Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons

Kenji Fukumizu, Shun-ichi Amari
Brain Science Institute
The Institute of Physical and Chemical Research (RIKEN)
E-mail: {fuku,amari}@brain.riken.go.jp

October 22, 1999

Abstract

Local minima and plateaus pose a serious problem in learning of neural networks. We investigate the hierarchical geometric structure of the parameter space of three-layer perceptrons in order to show the existence of local minima and plateaus. It is proved that a critical point of the model with $H-1$ hidden units always gives many critical points of the model with H hidden units. These critical points consist of many lines in the parameter space, which can cause plateaus in learning of neural networks. Based on this result, we prove that a point in the critical lines corresponding to the global minimum of the smaller model can be a local minimum or a saddle point of the larger model. We give a necessary and sufficient condition for this, and show that this kind of local minima exist as a line segment if any. The results are universal in the sense that they do not require special properties of the target, loss functions, and activation functions, but only use the hierarchical structure of the model.

1 Introduction

It has been believed that the error surface of multilayer perceptrons (MLP) has in general many local minima. This has been regarded as one of the disadvantages of neural networks, and a great deal of effort has been paid to find good methods of avoiding them and achieving the global minimum.

There have been no rigorous results, however, to prove the existence of local minima. Even in the simple example of the XOR problem, existence of local minima had been a controversial problem. Lisboa and Perantonis

([1]) elucidated all the critical points of the XOR problem and asserted with a help of numerical simulations that some of them are local minima. Recently, Hamney ([2]) and Sprinkhuizen-Kuyper and Boers ([3],[4]) rigorously proved that what have been believed to be local minima in [1] correspond to local minima with infinite parameter values, and that there always exists a strictly decreasing path from each finite point to the global minimum. Thus, there are no local minima in the finite weight region for the XOR problem. Existence of local minima in general cases has been an open problem in the rigorous mathematical sense.

It is also difficult to derive meaningful results on local minima from numerical experiments. In practical applications, we often see extremely slow dynamics around a point that differs from the global minimum. However, it is not easy to tell rigorously whether it is a local minimum. It is known that a typical learning curve shows a *plateau* in the middle of training, which causes very slow decrease of the training error for a long time before a sudden exit from it (e.g. [5],[6]). A plateau can be easily misunderstood as a local minimum in practical problems.

This paper discusses critical points of the MLP model, which are caused by the hierarchical structure of the models having a smaller number of hidden units. For simplicity, we discuss only the MLP model with a one-dimensional output in this paper. The input-output function space of networks with $H-1$ hidden units is included in the function space of networks with H hidden units. However, the relation between the parameter spaces of these two models is not so simple (see [7],[8]). Sussmann ([9]) elucidated the condition that a function described by a network with H hidden units can be realized by a network with $H-1$ hidden units in the case of tanh activation function. In this paper, we further investigate the geometric structure of the parameters of networks which are realizable by a network with $H-1$ hidden units. In particular, we elucidate how they can be embedded in the parameter space of H hidden units. Based on the geometric structure, we show that a critical point of the error surface for the MLP model with $H-1$ hidden units gives a set of critical points in the parametric space of the MLP with H hidden units.

The main purpose of the present paper is to show that a subset of critical points corresponding to the global minimum of a smaller network can be local minima or saddles of the larger network. More precisely, the subset of critical points on which the input-output behavior is the same is divided into two parts, one consisting of local minima and the other saddle points. We give an explicit condition when this occurs. This gives a formal proof of the existence of local minima for the first time. Moreover, the coexistence

of local minima and saddles in one equivalent set of critical points explains a serious mechanism of plateaus: when such is the case, the network parameters are attracted in the part of local minima, stay and walking randomly for a long time in that flat region on which the performance is the same, but eventually go out from the part of saddles in the region. This is a new type of critical points in nonlinear dynamics given rise to by the hierarchical structure of a model.

This paper is organized as follows. In Section 2, after showing necessary definitions and terminologies, we elucidate the geometric or topological structure of the parameter space. Section 3 discusses critical points of the error surface. In Section 4, we mathematically prove the coexistence of local minima and saddles under one condition. This shows not only the existence of local minima but also a possible mechanism of plateaus. We also show the results of numerical simulations realizing local minima. Section 5 contains conclusion and discussion.

2 Geometric structure of the parameter space

2.1 Basic definitions

In this paper, we consider a three-layer perceptron with one linear output unit and L input unit. The input-output relation of a network with H hidden units is described by the function

$$f^{(H)}(\boldsymbol{x};\boldsymbol{\theta}^{(H)}) = \sum_{j=1}^{H} v_j \, \varphi\left(\boldsymbol{w}_j^T \boldsymbol{x} + w_{j0}\right) + v_0, \tag{1}$$

where T denotes transposition, $\boldsymbol{x} = (x_1, \ldots, x_L)^T \in \mathbb{R}^L$ is an input vector, $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jL})^T \in \mathbb{R}^L$ $(1 \leq j \leq H)$ is the weight vector of the jth hidden unit, and $\boldsymbol{\theta}^{(H)} = (v_0, v_1, \ldots, v_H, w_{10}, \boldsymbol{w}_1^T, \ldots, w_{H0}, \boldsymbol{w}_H^T)^T$ summarizes all the parameters in one large vector. The function $\varphi(t)$ is called an activation function. In this paper, we use *tanh* for φ. Introducing the notations

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} \in \mathbb{R}^{L+1}, \qquad \tilde{\boldsymbol{w}}_j = \begin{pmatrix} w_{j0} \\ \boldsymbol{w}_j \end{pmatrix} \in \mathbb{R}^{L+1}, \quad (1 \leq j \leq H), \tag{2}$$

for simplicity, we can write

$$f^{(H)}(\boldsymbol{x};\boldsymbol{\theta}^{(H)}) = \sum_{j=1}^{H} v_j \, \varphi\left(\tilde{\boldsymbol{w}}_j^T \tilde{\boldsymbol{x}}\right) + v_0. \tag{3}$$

Using the result of Kůrkova & Kainen ([10]), all the theorems obtained in this paper are applicable for other sigmoid- or bell-shaped activation functions with necessary modifications. We will show this later. We use the linear activation in the output layer. However, all the results are easily extended to a model with a monotone nonlinear output unit, because this causes only a nonlinear rescaling of the output data.

Given N input-output training data $\{(\boldsymbol{x}^{(\nu)}, y^{(\nu)}) | \nu = 1, \ldots, N\}$, we use a MLP model to realize the relation expressed by the data. The objective of training is to find the parameter that minimizes the error function defined by

$$E_H(\boldsymbol{\theta}^{(H)}) = \sum_{\nu=1}^{N} \ell(y^{(\nu)}, f(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}^{(H)})), \tag{4}$$

where $\ell(y, z)$ is a loss function such that $l(y, z) \geq 0$ and the equality holds if and only if $y = z$. When $l(y, z) = \frac{1}{2}\|y - z\|^2$, the objective function is the mean square error. We can use other loss functions such as L_p norm $l(y, z) = \frac{1}{p}\|y - z\|^p$ and the cross entropy $l(y, z) = -\sigma(y) \log \sigma(z) - (1 - \sigma(y)) \log(1 - \sigma(z))$, where $\sigma(t)$ is a sigmoidal function for the nonlinearity of the output unit. The results in this paper are independent of the choice of a loss function.

2.2 Hierarchical structure of MLP

The parameter vector $\boldsymbol{\theta}^{(H)}$ consists of a $LH + 2H + 1$ dimensional Euclidean space Θ_H. Each $\boldsymbol{\theta}^{(H)}$ gives a nonlinear function eq.(1) of $\boldsymbol{x}$, so that the set of all the functions realized by Θ_H is a function space described by

$$\mathcal{S}_H = \{f^{(H)}(\boldsymbol{x}; \boldsymbol{\theta}^{(H)}) : \mathbb{R}^L \to \mathbb{R} \mid \boldsymbol{\theta}^{(H)} \in \Theta_H\}. \tag{5}$$

We denote the mapping from Θ_H onto $\mathcal{S}_H$ by

$$\pi_H : \Theta_H \longrightarrow \mathcal{S}_H, \qquad \boldsymbol{\theta}^{(H)} \mapsto f(\boldsymbol{x}; \boldsymbol{\theta}^{(H)}). \tag{6}$$

We sometimes write $f^{(H)}_{\boldsymbol{\theta}}$ for $\pi_H(\boldsymbol{\theta})$.

It is important to note that π_H is *not* one-to-one, that is, different $\boldsymbol{\theta}^{(H)}$ may give the same input-output function. The interchange between $(v_{j_1}, \boldsymbol{w}_{j_1})$ and $(v_{j_2}, \boldsymbol{w}_{j_2})$ does not alter the image of π_H. In the case of tanh activation function, Chen et al. ([7]) showed that any analytic map $T : \Theta_H \to \Theta_H$ such that $f^{(H)}(\boldsymbol{x}; T(\boldsymbol{\theta}^{(H)})) = f^{(H)}(\boldsymbol{x}; \boldsymbol{\theta}^{(H)})$ is a composition of hidden unit weight interchanges and hidden unit weight sign flips, which

latter are defined by $(v_j, \tilde{\boldsymbol{w}}_j) \mapsto (-v_j, -\tilde{\boldsymbol{w}}_j)$. These transforms consist of an algebraic group G_H, which is isomorphic to a direct product of Weyl groups. We write T_g for the transform given by $g \in G_H$

The function spaces $\mathcal{S}_H$ $(H = 0, 1, 2, \dots)$ have a trivial hierarchical structure:

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots \subset \mathcal{S}_{H-1} \subset \mathcal{S}_H \subset \cdots . \tag{7}$$

The inclusion is denoted by $\iota_{H-1} : \mathcal{S}_{H-1} \hookrightarrow \mathcal{S}_H$. On the other hand, the parameter space of the smaller networks is not canonically included in Θ_H. Given a function $f^{(H-1)}_{\boldsymbol{\theta}}$ realized by a network with $H-1$ hidden units, there are a family of networks with H hidden units and parameters $\boldsymbol{\theta}^{(H)}$ that realizes the same function $f^{(H-1)}_{\boldsymbol{\theta}}$. In other words, a map from Θ_{H-1} to Θ_H that commutes the following diagram is not unique.

$$\begin{array}{ccc} \Theta_{H-1} & \longrightarrow & \Theta_H \\ {\scriptstyle \pi_{H-1}}\downarrow & & \downarrow{\scriptstyle \pi_H} \\ \mathcal{S}_{H-1} & \underset{\iota_{H-1}}{\longrightarrow} & \mathcal{S}_H \end{array} \tag{8}$$

The set of all the parameters $\boldsymbol{\theta}^{(H)}$ that realize the input-output functions of networks with $H-1$ hidden units is denoted by

$$\Omega_H = \pi_H^{-1}(\iota_{H-1}(\mathcal{S}_{H-1})). \tag{9}$$

From Sussmann's result ([9], Theorem 1 and its corollary), the parameter set Ω_H is the union of the following submanifolds of Θ_H (see Figure 1);

$$\mathcal{A}_j = \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid v_j = 0\} \quad (1 \le j \le H), \tag{10}$$

$$\mathcal{B}_j = \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \boldsymbol{w}_j = \boldsymbol{0}\} \quad (1 \le j \le H), \tag{11}$$

$$\mathcal{C}^{\pm}_{j_1 j_2} = \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \tilde{\boldsymbol{w}}_{j_1} = \pm \tilde{\boldsymbol{w}}_{j_2}\} \quad (1 \le j_1 < j_2 \le H). \tag{12}$$

Here, $\mathcal{A}_j$ is the set of parameters where $v_j = 0$ so that the jth hidden units plays no role. Similarly, the jth hidden unit has 0 weight in $\mathcal{B}_j$ so that it outputs only a constant bias term. In $\mathcal{C}^{\pm}_{j_1 j_2}$, the j_1th hidden unit and j_2th hidden unit have the same (or opposite) wight vector and bias, so that their behaviors are the same (opposite). They may be integrated into one unit, where $v_1 \pm v_2$ is the weight of the new unit to the output unit. From the viewpoint of mathematical statistics, it is also proved by Fukumizu ([11]) that Ω is the set of all the points at which the Fisher information matrix is singular.

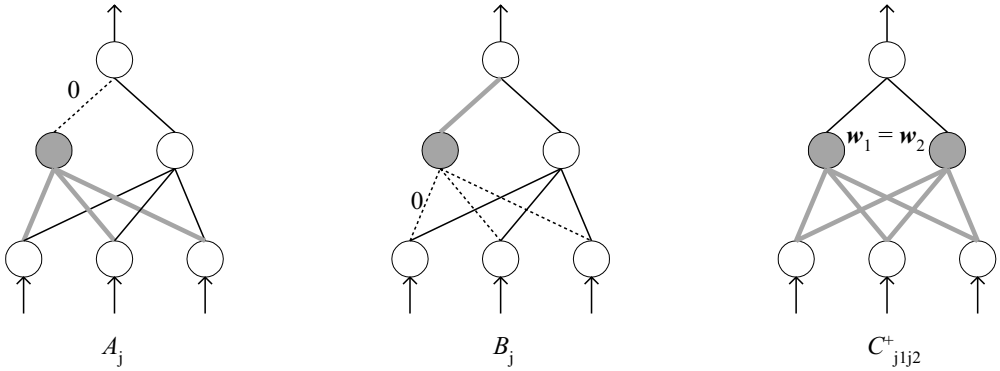

Figure 1: A network given by a parameter in $\mathcal{A}_j$, $\mathcal{B}_j$ and $\mathcal{C}^{\pm}_{j_1 j_2}$.

We further investigate how each function in $\mathcal{S}_{H-1}$ is. Let $f^{(H-1)}_{\boldsymbol{\theta}^{(H-1)}}$ be a function in $\mathcal{S}_{H-1} - \mathcal{S}_{H-2}$. To distinguish Θ_{H-1} from Θ_H, we use different parameter variables and indexing:

$$f^{(H-1)}(\boldsymbol{x}; \boldsymbol{\theta}^{(H-1)}) = \sum_{j=2}^{H} \zeta_j \varphi(\tilde{\boldsymbol{u}}_j^T \tilde{\boldsymbol{x}}) + \zeta_0. \tag{13}$$

Let $\Omega_H(\boldsymbol{\theta}^{(H-1)})$ be the set of parameters $\boldsymbol{\theta}^{(H)}$ that realizes a given $f^{(H-1)}_{\boldsymbol{\theta}^{(H-1)}}$;

$$\Omega_H(\boldsymbol{\theta}^{(H-1)}) = \pi_H^{-1}(\iota_{H-1}(f^{(H-1)}_{\boldsymbol{\theta}^{(H-1)}}). \tag{14}$$

Then, $\Omega_H(\boldsymbol{\theta}^{(H-1)})$ is the union of the submanifolds in each of $\mathcal{A}_j$, $\mathcal{B}_j$ and $\mathcal{C}^{\pm}_{j_1 j_2}$. For simplicity, we show only an example of the submanifolds of $\mathcal{A}_1$, $\mathcal{B}_1$ and $\mathcal{C}^{+}_{12}$;

$$\begin{aligned}
\Lambda &= \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid v_1 = 0, v_0 = \zeta_0, v_j = \zeta_j, \tilde{\boldsymbol{w}}_j = \tilde{\boldsymbol{u}}_j, (2 \le j \le H)\} \\
\Xi &= \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \boldsymbol{w}_1 = \boldsymbol{0}, v_1 \varphi(w_{10}) + v_0 = \zeta_0, v_j = \zeta_j, \tilde{\boldsymbol{w}}_j = \tilde{\boldsymbol{u}}_j, (2 \le j \le H)\} \\
\Gamma &= \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \tilde{\boldsymbol{w}}_1 = \tilde{\boldsymbol{w}}_2 = \tilde{\boldsymbol{u}}_2, v_0 = \zeta_0, v_1 + v_2 = \zeta_2, \\
&\qquad v_j = \zeta_j, \tilde{\boldsymbol{w}}_j = \tilde{\boldsymbol{u}}_j, (3 \le j \le H)\}.
\end{aligned} \tag{15}$$

The submanifold Λ is an $L+1$ dimensional affine space parallel to the $\tilde{\boldsymbol{w}}_1$-plane, because $\tilde{\boldsymbol{w}}_1$ may take arbitrary values in it, but all the other components of $\boldsymbol{\theta}^{(H)}$ are determined by prescribed $\boldsymbol{\theta}^{(H-1)}$. The set Ξ is a 2 dimensional submanifold defined by a nonlinear equation

$$v_1 \varphi(w_{10}) + v_0 = \zeta_0, \tag{16}$$

where v_0, v_1, and w_{10} can take arbitrary values provided they satisfy the above. The set Γ is a line in the v_1v_2-plane, defined by $v_1 + v_2 = \zeta_2$. It is known that all the other components in $\Omega(\boldsymbol{\theta}^{(H-1)})$ are obtained as the transforms of Λ, Ξ, and Γ by $g \in G_H$ ([9],[10]). For example, the image of Γ by the sign flip about the second hidden unit is given by

$$\Gamma^{(-1)} = \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \tilde{\boldsymbol{w}}_1 = -\tilde{\boldsymbol{w}}_2 = \tilde{\boldsymbol{u}}_2, v_0 = \zeta_0, v_1 - v_2 = \zeta_2, \\ v_j = \zeta_j, \tilde{\boldsymbol{w}}_j = \tilde{\boldsymbol{u}}_j, (3 \leq j \leq H)\}. \quad (17)$$

The image of Λ, Ξ, and Γ by a hidden-unit interchange is trivial. Thus, each function of a smaller network is realized not by discrete points but by high-dimensional submanifolds in Θ_H.

In order to make analysis more concrete, we give a definite correspondence between Θ_{H-1} and Θ_H that realize the same function. We define the following canonical embeddings of Θ_{H-1} into Θ_H, which commute the diagram (8), using $\tilde{\boldsymbol{w}} \in \mathbb{R}^{L+1}$, $(v, w) \in \mathbb{R}^2$, and $\lambda \in \mathbb{R}$ as their parameters;

$$\begin{aligned}
\alpha_{\tilde{\boldsymbol{w}}} &: \Theta_{H-1} \longrightarrow \Theta_H, \quad \boldsymbol{\theta}^{(H-1)} \mapsto (\zeta_0, 0, \zeta_2, \ldots, \zeta_H, \tilde{\boldsymbol{w}}^T, \tilde{\boldsymbol{u}}_2^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T, \\
\beta_{(v,w)} &: \Theta_{H-1} \longrightarrow \Theta_H, \quad \boldsymbol{\theta}^{(H-1)} \mapsto (\zeta_0 - v\varphi(w), v, \zeta_2, \ldots, \zeta_H, (w, \mathbf{0}^T), \tilde{\boldsymbol{u}}_2^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T, \\
\gamma_\lambda &: \Theta_{H-1} \longrightarrow \Theta_H, \\
&\boldsymbol{\theta}^{(H-1)} \mapsto (\zeta_0, \lambda\zeta_2, (1-\lambda)\zeta_2, \zeta_3, \ldots, \zeta_H, \tilde{\boldsymbol{u}}_2^T, \tilde{\boldsymbol{u}}_2^T, \tilde{\boldsymbol{u}}_3^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T.
\end{aligned} \quad (18)$$

These maps are illustrated in Figures 2, 3, and 4. If we change the parameter of each embedding, the images of $\boldsymbol{\theta}^{(H-1)}$ span the components, Λ, Ξ, and Γ, of $\Omega_H(\boldsymbol{\theta}^{(H-1)})$; that is

$$\begin{aligned}
\Lambda &= \{\alpha_{\tilde{\boldsymbol{w}}}(\boldsymbol{\theta}^{(H-1)}) \mid \tilde{\boldsymbol{w}} \in \mathbb{R}^{L+1}\}, \\
\Xi &= \{\beta_{(v,w)}(\boldsymbol{\theta}^{(H-1)}) \mid (v, w) \in \mathbb{R}^2\}, \\
\Gamma &= \{\gamma_\lambda(\boldsymbol{\theta}^{(H-1)}) \mid \lambda \in \mathbb{R}\}.
\end{aligned} \quad (19)$$

3 Critical points of the MLP model

3.1 Learning and critical points

Generally, the optimum parameter cannot be calculated analytically when the model is nonlinear. Some numerical optimization method is needed to

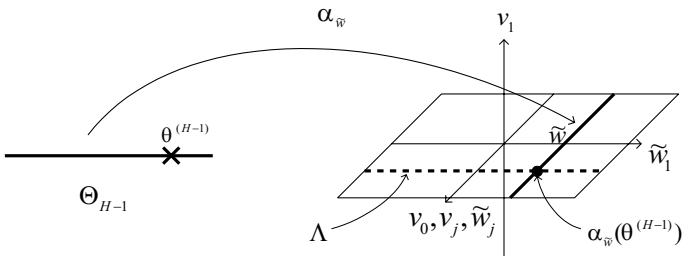

Figure 2: Embedding $\alpha_{\tilde{\boldsymbol{w}}}$.

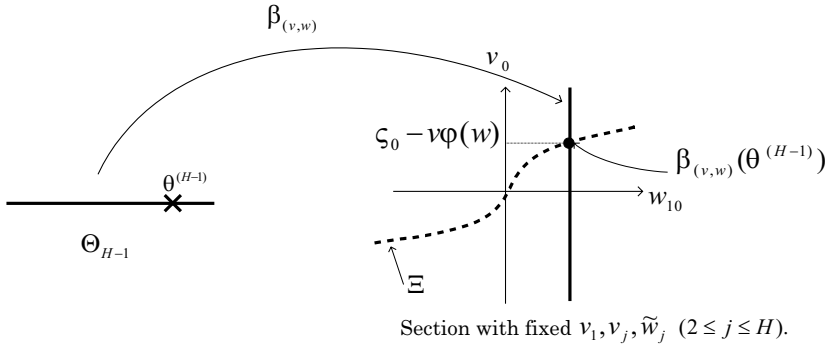

Figure 3: Embedding $\beta_{(v,w)}$.

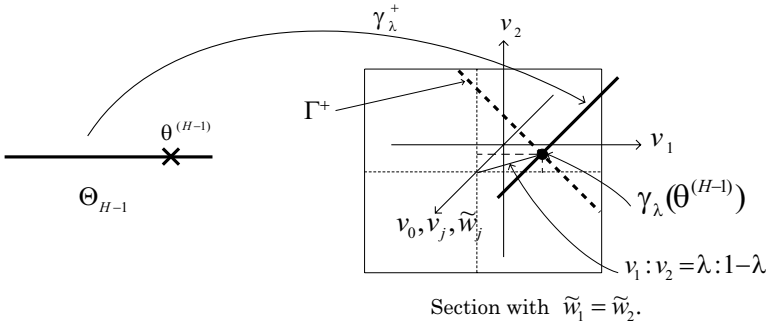

Figure 4: Embedding γ_λ^+.

obtain its approximation. One widely-used method is the steepest descent method, which leads to a learning rule given by

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \delta \frac{\partial E_H(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}, \tag{20}$$

where δ is a learning rate. If $\boldsymbol{\theta}_*$ is the global minimum, $\frac{\partial E_H}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_*) = \mathbf{0}$ holds and the above learning rule stops there. However, we cannot always obtain the global minimum, since all the points that satisfy $\frac{\partial E_H}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathbf{0}$ are stationary points of eq.(20). Such a point is called a *critical point* of E_H.

There are three types of critical point: a local minimum, a local maximum, and a saddle point. A critical point $\boldsymbol{\theta}_0$ is called a *local minimum* (*maximum*) if there exists a neighborhood around $\boldsymbol{\theta}_0$ such that for any point $\boldsymbol{\theta}$ in the neighborhood $E_H(\boldsymbol{\theta}) \geq E_H(\boldsymbol{\theta}_0)$ ($E_H(\boldsymbol{\theta}) \leq E_H(\boldsymbol{\theta}_0)$) holds, and called a *saddle* if it is neither a local minimum nor a local maximum, that is, if in an arbitrary neighborhood of $\boldsymbol{\theta}_0$ there exist a point at which E_H is smaller than $E_H(\boldsymbol{\theta}_0)$ and a point at which E_H is larger than $E_H(\boldsymbol{\theta}_0)$. It is well known that if the Hessian matrix at a critical point is strictly positive (negative) definite, the critical point is a local minimum (maximum), and if the Hessian has both positive and negative eigenvalues, it is a saddle.

3.2 Existence of critical points

It is very natural to look for a critical point of E_H in the set $\Omega_H = \pi_H^{-1}(\iota_{H-1}(\mathcal{S}_{H-1}))$, because a critical point of E_{H-1} already satisfies some of the conditions of a critical point of E_H.

Let $\boldsymbol{\theta}_*^{(H-1)} = (\zeta_{0*}, \zeta_{2*}, \ldots, \zeta_{H*}, \tilde{\boldsymbol{u}}_{2*}^T, \ldots, \tilde{\boldsymbol{u}}_{H*}^T)^T \in \Theta_{H-1} - \Theta_{H-2}$ be a critical point of E_{H-1}. It really exists if we assume that the global minimum of E_{H-1} is isolated, which means it is not included in Θ_{H-2}. This assumption is practically plausible, because, for a set of data which is fitted well with H hidden units, the optimum network with $H-1$ hidden units has no redundant hidden units in general.

At the critical point, the following equations hold for $2 \leq j \leq H$;

$$\begin{aligned}
&\frac{\partial E_{H-1}}{\partial \zeta_0}(\boldsymbol{\theta}_*^{(H-1)}) = \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\left(y^{(\nu)}, f^{(H-1)}(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}_*^{(H-1)})\right) = 0, \\
&\frac{\partial E_{H-1}}{\partial \zeta_j}(\boldsymbol{\theta}_*^{(H-1)}) = \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\left(y^{(\nu)}, f^{(H-1)}(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}_*^{(H-1)})\right)\varphi(\tilde{\boldsymbol{u}}_{j*}^T \tilde{\boldsymbol{x}}^{(\nu)}) = 0, \\
&\frac{\partial E_{H-1}}{\partial \tilde{\boldsymbol{u}}_j}(\boldsymbol{\theta}_*^{(H-1)}) = \zeta_{j*} \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\left(y^{(\nu)}, f^{(H-1)}(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}_*^{(H-1)})\right)\varphi'(\tilde{\boldsymbol{u}}_{j*}^T \tilde{\boldsymbol{x}}^{(\nu)})\tilde{\boldsymbol{x}}^{(\nu)T} = \mathbf{0}^T,
\end{aligned} \tag{21}$$

We have two kinds of critical points.

Theorem 1. *Let γ_λ be as in eq.(18). Then, for any $\lambda \in \mathbb{R}$, the point $\gamma_\lambda(\boldsymbol{\theta}_*^{(H-1)})$ is a critical point of E_H.*

Theorem 2. *Let $\beta_{(v,w)}$ be as in eq.(18). Then, for any $w \in \mathbb{R}$, the point $\beta_{(0,w)}(\boldsymbol{\theta}_*^{(H-1)})$ is a critical point of E_H.*

These theorems are easily obtained if we consider the partial derivatives of E_H, which are given by

$$\begin{aligned}
\frac{\partial E_H}{\partial v_0}(\boldsymbol{\theta}) &= \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\big(y^{(\nu)}, f^{(H)}(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})\big), \\
\frac{\partial E_H}{\partial v_j}(\boldsymbol{\theta}) &= \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\big(y^{(\nu)}, f^{(H)}(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})\big)\varphi(\tilde{\boldsymbol{w}}_j^T \tilde{\boldsymbol{x}}^{(\nu)}), \quad (1 \le j \le H), \\
\frac{\partial E_H}{\partial \tilde{\boldsymbol{w}}_j}(\boldsymbol{\theta}) &= v_j \textstyle\sum_{\nu=1}^N \frac{\partial \ell}{\partial z}\big(y^{(\nu)}, f^{(H)}(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})\big)\varphi'(\tilde{\boldsymbol{w}}_j^T \tilde{\boldsymbol{x}}^{(\nu)})\tilde{\boldsymbol{x}}^{(\nu)T}, \quad (1 \le j \le H).
\end{aligned} \tag{22}$$

Note that $f^{(H)}(\boldsymbol{x};\boldsymbol{\theta}) = f^{(H-1)}(\boldsymbol{x};\boldsymbol{\theta}_*^{(H-1)})$ for $\boldsymbol{\theta} = \gamma_\lambda(\boldsymbol{\theta}_*^{(H-1)})$ or $\boldsymbol{\theta} = \beta_{(0,w)}(\boldsymbol{\theta}_*^{(H-1)})$. It is easy to check that the conditions eq.(21) make all the above derivatives zero.

The critical points in Theorems 1 and 2 consist of a line in Θ_H if we move $\lambda \in \mathbb{R}$ and $w \in \mathbb{R}$, respectively. Note that $\alpha_{\tilde{\boldsymbol{w}}} = \beta_{(0,w)}$ if $\tilde{\boldsymbol{w}} = (w, \mathbf{0}^T)^T$. Thus, these two embeddings give the same critical point set. If $\boldsymbol{\theta}$ is a critical point of E_H, so is $T_g(\boldsymbol{\theta})$ for any $g \in G_H$. We have many critical lines in Θ_H.

The critical points in Theorems 1 and 2 do not cover all of the critical points of E_H. We consider the special subset of the critical points, which appears because of the hierarchical structure of the model.

4 Local minima of the MLP model

4.1 A condition for the existence of local minima

In this section, we show a condition that a critical point in Theorem 1 is a local minimum or a saddle point. The usual sufficient condition of a local minimum using the Hessian matrix cannot be applied for a critical point in Theorem 1 and 2. The Hessian is singular, because a one-dimensional set including the point shares the same value of E_H in common.

Let $\boldsymbol{\theta}_*^{(H-1)}$ be a point in Θ_{H-1}. We define the following $(L+1)\times(L+1)$ symmetric matrix:

$$A_2 = \zeta_{2*} \sum_{\nu=1}^{N} \frac{\partial \ell}{\partial z}(y^{(\nu)}, f^{(H-1)}(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}_*^{(H-1)}))\varphi''(\tilde{\boldsymbol{u}}_{2*}^T \tilde{\boldsymbol{x}}^{(\nu)})\tilde{\boldsymbol{x}}^{(\nu)}\tilde{\boldsymbol{x}}^{(\nu)T}. \quad (23)$$

Theorem 3. *Let $\boldsymbol{\theta}_*^{(H-1)}$ be a local minimum of E_{H-1} such that the Hessian matrix at $\boldsymbol{\theta}_*^{(H-1)}$ is positive definite. Let γ_λ be defined by eq.(18), and $\Gamma := \{\boldsymbol{\theta}_\lambda \in \Theta_H | \boldsymbol{\theta}_\lambda = \gamma_\lambda(\boldsymbol{\theta}_*^{(H-1)}), \lambda \in \mathbb{R}\}$. A matrix A_2 is defined by eq.(23). If A_2 is positive (negative) definite, any point in the set $\Gamma_0 = \{\boldsymbol{\theta}_\lambda \in \Gamma \mid \lambda(1-\lambda) > 0 \ (< 0)\}$ is a local minimum of E_H, and any point in $\Gamma - \Gamma_0$ is a saddle. If A_2 has both positive and negative eigenvalues, all the points in Γ are saddle points.*

For the proof, see Appendix. Local minima given by Theorem 3, if any, appear as one or two segments in a line. It is interesting that such a local minimum can be changed into a saddle point without altering the function $f_{\boldsymbol{\theta}}^{(H)}$, when the point moves in the segment. Figure 5 illustrates the error surface around this critical set. We show only two coordinate axes for variables: one is the direction along Γ and the other is the direction that attains the minimum and maximum values at the points on Γ. Each point that looks like a maximum in the figure is a saddle point in reality.

Note that, if $\boldsymbol{\theta}$ is a local minimum given by Theorem 3, the image of the point by any transform in G_H is also a local minimum. This can be easily proved because the local property of E_H around the point does not change by the transform. Therefore, the error function has many line segments of local minima, if the condition of Theorem 3 holds.

The critical points in Theorem 2 do not give local minima.

Theorem 4. *Any critical point given by Theorem 2 is a saddle.*

For the proof, see Appendix.

The statements of Theorems 3 and 4 are also valid even if we consider the transform of the embedded point by any $g \in G_H$. This can be proved easily because the local property around the point is not changed any transform in G_H. There are many saddle line segments, and line segments of local minima if any.

4.2 Plateaus

We have proved that, when A_2 is positive or negative definite, there exists a one-dimensional submanifold Γ of critical points. The output function is

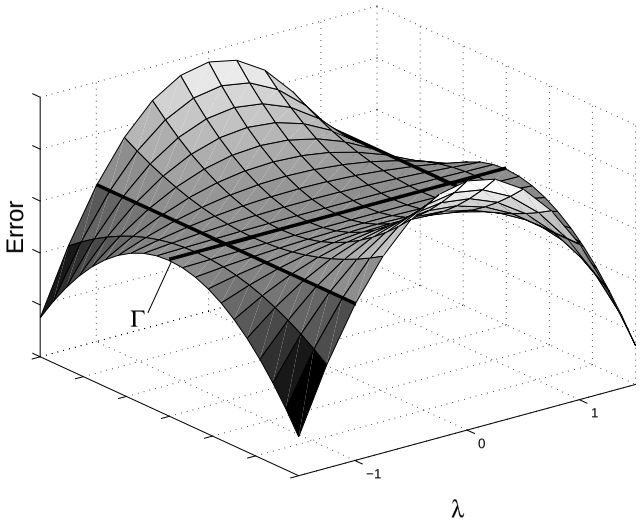

Figure 5: Critical set with local minima and plateaus

the same in Γ. The set Γ is divided into two parts Γ_0 and $\Gamma - \Gamma_0$, where Γ_0 consists of local minima and $\Gamma - \Gamma_0$ saddles.

If we map Γ to the function space, $\pi_H(\Gamma)$ consists of a single point which is the common function $f^{(H)}_{\boldsymbol{\theta}^{(H)}} \in \mathcal{S}_H$. Therefore, if we consider the cost function E_H as a function on $\mathcal{S}_H$, $\pi_H(\Gamma)$ is a saddle, because E_H takes both larger and smaller values than $E_H(\boldsymbol{\theta}^{(H)})$ in any neighborhood of $f^{(H)}_{\boldsymbol{\theta}^{(H)}}$ in $\mathcal{S}_H$.

It is interesting to see that Γ_0 is attractive in its neighborhood. Hence, any point in its small neighborhood is attracted to Γ_0. However, if we use on-line learning, in which the parameter is updated with a training datum presented one by one, the point attracted to Γ_0 fluctuates randomly along Γ_0 by learning dynamics (20). It eventually escapes from Γ when it reaches $\Gamma - \Gamma_0$. This takes a long time because of the nature of random fluctuation. This explains that this type of critical points are serious plateaus. This is a new type of saddle which has so far not remarked in nonlinear dynamics. This type of "intrinsic saddle" is given rise to by the singular structure of the topology of $\mathcal{S}_H$.

4.3 Remarks

The only property of tanh used in this paper is that it is odd. Kůrkova & Kainen ([10], Theorem 8) introduced the notion of affinely recursive functions, and proved that if the activation function is odd or even and is not affinely recursive, the functionally equivalent parameters are given by interchanges and sign flips, neglecting compensation of constant. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is not affinely recursive if and only if it has a non-trivial affine relation $\varphi(t) = a\varphi(wt + u) + b$ for $a, w \neq 0$ and an affine relation $\sum_{j=1}^{m} a_j \varphi(w_j t + u_j) + b = 0$ of more than three components can be decomposed into affine relations of two components (For the precise definition, see [10]). Using this result, we can deduce that Ω_H is the same as for MLP models with an odd or even activation function that is not affinely recursive, and can determine the transform group for such MLP models. The group is still the same as G_H, while we must replace the definition of a sign flip by $(v_j, \boldsymbol{w}_j) \mapsto (v_j, -\boldsymbol{w}_j)$ for an even activation function. Similar arguments in Section 2.2 are valid, and Theorems 1–4 also hold with necessary modifications of the statements. A typical activation function like the logistic function and Gaussian function can be converted by an affine transform to an odd or even function that is not affinely recursive. Therefore, the results obtained in the above are applicable to a wide class of three-layer models.

4.4 Numerical simulations

We have tried numerical simulation to exemplify local minima given by Theorem 3 and plateaus described in 4.2.

In the first simulation, We use a network with 1 input unit, 1 output unit, and 2 hidden units. We do not use bias terms for simplicity. Note that there always exist local minima in this case, since A_2 is a scalar. We use the logistic function $\varphi(t) = \frac{1}{1+e^{-t}}$ as the activation function, and the mean square error $l(y, z) = \frac{1}{2}\|y - z\|^2$ for the loss function. To obtain training data, 100 input data are generated using a normal distribution with 0 as its mean and 4.0 as its variance, and corresponding output data are obtained as $y = f(x) + Z$, where $f(x) = 2\varphi(x) - \varphi(4x)$ and Z is a random variable subject to the normal variable with 0 as its mean and 10^{-4} as its variance. For a fixed set of training data, we numerically calculate the global minimum of MLP with 1 hidden unit using the steepest descent method. We update the parameter 20000 times, and use the final state as the global minimum. Even if we try several different initial conditions, obtained results are almost the same. Therefore, we can consider it as an approximation of the global

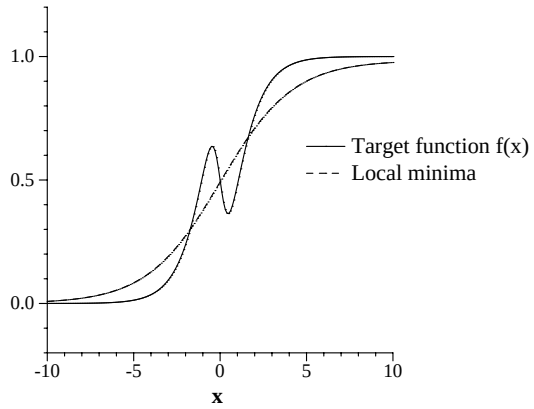

Figure 6: A local minimum in MLP (L=1, H=2).

minimum $\boldsymbol{\theta}_*^{(1)}$ with high accuracy. The parameter is given by $\zeta_{2*} = 0.984$ and $u_{2*} = 0.475$. In this case, we have $A_2 = 1.910 > 0$. Then, any point in the set $\Gamma_0 = \{(v_1, v_2, w_1, w_2) \mid v_1 + v_2 = \zeta_{2*}, v_1 v_2 > 0, w_1 = w_2 = u_{2*}\}$ is a local minimum. We set $v_1 = v_2 = \zeta_{2*}/2$ as $\boldsymbol{\theta}_\lambda$ ($\lambda = 1/2$), and evaluate the values of E_2 at 1 million points around $\boldsymbol{\theta}_\lambda$, which are generated using a 4 dimensional normal distribution with $\boldsymbol{\theta}_\lambda$ as its mean and $10^{-6} I_4$ as its variance-covariance matrix. As a result, all these values are larger than $E(\boldsymbol{\theta}_\lambda)$. This experimentally verifies that $\boldsymbol{\theta}_\lambda$ is a local minimum. The graphs of the target function $f(x)$ and the function given by the local minimum $f(x; \boldsymbol{\theta}_*^{(1)})$ are shown in Figure 6.

In the second simulation, we use a network with 2 input units, 1 output unit, and 3 hidden units. We do not use bias terms also in this simulation. The 2×2 matrix A_2 can have both of a negative and a positive eigenvalue at the same time. The activation function is tanh, and the set of input data is 100 independent samples from the normal distribution with 0 as its mean and $25 \times I_2$ as its covariance matrix. The target function is given by a function in the model, which is defined by $v_1 = v_2 = v_3 = 1$, $\boldsymbol{w}_1 = (2, 1)^T$, $\boldsymbol{w}_2 = (1, -1)^T$, and $\boldsymbol{w}_3 = (0.5, 0)^T$. We numerically obtain the global minimum of the model with two hidden units, $\boldsymbol{\theta}_*^{(2)}$, in the similar method to the first simulation. There are many cases in which the matrix A_2 has a negative and a positive eigenvalue, but for some sets of training data and initial parameters we can find the matrix positive or negative definite. Figure 7 shows the graph of a function given by one of such local minima and the graph of the target function. The parameter of this local minimum is $\zeta_{1*} =$

1.864, $\zeta_{2*} = -1.158$, $\boldsymbol{u}_{1*} = (-0.680, 0.247)^T$, and $\boldsymbol{u}_{2*} = (-0.905, -1.158)^T$. We numerically confirmed in the same way as the first simulation that this is really a local minimum.

Next, we have tried to verify that this local minimum causes a plateau. In this simulation, we use online learning, in which the parameter is updated with respect to only one training data that is selected at that time. All of the training data are used by turns, and training is repeated cyclically. We observe the behavior of the parameter after setting it close to the one that gives the local minimum. Figure 8 is the graph of the value of error function $E_3(\boldsymbol{\theta})$ during learning, which shows a typical plateau until about 50000 iterations. One sequence of presenting all data is counted as one iteration in this figure. We can see a very long time interval, in which the error function decreases very slowly, and a sudden steep decrease of the training error. Figure 9 shows the behavior of the parameter $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. They move close to the parameter $\boldsymbol{u}_{1*}$, which gives the local minimum, and suddenly go away from it. This simulation verifies that local minima given by Theorem 3 can give rise to plateaus as we discussed in 4.2.

5 Conclusion

We investigated the geometric structure of the parameter space of multilayer perceptrons with $H-1$ hidden units embedded in the parameter space of H hidden units. Based on the structure, we found a finite family of critical point sets of the error surface. We showed that a critical point of a smaller network can be embedded into the parameter space as a critical point set of a one-dimensional affine space in two ways. We further elucidated a condition that a point in the image of one embedding is a local minimum, and showed that the image of the other embedding is a saddle. From this result, we see that under one condition there exist local minima as line segments in the parameter space, which cause serious plateaus because all points around the set of local minima once converge to it and have to escape from it by random fluctuation. These results are not dependent on the specific form of activation functions nor the loss functions.

We consider only networks with one output unit. The extension of the result on existence of local minima is not straightforward. The image of the embedding γ_λ form a critical line even in the M dimensional output case. However, the critical line is contained in the M dimensional affine space defined by $\boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{\zeta}_{2*}$, in which a point does not give a critical point in general, but defines the same input-output function as the critical line.

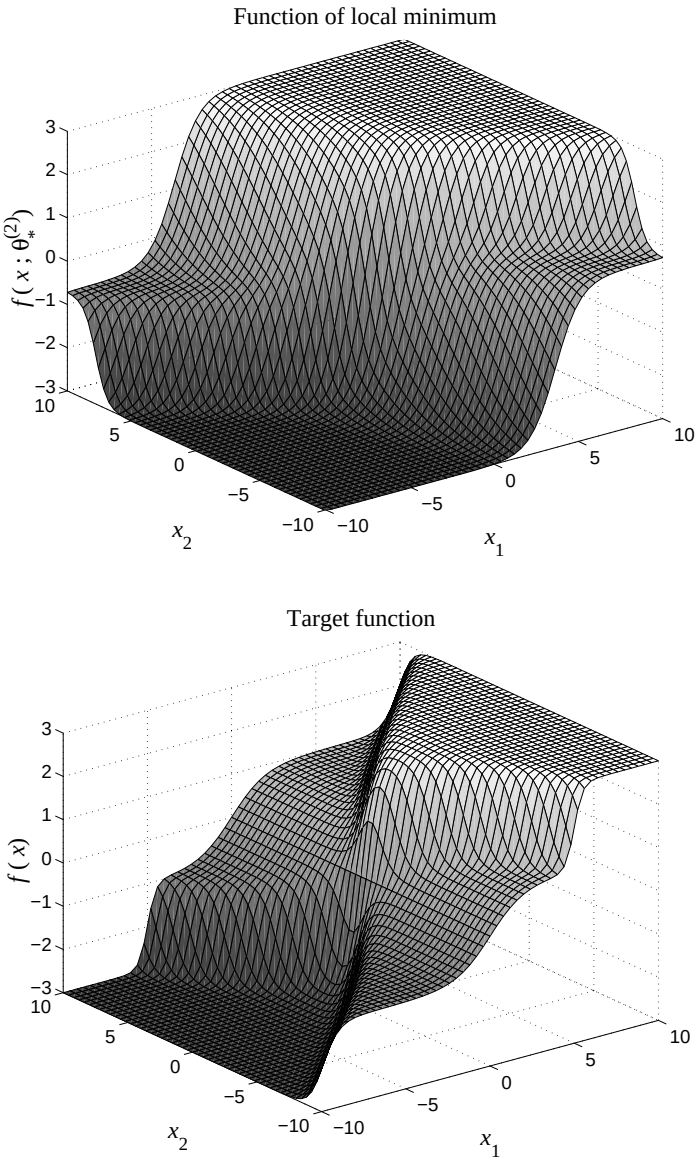

Figure 7: A local minimum and the target in MLP (L=2, H=3).

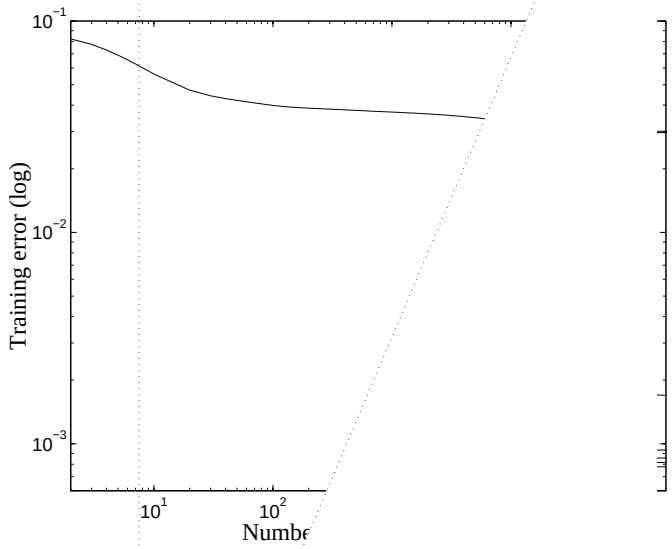

Figure 8: Value of th g

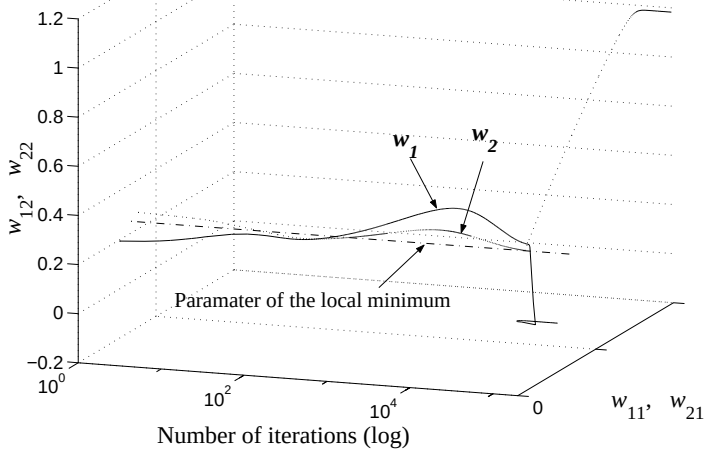

Figure 9: Behavior of parameters

From Lemma 1 in Appendix, we see that any point in the critical line is a saddle. We have not yet known about existence of local minima in the case of multiple output units.

Theorems 3 and 4 mean that the critical points are saddles in many cases. It is very important to know a condition on the positive or negative definiteness of A_2. This is a difficult problem, because it deeply depends on the relation between the global minimum in Θ_{H-1} and the target, and the randomness of training data. From the practical point of view, it is meaningful to see whether the saddle points in Theorem 3 and 4 are the only reason of plateaus. If this is true, we can effectively avoid them by the method of natural gradient ([6],[12],[13],[14]), because it enlarges the gradient of the repulsive direction from Γ by multiplying the inverse of the almost singular Fisher information matrix. However, all of the above problems are left open.

References

[1] P.J.G. Lisboa and S.J. Perantonis. Complete solution of the local minima in the XOR problem. *Network*, 2:119-124, 1991.

[2] L.G.C. Hamney. XOR has no local minima: a case study in neural network error surface analysis. *Neural Networks*, 11(4):669–682, 1998.

[3] I.G. Sprinkhuizen-Kuyper and E.J.W. Boers. The error surface of the 2-2-1 XOR network: the finite stationary points. *Neural Networks*, 11(4):683–690, 1998.

[4] I.G. Sprinkhuizen-Kuyper and E.J.W. Boers. The error surface of the 2-2-1 XOR network: stationary points with infinite weights. Technical Report 96-10, Dept. of Computer Science, Leiden University. 1996.

[5] D. Saad and S. Solla. On-line learning in soft committee machines. *Physical Review E*, 52:4225–4243, 1995.

[6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[7] A.M. Chen, H. Lu, & R. Hecht-Nielsen. On the geometry of feedforward neural network error surfaces. *Neural Computation*, 5:910–927, 1993.

[8] S.M. Rüger, & A. Ossen. The metric of weight space. *Neural Processing Letters*, 5:63–72, 1997.

[9] H.J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5:589–593, 1992.

[10] V. Kůrkova & P.C. Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6:543–558, 1994.

[11] K. Fukumizu. A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, 9(5):871–879, 1996.

[12] H.H. Yang & S. Amari. Complexity issues in natural gradient descent method for training multilayer perceptrons. *Neural Computation*, 10:2137–2157, 1998.

[13] S. Amari, H. Park, & K. Fukumizu. Adaptive method of realizing natural gradient learning for multilayer perceptrons. Submitted to *Neural Computation*. 1999.

[14] M. Rattray, D. Saad, & S. Amari. Natural gradient descent for on-line learning. *Physical review letters* 81:5461–5465. 1998.

Appendix

A Proof of Theorem 3

Proof. For simplicity, we change the order of the components of $\boldsymbol{\theta}^{(H)}$ and $\boldsymbol{\theta}^{(H-1)}$ as $(v_1, v_2, \tilde{\boldsymbol{w}}_1^T, \tilde{\boldsymbol{w}}_2^T, v_0, v_3, \ldots, v_H, \tilde{\boldsymbol{w}}_3^T, \ldots, \tilde{\boldsymbol{w}}_H^T)^T$ and $(\zeta_2, \tilde{\boldsymbol{u}}_2^T, \zeta_0, \zeta_3, \ldots, \zeta_H, \tilde{\boldsymbol{u}}_3^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T$ respectively. We introduce a new coordinate system of Θ_H to see the embedding γ_λ more explicitly. Let $(\xi_1, \boldsymbol{\eta}^T, \xi_2, \boldsymbol{b}^T, v_0, v_3, \ldots, v_H, \tilde{\boldsymbol{w}}_3^T, \ldots, \tilde{\boldsymbol{w}}_H^T)^T$ be a coordinate system of $\Theta_H - \{\boldsymbol{\theta}^{(H)} \mid v_1 + v_2 = 0\}$, where

$$\begin{aligned} \xi_1 &= v_1 - v_2, \\ \boldsymbol{\eta} &= \frac{1}{v_1 + v_2}\big(\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}_2\big), \\ \xi_2 &= v_1 + v_2, \\ \boldsymbol{b} &= \frac{v_1}{v_1 + v_2}\tilde{\boldsymbol{w}}_1 + \frac{v_2}{v_1 + v_2}\tilde{\boldsymbol{w}}_2. \end{aligned} \tag{24}$$

This is well-defined as a coordinate system, since the inverse is given by

$$\begin{aligned} v_1 &= \frac{1}{2}\xi_1 + \frac{1}{2}\xi_2, \\ v_2 &= -\frac{1}{2}\xi_1 + \frac{1}{2}\xi_2, \\ \tilde{\boldsymbol{w}}_1 &= \boldsymbol{b} + \frac{-\xi_1 + \xi_2}{2}\boldsymbol{\eta}, \\ \tilde{\boldsymbol{w}}_2 &= \boldsymbol{b} - \frac{\xi_1 + \xi_2}{2}\boldsymbol{\eta}. \end{aligned} \tag{25}$$

Using this coordinate system, the embedding γ_λ is expressed as

$$\begin{aligned} \gamma_\lambda : (\zeta_2, \tilde{\boldsymbol{u}}_2^T, \zeta_0, \zeta_3, \ldots, \zeta_H, \tilde{\boldsymbol{u}}_3^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T \\ \mapsto ((2\lambda - 1)\zeta_2, \boldsymbol{0}^T, \zeta_2, \tilde{\boldsymbol{u}}_2^T, \zeta_0, \zeta_3, \ldots, \zeta_H, \tilde{\boldsymbol{u}}_3^T, \ldots, \tilde{\boldsymbol{u}}_H^T)^T. \end{aligned} \tag{26}$$

Note that in this definition we use the order of the components introduced at the beginning of the proof.

Let $(\zeta_{2*}, \tilde{\boldsymbol{u}}_{2*}^T, \zeta_{0*}, \zeta_{3*}, \ldots, \zeta_{H*}, \tilde{\boldsymbol{u}}_{3*}^T, \ldots, \tilde{\boldsymbol{u}}_{H*}^T)^T$ be the component of $\boldsymbol{\theta}_*^{(H-1)}$. The critical point set Γ is a one-dimensional affine space parallel to ξ_1-axis with $\boldsymbol{\eta} = \boldsymbol{0}$, $\xi_2 = \zeta_{2*}$, $\boldsymbol{b} = \tilde{\boldsymbol{u}}_{2*}$, $v_0 = \zeta_{0*}$, $v_j = \zeta_{j*}$ $(3 \leq j \leq H)$, and $\tilde{\boldsymbol{w}}_j = \tilde{\boldsymbol{u}}_{j*}$ $(3 \leq j \leq H)$.

Let ξ_{1*} be the ξ_1 component of $\boldsymbol{\theta}_\lambda$, and $V_{\xi_{1*}}$ be a complement of Γ defined by

$$V_{\xi_{1*}} := \{(\xi_1, \boldsymbol{\eta}^T, \xi_2, \boldsymbol{b}^T, v_0, v_3, \ldots, v_H, \tilde{\boldsymbol{w}}_3^T, \ldots, \tilde{\boldsymbol{w}}_H^T)^T \in \Theta_H \mid \xi_1 = \xi_{1*}\}. \tag{27}$$

We have $\Gamma \cap V_{\xi_{1*}} = \boldsymbol{\theta}_\lambda$. If $\boldsymbol{\theta}_\lambda$ is a local minimum in $V_{\xi_{1*}}$ for an arbitrary $\boldsymbol{\theta}_\lambda \in \Gamma_0$, it is a local minimum also in Θ_H, since E_H has the same value on each point of Γ. It is trivial that if $\boldsymbol{\theta}_\lambda$ is a saddle point in $V_{\xi_{1*}}$, it is a saddle also in Θ_H. Thus, we can reduce the problem to the Hessian of E_H restricted on $V_{\xi_{1*}}$. We write it by $\mathcal{G}_{\xi_{1*}}$.

From the definition of $\boldsymbol{\eta}$ and ξ_1, we have

Lemma 1. *For any* $\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \boldsymbol{\eta} = \mathbf{0}\}$,

$$\frac{\partial f}{\partial \boldsymbol{\eta}}(\boldsymbol{x};\boldsymbol{\theta}) = \mathbf{0} \quad \textit{and} \quad \frac{\partial f}{\partial \xi_1}(\boldsymbol{x};\boldsymbol{\theta}) = 0 \tag{28}$$

hold.

From eq.(26), we have also $\frac{\partial f}{\partial \boldsymbol{b}}(\boldsymbol{\theta}_\lambda) = \mathbf{0}$ and $\frac{\partial f}{\partial \xi_2}(\boldsymbol{\theta}_\lambda) = 0$ (this is another proof of Theorem 2). Therefore, the second derivative of E_H at $\boldsymbol{\theta}_\lambda$ can be written as

$$\nabla\nabla E_H(\boldsymbol{\theta}_\lambda) = \sum_{\nu=1}^{N} \frac{\partial \ell}{\partial z}(\boldsymbol{y}^{(\nu)}, f(\boldsymbol{x}^{(\nu)}, \boldsymbol{\theta}_\lambda))\nabla\nabla f(\boldsymbol{x}^{(\nu)}, \boldsymbol{\theta}_\lambda). \tag{29}$$

Let ω represent one of the coordinate components in $(\xi_1, \boldsymbol{\eta}^T, \xi_2, \boldsymbol{b}^T, v_0, v_3, \ldots, v_H, \tilde{\boldsymbol{w}}_3^T, \ldots, \tilde{\boldsymbol{w}}_H^T)$. From Lemma 1, at any point $\boldsymbol{\theta} \in \{\boldsymbol{\eta} = \mathbf{0}\}$, the second derivative $\frac{\partial^2 f}{\partial \xi_1 \partial \omega}(\boldsymbol{\theta}) = \mathbf{0}$ and $\frac{\partial^2 f}{\partial \boldsymbol{\eta} \partial \omega}(\boldsymbol{\theta}) = \mathbf{0}$ unless $\omega = \eta_j$ $(1 \leq j \leq L+1)$. Combining this fact with the expression of eq.(26), we have

$$\mathcal{G}_{\xi_{1*}} = \begin{pmatrix} \frac{\partial^2 E_H}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{\theta}_\lambda) & O \\ O & \frac{\partial^2 E_{H-1}}{\partial \boldsymbol{\theta}^{(H-1)} \partial \boldsymbol{\theta}^{(H-1)}}(\boldsymbol{\theta}_*^{(H-1)}) \end{pmatrix}. \tag{30}$$

By simple calculation, we can derive the following

Lemma 2. *For any* $\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{(H)} \in \Theta_H \mid \boldsymbol{\eta} = \mathbf{0}\}$,

$$\frac{\partial^2 f}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{x}, \boldsymbol{\theta}) = v_1 v_2 \frac{\partial^2 f}{\partial \boldsymbol{b} \partial \boldsymbol{b}}(\boldsymbol{x}, \boldsymbol{\theta}) = v_1 v_2 \xi_2 \varphi''(\boldsymbol{b}^T \tilde{\boldsymbol{x}}) \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^T \tag{31}$$

holds.

From this lemma, we have $\frac{\partial^2 E_H}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{\theta}_\lambda) = \lambda(1-\lambda)\zeta_{2*}^2 A_2$. From the assumption, all the eigenvalues of $\frac{\partial^2 E_{H-1}}{\partial \boldsymbol{\theta}^{(H-1)} \partial \boldsymbol{\theta}^{(H-1)}}(\boldsymbol{\theta}_*^{(H-1)})$ are positive, and $\zeta_{2*} \neq 0$. Thus, if A_2 is positive or negative definite, all the eigenvalues of $\mathcal{G}_{\xi_{1*}}$ at a point in Γ_0 are positive, which means $\boldsymbol{\theta}_\lambda$ is a local minimum in Θ_H. If A_2

has positive and negative eigenvalues, so does $\mathcal{G}_{\xi_{1*}}$ except for two points given by $\lambda = 0, 1$. Then, all the points in $\Gamma - \bar{\Gamma}_0$ are saddle points. As for the two boundary points of Γ_0, any neighborhood of them contains a point of $\Gamma - \bar{\Gamma}_0$. Thus, the neighborhood includes a point attaining larger E_H than $E_H(\boldsymbol{\theta}_\lambda)$ and a point attaining smaller E_H than $E_H(\boldsymbol{\theta}_\lambda)$. Thus, they are also saddle points, and this completes the proof. □

B Proof of Theorem 4

Proof. First, we show the following lemma.

Lemma 3. *Let $E(\boldsymbol{\theta})$ be a function of class C^1, and $\boldsymbol{\theta}_*$ be a critical point of $E(\boldsymbol{\theta})$. If in any neighborhood of $\boldsymbol{\theta}_*$ there exists a point $\boldsymbol{\theta}$ such that $E(\boldsymbol{\theta}) = E(\boldsymbol{\theta}_*)$ and $\frac{\partial E}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) \neq \mathbf{0}$, then $\boldsymbol{\theta}_*$ is a saddle point.*

Proof. Let U be a neighborhood of $\boldsymbol{\theta}_*$. From the assumption, we have a point $\boldsymbol{\theta}_1 \in U$ such that $E(\boldsymbol{\theta}_1) < E(\boldsymbol{\theta}_*)$ and a point $\boldsymbol{\theta}_2 \in U$ such that $E(\boldsymbol{\theta}_2) > E(\boldsymbol{\theta}_*)$. This means $\boldsymbol{\theta}_*$ is a saddle point. □

Back to the proof of Theorem 4, note that $\beta_{(0,w)}(\boldsymbol{\theta}_*^{(H-1)}) \in \{\alpha_{\tilde{\boldsymbol{w}}}(\boldsymbol{\theta}_*^{(H-1)}) \mid \tilde{\boldsymbol{w}} \in \mathbb{R}^{L+1}\}$. In other words, the critical line in Theorem 2 is embedded in an $L+1$ dimensional plane that gives the same function as the critical line. However, the point $\alpha_{\tilde{\boldsymbol{w}}}(\boldsymbol{\theta}_*^{(H-1)})$ is not a critical point for $\boldsymbol{w} \neq \mathbf{0}$, because $\frac{\partial E_H}{\partial v_1} \neq 0$ in general. Thus, $\beta_{(0,w)}(\boldsymbol{\theta}_*^{(H-1)})$ satisfies the assumption of Lemma 3. □