ニューラルネットの推定理論
— モデルの対称性と識別不能性 —

福水健次
統計数理研究所

要旨

本論文は、多層ニューラルネットがモデルの構造として持つ対称性に注目し、そこから生じるパラメータの識別不能性に関する 2 つの問題を論じる。第 1 は、ニューラルネットを用いた場合の経験損失関数の臨界点やローカルミニマについてである。ニューラルネットのような対称性を持つモデルでは、構造上臨界点が必ず存在することを示し、それが極小点になるための十分条件を示す。第 2 は、真のパラメータが識別不能となる場合の尤度比の漸近論である。識別不能性を持つモデルは有限混合分布や ARMA など多く見られるが、これらを一般的に扱う枠組みとして局所錐型モデルを紹介し、データ数に対する尤度比のオーダーが通常よりも大きいための十分条件を示す。この結果をニューラルネットに応用し、さらに精密なオーダーの下界を求める。

1 はじめに

ニューラルネットモデルは、特に多層パーセプトロンの提案以来、工業製品への応用をはじめ時系列解析，パターン識別など多くの分野に適用されている。後述するように、ニューラルネットはパラメトリックな非線形回帰のひとつとして定式化することができるが、それを用いる態度としては、個々の問題の構造を分析してモデル化を行うというよりも汎用な関数系としてさまざまな問題に適用するという点に重きがおかれている。このことは、問題の構造を把握しにくい文字認識（[11]）などによく用いられる点にも現れている。汎用的な関数近似系としてのニューラルネットは、多項式など線形の関数系などよりは複雑な構造を持っており、そこから興味深い性質が生じる。

階層型ニューラルネットが構造的に持つ興味深い点のひとつは、モデルの定義式が持つ対称性である。多少単純化すると、階層型ニューラルネットは、パラメータ $\boldsymbol{w}$ を持つある非線形関数 $h(\boldsymbol{x};\boldsymbol{w})$ を用いて定義される、

$$\varphi(x;\boldsymbol{\theta}) = \textstyle\sum_{j=1}^{H} b_j\, h(\boldsymbol{x};\boldsymbol{w}_j) \tag{1}$$

という関数系である。ここで $\boldsymbol{\theta} = (\boldsymbol{w}_1, b_1, \ldots, \boldsymbol{w}_H, b_H)$ はパラメータである。すぐに気づくように、(1) 式は 2 つの j の交換（中間素子の交換）に対して不変である。従ってパラメータ空間には同一の関数を定義する領域が $H!$ 個存在する。さらに興味深いのはこれら領域の境界、すなわち $a_1 = a_2$ などを満たすパラメータ集合である。この集合上では b_1, b_2 の個別の値は意味をなさず $b_1 + b_2$ の値のみが関数を決めるのに有効である。従って、ひとつの関数を与えるパラメータが連続集合として存在している。このようなパラメータは識別不能であると呼ばれる。

モデルの持つこのような対称性から生まれる問題として、本論文では、パラメータ推定に用いる経験損失関数の臨界点と極小点に関する話題と、真のパラメータが識別不能な場合の尤度比の漸近的挙動に関して論じる。

まず、(1) 式のような対称性を持つモデルの経験損失関数は、その対称の境界上に臨界点を持つことを示し、さらにその臨界点が極小点であるための十分条件を示す ([7])。ニューラルネットのような非線形モデルのパラメータ推定では、経験損失関数の最小値を求めるのに数値的最適化を要する場合が多く、臨界点やローカルミニマは大きな問題である。しかし、ローカルミニマを理論的に議論するのは難しく、その存在すらも未解決の部分が多い。本論文で示す結果は、ローカルミニマや臨界点の存在に対する理論的結果の一つである。また (1) 式の関数形は有限混合モデルの密度関数と酷似しており、類似の議論が有限混合モデルにも適用可能である。

次に尤度比の漸近論に関しては、真のパラメータが識別不能な場合に、ニューラルネットの尤度比が $O_p(1)$ よりも大きいオーダーを持つことを示す。識別不能性は、ニューラルネットに限らず有限混合モデル ([5])，ARMA([16])，変化点問題 ([3]) など重要なモデルの多くに見られるが、真のパラメータが識別不能であると最尤推定量の漸近正規性などは成立せず、モデル選択をはじめ多くの数理統計的手法を再考する必要が生じる。本論文では、識別不能性を扱うための一般的な枠組みである局所錐型モデル ([5]) を紹介し、識別不能な状況下で尤度比のオーダーが $O_p(1)$ より大きくなるための一般的な十分条件を示し、ニューラルネットに応用する。

2　多層ニューラルネットワーク

本論文では出力が 1 次元の 3 層ネットワークのみを扱う。中間素子の非線形関数としてパラメータ $\boldsymbol{w}$ を持つ関数 $h(\boldsymbol{x};\boldsymbol{w})$ を用意する。中間素子を H 個持つ**３層ニューラルネット**は、

$$\varphi(\boldsymbol{x};\boldsymbol{\theta}) = \textstyle\sum_{j=1}^{H} b_j\, h(\boldsymbol{x};\boldsymbol{w}_j) + d \tag{2}$$

により定まる関数族 $\{\varphi(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\boldsymbol{w}_1, b_1, \ldots, \boldsymbol{w}_H, b_H, d)\}$ として定義される。中間素子を H 個持つモデルのパラメータ空間を以降 Θ_H で表すことにする。

３層パーセプトロンモデルとは、中間素子の関数 $h(\boldsymbol{x};\boldsymbol{w})$ として特に

$$h(\boldsymbol{x};\boldsymbol{w}) = \tfrac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x}-c)} \tag{3}$$

（$\boldsymbol{w} = (\boldsymbol{a}, c)$ ）を用いたモデルである（図 1）。このモデルが「ニューラルネットワーク」と呼ばれるのは、もともと脳の神経細胞の数理モデルを単純化したものとして提案されたためである。また、ガウス型の関数 $h(\boldsymbol{x};\boldsymbol{w}) = \exp\bigl\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{a}\|^2\bigr\}$ （$\boldsymbol{w} = (\boldsymbol{a}, \sigma)$）を用いた、Radial Basis Functions (RBF) と呼ばれるモデルもよく用いられる。

多層ニューラルネットでは、中間素子の非線形関数に関する一定の条件のもと、3 層モデルを用いて中間素子の数を増やしていけば、コンパクト集合上の任意の連続関数が sup ノルムに関して任意の精度で近似可能であることが知られており ([4])、3 層モデルがよく用いられる。

関数族 $\{\varphi(\boldsymbol{x};\boldsymbol{\theta})\}$ として定義されたニューラルネットを統計的な枠組みで議論するには、出力 y に対する適当な統計モデル $r(y|s)$ を用意し、固定された $\boldsymbol{x}$ の分布 $q(x)dx$ とともに、$(\boldsymbol{x}, y)$ の同時分布の密度関数 $f(\boldsymbol{x}, y;\boldsymbol{\theta})$ を次式で定義する。

$$f(\boldsymbol{x}, y;\boldsymbol{\theta}) = r(y|\varphi(\boldsymbol{x};\boldsymbol{\theta}))q(x). \tag{4}$$

これにより、ニューラルネットはパラメータ $\boldsymbol{\theta}$ を持つ非線形回帰モデルとして扱うことが出来る。$r(y|s)$ としては、正規雑音を仮定した $r(y|s) = \frac{1}{\sqrt{2\pi}}\exp\bigl\{\frac{1}{2}(y-s)^2\bigr\}$ や、識別問題など二値出力 $y \in \{0, 1\}$ の場合によく用いられる $r(y|s) = \frac{e^{ys}}{1+e^s}$ などが代表的である。

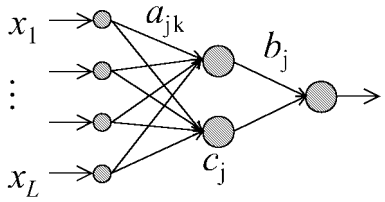

図 1: 3層パーセプトロンモデル

与えられたサンプル（学習データ）$(X_1, Y_1), \ldots, (X_n, Y_n)$ に対して推定量 $\boldsymbol{\theta}$ を得るためには、**損失関数**と呼ばれる 2 変数関数 $\ell(y, s)$ と学習データに対して定義される

$$\ell_n(\boldsymbol{\theta}) = \textstyle\sum_{i=1}^{n} \ell(Y_i, \varphi(X_i, \boldsymbol{\theta})) \tag{5}$$

を最小にするパラメータを求める。$\ell_n(\boldsymbol{\theta})$ のことを**経験損失関数**と呼ぶことにする。$\ell(y, s) = -\log r(y|s)$ とおくと、これは最尤推定に一致する。

(5) 式の最小化問題は、ニューラルネットの非線形性のため解析的に解を求めることは困難であり、最急降下法をはじめとする数値的最適化手法が用いられる。パラメータが逐次的によくなる様子から、最適化の過程を「学習」と呼ぶことも多い。ニューラルネットのような複雑な非線形モデルの経験損失関数は一般にローカルミニマを持つ可能性があり、勾配法に基づく数値的最適化手法を用いると、局所解にとらわれ得るという問題がある。そこで、ローカルミニマをいかにうまく避け学習を高速に行うかといった研究が数多くなされている ([14], [12])。

3　ニューラルネットワークの対称性と識別不能性

3.1　一般の３層ニューラルネットの対称性と識別不能性

(2) 式で定義された３層ニューラルネットは、「はじめに」で述べたように「中間素子の交換に対して関数が不変」という対称性を持つ。この交換によりパラメータを同一視して同値類を取るとこの冗長性は除去できるが、得られた同値類の空間には $\boldsymbol{w}_{j_1} = \boldsymbol{w}_{j_2}$ （$j_1 \neq j_2$）を満足する集合（縁に相当する部分）に特異点が生じる。実際、ほとんどすべてのパラメータに対して同値類は有限集合になるが、$\boldsymbol{w}_{j_1} = \boldsymbol{w}_{j_2}$ を満たす集合内では $b_{j_1} + b_{j_2} =$ 定数 を満たす直線が同一の関数を定義するため、同値類は連続集合からなり、他の点より次元が退化している。さらに、ある j に対して $b_j = 0$ を満たすパラメータに対しては、任意の $\boldsymbol{w}_j$ が同一の関数を定める。すなわち、アフィン平面が同じ関数を定義している。

上の２つの場合、定義される関数は $H - 1$ 個の中間素子で実現可能である。すなわち、ひとつ小さいサイズのネットワークで実現できる関数

$$\varphi_0(\boldsymbol{x}) = \textstyle\sum_{j=2}^{H} \zeta_j^0 h(\boldsymbol{x}; \boldsymbol{u}_j^0) + \delta^0 \tag{6}$$

（添え字のつけ方に注意せよ）に対して、

$$\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{u}_2^0, \qquad b_1 + b_2 = \zeta_2^0, \qquad d = \delta^0, \quad \boldsymbol{w}_j = \boldsymbol{u}_j^0, \quad b_j = \zeta_j^0 \quad (3 \leq j \leq H) \tag{7}$$

によって定義される直線上のパラメータ点と、

$$b_1 = 0, \qquad \boldsymbol{w}_j = \boldsymbol{u}_j^0, \quad b_j = \zeta_j^0 \quad (2 \leq j \leq H), \qquad \boldsymbol{w}_1 : \text{フリー} \tag{8}$$

で定義されるアフィン平面上のパラメータ点は全て $\varphi_0(\boldsymbol{x})$ を定める。第 1 のケースでは任意の 2 つの中間素子の組、第 2 のケースでは任意の中間素子に対して同様の連続集合が定義できる。

一般に統計モデルのパラメータ $\boldsymbol{\theta}$ に対して、パラメータ集合の 1 次元以上の部分多様体が存在して、それが $\boldsymbol{\theta}$ を含み、かつその任意の点が同一の関数を定めるとき、$\boldsymbol{\theta}$ は**識別不能**であると呼ぶことにする。3 層ニューラルネットでは、中間素子の非線形関数に依らず (7) 式と (8) 式で与えられるパラメータ点は識別不能である。また、(2) 式の関数系は有限混合モデルの密度関数の形と類似しており、有限混合モデルにおいても対称性に由来する全く同様の識別不能性が存在する。

3.2 3 層パーセプトロンの識別不能性

ここでは、中間素子の非線形関数が $h(\boldsymbol{x}; \boldsymbol{w}) = \tanh(\boldsymbol{a}^T \boldsymbol{x} + c)$ の場合、すなわち

$$\varphi(\boldsymbol{x}; \boldsymbol{\theta}) = \textstyle\sum_{j=1}^{H} b_j \tanh(\boldsymbol{a}_j^T \boldsymbol{x} + c_j) + d. \tag{9}$$

を考える。ロジスティック関数 $\frac{1}{1+\exp(-t)}$ と $\tanh(t)$ とは t のアフィン変換によって移りあうので、定義される関数族は (3) 式によるものと同一である。

この関数族では、前節で述べた 2 種類の識別不能性に加えて、ある j に対して $a_j = 0$ なるパラメータも識別不能である。実際、$b_j \tanh(c_j) + d = $ 定数 を満たすパラメータは同一の関数を定める。これらをまとめると、

[1] 相異なる j_1, j_2 が存在して、$(a_{j_1}, c_{j_1}) = \pm(a_{j_2}, c_{j_2})$[1].
[2] ある j に対し $a_j = 0$.
[3] ある j に対し $b_j = 0$.

の 3 種の集合上の点は識別不能である（図 2)。ここで注意すべき点は、以上述べた 3 種類の識別不能なパラメータ点は、すべて $H-1$ 個の中間素子で実現可能な関数を定義している点である。さらに次の定理が成立する。

定理 1 ([15],[1],[7]). *(9)* 式で定義される H 個の中間素子を持つ 3 層パーセプトロンにおいて、パラメータが識別不能であるための必要十分条件は、そのパラメータで定義される関数が $H-1$ 個の中間素子を持つ 3 層パーセプトロンで実現できることであり、さらにこれは、上の *[1]-[3]* の条件が成り立つことと同値である。

4 ニューラルネットの臨界点とローカルミニマ

4.1 3 層ニューラルネットの臨界点

前章で見たように、3 層ネットワークのパラメータ空間には、より小さいサイズのモデルによって定まる関数のパラメータが複雑な構造を持って埋め込まれている。このことを使って、3 層ニューラルネットの経験損失関数の構造を探っていく。

[1] tanh が奇関数なので (1) で $(a_{j_1}, c_{j_1}) = -(a_{j_2}, c_{j_2})$ も許される

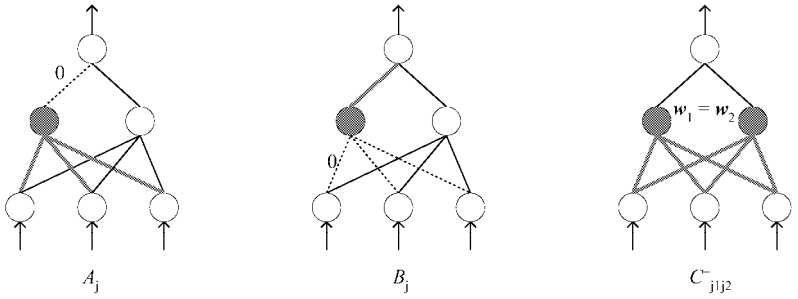

図 2: 識別不能な 3 つの場合

本章では中間素子の個数を強調するため、中間素子を H 個持つ 3 層ネットワークモデルを $\varphi^{(H)}(x;\boldsymbol{\theta})$、その経験損失関数を $\ell_n^{(H)}(\boldsymbol{\theta})$ で表す。経験損失関数の最小値問題の解は、適当な正則化条件のもとで次の推定方程式を満たす。

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ell_n^{(H)}(\boldsymbol{\theta}) = 0. \tag{10}$$

この方程式は最小値問題の解の必要条件であって、一般には十分条件ではない。実際、上の方程式の解は経験損失関数 $\ell_n^{(H)}(\boldsymbol{\theta})$ の臨界点に過ぎず、鞍点、極小点、極大点[2]のいずれなのかはわからない。勾配に基づく数値的最適化手法を用いた場合、ローカルミニマが特に問題となる。

いま、$H-1$ 個の中間素子を持つ 3 層ネットワークを

$$\varphi^{(H-1)}(\boldsymbol{x};\boldsymbol{\omega}) = \textstyle\sum_{j=2}^{H}\zeta_j h(\boldsymbol{x};\boldsymbol{u}_j) + \delta \tag{11}$$

とし、経験損失関数 $\ell_n^{(H-1)}(\boldsymbol{\omega})$ の臨界点を $\boldsymbol{\omega}^* = (\boldsymbol{u}_2^*, \zeta_2^*, \ldots, \boldsymbol{u}_H^*, \zeta_H^*, \delta^*)$ とおく。すると

$$\frac{\partial}{\partial\boldsymbol{\omega}}\ell_n^{(H-1)}(\boldsymbol{\omega}^*) = 0 \tag{12}$$

が成り立つ。小さいモデルは大きいモデルの中に埋め込まれているから、(12) 式の条件は、Θ_H 内で $\boldsymbol{\omega}^*$ に対応する点における $\ell_n^{(H)}$ の微分に関しても多くの情報を有しているはずである。一般には、次元の低い集合上の臨界点が大きい空間内でも臨界点であることは期待できない。余次元方向の方向微分に関しては一般には情報がないからである。しかし 3 層ニューラルネットの場合には $\ell_n^{(H-1)}(\boldsymbol{\omega})$ の臨界点は H 個の中間素子を持つパラメータ空間の中で特殊な構造を持ち、各点が臨界点からなる直線を形成している。

定理 2 (Fukumizu and Amari [7]). *(11)* 式で定義される $H-1$ 個の中間素子を持つ 3 層ニューラルネットの経験損失関数 $\ell_n^{(H-1)}(\boldsymbol{\omega})$ の臨界点 $\boldsymbol{\omega}^*$ が $\zeta_2^* \neq 0$ を満たすとする。$\lambda \in \mathbb{R}$ に対し、$\varphi^{(H-1)}(x;\boldsymbol{\omega}^*)$ と同一の関数を与える Θ_H の点 $\boldsymbol{\theta}(\lambda)$ を

$$a_1 = a_2 = u_2^*, \quad b_1 = \lambda\zeta_2^*, \quad b_2 = (1-\lambda)\zeta_2^*, \quad a_j = u_j^*, \quad b_j = \zeta_j^* \quad (3 \le j \le H) \tag{13}$$

により定める。このとき、任意の $\lambda \in \mathbb{R}$ に対し $\boldsymbol{\theta}(\lambda)$ は $\ell_n^{(H)}(\boldsymbol{\theta})$ の臨界点である。

[2]関数 $F(\boldsymbol{\theta})$ の臨界点とは、$\frac{\partial}{\partial\boldsymbol{\theta}}F(\boldsymbol{\theta}) = 0$ を満たす $\boldsymbol{\theta}$ のことをいう。F の臨界点 $\boldsymbol{\theta}_0$ が極小 (大) 点であるとは、$\boldsymbol{\theta}_0$ のある近傍があって、その上の任意の点 $\boldsymbol{\theta}$ で $F(\boldsymbol{\theta}) \ge (\le) F(\boldsymbol{\theta}_0)$ が成り立つことであり、$\boldsymbol{\theta}_0$ が F の鞍点であるとは、$\boldsymbol{\theta}_0$ の任意の近傍が $F(\boldsymbol{\theta}_1) > F(\boldsymbol{\theta}_0)$ と $F(\boldsymbol{\theta}_2) < F(\boldsymbol{\theta}_0)$ を満たす $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ を含むことをいう。

略証. $\{\boldsymbol{\theta} \in \Theta_H \mid b_1 + b_2 \neq 0\}$ なる空間に新しい座標系 $(\xi_1, \boldsymbol{\eta}, \xi_2, \boldsymbol{\beta}, d, b_3, \boldsymbol{w}_3, \ldots, b_H, \boldsymbol{w}_H)$ を

$$b_1 = \tfrac{1}{2}(\xi_1 + \xi_2),\ \ b_2 = \tfrac{1}{2}(-\xi_1 + \xi_2),\ \ \boldsymbol{w}_1 = \boldsymbol{\beta} + \tfrac{1}{2}(-\xi_1 + \xi_2)\boldsymbol{\eta},\ \ \boldsymbol{w}_2 = \boldsymbol{\beta} - \tfrac{1}{2}(\xi_1 + \xi_2)\boldsymbol{\eta} \quad (14)$$

により導入する。$\lambda \in \mathbb{R}$ に対し $\xi_1 = (2\lambda - 1)\xi_2, \eta = 0$ によって定義されるアフィン部分空間は自然に Θ_{H-1} と同一視され、この同一視のもと $\ell_n^{(H-1)}(\boldsymbol{\omega})$ と $\ell_n^{(H)}(\boldsymbol{\theta})$ が同じ値をとることも容易に確認できる。したがって、$\boldsymbol{\theta}(\lambda)$ が $\ell_n^{(H)}$ の臨界点であることをいうためには $\frac{\partial \ell_n^{(H)}(\boldsymbol{\theta}(\lambda))}{\partial \xi_1} = 0$, $\frac{\partial \ell_n^{(H)}(\boldsymbol{\theta}(\lambda))}{\partial \boldsymbol{\eta}} = \boldsymbol{0}$ を示せばよいが、これは次の命題から容易に従う。 □

命題 1. 集合 $\{\boldsymbol{\theta} \in \Theta_H \mid \boldsymbol{\eta} = 0\}$ の任意の点 $\boldsymbol{\theta}$ に対して次式が成立する。

$$\tfrac{\partial}{\partial \boldsymbol{\eta}} \varphi^{(H)}(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{0}, \qquad \tfrac{\partial}{\partial \xi_1} \varphi^{(H)}(\boldsymbol{x}; \boldsymbol{\theta}) = 0. \quad (15)$$

証明. 直接微分することによりすぐに確認できる。 □

定理 2 の証明は損失関数や中間素子関数の具体的な形には全く依存しないことに注意されたい。(2) 式の関数系は有限混合モデルの密度関数の関数系と同様であることから、定理 2 と全く同様の定理が有限混合モデルに対しても成立する。

4.2 3層パーセプトロンのローカルミニマ

本節では、前節の臨界点がローカルミニマとなるための十分条件を求める。この十分条件は $H-1$ 個の中間素子を持つネットワークに対して定義される行列

$$A = \zeta_2^* \textstyle\sum_{i=1}^n \frac{\partial \ell}{\partial s}(Y_i, \varphi^{(H-1)}(X_i; \boldsymbol{\omega}^*)) \frac{\partial^2 h}{\partial \boldsymbol{w} \partial \boldsymbol{w}}(\boldsymbol{x}; \boldsymbol{u}_2^*) \quad (16)$$

だけによって記述される点が興味深い。

定理 3 (Fukumizu and Amari [7]). $\boldsymbol{\omega}_*$ を $\ell_n^{(H-1)}(\boldsymbol{\omega})$ の極小点で、*Hesse* 行列が正定値なものとする。もし *(16)* 式で定義される行列 A が正定値 *[*負定値*]* であるならば、定理 *2* の $\boldsymbol{\theta}(\lambda)$ は、$\lambda(1-\lambda) > 0$ $[< 0]$ において $\ell_n(\boldsymbol{\theta})$ の極小点であり、$\lambda(1-\lambda) \leq 0$ $[\geq 0]$ において鞍点である。行列 A が正負両方の固有値を持つ場合は、任意の λ に対して $\boldsymbol{\theta}(\lambda)$ は鞍点である。

証明. 証明略。 命題 1 を使って $\ell_n^{(H)}(\boldsymbol{\theta})$ の Hesse 行列を計算する。 □

この定理によると、サイズの一つ小さいネットワークの極小点を用意して、その点での行列 A の固有値を調べてそれらがすべて同符号ならば、その極小点を埋め込んだ Θ_H の線分が極小点となる。サイズの増加による関数の自由度の増加を考えると、この点は $\ell_n^{(H)}(\boldsymbol{\theta})$ の最小値ではない局所極小である場合が多いであろう。埋め込み方の数は $\binom{H}{2}$ 通りあるので、このような場合はローカルミニマが多数の線分として存在する。

5 真のパラメータが識別不能な場合の尤度比の漸近論

5.1 局所錐型モデル

本章では、真のパラメータが識別不能な場合の尤度比の漸近的挙動を論じる。モデルとしては主にニューラルネットを念頭におくが、一般論を展開する道具として局所錐型モデルを定義する。これは Dacunha-Castelle & Gassiat ([5]) が導入したものを少し修正したものである。

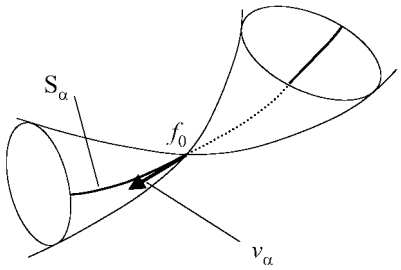

図 3: 局所錐型モデル

A_0 を $(d-1)$ 次元の（境界付き）微分可能多様体、Θ を $A_0 \times \mathbb{R}$ の開集合とする。測度空間 $(\mathcal{Z}, \mathcal{B}, \mu)$ 上の統計モデル $S = \{f(z;\theta) \mid \theta \in \Theta\}$ と $f_0 \in S$ が与えられているとする。パラメータ $\theta \in \Theta$ を $A_0 \times \mathbb{R}$ の分解にあわせて、$\theta = (\alpha, \beta)$ と書く。このとき、統計モデル S が f_0 において**局所錐型**であるとは次の 4 条件が満たされることをいう。

1. $f(z;(\alpha,\beta))$ は、確率 $f_0\mu$ に関してほとんどすべての z に対し、β について微分可能である。
2. $\alpha \in A_0$ に対し $\Theta(\alpha) = \Theta \cap (\{\alpha\} \times \mathbb{R})$ とおくと、$\Theta = \bigcup_{\alpha \in A_0} \Theta(\alpha)$ が成り立つ。
3. 密度関数 f_0 を与えるパラメータ集合は $\Theta_0 = \Theta \cap (A_0 \times \{0\})$ に等しい。すなわち、
$$f(z;(\alpha,\beta))\mu = f_0(z)\mu \iff \beta = 0.$$
4. 任意の $\alpha \in A_0$ に対し $\left\| \frac{\partial \log f(z;\alpha,0)}{\partial \beta} \right\|_{L^2(f_0\mu)} = 1$.

A_0 の次元が 1 以上であれば、f_0 を定めるパラメータは識別不能である。直感的に言うと、確率密度関数全体の空間の中で局所錐型モデルは d 次元の集合をなしているが、f_0 が特異点となっている (図 3)。各 $\alpha \in A_0$ に対して 1 次元部分モデル $S_\alpha = \{f(z;\theta) \mid \theta \in \Theta(\alpha)\}$ は $\beta = 0$ のみで f_0 を与える識別可能なモデルであり、S_α の $\beta = 0$ におけるスコア関数

$$v_\alpha(z) = \frac{\partial}{\partial \beta} \log f(z;(\alpha,0)) \tag{17}$$

は S_α に沿った $L^2(f_0\mu)$-ノルム 1 の接ベクトルだと考えられる。このような接ベクトル全体の集合を $C = \{v_\alpha \mid \alpha \in A_0\}$ で表すと、集合 C は特異点 f_0 における接錐を定める。そこで集合 C を**接錐の底**と呼ぶことにする。接錐の底は以降の議論で重要な役割をする。

確率 $f_0\mu$ に従う i.i.d. サンプル $Z_1, \ldots, Z_n$ に対して、**尤度比** $L_n(\theta)$ は次式で与えられる。

$$L_n(\theta) = \textstyle\sum_{i=1}^n \log \frac{f(Z_i;\theta)}{f_0(Z_i)}. \tag{18}$$

最尤推定量 $\hat{\theta}$ は $L_n(\theta)$ の最大値を取る点である。f_0 を与える真のパラメータが 1 点からなる場合は、適当な正則条件のもと、$L_n(\theta)$ は自由度 d のカイ 2 乗分布に法則収束することはよく知られている。また、もし接錐の底が $L^2(f_0\mu)$ のある有限次元部分空間に含まれていれば、尤度比の極限分布は本質的に有限次元の問題であり、Chernoff ([2]) で述べられているように、モデルが錐を成すという制約下での正規分布の位置母数の推定の話に漸近的には帰着される。その場合、多項分布の混合モデルなど、精密な漸近分布論まで展開できるケースもある ([13])。

本論文では、接錐の底が有限次元の部分空間に含まれない場合を考える。このような場合の現象は複雑である。例えば Hartigan([9]) は、2 コンポーネントからなる正規混合モデルで真の確率が 1 コンポーネントで表せる場合には、尤度比は $n \to \infty$ において発散することを示し

ている。また、Hagiwara et al. ([8]) は真の関数が定数 0 でガウスノイズモデルを仮定した場合に、2 個以上の中間素子を持つ 3 層パーセプトロンの尤度比のオーダーが $O_p(\log n)$ 以上であることを示している。本論文は、後でこれらの結果を一般化する。

5.2 局所錐型モデルの最尤推定

統計モデル $S = \{f(z;(\alpha,\beta))\}$ は $f_0 \in S$ において局所錐型とし、各 $\alpha \in A_0$ に対して 1 次元部分モデル S_α の最尤推定量が存在するとして、それを $\hat{\beta}_\alpha$ と書く。このとき、モデル S の最尤推定量の尤度比は次式で表される。

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha\in A_0} L_n(\alpha, \hat{\beta}_\alpha). \tag{19}$$

各部分モデル S_α が漸近有効性の正則条件を満たすと仮定する。このとき、Taylor 展開による標準的議論により、各 α に対し

$$L_n(\alpha, \hat{\beta}_\alpha) = \tfrac{1}{2} U_n(\alpha)^2 + o_p(1) \tag{20}$$

を得る。ここで $U_n(\alpha)$ は、接錐の底の要素 v_α （(17) 式）を用いて

$$U_n(\alpha) = \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n v_\alpha(Z_i) \tag{21}$$

により与えられる。

仮定から確率変数 $U_n(\alpha)$ は $n \to \infty$ のとき標準正規分布に法則収束するが、すべての α を考えると U_n は接錐の底 C 上の経験過程と見なせる。モデル S の最尤推定の尤度比は

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha\in A_0} \{\tfrac{1}{2} U_n(\alpha)^2 + o_p(1)\} \tag{22}$$

で与えられる。Dacunha-Castelle & Gassiat ([5]) は (22) 式の高次項 $o_p(1)$ が α に関して一様で、かつ U_n があるガウス過程 W に一様ノルムの意味で収束する場合を議論した。その場合、

$$\sup_{\theta\in\Theta} L_n(\theta) = \sup_{\alpha\in A_0} \tfrac{1}{2} W^2 + o_p(1) \tag{23}$$

とすることができ、漸近的には尤度比検定は $|W|$ の sup の分布を計算する問題に帰着される。我々は以下で、(23) 式のような一様性が成り立たない場合を議論する。

5.3 一様な収束をしない場合の尤度比

正規混合モデルの場合に Hartigan ([9]) が行った議論は以下のようなものであった。接錐の底 C 内の有限個の関数 $v_1, \ldots, v_m$ に対して U_n の周辺分布は m 変数正規分布に法則収束し、その共分散は $E_P[v_i v_j]$ で与えられる。そこで、もし任意の $m \in \mathbb{N}$ に対し C 内に m 個の「ほとんど無相関」な変数が存在すれば、それらの上での $U_n(\alpha)$ の最大値は標準正規分布からの m 個の独立な標本の最大値で近似でき、その値はおよそ $\sqrt{2\log m}$ である。m は任意であるから尤度比は有限の値に収まることはない。このアイデアを拡張することにより以下の定理を得る。

定理 4 (Fukumizu [6]). 統計モデル $S = \{f(z;(\alpha,\beta))\}$ は $f_0 \in S$ で局所錐型であるとし、$C = \{v_\alpha(z) = \frac{\partial}{\partial\beta} f(z;(\alpha,0))\}$ をその接錐の底とする。また、任意の $\alpha \in A_0$ に対する部分モ

デル $\{f(z;\alpha,\beta)\mid\beta\}$ は漸近正規性を満たすとする。このとき、もし C 内の系列で、$f_0\mu$ に関して 0 に確率収束するものが存在すれば、任意の $M>0$ に対して次式が成り立つ。

$$\lim_{n\to\infty}\mathrm{Prob}\big(\sup_{(\alpha,\beta)}L_n(\alpha,\beta)\le M\big)=0. \tag{24}$$

略証. 以下の 命題 2 により、任意の $\varepsilon>0$ と $m\in\mathbb{N}$ に対して $v_1,\dots,v_m\in C$ が存在して $|E[v_iv_j]|<\varepsilon$ $(i\neq j)$ が成り立つ。あとの証明は Hartigan の議論と同様である。 □

命題 2. $(\Omega,\mathcal{B},P)$ を確率空間とし、$\{v_n\}_{n=1}^{\infty}$ を $L^2(P)$-ノルムが全て 1 の確率変数列とする。もし v_n が 0 に確率収束するならば、任意の $\varepsilon>0$ に対し、ある部分列 $\{v_{n(k)}\}_{k=1}^{\infty}$ が存在し、異なる k,h に対し $E_P|v_{n(k)}v_{n(h)}|<\varepsilon$ が成立する。

証明. 略。Fukumizu ([6]) 参照。 □

5.4 3層パーセプトロンの尤度比

前節の結果を 3 層パーセプトロンに応用する。本節では x が 1 次元で、中間素子の関数が $h(x;w)=\tanh(ax+c)$ の場合を考える。

$0\le K<H$ として、H 個の中間素子を持つ 3 層パーセプトロンモデルが、K 個の中間素子で実現可能な関数において局所錐型であることを示そう。パラメータ空間 Θ_H を少し制限し、$\Theta_H^*=\{\theta=(a_1,\dots,a_H,b_1,\dots,b_H,c_1,\dots,c_H,d)\in\Theta_H\mid a_j\neq0,\ b_j\neq0\ (1\le j\le H),\ (a_j,c_j)\neq\pm(a_h,c_h)\ (1\le j<h\le H)\}$ と定義する。定理 1 より、Θ_H^* は、中間素子 H 個のパーセプトロンで書けるが H より小さい中間素子数では実現不可能な関数全体を定める。

$\varphi_0(x)$ を K 個の中間素子で実現可能な関数

$$\varphi_0(x)=\textstyle\sum_{k=1}^{K}b_k^0\tanh(a_k^0x+c_k^0)+d^0 \tag{25}$$

とする。ここで $(a_1^0,\dots,a_K^0,b_1^0,\dots,b_K^0,c_1^0,\dots,c_K^0,d^0)\in\Theta_K^*$ とする。与えられた $\varphi_0(x)$ に対して、さらにパラメータ空間を少し制限し、$\Theta_H^{**}=\{\theta\in\Theta_H^*\mid(a_j,c_j)\neq\pm(a_k^0,c_k^0)\ (1\le k\le K,K+1\le j\le H)\}$ を考える。このような制限を行っても、最尤推定量は確率 1 で Θ_H^{**} に入るので、最尤推定を考える際には問題がない。さらに $\theta\in\Theta_H^{**}$ に対し、以下のような新しいパラメトリゼーションを導入する。ただし $1\le k\le K$, $K+1\le j\le H$ である。

$$\begin{aligned}\xi_k&=\tfrac{1}{\beta}(a_k-a_k^0),\quad \eta_k=\tfrac{1}{\beta}(b_k-b_k^0),\quad \zeta_k=\tfrac{1}{\beta}(c_k-c_k^0),\quad \delta=\tfrac{1}{\beta}(d-d^0)\\ \xi_j&=a_j,\quad \eta_j=\tfrac{b_j}{\beta},\quad \zeta_j=c_j,\quad \beta=\mathrm{sgn}(b_{K+1})\sqrt{b_{K+1}^2+\cdots+b_H^2}.\end{aligned} \tag{26}$$

新しいパラメータ空間は $\Pi_H=\{\omega=(\xi_1,\dots,\xi_H,\eta_1,\dots,\eta_H,\zeta_1,\dots,\zeta_H,\delta,\beta)\mid a_k^0+\beta\xi_k\neq0\ (1\le k\le K),\xi_j\neq0\ (K+1\le j\le H),\ (a_k^0+\beta\xi_k,c_k^0+\beta\zeta_k)\neq\pm(a_h^0+\beta\xi_h,c_h^0+\beta\zeta_h)\ (1\le k<h\le K),\ (a_k^0+\beta\xi_k,c_k^0+\beta\zeta_k)\neq\pm(\xi_j,\zeta_j)\ (1\le k\le K,K+1\le j\le H),\ (\xi_j,\zeta_j)\neq\pm(\xi_i,\zeta_i)\ (K+1\le j<i\le H),\ (\xi_j,\zeta_j)\neq\pm(a_k^0,c_k^0)\ (1\le k\le K,K+1\le j\le H),b_k^0+\beta\eta_k\neq0\ (1\le k\le K),\ \sum_{j=K+1}^{H}\eta_j^2=1,\ \eta_j\neq0\ (K+1\le j\le H),\ \eta_{K+1}>0,\ \beta\in\mathbb{R}\}$ で与えられる。また $\Pi_H^{**}=\{\omega\in\Pi_H\mid\beta\neq0\}$ とおく。すると、3 層パーセプトロンは

$$\psi(x;\omega)=\sum_{k=1}^{K}(b_k^0+\beta\eta_k)\tanh\big((a_k^0+\beta\xi_k)x+(c_k^0+\beta\zeta_k)\big)+\sum_{j=K+1}^{H}\beta\eta_j\tanh(\xi_jx+\zeta_j)+\beta\delta \tag{27}$$

と表現することが出来る。Π_H^{**} と Θ_H^{**} は (26) 式の変換で 1 対 1 に移りあい、この対応に対して $\varphi(x;\theta) = \psi(x;\omega)$ が成り立つことは容易に確かめられる。Π_H で定まる関数族は、Θ_H^{**} で定義される丁度 H 個の中間素子数で実現される関数と、$\beta = 0$ に対応する φ_0 とから成る。

統計モデル $S_H = \{f(x,y;\omega) \mid \omega \in \Pi_H\}$ を

$$f(x,y;\omega) = r(y|\psi(x;\omega))q(x) \tag{28}$$

によって定義し、$\varphi_0(x)$ に対応する密度関数を $f_0(x,y)$ とする。ω の要素の内 $(\xi_1, \ldots, \zeta_H, \delta)$ を α で表すとき、以下の定理が成立する。

定理 5 (Fukumizu [6]). *S_H を (27),(28) 式で定義される、中間素子を H 個持つ 3 層パーセプトロンモデルとする。ノイズモデル $r(y|s)$ に関する適当な正則条件のもとで、S_H は f_0 において局所錐型である。*

略証. 局所錐型の定義の $1-3$ を満たすことは Π_H の条件から示される。$\frac{\partial}{\partial\beta}\log f(x,y;(\alpha,0))$ の L^2 ノルムを $N(\alpha)$ とおくと、適当な条件のもと $0 < N(\alpha) < \infty$ になることが示せるので、β の代わりに $\beta N(\alpha)$ を用いれば、定義の 4 を満足する。 □

この局所錐型モデルは 定理 4 の仮定を満足し、次の定理が得られる。

定理 6 (Fukumizu [6]). 中間素子を H 個持つ 3 層パーセプトロンモデルに対し、学習データを発生させる真の関数が中間素子 K 個（$K < H$）で実現できたとする。このときノイズモデル $r(y|s)$ に対する適当な正則条件のもと、任意の $M > 0$ に対し次式が成立する。

$$\lim_{n\to\infty} \mathrm{Prob}\big(\sup_\theta L_n(\theta) \le M\big) = 0. \tag{29}$$

Remark. この定理より、真の関数を表現するのに過剰な中間素子を持つネットワークを用いると、3 層パーセプトロンの尤度比は $O_p(1)$ より真に大きいオーダーを持つことがわかる。

略証. 部分モデル $g(z;t,c,\beta) = r(y|\varphi_0(x) + \beta w(x;t,c))q(x)$ を考えれば十分である。ここで $w(x;t,c) = \frac{1}{\sqrt{B(t,c)}}\sigma(x;c^2,t+\frac{1}{c})$ ただし、$\sigma(x;\xi,h) = \frac{1}{2}\{1+\tanh(-\frac{1}{2}\xi(x-h))\} = \frac{1}{1+\exp\{\xi(x-h)\}}$, $B(t,c)$ は接ベクトルの L^2 ノルムの正規化定数である。この部分モデルに対し (c,t) を固定した 1 次元モデルは適当な条件のもと漸近正規性を満たす。また、接錐の底は $v(x,y;t,c) = \frac{1}{\sqrt{B(t,c)}}\frac{\partial \log r(y|\varphi_0(x))}{\partial s}\sigma(x;c^2,t+\frac{1}{c})$ という形の関数からなるが、$t_n \to \infty$, $c_n \to \infty$ なる列 (t_n,c_n) をうまく取ると $v(x,y;t_n,c_n)$ が 0 に概収束する。よって定理 4 が使える。 □

もし $K \le H-2$ ならば、上の証明に用いたものと違うタイプの関数列で 0 に確率収束するものを構成できる。ます関数族 $\mathcal{W} = \{w(x;\xi,h,t)\}$ を $w(x;\xi,h,t) = \frac{1}{\sqrt{A(\xi,h,t)}}\frac{1}{2}\{\tanh(\xi(x-t+h)) - \tanh(\xi(x-t-h))\}$, によって定める。ここで $A(\xi,h,t)$ は、以下の $v(z;\xi,h,t)$ の L^2 ノルムの正規化定数である。3 層パーセプトロンの部分モデルを $\psi(x;\xi,h,t,\beta) = \varphi_0(x) + \beta w(x;\xi,h,t)$ によって定めると、接錐の底は

$$v(z;\xi,h,t) = \frac{\partial \log r(y|\varphi_0(x))}{\partial s} w(x;\xi,h,t) \tag{30}$$

という形の関数よりなる。この関数に対して $\xi_n \to \infty$, $h_n \to 0$ なる点列 (ξ_n,h_n,t_n) をうまく取ると、$v(z;\xi_n,h_n,t_n)$ は 0 に概収束することが示される。さらに、この関数族を使うと、$K \le H-2$ の場合に尤度比のオーダーの下界は次のように改良される。

定理 7 (Fukumizu [6]). H 個の中間素子を持つ 3 層パーセプトロンモデルに対して、データを発生する真の関数が K 個の中間素子で実現可能だと仮定する。もし $K \leq H-2$ ならば、ノイズモデル $r(y|s)$ に関する適当な正則条件のもと、ある $\delta > 0$ が存在して次が成立する。

$$\liminf_{n\to\infty} \mathrm{Prob}\Big(\frac{\sup_\theta L_n(\theta)}{\log n} \geq \delta\Big) \quad > \quad 0. \tag{31}$$

証明の概略. サンプル数が n のときに、接錐の底 C 内に、n^γ（$\gamma > 0$）個のほとんど無相関な関数が存在することを示す。そのために、関数族 $\mathcal{W}$ によって定義される部分モデルで考える。

任意の閉区間 I に対し非負実数 $M(I)$ を $M(I) = E_{f_0\mu}\Big[\Big(\frac{\partial \log r(y|\varphi_0(x))}{\partial s}\Big)^2 \chi_I(x)\Big]$ により定義する。実数直線上に互いに交わらない区間を $m = n^\gamma$ 個取り、$r(y|\varphi_0(x) + \beta\frac{1}{\sqrt{M(I_k)}}\chi_{I_k}(x))q(x)$（$1 \leq k \leq m$）により定義される 1 次元モデルたちを考えると、それらの接ベクトル $u_k(z) = \frac{1}{\sqrt{M(I_k)}}\frac{\partial \log r(y|\varphi_0(x))}{\partial u}\chi_{I_k}(x)$ は互いに無相関である。そこで m 次元確率ベクトル $V_n = \big(\frac{1}{\sqrt{n}}\sum_{i=1}^n u_1^{[m]}(Z_i), \ldots, \frac{1}{\sqrt{n}}\sum_{i=1}^n u_m^{[m]}(Z_i)\big)$ を考えると、ノイズモデル $r(y|s)$ に関する適当な正則条件のもとで、γ を十分小さく取ると、V_n の分布と m 次元標準正規分布とは一様に近いことを示すことが出来る。よって、$|V_n|$ の最大値は $\sqrt{2\log m} = \sqrt{2\gamma \log n}$ に近い。ところが、$\frac{1}{\sqrt{M(I)}}\chi_I(x)$ は $\mathcal{W}$ によっていくらでもよく近似できるので、$\mathcal{W}$ 内の n^γ 個の関数があって、その上で (22) 式の値が $\log n$ のオーダーになる。詳細は Fukumizu ([6]) を参照されたい。 □

上の定理は、3 層パーセプトロンにおいてモデルに冗長な中間素子が 2 個以上あれば、尤度比のオーダーが $O_p(\log n)$ 以上であることを意味している。このオーダーの下界は、真の関数が 0 定数関数でノイズモデルがガウス分布の場合には Hagiwara et al. ([8]) が既に求めている。上の結果は、ノイズモデルと真の関数に関して一般化したものとなっている。

今までの議論からも明らかなように、局所錐型モデルの尤度比の漸近分布は接錐の底の性質に深く依存する。実際、オーダーが $O_p(\log n)$ よりも小さい例も知られている。例えば、Hartigan ([9]) は、正規混合モデルにおいて、モデルが 2 コンポーネントで真の分布が 1 コンポーネントからなる場合の尤度比は $O_p(\log\log n)$ だと予想している。また、ステップ関数を中間素子の関数に持つ 1 個の中間素子からなる 3 層ネットワークに対して、真の関数が定数 0 でガウスノイズの場合には、尤度比のオーダーは $O_p(\log\log n)$ となる ([10])。この例は変化点問題とほぼ同等である ([3])。このようなオーダーの違いを規定しているものが何なのかはよくわかっていない。また、本論文では尤度比のオーダーの下界のみを議論したが、正確なオーダーがどのようなものであり、また漸近分布がどのようになるのかといった問題は、今後の課題である。

6 おわりに

本論文では、3 層ニューラルネットモデルの数理的、統計的な性質に関して、特にモデルが構造的に持つ対称性に焦点をあてて議論した。前半では、サイズの 1 つ小さいモデルでの経験損失関数の臨界点が大きいサイズでの臨界点として埋め込まれ得ること、および、小さいサイズでの極小点を埋め込んだものが大きいサイズでの極小点となるための十分条件を、小さいサイズに関する量のみで表した。

後半では、真のパラメータが識別不能な場合の最尤推定を議論する枠組みとして局所錐型モデルを紹介し、尤度比が通常の $O_p(1)$ のオーダーよりも大きくなるための簡単な十分条件を与えた。また、この結果を 3 層パーセプトロンに応用して、真の関数を表現するのに冗長な中間

素子が存在する場合には、尤度比は $O_p(1)$ よりも真に大きくなり、さらに冗長な中間素子が 2 個以上存在すれば、尤度比が $O_p(\log n)$ 以上のオーダーを持つことを示した。

本論文では主として 3 層パーセプトロンをモデルとして説明したが、モデルの持つ対称性は有限混合モデルのそれとほとんど同じである。また局所錐型モデルの枠組みは識別不能性のある統計モデルの多くをカバーしている。これらモデルの対称性、識別不能性に関する問題には未解決のものが多く、尤度比の分布論なども含めて今後さらなる発展が期待される。

参考文献

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

[2] H. Chernoff. On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25:573–578, 1954.

[3] M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. John Wiley and Sons, 1996.

[4] G. Cybenco. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[5] D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models and application to mixture models. *ESAIM Probability and Statistics*, 1:285–317, 1997.

[6] K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. Research Memorandum 780, The Institute of Statistical Mathematics, 2001.

[7] K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.

[8] K. Hagiwara, K. Kuno, and S. Usui. On the problem in model selection of neural network regression in overrealizable scenario. In *Proc. of Intern. Joint Conf. on Neural Networks*, 2000.

[9] J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 807–810, 1985.

[10] T. Hayasaka, N. Toda, S. Usui, and K. Hagiwara. On the least square error and prediction square error of function representation with discrete variable basis. In *Proc. of Neural Networks for Signal Processing*, pages 72–81, 1996.

[11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufman, 1990.

[12] Y. LeCun, L. Bottou, G. B. Orr., and K.-R. Müller. Efficient backprop. In G. B. Orr and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, pages 9–50. Springer, Berlin, 1998.

[13] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics, California, 1995.

[14] R. D. Reed and R. J. Marks II. *Neural Smithing*. MIT Press, 1999.

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

[16] S. Veres. Asymptotic distributions of likelihood ratios for overparameterized arma processes. *Journal of Time Series Analysis*, 8(3):345–357, 1987.

連絡先： 福水 健次. 〒 106-8569 東京都港区南麻布 4-6-7. 統計数理研究所

Tel: 03-5421-8730. E-mail: fukumizu@ism.ac.jp