Online learning dynamics of multilayer perceptrons with unidentifiable parameters

2003 J. Phys. A: Math. Gen. 36 11753

(http://iopscience.iop.org/0305-4470/36/47/004)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 11753–11764 PII: S0305-4470(03)63698-X

Online learning dynamics of multilayer perceptrons with unidentifiable parameters

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

[1] Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[2] Department of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan
[3] 'Intelligent Cooperation and Control', PRESTO, JST, c/o RIKEN BSI, Saitama 351-0198, Japan

E-mail: hypark@brain.riken.go.jp, minoue@brain.riken.go.jp and okada@brain.riken.go.jp

Received 20 May 2003
Published 12 November 2003
Online at stacks.iop.org/JPhysA/36/11753

Abstract

In the over-realizable learning scenario of multilayer perceptrons, in which the student network has a larger number of hidden units than the true or optimal network, some of the weight parameters are unidentifiable. In this case, the teacher network consists of a union of optimal subspaces included in the parameter space. The optimal subspaces, which lead to singularities, are known to affect the estimation performance of neural networks. Using statistical mechanics, we investigate the online learning dynamics of two-layer neural networks in the over-realizable scenario with unidentifiable parameters. We show that the convergence speed strongly depends on the initial parameter conditions. We also show that there is a quasi-plateau around the optimal subspace, which differs from the well-known plateaus caused by permutation symmetry. In addition, we discuss the property of the final learning state, relating this to the singular structures.

PACS number: 05.20.−y

1. Introduction

When we design a neural network for approximating a true function that generates observed samples, we need to choose an appropriate network size. Since the optimal network size is unknown, a sufficiently large network is usually used as the first stage. For this reason, the learning network in practical applications, which we call a student network, is larger than the optimal or true network. This is called an over-realizable scenario. When we use a multilayer perceptron, an over-realizable scenario means the case that the number of hidden nodes in the student network exceeds the optimal number of hidden nodes.

The over-realizable learning scenario of multilayer perceptrons connotes a severe problem, which is called a problem of unidentifiability or a singularity problem [1]. In the over-realizable case, the optimal network is represented not by a unique point but by a union of subspaces in the parameter space of the student network, and thus the optimal parameter is unidentifiable. In addition, all the points in the optimal subspaces are singular, at which the Fisher information matrix degenerates. For this reason, the classical statistical theory cannot be applied for analysing the learning behaviour of multilayer perceptrons in the over-realizable scenario. Since the optimal estimator is not a point, the asymptotic normality about an optimal estimator is destroyed. Since the Fisher information matrix degenerates at the optimum, the Cramér–Rao paradigm cannot be applied to analyse the estimation performance of the network.

Recently, the importance of this singularity problem has become apparent, and a number of studies on building a new statistical theory for nonregular (singular) models have been performed. Hagiwara *et al* [2] have showed that the classical model selection criterion, the Akaike information criterion (AIC), does not work on multilayer perceptrons, and suggested that this is due to the singularity of the optimal point. Hagiwara [3] used simple models to show that the least square error of the estimator does not obey the conventional asymptotic rule. Fukumizu [4] gave a general analysis of the maximum likelihood estimators in singular statistical models that include multilayer perceptrons. Watanabe [5] applied algebraic geometry to elucidate the behaviour of the Bayesian predictive estimator of multilayer perceptrons in the over-realizable scenario. These studies have shown that the properties of networks in the singular (over-realizable) case are strictly different from those in the regular (realizable and unrealizable) case.

Until now, however, most of the work related to the singularity problem has been done through asymptotic statistical analysis of the estimators. Although theoretical analysis of network models is indispensable, it is also important to investigate the learning behaviour of network models. In practical applications, the finally obtained parameters often depend on the learning trajectories of the learning network. In the over-realizable scenario, some parameters are unidentifiable, and the optimal destination of learning dynamics is not one point, but some subspaces in the parameter space. Therefore, the dynamics has much more freedom than in the realizable and unrealizable scenarios, and clarifying the network behaviour is an interesting objective.

In this paper, we investigate the dynamics of the online learning of multilayer perceptrons in an over-realizable scenario with unidentifiable optimal parameters. To do this, we take the statistical mechanical approach [6, 7], which can be used to analyse the dynamics of gradient descent learning algorithms at the large limit of the input dimension. Using such a framework, Saad and Solla [7] have shown that the plateaus in the learning of soft committee machines are closely related to the singularity in the parameter space. Inoue *et al* [8] also analysed the relationship between the length of plateaus and the position of the optimal point. Especially, Biehl and Schwarze [9] and Biehl *et al* [10] discussed the dynamics of the over-realizable scenario. However, these studies were done using soft committee machines, which always have a unique optimal point in the parameter space and are not involved in the problem of unidentifiability. Riegler and Biehl [6] investigated the dynamics of general multilayer perceptrons, but they mainly investigated the realizable case, in which the optimal point is uniquely determined. Even though many interesting dynamical properties of multilayer perceptrons have been investigated through these studies, the dynamics related to the intrinsic singularity in the parameter space has not been dealt with so far. The work reported here is the first step towards solving the challenging problem.

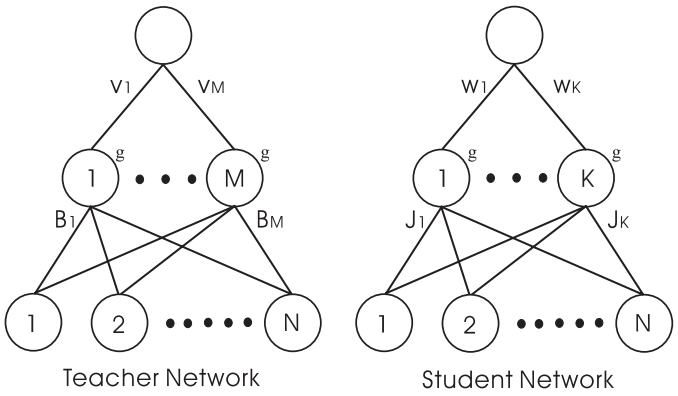

Figure 1. Architecture of the teacher and student networks.

2. Multilayer perceptrons with unidentifiable parameters

The multilayer perceptron discussed in this paper is defined as

$$\zeta' = f_{\boldsymbol{J},w}(\boldsymbol{\xi}) = \sum_{i=1}^{K} w_i g(\boldsymbol{J}_i \cdot \boldsymbol{\xi}). \tag{1}$$

Here, $\boldsymbol{\xi} \in \Re^N$ denotes the input vector; $\boldsymbol{J}_i \in \Re^N$ and $w_i \in \Re$ denote the weight parameters connected to the ith hidden unit; N denotes the number of input nodes and $g(\cdot)$ denotes an activation function. We assumed that the true function generating the training data $(\boldsymbol{\xi}, \zeta)$ can also be described by a teacher multilayer perceptron with the same architecture, such as

$$\zeta = f_{\boldsymbol{B},v}(\boldsymbol{\xi}) = \sum_{n=1}^{M} v_n g(\boldsymbol{B}_n \cdot \boldsymbol{\xi}) \tag{2}$$

where $\boldsymbol{B}_n$ and v_n are the true (optimal) parameters to be estimated through learning. Figure 1 shows the architecture of the networks.

When the number of hidden nodes of the student network exceeds that of the true network, $K > M$, $K - M$ hidden units in the student network are redundant, and some parameters related to the redundant units become unidentifiable. As a simple example, let us consider the case of $K = 2$, $M = 1$. The student network can realize the teacher network when one of the following conditions is satisfied:

Condition 1 $\quad \boldsymbol{J}_1 = \boldsymbol{J}_2 = \boldsymbol{B} \qquad w_1 + w_2 = v$

Condition 2 $\quad \begin{cases} \boldsymbol{J}_1 = \boldsymbol{B} & w_1 = v & w_2 = 0 \\ \boldsymbol{J}_2 = \boldsymbol{B} & w_2 = v & w_1 = 0. \end{cases}$

In the case of condition 1, parameters w_1 and w_2 are unidentifiable under the restriction of $w_1 + w_2 = v$. In the case of condition 2, parameter $\boldsymbol{J}_1$ or $\boldsymbol{J}_2$ is unidentifiable. Note here that this kind of unidentifiability does not exist in soft committee machines. Because w_1 and w_2 are set to a fixed value in soft committee machines, the optimal solution is uniquely determined as $\boldsymbol{J}_1 = \boldsymbol{B}$ and $\boldsymbol{J}_2 = \boldsymbol{0}$. (Here, we ignore the trivial permutation, $\boldsymbol{J}_1 = \boldsymbol{0}$ and $\boldsymbol{J}_2 = \boldsymbol{B}$.)

The parameters satisfying the optimal conditions constitute a set of optimal subspaces in the parameter space. Figure 2 shows a part of the optimal subspaces in the parameter

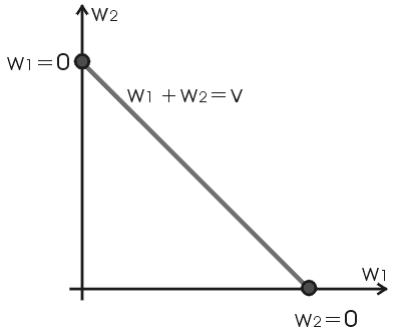

Figure 2. Optimal subspace in parameter space (w_1, w_2).

space (w_1, w_2). Condition 1 corresponds to the thick solid line in the figure, and condition 2 corresponds to the endpoints of the line as expressed by the two dots on the axes of w_1 and w_2. This kind of optimal subspace does not exist in the parameter space of soft committee machines, even though it affects learning dynamics seriously as we shall discuss in section 4.

Since the optimal parameters form subspaces, the possible goal of the learning trajectory is not unique, but can be any point within the optimal subspaces. In statistical analysis [2–5, 11], all the points in the subspaces are considered to represent the same conditional probability distribution, and the differences among them do not matter. However, condition 1 and condition 2 have intrinsically different singular structures in the parameter space, and thus influence the estimation properties differently. Therefore, when we take the dynamics into account, it is interesting to check which type of optimal condition is finally satisfied in the learning. In addition, since optimal condition 1 is closely related to the permutation symmetry that causes plateaus in the learning of soft committee machines, it is also interesting to see how the subspaces of optimal condition 1 influence the learning dynamics. In this paper, we discuss our investigation of the dynamical behaviour under an online learning scheme to clarify the properties of the learning trajectories.

3. Analysis of online learning dynamics

3.1. Gradient descent learning

We investigated the dynamics of standard gradient descent learning, which is also called backpropagation learning. At each learning step, new training data $(\boldsymbol{\xi}, \zeta)$ are generated from the teacher network. The parameter is updated to decrease the squared error,

$$e_{\boldsymbol{J},w}(\boldsymbol{\xi}, \zeta) = \frac{1}{2}[f_{\boldsymbol{B},v}(\boldsymbol{\xi}) - f_{\boldsymbol{J},w}(\boldsymbol{\xi})]^2 \tag{3}$$

$$= \frac{1}{2}\left[\sum_{i=1}^{K} w_i g(x_i) - \sum_{n=1}^{M} v_n g(y_n)\right]^2 \tag{4}$$

where

$$x_i = \boldsymbol{J}_i \cdot \boldsymbol{\xi} \qquad y_n = \boldsymbol{B}_n \cdot \boldsymbol{\xi}.$$

When we use the gradient descent learning method, the update term is given by

$$\Delta \boldsymbol{J}_i = -\frac{\eta}{N}\frac{\partial e_{\boldsymbol{J},w}}{\partial \boldsymbol{J}_i} = -\frac{\eta}{N}\delta_i \boldsymbol{\xi} \tag{5}$$

$$\Delta w_i = -\frac{\eta}{N}\frac{\mathrm{d}e_{\boldsymbol{J},w}}{\mathrm{d}w_i} = -\frac{\eta}{N} g(x_i)\left[\sum_{j=1}^{K} w_j g(x_j) - \sum_{n=1}^{M} v_n g(y_n)\right] \tag{6}$$

where

$$\delta_i = w_i g'(x_i)\left[\sum_{j=1}^{K} w_j g(x_j) - \sum_{n=1}^{M} v_n g(y_n)\right]$$

and η is the learning rate. The estimation accuracy of the network is evaluated using the generalization error, which is defined as

$$E^{\mathrm{gen}} = \left\langle \frac{1}{2}\{f_{\boldsymbol{B},w}(\boldsymbol{\xi}) - f_{\boldsymbol{J},w}(\boldsymbol{\xi})\}^2 \right\rangle_{\{\xi\}} \tag{7}$$

where $\langle\cdot\rangle_{\{\xi\}}$ denotes the expectation with respect to the input random value $\boldsymbol{\xi}$.

3.2. Statistical mechanics for analysing dynamics

To see the dynamics of learning, we need to trace the evolution of the weight parameters $\boldsymbol{J}_i$ and w_i. This is almost impossible when the input dimension N is large. In addition, the randomness of data makes the analysis of online learning dynamics difficult. To solve these difficulties, we use the statistical mechanical framework, and analyse the dynamics at the thermodynamic limit; i.e., the limit of $N \to \infty$. Within the framework, new order parameters are defined and used to describe the generalization error. Saad and Solla [7], Biehl and Schwarze [9] and Biehl *et al* [10] applied such a method to analyse the dynamics of gradient learning in soft committee machines and other simple two-layer networks. Riegler and Biehl [6] extended it to multilayer perceptrons. In this section, we briefly review the analysing method including the order parameters and their motion equations for the multilayer perceptron (1), which was derived in [6]. (For details, one can also refer to [12].)

To describe the learning dynamics, we can use the order parameters representing the correlations between weight vectors $\boldsymbol{J}_i$ and $\boldsymbol{B}_n$, instead of using the N-dimensional weight vector. The order parameters are defined as

$$Q_{ij} = \boldsymbol{J}_i \cdot \boldsymbol{J}_j \qquad i, j = 1, \ldots, K \tag{8}$$

$$R_{in} = \boldsymbol{J}_i \cdot \boldsymbol{B}_n \qquad i = 1, \ldots, K \quad n = 1, \ldots, M. \tag{9}$$

These parameters are updated through learning of the weight vectors, $\boldsymbol{J}_i (i = 1, \ldots, K)$. We also have similar values, $T_{nm} = \boldsymbol{B}_n \cdot \boldsymbol{B}_m$ $(n, m = 1, \ldots, M)$ which are determined by the teacher network and are fixed during learning. In the case of soft committee machines, these order parameters are sufficient to describe the learning dynamics and the generalization error. In the case of multilayer perceptrons, we need one more set of parameters w_i $(i = 1, \ldots, K)$, and fixed variables v_n $(n = 1, \ldots, M)$.

Under the assumption that all elements of the input vector are independent and identically different random variables with zero mean and unit variance, the generalization error at the thermodynamic limit can be determined using parameters R_{in}, Q_{ij} and w_i, and the fixed values $T_{i,j}$ and v_i. Furthermore, if we define the activation function g as $g(u) = \mathrm{erf}(u/\sqrt{2})$, we can get the explicit form of the generalization error of (7);

$$E^{\text{gen}} = \frac{1}{\pi}\left[\sum_{i,j}^{K} w_i w_j \arcsin\left(\frac{Q_{ij}}{\sqrt{1+Q_{ii}}\sqrt{1+Q_{jj}}}\right) + \sum_{m,n}^{M} v_m v_n \arcsin\left(\frac{T_{mn}}{\sqrt{1+T_{mm}}\sqrt{1+T_{nn}}}\right) - 2\sum_{i}^{K}\sum_{n}^{M} w_i v_n \arcsin\left(\frac{R_{in}}{\sqrt{1+Q_{ii}}\sqrt{1+T_{nn}}}\right)\right]. \tag{10}$$

As shown in (10), the time evolution of the generalization error can be obtained through the evolution of order parameters R_{in} and Q_{ij}, and parameter w_i. From the learning rule of $\boldsymbol{J}_i$ in (5) and the definition of the order parameters (8), (9), we can obtain

$$\Delta R_{in} = \Delta \boldsymbol{J}_i \cdot \boldsymbol{B}_n = -\frac{\eta}{N}\delta_i y_n \tag{11}$$

$$\Delta Q_{ij} = \Delta \boldsymbol{J}_i \cdot \boldsymbol{J}_j + \Delta \boldsymbol{J}_j \cdot \boldsymbol{J}_i + \Delta \boldsymbol{J}_i \cdot \Delta \boldsymbol{J}_j \tag{12}$$

$$= -\frac{\eta}{N}(\delta_i x_j + \delta_j x_i) + \frac{\eta^2}{N^2}\delta_i \delta_j (\boldsymbol{\xi} \cdot \boldsymbol{\xi}). \tag{13}$$

At the thermodynamic limit $N \to \infty$, we introduce a new time variable α ($\Delta\alpha = 1/N$), which can be interpreted as a continuous time variable. The motion equations of the order parameters are then obtained by taking the expectation of all terms in (11), (13) and (6) with respect to input $\boldsymbol{\xi}$:

$$\frac{\mathrm{d}R_{in}}{\mathrm{d}\alpha} = \eta w_i \left[\sum_{m=1}^{M} v_m \langle g'(x_i) g(y_m) y_n\rangle - \sum_{j=1}^{K} w_j \langle g'(x_i) g(x_j) y_n\rangle\right] \tag{14}$$

$$\begin{aligned}\frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha} = {} & \eta w_i \left[\sum_{m=1}^{M} v_m \langle g'(x_i) g(y_m) x_j\rangle - \sum_{k=1}^{K} w_k \langle g'(x_i) g(x_k) x_j\rangle\right] \\ & + \eta w_j \left[\sum_{m=1}^{M} v_m \langle g'(x_j) g(y_m) x_i\rangle - \sum_{k=1}^{K} w_k \langle g'(x_j) g(x_k) x_i\rangle\right] \\ & + \eta^2 w_i w_j \left[\sum_{n=1}^{M}\sum_{m=1}^{M} v_n v_m \langle g'(x_i) g'(x_j) g(y_n) g(y_m)\rangle \right. \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K} w_k w_l \langle g'(x_i) g'(x_j) g(x_k) g(x_l)\rangle \\ & \left. - 2\sum_{k=1}^{K}\sum_{m=1}^{M} w_k v_m \langle g'(x_i) g'(x_j) g(x_k) g(y_m)\rangle\right]\end{aligned} \tag{15}$$

$$\frac{\mathrm{d}w_i}{\mathrm{d}\alpha} = \eta\left[\sum_{m=1}^{M} v_m \langle g(x_i)(y_m)\rangle - \sum_{j=1}^{K} w_j \langle g(x_i) g(x_j)\rangle\right]. \tag{16}$$

The terms $\langle g'(z_1)g(z_2)z_3\rangle$, $\langle g'(z_1)g'(z_2)g(z_3)g(z_4)\rangle$ and $\langle g(z_1)g(z_2)\rangle$ denote the expectation with respect to the input $\boldsymbol{\xi}$. Here, we assume that $\boldsymbol{\xi}$ has zero mean and unit variance. Then, if we take the thermodynamic limits, the random variables z_i are subject to the Gaussian distribution with zero mean, and these expectations can be determined from their variance–covariance matrix. In the case of $g(u) = \mathrm{erf}(u/\sqrt{2})$ especially, these expectations can be analytically calculated, and the motion equations can be given in a compact form using

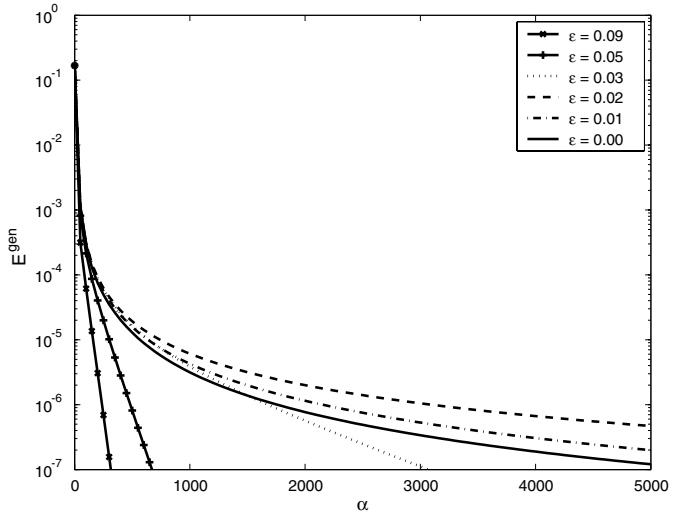

Figure 3. Time evolutions of generalization errors for various initial states.

parameters R_{in}, Q_{ij} and w_i, and the fixed values T_{nm} and v_n. (The explicit forms of the expectations are given in [12].)

These motion equations and the generalization error are general in the number of K and M so that they can be applied to the analysis of the dynamics under the over-realizable scenario. However, the previous studies mainly investigated the realizable case with no unidentifiable parameter. In this paper, we discuss the dynamics of the multilayer perceptrons under the over-realizable scenario, in which the optimal networks constitute subspaces in the parameter space of the student networks.

4. Results

4.1. Dynamics at the thermodynamic limit

We investigated a simple over-realizable scenario, $K = 2, M = 1$. To describe the learning dynamics, we need five order parameters (R_{11}, R_{21}, Q_{11}, Q_{22} and $Q_{12} = Q_{21}$), two hidden-output weight parameters (w_1, w_2) and the fixed values related to the teacher network (T, v). For the teacher network, we set $T = 1$ and $v = 1$. The student network then realizes the teacher network when one of the following conditions is satisfied;

$$(1)\ R_{11} = R_{21} = Q_{11} = Q_{22} = Q_{12} = 1 \qquad w_1 + w_2 = 1 \tag{17}$$

$$(2.1)\ R_{11} = Q_{11} = 1 \qquad w_1 = 1 \qquad w_2 = 0 \qquad Q_{12} = R_{21} \tag{18}$$

$$(2.2)\ R_{21} = Q_{22} = 1 \qquad w_1 = 0 \qquad w_2 = 1 \qquad Q_{12} = R_{11}. \tag{19}$$

The initial state for each order parameter was set as

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & Q_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} R_{11} \\ R_{21} \end{bmatrix} = \begin{bmatrix} 2 \times 10^{-2} \\ 10^{-2} \end{bmatrix}. \tag{20}$$

For the parameters (w_1, w_2), we tried various initial states: $(w_1, w_2) = (0.1, 0.1 - \varepsilon)$, $\varepsilon = 0, 0.01, 0.02, 0.03, 0.05, 0.09$.

The evolution of the generalization error for each initial value of ε is shown in figure 3. From the figure, we can see that the convergence speeds were strongly dependent on the initial

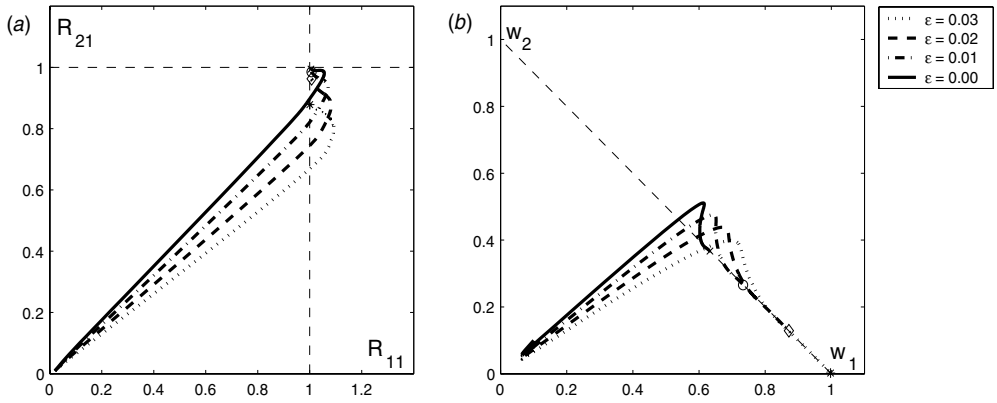

Figure 4. Parameter trajectories: (*a*) within the space of (R_{11}, R_{21}) and (*b*) within the space of (w_1, w_2).

relation between w_1 and w_2. When the difference between w_1 and w_2 (i.e., ε) was larger than 0.03, the learning converged rapidly. However, as ε decreased, the convergence speeds slowed down remarkably, and plateau-like phenomena appeared.

To find the cause of the slow convergence, we investigated the learning trajectories of order parameters (R_{11}, R_{21}) and parameters (w_1, w_2). Figure 4(*a*) shows the trajectories of (R_{11}, R_{21}) with different values of ε. The two dashed straight lines show the optimal subspaces. The crossing point of the two lines ($R_{11} = R_{21} = 1$) corresponds to optimal condition 1 defined in (17), and the other parts of the lines correspond to optimal condition 2 defined in (18) and (19). For a small ε, the parameters clearly approached optimal condition 1. From this, we conjecture that optimal condition 1 is related to the convergence speed. Figure 4(*b*), which shows the trajectories of (w_1, w_2), also supports our conjecture. In the space of (w_1, w_2), optimal condition 1 makes a line ($w_1 + w_2 = 1$), and the end points of the line in the figure ((1,0) and (0,1)) correspond to optimal condition 2. As shown in the figures, when $\varepsilon = 0.03$ (and also when $\varepsilon > 0.03$), the learning rapidly converged to the optimal subspace of condition 2. When $\varepsilon < 0.03$, slow dynamics was observed around the subspace $w_1 + w_2 = 1$, resulting in the slow convergence.

From these results, we can say that there is a quasi-plateau around the subspace satisfying optimal condition 1, at which the convergence speed evidently slows down. Additionally, note that the quasi-plateau differs from the plateau caused by permutation symmetry that was reported in [7] with regard to two points. First, the quasi-plateau occurs around the minimum, whereas the conventional plateau occurs around the saddle point. Second, the quasi-plateau is observed under specific initial conditions, whereas the conventional plateau is a typical phenomenon in the learning of neural networks. We should also remark that the optimal subspace causing the quasi-plateau does not exist in the parameter space of soft committee machines. This is the reason why the quasi-plateau has not been observed in previous works on the over-realizable scenario of soft committee machines [9, 10].

However, it is important to know the point of convergence of the learning dynamics, since the two different optimal subspaces have intrinsically different singular structures. From figure 4(*b*), we can see that the parameters moved along the subspace of optimal condition 1. However, it is difficult to decide from the trajectories whether the learning was slowly moving towards optimal condition 2 or stopping at optimal condition 1. Regarding the convergence point, one interesting point should be noted. In figure 3, within the range of sufficiently large value of ε ($\varepsilon > 0.03$), the convergence of learning slowed down as the value of ε decreased.

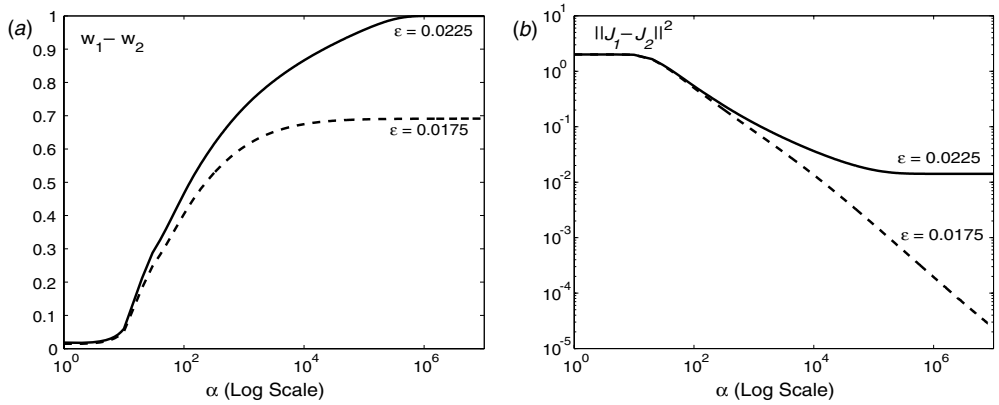

Figure 5. Time evolutions of $w_1 - w_2$ and $\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|^2$.

An interesting thing is that there was a turning point around $\varepsilon = 0.02$, at which the convergence again sped up.

To find what occurs around the turning point, we traced the motion equations with two different initial values of ε around 0.02: 0.0225 and 0.0175. Figure 5(a) shows the evolution of the difference between w_1 and w_2, and figure 5(b) shows the evolution of the sizes of difference vector $\boldsymbol{J}_1 - \boldsymbol{J}_2$. For $\varepsilon = 0.0225$, the value of $w_1 - w_2$ finally converged to 1, and the value of $\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|^2$ remained around 0.01; this implies optimal condition 2. In contrast, for $\varepsilon = 0.0175$, the value of $w_1 - w_2$ remained near 0.7, and the value of $\|\boldsymbol{J}_1 - \boldsymbol{J}_2\|^2$ linearly decreased to zero; this implies optimal condition 1. From these results, we can say that there is a critical value of ε around 0.02, at which the terminal condition of learning changes.

4.2. Numerical simulations with finite input dimension

We confirmed the validity of our theoretical results at the thermodynamic limits through numerical simulations of online learning with finite input dimensions. In the simulations, we used $(w_1, w_2) = (0.1, 0.08)$ as the initial values (i.e., $\varepsilon = 0.02$). To set the value of $\boldsymbol{B}$ and the initial values of $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$, we randomly and independently choose each element of the vectors subject to the normal distribution $\aleph(0, 1/N)$, and slightly changed the values so as to satisfy the initial conditions for R and Q defined in (20).

Figure 6 shows the simulation results for various input dimensions: $10, 10^2, 10^3$ and 10^4. The dashed curve in each figure shows the evolution of the generalization error obtained from the motion equations at the thermodynamic limit. Other curves were obtained from independent runs of online learning with randomly selected input $\boldsymbol{\xi} \sim \aleph(\boldsymbol{0}, \boldsymbol{I})$. Even though the curves from the online learning were scattered when the input dimension was small, we found that they approached close to the curve of the thermodynamic limit as N increased. The scatter of the online learning curves is an interesting phenomenon which was due to the large degree of freedom of the optimal points and the existence of the quasi-plateau in the neighbourhood of some optimal subspaces. Another interesting point arising from the simulations was that the theoretical curve was not located near the median of the simulation curves, but was located at the top. In the case of small N, we can speculate that the stochastic properties of online learning can cause the learning to avoid the quasi-plateau.

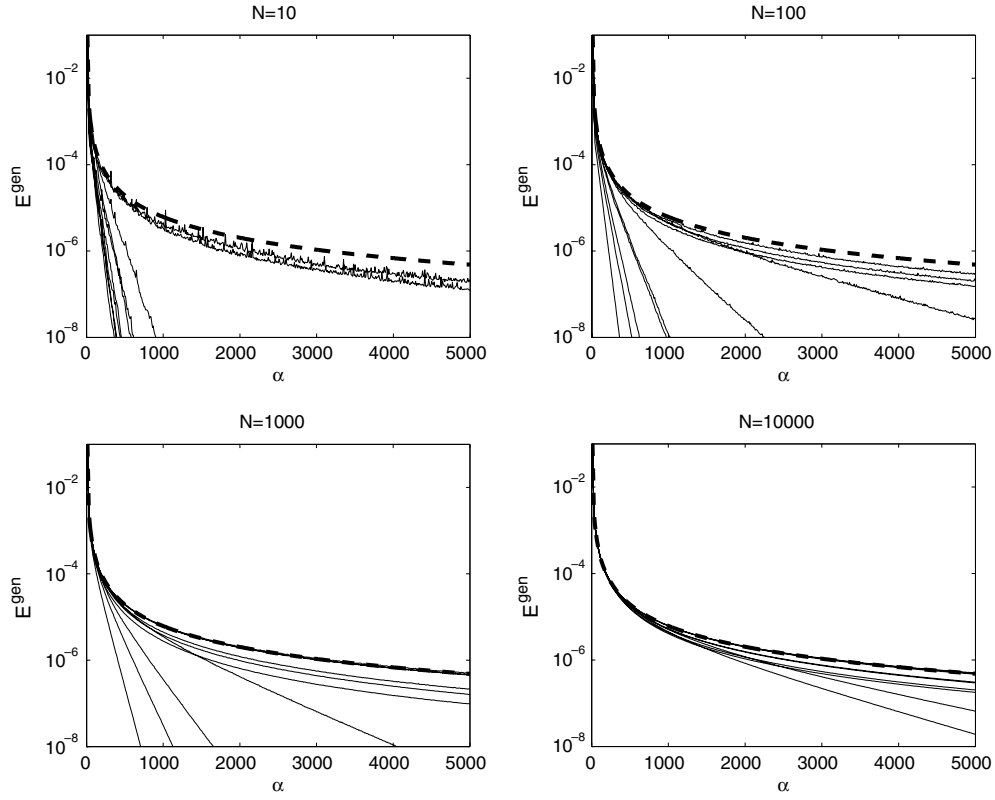

Figure 6. Numerical simulations with different values for input dimension (N).

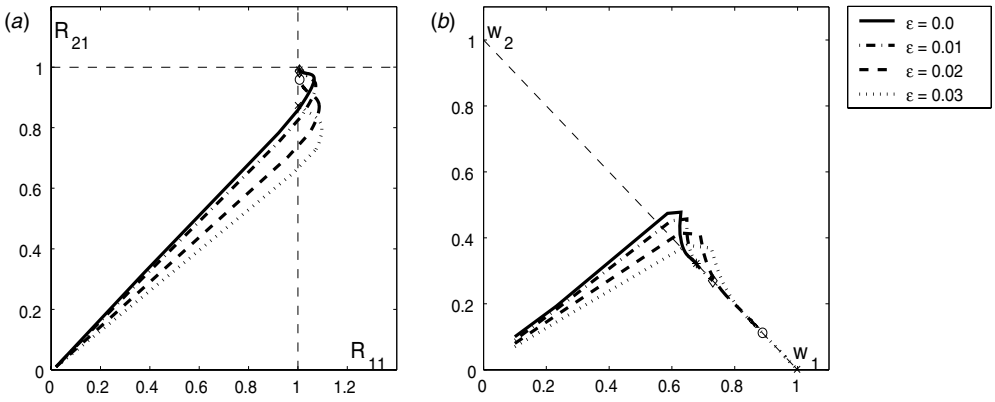

Figure 7. Numerical simulations for various initial states (ε).

Figure 7 shows the trajectories of parameters obtained from the numerical simulations with input dimension $N = 10^4$ for various initial values of ε. From the results, we can see that the trajectories at the thermodynamic limits in figure 4 fit well with the trajectories in the case of a finite input dimension.

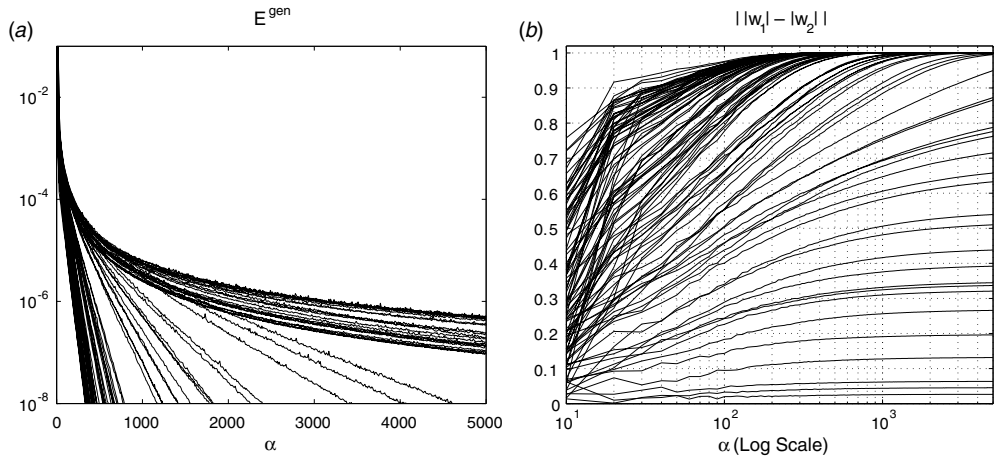

Figure 8. Time evolution of the generalization error and $||w_1| - |w_2||$ for random initial conditions.

5. Conclusions and discussions

Even though the singularity problem of multilayer perceptrons has been actively studied in recent years, the influence of singularity on learning dynamics has not been thoroughly discussed. By using a statistical mechanical framework, we investigated the convergence properties of learning dynamics when a teacher network is on the singularities in the parameter space of a student network. From the analysis, we first showed that the convergence speed was strongly dependent on an initial condition as well as that a quasi-plateau existed. In addition, we discussed the terminal point of learning. From the simulations, we showed that there is a critical initial state at which the terminal condition of learning changed. We also showed that learning converges to optimal condition 2, unless the difference between two hidden-output weight parameters is sufficiently small.

In addition, we checked whether the results obtained at the thermodynamic limit with specific initial conditions were still valid in practical situations where typical initial conditions held. To do this, we investigated the learning behaviour through numerical simulations. We conducted online learning for a student network with $N = 100$. For the parameter of the teacher network, we first randomly selected each element of $\boldsymbol{B}$ subject to $\aleph(0, 1/N)$, and then normalized $\boldsymbol{B}$ so as to satisfy $T = 1$. The value of v was set to 1. Training data were generated from the teacher network with input $\boldsymbol{\xi}$ subject to a standard multivariate normal distribution. We conducted 100 trials of the learning process with different initial values of $\boldsymbol{J}$ and w. For the initial values of $\boldsymbol{J}_i$ $(i = 1, 2)$, we randomly selected each element of the vector $\boldsymbol{J}_i$ subject to the normal distribution $\aleph(0, 1/N)$. The initial value of w_i was also randomly selected subject to the uniform distribution in $(-0.5, 0.5)$.

For the evolution of the generalization errors in figure 8(a), we confirmed the dependence of the convergence speed on the initial conditions. Figure 8(b) shows the evolution of $||w_1| - |w_2||$ for 100 trials with different initial values of the weight parameters. From the figure, we can see that $||w_1| - |w_2||$ mainly converged to 1 where the optimal subspace of condition 2 was satisfied. This result corresponds to the simulation results at the thermodynamic limit under specific initial conditions. From these results, we can say that the two observations—the large dispersion in convergence speed and the tendency to converge to the optimal condition 2—are still valid under the usual initial conditions. In a practical sense,

the convergence to optimal condition 2 has an important meaning, because it simplifies the structure optimization task by automatically pruning some unnecessary hidden nodes.

The singular structure of the parameter space of multilayer perceptrons greatly influences the learning dynamics in various situations, and this has not been clarified thoroughly. This paper is a preliminary step towards a better understanding of the rich dynamics related to singularities.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas no 14084212 and Grant-in-Aid for Scientific Research (C) no 14580438.

References

[1] Amari S, Ozeki T and Park H 2003 Learning and inference in hierarchical models with singularities *Syst. Comput. Japan* **34** 34–42
[2] Hagiwara K, Kuno K and Usui S 2000 On the problem in model selection of neural network regression in overrealizable scenario *Proc. Int. Joint Conf. of Neural Networks, IV* pp 461–6
[3] Hagiwara K 2002 On the problem in model selection of neural network regression in overrealizable scenario *Neural Comput.* **14** 1979–2002
[4] Fukumizu K 2003 Likelihood ratio of unidentifiable models and multilayer neural networks *Ann. Stat.* **3** 833–51
[5] Watanabe S 2001 Algebraic analysis for non-identifiable learning machines *Neural Comput.* **13** 899–933
[6] Riegler P and Biehl M 1995 On-line backpropagation in two-layered neural networks *J. Phys. A: Math. Gen.* **28** L507–13
[7] Saad D and Solla A 1995 On-line learning in soft committee machines *Phys. Rev.* E **52** 4225–43
[8] Inoue M, Park H and Okada M 2003 On-line learning theory of soft committee machines with correlated hidden units—steepest gradient descent and natural gradient descent *J. Phys. Soc. Japan* **72** 805–10
[9] Biehl M and Schwarze H 1995 Learning by on-line gradient descent *J. Phys. A: Math. Gen.* **28** 643–56
[10] Biehl M, Riegler P and Wöhler C 1996 Transient dynamics of on-line learning in two-layered neural networks *J. Phys. A: Math. Gen.* **29** 4769–80
[11] Amari S, Park H and Ozeki T 2003 Geometrical singularities in the neuromanifold of multilayer perceptrons *Adv. NIPS* **14** 343–50
[12] Riegler P 1997 Dynamics of on-line learning in neural networks *PhD* dissertation Bayerische Julius-Maximilians Universität, Würzburg