On-line learning of a mixture-of-experts neural network

On-line learning of a mixture-of-experts neural network

N-J Huh, J-H Oh and K Kang

Department of Physics, Pohang University of Science and Technology, Pohang, 790-784, Korea

Received 19 September 2000

Abstract. The on-line learning of a mixture-of-experts system is studied in the framework of statistical physics. The time dependence of the overlap-order parameters during training is calculated analytically in the thermodynamic limit. When the number of training examples is small each expert is in a symmetric state. As the number of time steps approaches a critical point, the symmetric state begins to disintegrate. This symmetry-breaking behaviour is accounted for by means of a gating network. In the symmetric state the gating network has little effect on the learning, but when the symmetry is broken the gating network assigns the experts to appropriate subspaces in the input space. A generalization curve shows a plateau between the symmetric- and broken-symmetry states. We also find that the learning curves show different behaviours depending on the stiffness of the gating function.

1. Introduction

A neural network can be trained using a set of examples. It learns the input–output relations of the example set and subsequently is able to assign an output to a novel input. Such generalization ability makes it possible to use neural networks for classification and regression tasks.

Since a simple perceptron was studied by Sompolinsky *et al* and György *et al* [1–5] there have been many studies on neural networks with various architectures, such as multilayer perceptrons and modular networks. Investigation of neural network generalization has been successfully performed using statistical mechanics tools. With this approach learning is regarded as a stochastic process generating a Gibbs distribution in the network weight-space. Kang *et al* [3] showed that symmetry-breaking phenomena exist in the learning of a committee machine. Such symmetry-breaking phenomena have been found repeatedly in multilayer networks.

In recent years there has been significant interest in on-line learning of neural networks and learning dynamics.

In on-line learning, examples are given one by one and the weights are updated according to the most recent example. On-line learning studies can give a better understanding of realistic situations because of the obvious similarity to a practical implementation of a back-propagation algorithm. On-line learning is described as a dynamic evolution of weights, and the generalization error is calculated numerically by solving the differential equations of the order parameters. Since Biehl and Schwarze's work [6] on perceptrons and simplified models of multilayer networks, the dynamics of on-line learning has been widely investigated [7–11]. Saad *et al* showed that a soft-committee machine can be analysed using this on-line learning approach and since then much related research has been carried out.

Among the various kinds of multilayer networks, modular networks are of considerable interest due to their productive generalization capability [4, 5, 12–15]. In a modular network, parts of the neural network become modules in a larger structure. A complex task is divided into relatively simple ones and each of the resulting subtasks can be solved effectively by a module. The mixture-of-experts system is a well known example that elegantly implements the philosophy of divide-and-conquer [4, 12, 13]. Some gating networks divide tasks into smaller ones, and the subtasks are then assigned to the appropriate experts. Such a strategy can be a powerful tool for solving a mixed task with different local rules. Whilst modular networks have been mostly interpreted in the framework of statistics, a statistical physics approach has also been found to be successful [4, 5]. In a previous paper, we studied the generalization capability of a mixture of experts using equilibrium statistical physics [4]. The learning curve as a function of off-line examples shows an interesting phase transition that is related to permutation-symmetry breaking.

Here we present theoretical results for on-line learning of a mixture of experts. We find that the learning curve has interesting features, such as a plateau that is related to a permutation-symmetric fixed point similar to the one that was found in the learning curves of a soft-committee machine. We study the behaviour of the learning curves according to the stiffness of the gating function, and find that the stiffness is an important parameter determining the characteristics of learning, both in the symmetric and the broken-symmetry states.

In section 2, the model and the on-line learning rules of a mixture-of-experts network are introduced. The generalization error and overlap-order parameters are also defined. In section 3, we investigate the properties of this network using specific examples and analytic results are compared with simulated results. In section 4 we discuss the results and propose future studies that could lead to a better understanding of mixture-of-experts systems.

2. Learning and generalization

In on-line learning, the weights of a student network are updated following the error gradient corresponding to the latest in a sequence of examples. The student network is trained using the examples given by a teacher network with the same architecture as the student network. The resulting changes in the weights are represented as difference equations of order parameters that show the degree of overlap between the weights of the student network and the teacher network. In the thermodynamic limit, where the input dimension goes to infinity, these difference equations can be considered as differential equations, and the examples play the role of continuous time. Solving these differential equations leads to the generalization ability of the student network.

2.1. Model and learning rules

The mixture-of-experts system [12] is a tree consisting of expert and gating networks. The expert networks sit on the leaves of the tree, whilst the gating networks sit at the tree's branching points and assign weights to the outputs of the experts. For simplicity, we consider a network with one gating network and two experts. The output of each expert is given by

$$\mu_j = f(\boldsymbol{W}_j \cdot \boldsymbol{x}) \qquad j = 1, 2 \tag{1}$$

where $\boldsymbol{W}_j$ is the weight vector of the jth expert, and $f(x) = \mathrm{erf}(x/\sqrt{2})$ is a nonlinear, differentiable, transfer function. The principle of divide-and-conquer is implemented by assigning each expert to the subspace of an input space with different local rules. A gating network partitions the input space and assigns each expert a weighing factor:

$$\nu_j = g(\boldsymbol{V}_j \cdot \boldsymbol{x}) \qquad j = 1, 2 \tag{2}$$

where $\boldsymbol{V}_j$ is a weight vector corresponding to the jth output node of the gating network. Here we use the gating function $g(x) = \frac{1}{2}(1 + \mathrm{erf}(d\frac{x}{\sqrt{2}}))$ and d controls the stiffness of the gating function. For two experts, this gating function provides a boundary between the two subspaces that is perpendicular to the vector $\boldsymbol{V}_1 = -\boldsymbol{V}_2 = \boldsymbol{V}$. The sum of the weights is unity and the output of the gating network can be regarded as the probability of selecting each expert. If the slope d is large, the function forms a sharp boundary between the two subspaces. If d is small, the boundary is rather soft and there is an intermediate region where both expert networks contribute. In the limit, as d goes to zero, $\nu_1 = \nu_2 = 1/2$ results for each input $\boldsymbol{x}$. The network then has the same properties as a soft-committee machine [7]. Therefore, d regulates the overlap between subspaces. Now, the weighted output from the mixture of experts is written as

$$\sigma(\boldsymbol{V}, \{\boldsymbol{W}_j\}; \boldsymbol{x}) = \sum_{j=1}^{2} g(\boldsymbol{V}_j \cdot \boldsymbol{x}) f(\boldsymbol{W}_j \cdot \boldsymbol{x}). \tag{3}$$

The network learns rules from training examples generated by a teacher network that has the same architecture:

$$\sigma^0(\boldsymbol{V}^0, \{\boldsymbol{W}_j^0\}; \boldsymbol{x}) = \sum_{j=1}^{2} g(\boldsymbol{V}_j^0 \cdot \boldsymbol{x}) f(\boldsymbol{W}_j^0 \cdot \boldsymbol{x}) \tag{4}$$

where $\boldsymbol{V}_j^0$ and $\boldsymbol{W}_j^0$ are the jth weight vectors of the gating network and of the expert of the teacher network respectively. When using a mixture-of-experts system learning has probabilistic interpretations, where the learning algorithm is considered as a maximum-likelihood estimation. Statistical methods such as the expectation-maximization (EM) algorithm are often used. However, if we assume a Gaussian distribution, a maximum-likelihood estimation can be made by minimizing the usual quadratic-error function

$$\epsilon(\boldsymbol{V}, \{\boldsymbol{W}_j\}; \boldsymbol{x}) = \tfrac{1}{2}[\sigma(\boldsymbol{V}, \{\boldsymbol{W}_j\}; \boldsymbol{x}) - \sigma^0(\boldsymbol{V}^0, \{\boldsymbol{W}_j^0\}; \boldsymbol{x})]^2. \tag{5}$$

The on-line learning rule for the student weight vectors [6] $\boldsymbol{V}_j$ and $\boldsymbol{W}_j$ is written as

$$\boldsymbol{V}^{p+1} = \boldsymbol{V}^p - \frac{\eta}{N}[\sigma(\boldsymbol{V}^p, \{\boldsymbol{W}_j^p\}; \boldsymbol{x}^p) - \sigma^0(\boldsymbol{V}^0, \{\boldsymbol{W}_j^0\}; \boldsymbol{x}^p)] g'(t^p)[f(y_1^p) - f(y_2^p)]\boldsymbol{x}^p \tag{6}$$

$$\boldsymbol{W}_j^{p+1} = \boldsymbol{W}_j^p - \frac{\eta}{N}[\sigma(\boldsymbol{V}^p, \{\boldsymbol{W}_j^p\}; \boldsymbol{x}^p) - \sigma^0(\boldsymbol{V}^0, \{\boldsymbol{W}_j^0\}; \boldsymbol{x}^p)] f'(y_j^p) g((-1)^{j-1} t^p)\boldsymbol{x}^p \tag{7}$$

where $y_j^p = \boldsymbol{W}_j^p \cdot \boldsymbol{x}^p$ and $t^p = \boldsymbol{V}^p \cdot \boldsymbol{x}^p$ are the internal fields of the jth expert, and the gating network for the pth pattern, respectively. The learning rate η is scaled with the network size N. In the thermodynamic limit where N goes to infinity, we may consider $\alpha = p/N$ as a continuous time. Equations (6) and (7) can be transformed to differential equations for the corresponding order parameters. The generalization error is obtained from the numerical solutions of these differential equations.

2.2. Generalization error and overlap-order parameters

We measure the performance of the student network by the generalization error of this network. The generalization error is defined as

$$\epsilon_g(\boldsymbol{V}, \{\boldsymbol{W}_j\}) = \langle \epsilon(\boldsymbol{V}, \{\boldsymbol{W}_j\}; \boldsymbol{x}) \rangle_{\boldsymbol{x}} \tag{8}$$

where $\langle \cdots \rangle_{\boldsymbol{x}}$ represents an average over all possible input vectors. Input vectors are drawn independently from the Gaussian distribution with zero mean and unit variance.

After equation (8) is averaged over the input, the generalization error $\epsilon_g(\boldsymbol{V}, \boldsymbol{W}_j)$ can be expressed as a function of the overlap-order parameters between the student and the teacher weight vectors:

$$\begin{aligned}\epsilon_g(\boldsymbol{V},\{\boldsymbol{W}_j\}) = \frac{1}{2\pi^2}\Bigg[\Bigg[&\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)+\frac{\pi}{2}\right\}\\ &\times\left\{\sum_{i=1}^{2}\sin^{-1}\left(\frac{Q_i}{1+Q_i}\right)-2\sin^{-1}\left(\frac{q}{\sqrt{(1+Q_1)(1+Q_2)}}\right)\right\}\\ &-2\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)+\frac{\pi}{2}\right\}\\ &\times\left\{\sum_{i=1}^{2}\sin^{-1}\left(\frac{R_i}{\sqrt{2(1+Q_i)}}\right)-\sum_{i=1}^{2}\sin^{-1}\left(\frac{r_i}{\sqrt{2(1+Q_i)}}\right)\right\}\\ &+2\pi\left\{\sin^{-1}\left(\frac{q}{\sqrt{(1+Q_1)(1+Q_2)}}\right)-\sum_{i=1}^{2}\sin^{-1}\left(\frac{r_i}{\sqrt{2(1+Q_i)}}\right)\right\}\\ &+\frac{\pi}{3}\left\{\sin^{-1}\left(\frac{d_{\mathrm{t}}^2}{1+d_{\mathrm{t}}^2}\right)+\frac{\pi}{2}\right\}\Bigg]\Bigg]\end{aligned}\qquad(9)$$

where the overlap-order parameters are defined as

$$\begin{array}{lll} R_j = \boldsymbol{W}_j\cdot\boldsymbol{W}_j^0 & Q_j = \boldsymbol{W}_j\cdot\boldsymbol{W}_j & \\ r_1 = \boldsymbol{W}_1\cdot\boldsymbol{W}_2^0 & r_2 = \boldsymbol{W}_2\cdot\boldsymbol{W}_1^0 & q = \boldsymbol{W}_1\cdot\boldsymbol{W}_2 \\ S = \boldsymbol{V}\cdot\boldsymbol{V}^0 & T = \boldsymbol{V}\cdot\boldsymbol{V}. & \end{array}$$

In the thermodynamic limit, α/N can be considered as a continuous variable. The subsequent difference equations (6) and (7), can be rewritten as differential equations using the overlap-order parameters:

$$\begin{array}{ll} \dfrac{\mathrm{d}R_j}{\mathrm{d}\alpha} = \eta\langle\delta_j z_j\rangle & \dfrac{\mathrm{d}Q_j}{\mathrm{d}\alpha} = 2\eta\langle\delta_j y_j\rangle + \eta^2\langle\delta_j^2\rangle \\ \dfrac{\mathrm{d}r_1}{\mathrm{d}\alpha} = \eta\langle\delta_1 z_2\rangle & \dfrac{\mathrm{d}r_2}{\mathrm{d}\alpha} = \eta\langle\delta_2 z_1\rangle \\ \dfrac{\mathrm{d}q}{\mathrm{d}\alpha} = \eta\langle\delta_2 y_1\rangle + \eta\langle\delta_1 y_2\rangle + \eta^2\langle\delta_1\delta_2\rangle & \\ \dfrac{\mathrm{d}S}{\mathrm{d}\alpha} = \eta\langle\delta s\rangle & \dfrac{\mathrm{d}T}{\mathrm{d}\alpha} = 2\eta\langle\delta t\rangle + \eta^2\langle\delta^2\rangle \end{array}\qquad(10)$$

where the internal fields of the teacher network z_j, s and other related functions are defined as

$$\begin{aligned}\delta_j &= [\sigma^0(\boldsymbol{V}^0,\{\boldsymbol{W}_j^0\};\boldsymbol{x})-\sigma(\boldsymbol{V},\{\boldsymbol{W}_j\};\boldsymbol{x})]f'(y_j)g((-1)^{j-1}t)\\ \delta &= [\sigma^0(\boldsymbol{V}^0,\{\boldsymbol{W}_j^0\};\boldsymbol{x})-\sigma(\boldsymbol{V},\{\boldsymbol{W}_j\};\boldsymbol{x})]g'(t)(f(y_1)-f(y_2))\\ z_j &= \boldsymbol{W}_j^0\cdot\boldsymbol{x} \qquad s = \boldsymbol{V}^0\cdot\boldsymbol{x}.\end{aligned}$$

For the order parameter R_j, r_j and S, the average in equation (10) can be calculated exactly, as seen from equations (A.1), (A.2) and (A.5) in the appendix. Here, we consider the condition that the learning rate η is small, so that the equations for the order parameters Q_j, q and T can be approximated by the first-order terms of η. We can subsequently calculate analytical solutions for these differential equations, as seen from equations (A.3), (A.4) and (A.6) in the appendix. The time evolution of the generalization error, equation (9), can be calculated using the numerical values of the order parameters.

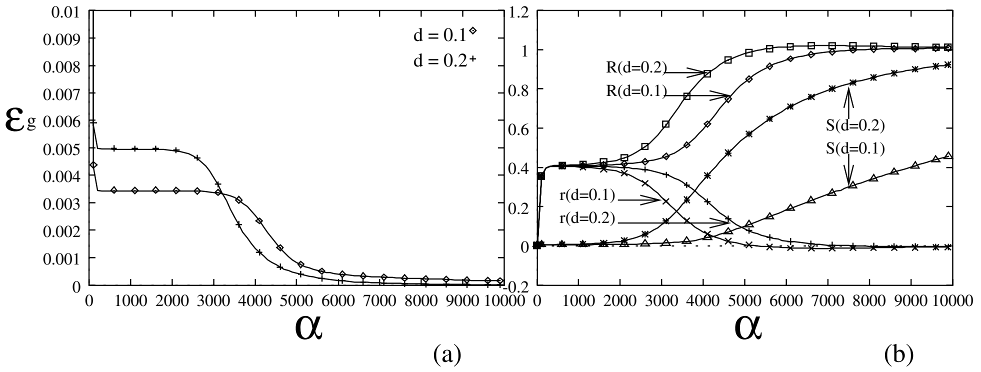

Figure 1. Learning curves of the mixture-of-experts system with $d = 0.1$ and $d = 0.2$: lines show the analytical results and symbols show the simulated results. (a) The generalization error (ϵ_g versus α for the learning rate $\eta = 0.1$ with network size $N = 100$). (b) Time evolution of the overlapping order parameters R_1, r_1 and S.

From these order parameters and the generalization error, we find the typical behaviour of learning curves in on-line learning of the mixture-of-experts system. In the next section, we present several examples and compare these analytical results with simulated results.

3. Results

3.1. The case with the same stiffness ($d_s = d_t = d$)

We consider the case where the stiffnesses of the gating functions of the student and teacher networks are the same ($d_s = d_t$), and show learning curves from analytical calculations and from numerical simulation for a large network. The simulations were performed with a network of size $N = 100$. We find that the analytical results agree well with the simulated results.

Figure 1(a) shows the generalization error as a function of α with learning rate $\eta = 0.1$. The behaviour of the order parameters is shown in figure 1(b). Figure 1(a) shows two states of the learning process: the permutation-symmetric state; and the broken-symmetry state. Before the symmetry breaks, the generalization error converges to a relatively high value, and the curve shows a plateau where the error remains constant with increasing α. In this symmetric state, the order parameters S and T are nearly zero, and the order parameters related to the experts show symmetric behaviour ($R_i = r_i$ and $Q_i = q$). Therefore, the gating network has no effect on the learning, and the role of each expert has not yet become specialized.

A mixture-of-experts system does not use the advantages of a modular network in the symmetric phase. When α reaches a critical point, the symmetry between experts begins to break and the gating network learns how to divide the input space. As α increases, the order parameters S and T approach 1, indicating that the gating network has learnt the rule perfectly. The order parameters R_i and r_i branch at the critical point and approach unity and zero, respectively. This result reveals that the gating network divides the input space appropriately and assigns each expert the corresponding subspace. In the broken-symmetry state, each expert learns its local rule and this mixture-of-experts system successfully performs the divide-and-conquer strategy.

We next compare two networks that have different stiffness. The network with a larger d approaches the critical point at smaller α and the generalization error is smaller in the large α

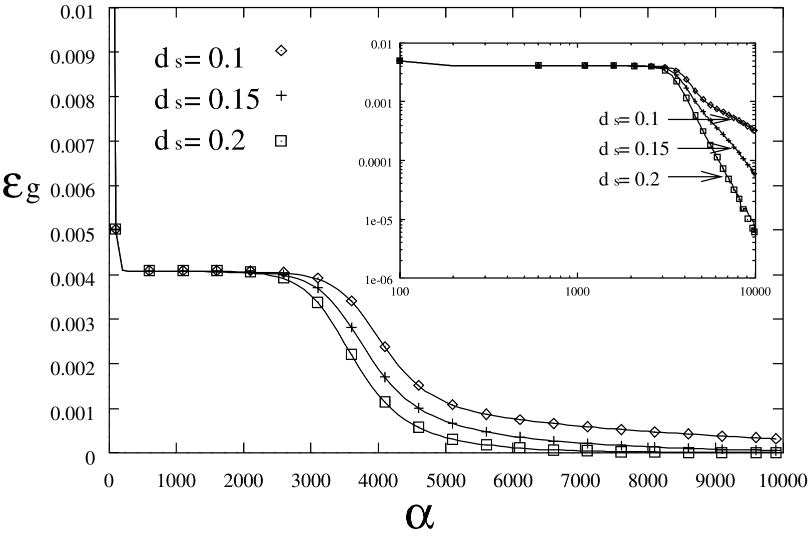

Figure 2. Learning curves of the mixture-of-experts system with $d_s = 1.0$, $d_s = 1.5$ and $d_s = 2.0$ for fixed $d_t = 0.15$, where d_s and d_t mean the stiffness of a student and of a teacher gating network, respectively. Lines show the analytic results and symbols show the simulated results. (Inset: the same curves are plotted using a log scale.)

limit. This is related to the role of the gating network in the learning process of the mixture-of-experts system. When d is large, the subspaces assigned to experts are easy to discriminate, so that the broken-symmetry state appears at small α and the generalization error is also small.

3.2. The case with different stiffness ($d_s \neq d_t$)

We now study examples where the stiffnesses of the gating functions of the student and teacher networks are different ($d_s \neq d_t$). These examples show the effect of the mixture-of-experts gating function when we do not know the stiffness of the teacher gating network.

In figure 2, we see that the learning curves show different behaviour according to the stiffness of the student network, although the teacher network is the same in all cases. In the symmetric state, the generalization error is the same, but the length of the plateau becomes shorter as d_s increases. As explained above, the gating networks of the students have no effect on learning in the symmetric state, so that the generalization error at the plateau is solely determined by the teacher network. When d is large, the critical point at the beginning of the broken-symmetry state appears when α is small because student networks with large d easily discriminate the subspace assigned to each expert. In the broken-symmetry state, as shown in the inset of figure 2, the learning curve becomes steeper as d_s increases.

From these results, we find that student networks with larger d_s learn the rules of the teacher network better. We also find, however, the existence of local minima in the learning of a mixture-of-experts system with large d, indicating that the selection of an optimal stiffness is necessary for efficient learning.

3.3. Local minima

To understand the nature of local minima, we explicitly analyse the time evolution of a small network. The network size is $N = 3$ and the initial weights are randomly chosen from a Gaussian

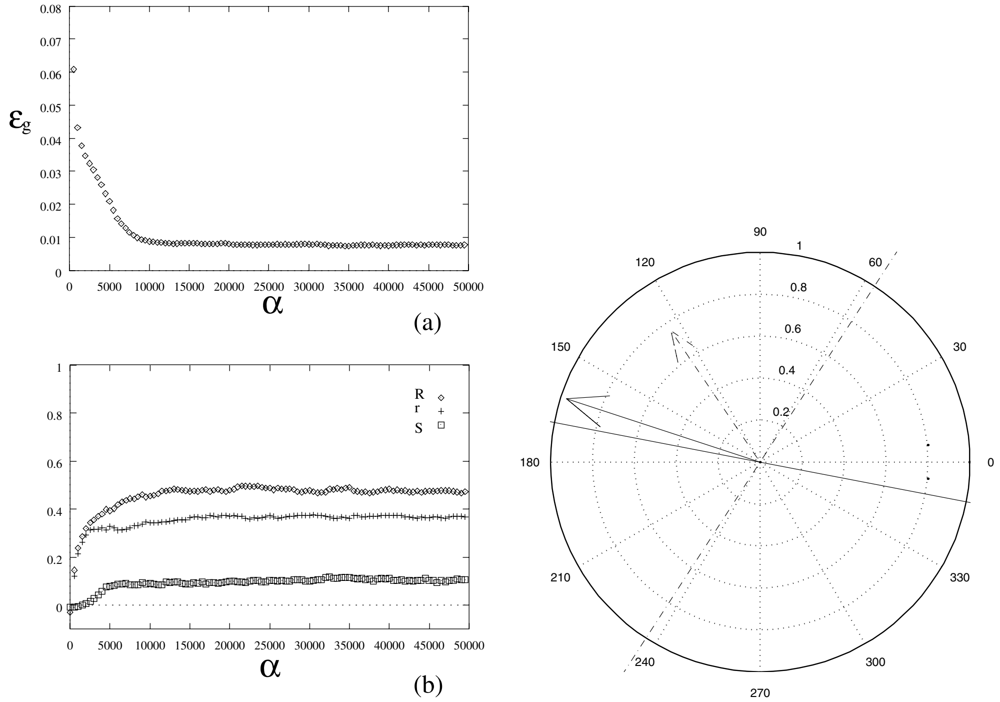

Figure 3. Learning curves of the mixture-of-experts system trapped in a local minimum (learning rate $\eta = 0.005$ with network size $N = 3$ and $d_s = d_t = 1.0$): a circle indicates a weight space projected into two dimensions. In the circle, the arrow with a solid line is the weight vector of the teacher expert and the arrow with a dash-dot line is that of the student expert. The solid line and the dash-dot line are the boundaries between subspaces given by the gating network of the teacher and the student, respectively. (*a*) The generalization error (ϵ_g versus α). (*b*) The overlapping order parameters R_1, r_1 and S are shown for the same case.

distribution. In figure 3, we show an example trapped in a local minimum, where the generalization error approaches a value larger than zero, and several different broken symmetry fixed points are also found. Each fixed point corresponds to a different final value of the residual error.

The circle in figure 3 is a weight space projected into two dimensions. Lines in the circle are the boundaries between two subspaces assigned to the experts, which are determined by the gating networks of the student and the teacher. Different boundaries show that the student gating network does not learn the rule of the teacher gating network. The arrows are the weight vectors of the student and the teacher. Since the student gating network assigns the wrong subspaces to the experts, the student experts cannot learn the rules of the corresponding teacher experts. Thus, the learning curve does not approach the perfect broken-symmetry state, as shown in figures 3(*a*) and (*b*).

Local minima occur in cases that have the same or different stiffnesses of the gating functions. These phenomena may be explained by the analysis of fixed points and their stabilities in the dynamics of the order parameters. When the slope d is small, the forms of the output function and the error surface are smooth. Only the global minimum becomes a stable fixed point in the learning of a mixture-of-experts system when d is small. As the stiffness increases, other stable fixed points, i.e. local minima, occur.

4. Conclusion

We find the analytical solutions for on-line learning of a mixture-of-experts system, and apply these models to several examples. We find that a symmetric and a broken-symmetry state exist in a learning curve. The generalization error and the overlap-order parameters show plateaus in the learning curves, as is the case for a soft-committee machine. However, in the mixture-of-experts system the symmetry is broken by the gating network, which assigns the appropriate subspaces to the experts.

These symmetric phenomena are also found in batch learning of a mixture-of-experts system. In batch learning, however, it is difficult to study the effects of gating networks in the training of a mixture-of-experts system. As a result, we must study on-line learning to analyse the role of the gating network.

The effects of the learning rate have previously been investigated mainly in on-line learning of neural networks with various architectures, i.e. a simple perceptron, a soft-committee machine and a radial basis function (RBF) network. In a mixture-of-experts system, however, the stiffness of the gating function is an important parameter that determines the behaviour of the learning curves, as well as the learning rate. In this paper, we investigated the role of the gating function in the on-line learning of a mixture-of-experts system.

In cases where a student and a teacher gating network have the same or different stiffnesses, the behaviour of the learning curve and the critical point both depend on the magnitude of the slope d. The properties in the symmetric state are determined by the teacher network. In the broken-symmetry state, the student network determines the steepness of the learning curve. We find that the student network can easily learn the rules when d is large (the overlap between subspaces assigned to experts is small). This reveals that the stiffness of the gating function is important in the learning of the mixture-of-experts system. As local minima may occur when d is large, the selection of optimal stiffness is essential in the learning of a mixture-of-experts system.

To find an optimal stiffness of a gating function, it is necessary to investigate the locations of local minima and their stabilities using analysis of fixed points in the dynamics of the order parameters. In future work we will study the analysis of the fixed points in order to find optimal parameters for efficient learning. This study may be expanded to a piece-wise linear model with a linear activation unit for an expert network. It would be interesting to study a hierarchical mixture of experts, which is applicable to more complex tasks.

Acknowledgment

This research was funded by the Brain Science & Engineering Research Programme (the Ministry of Science and Technology) in Korea.

Appendix

The order parameters related to experts are R_i, r_i, Q_i and q. The overlaps R_i and r_i between a teacher and a student expert are given by

$$\frac{\mathrm{d}R_i}{\mathrm{d}\alpha} = \frac{\eta}{\pi^2(1+Q_i)}\left[\left\{\sin^{-1}\left(\frac{d_\mathrm{s}^2 T}{1+d_\mathrm{s}^2 T}\right)+\frac{\pi}{2}\right\}\frac{-R_i}{\sqrt{1+2Q_i}}\right.$$
$$+\left\{\sin^{-1}\left(\frac{d_\mathrm{s}d_\mathrm{t}S}{\sqrt{(1+d_\mathrm{t}^2)(1+d_\mathrm{s}^2 T)}}\right)+\frac{\pi}{2}\right\}\frac{1+Q_i-R_i^2}{\sqrt{2(1+Q_i)-R_i^2}}$$

$$\begin{aligned}
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)-\frac{\pi}{2}\right\}\frac{r_iR_i}{\sqrt{2(1+Q_i)-r_i^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)-\frac{\pi}{2}\right\}\frac{r_j(1+Q_i)-R_iq}{\sqrt{(1+Q_i)(1+Q_j)-q^2}}\Bigg]\qquad i\neq j \qquad\text{(A.1)}
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}r_i}{\mathrm{d}\alpha}=\frac{\eta}{\pi^2(1+Q_i)}\Bigg[\Bigg[&\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)+\frac{\pi}{2}\right\}\frac{-r_i}{\sqrt{1+2Q_i}}\\
&-\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)+\frac{\pi}{2}\right\}\frac{r_iR_i}{\sqrt{2(1+Q_i)-R_i^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)-\frac{\pi}{2}\right\}\frac{r_i^2-(1+Q_i)}{\sqrt{2(1+Q_i)-r_i^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)-\frac{\pi}{2}\right\}\frac{R_j(1+Q_i)-r_iq}{\sqrt{(1+Q_i)(1+Q_j)-q^2}}\Bigg]\qquad i\neq j. \qquad\text{(A.2)}
\end{aligned}$$

The magnitudes of the student weight vectors Q_i are given by

$$\begin{aligned}
\frac{\mathrm{d}Q_i}{\mathrm{d}\alpha}=\frac{2\eta}{\pi^2(1+Q_i)}\Bigg[\Bigg[&\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)+\frac{\pi}{2}\right\}\frac{-Q_i}{\sqrt{1+2Q_i}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)+\frac{\pi}{2}\right\}\frac{R_i}{\sqrt{2(1+Q_i)-R_i^2}}\\
&-\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)-\frac{\pi}{2}\right\}\frac{r_i}{\sqrt{2(1+Q_i)-r_i^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)-\frac{\pi}{2}\right\}\frac{q}{\sqrt{(1+Q_i)(1+Q_j)-q^2}}\Bigg]. \qquad\text{(A.3)}
\end{aligned}$$

The overlap q between weight vectors of different student experts is given by

$$\begin{aligned}
\frac{\mathrm{d}q}{\mathrm{d}\alpha}=\frac{\eta}{\pi^2(1+Q_1)}\Bigg[\Bigg[&\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)+\frac{\pi}{2}\right\}\frac{-q}{\sqrt{1+2Q_1}}\\
&-\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)+\frac{\pi}{2}\right\}\frac{-r_2(1+Q_1)+qR_1}{\sqrt{2(1+Q_1)-R_1^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)-\frac{\pi}{2}\right\}\frac{-R_2(1+Q_1)+qr_1}{\sqrt{2(1+Q_1)-r_1^2}}\\
&-\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)-\frac{\pi}{2}\right\}\frac{-Q_2(1+Q_1)+q^2}{\sqrt{(1+Q_1)(1+Q_2)-q^2}}\Bigg]\\
&+\frac{\eta}{\pi^2(1+Q_2)}\Bigg[\Bigg[\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)+\frac{\pi}{2}\right\}\frac{-q}{\sqrt{1+2Q_2}}\\
&-\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)+\frac{\pi}{2}\right\}\frac{-r_1(1+Q_2)+qR_2}{\sqrt{2(1+Q_2)-R_2^2}}\\
&+\left\{\sin^{-1}\left(\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)}}\right)-\frac{\pi}{2}\right\}\frac{-R_1(1+Q_2)+qr_2}{\sqrt{2(1+Q_2)-r_2^2}}
\end{aligned}$$

$$-\Bigg\{\sin^{-1}\left(\frac{d_{\mathrm{s}}^2T}{1+d_{\mathrm{s}}^2T}\right)-\frac{\pi}{2}\Bigg\}\frac{-Q_1(1+Q_2)+q^2}{\sqrt{(1+Q_1)(1+Q_2)-q^2}}\Bigg]. \tag{A.4}$$

The order parameters related to gating networks are S and T. The overlap S between a teacher and a student gating network is given by

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}\alpha}=\frac{\eta}{\pi^2(1+d_s^2T)}\Bigg[\Bigg[\Bigg\{\sum_{i=1}^{2}\left(\sin^{-1}\left(\frac{R_i}{\sqrt{2(1+Q_i)}}\right)-\sin^{-1}\left(\frac{r_i}{\sqrt{2(1+Q_i)}}\right)\right)\Bigg\}\\ \times\frac{d_{\mathrm{s}}d_{\mathrm{t}}(1+d_{\mathrm{s}}^2T)-d_{\mathrm{s}}^2d_{\mathrm{t}}^2S^2}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)-d_{\mathrm{s}}^2d_{\mathrm{t}}^2S^2}}\\ +\Bigg\{\sum_{i=1}^{2}\sin^{-1}\left(\frac{Q_i}{1+Q_i}\right)-2\sin^{-1}\left(\frac{q}{\sqrt{(1+Q_1)(1+Q_2)}}\right)\Bigg\}\frac{-d_{\mathrm{s}}^2S}{\sqrt{1+2d_{\mathrm{s}}^2T}}\Bigg].\end{aligned} \tag{A.5}$$

The magnitude of a student weight vector T is given by

$$\begin{aligned}\frac{\mathrm{d}T}{\mathrm{d}\alpha}=\frac{2\eta}{\pi^2(1+d_s^2T)}\Bigg[\Bigg[\Bigg\{\sum_{i=1}^{2}\left(\sin^{-1}\left(\frac{R_i}{\sqrt{2(1+Q_i)}}\right)-\sin^{-1}\left(\frac{r_i}{\sqrt{2(1+Q_i)}}\right)\right)\Bigg\}\\ \times\frac{d_{\mathrm{s}}d_{\mathrm{t}}S}{\sqrt{(1+d_{\mathrm{t}}^2)(1+d_{\mathrm{s}}^2T)-d_{\mathrm{s}}^2d_{\mathrm{t}}^2S^2}}\\ +\Bigg\{\sum_{i=1}^{2}\sin^{-1}\left(\frac{Q_i}{1+Q_i}\right)-2\sin^{-1}\left(\frac{q}{\sqrt{(1+Q_1)(1+Q_2)}}\right)\Bigg\}\frac{-d_{\mathrm{s}}^2T}{\sqrt{1+2d_{\mathrm{s}}^2T}}\Bigg].\end{aligned} \tag{A.6}$$

References

[1] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056

[2] Györgyi G 1990 *Phys. Rev. Lett.* **64** 2957
Györgyi G 1990 *Phys. Rev.* A **41** 7097

[3] Kang K and Oh J-H 1993 *Phys. Rev.* E **48** 4805

[4] Oh J-H and Kang K 1997 *Advances in Neural Information Processing Systems* vol 9 (Cambridge, MA: MIT Press) p 183

[5] Kang K and Oh J-H 1997 *Phys. Rev.* E **55** 3257

[6] Biehl M and Schwarze H 1995 *J. Phys. A: Math. Gen.* **28** 643

[7] Saad D and Solla S A 1995 *Phys. Rev. Lett.* **74** 4337

[8] Saad D and Rattray M 1998 *Phys. Rev.* E **57** 2170

[9] Rattray M and Saad D 1998 *Phys. Rev.* E **58** 6379

[10] Sompolinsky H and Kim J W 1998 *Phys. Rev.* E **58** 2335

[11] Kim J W and Sompolinsky H 1998 *Phys. Rev.* E **58** 2348

[12] Jordan M I and Jacobs R A 1994 *Neural Comput.* **6** 181

[13] Jacobs R A, Peng F and Tanner M A 1997 *Neural Network* **10** 231

[14] Krogh A and Sollich P 1997 *Phys. Rev.* E **55** 811

[15] Wolpert D 1992 *Neural Network* **5** 241