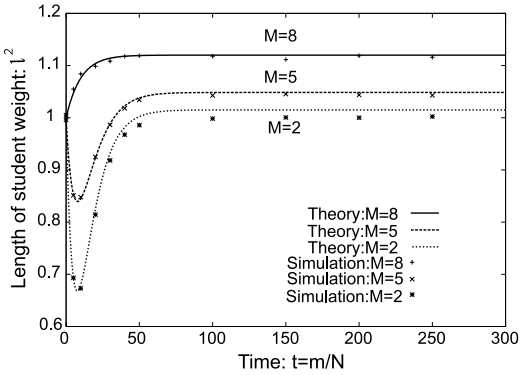

Fig. 2 Learning time dependence of the student length.

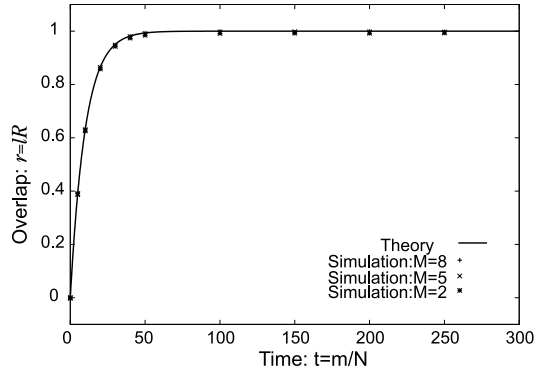

Fig. 3 Learning time dependence of overlap between teacher and student.

$$= \frac{\eta\sigma^2 M(M+2)(M+4)}{8(2-(M+2)\eta)} + \left(\varepsilon_g^{(0)} - \frac{\eta\sigma^2 M(M+2)(M+4)}{8(2-(M+2)\eta)}\right) e^{-\eta(2-(M+2)\eta)t}. \quad (31)$$

From Eqs. (31) and (29), we therefore show that $(l^{(t)})^2$ becomes larger due to cross-talk noise from other outputs, and so the generalization error worsens as output number M increases. Moreover, the convergence condition of NP-learning of $0 < \eta < 2/(M+2)$ is given by Eq. (31). This condition can be rewritten as

$$M < \frac{2(1-\eta)}{\eta}. \quad (32)$$

Consequently, we calculate the asymptotic value of the generalization error by substituting $t \to \infty$ into Eq. (31).

$$\varepsilon_g^{(\infty)} = \frac{\eta\sigma^2 M(M+2)(M+4)}{(8(2-(M+2)\eta)}. \quad (33)$$

The asymptotic value of the generalization error is vanished in the limit $\eta \to 0$.

Next, we compare the analytical results with those of computer simulation to examine the validity of analysis. We show results for the student length $(l^{(t)})^2$ (**Fig. 2**), the overlap $r^{(t)} = R^{(t)}l^{(t)}$ (**Fig. 3**) and the generalization error $\varepsilon_g^{(t)}$ (**Fig. 4**). We set the output unit number $M = 2$, 5, or 8; set the standard deviation $\sigma = 0.2$; and set the learning rate $\eta = 0.1$. The results were obtained

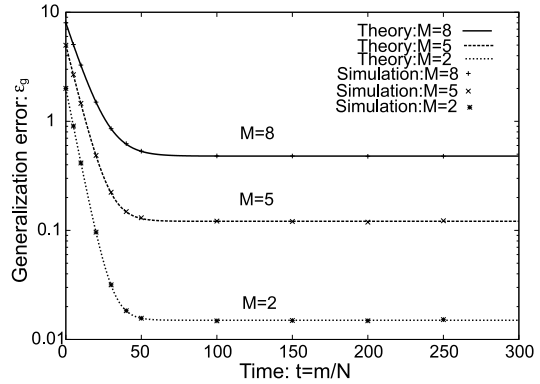

Fig. 4 Learning time dependence of the generalization error.

through computer simulation with $N = 1{,}000$ and each point obtained through averaging over 50 trials. Each element of the teacher and initial student weight vectors was independently drawn from the distribution of $N(0,1)$, and the element of input is drawn from $N(0,1/N)$. In the figures, the theoretical results for $M = 2$ are shown by a dotted line, the results for $M = 5$ are shown by a dashed line, and the results for $M = 8$ are shown by a solid line. The simulation results for $M = 2$ are represented by "$*$", those for $M = 5$ by "$\times$", and those for $M = 8$ by "+". The horizontal axis of each figure is normalized time $t = m/N$, where

m is the number of learning iterations.

Figure 2 shows the learning time dependence for student length l^2. The vertical axis is the student length. As shown, the analytical results agreed with those of the computer simulations, confirming the validity of the theoretical assumptions. The student length becomes longer as the output number M increases, and this result supports the analytical observations.

Figure 3 shows the time dependence of overlap between the teacher and student $r = Rl$. The vertical axis is the overlap r. As shown, the analytical results again agreed with those of the computer simulations. Moreover, the overlap between teacher and student does not depend on output number M, and this supports the analytical observations.

Figure 4 shows time dependence of the generalization error. The vertical axis is the generalization error ε_g. As shown, analytical results agreed with those of the computer simulations. The generalization error becomes larger as the output number M increases, and this phenomenon is due to the M dependence of l^2. This result supports the analytical observations. The above results show the validity of the analytical solutions.

5. Discussion

In the present paper, we derived the generalization error by using a statistical mechanics method. On the other hand, Werfel, et al. [3] derived the discrete evolution equation of the squared error as a learning curve by averaging the squared error over the possible inputs. These two methods use different techniques, so it is useful to clarify the relation between the methods.

For comparison, we re-formulate Werfel's formulation of the norm of the input $||\boldsymbol{x}||$ as $O(1)$, while it was originally formulated as $O(\sqrt{N})$, and re-derive Werfel's learning curve by following their analysis. (The derivation is shown in the Appendix.)

$$\left\langle E^{(m)} \right\rangle = \frac{\eta\sigma^2(M+2)(M+4)M}{8\left(2-(M+2)\eta\right)} + \left(E^{(0)} - \frac{\eta\sigma^2(M+2)(M+4)M}{8\left(2-(M+2)\eta\right)} \right) e^{-\eta(2-\eta(N+2))\frac{m}{N}}. \quad (34)$$

If we substitute $t = m/N$ and $E^{(0)} = \varepsilon_g^{(0)}$, Eq. (34) becomes identical to Eq. (31).

Next, we compare results obtained through the statistical mechanics method with those of Werfel, et al. [3]. From the statistical mechanics results, Eq. (29) shows that cross-talk noise is the cause of the longer student length as output number M increases. Equation (28) shows that the overlap r remains constant when output number M changes. Therefore, as we have shown, when the output number M becomes larger, the cross-talk noise from other outputs will increase the student vector length and the generalization error will become larger. Our result describes the deterministic behavior of the system while Werfel's result is just an average of the squared error. Moreover, statistical mechanical analysis can treat the case of nonlinear output functions through statistical mechanical methods while Werfel's analysis method cannot. On the other hand, when the input number is finite, we cannot analyze the system because we assume an infinite number N in the statistical mechanical method. The previous study [3] also showed that the generalization error becomes larger for a larger output number M, and their approach enables analysis of the case of finite input number N. However, it cannot show the cause of the larger generalization error for a larger output number M. The previous studies cannot analyze the case using nonlinear output functions.

6. Conclusion

We have analyzed node perturbation learning (NP-learning) using a statistical mechanical method within the framework of on-line learning. NP-learning is a kind of stochastic gradient method and it can be widely used in machine learning. We formulated NP-learning by using the teacher-student formulation, and we assumed the thermodynamic limit of $N \to \infty$. We showed that NP-learning can be formulated as noisy perceptron learning, and then derived the differential equations of order parameters that depict the learning process. The order parameters of NP-learning are the length of the student weight vector l_k and the overlap between teacher and student R_k. We derived these differential equations using a statistical mechanics method and solved them analytically. We then derived dynamics of the generalization error using these order parameters. Consequently, we showed that when the output number M becomes larger, the cross-talk noise from other outputs increases the student vector length and the

generalization error becomes larger. We also showed that the effect of cross-talk noise for larger number of outputs can be canceled out by decreasing the learning rate η. Our future work will include the analysis of NP-learning with a non-linear output function.

References

1) Widrow, B. and Lehr, M.A.: 30 years of adaptive neural networks: Perceptron, Madaline, and Backpropagation, *Proc. IEEE*, Vol.78, No.9, pp.1415–1442 (1990).
2) Williams, R.J.: Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning*, Vol.8, pp.229–256 (1992).
3) Werfel, J., Xie, X. and Seung, H.S.: Learning Curves for Stochastic Gradient Descent in Linear Feedforward Networks, *Neural Computation*, Vol.17, pp.2699–2718 (2005).
4) Sprekeler, H., Hennequin, G. and Gerstner, W.: Code-specific policy gradient rules for spiking neurons, *Proc. Neural Information Processing Systems 2010*, The MIT press, pp.19–28 (2010).
5) Fiete, I.R. and Seung, H.S.: Gradient Learning in Spiking Neural Networks by Dynamic Perturbation of Conductances, *Physical Review Letters*, Vol.97, p.048104 (2006).
6) Fiete, I.R., Fee, M.S. and Seung, H.S.: Model of Birdsong Learning Based on Gradient Estimation by Dynamic Perturbation of Neural Conductances, *Journal of Neurophysiology*, Vol.98, pp.2038–2057 (2007).
7) Reents, G. and Urbanczik, R.: Self-Averaging and On-line Learning, *Physical Review Letters*, Vol.80, pp.5445–5448 (1998).
8) Biehl, M. and Riegler, P.: On-Line Learning with a Perceptron, *Europhysics Letters*, Vol.28, No.7, pp.525–530 (1994).
9) Engel, A. and den Broeck, C.V.: *Statistical Mechanics of Learning*, Cambridge University Press, Cambridge, UK, 1st edition (2001).
10) Hara, K. and Okada, M.: Ensemble Learning of Linear Perceptrons: On-Line Learning Theory, *Journal of the Physical Society of Japan*, Vol.74, No.11, pp.2966–2972 (2005).
11) Nishimori, H.: *Spin Glass Theory and Statistical Mechanics of Information*, Oxford University Press, Oxford, UK (2001).
12) Saad, D.: *On-Line Learning in Neural Networks*, Cambridge University Press, Cambridge, UK (1992).
13) Krogh, A.: Learning with noise in a linear perceptron, *Journal of Physics A: Mathematical and General*, Vol.25, No.5, pp.1119–1133 (1992).
14) Krogh, A. and Hertz, J.A.: Generalization on a linear perceptron in the presence of noise, *Journal of Physics A: Mathematical and General*, Vol.25, No.5, pp.1135–1147 (1992).

Appendix

A.1 Analysis of Linear Perceptron Learning with Noise

In linear perceptron learning with noise, noise $\xi^{(m)}$ is added on the teacher output or student output in the learning process. The learning equation is

$$\boldsymbol{w}^{(m+1)} = \boldsymbol{w}^{(m)} + \eta(d^{(m)} - y^{(m)} - \xi^{(m)})\boldsymbol{x}^{(m)} = \boldsymbol{w}^{(m)} + \eta f^{(m)}\boldsymbol{x}^{(m)}. \quad (35)$$

Here, the noise ξ^m is drawn from $N(0, \sigma^2)$.

First, we derive the differential equation of student length l. We can rewrite Eq. (6) as $\boldsymbol{w} \cdot \boldsymbol{w} = Nl_k^2$, and we square both sides of Eq. (35).

$$N(l^{(m+1)})^2 = N(l^{(m)})^2 + 2\eta f^{(m)} y^{(m)} + \eta^2 (f^{(m)})^2. \quad (36)$$

Note that $\boldsymbol{x}$ and $\boldsymbol{w}$ are random variables, so the equation becomes a random recurrence formula. We formulate the size of input $||\boldsymbol{x}||$ as $O(1)$ and the size of student weight vector $||\boldsymbol{w}||$ as $O(\sqrt{N})$, so the length of the weight vector has a self-averaging property. Equation (36) is then rewritten as

$$N(l^{(m+1)})^2 = N(l^{(m)})^2 + 2\eta \langle fy \rangle + \eta^2 \langle f^2 \rangle . \quad (37)$$

Here, $\langle \cdot \rangle$ denotes the average over possible inputs. Next, we rewrite m as $m = Nt$, and represent the learning process using continuous time t in the thermodynamic limit of $N \to \infty$. At the limit, Eq. (37) becomes a differential equation, so we put $l^{(m)} \to l$, $l^{(m+1)} \to l + dl_k$, $1/N \to dt$, and then obtain the deterministic differential equation of l_k:

$$\begin{aligned} N(l + dl)^2 &= Nl^2 + 2\eta \langle fy \rangle + \eta^2 \langle f^2 \rangle , \\ \frac{dl^2}{dt} &= 2\eta \langle fy \rangle + \eta^2 \langle f^2 \rangle . \end{aligned} \quad (38)$$

Here, time t is omitted from functions l, y, and f for the sake of simplicity.

Next, we derive the differential equation of R that is the second order parameter of the system. R is the direction cosine (overlap) between teacher weight vector $\boldsymbol{w}^*$ and student weigh vector $\boldsymbol{w}$ defined as

$$R \equiv \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}}{||\boldsymbol{w}^*||||\boldsymbol{w}||} = \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}}{Nl}. \quad (39)$$

The differential equation of overlap R is derived by calculating the product of $\boldsymbol{w}^*$

and Eq. (35), and we then obtain the term of the equation using the distribution of $P(d, y)$. We then get

$$Nr^{(m+1)} = Nr^{(m)} + \eta \langle fd \rangle . \tag{40}$$

Here, we write overlap between the teacher weight vector and the student weight vector as r and $r = Rl$ for the sake of convenience. The overlap R also has a self-averaging in the thermodynamic limit, and the deterministic differential equation of r is then obtained through a calculation similar to that used for l.

$$\frac{dr}{dt} = \eta \langle fd \rangle . \tag{41}$$

Here, time t is omitted from functions r, d, and f for the sake of simplicity. $\langle \cdot \rangle$ denote the average over the possible inputs.

The three averages in Eqs. (38) and (41) are calculated as

$$\langle f \cdot y \rangle = \langle (d - y - \xi) y \rangle = \langle dy \rangle - \left\langle (y)^2 \right\rangle - \langle \xi y \rangle = r - l^2, \tag{42}$$

$$\begin{aligned} \left\langle f^2 \right\rangle &= \left\langle (d - y - \xi)^2 \right\rangle \\ &= \left\langle (d)^2 \right\rangle + \left\langle (y)^2 \right\rangle + \left\langle (\xi)^2 \right\rangle - 2 \langle dy \rangle + 2 \langle y\xi \rangle - 2 \langle d\xi \rangle \\ &= 1 - 2r + l^2 + \sigma^2, \end{aligned} \tag{43}$$

$$\langle f \cdot d \rangle = \langle (d - y - \xi) d \rangle = \left\langle (d)^2 \right\rangle - \langle yd \rangle - \langle \xi d \rangle = 1 - r. \tag{44}$$

As a result, differential equations of the order parameters are give as

$$\frac{dl}{dt} = 2\eta(r - l^2) + \eta^2(1 - 2r + l^2 + \sigma^2), \tag{45}$$

$$\frac{dr}{dt} = \eta(1 - r). \tag{46}$$

A.2 Derivation of Learning Curve of NP-learning

To compare statistical mechanics results with that of Werfel's, we change the formulation of norm of input $||\boldsymbol{x}||$ from $O(\sqrt{N})$ to $O(1)$, and derive the learning curve of NP-learning by following the derivation of Werfel, et al. [3]. We rewrite Eq. (7) as

$$E = \frac{1}{2}||\boldsymbol{d} - \boldsymbol{y}||^2 = \frac{1}{2}||(\boldsymbol{w}^* - \boldsymbol{w}) \cdot \boldsymbol{x}||^2 = \frac{1}{2}||\boldsymbol{W} \cdot \boldsymbol{x}||^2. \tag{47}$$

The learning curve shows the behavior of the squared error between the teacher outputs and those of the students. We calculate an ensemble of averages of the squared error $\langle E \rangle$ by expanding Eq. (7).

$$\begin{aligned} \left\langle E^{(m)} \right\rangle &= \frac{1}{2} \left\langle ||\boldsymbol{W}^{(m)} \cdot \boldsymbol{x}^{(m)}||^2 \right\rangle = \frac{1}{2} \left\langle \sum_{i=1}^{N} \left(\sum_{k=1}^{M} W_{ik}^{(m)} x_k^{(m)} \right)^2 \right\rangle \\ &= \frac{1}{2} \sum_{i=1}^{N} \left[\left\langle \sum_{k=1}^{M} W_{ik}^{(m)} x_k^m \sum_{n \neq k} W_{in}^{(m)} x_n^{(m)} \right\rangle + \left\langle \sum_{k=1}^{M} \left(W_{ik}^{(m)} \right)^2 \left(x_k^{(m)} \right)^2 \right\rangle \right] \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \left\langle \left(W_{ik}^{(m)} \right)^2 \right\rangle \frac{1}{N}. \end{aligned} \tag{48}$$

Here, we assume that an input element x_i is drawn from a probabilistic distribution with zero mean and $\frac{1}{N}$ variance. To calculate Eq. (48), we need to obtain the time evolution of the ensemble average of the square of student weight vector $\left\langle \left(W_{ik}^{(m)} \right)^2 \right\rangle$. The square of the weight vector of one-step update $W_{ik}^{(1)}$ is given by Eq. (9) as follows;

$$\left(W_{ij}^{(1)} \right)^2 = \left(W_{ij}^{(0)} + \Delta W_{ij}^{(0)} \right)^2 = \left(W_{ij}^{(0)} - \frac{\eta}{\sigma^2} \left(E_{NP}^{(0)} - E^{(0)} \right) \xi_i^{(0)} x_j^{(0)} \right)^2 . \tag{49}$$

By substituting Eqs. (7) and (8) into Eq. (49) and then averaging the term over the possible input, we get

$$\begin{aligned} \left\langle \left(W_{ij}^{(1)} \right)^2 \right\rangle = \left\langle \left(W_{ij}^{(0)} \right)^2 \right\rangle \left(1 - \frac{2\eta}{N} \right) + \frac{\eta^2}{N^2} \left\langle \left(W_{ij}^{(0)} \right)^2 \right\rangle (MN + 2N + 2M + 4) \\ + \frac{\eta^2 \sigma^2}{4N} (M + 2)(M + 4). \end{aligned} \tag{50}$$

Here, we have used $\left\langle x_i^2 \right\rangle = 1/N$, $\left\langle \xi^2 \right\rangle = \sigma^2$, $\left\langle \xi^4 \right\rangle = 3\sigma^4$, $\left\langle \xi^6 \right\rangle = 15\sigma^6$, $\left\langle \xi^3 \right\rangle = \left\langle \xi^5 \right\rangle = 0$. Applying Eq. (50) for the m-th iteration, summing up for i and j, and then substituting them into Eq. (48), we get the learning equation as

$$\begin{aligned} \left\langle E^{(t)} \right\rangle = \frac{\eta \sigma^2 (M + 2)(M + 4) M}{8 \left(2 - (M + 2)\eta \right)} \\ + \left(E^{(0)} - \frac{\eta \sigma^2 (M + 2)(M + 4) M}{8 \left(2 - (M + 2)\eta \right)} \right) e^{-\eta(2 - \eta(N+2))t}. \end{aligned} \tag{51}$$

Here, time t is defined as $t = m/N$, and the ensemble average of the squared error at t is $\left\langle E^{(t)} \right\rangle$, and that of the squared error at $t = 0$ is $E^{(0)}$.

(Received August 27, 2010)
(Revised October 9, 2010)
(Accepted October 22, 2010)

Kazuyuki Hara received his B.Eng. and M.Eng. degrees from Nihon University in 1979 and 1981 respectively and Ph.D. degree from Kanazawa University in 1997. He was involved in NEC Home Electronics Corporation from 1981 until 1987. He joined to Toyama Polytechnic College in 1987 where he was a lecturer. He joined Tokyo Metropolitan College of Technology in 1998 where he was an associate professor and became a professor in 2005. He became a professor at Nihon University in 2010. His current research interests include statistical mechanics of on-line learning.

Kentaro Katahira received his B.S. degree from Chiba University in 2002 and M.S. and Ph.D. degrees from The University of Tokyo in 2004, 2009, respectively. From 2004 to 2005, he worked at Yamaha Corporation. Currently, he is an researcher of Japan Science Technology Agency, ERATO, OKANOYA Emotional Information Project. His research interests include statistical data modeling, decision making and statistical learning.

Kazuo Okanoya received the B.S. degree from Keio University in 1983. He received his M.S. and Ph.D. degrees from University of Maryland in 1986 and 1989, respectively. From 1989 to 1990, he was a post-doctoral fellow of JSPS. From 1990 to 1993, he was a post-doctoral fellow of JST. From 1994 to 2005, he was an associate professor at Department of Cognitive and Information Sciences, Faculty of Letters, Chiba University. Currently, he is a laboratory head of laboratory for Biolinguistics, RIKEN Brain Science Institute, a research director of OKANOYA Emotional Information Project, JST ERATO, and a professor of Graduate School of Arts and Sciences, The University of Tokyo. His research interest is in biological inquiry into the origin of language and emotion.

Masato Okada received his B.Sc. degree in physics from Osaka City University in 1985, M.Sc. degree in physics and Ph.D. degree in science from Osaka University, Osaka, Japan, in 1987, and 1997, respectively. From 1987 to 1989, he worked at Mitsubishi Electric Corporation, and from 1991 to 1996, he was a research associate at Osaka University. He was a researcher on the Kawato Dynamic Brain Project until 2001. He was a deputy head of the Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama, Japan, and a PRESTO researcher on intelligent cooperation and control at the Japan Science and Technology Agency until 2004. Currently, he is a professor in the Graduate School of Frontier Science, The University of Tokyo. His research interests include the computational aspects of neural networks and statistical mechanics for information processing.