誤差確率分布を考慮した誤差逆伝播学習

鈴 木 昇 一

Maximum-Likelihood Error Back-Propagation Learning Algorithm

Shoichi Suzuki

Abstract

This paper proposes a new theoretical framework of error backpropagation learning by applying a method of maximum likelihood to a probability distribution of errors between actual and desired outputs in a multilayered feedforward neural network. The synaptic weights that connect neurons in one layer with neurons in the other layer are obtained as maximum-likelihood-type estimators through a kind of supervised learning. It will be shown that the learning with the network presented here means finding out a set of synaptic weights that minimize a weighted adaptive error.

The distribution of errors is assumed to be the normal, the exponential, the student's distribution, or a combinatory distribution so that it is Gaussian in the middle and Laplacian at the tails with much large variance. The corresponding analyses are carried out. These results of this paper that have been successfully generalized for a highly robust learnig on condition that the error distribution is suitably selected out for the learning environment provide a direct generalization of the Perceptron Learning procedure, the ordinary learning algorithm proposed by Rumelhart et al., etc..

要　　約

本論文では，理想出力からのズレ（誤差）の確率分布に対し最尤法を適用し，階層形ニューラルネット誤差逆伝播学習の新しい理論的枠組が提案される。一つの層のニューロンと次の層のニューロンとを結ぶシナプス結合の重みは教師あり学習を介し，最尤形推定量として得られる。本研究で提案されるネットワークの学習とは，ある重みつき適応誤差を極小ならしめるシナプス結合重みを発見することを意味する事実が示される。

誤差分布が各々，正規分布，指数分布，t分布，ガウス・ラプラシアン組合せ分布である場合が解析され，誤差分布が学習環境に対し適切に選ばれると頑健な学習をもたらすこれらの結果はパーセプトロン学習，ルーメルハート等の通常の学習アルゴリズム，その他諸々の直接的な一般化を提供する。

1．まえがき

現実出力と理想出力との間の差（誤差）を極小ならしめるニューラルネット学習方法[12),(13)]として，誤差逆伝播学習アルゴリズム（error back-propagation learning algorithm ; BPLA）[(14)]がよく知られているが，理想出力からのズレに対し頑健な（robust）な学習法として，

最尤法（method of maximum-likelihood）

を誤差確率分布に対し適用して得られた

最尤型誤差逆伝播学習アルゴリズム MLBPLA (maximum-likelihood error back-propagation learning algorithm)

が本研究で提案される。従来の BPLA は誤差確率分布を仮定しない（distribution-free）推定法であるが，

MLBPLA において，誤差分布として等分散の正規分布

を採用したものに一致することが示される。

BPLA は前進形の多層ネット（multi-layer feedforward network）において最小自乗法（method of least squares）を適用した学習法であり，得られるニューロン間シナプス結合（synaptic connection）の重み（weight）の組は最小自乗推定量（least squares estimator）である。単層あるいは2層ネットとしてのパーセプトロン[(24)]（Perceptron）での

学習法（Perceptron learning algorithm ； PLA）

の一般化である。

それで，先ず，Perceptron[12),(13)]について復習しよう。

入力ベクトル x を

$$x = \{x_\ell | \ell = 0, 2, \cdots, n\}$$

ここに，$x_0 \equiv 1, x_\ell \in \{0, 1\} (1 \leqq \ell \leqq n)$ とし，

また，重み W を

$$W = \{W_\ell | \ell = 0 \sim n\}$$

ここに，W_ℓ は任意の実数

として，

$$z = g(y) \in \{0, \frac{1}{2}, 1\}$$

を出力するシステムが単純 Perceptron (simple Perceptron) である。

ここに，

$$y = W \cdot x = \Sigma_{\ell=1}^{n} W_\ell \cdot x_\ell = W_0 + \Sigma_{\ell=1}^{n} W_\ell \cdot x_\ell$$

$g(u) = 0$ *if* $u < 0, = 2^{-1}$ if $u = 0$,
$= 1$ *if* $u > 0$.

単純 Perceptron の訓練・学習定理[12]

n 個の，順序づけられた自由変項（free variable）$x_1, x_2, \cdots, x_n$ を含む二つの式

$E^-(x') \equiv E^-(x_1, x_2, \cdots, x_n)$
$E^+(x') \equiv E^+(x_1, x_2, \cdots, x_n)$

を各々，

$W_0 + \sum_{\ell=1}^{n} W_\ell \cdot x_\ell < 0$， $W_0 + \sum_{\ell=1}^{n} W_\ell \cdot x_\ell > 0$.

と定義する。ここに，

$x' = (x_1, x_2, \cdots, x_n) = x - \{x_0\}$.

さて，$E^-(x'), E^+(x')$ という性質をもつすべての n 項 x' の集合（抽象集合）を各々，

$\lambda x'.E^-(x')$， $\lambda x'.E^+(x')$

で表し，これを各々

$E^-(x')$，$E^+(x')$ の抽象（abstract）

といい，

$\lambda x'(= \lambda x_1, x_2', \cdots x_n)$

を抽象化オペレータ（abstraction operator）と称する。ある $a = (a_1, a_2, \cdots, a_n)$ がこの抽象集合に属することを

$a \in \lambda x', E^{\pm}(x')$

という記号で表す。これを

$[\lambda x'.E^{\pm}(x')]a$ あるいは $E^{\pm}(a)$ と表すこともある。

a を digitized pattern（計数型パターン）といい，

$\lambda x'.E^-(x')$， $\lambda x'.E^+(x')$

を各々，

カテゴリ $\mathfrak{C}^-$，カテゴリ $\mathfrak{C}^+$

という。

個々のパターン a は具体例（instance）であるが，その集まり $\mathfrak{C}^-$，$\mathfrak{C}^+$ は各々，ある性質を共通にもっていることに注目し分類して得られただけで，抽象的な存在である。これは，Taro, Hanako は抽象的な存在ではないが，その集まりとしての男性，女性という集合概念は抽象的存在であることを想い起こせば，理解できよう。

ある重みの組 W が存在し，

$\mathfrak{C}^- = \lambda x'.E^-(x') \subset I^n$ (n 次元超単位立方体)
$\mathfrak{C}^+ = \lambda x'.E^+(x') \subset I^n$

に対し,
$\mathfrak{C}^- \cap \mathfrak{C}^+ = \phi$ (空集合)

が成立するものとする。これが線形分離 (linearly separable) という仮定である。
さて,

$X^- \subset \mathfrak{C}^-, X^+ \subset \mathfrak{C}^+$

なる二つの有限部分集合を $X^{\pm}$ を考える。$X^- \cup X^+$ に属するすべてのパターンを適当な順序に並べ, それを無限に繰り返して出来るパターンの列を

$x'(1), x'(2), \cdots, x'(k) = (x_1(k), x_2(k), \cdots x_n(k)), \cdots \in I^n$

とする。

$x(k) = (x_0(k), x_1(k), \cdots, x_n(k))$,

ここに, $x_0(k) \equiv 1$
なる $x(k)$ も (一般化 digitized) pattern ということにする。
(学習の) 第 k ステップで, パターン $x(k) \in I^{n+1}$ を提示するものとし, 第 k ステップでの重み $W(k)$ を,
初期値 $W(k)|_{k=0}$
が与えられたとき,

$$W(k+1) = \begin{cases} W(k) & if \quad x'(k) \in X^- \wedge W(k) \cdot x(k) < 0 \\ W(k) & if \quad x'(k) \in X^+ \wedge W(k) \cdot x(k) > 0 \\ W(k) - x(k) & if \quad x'(k) \in X^- \wedge W(k) \cdot x(k) \geqq 0 \\ W(k) + x(k) & if \quad x'(k) \in X^+ \wedge W(k) \cdot x(k) \leqq 0 \end{cases} \qquad (1.\ 1)$$

と変更していけば,
ある有限の k が存在して,

$W(k) = W(k+1) = W(k+2) = \cdots$

が成り立ち, この $W(k)$ は連立不等式

$[\forall\, x \in X^- \subset \mathfrak{C}^-, W(k) \cdot x < 0] \wedge$
$[\forall\, x \in X^+ \subset \mathfrak{C}^+, W(k) \cdot x > 0]$

の解になっている[12]。 □

上記の定理は, 弛緩法 (relaxation procedure) の適用下で, $\mathfrak{C}^-$, $\mathfrak{C}^+$ を線形分離する重み W を修正していく学習 (learning) 法を提示しており, この意味で, パーセプトロンは $\mathfrak{C}^-$, $\mathfrak{C}^+$ を二つのカテゴリとする
two-category classifier[5]

である。式（1．1）がパーセプトロンでの学習規則（learning rule）といわれるものである。

ここで，パーセプトロンの識別器としての役割に関する上記の学習法とは別に，

重要な幾何学的な性質は必ず，ある変換群の下での不変量（invariants）になっている

という“Felix Klein の数学的立場”に立ち，

図形からその特徴を抽出する“特徴抽出器”（feature-extractor）としての役割に目を向けてみよう。

関数 $\varphi_\ell : I^n \to I^1$ を想定し，x_ℓ の代りに $\varphi(x') = \varphi_\ell(x_1, x_2, \cdots, x_n)$ を考えると，$z = g(y)$ は

$$z = g(\sum_{\ell=1}^{n} W_\ell \cdot \varphi_\ell(x') + W_0)$$

の形になる。今，3値関数 g の代りに，一変数 u の2値関数

$$psn(u) = 1 \quad if \quad u \geqq 0, = 0 \quad if \quad u < 0$$

を考え，

$$\theta = -W_0$$

とおくと，パーセプトロン出力 z は

$$z = psn(\sum_{\ell=1}^{n} W_\ell \cdot \varphi_\ell(x') > \theta)$$

となり，これは

$$z = \lceil \sum_{\ell=1}^{n} W_\ell \cdot \varphi_\ell(x') > \theta \rceil \qquad (1．2)$$

とも書ける。ここに，

$\lceil A \rceil$ は命題 A が真ならば1，偽ならば0をとる

と規約されたものである。

式（1．2）が Minsky-Papert のパーセプトロン理論[12]でのパーセプトロンの表現であり，それは精確には，

φ_ℓ の集合を Φ として，重み W_ℓ を $W(\varphi)$ と書いて，

$$\psi(x') = \lceil \sum_{\varphi \in \Phi} W(\varphi) \cdot \varphi(x') > \theta \rceil \qquad (1，3)$$

と書かれる。

Minsky-Papert の理論での基本定理は**群不変定理**（group invariance theorem）といわれるものであり，

x' のある集合 X' を図形とみなす

と，図形の群不変な特徴に対し，反応する重み係数 $W(\varphi)$ の一意的な存在を指摘したものである。

群不変定理

次の3条件(1)，(2)，(3)を考える。

(1)G は $x' = (x_1, x_2, \cdots x_n) \in Y'$ の上で定義された有限な変換群である，つまり，

$x' \in Y'$ ならば $gx' \in Y'$

が成り立つ。ちなみに，Gが群であるとは，

$g_1, g_2 \in G$に対し$g_1 \cdot g_2 \in G$

なる演算・が定義され，次の3条件1．i）～1．iii）が成り立つことである：

1．i）（左単位元eの存在）

$\exists\, e \in G, \forall\, g \in G, e \cdot g = g.$

1．ii）（結合律）$\forall\, g_1, \forall\, g_2, \forall\, g_3 \in G, g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3 \in G.$

1．iii）（左逆元g^{-1}の存在）$\forall\, g \in G$に対し，gの左逆元といわれる$g^{-1} \in G$が存在し，$g^{-1} \cdot g = e \in G$。

備考2．1：群Gの左単位元はまた右単位元であり，[$\exists\, e \in G$，$\forall\, g \in G$，$g \cdot e = g$] が成り立つ。また，元$g \in G$の左逆元は同時に右逆元であり，[$\forall\, g \in G$，$g \cdot g^{-1} = e \in G$] が成り立つ。さらに，元$g^{-1} \in G$の逆元は$g \in G$に一致し，$(g^{-1})^{-1} = g$が成立する。 □

(2) φの集合Φは変換群Gの下で閉じている：$\forall\, \varphi \in \Phi$，$\forall\, g \in G$，$\varphi g \in \Phi$，

ここに，$\forall\, x' \in Y'$，$\varphi g(x') \equiv \varphi(g(x'))$

(3) $\sum_{\varphi \in \Phi} v(\varphi) \cdot \varphi$は$G$の下で不変である：$\sum_{\varphi \in \Phi} v(\varphi) \cdot \varphi(x')$

$= \sum_{\varphi \in \Phi} v(\varphi g) \cdot \varphi g(x')$ for any $x' \in Y'$.

上の3条件(1)～(3)の下では

$\phi(x') = [\sum_{\varphi \in \Phi} W(\varphi) \cdot \varphi(x') > \theta]$

内の重み係数$W(\varphi)$に関し，群不変性

$\forall\, x \in X, \exists\, g \in G, \varphi g(x) = \varphi'(x)$

ならば，$W(\varphi) = W(\varphi')$

が成立し，実際，$W(\varphi)$は次のように与えられる：$W(\varphi) = \sum_{g \in G} v(\varphi g)$。 □

2．これ迄の，S. Suzukiのニューラルネット研究

S. Suzukiは従来の，有限次元空間で展開されているニューラルネット情報処理[2),(12),(13),(14),(16),(17),(18)]の典型的3手法に対応した手法を，収縮写像（モデル構成作用素）を適用し，無限次元空間としての可分な一般抽象Hilbert空間$\mathfrak{H}$上で構築しようとしている。[(11)の17部～23部;(9),(10)]これらは次の様に位置づけられるであろう。

(i) Rosenblatt (1961)[(18)]の提案した，2カテゴリ分類器としての単純パーセプトロン（第1章を参照）に対応する空間パーセプトロン[(4),(5)～(8)]

(ii) Hopfield (1982)[(16)]の提案した，エネルギーの極小値をもたらすパターンへの変換を計算アルゴリズム（computational algorithm）として持つHopfield net[(25)]に対応するモデル[(11)の17部～20部]

なお，Hopfield netを確率的動作の下で稼働させる形式としてのBoltzmann machine[(17)]についても発表を予定していることを付言しておく。

(iii) Rumelhart et al. (1986)の提案した，誤差逆伝播学習アルゴリズム[(14)]下での階層形ネット

(multilayered net) に対応するモデル[11]の(21)部～(23)部

なお，従来の階層形ネットに Rosenfeld 型の確率的弛緩変換[19]を導入した研究[3]も発表している。

これらの諸研究について少し紹介しておこう。

S. Suzuki は，内積，ノルムを各々，$(,)$，$\|\cdot\|=\sqrt{(\cdot,\cdot)}$ とする可分な一般抽象 Hilbert 空間 $\mathfrak{H}$ の部分集合 Φ を，処理の対象とするパターン φ 集合と考え，

$G, H : \mathfrak{H}$ での二つの自己共役作用素

$\theta_\ell(H) : H$ の関数としての，第 $\ell \in L$ 番目の射影作用素

として，パーセプトロン形作用素（Perceptron-like operator）

$$Q \equiv \Sigma_{\ell L} W_\ell \cdot \theta_\ell(H) \qquad (2.\ 1)$$

を提案した。ここに，実係数 W_ℓ は第 $\ell \in L$ 番目の重みであり，射影作用素の系

$$\{\theta_\ell(H) \mid \ell \in L\}$$

は3条件

$$\forall\ \ell \in L,\ \theta_\ell(H) \neq 0,\ \neq I \text{（恒等作用素）}$$
$$\theta_k(H) \cdot \theta_\ell(H) = 0\ (k \neq \ell) \text{（直交性）}$$
$$\Sigma_{\ell \in L}\, \theta_\ell(H) = I \qquad (2.\ 2)$$

を満たしていることが望ましい。式（2，1）の Q は自己共役作用素であり，パーセプトロン形空間回路（Perceptron-like spatial circuit）と称されることがある。

パターン $\varphi \in \Phi$ は作用素 Q により

$$\varphi \rightarrow \phi \equiv Q\varphi = \Sigma_{\ell \in L} W_\ell \cdot \theta_\ell(H)\varphi$$

と変換され，それに伴って，パターン φ の特徴量（測度的不変量[1],[21]；metrical invariant's）

$$(G\varphi, \varphi)$$

は

$$(G\varphi, \varphi) \rightarrow (G\phi, \phi) = \Sigma_{k \in L} W_k \cdot (G \cdot \theta_k(H)\varphi, \phi)$$
$$= \Sigma_{k \in L} \Sigma_{m \in L} a_{km} \cdot W_k \cdot W_m$$

，ここに，

$$a_{km} \equiv (G \cdot \theta_k(H)\varphi, \theta_m(H)\varphi)$$

と変換されることに注意しておく。

SS 理論[11]での収縮写像

$$T : \Phi \rightarrow \Phi$$

を導入し，再び，パターン変換

$$\varphi \rightarrow T\varphi \rightarrow QT\varphi = \Sigma_{\ell \in L} W_\ell \cdot \theta_\ell(H) T\varphi$$

を考えよう。ここに，$T\varphi \in \Phi$ はパターン $\varphi \in \Phi$ から特徴抽出した結果を反映するように構成可能[1),(4),(11)]であり，パターン φ の構造モデル（structural model）と称されてよい。

自己共役作用素 Q とモデル $T\varphi$ とのなす測度的不変量

$$(QT\varphi, T\varphi)/(T\varphi, T\varphi) \qquad (2.\ 3)$$

を計算するとしよう。

$$v_\ell(\varphi) \equiv (\theta_\ell(H)\ T\varphi,\ T\varphi)/(T\varphi, T\varphi) \qquad (2.\ 4)$$

として

$$0 \leqq v_\ell(\varphi) \leqq \wedge \Sigma_{\ell \in L} v_\ell(\varphi) = 1 \qquad (2.\ 5)$$

が成立しているが，このとき，表現

$$(QT\varphi, T\varphi)/(T\varphi, T\varphi) = \Sigma_{\ell \in L} W_\ell \cdot v_\ell(\varphi) \qquad (2.\ 6)$$

が成立するから，式（1．2）に対応する two-category classifier の表現として，

$$\lceil (QT\varphi, T\varphi)/(T\varphi, T\varphi) > h \rceil \qquad (2.\ 7)$$

が採用できる。これが SS 理論[11]でいう空間パーセプトロンの表現であり[5]，特徴抽出・識別なる二つの働きを同時に兼ね備えていることが従来のパーセプトロンに比し異なる点である。

各入力に重みを乗じ，その総和をシグモイド関数（sigmoid function）などの神経ユニット発火関数で変換して各出力を導く計算システムとしてのニューラルネット（neural network）による情報処理は，記号列処理と異なり，ヒトの脳内情報処理を手本としていることである。n 個の神経ユニットが相互に結合されている形式としてのニューラルネットの動作とは，第 j 神経ユニットへの重みつき入力の総和

$$u_j \equiv \Sigma_{i=1}^{n} W_{ij} \cdot x_i \qquad (2.\ 8)$$

を計算して，例えば

$$g(u) = \frac{d - c}{1 + \exp[-(u - h)/a]} + c \qquad (2.\ 9)$$

ここに，$c < d \wedge 0 < a$，h は実数しきい値

なるシグモイド関数などの神経ユニット発火関数 $g : R \rightarrow R$（R は実数全体）で

$$x_j = g(u_j)$$

と変換して得る第 j 神経ユニット出力 x_j を，$j = 1 \sim n$ につき時々刻々求めることである。

その学習過程とは，希望する入出力関係を満たすように，入力層内神経ユニットにあるデータ

を与え，出力層内神経ユニットから得られる稼働出力と与えられた希望出力との差を限りなく零にするように，第 i 神経ユニットから第 j 神経ユニットへの結合の重み W_{ij} を

$$W_{ij}(t) \to W_{ij}(t+\Delta t) = W_{ij}(t) + \Delta W_{ij}(t) \qquad (2.\ 10)$$

という様式で少量ずつ変更してゆき，最終的には，今迄入力されなかった未知の入力（パターン）に対しても所要の正しい出力を得るようにすること（汎化能力）である。

入力と出力とを結ぶ関係情報を重み W_{ij} の組に圧縮して記憶しているという“情報圧縮性”は在来のニューラルネットの一大特色である。S. Suzuki のニューラルネット理論は，文献(11)での axiom 1 からわかるように，ニューラルネット構造を決める写像（収縮写像，モデル構成作用素）

$$T\cdot = \Sigma_{\ell\in L}\, u(\cdot,\ell)\cdot\theta_\ell(H)\,\xi\|\xi\|^{-1}$$

ここに，$u(\varphi,\ell)\in R$ はパターン $\varphi\in\Phi$ から抽出される第 $\ell\in L$ 番目の特徴量　（2．11）

がベキ等性（idempotent property）

$$TT = T$$

を満たしていることにより，重み W_{ij} の組による情報圧縮性のみならず，ニューラルネット構造自体が入力情報（パターン）を変換しながら，情報（パターンのもつ特徴量）を圧縮し再表現していることが基調となっていることである[27]。

これは，モデル構成作用素 T に依存した，S. Suzuki の提唱する次のニューラルネット構造形式（2．16），（2，18）からもたらされることから得心がいくかも知れない。

まずパターン $\varphi\in\Phi$ のモデル $T\varphi\in\Phi$ は

$$T\varphi = [\Sigma_{\ell\in L}\, u(\varphi,\ell)\cdot\theta_\ell(H)]\,\xi\|\xi\|^{-1} \qquad (2.\ 12)$$

と書けるから，これは，パターン $\xi\|\xi\|^{-1}\in\Phi$ にパーセプトロン形作用素

$$Q(\varphi) \equiv \Sigma_{\ell\in L}\, u(\varphi,\ell)\cdot\theta_\ell(H) \qquad (2.\ 13)$$

を作用させたものであり，

$$T\varphi = Q(\varphi)\,\xi\|\xi\|^{-1}$$

が成立することに注意しておく[21]。ここに，パターン $\xi\in\Phi$ については，

モデル $T\varphi$ に処理対象としてのパターン集合 Φ の形状を反映させるためには，例えば，

$$\xi = \Sigma_{\ell\in J}\, P(\mathfrak{C}_j)\cdot\omega_j\|\omega_j\|^{-1} \qquad (2.\ 14)$$

とおくことができる。各パターン $\varphi\in\Phi$ はカテゴリ集合

$$\mathfrak{C} = \{\mathfrak{C}_j \mid j\in J\}$$

内のいずれか一つのカテゴリに帰属しているとして，第 $j\in J$ 番目のカテゴリ $\mathfrak{C}_j$ の代表パターンを $\omega_j\in\Phi$ としており，$\mathfrak{C}_j$ の生起確率を $P(\mathfrak{C}_j)$ としている。

$$[\forall j \in J, 0 < P(\mathfrak{C}_j) < 1] \wedge \Sigma_{j \in J} P(\mathfrak{C}_j) = 1 \qquad (2.15)$$

であることが望ましい。このとき，ξ は全カテゴリ集合 $\mathfrak{C}$ 上の平均化パターン（average pattern）と呼ばれることがある。

S. Suzuki 理論[11]では，n 個の神経ユニット（ニューロン，計算素子）から成る非階層システム（nonhierarchical system）での，第 j 神経ユニットからの出力

$$\eta_j \equiv \Sigma_{\ell \in L} g(\Sigma_{i=1}^{n} W_{ij}(\ell) \cdot u(\eta_i, \ell) + a_j(\ell) - v_j(\ell)) \cdot \theta_\ell(H)\, \xi \| \xi \|^{-1} \in \Phi \qquad (2.16)$$

は，記憶内容 $\xi \| \xi \|^{-1} \in \Phi$ に，式（2．1）のパーセプトロン形作用素 Q の特別な形式としての連想オペレータ（自己共役作用素）

$$\Sigma_{\ell \in L} g(\Sigma_{i=1}^{n} W_{ij}(\ell) \cdot u(\eta_i, \ell) + a_j(\ell) - v_j(\ell)) \cdot \theta_\ell(H) \qquad (2.17)$$

を作用させて得られる連想内容（パターン）である。ここに，

$g : R \to R$：神経ユニット発火関数

$u : \Phi \times L \to R$：特徴抽出写像

$u(\eta_i, \ell) \in R$ はパターン η_i（神経ユニット i からの出力パターン）$\in \Phi$ の，第 $\ell \in L$ 番目の特徴量

$W_{ij}(\ell)$：神経ユニット i から神経ユニット j への結合に関する重み W_{ij} の第 $\ell \in L$ 成分

$a_j(\ell)$：神経ユニット j への外部（からの）入力 a_j の第 $\ell \in L$ 成分

$v_j(\ell)$：神経ユニット j のしきい値 v_j の第 $\ell \in L$ 成分。 □

なお，式（2．17）のパーセプトロン形作用素は，$g(u) \in \{0, 1\}$ なるごとく選んでおけば射影作用素となる[11]ことに注意しておこう。

上述の式（2．16）で示されるニューラルネットは従来の Hopfield neural net の，Hilbert 空間 $\mathfrak{H}$ 上への一般化[11]の第17部～第20部であるが，以下の式（2，18）で示されるニューラルネットは Rumelhart (et al.) neural net の同様な一般化である。[11]の第21部～第23部

S. Suzuki は，これ迄通り $u(\varphi, \ell) \in R$ をパターン $\varphi \in \Phi$ から抽出される第 $\ell \in L$ 番目の特徴量とすると，第 k 層内第 j 神経ユニット出力 η_j^k が

$$\eta_j^k = \Sigma_{\ell \in L} g(S_j^k(\ell)) \cdot \theta_\ell(H)\, \xi \| \xi \|^{-1} \qquad (2.18)$$

ここに，$g : R \to R$ はユニット発火関数

$W_i^{k-1}{}_j^k(\ell)$ は

第 $(k-1)$ 層内第 i ユニットから第 k 層内第 j ユニットへのシナプス結合の第 $\ell \in L$ 番目の重み

として，

$$S_j^k(\ell) = \Sigma_{i=1}^{n(k-1)} W_i^{k-1}{}_j^k(\ell) \cdot u(\eta_i^{k-1}, \ell) \qquad (2.19)$$

と表現される階層形ニューラルネット（hierarchical neural net）を考え，このネットが在来のニューネルネットの定義域を

実数ベクトルの部分集合から，ヒルベルト空間 $\mathfrak{H}$ の部分集合 Φ へと拡張している[27]

事実を示し，誤差逆転播学習アルゴリズムを提案したが，この提案は従来のこの種の学習アルゴリズムを

multichannel の場合

に自然に拡張していることが明らかにされている。

式（2．18）の想起内容 η_j^k もまた，パターン $\xi\|\xi\|^{-1} \in \Phi$ にパーセプトロン形作用素

$$\sum_{\ell\in L} g(S_j^k(\ell)) \cdot \theta_\ell(H) \tag{2．20}$$

を作用させて得られていることに注意しておく。

最後に，計算機シミュレーション済の，S. Suzuki の提案による recurrent neural net の例[9]をあげておこう。[20]

$$\Im_\ell(\varphi) = (f(H) \cdot \theta_\ell(H)\,\varphi,\ \varphi)/(\varphi,\varphi) \tag{2．21}$$

ここに，$f(H)$ は自己共役作用素 H の関数としての正値自己共役作用素

はパターン $\varphi \in \Phi$ から抽出される第 $\ell \in L$ 番目の測度的不変量[1],[21]であるが，しきい値 e_ℓ が不等式

$$0 < e_\ell \leqq (\xi\|\xi\|^{-1} \tag{2．22}$$

を満たせば，構造化モデル写像[21]

$$T\cdot = \sum_{\ell\in L} psn(\Im_\ell(\cdot) - e_\ell) \cdot \theta_\ell(H)\,\xi\|\xi\|^{-1} \tag{2．23}$$

は文献[11]での axiom 1 を満たし，ベキ等性 $TT = T$ が成立しており，収縮写像である[11]。

パターン $\varphi \in \Phi$ のモデル $T\varphi \in \Phi$ は式（2．23）からわかるように，式（2．13）のパーセプトロン形作用素 $Q(\varphi)$ において，φ から抽出される第 $\ell \in L$ 番目の特徴量 $u(\varphi,\ell)$ を

$$u(\varphi,\ell) = psn(\Im_\ell(\varphi) - e_\ell) \tag{2．24}$$

とおいて得られる空間回路をパターン $\xi\|\xi\|^{-1}$ に作用させて得られることに留意しておく。この式（2．24）の $u(\varphi,\ell)$ を使えば，

$$\forall\ \varphi \in \Phi,\ \forall\ \ell \in \mathrm{L},\quad u(T\varphi,\ell) = u(\varphi,\ell) \tag{2．25}$$

が成立していることも証明されている[21]。この式（2．25）の成立が実は，式（2．23）のモデル構成作用素 T がベキ等性 $TT = T$ を満たすことの証明となっている。

U を自己共役作用素 H と可換なウニタリ作用素とすると，U 不変性

$$\forall\varphi\in\Phi,\ TU\varphi = T\varphi \tag{2．26}$$

も成立している。[1],[21]この U 不変性は，パターン φ の持っている性質のある種の座標の選択に関係のない，従ってそのパターン φ の性質を実際に表しているもの（特徴）のみを抽出して，パターン φ の構造化モデル $T\varphi$ が得られている事実を指摘している。

さて，次のシステム方程式（2．27）で記述される非線形連想形記憶器（nonlinear associator）としての再帰形ニューラルネットに注目しよう[9]。

$$\phi_{t,\tau}=\begin{cases} TT\eta_\tau \quad \text{if} \quad \sigma^2<\|T\eta_\tau\|^2 \\ T(T\phi_{t,\tau}+T\eta_\tau) \\ \quad \text{if} \quad \sigma^2\geqq\|T\eta_\tau\|^2 \end{cases} \tag{2．27}$$

ここに，

$$\sigma^2\geqq\|T\eta_\tau\|^2 \quad \text{ならば} \quad T\eta_\tau=0 \tag{2．28}$$

を満たすように，正数 σ^2 を

$$\sigma^2<inf_{\ell\in L}\|\theta_\ell(H)\,\xi\|\xi\|^{-1}\|^2 \tag{2．29}$$

と選んでいるから[9]，結局，式（2，27）は

$$\phi_{t,\tau}=\begin{cases} T\eta_\tau \quad \text{if} \quad \sigma^2<\|T\eta_\tau\|^2 \\ T\phi'_{t,\tau} \quad \text{if} \quad \sigma^2\geqq\|T\eta_\tau\|^2 \end{cases} \tag{2．30}$$

と書ける。

$$\mathfrak{G}_\ell(t,\tau)=\Sigma_{n=1}^{N}\Sigma_{k\in L}a_{\ell k}(n;t)\cdot psn(\mathfrak{F}_k(\phi_{t,\tau-n})-e_k) \tag{2．31}$$

を導入して，2式（2．27），（2．30）内の $\phi'_{t,\tau}$ は，パターン $\xi\|\xi\|^{-1}\in\Phi$ にパーセプトロン形空間回路

$$\Sigma_{\ell\in L}psn(\mathfrak{G}_\ell(t,\tau))\cdot\theta_\ell(H) \tag{2．32}$$

を作用させて得られ，

$$\phi'_{t,\tau}\equiv\Sigma_{\ell\in L}psn(\mathfrak{G}_\ell(t,\tau))\cdot\theta_\ell(H)\,\xi\|\xi\|^{-1} \tag{2．33}$$

と定義されている。

このシステム方程式（2．30）は次の2事柄(a)，(b)を記述している：

(a) $\sigma^2<\|T\eta_\tau\|^2$ が成立し，η_τ が雑音勢力 σ^2 に打ち勝って時刻 τ に存在すれば，入力 η_τ のモデル $T\eta_\tau$ が $\phi_{t,\tau}$ として強制出力される。

(b)実は，文献(11)の第15部，補助定理4，3を適用すれば，式（2．30）内の $T\phi'_{t,\tau}$ に関し，T 不変性

$$T\phi'_{t,\tau}=\phi'_{t,\tau} \tag{2．34}$$

が常に成立している。何故ならば，

$$\forall\,\ell\in L,\ psn(\mathfrak{F}_\ell(\phi'_{t,\tau})-e_\ell)\cdot=psn(\mathfrak{G}_\ell(t,\tau)) \tag{2．35}$$

が成立するからである。よって，式（2．30）の $\phi_{t,\tau}$ に関し，

$$\underline{\phi_{t,\tau}=\phi'_{t,\tau} \quad \text{if} \quad \sigma^2\geqq\|T\eta_\tau\|^2} \tag{2．36}$$

が成立し，

$\sigma^2\geqq\|T\eta_\tau\|^2$ が満たされ，強制入力 η_τ が無視できる程小であれば，

時刻 τ に，過去の時刻 $\tau-n(n=1,2,\cdots,N)$ での出力 $\phi_{t,\tau-n}\in\Phi$ の2値化特徴量（binarized

feature)

$$psn(\mathfrak{F}_k(\phi_{t,\tau-n}) - e_k), k \in L$$

を $a_{\ell k}(n;t)$で重み付けて得られる1次結合量 $\mathfrak{G}_\ell(t,\tau)$ の2値量

$$psn(\mathfrak{G}_\ell(t,\tau))$$

を第 $\ell \in L$ 番目の2値化特徴量として持つパターン $\phi'_{t,\tau}$ を $\phi_{t,\tau}$ として自由再生的に出力する。□

要約すれば，$\phi_{t,\tau} \in \Phi$ は，時刻においてパターン $\eta_\tau \in \Phi$ を受け入れ，時刻 t 迄に得られている重みの組

$$a_{\ell k}(n;t), \quad \ell, k \in L, n = 1 \sim N$$

の下で，時刻 τ において得られる連想出力パターンである。

動作が式（2．27）で記述されるこの連想形記憶器（associative memorizer）が

$$/a/,/i/,/u/,/e/,/o/ \qquad (2,\ 37)$$

という生起順序をもつ日本語単独母音系列をこの順序で記憶し，自由再生（free recall）する機能があることが計算械シミュレーションで確かめられている。[9]

重み $a_{\ell k}(n;t)$ を少量ずつ変更しながら決定していく学習規則は，$|n|$, $|\ell - k|$ 双方の単調非減少関数

$$\Delta_{\ell k}(n) = \frac{1}{(|\ell - k| + 1) \cdot |n| \cdot N \cdot \#(L)} \qquad (2.\ 38)$$

ここに，記号 # は the number of の意を導入し，

$$a_{\ell k}(n;t)|_{t=0} = 0$$
$$a_{\ell k}(n;t+1) = a_{\ell k}(n;t) \cdot + \Delta a_{\ell k}(n;t)$$

ここに，

$$\Delta a_{\ell k}(n;t) = [psn(\mathfrak{F}_\ell(\phi_{t,t}) - e_\ell) - psn(\mathfrak{G}_\ell(t,t)] \cdot psn(\mathfrak{G}_k(\phi_{t-n,t-n}) - e_k) \cdot \Delta_{\ell k}(n) \qquad (2.\ 39)$$

としている。この学習規則（2．39）は，

$$psn(\mathfrak{G}_k(\phi_{t-n,t-n}) - e_k) = 1$$

である限り，

$$\forall t(=0,1,2,\cdots), \forall \ell \in L, psn(\mathfrak{F}_\ell(\phi_{t,t}) - e_\ell) = psn(\mathfrak{G}_{\ell(t,t)}) \qquad (2.\ 40)$$

が達成されるように，重み $a_{\ell k}(n;t)$ を変更・自己組織化して行くようなものである。

システム階数（何単位時間迄過去にさかのぼってパターンを記憶するかという整数値）N に関し，次の事実が認められた：

記憶すべきパターン系列の周期 p（上記の日本語単独母音系列（2．37）では $p=5$）に対し，

不等式

$$N \geqq p/2 \qquad (2.\ 41)$$

を満たせば，

式（2．40）がほぼ成立するという意味で学習は約 $6N$ 期間でほぼ完了し，この直後の $(N+1)$ 個の時点において正しく自由再生する。また，N 個の時点にわたって連続的に強制入力（上記の事項(a)で示したように，不等式 $\sigma^2 < \|T\eta_\tau\|^2$ を満たすパターン η_τ を入力することの意）した直後においては，一周期 p 位の期間にわたり，正しく自由再生する。

3．誤差確率分布を考慮した誤差逆伝播学習アルゴリズム MLBPLA の定式化

一般回帰問題（general regression problem）とは，有限個の観測値から，統計的モデル内の諸パラメータを推定することである。[22] 結果としては回帰推定値（regression estimator）として，階層形ニューラルネットの重み $W^{k-1\ k}_{\ i\ j}$ が得られるように，希望出力からのズレに対し頑健な（robust）学習法を，

最尤法（method of maximum likelihood）

を介し，研究しよう。誤差分布（error distribution）として，平均値 0，等分散の正規分布を採用すると，従来の Rumelhart et al. の学習公式[2],[14]が得られる様な一般的な結果が示される。

3．1　従来の回帰推定と最急上昇法

例えば，線形回帰モデル（linear regression model）

$$y_i = \sum_{j=1}^{p} \beta_j \cdot x_{ji} + \varepsilon_i,\ i = 1, 2, \cdots, n$$

を考え，各誤差項 ε_i が互いに統計的に独立に，確率密度関数 $f(\varepsilon)$ をもつ同一の確率分布に従うものとする。このとき，尤度関数（likelihood function）K は

$$K = \Pi_{i=1}^{n} f\left(y_i - \sum_{j=1}^{p} \beta_j \cdot x_{ji}\right)$$

であり，回帰係数（regression）β_j の最尤推定量（maximum likelihood estimate）は対数尤度関数 $\log K$ の値が最大になる解として与えられる。そのためには，対数尤度方程式，つまり，連立方程式

$$\frac{\partial}{\partial \beta_j} \log K = -\sum_{i=1}^{n} \frac{f'\left(y_i - \sum_{j=1}^{p} \beta_j \cdot x_{ji}\right)}{f\left(y_i - \sum_{j=1}^{p} \beta_j \cdot x_{ji}\right)} \cdot x_{ji} = 0$$
$$, j = 1, 2, \cdots p$$

を解けばよい[15]。ただし，$f'(u)$ は $f(u)$ の導関数である。

これは正に，神経ユニット発火関数 $g(u)$ を $g(u) = u$ とした場合の 2 層 neural net の重み β_j の組の一括決定法である。

本論文では，逐次決定法的に m 層 neural net の重み $W^{k-1\ k}_{\ i\ j}$ を決める学習法が，上述に hint を得て，研究される。いいかえれば，解析的に解を求めるのが困難とみて，最急上昇法（hill-climbing method）

$$d\beta_j/dt = (\partial/\partial\beta_j) \log K, j = 1 \sim p \qquad (3.\ 1)$$

により逐次的に解を求める学習過程

$$\beta_j(t+\Delta t)=\beta_j(t)+\Delta\beta_j(t)$$

を導入すればよい。何故ならば

$$\frac{d\log K}{dt}=\sum_{j=1}^{p}\frac{\partial\log K}{\partial\beta_j}\cdot\frac{d\beta_j}{dt}$$
$$=\sum_{j=1}^{p}[(\partial/\beta_j)\log K]^2\geqq 0 \qquad (3.\ 2)$$

が成立し，$\log K$ は微分方程式（3．1）の解曲線の上で決して減少しないからである。この微分方程式（3．1）を解き，十分時間が経過したときの β_j を求めれば，$\log K$ を最大とする回帰係数 β_j が得られるからである。

3．2　最急降下法と m 層ニューラルネットの学習方程式

第 k 層内第 j 神経ユニットからの出力 x_j^k が，第 $(k-1)$ 層内第 i 神経ユニットからの出力 x_i^{k-1} に重み $W_{i\ j}^{k-1\ k}$ をかけその総和

$$u_j^k=\sum_{i=1}^{n(k-1)}W_{i\ j}^{k-1\ k}\cdot x_i^{k-1} \qquad (3.\ 3)$$

が入力された，神経ユニット発火関数

$$g:R\rightarrow R \qquad (3.\ 4)$$

からの出力として，

$$x_j^x=g(u_j^k),\ j=1\sim n(k) \qquad (3.\ 5)$$

と定まる，m 層から成る階層形ニューラルネットを考えよう $(k=1\sim m)$。$W_{i\ j}^{k-1\ k}$ は第 $(k-1)$ 層内第 i ユニットから第 k 層内第 j ユニットへの結合の重み（a weight of the connection from i-th unit in $(k-1)$-th layer to j-th unit in k-th layer）である。

ある時刻 t において，入力層（第1層）に入力

$$s=\{s_j\,|\,j=1\sim n(1)\} \qquad (3.\ 6)$$

が加えられたとき，出力層（第 m 層）に出力（希望出力，理想出力；desired output）

$$y=\{y_j\,|\,j=1\sim n(m)\} \qquad (3.\ 7)$$

が得られることを要求しよう。このとき，時刻 t での重みの組を

$$W\equiv W(t)\equiv\{W_i^{k\ k+1}{}_j(t)\,|\,i=1\sim n(k),\ j=1\sim n(k+1),\ k=1\sim m-1\} \qquad (3.\ 8)$$

と表記すると，

$u_j^1=s_j,\ j=1\sim n(1)$ を入力し，このニューラルネットを実際に稼働させて得られる出力（現実出力；actual output）

は，重み W と入力 s の関数として，

$$x_j^m(W,s) \qquad (3.\ 9)$$

と書ける。

$$\langle s, y \rangle \tag{3．10}$$

は入力 s とその対応する希望出力 y とのなす対（pair）であり，時刻 t において入力される訓練例（training example）と呼ばれる。

$$z_j = x_j^{\mathrm{m}}(W, s) - y_j \tag{3．11}$$

は，第 m 層内第 j ユニットからの出力誤差（error between the actual output and the desired output）である。このとき，誤差 z_j の確率密度関数（probability density function）

$f_j^m(z_j)$，ここに，

$$[\forall u, f_j^m(u) \geqq 0] \wedge \int_{-\infty}^{+\infty} du f_j^m(u) = 1 \tag{3．12}$$

を想定し，符号反転型対数尤度

$$\begin{aligned} E &\equiv E(W) \equiv E(W, < s, y >) \\ &\equiv \sum_{j=1}^{n(m)} \log e \, [f_j^m(z_j)]^{-1} \\ &\equiv - \sum_{j=1}^{n(m)} \log e \, f_j^m(z_j) \end{aligned} \tag{3．13}$$

を最小とするように，重み $W_{i\ j}^{k-1\ k}(t)$ を

$$W_{i\ j}^{k-1\ k}(t + \Delta t) \equiv W_{i\ j}^{k-1\ k}(t) + \Delta W_{i\ j}^{k-1\ k}(t) \tag{3．14}$$

という方式で小量ずつ変更していくとしよう（更新ルール； update rule）。

この更新方式は，ε' を十分小さい正数として，$W_{i\ j}^{k-1\ k}(t)$ の時間変動を記述する微分方程式（学習方程式）

$$\frac{d}{dt} W_{i\ j}^{k-1\ k}(t) = - \varepsilon' \cdot \frac{\partial E}{\partial W_{i\ j}^{k-1\ k}(t)} \tag{3．15}$$

を導入すれば，

$$\begin{aligned} \frac{d}{dt} E &= \sum_{k=2}^{m} \sum_{i=1}^{n(k-1)} \sum_{j=1}^{n(k)} \frac{\partial E}{\partial W_{i\ j}^{k-1\ k}(t)} \cdot \frac{dW_{i\ j}^{k-1\ k}(t)}{dt} \\ &= - \varepsilon' \cdot \sum_{k=2}^{m} \sum_{i=1}^{n(k-1)} \sum_{j=1}^{n(k)} \left(\frac{\partial E}{\partial W_{i\ j}^{k-1\ k}(t)} \right)^2 \leqq 0 \end{aligned} \tag{3．16}$$

となり，E は学習方程式（3．14）の解曲線の上で決して増加しないこが判明するから，この学習方程式（3．14）を解き，十分時間が経過したときの $W_{i\ j}^{k-1\ k}(t)$ の値を求めることの近似である。実際の学習過程は，式（3．14）において各時刻に各々相異なる組〈入力，理想出力〉を与えることになるのであるが。

従って，最急降下法（gradient descent scheme）によれば，式（3．14）内の変更分 $\Delta W_{i\ j}^{k-1\ k}(t)$ を

$$\Delta W_{i\ j}^{k-1\ k} = - \varepsilon \cdot \frac{\partial E}{\partial W_{i\ j}^{k-1\ k}(t)}, \quad \text{ここに} \varepsilon = \varepsilon' \cdot \Delta t \tag{3．17}$$

と与えればよい。

通常，推定誤差 z_j の絶対値 $|z_j|\rightarrow$ 小であれば，z_j の生起確率 $f_j^m(z_j)\cdot\Delta z_j\rightarrow$ 大になり，$E\rightarrow$ 小という関係が得られるような $f_j^m(z_j)$，例えば $|z_j|$ の減少関数であるような $f_j^m(z_j)$ を考えていることになる。

$\log[f_j^m(z_j)']^{-1}$ は一種の対数尤度関数であるから，この学習法は

最尤法を出力層の各神経ユニットに対し適用し，$W_{ij}^{k-1\,k}$ の最尤推定量（maximum-likelihood-type estimator）

を求めている。これが

現在の重み $W_{ij}^{k-1\,k}(t)$ を修正して新しい重み $W_{ij}^{k-1\,k}(t+\Delta t)$ を得るという学習規則（learning rule）としての最尤型誤差逆伝播学習アルゴリズム MLBPLA

を提供する。MLBPLAは

時刻 t での，式（3．14）で示される重み $W_{ij}^{k-1\,k}(t)$ の更新を，入力・理想出力の集合（訓練例の集合；a set of training examples）

$$\{<s^{(q)},y^{(q)}>\,|\,q=1\sim N\} \qquad (3．18)$$

の中から一つの訓練例 $\langle s,y\rangle$ をランダムに選び，各時刻にわたり繰り返し実行すること

（3．19）

で構成される。MLBPLA の適用によって，誤差分布に適応した重みの組 W が決定されるということになる。

4．MLBPLA の具体化

前章においては最尤型誤差逆伝播学習アルゴリズム MLBPLA が定式化された。本章では，このアルゴリズム MLBPLA を具体化するために，学習方程式（3．14）の重み $W_{ij}^{k-1\,k}(t)$ の更新式内の更新分 $\Delta W_{ij}^{k-1\,k}(t)$ である式（3．17）の表現をあらかじめ，求めておくことにしよう。

4．1　更新分 $\Delta W_{ij}^{k-1\,k}(t)$ の具体的表現

式（3．17）内の微分係数

$$\partial E/\partial W_{ij}^{k-1\,k}(t)$$

を計算しよう。

さて，

重み $W_{ij}^{k-1\,k}(t)$ の変化は，式（3．3）の

$$u_j^k=\sum_{i=1}^{n(k-1)}W_{ij}^{k-1\,k}(t)\cdot x_i^{k-1}$$

の変化をもたらし，この u_j^k の変化が式（3．13）の $E(W(t))$ の変化をもたらすから，

$$\frac{\partial E(W(t))}{\partial W_{ij}^{k-1\,k}(t)}=\frac{\partial E(W(t))}{\partial u_j^k}\cdot\frac{\partial u_j^k}{\partial W_{ij}^{k-1\,k}}=d_j^k\cdot\frac{\partial u_j^k}{\partial W_{ij}^{k-1\,k}},$$

ここに，$d_j^k=\partial E(W(t))/\partial u_j^k$ （4．1）

である。ところが

$$\partial u_j^k / \partial W_{ij}^{k-1\,k} = x_i^{k-1} \qquad (4.\ 2)$$

が成立しているから，結局

$$\partial E(W(t)) / \partial W_{ij}^{k-1\,k}(t) = d_j^k \cdot x_i^{k-1} \qquad (4.\ 3)$$

である。この式（4．3）を式（3．17）に代入すれば，更新分 $\Delta W_{ij}^{k-1\,k}(t)$ の表現

$$\Delta W_{ij}^{k-1\,k}(t) = -\varepsilon \cdot d_j^k \cdot x_i^{k-1},\ i = 1 \sim n(k-1),\ j = 1 \sim n(k),\ k = 2 \sim m \qquad (4.\ 4)$$

が得られる。

式（4．1）での d_j^k をより具体的に計算しよう。

4．2　$k = 1 \sim m-1$ のときの d_j^k の計算

式（4．1）での d_j^k は

u_j^k の変化が式（3．5）の $x_j^k = g(u_j^k)$ の変化をもたらし，x_j^k の変化が式（3．3）の意味する

$$u_i^{k+1} = \sum_{j=1}^{n(k)} W_{ji}^{k\,k+1} \cdot x_j^k$$

の変化をもたらし，u_i^{k+1} の変化が式（3．13）の $E(W(t))$ の変化を $i = 1 \sim n(k+1)$ にわたってもたらす

から，

$$d_j^k = \partial E(W(t)) / \partial u_j^k$$

$$= \sum_{i=1}^{n(k+1)} \frac{\partial E(W(t))}{\partial u_i^{k+1}} \cdot \frac{\partial u_i^{k+1}}{\partial x_j^k} \cdot \frac{\partial x_j^k}{\partial u_j^k}$$

$$= \left[\sum_{i=1}^{n(k+1)} d_i^{k+1} \cdot \frac{\partial u_i^{k+1}}{\partial x_j^k}\right] \cdot \frac{\partial x_j^k}{\partial u_j^k} \qquad (4.\ 5)$$

と計算される。ここで，

$$\partial x_j^k / \partial u_j^k = (dg(u)/du)\big|_{u=u_j^k} \qquad (4.\ 6)$$

$$\partial u_i^{k+1} / \partial x_j^k = W_{ji}^{k\,k+1} \qquad (4.\ 7)$$

が成立しているから，2式（4．6），（4．7）を式（4．5）に代入すれば，具体的に

$$d_j^k = \left[\sum_{i=1}^{n(k+1)} W_{ji}^{k\,k+1}(t) \cdot d_i^{k+1}\right] \cdot \left.\frac{dg(u)}{du}\right|_{u=u_j^k}$$

$$k = 1 \sim m-1,\ j = 1 \sim n(k) \qquad (4.\ 8)$$

が得られる。

4．3　微分係数 d_j^m の計算

$k = m$ のときの d_j^k つまり微分係数 d_j^m を計算しよう。

さて，

u_j^m の変化が式（3．3）の $x_j^m = g(u_j^m)$ の変化をもたらし，x_j^m の変化が式（3．13）の $E(W(t))$ の変化をもたらす

から，$k = m$ のときの $d_j^k = d_j^m$ は，式（4．1）の定義から，

$$d_j^m = \frac{\partial E(W(t))}{\partial u_j^m} = \frac{\partial E(W(t))}{\partial x_j^m} \cdot \frac{\partial x_j^m}{\partial u_j^m} \qquad (4．9)$$

と計算される。ここで，

$$\partial x_j^m / \partial u_j^m = (dg(u)/du)\,|_{u=j^m} \qquad (4，10)$$

であり，$\partial E(W(t))/\partial x_j^m$ は2式（3．11），（3．13）から

$$\begin{aligned}
&\partial E(W(t))/\partial x_j^m \\
&= (\partial/\partial x_j^m) \sum_{\ell=1}^{n(m)} - \ell og_e f_\ell^m (x_\ell^m (x_\ell^m(W, s) - y_\ell) \\
&= -(\partial/\partial x_j^m)\, \ell og_e f_j^m (x_j^m(W, s) - y_j) \\
&= -\frac{1}{f_j^m(x_j^m(W, s) - y_j)} \cdot \frac{df_j^m(u)}{du}\Big|_{u=x_j^m(W,s)-y_j} \qquad (4．11)
\end{aligned}$$

と計算されるから，2式（4．10），（4．11）を式（4．9）に代入すれば，最終的に，d_j^m は

$$d_j^m = -\frac{1}{f_j^m(x_j^m(W, s) - y_j)} \cdot \frac{df_j^m(u)}{du}\Big|_{u=x_j^m(W,s)-y_j} \cdot \frac{dg(u)}{du}\Big|_{u=u_j^m}, j = 1 \sim n(m) \qquad (4，12)$$

と求められる。

MLBPLA の最終的な表現である（3．19）における，重み $W_{i\ j}^{k-1\ k}$ の，時刻 t での更新式（3．14）は，式（4．4）の更新分 $\Delta W_{i\ j}^{k-1\ k}(t)$ において，式（4，8）の微分係数 d_j^k，式（4．12）の微分係数 d_j^m を代入すれば，計算されることに注意しておこう。

4．4　加重回帰的推定

加重平方和（以下の式（4．13））を最小にするように回帰係数（式（3．3）内の重み $W_{i\ j}^{k-1\ k}$）を推定する手法のことを

加重回帰（weighted regression）

というが，加重つき適応誤差（weighted adaptive error）

$$\sum_{j=1}^{n(m)} v_j \cdot [x_j^m(W, s) - y_j]^2 \qquad (4．13)$$

を考え（式（3．11）をみよ），式（3．13）の符号反転型対数尤度 $E(W, <s, y>)$ を

$$\begin{aligned}
&E = E(W) = E(W, <s, y>) \\
&= \sum_{j=1}^{n(m)} \ell og_e[f_j^m(x_j^m(W, s) - y_j)]^{-1} \\
&= \sum_{j=1}^{n(m)} v_j \cdot [x_j^m(W, s) - y_j)]^2 \qquad (4．14)
\end{aligned}$$

とおいてみよう。このとき，

$$\partial E(W(t))/\partial x_j^m$$
$$= 2v_j \cdot [x_j^m(W,s) - y_j] \qquad (4.15)$$

となる。よって，式（4．11）に注目し，

$$2v_j = -\frac{1}{[x_j^m(W,s) - y_j]} \cdot \frac{1}{f_j^m(x_j^m(W,s) - y_j)} \cdot \frac{d}{du} f_j^m(u)\Big|_{u=x_j^m(W,s)-y_j} \qquad (4.16)$$

とおくと，

$$\partial E(W(t))/\partial x_j^m$$
$$= -\frac{1}{f_j^m(x_j^m(W,s) - y_j)} \cdot \frac{d}{du} f_j^m(u)\Big|_{u=x_j^m(W,s)-y_j} \qquad (4.17)$$

を得，式（4．11）と一致することが知れる。

上述は次の事実を意味する：本研究で提案する最尤型誤差逆伝播学習アルゴリズム MLBPLA は，m 層階層型ニューラルネットでの出力層（第 m 層）内の第 j 神経ユニットからの推定誤差である式（3．11）の

$$z_j = x_j^m(W,s) - y_j$$

に，式（4．16）でいう v_j を式（4．14）のごとく重みとして採用した場合に相当し，加重回帰推定量として，式（3．3）内のニューラルネットのシナプス結合の重み $W_i^{k-1}{}_j^k(t)$ を学習している方式である。 □

なお，小池・田辺[23]は階層形ニューラルネットの学習での自乗誤差評価関数 E として，

$$E = \frac{1}{2}\Sigma_j h(\widehat{y_j} \cdot (y_j - \widehat{y_j})^2$$

を用いている。ここに，

y_j はユニット j の出力

$\widehat{y_j}$ は y_j に対応する教師信号（理想出力）

であり，

$h(\widehat{y_j})$ は $0 \leqq \widehat{y_j} \leqq 1$ なる $\widehat{y_j}$ について単調増加する重み関数

とみなし，

$$k(\widehat{y_j}) = \widehat{y_j}^2$$

を採用しているが，この事態は，式（4．14）において，重み v_j を天下り的に

$$v_j = 2^{-1} \cdot h(\widehat{y_j}) = 2^{-1} \cdot \widehat{y_j}^2$$

としたものに相当する。式（4．16）での重み v_j の表現式を勘案し理解できるように，このような天下り的設定が果して妥当であるかどうかは理論的には疑問の余地があろう。これは次章の解析を見ればを勘案すれば了解できよう。

5．正規分布，指数分布，t 分布における学習

本章では，式（3．12）での，誤差確率分布の密度関数 $f(u)=f_j^m(u)$ として次の3種 i，ii，iii を採用した場合を解析する。具体的には，式（4．11）の偏微分係数

$$\partial E(W(t))/\partial x_j^m$$

を各々の場合に計算することになる。

(i) 正規分布（normal distribution）$N(a, \sigma^2)$ の確率密度

$$f(u)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot exp\left[-\frac{(u-a)^2}{2\sigma^2}\right],\quad -\infty<u<+\infty$$

(ii) 指数分布（exponential distribution）$L(\lambda, \alpha)$ の確率密度

$$f(u)=\frac{1}{2\alpha}\cdot exp\left[-\frac{|u-\lambda|}{\alpha}\right],\quad -\infty<u<+\infty$$

(iii) 自由度 $n(\geqq 1)$ の t 分布（student's distribution）の確率密度

$$f(u)=C\cdot\frac{1}{(1+u^2/n)^{\frac{n+1}{2}}},\quad -\infty<u<+\infty$$

備考5．1（χ^2（カイ自乗）分布）

$X_1, X_2, \cdots, X_n$ が独立で，同一の正規分布 $N(0,1)$ を持つならば，

$$Y_n=X_1+X_2+\cdots+X_n$$

の確率密度 $k_n(x)$ は

$$k_n(x)=\begin{cases}0 & \text{if}\quad x\leqq 0\\ C\cdot x^{\frac{n}{2}-1}\cdot exp\left[-\frac{x}{2}\right] & \text{if}\quad x>0\end{cases}$$

であり，

Y_n の期待値 $E(Y_n)=n$

Y_n の分散 $\sigma^2(Y_n)=E((Y_n-E(Y_n)^2)=2n$

が成立する。確率密度 $k_n(x)$ を持つ分布を自由度 n の χ^2 分布という[26]。なお，定数 c は

$$\int_0^\infty du\, k_n(u)=1$$

となるように定める。

備考5．2（コーシー分布）

コーシー分布（Cauchy's distribution）$C(\lambda, \alpha)$ 確率密度 $g(x)$ は

$$g(x)=\frac{1}{\pi}\cdot\frac{2}{\alpha^2+(x-\lambda)^2},\quad -\infty<x<+\infty$$

である。

$$\int_{-\infty}^{x} du\, g(u)=\frac{1}{2}$$

を満たす x（中央値，メジアン）は $x=\lambda$ である[26]。X_0 が $C(0,1)$ なる分布を定めるとき，$Y=\alpha\cdot X_0+\lambda(\alpha>0)$ の分布は $C(\lambda,\alpha)$ であることが知られている。

備考５．３（t 分布）

上述の２備考５．１，５．２での χ^2 分布，コーシー分布と正規分布と関連して，t 分布を説明しておこう[26]。

X, Y_n が独立で，各々正規分布 $N(0,1)$，自由度 n の χ^2 分布を持つならば，

$$z_n=X/\sqrt{Y_n/n}$$

の確率密度は

$$s_n(z)=C\cdot\frac{1}{\left(1+\frac{z^2}{n}\right)^{\frac{n+1}{2}}}\ (n\geqq 1)$$

である。定数 C は

$$\int_{-\infty}^{+8} dz\, s_n(z)=1$$

となる様に求める。このとき，自由度 n の $t-$ 分布が定義されている。

ここで，$n=1$ とすれば，

$$s_1(z)=c/(1+z^2)$$

となり，これはコーシー分布 $C(0,1)$ の確率密度である。また，

$$\lim_{n\to\infty}\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}=\exp\left[\frac{x^2}{2}\right]$$

から予想されるように，

$n\to\infty$ とすれば，$s_n(z)$ は正規分布 $N(0,1)$ の確率密度に収束する。実用上は，$n>30$ で，$s_n(z)$ と標準正規曲線とは一致するものと考えて，ほとんど差支えない[26]。 □

５．１　正規誤差分布における学習

第 m 層（出力層）内第 j ユニットの適応誤差である式（３．11）の $z_j=x_j^m(W,s)-y_j$ が正規分布

$$N(a_j^m,(\sigma_j^m)^2)$$

に従う場合，その確率密度関数 $f_j^m(z_j)$ は

$$f_j^m(z_j) = \frac{1}{\sqrt{2\pi(\sigma_j^m)^2}} \cdot \exp\left[-\frac{(z_j - a_j^m)^2}{2(\sigma_j^m)^2}\right], \quad -\infty < z_j < +\infty \tag{5．1}$$

であるから，

$$-\frac{1}{f_j^m(z_j)} \cdot \frac{df_j^m(z_j)}{dz_j} = \frac{(z_j - a_j^m)}{(\sigma_j^m)^2} \tag{5．2}$$

が成り立つ。

よって，式（4．16）の $2v_j$ は

$$2v_j = \frac{1}{[x_j^m(W, s) - y_j]} \cdot \frac{[x_j^m(W, s) - y_j - a_j^m]}{(\sigma_j^m)^2} \tag{5．3}$$

と表現され，式（4．17）の偏微分係数 $\partial E(W(t))/\partial x_j^m$ は

$$\frac{\partial E(W(t))}{\partial x_j^m} = \frac{[x_j^m(W, s) - y_j - a_j^m]}{(\sigma_j^m)^2} \tag{5．4}$$

と表現されることが判明し，

$$a_j^m = 0, \quad 1/(\sigma_j^m)^2 = 2 \tag{5．5}$$

の場合は，

$$v_j = 1 \tag{5．6}$$

を得て，式（4．14）からわかるように，最小自乗法（method of least squares）による結果と一致する。即ち，学習方程式（3．14）を繰り返し適用し求められるニューラルネットのシナプス結合の重み $W_i^{k-1}{}_j^k(t)$ は最小2乗推定量（least squares estimator）に収束することが知れる。

適応すべき誤差 z_j が正規分布に従うとき，最小2乗推定量が最尤推定量になることを意味している。

一般に，推定結果が少数個の誤差項の大きい訓練入力 $\langle s, y\rangle$ に引っ張られるのは好ましくはない。正規分布より裾野が広い誤差確率分布の場合には，絶対値の大なる誤差が出現しやすい。そのため，等分散形誤差正規分布のように，等ウェイトの最小2乗推定を行うと（これが従来の Rumelhart et al. の誤差逆伝播学習[14]である），絶対値の大きい誤差項をもつ訓練入力が推定結果を左右しがちなことに留意しておかねばならない。

5．2　指数誤差分布における学習

出力層内第 j 神経ユニット出力の適応誤差である式（3．11）の z_j が指数分布 $L(a_j^m, \alpha_j^m)$ に従う場合，その確率密度関数 $f_j^m(z_j)$ は

$$f_j^m(z_j) = \frac{1}{2\alpha_j^m} \cdot \exp\left[-\frac{|z_j - a_j^m|}{\alpha_j^m}\right], \quad -\infty < z_j < +\infty \tag{5．7}$$

であり，

$$\mathrm{sgn}(x) = +1 \quad \text{if} \quad x > 0, = 0 \quad \text{if} \quad x = 0,$$

$= -1 \quad \text{if} \quad x < 0$

また，$h(x) = \exp[-|x|]$ （5．8）

として，

$$(d/dx)h(x) = -\operatorname{sgn}(x) \cdot h(x) \tag{5．9}$$

が成り立つことを使えば，

$$\frac{b}{dz_j} f_j^m(z_j) = -\operatorname{sgn}(z_j - a_j^m) \cdot f_j^m(z_j) \cdot \frac{1}{\alpha_m} \tag{5．10}$$

がいえ，結局

$$-\frac{1}{f_j^m(z_j)} \cdot \frac{b}{dz_j} f_j^m(z_j) = \operatorname{sgn}(z_j - a_j^m) \cdot \frac{1}{\alpha_m} \tag{5．11}$$

が成り立つ。よって，重みつき適応自乗誤差である式（4．14）での$E(W)$内の，式（4．16）の重み$2v_j$は

$$2v_j = \frac{1}{[x_j^m(W,s) - y_j]} \cdot \operatorname{sgn}(x_j^m(W,s) - y_j - a_j^m) \cdot \frac{1}{\alpha_j^m} \tag{5．12}$$

と表現され，

任意の実数zは$z = \operatorname{sgn}(z) \cdot |z|$と表現される

ことを使用すれば，

$a_j^m = 0$の場合

$$2v_j = \frac{1}{\alpha_j^m} \cdot |x_j^m(W,s) - y_j|^{-1} \tag{5．13}$$

となる。また，式（4．17）の偏微分係数$\partial E(W(t))/\partial x_j^m$は

$$\frac{\partial E(W(t))}{\partial x_j^m} = \operatorname{sgn}(x_j^m(W,s) - y_j - a_j^m) \cdot \frac{1}{\alpha_j^m} \tag{5．14}$$

と表現される。

さて，式（5．13）の$2v_j$を，式（4．14）の$E(W)$に代入してみると，

$$\begin{aligned} &E(w) \\ &= \textstyle\sum_{j=1}^{n(m)} v_j \cdot [x_j^m(W,s) - y_j]^2 \\ &= \textstyle\sum_{j=1}^{n(m)} \frac{1}{2} \cdot (\alpha_j^m)^{-1} \cdot |x_j^m(W,s) - y_j| \end{aligned}$$

，ここに，$a_j^m = 0$ （5．15）

と再表現され，これは出力$x_j^m(W,s)$の理想出力y_jについての，絶対偏差（absolute deviation）である。よって，

誤差が指数分布に従うとき，最小絶対偏差推定量（least absolute deviation estimator）が最尤推定量になる

という，多変量解析法（multivariate analysis）における通常の回帰分析（regression analysis）と一致することが示された。

5．3　t 分布における学習

式（3．11）で示される出力層第 j 神経ユニットの適応誤差 z_j が自由度 $q(\geqq 1)$ の t 分布に従う場合，その確率密度関数 $f_j^m(z_j)$ は

$$f_j^m(z_j) = C \cdot \frac{1}{\left[1 + \frac{(z_j - a_j^m)^2}{q_j}\right]^{\frac{q+1}{2}}}, \quad -\infty < z_j < +\infty \tag{5．16}$$

である。

$q_j = 1$ の場合

$$f_j^m(z_j) = C \cdot \frac{1}{1 + (z_j - a_j^m)^2} \tag{5．17}$$

となり，$C = 1/\pi$ と採ると，コーシー分布 $C(a_j^m, 1)$ の確率密度となる。また，$a_j^m = 0$ とすれば，式（5．16）の $f_j^m(z_j)$ は $q_j \to \infty$ のとき，正規分布 $N(0, 1)$ の確率密度

$$f_j^m(z_j) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[\frac{z_j^2}{2}\right], \quad -\infty < z_j < +\infty \tag{5．18}$$

に漸近し，収束する。

簡単な計算から，

$$-\frac{1}{f_j^m(z_j)} \cdot \frac{d}{dz_j} f_j^m(z_j) = (q_j + 1) \cdot \frac{(z_j - a_j^m)}{q_j + (z_j - a_j^m)^2} \tag{5．19}$$

が知れ，よって，式（4．14）の重みつき自乗誤差 $E(W)$ 内の式（4．16）の重み $2v_j$ は

$$2v_j = \frac{1}{[x_j^m(W, s) - y_j]} \cdot (q_j + 1) \cdot \frac{[x_j^m(W, s) - y_j - a_j^m]}{q_j + [x_j^m(W, s) - y_j - a_j^m]^2} \tag{5．20}$$

と表わされる。さらに，式（4．17）の偏微分係数については

$$\frac{\partial E(W(t))}{\partial x_j^m} = (q_j + 1) \cdot \frac{[x_j^m(W, s) - y_j - a_j^m]}{q_j + [x_j^m(W, s) - y_j - a_j^m]^2} \tag{5．21}$$

と求められる。

自由度 q_j が小さいときは，重み v_j を

$$|x_j^m(W, s) - y_j - a_j^m|^{-2} \tag{5．22}$$

の割合で小さくすることが学習にとって望ましい。逆に，自由度 q_j が十分大きいときは，

$2v_j =$ 一定　for any j

でも差支えない。

このように，誤差が裾野の広い分布に従う場合に，頑健性（robustness）が保持されるためには（学習効果が損なわれないためには），重み v_j は

$$|x_j^m(W, s) - y_j - a_j^m| \qquad (5.\ 23)$$

の減少関数であることが望ましい。すなわち，絶対値の大きい誤差をもつような現実出力 $x_j^m(W, s)$ に対する重み v_j を相対的に小さくすることが望ましい。最小絶対偏差推定法の重み v_j は式（5．13）からわかるように，確かにこの条件を満たしていることに注意しておこう。

6．むすび

例えば，式（3．11）で示される出力層内第 j 神経ユニットの適応誤差 z_j が確率 $1-\varepsilon_j$ を持ち，その誤差分布（error distribution）の確率密度関数（density function）$f_j^m(u)$ が

$$\rho_j(u) = \begin{cases} u^2/2 & \text{if} \quad |u| < a_j \\ a_j \cdot |u| - a_j^2/2 & \text{if} \quad |u| \geqq a_j \end{cases}$$

として，

$$f_j^m(z_j) = C \cdot (1-\varepsilon_j)(1/\sqrt{2\pi}) \cdot \exp[-\rho_j(z_j)], \quad a_j > 0, c_j > 0, 0 \leqq \varepsilon_j \leqq 1$$

と表現される場合[22]を想定してみよう。これは，

a combinatory distribution so that it is Gaussian in the middle and Laplacian at the tails with much large variance

を想定したことに相当する。このとき，

$$-\frac{1}{f_j^m(z_j)} \cdot \frac{d}{dz_j} \cdot f_j^m(z_j) = \frac{d}{dz_j} \cdot \rho_j(z_j)$$

ここに，$(d/dz_j)\rho_j(z_j)$

$= \max\{-a_j, \min\{z_j, a_j\}\}$

が成立し，式（4．14）で示される重みつき自乗誤差 $E(W)$ 内の重みである v_j は，式（4．16）から，

$$2v_j = \frac{1}{[x_j^m(W,s) - y_j]} \cdot \frac{d}{du}\rho_j(u)\,|_{u=x_j^m(w,s)-y_j}$$

と表現され，式（4．17）の偏微分係数

$\partial E(W(t))/\partial x_j^m$ は

$\partial E(W(t))/\partial x_j^m$

$$= \frac{d}{du}\rho_0(u)\,|_{u=x_j^m(w,s)-y_j}$$

と表される。本最尤形誤差逆伝播学習アルゴリズム（3．19）はこの場合，有効に機能することが期待される。

階層形ニューラルネットは前進形の多層ネット（multi-layer feedforward network）とも呼ばれるが，この様なネットのシナプス結合の重み $W_i^{k-1}{}_j^k$ の組を推定するのに，多変量解析法での回帰分析を本格的に適用することが望まれていたことは文献㉓での天下り的な重みの設定法からも理解できよう（4．4節を参照）。本最尤法で得られる重み $W_i^{k-1}{}_j^k$ の組は，加重つき適応誤差を極小にするように回帰係数として $W_i^{k-1}{}_j^k$ を推定する加重回帰法を適用したものと同一になった。

最小2乗推定量は，線形で不偏な推定量のクラスの中で最小分散をもつという意味で最良である。さらに，一層強力な非線形な推定量を含めたすべての不偏推定量のクラスの中で最小2乗推定量が最良であるためには，各誤差が互いに独立で正規分布 $N(0, \sigma^2)$ に従うことを仮定しなければならないことはよく知られている。

観測誤差の分布が正規分布によって良く近似できることはガウスによって論証され，ガウスが最小2乗法を発案するに至ったのも正規分布を仮定したからである。Rumelhart et al. などの従来の単純な誤差逆伝播法[14]，[23]は誤差分布を仮定していなくて，最小2乗法を適用して，ニューラルネットの各層間結合の重み，$W_i^{k-1}{}_j^k$ を逐次学習で求める手法である。この従来の学習法は誤差分布として，等分散の正規分布を仮定していることに相当することは5．1節で示された。

本手法は唯単純に最尤法を適用したのではない。単純に，情報量密度を極値ならしめる手法としての最尤法を適用するならば，式（3．13）の $E(W)$ は

$$f = f_j^m \text{ for any } j$$

として，

$$E(W) = \Pi_{j=1}^{n(m)} \log f(z_j)$$

と設定されねばならない。本研究では，そうしなくて，式（3．11）で示される適応残差（adaptive residuals）z_j の，$j = 1 \sim n(m)$ にわたる組を考慮して，各残差 z_j の生起確率の積

$$F(W) = \Pi_{j=1}^{n(m)} f_j^m(z_j)$$

の対数

$$\log F(W) = \Sigma_{j=1}^{n(m)} \log f_j^m(z_j)$$

を尤度関数と考え，この対数尤度の最大を達成するような逐次学習法を提案していることに留意しておかねばならないだろう。また，

2乗誤差の the local minima の一つに達してしまい，the global minimum に決して達することのない

なる如き事態が生じない層構成法[14]については研究しなかったが，the model fitting problem を the general robust regression problem

とみて，逐次学習アルゴリズムを研究した本論文によって，見通しの良い形でこれ以上ない一般

化がもたらされ，しかも誤差逆伝播法に関連したニューラルネット逐次学習法は統一されたといえよう。

maximum-likelihood-type-estimator

のクラスは robust procedure を提供することは良く知られており，勿論，本研究での最尤型誤差逆伝播学習アルゴリズム MLBPLA はこのクラスに属するものであり，冒頭で説明し一例を提示したように，誤差分布として適切なものが採用されれば，

highly robust learning

が本ニューラルネットによってなされることは大いに期待される。

文　献

(1) 鈴木昇一：認識工学（上），柏書房，1975

(2) 鈴木昇一：半順序と情報処理，情報研究（文教大学情報学部），Vol. 12，pp. 121—174，1991—12

(3) 鈴木昇一：分析的／全体的処理と STOCHASTIC NEURO-COMPUTER(1)誤差逆伝播モデル，電子通信学会技術研究報告〔ニューロコンピューティング〕，Vol. 90，No. 483，NC 90—68，pp. 1—6，1991—03

(4) 鈴木昇一：パターン認識における構造化モデルの4性質とその応用，電子通信学会論文誌，Vol. 60—D，No. 9，pp. 710—717，1977—09

(5) 鈴木昇一：線形空間回路網のパーセプトロン形構造変化による情報パターン集合の2分割法，電子通信学会オートマトン研究会資料，A 71—80，1971—12

(6) 鈴木昇一，飛沢兼夫，五十嵐彰一，安藤孝男：生体視覚系観測機構と空間パーセプトロンによる手書きひらがな文字の識別実験，電子通信学会医用電子生体工学研究会資料，MBE 72—11（1972—07）

(7) 鈴木昇一，磯谷修平：位相不変量子空間パーセプトロンの帰納的類別能力と手書き文字に対するその計算機シミュレーション，電子通信学会パターン認識と学習研究会資料，PRL 73—15，1973—05

(8) 鈴木昇一：位相不変連続空間パーセプトロン SPAPの手書き漢字に対する観測不変帰納類別計算機シミュレーション，電子通信学会パターン認識と学習研究会資料，PRL 73—43，1973—07

(9) 鈴木昇一：連想形記憶器 MEMOTRON と日本語母音系列の再生に関する計算機シュミレーション，情報研究（文教大学情報学部），Vol. 7，pp. 14—29（1986—12）

(10) 鈴木昇一：多変量解析に基づく大分類関数の決定とその計算機シミュレーション，情報研究（文教大学情報学部），Vol.10，pp.35—49，1989—12

(11) 鈴木昇一：パターン認識の数学的理論，

第Ⅰ部（考え方，PRL 84—6，pp. 1—10，1984—05），

第Ⅱ部（認識抽象と公理系，定理系，PRL 84—30，pp. 65—74，1984—09），

第Ⅲ部（認識抽象と不動点諸定理，PRL 84—38，pp. 65—73，1984—09），

第Ⅳ部（パターンの素領域，PRL 85—27，pp. 1—10，1985—09），

第Ⅴ部（認識停止と認識同値，PRU 86—8，pp. 65—74，1986—05），

第Ⅵ部（類似度関数の三構成法，PRU 86—35，pp. 51—60，1986—07），

第Ⅶ部（類似度関数の実現と解析，PRU 86—69，pp. 1—8，1986—12），

第Ⅷ部（大分類関数の自己組織化，PRU 87—1，pp. 1—8，1987—05），

第Ⅸ部（帰属関数あいまい度と認識情報量，PRU 87—28，pp. 1—10，1987—07），

第Ⅹ部（mixture 条件の研究，PRU 88—30，pp. 1—8，1988—07），

第Ⅺ部（認識プログラム FERT の近似の鎖，PRU 89—1，pp. 1—8，1989—05），

第XII部（ボテンシャル関数による認識過程の評価，PRU 89—27，pp. 1—8，1989—07），
第XIII部（認識プログラム $FERT_D$ の不動点認識定理，PRU 89—40，pp. 1—8，1989—09），
第XIV部（線形帰属係数法と諸基本定理，PRU 89—66，pp. 1—8，1989—11），
第XV部（パターンの構造的類似性をもたらす4種類の収縮写像，PRU 89—77，pp. 1—8，1989—12），
第XVI部（コネクショニスト・モデルと収縮写像，PRU 89—136，pp. 9—16，1990—03），
第XVII部（ホップフィールドネットワーク2値モデルと収縮写像(1)，PRU 90—5，pp. 1—8，1990—05），
第XVIII部（ホップフィールドネットワーク2値モデルと収縮写像(2)，PRU 90—15，pp. 1—8，1990—06），
第XIX部（ホップフィールドネットワークの連続モデルと2種類と収縮写像(1)，PRU 90—29，pp. 9—16，1990—07），
第XX部（ホップフィールドネットワークの連続モデルと2種類と収縮写像(2)，PRU 90—125，pp. 1—8，1991—02），
第XXI部（誤差逆伝播ニューラルネットモデルと特徴抽出(1)，PRU 91—1，pp. 1—8，1991—05），
第XXII部（誤差逆伝播ニューラルネットモデルと特徴抽出(2)，PRU 91—29，pp. 23—28，1991—06），
第XXIII部（誤差逆伝播ニューラルネットモデルと特徴抽出(3)，PRU 91—42，pp. 1—8，1991—07），
第XXIV部（再帰領域方程式と標本化，PRU 92—1，pp. 1—8，1992—05），
第25部（画像前処理，PRU 92—18，pp. 1—8，1992—06），
第26部（線形歪を持った多次元パターンの，モーメントによる正規化，PRU 92—25，pp. 1—8，1992—09）
第27部（モデル構成作用素による Extended Dynamic Axes Warping(1)，PRU 92—89，pp. 1—8，1992—12），
電子（情報）通信学会技術研究報告[パターン認識と学習，パターン認識と理解]

(12) 松本元・大津展之共編（大津・上坂・乾・村岡・古谷・星野執筆）：ニューロコンピューティングの周辺（脳とコンピュータ2），培風館，第2章（学習機械の理論，上坂，pp. 43—82），1991—07

(13) 甘利・中野・上坂・倉田・川人：ニューロンコンピューティングの基礎理論，㈱日本工業技術振興協会ニューロコンピュータ研究部会，海文堂出版株式会社，1990—12

(14) Marco Gori and Alberto Tesi ： On the Problem of Local Minima in Backpropagation, IEEE TRANS. ON PAMI, Vol. 14, No. 1, pp. 76—86, 1992–01

(15) 佐波隆光： 回帰分析，朝倉書店，1979—04

(16) Hopfield J. J. ： Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci, 79, pp. 2554–2558, 1982

(17) DAVID H. ACKLEY, GEOFFREY E. HINTON, TERRENCE J. SEJNOWSKI ： A Learning Algorithm for Boltzmann Machines, Cognitive Science,Vol. 9, pp. 147–169, 1985

(18) F. Rosenblatt ： Principles of Neurodynamics, Washington, D. C., Spartan Books, 1961

(19) 鈴木昇一：Rosenfeld 型の確率的弛緩ラベリング法の基本的諸性質，情報研究（文教大学情報学部），Vol. 11, pp. 163—181，1990—12

(20) 鈴木昇一：連想形記憶器内荷重関数の最小自乗法，自己組織化法による決定，情報研究（文教大学情報学部），Vol. 5，pp. 16—28，1984—12

(21) 鈴木昇一：測度的不変量検出形認識系の構成理論，電子通信学会論文誌，Vol. 55—D， No. 8，pp. 531—538，1972—08

(22) Xinhua Zhuang, Tao Wang, and Peng Zhang ： A Highly Robust Estimator through Partially Likelihood Function Modeling and Its Application in Computer Vision, IEEE TRANS ON PAMI, Vol. 14, No. 1, pp. 19–35, 1992–01

(23) 小池義昌，田辺雅秋：テンプレート補間および重み付き学習型ニューラルネットワークを用いた濃淡画像の照合法，電子情報通信学会論文誌 D—II，Vol. J 75—D—ＩＩ，No. 7，pp. 1151—1159，1992—07
(24) M. Minsky, S. Papert：パーセプトロン，斎藤正男訳，東京大学出版会，1971—08
(25) 西山清，後藤治英：冗長なニューロンをもつ Hopfield ニューラルネットワークに基づく連想記憶モデル，電子情報通信学会論文誌 D—ＩＩ，Vol. J 75—D—ＩＩ，No. 7，pp. 1241—1250，1992—07
(26) 河田敬義，丸山文行：数理統計，裳華房，1963—08
(27) 鈴木昇一：視聴覚空間神経系のモデルと連想記憶能力に就て，電子通信学会医用電子・生体工学研究会，MBE 70—34，1971—01

（鈴木昇一，誤差確率分布を考慮した誤差逆伝播学習，情報研究 No. 13投稿論文，1992年 9 月24日）