DE GRUYTER OPEN

JAISCR, 2015, Vol. 5, No. 1, pp. 71 – 77
10.1515/jaiscr-2015-0020

# MUTUAL LEARNING USING NONLINEAR PERCEPTRON

Daisuke Saitoh[1], Kazuyuki Hara[2]

[1]*Graduate School of Industrial Technology, Nihon University*
*1-2-1 Izumi-cho, Narashino, Chiba, Japan 275-8575*
*e-mail: cida14004@g.nihon-u.ac.jp*

[2]*College of Industrial Technology, Nihon University*
*1-2-1 Izumi-cho, Narashino, Chiba, Japan 275-8575*
*e-mail: hara.kazuyuki@nihon-u.ac.jp*

**Abstract**

We propose a mutual learning method using nonlinear perceptron within the framework of online learning and have analyzed its validity using computer simulations. Mutual learning involving three or more students is fundamentally different from the two-student case with regard to variety when selecting a student to act as the teacher. The proposed method consists of two learning steps: first, multiple students learn independently from a teacher, and second, the students learn from others through mutual learning. Results showed that the mean squared error could be improved even if the teacher had not taken part in the mutual learning.

## 1 Introduction

Kinzel proposed mutual learning [1]-[3] within the framework of online learning [4]-[6] as a model to explore interactions between students. His model employs two students, and a student learns with the other student acting as a teacher. The target of his model is to obtain the same networks through the learning as the common key of cryptography.

In terms of the learning problem, how the student approaches the teacher is important. In our previous work [7], we showed that the generalization error of the students becomes smaller through mutual learning, even if the teacher does not take part in the mutual learning. In that work, we used linear perceptron. However, nonlinear perceptron has several advantages over linear perceptron, such as the ability to use nonlinear outputs, learnability, storage capacity, and so forth. Thus, the learning behavior of a nonlinear perceptron is of interest.

In the current work, we explore mutual learning for nonlinear perceptron. The learning settings are formulated similar to those of statistical mechanics because we intend to construct a theory for the proposed method to utilize in future research. We demonstrate the validity of the proposed method by computer simulations.

## 2 Formulations

In this work, we employ a teacher-student formulation and assume the existence of a teacher network that produces the desired output for the student network. By introducing the teacher network, we can directly measure the similarity of the student weight vector against that of the teacher. Here, first we formulate a teacher network and a student network and then we introduce the gradient descent algorithm.

The teacher is nonlinear perceptron with N input units and an output and student networks are $n_s$ nonlinear perceptron, as shown in Figure 2. Here, $n_s$ denotes the number of students. We assume the teacher and student networks receive $N$-dimensional input $x^{(m)} = (x_1^{(m)}, \ldots, x_N^{(m)})$ at the $m$-th learning iteration, as shown in Figure 2. $x_i^{(m)}$ of the independently drawn input $x^{(m)}$ are uncorrelated random variables with zero mean and $1/N$ variance. $f$ in Figure 2 is the output function defined by

$$f(\xi) = \mathrm{erf}\left(\frac{\xi}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\xi}^{\xi} e^{-\frac{t^2}{2}} dt. \quad (1)$$

The output function is similar to sigmoid function. Teacher output $v^{(m)}$ is calculated using $v^{(m)} = f(\sum_{j=1}^{N} B_j x_j^{(m)})$ and that of student $u_i^{(m)}$ is calculated using $u_i^{(m)} = f(\sum_{j=1}^{N} J_{ij}^{(m)} x_j^{(m)})$. Here, $B = (B_1, \ldots B_N)$ is the teacher weight vector and $J_i^{(m)} = (J_{i1}^{(m)}, \ldots, J_{iN}^{(m)})$ are the weight vector of the $i$th student, where $m$ denotes learning iterations. Learning iteration $m$ is ignored in the figures. Each element $B_j,\ j = 1 \sim N$ is drawn from a probability distribution with zero mean and unit variance. We also assume that each element of $J_{ij}^{(0)}$, which is the initial value of the student vector $J_i^{(0)}$, is drawn from a probability distribution with zero mean and unit variance.

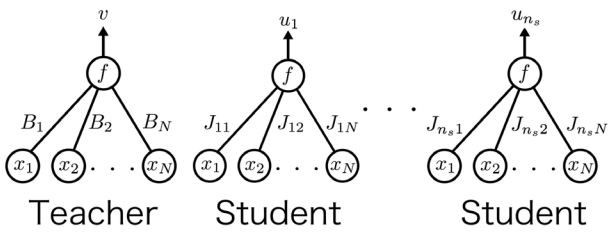



**Figure 1**. Architecture of teacher and student networks.

Next, we formulate the squared error. The squared error for the $i$-th student $E_i$ is given by

$$E_i^{(m)} = \frac{1}{2}\left(v^{(m)} - u_i^{(m)}\right)^2. \quad (2)$$

In the computer simulation, we use the mean squared error $E_{Mi}^{(m)}$ that calculated by average of $E_i^{(m)}$ over $K$ inputs.

Next, we formulate the overlap. The overlap is the direction cosine between the teacher weight vector and the $i$th student weight vector given by

$$R_i = \frac{B \cdot J_i}{\|B\|\|J_i\|}. \quad (3)$$

Next, we formulate the similarity between the teacher weight vector and the student weight vector. The similarity is a short form of the similarity between the teacher weight vector and the student weight vector[8]. The similarity is given by

$$\|B - J_i\|^2 = \sum_{j=1}^{N} (B_j - J_{ij})^2. \quad (4)$$

Here, the similarity becomes zero when both the direction cosine and the norm of the student weight vector are unity at the same time.

# 3 Proposed method

In this section, we formulate a mutual learning algorithm for nonlinear perceptron. This algorithm is composed of two parts: initial learning and mutual learning. In the initial learning, a student learns from the teacher, and we use the gradient descent algorithm to modify each student's weight vector $J_i^{(m)}$:

$$J_i^{(m+1)} = J_i^{(m)} + \eta\left(v^{(m)} - u_i^{(m)}\right)x^{(m)}. \quad (5)$$

An identical input $x^{(m)}$ is applied to all students in the same order. Equation (5) shows that the initial learning is carried out between the teacher and one of the students. All the students independently learn the relationship between the input $x^{(m)}$ and the target $v^{(m)}$ given in every iteration. Therefore, after the initial learning, there is some correlation between the teacher and students. This means that the students have the portion of the information which the teacher has. When the mean squared error reaches $E_{Mi}^{(m)} = E_R$, we switch the initial learning to the mutual learning. The learning equation of the mutual learning is

$$J_i^{(m+1)} = J_i^{(m)} + \eta\left(u_{i'(m)} - u_i^{(m)}\right)x^{(m)}. \quad (6)$$

Here, subscript $i$ denotes a student and subscript $i'(m)$ denotes a student acting as a teacher at the $m$-th iteration.

# 4 Results

In this section, we discuss the behavior of the initial learning and mutual learning through the computer simulations. The element of teacher weight vector $B_i$ is sampled from $\mathcal{N}(0,1)$, that of student $J_i^{(0)}$ is sampled from $\mathcal{N}(0,1)$, and that of input $x_i^{(m)}$ is sampled from $\mathcal{N}(0,1/N)$, as described in Sec. 2. The learning step size is set to $\eta = 0.1$, the input dimension is set to $N = 1000$, and the results are averages obtained over 10 trials.

## 4.1 Initial learning

The results of the initial learning are presented in this subsection. In the initial learning, equation (5) is used as the learning equation. The results of student 1 are used. In Figs. 2, 3, and 4, the horizontal axis is time $t = m/N$, where $m$ is the number of learning iterations. The vertical axis of Figure 2 is the overlap $R_1$, that of Figure 3 is the mean squared error $E_{M1}$, and that of Figure 4 is the similarity $\|B - J_1\|^2$. In these figures, the $m$ of $B$ and $J_1$ has been omitted for simplicity. We set $E_R = 0.001$, and the initial learning stopped when $E_{M1} = 0.001$. As shown in Figure 2, $R_1$ started from $R_1^{(0)} = 0$, increased along $t$, and was eventually almost 1 beyond $t = 75$. This demonstrates that when we use a smaller $E_R$, the angle between the student weight vector and that of the teacher will be smaller. The time course of $E_{M1}$ and $\|B - J_1\|^2$ are shown in Figs. 3 and 4. Both exhibit similar curves and approach zero beyond $t = 75$. Note that the meanings of $\|B - J_1\|^2 \sim 0$ and $E_{M1} \sim 0$ are different: the similarity directly measures the difference between the teacher weight vector and the student weight vector while the mean squared error measures the output difference of between the teacher weight vector and that of the student. Therefore, from Figure 4, it is shown that by using a smaller $E_R$, the student weight vector becomes more similar to the teacher weight vector. Note that the similarity can be measured when we use the teacher-student formulation.

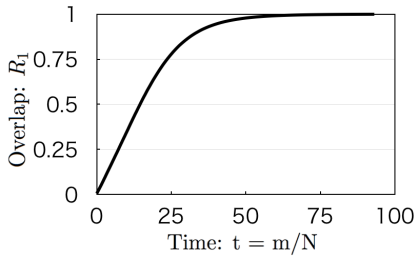



**Figure 2**. Time course of direction cosine of teacher weight vector and student weight vector $R_1$.

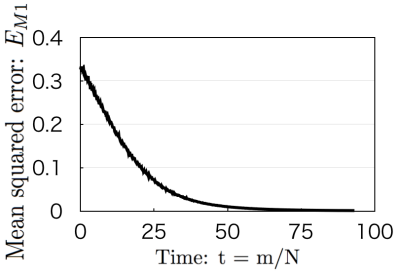



**Figure 3**. Time course of mean squared error $E_{M1}$ .

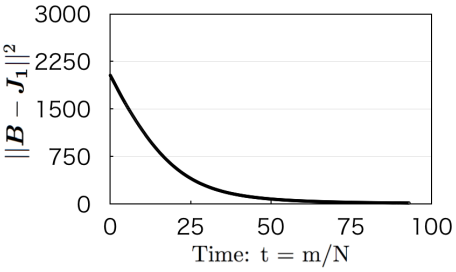



**Figure 4**. Time course of similarity $\|B - J_1\|^2$ .

## 4.2 Mutual learning with two students

The results of mutual learning with two students are presented in this subsection. In mutual learning, equation (6) is used as the learning equation. The initial learning is done beforehand with mutual learning. After the initial learning, the two students achieved $E_{M1} = E_{M2}$. In mutual learning, students take turns acting as a teacher.

Computer simulation results of mutual learning with two students are shown in Figs. 5, 6, and 7, where the horizontal axis is time $t = m/N$. The vertical axis of Figure 5 is the overlap $R_1$, that of Figure 6 is the mean squared error $E_{M1}$, and that of Figure 7 is the similarity $\|B - J_1\|^2$. In our computer simulations, mutual learning stopped at $t = 749$. We identified the stopping times at which students had the

best performances. Results obtained are the averages over 10 trials. In Figs. $5-7$, results are given using the teacher network and student network 1.

Figure 5 shows the dependence of overlap $R_1$ according to the switching error $E_R$. Overlap at $t = 749$ depends on $E_R$. As shown in the figure, a smaller $E_R$ resulted in a larger $R_1$ through mutual learning. Note that $R_1$ is improved even with a small switching error, e.g., $E_R = 0.001$. Moreover, the overlap was enlarged through mutual learning only, without using the teacher network. The dependence of mean squared error $E_{M1}$ according to the switching error $E_R$, shown in Figure 6, indicates that the mean squared error $E_{M1}$ can be reduced by mutual learning, and that this tendency depends on the switching error $E_R$. From these results, as with the overlap, the mean squared error was reduced through mutual learning without the teacher network. Figure 7 shows the dependence of the similarity $\|B-J_1\|^2$ according to the switching error $E_R$. The tendency of the similarity is similar to that of the mean squared error. This demonstrates that mutual learning improved the similarity and mean squared error at the same time without the teacher network. This demonstrates that mutual learning is effective.

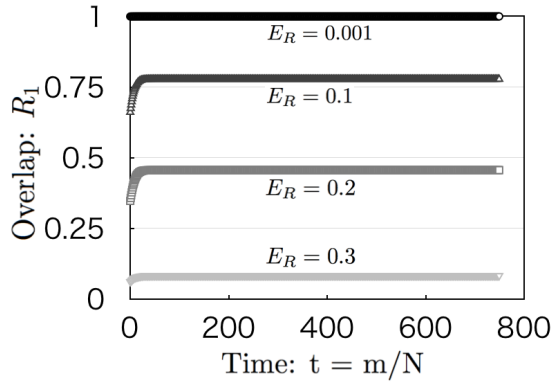



**Figure 5**. Dependence of overlap $R_1$ according to switching error $E_R$.

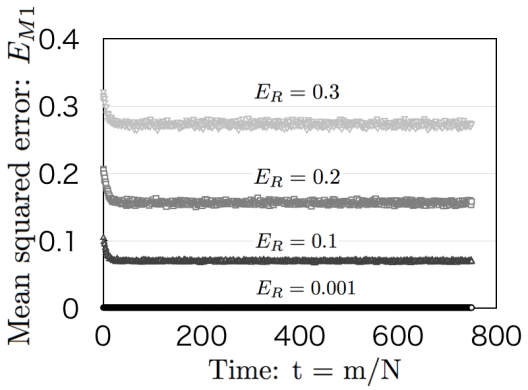



**Figure 6**. Dependence of mean squared error $E_{M1}$ according to switching error $E_R$ .

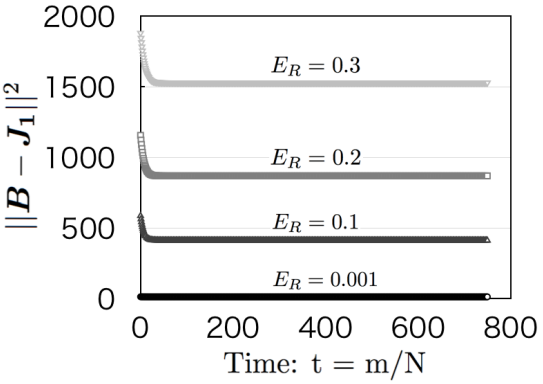



**Figure 7**. Dependence of similarity $\|B-J_1\|^2$ according to switching error $E_R$ .

Next, we discuss the effect of mutual learning in more detail. Figures 8, 9, and 10 show the results of analysis of Figs. 5, 6, and 7, where the horizontal axis is the switching error $E_R$. The vertical axis of Figure 8 is the ratio of $R_1^*/R^{(0)}$. Here, $R_1^*$ means $R_1$ at $t = 749$. The vertical axis of Figure 9 is the ratio of $E_{M1}^*/E_{M1}^{(0)}$ and that of Figure 10 is the ratio of $\|B-J_1^*\|^2/\|B-J_1^{(0)}\|^2$. Here, $E_{M1}^*$ means $E_{M1}$ at $t = 749$ and $J_1^*$ means $J_1$ at $t = 749$. Figure 8 shows the dependence of the ratio $R_1^*/R_1^{(0)}$ according to the switching error $E_R$. As shown in the figure, the ratio $R_1^*/R_1^{(0)}$ is proportional to $E_R$ and the biggest improvement of the ratio $R_1^*/R_1^{(0)} \sim 1.38$ is achieved when $E_R = 0.3$. Note that the ratio is not improved when $E_R = 0.001$. Figure 9 shows the dependence of the ratio $E_{M1}^*/E_{M1}^{(0)}$ according to the switching error $E_R$. In this case, the biggest improvement of ratio $E_{M1}^*/E_{M1}^{(0)} \sim 0.7$ is achieved when $E_R = 0.1$, and the ratio is proportional to $E_R$

except for $E_R = 0.001$. Note that the ratio $E^*_{M1}/E^{(0)}_{M1}$ is improved when $E_R = 0.001$. These results differ from those of ratio $R^*_1/R^{(0)}_1$. Figure 10 shows the dependence of the ratio $\|B-J^*_1\|^2/\|B-J^{(0)}_1\|^2$ according to the switching error $E_R$. The tendency of ratio $\|B-J^*_1\|^2/\|B-J^{(0)}_1\|^2$ is similar to that of ratio $E^*_{M1}/E^{(0)}_{M1}$. Note that the ratio is also improved when $E_R = 0.001$. These results also differ from those of ratio $R^*_1/R^{(0)}_1$. This mean that the direction cosine between the teacher weight vector and the student weight vector is not changed, although the norm of the student vector may approach that of the teacher vector. Next, we discuss the difference between $R_1$, $E_{M1}$, and $\|B-J_1\|^2$ based on their definitions. $R_1$ shows the direction cosines of the teacher and student weight vectors specifically, only the overlap tendency of the two vectors. However, we must consider both the direction cosine and the magnitude difference between the teacher and student weight vectors. For this purpose, we must use $\|B-J_1\|^2$, which measures the direction cosine and the magnitude difference at the same time. $E_{M1}$ is also a good measure, but $E_{M1}$ is calculated by using the output of the teacher network and the student networks, so it depends on the input $x^{(m)}$ and also on the output function $f(\cdot)$. Therefore, $\|B-J_1\|^2$ is the most useful in our case.


**Figure 8**. Dependence of ratio $R^*_1/R^{(0)}_1$ according to $E_R$.


**Figure 9**. Dependence of ratio $E^*_{M1}/E^{(0)}_{M1}$ according to $E_R$.

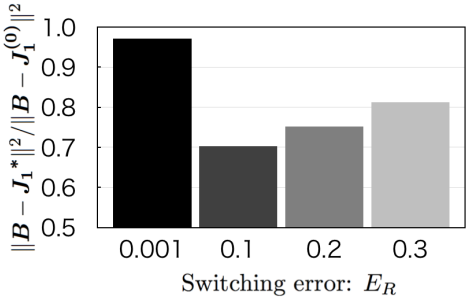



**Figure 10**. Dependence of ratio $\|B-J_1\|^2$ according to $E_R$.

### 4.3 Mutual learning with more than three students

Mutual learning involving three or more students is fundamentally different from the two-student case in terms of variety when selecting a student to act as teacher. Figures 11 and 12 show the results obtained through learning in a cyclic order of A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ A. Here, A learns B is referred to as A $\rightarrow$ B. In these figures, the horizontal axis is the number of students $n_s$. The vertical axis of Figure 11 (top) is the overlap at $t = 749$, $R_1{}^*$, that of Figure 11 (bottom) is the mean squared error at $t = 749$, $E^*_{M1}$, and that of Figure 12 is the similarity at $t = 749$, $\|B-J^*_1\|^2$. We used $n_s = 2$, 3, 5, and 10 and set $E_R = 0.001$, 0.1, 0.2, and 0.3. The symbols "$\bigcirc$", "$\triangle$", "$\square$", and "$\bigtriangledown$" show the results for $E_R = 0.001, 0.1, 0.2$, and 0.3, respectively. The results obtained are the averages over 10 trials. From Figs. 11 and 12, $R^*_1$ is proportional to the number of students for all $E_R$s and is particularly improved when $E_R = 0.2$, and $E^*_{M1}$ and $\|B-J^*_1\|^2$ are inversely proportional to the number of students $n_s$. Note that for all $E_R$s, $R^*_1$, $E^*_{M1}$, and $\|B-J^*_1\|^2$ are not saturated at $n_s = 10$. From our previous study [7], we know

that $R_1^*$ and $E_{M1}^*$ are saturated at $n_s = 10$ due to using the linear output function. Therefore, in our future work, we will investigate the issue of nonlinearity of the output function for mutual learning.

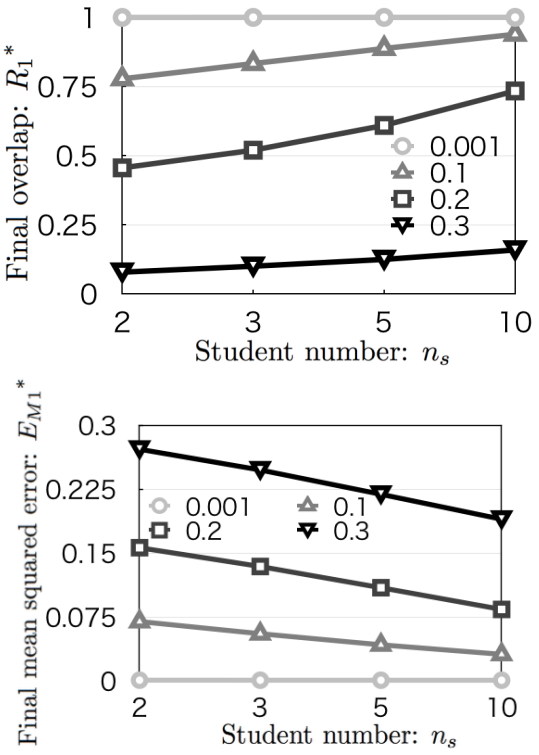



**Figure 11**. Dependence of $R_1{}^*$(top) and $E_{M1}{}^*$(bottom) according to number of students.

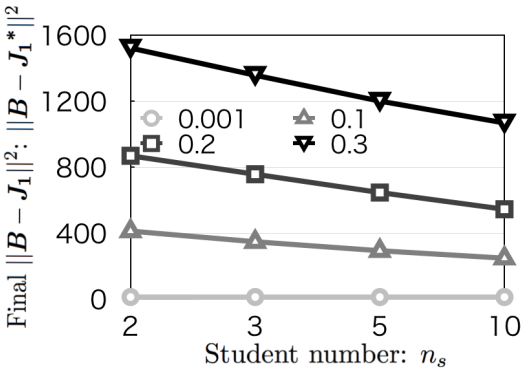



**Figure 12**. Dependence of $\|B - J_1^*\|^2$ according to number of students.

## 4.4 Trajectory of student weight vector through mutual learning

In the previous subsection, we showed that mutual learning can improve the performance of students without using a teacher. In this section, we clarify that the student weight vectors move toward the teacher weight vector through the mutual learning. We also clarify that the weight trajectories depending on the manner of selecting a student to act as teacher. For this purpose, we employ principle component analysis (PCA).

We obtained the trajectories of the student weight vectors by two steps: (1) obtain the first and second eigenvectors of matrix $H$, which consists of teacher weight vector $B$ and student weight vectors $J_i^{(m)}$ at every 100 iterations. Three students $J_1, J_2$, and $J_3$ are used.

$$H = \begin{bmatrix} B_1, & B_2, & \ldots, & B_N \\ J_{11}^{(0)}, & J_{12}^{(0)}, & \ldots, & J_{1N}^{(0)} \\ J_{11}^{(100)}, & J_{12}^{(100)}, & \ldots, & J_{1N}^{(100)} \\ \vdots & \vdots & \vdots & \vdots \\ J_{11}^{(M)}, & J_{12}^{(M)}, & \ldots, & J_{1N}^{(M)} \\ J_{21}^{(0)}, & J_{22}^{(0)}, & \ldots, & J_{2N}^{(0)} \\ J_{21}^{(100)}, & J_{22}^{(100)}, & \ldots, & J_{2N}^{(100)} \\ \vdots & \vdots & \vdots & \vdots \\ J_{21}^{(M)}, & J_{22}^{(M)}, & \ldots, & J_{2N}^{(M)} \\ J_{31}^{(0)}, & J_{32}^{(0)}, & \ldots, & J_{3N}^{(0)} \\ J_{31}^{(100)}, & J_{32}^{(100)}, & \ldots, & J_{3N}^{(100)} \\ \vdots & \vdots & \vdots & \vdots \\ J_{31}^{(M)}, & J_{32}^{(M)}, & \ldots, & J_{3N}^{(M)} \end{bmatrix}. \quad (7)$$

Here, $M$ is the number of iterations to stop the mutual learning. $J_i^{(0)}$ is the initial weight vector of $i$th student trained by mutual learning. (2) plots the trajectories of the student weight vectors at every 100 learning steps in the space spanned by the first and second eigenvectors. Figure 13 shows trajectories of the student weight vectors obtained during mutual learning involving three students respectively referred to as A, B, and C.

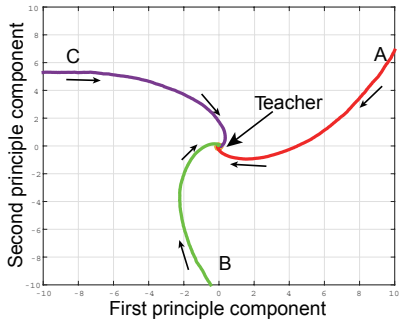
Figure 13(a) shows the results obtained during learning in cyclic order. The learning step size is $\eta = 0.1$ and the number of iterations to stop mutual learning is $M = 100 \times N$, where $N = 1000$. Initial value of overlap $R^{(0)}$ is set to 0.8. Figure 13(b) shows the results obtained through learning where one student is randomly selected to act as teacher for comparison. The symbol "o" at the center of each figure shows the weight vector of the teacher.

In these figures, the horizontal axis shows the first principle component and the vertical axis shows the second principle component.

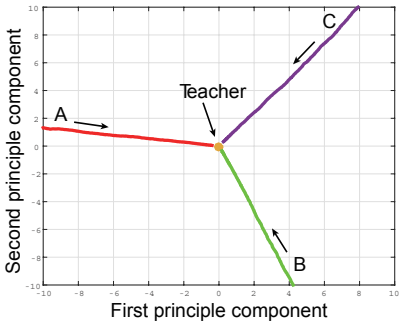
When the learning is cyclic, the trajectories of the students will head not toward the teacher but rather toward the student acting as teacher. As shown in Figure 13(a), the trajectory of A is heading toward B, not toward the teacher, in the early stage of learning. However, as the learning proceeds, all students head toward the teacher.

In random selected learning, one student is randomly selected to act as teacher. This student then learns the average of all students to act as teacher. As shown in Figure 13(b), the trajectories were all straight and heading toward the teacher weight vector.

From these results, we clarified that the student weight vectors head toward teacher weight vector through the mutual learning.




(a) Cyclic learning




(b) Random selected learning

**Figure 13**. Trajectory of student weight vector during mutual learning.

# 5 Conclusion

We have proposed a mutual learning method using nonlinear perceptron and demonstrated its validity through computer simulations. We have shown that the mutual learning improves the performance of student without a teacher. We also have shown that the performance achieved by mutual learning depends on the MSE of the initial learning. Moreover, the performance of students can be improved by using many students. In our future work, we intend to construct a theory of mutual learning through nonlinear perceptron and investigate the issue of nonlinearity of the output function for mutual learning.

The authors thank Professor Masato Okada for insightful discussions.

# References

[1] E. Klein, R. Mislovaty, I. Kanter, A. Ruttor, and W. Kinzel, " Synchronization of neural networks by mutual learning and its application to cryptography", *in Advance Neural Information Processing System*, **17** (MIT Press, Cambridge, MA, 2005).

[2] R. Metzler, W. Kinzel, and I. Kanter, "Interacting neural networks", Physical Review E **62**, 2, (2000) 2555.

[3] R. Mislovaty, E. Klein, I. Kanter, and W. Kinzel, "Public Channel Cryptography by Synchronization of Neural Networks and Chaotic Maps", Physical Review Letters, **91** (2003) 118701.

[4] D. Saad (Ed.), *Online Learning in Neural Networks*, (Cambridge University Press, Cambridge, UK., 1998).

[5] A. Krogh and P. Sollich, " Statistical mechanics of ensemble learning", Physical Review E, **55**, 1 (1997) 811.

[6] R. Beale and T. Jackson, *Neural Computing: An Introduction*, (Institution of Physic Publishing,1990).

[7] K. Hara, Y. Nakayama, S. Miyoshi, and M. Okada, "Statistical Mechanics of On-Line Mutual Learning with Many Linear Perceptrons", Journal of the Physical Society of Japan. **78** (2009) 114001.

[8] K. Hara, K. Katahira, K. Okanoya, and M. Okada, "Statistical Mechanics of Node-Perturbation Learning for Nonlinear Perceptron", Jounal of the Physical Society of Japan. **82** (2013) 054001.