Local Minima and Plateaus in Multilayer Neural Networks

Kenji Fukumizu and Shun-ichi Amari
Brain Science Institute, RIKEN
Hirosawa 2-1, Wako, Saitama 351-0198, Japan
E-mail: {fuku, amari}@brain.riken.go.jp

Abstract

Local minima and plateaus pose a serious problem in learning of neural networks. We investigate the 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 a critical point of the model with H hidden units. Based on this result, we prove that the critical point corresponding to the global minimum of a smaller model can be a local minimum or a saddle point of the larger model. We give a necessary and sufficient condition for this. The results are universal in the sense that they do not use special properties of 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. There have been no rigorous results, however, to prove the existence of local minima. Even in the XOR problem, existence of local minima had been controversial. 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 & Boers ([3]) rigorously proved that what have been believed to be local minima in [1] correspond to local minima with infinite parameter values, and that there are no local minima in the finite weight region for the XOR problem. Existence of local minima in general cases is still an open problem.

It is also difficult to derive meaningful results on local minima from numerical experiments. We often see extremely slow dynamics around a point in simulations. However, it is not easy to tell rigorously whether it is a local minimum. It is known ([4],[5]) that a typical learning curve shows a *plateau* in the middle of training, which causes almost no decrease of the training error. It can be easily misunderstood as a local minimum.

We mathematically investigate critical points of MLP, which are caused by the hierarchical structure of the model. We discuss only networks with one output unit in this paper. The 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 their parameter spaces is not so simple ([6],[7]). We investigate their geometric structure and elucidate how a parameter of a smaller network is embedded in the parameter space of larger networks. We show that a critical point of the error surface for the smaller model gives a set of critical points for the larger model.

The main purpose of this paper is to show that a subset of the critical points corresponding to the global minimum of the smaller model can be local minima of the larger model. The set of critical points is divided into two parts: local minima and saddles. 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 Moreover, the coexistence of local minima and saddles explains a serious mechanism of plateaus: when such is the case, the parameters are attracted in the part of local minima, walk randomly for a long time, but eventually go out from the part of saddles.

2 Geometric structure of the parameter space

2.1 Basic definitions

We consider a three-layer perceptron with one linear output unit and L input unit. The function of a network with H hidden units is defined by

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

where $\boldsymbol{x} \in \mathbb{R}^L$ is an input vector and $\boldsymbol{\theta}^{(H)} = (v_1, \ldots, v_H, \boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_H^T)^T$ is the parameter vector. We do not use bias terms for simplicity. The function $\varphi(t)$ is called an activation function. In this paper, we use *tanh* for φ. However, our results can be easily extended to a wider class of functions with necessary modifications.

Given N training data $\{(\boldsymbol{x}^{(\nu)}, y^{(\nu)})\}_{\nu=1}^N$, the objective of training is to find the parameter that minimizes the error function

$$E_H(\boldsymbol{\theta}) = \sum_{\nu=1}^{N} \ell(y^{(\nu)}, f(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})), \qquad (2)$$

where $\ell(y,z)$ is a loss function. If $\ell(y,z) = \frac{1}{2}\|y-z\|^2$, the objective function is the mean square error. Another popular choice is the cross-entropy. The results in this paper are independent of the choice of a loss function.

2.2 Hierarchical structure of MLP

The parameter $\boldsymbol{\theta}^{(H)}$ consists of a $(L+1)H$ dimensional Euclidean space Θ_H. All the functions eq.(1) realized by Θ_H consist of a function space;

$$\mathcal{S}_H = \{f^{(H)}(\boldsymbol{x};\boldsymbol{\theta}^{(H)}) : \mathbb{R}^L \to \mathbb{R} \mid \boldsymbol{\theta}^{(H)} \in \Theta_H\}. \qquad (3)$$

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

$$\pi_H : \Theta_H \to \mathcal{S}_H, \quad \boldsymbol{\theta}^{(H)} \mapsto f(\boldsymbol{x};\boldsymbol{\theta}^{(H)}). \quad (4)$$

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

A very important point is that π_H is *not* one-to-one, that is, different $\boldsymbol{\theta}^{(H)}$ may give the same function. It is easy to see that 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. For the tanh activation, Chen et al. ([6]) showed that any analytic transform T of Θ_H such that $f^{(H)}(\boldsymbol{x};T(\boldsymbol{\theta})) = f^{(H)}(\boldsymbol{x};\boldsymbol{\theta})$ is a composition of the interchanges and sign flips $(v_j, \boldsymbol{w}_j) \mapsto (-v_j, -\boldsymbol{w}_j)$. These transforms consist of an algebraic group G_H.

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

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \cdots \subset \mathcal{S}_{H-1} \subset \mathcal{S}_H \subset \cdots. \quad (5)$$

On the other hand, given a function $f_{\boldsymbol{\theta}^{(H-1)}}^{(H-1)}$ realized by a network with $H-1$ hidden units, there are a family of parameters $\boldsymbol{\theta}^{(H)} \in \Theta_H$ that realizes $f_{\boldsymbol{\theta}^{(H-1)}}^{(H-1)}$. Mathematically speaking, a map from Θ_{H-1} to Θ_H that commutes the following diagram is not uniquely determined.

$$\begin{array}{ccc} \Theta_{H-1} & \longrightarrow & \Theta_H \\ {\scriptstyle \pi_{H-1}}\downarrow & & \downarrow{\scriptstyle \pi_H} \\ \mathcal{S}_{H-1} & \xrightarrow[\iota_{H-1}]{} & \mathcal{S}_H \end{array} \qquad (6)$$

The set of all the parameters $\boldsymbol{\theta}^{(H)}$ that realize the functions of smaller networks is denoted by Ω_H; $\Omega_H = \pi_H^{-1}(\iota_{H-1}(\mathcal{S}_{H-1}))$. Sussmann ([7]) shows that Ω_H is the union of the following three kinds of submanifolds of Θ_H;

$$\begin{aligned} \mathcal{A}_j &= \{v_j = 0\} \quad (1 \le j \le H), \\ \mathcal{B}_j &= \{\boldsymbol{w}_j = \mathbf{o}\} \quad (1 \le j \le H), \\ \mathcal{C}_{j_1 j_2}^{\pm} &= \{\boldsymbol{w}_{j_1} = \pm \boldsymbol{w}_{j_2}\} \quad (j_1 < j_2). \end{aligned}$$

Fig.1 illustrates these parameters. In $\mathcal{A}_j$ and $\mathcal{B}_j$, the jth hidden unit plays no role in the value of the input-output function. In $\mathcal{C}_{j_1 j_2}^{\pm}$, the j_1th and j_2th hidden unit can be integrated into one, 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 known ([8]) that Ω_H is the set of all the points at which the Fisher information is singular.

Next, we will see how a specific function in the smaller model is embedded in the parameter space of the larger model. Let $f_{\boldsymbol{\theta}^{(H-1)}}^{(H-1)}$ be a function in $\mathcal{S}_{H-1} - \mathcal{S}_{H-2}$. To distinguish Θ_{H-1} and Θ_H, we use different parameter variables and indexing;

$$f^{(H-1)}(\boldsymbol{x};\boldsymbol{\theta}^{(H-1)}) = \sum_{j=2}^{H} \zeta_j \varphi(\boldsymbol{u}_j^T \boldsymbol{x}). \quad (7)$$

Then, given $\boldsymbol{\theta}^{(H-1)}$, the parameter set in Θ_H realizing $f_{\boldsymbol{\theta}^{(H-1)}}^{(H-1)}$ is the union of the submanifolds in each of $\mathcal{A}_j$, $\mathcal{B}_j$ and $\mathcal{C}_{j_1 j_2}^{\pm}$. For

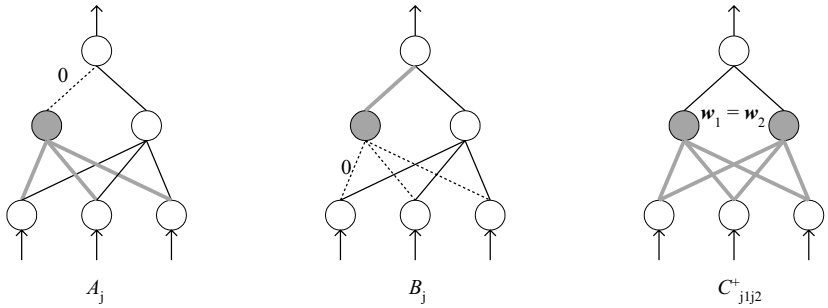

Figure 1: Networks given by $\mathcal{A}_j$, $\mathcal{B}_j$ and $\mathcal{C}^{\pm}_{j_1 j_2}$.

simplicity, we show only an example of the submanifolds of $\mathcal{A}_1$, $\mathcal{B}_1$ and $\mathcal{C}^{\pm}_{12}$;

$$\begin{aligned}\Lambda =&\{v_1 = 0, v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2),\\ &\boldsymbol{w}_1 : \text{free}\},\\ \Xi =&\{\boldsymbol{w}_1 = \boldsymbol{o}, v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2),\\ &v_1 : \text{free}\},\\ \Gamma^{\pm} =&\{\boldsymbol{w}_1 = \pm\boldsymbol{w}_2 = \boldsymbol{u}_2, v_1 \pm v_2 = \zeta_2,\\ &v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2)\}. \end{aligned} \tag{8}$$

All the other sets of parameters realizing $f^{(H-1)}_{\boldsymbol{\theta}^{(H-1)}}$ are obtained as the transform of Λ, Ξ, and $\Gamma^{\pm}$ by $T \in G_H$. The submanifold Λ is a L dimensional affine space parallel to the $\boldsymbol{w}_1$-plane, Ξ is a line with an arbitrary v_1, and $\Gamma^{\pm}$ is a line defined by $v_1 \pm v_2 = \zeta_2$ in the $v_1 v_2$-plane. Thus, each function of a smaller network is realized by high-dimensional submanifolds in Θ_H.

For further analysis, we define canonical embeddings of Θ_{H-1} into Θ_H, which commute the diagram (6);

$$\begin{aligned}\alpha_{\boldsymbol{w}} :&\boldsymbol{\theta}^{(H-1)} \mapsto (0, \zeta_2, \ldots, \zeta_H, \boldsymbol{w}, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_H),\\ \beta_v :&\boldsymbol{\theta}^{(H-1)} \mapsto (v, \zeta_2, \ldots, \zeta_H, \boldsymbol{o}, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_H),\\ \gamma^{\pm}_{\lambda} :&\boldsymbol{\theta}^{(H-1)} \mapsto (\lambda\zeta_2, \pm(1-\lambda)\zeta_2, \zeta_3, \ldots, \zeta_H,\\ &\boldsymbol{u}_2, \pm\boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_H). \end{aligned} \tag{9}$$

where $\boldsymbol{w} \in \mathbb{R}^L$, $v \in \mathbb{R}$, and $\lambda \in \mathbb{R}$ are their parameters. We see that the images of these embedding changing its parameter span the submanifolds Λ, Ξ, and $\Gamma^{\pm}$ respectively; that is,

$$\begin{aligned}\Lambda &= \{\alpha_{\boldsymbol{w}}(\boldsymbol{\theta}^{(H-1)}) \mid \boldsymbol{w} \in \mathbb{R}^L\},\\ \Xi &= \{\beta_v(\boldsymbol{\theta}^{(H-1)}) \mid v \in \mathbb{R}\},\\ \Gamma^{\pm} &= \{\gamma^{\pm}_{\lambda}(\boldsymbol{\theta}^{(H-1)}) \mid \lambda \in \mathbb{R}\}.\end{aligned}$$

3 Critical points of MLP

Generally, the optimum parameter cannot be calculated analytically, and some numerical optimization method is needed to obtain an approximation of the global minimum of E_H. One widely-used method is the steepest descent method, which leads to a learning rule defined by

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \delta \frac{\partial E_H(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}. \tag{10}$$

However, this learning rule stops at a *critical* point, which satisfies $\frac{\partial E_H}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = 0$, even if $\boldsymbol{\theta}$ is not the global minimum.

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 $\boldsymbol{\theta}$ 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 any neighborhood of $\boldsymbol{\theta}_0$ there exists 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 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.

We look for a critical point of E_H in Ω_H. Let $\boldsymbol{\theta}^{(H-1)}_* = (\zeta_{2*}, \ldots, \zeta_{H*}, \boldsymbol{u}_{2*}, \ldots, \boldsymbol{u}_{H*}) \in \Theta_{H-1} - \Theta_{H-2}$ be a critical point of E_{H-1}. It really exists if we assume that the global minimum of E_{H-1} is not included in Θ_{H-2}. Then, we have

$$\begin{aligned}&\textstyle\sum_{\nu=1}^N \partial_z \ell^{(\nu)} \varphi(\boldsymbol{u}^T_{j*} \boldsymbol{x}^{(\nu)}) = 0,\\ &\zeta_{j*} \textstyle\sum_{\nu=1}^N \partial_z \ell^{(\nu)} \varphi'(\boldsymbol{u}^T_{j*} \boldsymbol{x}^{(\nu)}) \boldsymbol{x}^{(\nu)} = 0, \end{aligned} \tag{11}$$

for $2 \leq j \leq H$, where we use

$$\partial_z \ell^{(\nu)} = \frac{\partial \ell}{\partial z}(y^{(\nu)}, f^{(H-1)}(\boldsymbol{x}^{(\nu)}; \boldsymbol{\theta}_*^{(H-1)})) \tag{12}$$

for simplicity. We have two kinds of critical points as follows.

Theorem 1. *Let β_v and $\gamma_\lambda^\pm$ be as in eq.(9). Then, $\gamma_\lambda^\pm(\boldsymbol{\theta}_*^{(H-1)})$ for all λ and $\beta_0(\boldsymbol{\theta}_*^{(H-1)})$ are critical points of E_H.*

The proof is easy. Noting that $f^{(H-1)}(\boldsymbol{x}; \boldsymbol{\theta}_*^{(H-1)}) = f^{(H)}(\boldsymbol{x}; \boldsymbol{\theta})$ for $\boldsymbol{\theta} = \gamma_\lambda^\pm(\boldsymbol{\theta}_*^{(H-1)})$ or $\beta_0(\boldsymbol{\theta}_*^{(H-1)})$, the condition of a critical point of E_H can be reduced to eq.(11).

Because $\alpha_{\mathbf{o}} = \beta_0$, they give the same critical point. The critical point $\gamma_\lambda^\pm(\boldsymbol{\theta}_*^{(H-1)})$ consist of a line in Θ_H if we move $\lambda \in \mathbb{R}$. If $\boldsymbol{\theta}$ is a critical point of E_H, so is $T(\boldsymbol{\theta})$ for all $T \in G_H$. We have many critical lines in Θ_H.

4 Local minima of MLP

4.1 A condition of local minima

In this section, we focus on the critical point $\gamma_\lambda^\pm(\boldsymbol{\theta}_*^{(H-1)})$, and show a condition that it is a local minimum or a saddle point. The usual sufficient condition using the Hessian matrix cannot be applied in this case. The Hessian is singular, because a line including the point has the same value of E_H in common.

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

$$A_2 = \zeta_{2*} \sum_{\nu=1}^{N} \partial_z \ell^{(\nu)} \varphi''(\boldsymbol{u}_{2*}^T \boldsymbol{x}^{(\nu)}) \boldsymbol{x}^{(\nu)} \boldsymbol{x}^{(\nu)T}. \tag{13}$$

Theorem 2. *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 γ_λ^+ or γ_λ^- in eq.(9), and $\Gamma = \{\boldsymbol{\theta}_\lambda \in \Theta_H | \boldsymbol{\theta}_\lambda = \gamma_\lambda(\boldsymbol{\theta}_*^{(H-1)}), \lambda \in \mathbb{R}\}$. If A_2 is positive (negative) definite, any point in the set $\Gamma_0 = \{\boldsymbol{\theta}_\lambda \in \Gamma \mid \lambda(1-\lambda) > 0 \quad (< 0)\}$ 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.*

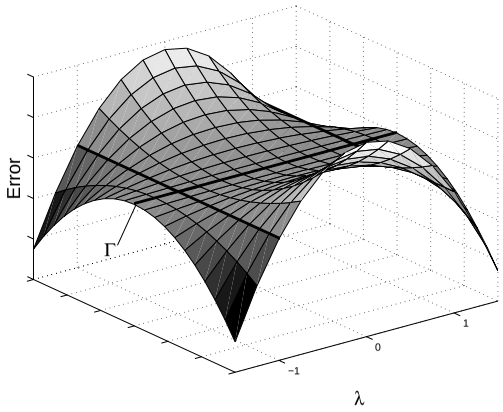

Figure 2: Error surface around local minima

For the proof, see Appendix. Local minima given by Theorem 2, if any, appear as line segments. Such a local minimum can be changed into a saddle through the line without altering the function $f_{\boldsymbol{\theta}}^{(H)}$. Fig.2 illustrates the error surface in this case.

4.2 Plateaus

We consider the case where A_2 is positive (negative) definite. If we map Γ to the function space, $\pi_H(\Gamma)$ consists of a single function $f_{\boldsymbol{\theta}^{(H)}}^{(H)} \in \mathcal{S}_H$. Therefore, if we regard 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 $f_{\boldsymbol{\theta}^{(H)}}^{(H)}$ 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, Γ_0 is neutrally stable in the direction of Γ_0, so that the point attracted to Γ_0 fluctuates randomly along Γ_0. It eventually escapes from Γ when it reaches $\Gamma - \Gamma_0$. This takes a long time because of the nature of random walk. 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 Numerical simulation

We have tried a numerical simulation to exemplify local minima given by Theorem 2, using MLP with 1 input, 1 output, and 2 hidden units. We use the logistic function

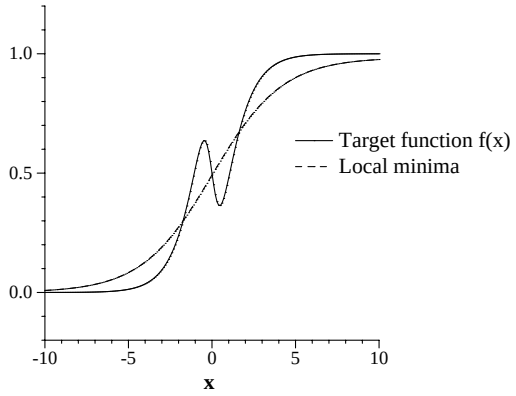

Figure 3: A local minimum in MLP.

$\varphi(t) = \frac{1}{1+e^{-t}}$ for the activation function, and $\ell(y,z) = \frac{1}{2}\|y-z\|^2$ for the loss function. For training data, 100 random input data are generated, and output data are obtained as $y = f(x)+Z$, where $f(x) = 2\varphi(x)-\varphi(4x)$ and Z is a Gaussian noise with 10^{-4} as its variance. Using back-propagation, we train the parameter of MLP with 1 hidden unit, and use it as the global minimum $\boldsymbol{\theta}_*^{(1)}$. In this case, we have $(\zeta_{2*}, u_{2*}) = (0.98, 0.47)$ and $A_2 = 1.91 > 0$. Then, any point in $\Gamma_0 = \{\gamma_\lambda^+(\boldsymbol{\theta}_*^{(1)}) | 0 < \lambda < 1\}$ is a local minimum. We set $v_1 = v_2 = \zeta_{2*}/2$ as $\boldsymbol{\theta}_\lambda$ $(\lambda = 1/2)$, and evaluate $E_2(\boldsymbol{\theta})$ at 1 million random points around $\boldsymbol{\theta}_\lambda$, which are generated by a normal distribution with 10^{-6} as its variance. As a result, all these values are larger than $E_2(\boldsymbol{\theta}_\lambda)$. This experimentally verifies that $\boldsymbol{\theta}_\lambda$ is a local minimum. The graphs of the target function and the function given by the local minimum $f(x; \boldsymbol{\theta}_*^{(1)})$ are shown in Fig.3.

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 set of critical points. We further elucidated a condition that a point in the image of one embedding is a local minimum. 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. It is important to see whether the critical sets in Theorem 2 are the only reason of plateaus. If this is the case, we can avoid them by the method of natural gradient ([5],[9]). However, this is still left as an open problem.

References

[1] P.J.G. Lisboa & 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 & 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] D. Saad & S. Solla. On-line learning in soft committee machines. *Physical Review E*, 52:4225–4243, 1995.

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

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

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

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

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

Appendix

A Proof of Theorem 2

We have only to prove the theorem on γ_λ^+, because the sign flip $(v_2, \boldsymbol{w}_2) \mapsto (-v_2, -\boldsymbol{w}_2)$ maps $\gamma_\lambda^-(\boldsymbol{\theta}_*^{(H-1)})$ to $\gamma_\lambda^+(\boldsymbol{\theta}_*^{(H-1)})$ preserving the local property of E_H.

For simplicity, we change the order of the components of $\boldsymbol{\theta}^{(H)}$ and $\boldsymbol{\theta}^{(H-1)}$ as $(v_1, v_2, \boldsymbol{w}_1^T, \boldsymbol{w}_2^T, v_0, v_3, \ldots, v_H, \boldsymbol{w}_3^T, \ldots, \boldsymbol{w}_H^T)$ and $(\zeta_2, \boldsymbol{u}_2^T, \zeta_0, \zeta_3, \ldots, \zeta_H, \boldsymbol{u}_3^T, \ldots, \boldsymbol{u}_H^T)$ respectively.

We introduce a new coordinate system of Θ_H. Let $(\xi_1, \boldsymbol{\eta}^T, \xi_2, \boldsymbol{b}^T, v_3, \ldots, v_H, \boldsymbol{w}_3^T, \ldots, \boldsymbol{w}_H^T)$ be a coordinate system of $\Theta_H - \{v_1 + v_2 = 0\}$, where

$$\begin{aligned} \xi_1 &= v_1 - v_2, \quad \boldsymbol{\eta} = \frac{1}{v_1 + v_2}(\boldsymbol{w}_1 - \boldsymbol{w}_2), \\ \xi_2 &= v_1 + v_2, \quad \boldsymbol{b} = \frac{v_1}{v_1 + v_2}\boldsymbol{w}_1 + \frac{v_2}{v_1 + v_2}\boldsymbol{w}_2. \end{aligned} \tag{14}$$

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

$$\begin{aligned} v_1 &= \frac{1}{2}\xi_1 + \frac{1}{2}\xi_2, \quad \boldsymbol{w}_1 = \boldsymbol{b} + \frac{-\xi_1 + \xi_2}{2}\boldsymbol{\eta}, \\ v_2 &= -\frac{1}{2}\xi_1 + \frac{1}{2}\xi_2, \quad \boldsymbol{w}_2 = \boldsymbol{b} - \frac{\xi_1 + \xi_2}{2}\boldsymbol{\eta}. \end{aligned} \tag{15}$$

Using this coordinate system, the embedding γ_λ is expressed as

$$\begin{aligned} &\gamma_\lambda : (\zeta_2, \boldsymbol{u}_2, \zeta_3, \ldots, \zeta_H, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_H) \mapsto \\ &((2\lambda - 1)\zeta_2, \mathbf{o}, \zeta_2, \boldsymbol{u}_2, \zeta_3, \ldots, \zeta_H, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_H). \end{aligned} \tag{16}$$

The critical point set Γ is a line parallel to the ξ_1-axis with $\boldsymbol{\eta} = 0$, $\xi_2 = \zeta_{2*}$, $\boldsymbol{b} = \boldsymbol{u}_{2*}$, $v_j = \zeta_{j*}$, and $\boldsymbol{w}_j = \boldsymbol{u}_{j*}$ $(3 \le j \le H)$.

Let ξ_{1*} be the ξ_1 component of $\boldsymbol{\theta}_\lambda$, and $V_{\xi_{1*}} = \{\boldsymbol{\theta} \in \Theta_H \mid \xi_1 = \xi_{1*}\}$ be a complement space of Γ. We have $\Gamma \cap V_{\xi_{1*}} = \boldsymbol{\theta}_\lambda$. If $\boldsymbol{\theta}_\lambda$ is a local minimum in $V_{\xi_{1*}}$ for any $\boldsymbol{\theta}_\lambda \in \Gamma_0$, it is a local minimum also in Θ_H, and if $\boldsymbol{\theta}_\lambda$ is a saddle 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 the Hessian by $\mathcal{G}_{\xi_{1*}}$.

From eq.(15), we have

Lemma 1. *For any $\boldsymbol{\theta} \in \{\boldsymbol{\eta} = \mathbf{o}\}$, we have*

$$\frac{\partial f}{\partial \boldsymbol{\eta}}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbf{o} \quad \textit{and} \quad \frac{\partial f}{\partial \xi_1}(\boldsymbol{x}; \boldsymbol{\theta}) = 0. \tag{17}$$

From eq.(16), $\frac{\partial f}{\partial \boldsymbol{b}}(\boldsymbol{\theta}_\lambda) = \mathbf{o}$ and $\frac{\partial f}{\partial \xi_2}(\boldsymbol{\theta}_\lambda) = 0$ also hold. Therefore,

$$\nabla\nabla E_H(\boldsymbol{\theta}_\lambda) = \sum_{\nu=1}^{N} \partial_z \ell^{(\nu)} \nabla\nabla f(\boldsymbol{x}^{(\nu)}, \boldsymbol{\theta}_\lambda). \tag{18}$$

From Lemma 1, for any $\boldsymbol{\theta} \in \{\boldsymbol{\eta} = \mathbf{o}\}$ $\frac{\partial^2 f}{\partial \xi_1 \partial \omega}(\boldsymbol{\theta}) = 0$ and $\frac{\partial^2 f}{\partial \boldsymbol{\eta} \partial \omega}(\boldsymbol{\theta}) = \mathbf{o}$ hold unless $\omega = \boldsymbol{\eta}$. Then, from eq.(16), we have

$$\mathcal{G}_{\xi_{1*}} = \begin{pmatrix} \frac{\partial^2 E_H}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{\theta}_\lambda) & O \\ O & \dfrac{\partial^2 E_{H-1}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}(\boldsymbol{\theta}_*^{(H-1)}) \end{pmatrix}. \tag{19}$$

By simple calculation, we can prove

Lemma 2. *For any $\boldsymbol{\theta} \in \{\boldsymbol{\eta} = \mathbf{o}\}$, we have*

$$\frac{\partial^2 f}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{x}, \boldsymbol{\theta}) = v_1 v_2 \xi_2 \varphi''(\boldsymbol{b}^T \boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^T. \tag{20}$$

From eq.(18) and Lemma 2, we have

$$\frac{\partial^2 E_H}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}(\boldsymbol{\theta}_\lambda) = \lambda(1 - \lambda)\zeta_{2*}^2 A_2. \tag{21}$$

Noting that $\frac{\partial^2 E_{H-1}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}}(\boldsymbol{\theta}_*^{(H-1)})$ is positive definite and $\zeta_{2*} \neq 0$, if $\lambda(1 - \lambda)A_2$ is positive definite, so is $\mathcal{G}_{\xi_{1*}}$, and if $\lambda(1 - \lambda)A_2$ has negative eigenvalues, $\mathcal{G}_{\xi_{1*}}$ has both positive and negative eigenvalues. This completes the proof.