On-Line Learning Theory of Soft Committee Machines with Correlated Hidden Units
—Steepest Gradient Descent and Natural Gradient Descent—

Masato Inoue [1,2,*], Hyeyoung Park [1,†] and Masato Okada [1,3]

[1] *Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351-0198*
[2] *Department of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto University, Kyoto 606-8507*
[3] *"Intelligent Cooperation and Control," PRESTO, JST, Saitama 351-0198*

The permutation symmetry of the hidden units in multilayer perceptrons causes the saddle structure and plateaus of the learning dynamics in gradient learning methods. The correlation of the weight vectors of hidden units in a teacher network is thought to affect this saddle structure, resulting in a prolonged learning time, but this mechanism is still unclear. In this paper, we discuss it with regard to soft committee machines and on-line learning using statistical mechanics. Conventional gradient descent needs more time to break the symmetry as the correlation of the teacher weight vectors rises. On the other hand, no plateaus occur with natural gradient descent regardless of the correlation for the limit of a low learning rate. Analytical results support these dynamics around the saddle point.

1. Introduction

One of the biggest problems of neural network learning is the plateau of the learning curve. Considering the gradient learning method and its generalization error, this plateau is mainly caused by the saddle structure of the error function. The permutation symmetry prevents the identification of the hidden units in multilayer perceptrons if they have the same weight vectors, and produces this saddle structure.[1,2] In the learning scenario of a teacher and a student network, the saddle is thought to be affected by the strength of the correlation of the hidden units in the teacher network, which may be closely related to the length of the plateau. More specifically, in the conventional gradient descent (GD), the weight vectors in the student network are known to approach the saddle before reaching their final states.[2] Since the saddle is located between the weight vectors of the teacher hidden units, their stronger correlation is supposed to force the student weight vectors closer to the saddle, resulting in a longer plateau.

Natural gradient descent (NGD), however, may be able to avoid the saddle because it can update the network parameters to the optimal direction in the Riemannian space.[3] NGD is a fairly general method for effectively adjusting the parameters of stochastic models, but its validity in multilayer perceptrons is uncertain because of three intrinsic problems: 1) NGD needs prior knowledge of the input distribution to calculate the Fisher information matrix, 2) NGD is unstable around the singular points of the Fisher information matrix, 3) matrix inversion is time consuming, which might be critical especially in real-time learning. The method proposed by Yang and Amari[4] can be used to calculate NGD efficiently in the case of a large input dimension in multilayer perceptrons. Also, the adaptive method can be used to approximate the inverse of the Fisher information matrix asymptotically without prior knowledge or matrix inversion.[5] In this paper, we discuss the problem of singularity; since the saddle is one of the singular points, how NGD works around there is one of our main topics.

On-line learning is one of the most popular forms of training. Analysis of the network dynamics in on-line learning is much easier than for batch learning because the state of the network and the learning samples are independent of each other. In this framework, the statistical mechanics method proposed by Saad and Solla can be used to analyze the GD dynamics exactly at the large limit of the input dimension.[2] Rattray and Saad extended this technique to NGD and reported that it works efficiently in multilayer perceptrons.[6] In this paper, we also use this method and contrast the dynamics for GD and NGD, focusing on the corrupted saddle structure under a strong correlation of the hidden units in the teacher network.

2. Model

Soft committee machines (Fig. 1) are considered where the teacher network has M hidden units while the student has K units. To apply NGD, Gaussian noise $n \sim \mathcal{N}(0, \sigma^2)$ is added to the output of the student;

$$\zeta \equiv f_{\boldsymbol{B}}(\boldsymbol{\xi}), \qquad f_{\boldsymbol{B}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{M} g(\boldsymbol{B}_k^T \boldsymbol{\xi}), \qquad (1)$$

$$\zeta' \equiv f_{\boldsymbol{J}}(\boldsymbol{\xi}) + n, \quad f_{\boldsymbol{J}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{K} g(\boldsymbol{J}_k^T \boldsymbol{\xi}), \qquad (2)$$

where $\boldsymbol{\xi} \in \mathbb{R}^N$ denotes the input vector while $\boldsymbol{B}_i \in \mathbb{R}^N$ and $\boldsymbol{J}_i \in \mathbb{R}^N$ are the ith weight vectors of the teacher and the student networks, respectively. Here, T means the transposition while g is an activation function.

---

* E-mail address: minoue@brain.riken.go.jp
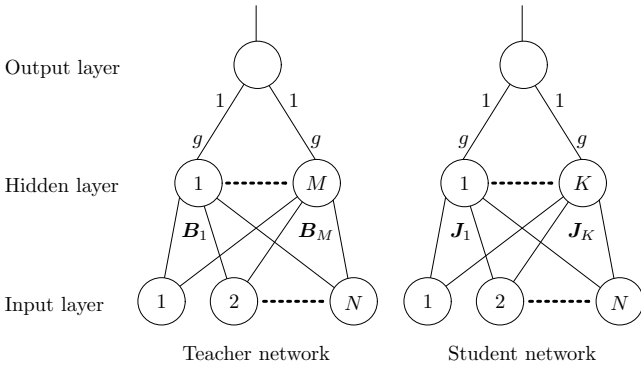† E-mail address: hypark@brain.riken.go.jp

Fig. 1. Teacher and student networks. Each weight between any hidden unit and the output is fixed to 1.

The joint probability distribution of the input $\boldsymbol{\xi}$ and the output ζ' of the student network is given by

$$p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta') \equiv p(\boldsymbol{\xi})p_{\boldsymbol{J}}(\zeta'|\boldsymbol{\xi}), \qquad (3)$$

$$p_{\boldsymbol{J}}(\zeta'|\boldsymbol{\xi}) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\{\zeta' - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2}{2\sigma^2}\right). \qquad (4)$$

The parameter vector of (3), $\boldsymbol{J} \equiv [\boldsymbol{J}_1^T, \boldsymbol{J}_2^T, ..., \boldsymbol{J}_K^T]^T \in \mathbb{R}^{KN}$, is updated iteratively to approximate the joint probability distribution of the input $\boldsymbol{\xi}$ and the output ζ of the teacher network,

$$p(\boldsymbol{\xi}, \zeta) \equiv p(\boldsymbol{\xi})\delta(\zeta - f_{\boldsymbol{B}}(\boldsymbol{\xi})), \qquad (5)$$

where δ is the delta function. The loss function for a given set of a learning sample $\{\boldsymbol{\xi}, \zeta\}$, defined using the logarithmic loss of the conditional probability distribution of (4), is

$$\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) \equiv -\ln p_{\boldsymbol{J}}(\zeta|\boldsymbol{\xi}) + c_0 = \frac{1}{2\sigma^2}\{\zeta - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2, \qquad (6)$$

where $c_0 \equiv -\ln\sqrt{2\pi\sigma^2}$ is constant. The generalization error is then defined as the expected loss:

$$\epsilon_g(\boldsymbol{J}) \equiv \langle\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta)\rangle_{\{\boldsymbol{\xi}, \zeta\}}. \qquad (7)$$

The definitions of (6) can be written, by applying (1) and (2), as

$$\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) = \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}) \equiv \frac{1}{2\sigma^2}\{f_{\boldsymbol{B}}(\boldsymbol{\xi}) - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2. \qquad (8)$$

We consider on-line learning in this paper, where the parameter vector $\boldsymbol{J}$ is updated for each set of an independently given sample $\{\boldsymbol{\xi}, \zeta\}$. The updating rule, the differential of $\boldsymbol{J}$, for GD is defined with a learning rate η as

$$\Delta\boldsymbol{J} = -\frac{\eta}{N}\nabla_{\boldsymbol{J}}\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta), \qquad (9)$$

where

$$\nabla_{\boldsymbol{J}i}\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) = -\frac{1}{\sigma^2}g'(\boldsymbol{J}_i^T\boldsymbol{\xi})\{f_{\boldsymbol{B}}(\boldsymbol{\xi}) - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}\boldsymbol{\xi}, \qquad (10)$$

where g' denotes the derivative of g. One for NGD is also defined as

$$\Delta\boldsymbol{J} = -\frac{\eta}{N}\boldsymbol{G}^{-1}\nabla_{\boldsymbol{J}}\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta), \qquad (11)$$

where $\boldsymbol{G}$ denotes the Fisher information matrix of the parameter vector $\boldsymbol{J}$:

$$\boldsymbol{G} \equiv \langle[\nabla_{\boldsymbol{J}}\ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')][\nabla_{\boldsymbol{J}}\ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')]^T\rangle_{\{\boldsymbol{\xi}, \zeta'\}}. \qquad (12)$$

The $\boldsymbol{G}$ can be written, in block form, as

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{1,1} & \cdots & \boldsymbol{G}_{1,K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{K,1} & \cdots & \boldsymbol{G}_{K,K} \end{bmatrix},$$

$$\boldsymbol{G}_{ij} = \frac{1}{\sigma^2}\langle g'(\boldsymbol{J}_i^T\boldsymbol{\xi})g'(\boldsymbol{J}_j^T\boldsymbol{\xi})\boldsymbol{\xi}\boldsymbol{\xi}^T\rangle_{\{\boldsymbol{\xi}\}}. \qquad (13)$$

In the case of the standard multivariate normal distribution input, $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the inverse of the Fisher information matrix is also given by

$$\boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}_{1,1}^{-1} & \cdots & \boldsymbol{G}_{1,K}^{-1} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{K,1}^{-1} & \cdots & \boldsymbol{G}_{K,K}^{-1} \end{bmatrix},$$

$$\boldsymbol{G}_{ij}^{-1} = \sigma^2\left[\theta_{ij}\boldsymbol{I} + \boldsymbol{J}'\boldsymbol{\Theta}_{ij}\boldsymbol{J}'^T\right], \qquad (14)$$

where $\boldsymbol{J}' \equiv [\boldsymbol{J}_1 \cdots \boldsymbol{J}_K]$ is a N by K matrix, while θ_{ij} is a scalar and $\boldsymbol{\Theta}_{ij}$ is a K by K matrix.[4]

3. Theory

3.1 Order parameters and generalization error

At the thermodynamics limit, the limit of $N \to \infty$, the dynamics of the network can be analyzed using statistical mechanics. Here, the order parameters that represent the correlations of the weight vectors are used instead of the N-dimensional vectors $\boldsymbol{\xi}$, $\boldsymbol{B}_i$, and $\boldsymbol{J}_i$. To make the present paper self-contained, we briefly summarize the derivation of the order parameter equations of the soft committee machine.[2, 6]

From here on, the input vector is assumed to obey a N-dimensional multivariate Gaussian noise with zero mean and a unit covariance matrix: $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The correlation between the input and each weight vector, denoted by $x_i \equiv \boldsymbol{J}_i^T\boldsymbol{\xi}$ and $y_i \equiv \boldsymbol{B}_i^T\boldsymbol{\xi}$, is then distributed as a normal distribution; $x_i \sim \mathcal{N}(0, \boldsymbol{J}_i^T\boldsymbol{J}_i)$ and $y_i \sim \mathcal{N}(0, \boldsymbol{B}_i^T\boldsymbol{B}_i)$, while each covariance of them is given by $\langle x_i x_j\rangle_{\{\boldsymbol{\xi}\}} = \boldsymbol{J}_i^T\boldsymbol{J}_j$, $\langle x_i y_j\rangle_{\{\boldsymbol{\xi}\}} = \boldsymbol{J}_i^T\boldsymbol{B}_j$, and $\langle y_i y_j\rangle_{\{\boldsymbol{\xi}\}} = \boldsymbol{B}_i^T\boldsymbol{B}_j$. Therefore, a new vector, defined as

$$\boldsymbol{z} \equiv [x_1, ..., x_K, y_1, ..., y_M]^T \in \mathbb{R}^{K+M}, \qquad (15)$$

is distributed as a multivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$:

$$p(\boldsymbol{z}) = \frac{1}{\sqrt{\{2\pi\}^{K+M}|\boldsymbol{C}|}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{C}^{-1}\boldsymbol{z}\right), \qquad (16)$$

where $\boldsymbol{C}$ is the variance-covariance matrix:

$$\boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{T} \end{bmatrix} \qquad (17)$$

with

$$\boldsymbol{Q} \equiv \boldsymbol{J}'^T \boldsymbol{J}' = \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,K} \\ \vdots & \ddots & \vdots \\ Q_{K,1} & \cdots & Q_{K,K} \end{bmatrix}, \qquad (18)$$

$$\boldsymbol{R} \equiv \boldsymbol{J}'^T \boldsymbol{B}' = \begin{bmatrix} R_{1,1} & \cdots & R_{1,M} \\ \vdots & \ddots & \vdots \\ R_{K,1} & \cdots & R_{K,M} \end{bmatrix}, \qquad (19)$$

$$\boldsymbol{T} \equiv \boldsymbol{B}'^T \boldsymbol{B}' = \begin{bmatrix} T_{1,1} & \cdots & T_{1,M} \\ \vdots & \ddots & \vdots \\ T_{M,1} & \cdots & T_{M,M} \end{bmatrix}, \qquad (20)$$

and $\boldsymbol{J}' \equiv [\boldsymbol{J}_1 \cdots \boldsymbol{J}_K]$, $\boldsymbol{B}' \equiv [\boldsymbol{B}_1 \cdots \boldsymbol{B}_M]$. Here, $\boldsymbol{Q}$ and $\boldsymbol{R}$ are the order parameters of this system.

Using these order parameters, the generalization error in (7), $\epsilon_g(\boldsymbol{J}) \equiv \langle \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta) \rangle_{\{\boldsymbol{\xi}, \zeta\}}$, can be calculated by

$$\epsilon_g(\boldsymbol{J}) = \int d\boldsymbol{z}\, p(\boldsymbol{z}) \frac{1}{2\sigma^2} \left\{ \sum_{k=1}^{M} g(y_k) - \sum_{k=1}^{K} g(x_k) \right\}^2. \quad (21)$$

If we define the activation function g as $g(x) = \mathrm{erf}(x/\sqrt{2})$ from here on, the generalization error is given by

$$\epsilon_g(\boldsymbol{J}) = \frac{1}{\pi\sigma^2} \left\{ -2 \sum_{i=1}^{K} \sum_{j=1}^{M} \arcsin \frac{R_{ij}}{\sqrt{Q_{ii}+1}\sqrt{T_{jj}+1}} \right.$$
$$+ \sum_{i,j=1}^{K} \arcsin \frac{Q_{ij}}{\sqrt{Q_{ii}+1}\sqrt{Q_{jj}+1}}$$
$$\left. + \sum_{i,j=1}^{M} \arcsin \frac{T_{ij}}{\sqrt{T_{ii}+1}\sqrt{T_{jj}+1}} \right\}, \quad (22)$$

which depends on only the order parameters.

3.2 Dynamics of the order parameters

Here we substitute the dynamics of the order parameters for those of the system. First, we can replace the updating rule (9) with

$$\Delta \boldsymbol{J}_i = -\frac{\eta}{\sigma^2 N} \delta_i \boldsymbol{\xi}, \qquad (23)$$

where

$$\delta_i \equiv -g'(x_i) \left\{ \sum_{k=1}^{M} g(y_k) - \sum_{k=1}^{K} g(x_k) \right\}. \qquad (24)$$

Thus, the updating rule of the order parameters is given by

$$\Delta R_{ij} = [\boldsymbol{J}_i + \Delta \boldsymbol{J}_i]^T \boldsymbol{B}_j - \boldsymbol{J}_i^T \boldsymbol{B}_j$$
$$= -\frac{\eta}{\sigma^2 N} \delta_i y_j, \qquad (25)$$

and

$$\Delta Q_{ij} = [\boldsymbol{J}_i + \Delta \boldsymbol{J}_i]^T [\boldsymbol{J}_j + \Delta \boldsymbol{J}_j] - \boldsymbol{J}_i^T \boldsymbol{J}_j$$
$$= -\frac{\eta}{\sigma^2 N} \{\delta_i x_j + \delta_j x_i\} + \frac{\eta^2}{\sigma^4 N^2} \delta_i \delta_j \boldsymbol{\xi}^T \boldsymbol{\xi}. \quad (26)$$

Here we introduce the time α; a short period, $\Delta\alpha = 1/N$, is defined to be consumed for each learning iteration. At the large limit of N, the differential dynamics of R_{ij} and Q_{ij} are calculated as

$$\frac{\partial R_{ij}}{\partial \alpha} = \lim_{\Delta\alpha \to 0} \frac{\Delta R_{ij}}{\Delta\alpha} = \lim_{N \to \infty} N \Delta R_{ij} = -\frac{\eta}{\sigma^2} \langle \delta_i y_j \rangle_{\{\boldsymbol{z}\}}$$
$$= -\frac{\eta}{\sigma^2} \phi_{ij} \qquad (27)$$

and

$$\frac{\partial Q_{ij}}{\partial \alpha} = -\frac{\eta}{\sigma^2} \{\psi_{ij} + \psi_{ji}\} + \frac{\eta^2}{\sigma^4} \upsilon_{ij}, \qquad (28)$$

where $\boldsymbol{\xi}^T \boldsymbol{\xi} \to N$ is applied, while the new variables are defined as

$$\phi_{ij} \equiv \langle \delta_i y_j \rangle_{\{\boldsymbol{z}\}}, \ \psi_{ij} \equiv \langle \delta_i x_j \rangle_{\{\boldsymbol{z}\}}, \ \upsilon_{ij} \equiv \langle \delta_i \delta_j \rangle_{\{\boldsymbol{z}\}}. \quad (29)$$

The dynamics for NGD can be provided in the same way:

$$\Delta \boldsymbol{J}_i = -\frac{\eta}{\sigma^2 N} \sum_{k=1}^{K} \delta_k \boldsymbol{G}_{ik}^{-1} \boldsymbol{\xi}. \qquad (30)$$

Thus, the dynamics of the order parameters are

$$\frac{\partial R_{ij}}{\partial \alpha} = -\eta \sum_{k=1}^{K} \{\theta_{ik} \phi_{kj} + \psi_{k\bullet} \boldsymbol{\Theta}_{ik}^T R_{\bullet j}\}, \qquad (31)$$

$$\frac{\partial Q_{ij}}{\partial \alpha} = -\eta \sum_{k=1}^{K} \{\theta_{ik} \psi_{kj} + \theta_{jk} \psi_{ki} + \psi_{k\bullet} \boldsymbol{\Theta}_{ik}^T Q_{\bullet j}$$
$$+ \psi_{k\bullet} \boldsymbol{\Theta}_{jk}^T Q_{\bullet i}\} + \eta^2 \sum_{k,l=1}^{K} \theta_{ik} \theta_{jl} \upsilon_{kl}, \qquad (32)$$

where $\psi_{k\bullet}$ denotes the kth row of the matrix $\{\psi_{ij}\}_{i,j=1,\ldots,K}$, while $R_{\bullet j}$ denotes the jth column of the matrix $\boldsymbol{R}$, and so on.[6]

4. Numerical Results

In this section, we discuss how the learning dynamics depend on the correlation of teacher weight vectors $T_{1,2}$. The results are also contrasted between GD and NGD.

We set the number of the hidden units and the lengths of the teacher weight vectors as follows:

$$K = M = 2, \quad T_{1,1} = T_{2,2}. \qquad (33)$$

We also restrict the initial conditions to

$$Q_{1,1} = Q_{2,2}, \quad R_{1,1} = R_{2,2}, \quad R_{1,2} = R_{2,1}. \qquad (34)$$

Because of the symmetry of the system, these restrictions are preserved throughout the learning. Specifically, we use

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{bmatrix}. \qquad (35)$$

Therefore, we have four free parameters $Q_{1,1}$, $Q_{1,2}$, $R_{1,1}$, and $R_{1,2}$ in this system. Note that $\boldsymbol{Q}$ and $\boldsymbol{T}$ are always symmetric matrices from the definitions of (18) and (20). Other parameters are set as $T_{1,2} = 1$, $\eta = 10^{-2}$, and
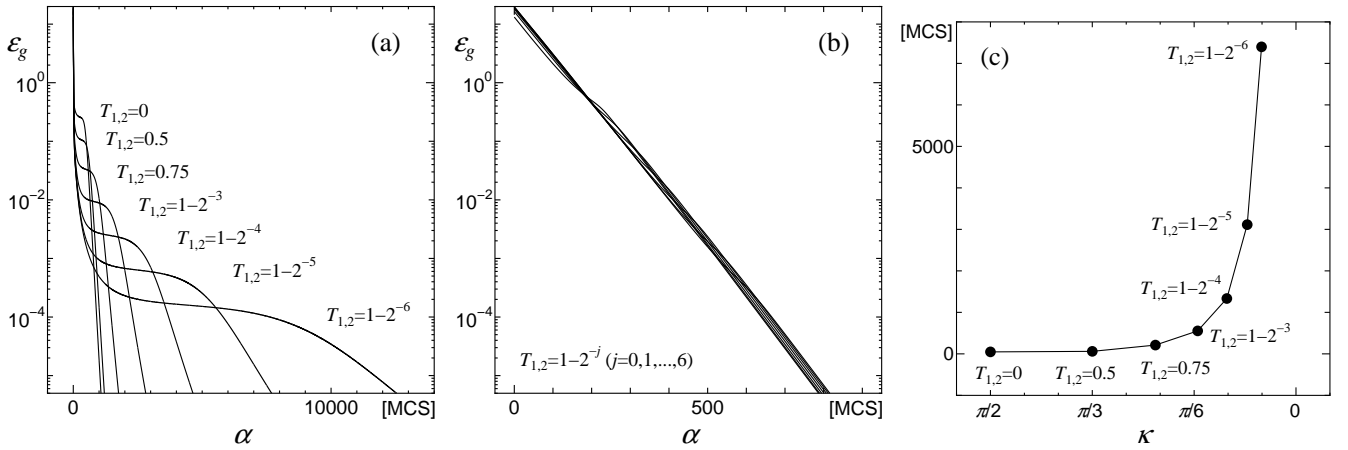
Fig. 2. Time evolution of the generalization error in GD (a) and NGD (b). All the trajectories are almost completely overlapped in (b). The plateau periods in (a) were measured and are shown in (c). MCS denotes the number of Monte Carlo steps.
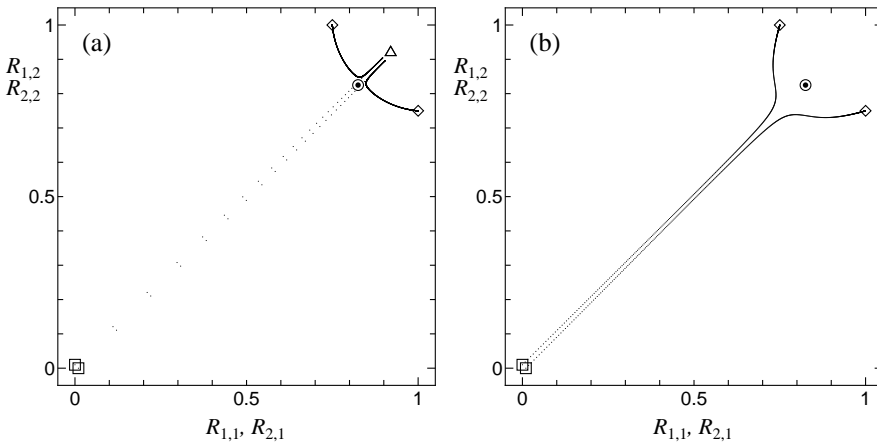

Fig. 3. Time evolution of the order parameters $(R_{1,1}, R_{1,2})$ and $(R_{2,1}, R_{2,2})$ in GD (a) and NGD (b). The correlation of the teacher weight vectors $T_{1,2} = 0.75$. Start points: □; turning points: △; the saddle: ⊙; and goals: ◇ .

$\sigma^2 = 5 \times 10^{-2}$. Various values for $T_{1,2}$ are employed to examine the influence of the correlation of the teacher hidden units. We sometimes use $\kappa \equiv \arccos \frac{T_{1,2}}{T_{1,1}}$, the angle of the teacher weight vectors, instead of $T_{1,2}$.

In this case, θ_{ij} and $\boldsymbol{\Theta}_{ij}$ in the inverse of the Fisher information matrix (14) can be simplified as

$$\theta_{1,1} = \theta_{2,2} = c\sqrt{a}, \quad \theta_{1,2} = \theta_{2,1} = -c\sqrt{b},$$

$$\boldsymbol{\Theta}_{1,1} = d\sqrt{a} \begin{bmatrix} 2a\{a-b\}-bQ_{1,1} & bQ_{1,2} \\ bQ_{1,2} & -bQ_{1,1} \end{bmatrix},$$

$$\boldsymbol{\Theta}_{2,2} = d\sqrt{a} \begin{bmatrix} -bQ_{1,1} & bQ_{1,2} \\ bQ_{1,2} & 2a\{a-b\}-bQ_{1,1} \end{bmatrix}, \quad (36)$$

$$\boldsymbol{\Theta}_{1,2} = \boldsymbol{\Theta}_{2,1}$$
$$= -d\sqrt{b} \begin{bmatrix} a\{Q_{1,1}+1\}-b^2 & aQ_{1,2} \\ aQ_{1,2} & a\{Q_{1,1}+1\}-b^2 \end{bmatrix},$$

where

$$a \equiv \{Q_{1,1}+1\}^2 - Q_{1,2}^2, \quad b \equiv 2Q_{1,1}+1,$$

$$c \equiv \frac{\pi}{2} \frac{\sqrt{ab}}{a-b}, \quad d \equiv \frac{c}{a^2-b^2}. \quad (37)$$

Here we summarize the order of each variable to N. Since the length of the input vector $\boldsymbol{\xi}$ is $O(\sqrt{N})$, x_i and y_i are $O(1)$. This guarantees that the arguments of the activation function g are $O(1)$. Therefore, the lengths of the weight vectors, $\sqrt{Q_{ii}}$ and $\sqrt{T_{ii}}$, are $O(1)$. If the direction of the initial $\boldsymbol{J}_i$ is chosen randomly, the size of R_{ii}, the correlation between $\boldsymbol{J}_i$ and $\boldsymbol{B}_i$, is $O(1/\sqrt{N})$. The initial numerical values in (35) are defined according to these sizes.

Figure 2 shows the time evolution of the generalization error. In the GD (Fig. 2a), the plateau was greatly prolonged as the correlation of the teacher weight vectors rose. In NGD (Fig. 2b), almost no plateau occurred at any $T_{1,2}$ if η was set small enough relative to the initial $R_{1,1}$, and the generalization error was exponentially decreased. The plateau periods of Fig. 1a were measured and are shown in Fig. 2c, where we defined a plateau as occurring if $\frac{\partial \ln \epsilon_g}{\partial \alpha} > -0.0005$. The order of the plateau lengths was about $O(\kappa^{-3})$ in GD.

Figure 3 shows the trajectories of the order parameters $(R_{1,1}, R_{1,2})$ and $(R_{2,1}, R_{2,2})$. Because of the symmetry, the latter plots are mirror images of the former. As $R_{1,1}$ is the correlation between the first student and the cor-
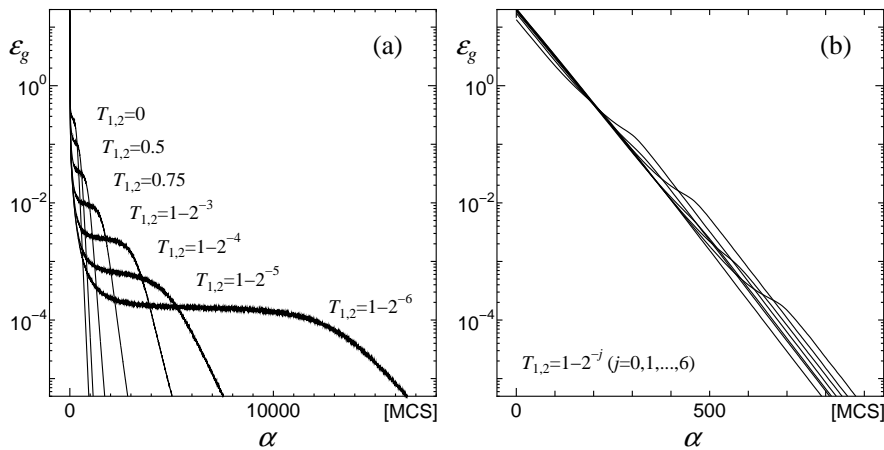
Fig. 4. Time evolution of the generalization error for the case of $N = 1000$ in GD (a) and NGD (b).
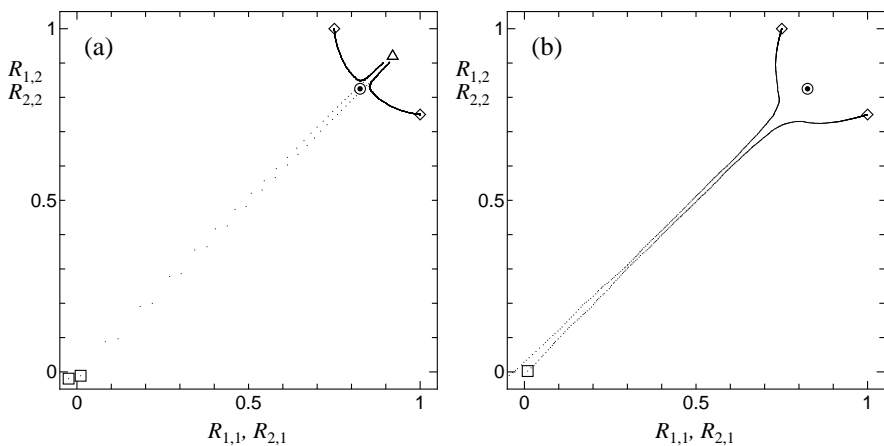

Fig. 5. Time evolution of the order parameters $(R_{1,1}, R_{1,2})$ and $(R_{2,1}, R_{2,2})$ for the case of $N = 1000$ in GD (a) and NGD (b) as in Fig. 3.

responding teacher, the initial value is almost 0 and the goal is 1; $R_{1,2}$ is the correlation between the first student and the not corresponding teacher, and the initial value is almost 0 and the goal is $T_{1,2}$. Therefore, the target location of the plots are $(1, T_{1,2})$ and $(T_{1,2}, 1)$, respectively (shown as $\diamondsuit$). The other order parameters $Q_{1,1}$ and $Q_{1,2}$ are not shown. In the case of GD (Fig. 3a), the plots start at $\square$, turn back at $\triangle$, then approach $\odot$ (the saddle, as explained in the next section), and finally reach $\diamondsuit$. Actually, the parameters never pass through the same place again because $Q_{1,1}$ and $Q_{1,2}$ are updated. In the case of NGD (Fig. 3b), the plots start at $\square$ and reach $\diamondsuit$ while avoiding $\odot$.

We performed a numerical simulation to confirm the dynamics at the above thermodynamics limit. The input dimension was $N = 1000$, the teacher weight vectors were set as

$$\boldsymbol{B}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \boldsymbol{B}_2 = \begin{bmatrix} \cos \kappa \\ \sin \kappa \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (38)$$

and every initial $\boldsymbol{J}_i$ was randomly and independently chosen from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/N)$ for each try. Thus, the order parameters $\boldsymbol{Q}$ and $\boldsymbol{R}$ were no longer limited by the restriction of (34). The learning was performed using these real weight vectors and the original equations: (9) for GD and (11) for NGD. Figures 4 and 5 show the time evolution of the generalization error and the trajectories of the order parameters in the same manner as Figs. 2 and 3, respectively. Both figures support the statistical dynamics well, which suggests the constraint of (34) is a rather minor problem and the system retains most of its generality even with that restriction.

5. Saddle

Here, we discuss why NGD is so effective even with a strong correlation between teacher hidden units. We consider the dynamics around the saddle of the generalization error under the conditions of (33) and (34). This point, where all the differentials of the order parameters are zero and the Hessian matrix is not positive definite nor negative definite, is shown as $\odot$ in Figs. 3 and 5:

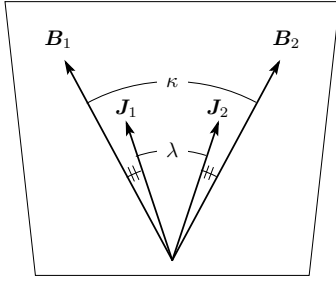$$Q_{1,1} = Q_{1,2} = \frac{T_{1,1} + T_{1,2}}{T_{1,1} - T_{1,2} + 2},$$

Fig. 6. The student weight vectors $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ belong to the plane made by the teacher weight vectors $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.

$$R_{1,1} = R_{1,2} = \frac{T_{1,1} + T_{1,2}}{\sqrt{2\{T_{1,1} - T_{1,2} + 2\}}}, \qquad (39)$$

where the term η^2 included in $\frac{\partial Q_{ij}}{\partial \alpha}$ can be ignored if the learning rate η is set small enough. This saddle is a special point because 1) it corresponds to the goal both in the case of $T_{1,1} = T_{1,2}$ (the teacher is a smaller network: $f_{\boldsymbol{B}}(\boldsymbol{\xi}) = 2g(\boldsymbol{B}_1^T\boldsymbol{\xi})$) and in the case that the student is a smaller network: $f_{\boldsymbol{J}}(\boldsymbol{\xi}) = 2g(\boldsymbol{J}_1^T\boldsymbol{\xi})$, 2) in GD, the plateau occurs around it, and in NGD the student vectors avoid it, 3) it coincides with one of the singular points of the Fisher information matrix since $Q_{1,1} = Q_{1,2}$.

We simplify the situation as shown in Fig. 6; the two student weight vectors belong to the plane made by the two teacher weight vectors. This simplification is useful because we are now interested in how fast the student vectors leave this point for the goals. The correlations are re-parameterized by κ and λ as

$$T_{1,2} = T_{1,1}\cos\kappa, \quad Q_{1,2} = Q_{1,1}\cos\lambda,$$
$$R_{1,1} = \sqrt{Q_{1,1}T_{1,1}}\cos\frac{\kappa-\lambda}{2},$$
$$R_{1,2} = \sqrt{Q_{1,1}T_{1,1}}\cos\frac{\kappa+\lambda}{2}. \qquad (40)$$

Now, we have only two free parameters $Q_{1,1}$ and λ. Since the first derivative of λ can be written with $Q_{1,1}$ and $Q_{1,2}$ as

$$\frac{\partial \lambda}{\partial \alpha} = \frac{\partial}{\partial \alpha}\arccos\frac{Q_{1,2}}{Q_{1,1}} = \frac{Q_{1,2}\frac{\partial Q_{1,1}}{\partial \alpha} - Q_{1,1}\frac{\partial Q_{1,2}}{\partial \alpha}}{Q_{1,1}\sqrt{Q_{1,1}^2 - Q_{1,2}^2}}, \quad (41)$$

we can formulate the angular velocity of λ at $0 < \lambda \ll 1$.

The angular velocity for GD is

$$\frac{\partial \lambda}{\partial \alpha} = c_1 \lambda \sin^2 \kappa, \qquad (42)$$

where $c_1 \equiv \frac{4\eta}{\pi\sigma^2}T_{1,1}\{T_{1,1}\{1-\cos\kappa\}+2\}^{-\frac{1}{2}}\{T_{1,1}\{3+\cos\kappa\}+2\}^{-\frac{3}{2}}$. We notice that the order of c_1 is not greatly changed by κ. The velocity converges to zero in the first order of λ. Moreover, it decreases as κ decreases. Therefore, this equation supports the simulation results

showing that the plateau is prolonged as the teacher correlation rises. The angular velocity for NGD is

$$\frac{\partial \lambda}{\partial \alpha} = c_2 \frac{1}{\lambda}\tan^2\frac{\kappa}{2}, \qquad (43)$$

where $c_2 \equiv 2\eta$. This velocity diverges to infinity as λ goes to zero. Although it decreases as κ decreases, this

effect would be canceled by λ^{-1} near the saddle. Therefore, this equation means that the student weight vectors are repelled by the saddle. In addition, this also supports the simulation results showing that the student weight vectors avoid the saddle and that the plateau does not occur even in the case of strongly correlated teacher hidden units.

In the case of $\lambda = 0$, these features of slow dynamics for GD and fast dynamics for NGD are still preserved. Specifically, dynamics for NGD (43) still holds, while the one for GD (42) must take the lower term into account:

$$\frac{\partial \lambda}{\partial \alpha} = c_1 \lambda \sin^2 \kappa + h_0(\lambda, \kappa)h_1(N), \ \left(\lim_{N\to\infty} h_1(N) = 0\right), \ (44)$$

where the second term of the right hand side represents fluctuation. Substituting $\lambda = 0$ in this eq., we obtain $\frac{\partial \lambda}{\partial \alpha} = h_0(\lambda, \kappa)h_1(N)$, which means that temporal evolution of λ is scaled as $O(h_1(N))$, and implies that the dynamics for GD is sufficiently slower than that for NGD.

6. Conclusion

We have studied the on-line learning of soft committee machines under correlated teacher hidden units. The plateau in GD is largely prolonged at about $O(\kappa^{-3})$ as the correlation of the teacher weight vectors rises, but almost no plateau occurs in NGD with a low learning rate η and this does not depend on the correlation. Our analytical results for around the saddle reveal that the NGD avoided the saddle, even though the strong correlation of the teacher weight vectors forced the student weight vectors close to the saddle where the Fisher information matrix is singular.

1) K. Fukumizu and S. Amari: Neural Networks **13** (2000) 317.
2) D. Saad and A. Solla: Phys. Rev. E **52** (1995) 4225.
3) S. Amari: Neural Comput. **10** (1998) 251.
4) H. H. Yang and S. Amari: Neural Comput. **10** (1998) 2137.
5) S. Amari, H. Park and K. Fukumizu: Neural Comput. **12** (2000) 1399.
6) M. Rattray and D. Saad: Phys. Rev. E **59** (1999) 4523.