Mutual Learning with Many Linear Perceptrons: On-Line Learning Theory

Kazuyuki Hara[1], Yoichi Nakayama[2], Seiji Miyoshi[3], and Masato Okada[4,5]

[1] Tokyo Metropolitan College of Industrial Technology, 1-10-40, Higashi-oi, Shinagawa Tokyo 140-0011, Japan
hara@s.metro-cit.ac.jp
[2] Tokyo Metropolitan College of Technology, 1-10-40, Higashi-oi, Shinagawa Tokyo 140-0011, Japan
[3] Faculty of Engineering Science, Kansai University, 3-3-35, Yamate-cho, Suita, Osaka, 564-8680, Japan
miyoshi@ipcku.kansai-u.ac.jp
[4] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba, 277-8561, Japan
okada@k.u-tokyo.ac.jp
[5] Brain Science Institute, Riken, 2-1, Hirosawa, Wako, Saitama, 351-0198, Japan

Abstract. We propose a new mutual learning using many weak learner (or student) which converges into the identical state of Bagging that is kind of ensemble learning, within the framework of on-line learning, and have analyzed its asymptotic property through the statistical mechanics method. Mutual learning involving more than three students is essential compares to two student case from a viewpoint of variety of selection of a student acting as teacher. The proposed model consists of two learning steps: many students independently learn from a teacher, and then the students learn from others through the mutual learning. In mutual learning, students learn from other students and the generalization error is improved even if the teacher has not taken part in the mutual learning. We demonstrate that the learning style of selecting a student to act as teacher randomly is superior to that of cyclic order by using principle component analysis.

1 Introduction

As a model incorporating the interaction between students, Kinzel proposed mutual learning within the framework of on-line learning[1,2]. Kinzel's model employs two students, and one student learns with the other student acting as a teacher. The target of his model is to obtain the identical networks through such learning. On the other hand, ensemble learning algorithms, such as bagging[3] and Ada-boost[4], try to improve upon the performance of a weak learning machine by using many weak learning machines; such learning algorithms have recently received considerable attention. We have noted, however, that the mechanism of integrating the outputs of many weak learners in ensemble learning is similar to that of obtaining the identical networks through mutual learning.

C. Alippi et al. (Eds.): ICANN 2009, Part I, LNCS 5768, pp. 171–180, 2009.

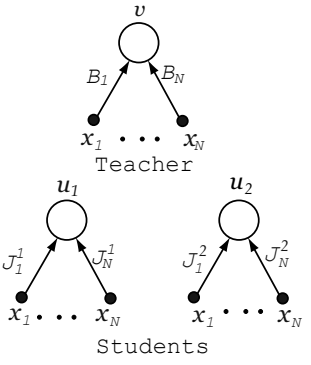


Fig. 1. Network structure of latent teacher and student networks, all having the same network structure

With regard to the learning problem, how the student approaches the teacher is important. However, Kinzel[1,2] does not deal with the teacher-student relation since a teacher is not employed in his model. In contrast to Kinzel's model, we have proposed mutual learning between two students who learn from a teacher in advance[5,6]. In our previous work[5,6], we showed that the generalization error of the students becomes smaller through the mutual learning even if the teacher does not take part in the mutual learning. Our previous work[5,6] treated a special case where the number of students is two. This paper treats the general case where the number of students is arbitrary. When the number of students becomes general, additional degrees of freedom associated with the selection of learning order are generated, and the problem settings become essentially different from the two students case. We formulate a new mutual learning algorithm, and then we analyze the asymptotic property of the proposed learning algorithm through statistical mechanics.

2 Formulation of Mutual Learning with a Latent Teacher

In this section, we formulate the latent teacher and student networks, and the mutual learning algorithms. We assume the latent 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. Learning iteration m is ignored in the figure.

The latent teacher network is a linear perceptron, and the student networks are K linear perceptrons. We also assume that the elements $x_i(m)$ of the independently drawn input $\boldsymbol{x}(m)$ are uncorrelated random variables with zero mean and $1/N$ variance; that is, the elements are drawn from a probability distribution $P(\boldsymbol{x})$. In this paper, the thermodynamic limit of $N \to \infty$ is assumed. Thermodynamic limit means that for the limit of system size N to be infinity, the law of large numbers and the central limit theorem are effected. We can then depict the system behavior by using a small number of parameters. At the limit, the size of input vector $\|\boldsymbol{x}\|$ then becomes one.

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

where $\langle \cdots \rangle$ denotes average, and $\| \cdot \|$ denotes the norm of a vector.

The latent teacher network is a linear perceptron, and is not subject to training. Thus, the weight vector is fixed in the learning process. The output of the latent teacher $v(m)$ for N-dimensional input $\boldsymbol{x}(m)$ at the m-th learning iteration is

$$v(m) = \sum_{i=1}^{N} B_i x_i(m) = \boldsymbol{B} \cdot \boldsymbol{x}(m), \tag{2}$$

$$\boldsymbol{B} = (B_1, B_2, \ldots, B_N), \tag{3}$$

where latent teacher weight vector $\boldsymbol{B}$ is an N-dimensional vector like the input vector, and each element B_i of the latent teacher weight vector $\boldsymbol{B}$ is drawn from a probability distribution of zero mean and unit variance. Assuming the thermodynamic limit of $N \to \infty$, the size of latent teacher weight vector $\|\boldsymbol{B}\|$ becomes $\sqrt{N}$.

$$\langle B_i \rangle = 0, \quad \langle (B_i)^2 \rangle = 1, \quad \|\boldsymbol{B}\| = \sqrt{N}. \tag{4}$$

The output distribution of the latent teacher $P(v)$ follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit of $N \to \infty$.

The K linear perceptrons are used as student networks that compose the mutual learning machine. Each student network has the same architecture as the latent teacher network. For the sake of analysis, we assume that each element of $\boldsymbol{J}^k(0)$ which is the initial value of the k-th student weight vector $\boldsymbol{J}^k$ is drawn from a probability distribution of zero mean and unit variance. The norm of the initial student weight vector $\|\boldsymbol{J}^k(0)\|$ is $\sqrt{N}$ in the thermodynamic limit of $N \to \infty$,

$$\langle J_i^k(0) \rangle = 0, \quad \langle (J_i^k(0))^2 \rangle = 1, \quad \|\boldsymbol{J}^k(0)\| = \sqrt{N}. \tag{5}$$

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

$$u_k(m) = \sum_{i=1}^{N} J_i^k(m) x_i(m) = \boldsymbol{J}^k(m) \cdot \boldsymbol{x}(m), \tag{6}$$

$$\boldsymbol{J}^k(m) = (J_1^k(m), J_2^k(m), \ldots, J_N^k(m)). \tag{7}$$

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

$$\|\boldsymbol{J}^k(m)\| = l_k(m)\sqrt{N}. \tag{8}$$

The distribution of the output of the k-th student $P(u_k)$ 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. After the students learn from a latent teacher, mutual learning is carried out[5]. The learning equation of the mutual learning is

$$\boldsymbol{J}^k(m+1) = \boldsymbol{J}^k(m) + \eta_k\Big(\ u_{k'} - u_k \Big)\boldsymbol{x}(m) \tag{9}$$

Here, k is a student and k' is a student to act as a teacher. m denotes the iteration number. We use the gradient descent algorithm in this paper, while another algorithm was used in Kinzel's work [1]. Equation (9) shows that mutual learning is carried out between two students. Therefore, the teacher used in the initial learning is called a latent teacher.

In the mutual learning, selection of a student to act as a teacher is important. In this paper, a student to act as a teacher is selected at random from all the students, then only the statistical effects is learned by a student and therefore a student tend to learn the average of all the students. Keeping this in mind, the learning equation is rewritten by the next equation.

$$\begin{aligned}\boldsymbol{J}^k(m+1) &= \boldsymbol{J}^k(m) + \eta_k\Big(\ \frac{1}{K}\sum_{i=1}^{K} u_i(m) - u_k(m) \ \Big)\boldsymbol{x}(m) \\ &= \boldsymbol{J}^k(m) + \eta_k(\overline{u}(m) - u_k(m))\boldsymbol{x}. \end{aligned} \tag{10}$$

Here, $\overline{u}$ is average of the student outputs and is to act as a teacher.

When the interaction between students is introduced, the performance of students may be improved if they exchange knowledge that each student has acquired from the latent teacher in the initial learning. In other words, two students approach each other through mutual learning, and tend to move towards the middle of the initial weight vectors. This tendency is similar to the integration mechanism of Bagging, so mutual learning may mimic this mechanism.

3 Theory

In this section, we first derive the differential equations of two order parameters which depict the behavior of mutual learning. After that, we derive an auxiliary order parameter which depicts the relationship between the teacher and students. We then rewrite the generalization error using these order parameters. We first derive the differential equation of the length of the student weight vector l_k. l_k is the first order parameter of the system. We modify the length of the student weight vector in Eq. (8) as $\boldsymbol{J}^k \cdot \boldsymbol{J}^k = N l_k^2$. To obtain a time dependent differential equation of l_k, we square both sides of Eq. (10). We then average the term of the equation using the distribution of $P(u_k, u_{k'})$. Note that $\boldsymbol{x}$ and $\boldsymbol{J}^k$ are random variables, so the equation becomes a random recurrence formula. We formulate the size of the weight vectors to be $O(N)$, and the size of input $\boldsymbol{x}$ is $O(1)$, so the length of the student weight vector has a self-averaging property. Here, we rewrite m as $m = Nt$, and represent the learning process using continuous time

t in the thermodynamic limit of $N \to \infty$. We then obtain the deterministic differential equation of l_k,

$$\frac{dl_k^2}{dt} = \frac{2\eta_k}{K}\left(\sum_{i\neq k}^{K} Q_{ik} - (K-1)l_k^2\right) + \frac{\eta_k^2}{K^2}\left\{(K-1)^2 l_k^2 + \sum_{i\neq k}^{K} K(l_i^2 - 2KQ_{ik}) + 2\sum_{i=1}^{K-1}\sum_{j>i}^{K} Q_{ij}\right\}. \tag{11}$$

Here, $k = 1 \sim K$. $Q_{kk'} = q_{kk'} l_k l_{k'}$, and $q_{kk'}$ is the overlap between $\boldsymbol{J}^k$ and $\boldsymbol{J}^{k'}$, defined as

$$q_{kk'} = \frac{\boldsymbol{J}_k \cdot \boldsymbol{J}_{k'}}{|\boldsymbol{J}^k|\ |\boldsymbol{J}^{k'}|} = \frac{\boldsymbol{J}^k \cdot \boldsymbol{J}^{k'}}{N l_k l_{k'}}, \tag{12}$$

$q_{kk'}$ is the second order parameter of the system. The overlap $q_{kk'}$ also has a self-averaging property, so we can derive the differential equation in the thermodynamic limit of $N \to \infty$. The differential equation is derived by calculating the product of the learning equation (eq. (9)) for $\boldsymbol{J}^k$ and $\boldsymbol{J}^{k'}$, and we then average the term of the equation using the distribution of $P(u_k, u_{k'})$. After that, we obtain the deterministic differential equation as

$$\frac{dQ_{kk'}}{dt} = \frac{1}{K}\left\{\left(l_k^2 + \sum_{i\neq k}^{K} Q_{ik}\right)(\eta_k - \eta_k\eta_{k'}) + \left(l_{k'}^2 + \sum_{i\neq k'}^{K} Q_{ik'}\right)(\eta_{k'} - \eta_k\eta_{k'})\right\} + \frac{\eta_k\eta_{k'}}{K^2}\left(\sum_{i=1}^{K} l_i^2 + 2\sum_{i=1}^{K-1}\sum_{i>j}^{K} Q_{ij}\right) - Q_{kk'}(\eta_k + \eta_{k'} - \eta_k\eta_{k'}). \tag{13}$$

Equations (11) and (13) form closed differential equations.

To depict the behavior of mutual learning with a latent teacher, we have to obtain the differential equation of overlap R_k, which is a direction cosine between latent teacher weight vector $\boldsymbol{B}$ and the k-th student weight vector $\boldsymbol{J}^k$ defined by eq. (14). We introduce R_k as the third order parameter of the system.

$$R_k = \frac{\boldsymbol{B} \cdot \boldsymbol{J}^k}{|\boldsymbol{B}|\ |\boldsymbol{J}^k|} = \frac{\boldsymbol{B} \cdot \boldsymbol{J}^k}{N l_k} \tag{14}$$

For the sake of convenience, we write the overlap between the latent teacher weight vector and the student weight vector as r_k and $r_k = R_k l_k$. The differential equation of overlap r_k is derived by calculating the product of $\boldsymbol{B}$ and eq. (9), and we then average the term of the equation using the distribution of $P(v, u_k, u_{k'})$. The overlap r_k also has a self-averaging property, and in the thermodynamic

limit, the deterministic differential equation of r_k is then obtained through a calculation similar to that used for l_k.

$$\frac{dr_k}{dt} = \frac{\eta_k}{K}\left(\sum_{i\neq k}^{K} r_i - (K-1)r_k\right) \tag{15}$$

The squared error for the k-th student ϵ^k is then defined using the output of the latent teacher and that of the student as given in eqs. (2) and (6), respectively.

$$\epsilon^k = \frac{1}{2}\Big(\boldsymbol{B}\cdot\boldsymbol{x} - \boldsymbol{J}^k\cdot\boldsymbol{x}\Big)^2 \tag{16}$$

The generalization error for the k-th student ϵ_g^k is given by the squared error ϵ^k in eq. (16) averaged over the possible input $\boldsymbol{x}$ drawn from a Gaussian distribution $P(\boldsymbol{x})$ of zero mean and $1/N$ variance.

$$\epsilon_g^k = \int d\boldsymbol{x} P(\boldsymbol{x})\ \epsilon^k = \frac{1}{2}\int d\boldsymbol{x} P(\boldsymbol{x})\Big(\boldsymbol{B}\cdot\boldsymbol{x} - \boldsymbol{J}^k\cdot\boldsymbol{x}\Big)^2. \tag{17}$$

This calculation is the N-th Gaussian integral with $\boldsymbol{x}$ and it is hard to calculate. To overcome this difficulty, we employ coordinate transformation from $\boldsymbol{x}$ to v and u_k in eqs. (2) and (6). Note that the distribution of the output of the students $P(u_k)$ 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 the latent teacher $P(v)$ follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit. Thus, the distribution $P(v,u_k)$ of latent teacher output v and the k-th student output u_k is

$$P(v,u_k) = \frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left[-\frac{(v,u_k)^T\Sigma^{-1}\ (v,u_k)}{2}\right], \tag{18}$$

$$\Sigma = \begin{pmatrix} 1 & r_k \\ r_k & l_k^2 \end{pmatrix}. \tag{19}$$

Here, T denotes the transpose of a vector, r_k denotes $r_k = R_k l_k$, and R_k is the overlap between the latent teacher weight vector $\boldsymbol{B}$ and the student weight vector $\boldsymbol{J}^k$ defined by eq. (14). Hence, by using this coordinate transformation, the generalization error in eq. (17) can be rewritten as

$$\epsilon_g^k = \frac{1}{2}\int dv du_k (v-u_k)^2 = \frac{1}{2}(1-2r_k+l_k^2). \tag{20}$$

Consequently, we calculate the dynamics of the generalization error by substituting the time step value of $l_k(t)$, $Q(t)$, and $r_k(t)$ into eq. (20).

4 Results

In this section, we discuss the dynamics of the order parameters and their asymptotic properties, and then discuss the relationship between mutual learning and

Bagging. For the sake of simplicity, the initial weight vectors of the students are homogeneously correlated. From the symmetry of the evolution equation for updating the weight vector, $l_k(t) = l(t), Q_{kk'}(t) = Q(t), r_k(t) = r(t)$ are obtained. We assume the learning step size $\eta_k = \eta$. By substitute above conditions into Eqs. (11), (13), and (15), we get

$$\frac{dl^2}{dt} = -\frac{K-1}{K}(l^2 - Q)(2\eta - \eta^2), \tag{21}$$

$$\frac{dQ}{dt} = \frac{1}{K}(l^2 - Q)(2\eta - \eta^2), \tag{22}$$

$$\frac{dr}{dt} = 0. \tag{23}$$

Here, Eqs. (21) and (22) form closed differential equations. These equations can be solved analytically.

$$l^2(t) = \frac{K-1}{K}(l^2(0) - Q(0))\exp(-(2\eta - \eta^2)t) + \frac{l^2(0) + (K-1)Q(0)}{K}, \tag{24}$$

$$Q(t) = -\frac{1}{K}(l^2(0) - Q(0))\exp(-(2\eta - \eta^2)t) + \frac{l^2(0) + (K-1)Q(0)}{K}, \tag{25}$$

where $l^2(0)$ is the initial value of $l^2(t)$, and $Q(0)$ is the initial value of $Q(t)$. From Eqs. (24) and (25), $l^2(t)$ and $Q(t)$ are diverged when $\eta \geq 2$, learning will not converge in this condition.

Equation (23) depicts dynamics of overlap between the teacher and the student. The analytical solution of Eq. (23) is easily given by

$$r(t) = r(0). \tag{26}$$

Here, $r(0)$ is the initial value of $r(t)$. By substituting Eqs. (24) and (26) into (20), we can rewrite the generalization error for K students.

$$\epsilon_g^K(t) = \frac{1}{2}\left(1 - 2r(0) + \frac{K-1}{K}(l^2(0) - Q(0))\exp(-(2\eta - \eta^2)t) + \frac{l^2(0) + (K-1)Q(0)}{K}\right) \tag{27}$$

The asymptotic property of the order parameters $l(\infty)$, $Q(\infty)$ and $r(\infty)$ is given by substituting $t \to \infty$ into Eqs. (24), (25) and (26),

$$l^2(\infty) = Q(\infty) = \frac{l^2(0) + (K-1)Q(0)}{K}, \tag{28}$$

$$r(\infty) = r(0). \tag{29}$$

Consequently, we calculate the asymptotic property of the generalization error by substituting $l_k(\infty)$, $Q(\infty)$, and $r_k(\infty)$ into eq. (20).

$$\epsilon_g(\infty)^K = \frac{1}{2}\left(1 - 2r(0) + \frac{l^2(0) + (K-1)Q(0)}{K}\right) \tag{30}$$

From Eq. (35), the generalization error of Bagging ϵ_g^B using K weak learners is given by

$$\epsilon_g^B = \frac{1}{2}\Big\{\ 1 - 2r + \frac{l^2 + (K-1)Q}{K}\Big\}. \tag{31}$$

By substituting $l = l(0)$, $Q = Q(0)$ and $r = r(0)$ into Eq. (31), the generalization error of mutual learning ϵ_g^K was identical to the one of Bagging ϵ_g^B, then mutual learning asymptotically converged into Bagging.

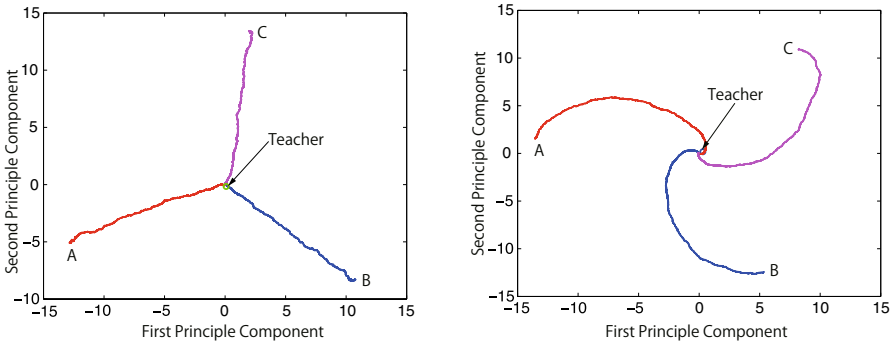
Moreover, in the limit of number of students is $K \to \infty$, the generalization error is

$$\epsilon_g(\infty)^\infty \sim \frac{1}{2}(1 - 2r(0) + Q(0)). \tag{32}$$

4.1 Learning Property through Computer Simulations

Mutual learning involving more than three students is the general case, compared to the two students case, with regard to the variety of students who can be selected to act as teacher. Figure 2 shows trajectories of the student weight vectors obtained by principle component analysis (PCA) during mutual learning involving three students. (The three students are respectively referred to as A, B and C, and A teaching B is indicated as A $\to$ B.) Figure 2(a) shows results obtained through learning where one student is randomly selected to act as teacher, and for comparison (b) shows results obtained through learning in a cyclic order of A$\to$ B $\to$ C $\to$ A. The symbol "o" at the center of each figure shows the weight vector of the latent teacher $\boldsymbol{B}$. In these figures, the horizontal axis shows the first principle component, and the vertical axis shows the second principle component. As shown, the trajectory of (a) converges with the minimum distance to the latent teacher, while that of (b) converges after a longer distance. This demonstrates that the learning style of (a) is superior to that of (b), confirming the validity of the proposed learning algorithm.

Next, we show the time dependence of the generalization error in Fig. 3. The number of students was 2, 3, 5, or 10, and the learning step size was $\eta = 0.1$. The initial conditions were $r(0) = 0.8$, $Q(0) = 0.6$, and $l(0) = 1$. The results were obtained through computer simulations with $N = 1000$ using random selection learning. In the figure, the horizontal axis is normalized time $t = m/N$, where m is the number of learning iterations. The vertical axis is the generalization error. The solid lines show the results using analytical solutions, and symbols "x", "+", "$*$", and "$\Box$" show the results for $K = 2, 3, 5$, and 10, respectively. We assumed a weight vector size of $O(\sqrt{N})$ and input size of $O(1)$ in the theoretical analysis. We kept these assumptions in the computer simulations, so if N is sufficiently large, the order parameters would have a self-averaging property. As shown, the



(a) Random selection learning (b) Cyclic order learning

Fig. 2. Trajectory of student weight vector during learning. Three students' case.

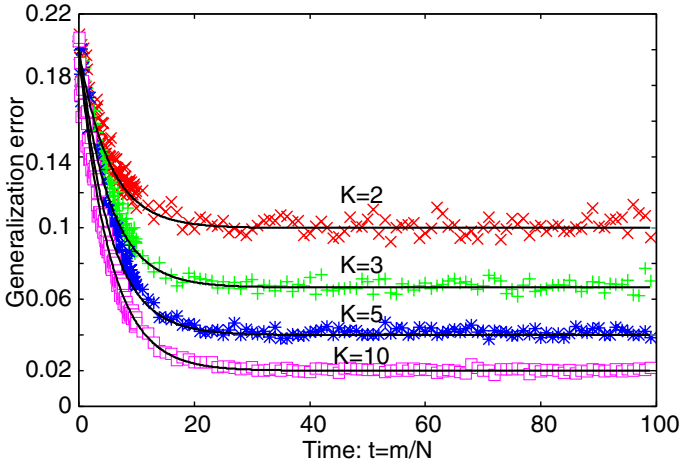


Fig. 3. Dependence of the mutual learning generalization error on the number of student networks K

analytical results agreed with those of the computer simulations, confirming the validity of the theoretical assumptions. We found that the generalization error decreased in proportion to $O(1/K)$. Moreover, the variance of the generalization error when using computer simulations tended to become smaller as the number of students K increased.

5 Conclusion

We have proposed a mutual learning algorithm using many students within the framework of on-line learning. From the results, analytical results are agreed with that of computer simulations, and the validity of the theoretical results

are shown. We showed that random selection of a student as a teacher is useful for mutual learning, and we found that the generalization error decreased in proportion to $O(1/K)$. Our future work is analyzing a mutual learning using many non-linear perceptrons.

Acknowledgment

This study has been supported by Grant-in-Aid for Scientific Research (C) No. 18500183.

References

1. Klein, E., et al.: Synchronization of neural networks by mutual learning and its application to cryptography. Proc. Neural Inf. Pro. Sys. 17, 689–696 (2004)
2. Mislovaty, R., Klein, E., Taunter, I., Kinzel, W.: Public channel cryptography by synchronization of neural networks and chaotic maps. Phys. Rev. Lett. 91(11), 118701 (2003)
3. Bran, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
4. Freund, Y., Shari'a, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139 (1997)
5. Hara, K., Okada, M.: Statistical mechanics of mutual learning with latent teacher. J. Phys. Soc. Jpn. 76, 014001 (2007)
6. Hara, K., Yamada, T.: Optimization of the asymptotic property of mutual learning involving an integration mechanism of ensemble learning. J. Phys. Soc. Jpn. 76, 024005 (2008)

A Bagging

Bagging is a learning method using many weak learning machines to improve upon the performance of a single weak learning machine[3]. Students learn from the teacher individually, and then an ensemble output $\overline{u}$ is calculated.

$$\overline{u} = \frac{\sum_{k=1}^{K} u_k}{K} = \frac{\sum_{k=1}^{K} \left(\boldsymbol{J}^k \cdot \boldsymbol{x} \right)}{K} = \boldsymbol{J}^B \cdot \boldsymbol{x}. \tag{33}$$

The length of the weight vector l^B and the overlap r^B are given by ,

$$(l^B)^2 = \frac{l^2 + (K-1)Q}{K}, \quad r^B = \frac{1}{K} \sum_{k=1}^{K} r_k = r. \tag{34}$$

Here, we assumed the conditions of $l_k = l$, $Q_k = Q$, and $r_k = r$, respectively. The generalization error of ensemble output ϵ_g^B is given by substituting Eqs. (34) into Eq. (20):

$$\epsilon_g^B = \frac{1}{2} \left\{ 1 - 2r + \frac{l^2 + (K-1)Q}{K} \right\}. \tag{35}$$