三層ニューラルネットワークのローカルミニマ

福水 健次（ＰＹ）, 甘利 俊一
理化学研究所　脳科学総合研究センター

Local minima of three-layer neural networks

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

Abstract

Although local minima pose a serious problem in multilayer networks, theoretical discussion on their existence is a difficult problem. We investigate the geometric structure of the parameter space of three-layer networks, and show the existence of critical points of the error surface. We further prove that the critical points can be local minima under some condition.

1. はじめに

多層パーセプトロンの誤差曲面には、一般に多くのローカルミニマが存在すると考えられ、それを避けるための工夫が提案されてきた。しかし、極小点の存在を理論的に解析するのは容易ではなく、XOR 問題ですら、ローカルミニマが存在しないことが最近ようやく証明された（[1]）。このように、ローカルミニマの存在に関しては依然未解決の部分が多い。

本稿では、3 層ネットワークの階層的な構造に着目し、誤差曲面の危点を解析する。3 層ネットワークでは、$H-1$ 個の中間素子で実現される関数は、H 個の中間素子を持つモデルのパラメータ空間の中に、部分多様体として埋め込まれている。この埋め込まれ方を考察し、$H-1$ 個の中間素子を持つモデルでの危点を埋め込んだ像が、やはり危点になっていることを示す。さらに、ある条件のもとで、この危点集合が極小点になることを示す。

2. パラメータ空間と関数空間

本稿では L 入力、1 出力の 3 層ネットワークを考察する。中間素子を H 個持つネットワークは、

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

とかける。$\boldsymbol{\theta}^{(H)} = (\boldsymbol{w}_j, v_j)$ はモデルの持つパラメータであり、簡単のため閾値は用いないことにする。$\varphi(t)$ としては tanh を考えるが、以下の結果は広いクラスの関数に拡張可能である。

N 個の学習データ $\{(\boldsymbol{x}_\nu, y_\nu)\}_{\nu=1}^N$ に対して、

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

キーワード：ローカルミニマ、多層パーセプトロン

を学習の目的関数とする。ここで $\ell(y,z)$ は損失関数であり、$\frac{1}{2}\|y-z\|^2$ などがよく使われる。

H 個の中間素子を持つモデルの「パラメータ」$\boldsymbol{\theta}^{(H)}$ の全体は $\mathbb{R}^{(L+1)H}$ をなすが、これを Θ_H で表わす。また、これにより実現される「関数」の全体を

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

で表わす。パラメータを一つ決めれば関数が一つきまるので、Θ_H から $\mathcal{S}_H$ へは自然な全射

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

が定まる。この対応 π_H が 1 対 1 でないことが重要である。たとえば、中間素子の交換 $(v_{j_1}, \boldsymbol{w}_{j_1}) \leftrightarrow (v_{j_2}, \boldsymbol{w}_{j_2})$ によって実現される関数は変化しない。また、$\varphi(t) = \tanh(t)$ の場合には符合反転 $(v_j, \boldsymbol{w}_j) \mapsto (-v_j, -\boldsymbol{w}_j)$ でも実現される関数は変らない。

3. モデルの階層構造

関数空間 $\mathcal{S}_H$ $(H = 0, 1, 2, \dots)$ には自然な包含関係

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

があるが、パラメータ空間 Θ_H の間の関係は単純ではない。$H-1$ 個の中間素子で実現できる関数 $f^{(H-1)}_{\boldsymbol{\theta}^{(H-1)}}$ に対して、H 個の中間素子を持つネットワークのパラメータでこれを実現するものは数多くあり、埋め込み写像 $\Theta_{H-1} \to \Theta_H$ は一意的には決まらない（[2]）。

$\mathcal{S}_{H-1} - \mathcal{S}_{H-2}$ に含まれる関数を

$$f^{(H-1)}(\boldsymbol{x};\boldsymbol{\theta}^{(H-1)}) = \sum_{j=2}^{H} \zeta_j \varphi(\boldsymbol{u}_j^T \boldsymbol{x}) \qquad (6)$$

と書くことすると、中間素子 H 個のモデルのパラメータ空間Θ_H の要素で $f^{(H-1)}(\cdot;\boldsymbol{\theta}^{(H-1)})$ を実現す

るものは、次の高次元部分多様体と、それらを中間素子の交換によって変換したもの全体の和集合になる。

$$\begin{aligned}\Lambda =&\{v_1 = 0, v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2), \boldsymbol{w}_1 : \text{free}\},\\ \Xi =&\{\boldsymbol{w}_1 = \mathbf{o}, v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2), v_1 : \text{free}\},\\ \Gamma^{\pm} =&\{\boldsymbol{w}_1 = \pm\boldsymbol{w}_2 = \boldsymbol{u}_2, v_1 \pm v_2 = \zeta_2,\\ & v_j = \zeta_j, \boldsymbol{w}_j = \boldsymbol{u}_j, (j \geq 2)\}. \end{aligned} \tag{7}$$

特に、$\Gamma^{\pm}$ は $v_1 \pm v_2 = \zeta_2$ で定まる直線となっている。

ここで、自然な埋め込み $\Theta_{H-1} \to \Theta_H$ として

$$\begin{aligned}\alpha_{\boldsymbol{w}} :&\boldsymbol{\theta}^{(H-1)} \mapsto (0, \zeta_2, \dots, \zeta_H, \boldsymbol{w}, \boldsymbol{u}_2, \dots, \boldsymbol{u}_H),\\ \beta_v :&\boldsymbol{\theta}^{(H-1)} \mapsto (v, \zeta_2, \dots, \zeta_H, \mathbf{o}, \boldsymbol{u}_2, \dots, \boldsymbol{u}_H),\\ \gamma_{\lambda}^{\pm} :&\boldsymbol{\theta}^{(H-1)} \mapsto (\lambda\zeta_2, \pm(1-\lambda)\zeta_2, \zeta_3, \dots, \zeta_H,\\ & \boldsymbol{u}_2, \pm\boldsymbol{u}_2, \boldsymbol{u}_3, \dots, \boldsymbol{u}_H)\end{aligned} \tag{8}$$

を定義する。$\boldsymbol{w} \in \mathbb{R}^L$, $v \in \mathbb{R}$, $\lambda \in \mathbb{R}$ は写像を定めるパラメータであり、これらを動かすと、上の 3 つの写像による $\boldsymbol{\theta}^{(H-1)}$ の像はそれぞれ Λ, Ξ, $\Gamma^{\pm}$ を張る。

4. 三層ネットワークの危点とローカルミニマ

$\boldsymbol{\theta}_*^{(H-1)} = (\zeta_{j*}, \boldsymbol{u}_{j*}) \in \Theta_{H-1} - \Theta_{H-2}$ を E_{H-1} の危点とする。このとき、(8) 式の埋め込みにより、2 種類の E_H の危点が生じる。

Theorem 1. $\gamma_{\lambda}^{\pm}(\boldsymbol{\theta}_*^{(H-1)})$ （$^{\forall}\lambda \in \mathbb{R}$）と$\beta_0(\boldsymbol{\theta}_*^{(H-1)})$ は E_H の危点である。

$\gamma_{\lambda}^{\pm}(\boldsymbol{\theta}_*^{(H-1)})$ は λ を動かすと直線を成すので、中間素子の交換で生じるものも考えれば、E_H には多くの危点集合（直線）が存在する。

この危点が極小点になるための条件を述べるため、$\boldsymbol{\theta}_*^{(H-1)} \in \Theta_{H-1}$ に対し、$L \times L$ 対称行列 A_2 を

$$A_2 = \zeta_{2*} \sum_{\nu=1}^{N} \frac{\partial \ell(y_\nu, f(\boldsymbol{x}_\nu; \boldsymbol{\theta}_*))}{\partial z} \varphi''(\boldsymbol{u}_{2*}^T \boldsymbol{x}_\nu) \boldsymbol{x}_\nu \boldsymbol{x}_\nu^T \tag{9}$$

で定義する。このとき、次の定理が成立する。

Theorem 2. $\boldsymbol{\theta}_*^{(H-1)}$ を E_{H-1} の極小点とし、$\boldsymbol{\theta}_*^{(H-1)}$ における *Hesse* 行列が正定値であると仮定する。γ_λ を *(8)* 式のγ_λ^+, γ_λ^- のいづれかとし、$\Gamma = \{\boldsymbol{\theta}_\lambda \in \Theta_H | \boldsymbol{\theta}_\lambda = \gamma_\lambda(\boldsymbol{\theta}_*^{(H-1)}), \lambda \in \mathbb{R}\}$ とおく。もし A_2 が正定値（負定値）ならば、$\Gamma_0 = \{\boldsymbol{\theta}_\lambda \in \Gamma \mid \lambda(1-\lambda) > 0 \ (< 0)\}$ の任意の点は E_H の極小点であり、$\Gamma - \Gamma_0$ の点は鞍点である。A_2 が正負両方の固有値を持つならば、Γ の任意の点は E_H の鞍点である。

この定理からわかるように、ローカルミニマは、存在するならば線分として出現し、関数を変化させること無く鞍点に変えることができる（図 1）。

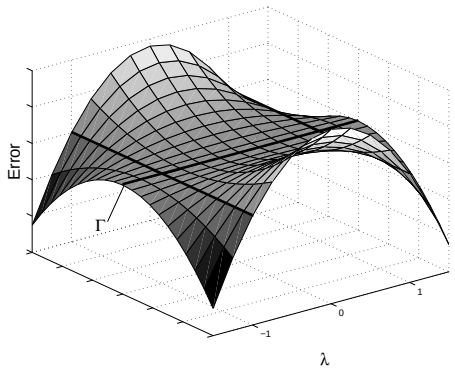

図 1: 極小点近傍の誤差曲面の様子

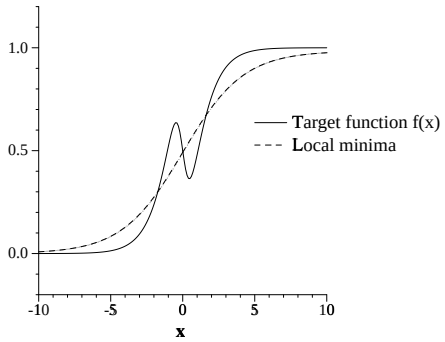

図 2: ローカルミニマを与える関数

定理の内容を実験的に確認するため数値実験を行なった。1 入力 1 出力、2 個の中間素子を持つネットワークを用い、正解の関数を $f(x) = 2\varphi(x) - \varphi(4x)$ として、出力にガウス雑音を加えた 100 個の学習データを用意した。中間素子 1 個のモデルでの最小点を BP 学習で求めると、$A_2 = 1.91 > 0$ となったので、$\Gamma_0 = \{\gamma_\lambda^+(\boldsymbol{\theta}_*^{(1)}) | 0 < \lambda < 1\}$ が極小点となる。実際、$\boldsymbol{\theta}_\lambda$ $(\lambda = 1/2)$ の近傍で $E_2(\boldsymbol{\theta})$ を評価したところ、$\boldsymbol{\theta}_\lambda$ がローカルミニマであることが数値的に確認できた。実現される関数 $f(x; \boldsymbol{\theta}_\lambda)$ の形は図 2 のようであった。

6. おわりに

3 層ネットワークのパラメータ空間の構造を解析することにより、誤差曲面には多くの危点集合が存在すること、および、ある条件のもとで、極小点が線分群として存在することを示した。これにより、ニューラルネットのローカルミニマが、モデルの構造に根差して広範囲に存在することが明らかとなった。

参考文献

[1] I.G. Sprinkhuizen-Kuyper & E.J.W. Boers. *Neural Networks*, 11(4):683–690, 1998.

[2] H.J. Sussmann. *Neural Networks*, 5:589–593, 1992.