社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ランク縮小写像のベイズ汎化誤差の解析

渡辺　一帆†　　渡辺　澄夫††

† 東京工業大学工学部情報工学科
†† 東京工業大学精密工学研究所
〒 226-8503 横浜市緑区長津田 4259

E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

あらまし　ランク縮小写像は，線形写像の推定において，写像のランクが一定値以下のものから最適なものを見出そうとする学習モデルであり，高次元入力から高次元出力への推論が実質的に低次元の自由度を持つと考えられる場合などに利用されている．また 3 層パーセプトロンにおいて中間ユニットが線形応答を持つ場合に相当するため，階層型の神経回路網の性質を数理的に考察するためのモデルとして研究されている．一方，ランク縮小写像は，多くの神経回路網や混合正規分布と同じく，その動作からパラメータを特定することができない「特定不能な学習モデル」であり，統計モデルとしては最尤推定量が漸近正規性を持つための正則条件を満たさないため，その学習の性質は，いまだに十分には解明されていない．本論文では，ランク縮小写像のベイズ学習を考察し，代数幾何学的な方法に基づいて確率的複雑さの上限を求め，ベイズ汎化誤差の上限を考察する．

キーワード　ランク縮小写像，ベイズ推測，汎化誤差，代数幾何，特異点

Bayes Generalization Errors of Reduced Rank Approximation

Kazuho WATANABE† and Sumio WATANABE††

† Department of Computer Science, Tokyo Institute of Technology
†† P&I Lab., Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku, Yokohama, 226-8503 Japan

E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

Abstract　Reduced rank approximation corresponds to a three-layer linear neural network with reduced hidden layer size. A lot of attention is paid to the properties of it and some basic questions can sometimes be answered analytically for its linearity. However, as is the case with other layered models such as neural networks and gaussian mixtures, its true parameter is not identifiable. Therefore some problems related to learning remain unsolved. In this paper, we analyze the generalization error of the reduced rank approximation in Bayesian estimation, and figure out its upper bound using an algebraic geometrical method.

Key words　Reduced rank approximation, Bayesian estimation, generalization error, algebraic geometry, singularities

1. はじめに

神経回路網や混合正規分布などの学習モデルが，パターン認識，時系列予測，システム制御などさまざまな実問題に利用され，その有用性について実験的に確認されつつある．しかしながら，これらの階層構造を持つ学習モデルにおいては，パラメータと学習モデルとの対応が 1 対 1 ではなく，学習モデルのふるまいからパラメータが一意的には定まらない．このため，サンプル数が十分に大きい場合でも統計的正則モデルの統計的漸近理論を適用することができず，学習モデルの性質を解明することや最適な設計を行うための数学的基盤が未だに十分には確立されていない．

この問題は，混合正規分布の検定において尤度比が通常のオーダーにならないことや（Hartigan, 1985)，神経回路網のモデル選択においては正則モデルの情報量規準は適用できないこと (萩原,1993) ，ベイズ事後分布が正規分布とは異なるものに近づくこと（渡辺,1997) などによって指摘され，神経回路網を含む非常に広い学習モデルにおいて一般的に生じている問題であることが認識されるようになった．

一般に，学習モデルの学習を行うためには，大別して次の二つの方法がある．ひとつは，対数尤度やその補正などで定義される損失関数を最小にする推定量を用いる方法である．この方法を特定不能な学習モデルに適用した場合の予測精度を解析するには，パラメータ空間を錘型に変形し，ガウス確率場の最大値の問題に帰着する方法が有効であることが，統計学ではよく知られている．(Dacunha − Castelle,1997; 栗木, 2001)．もうひとつは，パラメータ空間上に事前確率密度を用意して，事後分布を構成し，事後分布による平均を用いて予測を行う方法である．後者の方法はベイズ法と呼ばれ，その予測精度を解析する場合には，パラメータ空間の特異点を解消し，カルバック情報量のゼータ関数の極を求める方法が有効である (Watanabe, 1999)．

以上の二つの方法は学習システムを構成する場合の代表的な方法であり，応用を行う目的に応じて適するものが利用されている．予測精度を向上させるという目的に限定すれば，ベイズ推測に代表されるアンサンブル学習の方が，一個の推定量を用いる学習よりも優れた学習が実現できることが理論的にも実験的にも知られている．

本論文では，特定不能な学習モデルの代表例としてランク縮小写像を考察し，そのベイズ予測誤差を解析する．ランク縮小写像は，3 層パーセプトロンにおいて中間ユニットの応答が線形である場合に相当し，理論的にも応用上も様々な研究がなされている．このモデルでは，入力から出力への写像は線形写像であるが，対数尤度は，パラメータについては 4 次の多項式であり，原点に複雑な特異点を持つために，その学習は線形理論の枠組みでは取り扱うことができない．実際，最尤推定量を用いて予測を行うと，予測精度は（比例定数/学習例数）となるが，ここで比例定数は「パラメータ数/2」よりも大きな値になる (Fukumizu, 1999)．

本論文では，ランク縮小写像においてベイズ推測を適用する場合を考え，ブローアップによって特異点を変換することにより，その推定精度の上限を求め，比例定数が「パラメータ数/2」よりも遥かに小さな値であることを数学的に証明する．

本論文の構成は次の通りである．第 2 章では，ランク縮小写像とベイズ学習に関する一般的な説明を行う．第 3 章では，本論文の主定理を述べる．その証明は第 4 章で行う．第 5 章と第 6 章では，得られた結果に関する考察を述べ，結論を述べる．

2. ランク縮小写像とベイズ推測

ランク縮小写像とは，入力から出力への線形写像の中からもっとも相応しいランクの写像を学習によって見出すものである．$H \times M$ 行列 A，$N \times H$ 行列 B を，

$$A=\begin{pmatrix} a_{11} & \cdots\cdots & a_{1M} \\ \vdots & \ddots\ddots & \vdots \\ a_{H1} & \cdots\cdots & a_{HM} \end{pmatrix}, B=\begin{pmatrix} b_{11} & \cdots & b_{1H} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NH} \end{pmatrix} \quad (1)$$

とすると，入力 $\mathbf{x} \in R^M$ が与えられたときのランク縮小写像の出力 $\mathbf{y} \in R^N$ の分布は，

$$\mathbf{y} = BA\mathbf{x} + (\text{雑音})$$

と表される．特に雑音として，平均 0，分散共分散行列が $\sigma^2 I$(I は単位行列) の正規分布を考えると，入力 $\mathbf{x}$ が与えられたときの，出力 $\mathbf{y}$ の分布 $p(\mathbf{y}|\mathbf{x}, A, B)$ は，

$$p(\mathbf{y}|\mathbf{x}, A, B)=\frac{1}{\sqrt{2\pi\sigma^2}^N}\exp(-\frac{1}{2\sigma^2}\|\mathbf{y}-BA\mathbf{x}\|^2) \quad (2)$$

である．真のモデルが中間ユニット数 H_0 のモデルで表されると仮定する．行列 A_0,B_0 を，それぞれ $H_0 \times M$,$N \times H_0$ 行列で，

$$A_0=\begin{pmatrix} a^*_{11} & \cdots\cdots & a^*_{1M} \\ \vdots & \ddots\ddots & \vdots \\ a^*_{H_0 1} & \cdots\cdots & a^*_{H_0 M} \end{pmatrix}, B_0=\begin{pmatrix} b^*_{11} & \cdots & b^*_{1H_0} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ b^*_{N1} & \cdots & b^*_{NH_0} \end{pmatrix}$$

とすると，真の分布は，

$$q(\mathbf{y}|\mathbf{x})=\frac{1}{\sqrt{2\pi\sigma^2}^N}\exp(-\frac{1}{2\sigma^2}\|\mathbf{y}-B_0A_0\mathbf{x}\|^2) \quad (3)$$

となる．入力 $\mathbf{x}$ が密度関数 $q(x)$ を持つと仮定し，$q(\mathbf{x})q(\mathbf{y}|\mathbf{x})$ から，独立なサンプルが n 個得られたとする．パラメータ空間に事前分布を設定し，事前分布は原点の近傍で 0 にならないものとする．n 個の例と事前分布から作られるベイズ予測分布を $p_n(\mathbf{y}|\mathbf{x})$ とする．ベイズ汎化誤差 $G(n)$ は，真の分布からベイズ予測分布までのカルバック情報量をサンプルの出方について平均することによって定義される．

$$G(n) = E\Big[\int\int q(\mathbf{x})q(\mathbf{y}|\mathbf{x})\log\frac{q(\mathbf{y}|\mathbf{x})}{p_n(\mathbf{y}|\mathbf{x})}d\mathbf{x}d\mathbf{y}\Big]. \quad (4)$$

学習理論の課題のひとつは，汎化誤差を解明し，その結果に基づいて，より小さな汎化誤差を持つ学習法を構成することである．ランク縮小写像のように特異点を持つ学習モデルのベイズ汎化誤差については，最近，次のような解決が得られた (Watanabe,1999;Watanabe,2001)．

真の分布からパラメータ (A, B) を持つ学習モデルまでのカルバック情報量 $K(A, B)$ は

$$K(A, B)=\int\int q(\mathbf{x})q(\mathbf{y}|\mathbf{x})\log\frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x}, A, B)}d\mathbf{x}d\mathbf{y}$$

である．カルバック情報量と事前分布 $\varphi(A, B)$ により定まるゼータ関数を

$$J(z) = \int K(A, B)^z\varphi(A, B)dAdB$$

と定義すると，この関数は $Re(z) > 0$ では（複素関数として）正則であるが，複素平面全体まで有理型関数として解析接続することができ，その極はすべて負の実軸上にある有理数である．その極の中で，最も原点に近い極を $-\lambda$ とし，その位数を m とすると，汎化誤差 $G(n)$ が漸近展開できるならば

$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n\log n} + o(\frac{1}{n\log n})$$

が成り立つ．ここで，ゼータ関数の原点に一番近い極は，代数多様体 $\{(A, B); K(A, B) = 0\}$ の特異点を解消することにより見出すことができる．任意の代数多様体はブローアップの有限回の繰り返しにより，正規交差特異点だけを持つものに変換できることが知られている（広中の定理）が，特異点を解消する写像を具体的に構成することは，一般には，それほど容易ではない．しかしながら，ブローアップによってゼータ関数の極を見出すことは比較的容易であり，得られた極から，汎化誤差の上限が得られる．

本論文では，以上の方法に基づいて，ブローアップを適用することにより，ランク縮小写像のベイズ汎化誤差の上限値を求め，その値が正則な統計モデルよりも遥かに小さいことを証明する．

3. 主 定 理

[主定理 1] ランク縮小写像のベイズ汎化誤差の係数 λ は次の不等式を満たす.

$$\lambda \leqq \begin{cases} \frac{1}{4}\{(2M-1)H+(2M-2H_0+1)H_0\} \\ \qquad (M=N \text{ かつ } H-H_0 \text{ が偶数のとき}) \\ \frac{1}{4}\{(2M-1)H+(2M-2H_0+1)H_0+1\} \\ \qquad (M=N \text{ かつ } H-H_0 \text{ が奇数のとき}) \\ \frac{1}{2}\{\min(M,N)H+\max(M,N)H_0-H_0^2\} \\ \qquad (\text{それ以外のとき}) \end{cases}$$

4. 証 明

カルバック情報量 $K(A,B)$ は簡単な計算により

$$K(A,B)=\int \|B_0A_0\mathbf{x}-BA\mathbf{x}\|^2\frac{q(\mathbf{x})}{2\sigma^2}d\mathbf{x} \tag{5}$$

と一致する. ここで

$$\begin{aligned} &\|B_0A_0\mathbf{x}-BA\mathbf{x}\|^2 \\ &\quad=\sum_{i=1}^{N}\{\sum_{j=1}^{H}\sum_{k=1}^{M}b_{ij}a_{jk}x_k-\sum_{j=1}^{H_0}\sum_{k=1}^{M}b_{ij}^*a_{jk}^*x_k\}^2 \\ &\quad\leqq 2\sum_{i=1}^{N}\{\sum_{j=1}^{H_0}\sum_{k=1}^{M}(b_{ij}a_{jk}-b_{ij}^*a_{jk}^*)x_k\}^2 \\ &\qquad+2\sum_{i=1}^{N}\{\sum_{j=H_0+1}^{H}\sum_{k=1}^{M}b_{ij}a_{jk}x_k\}^2 \end{aligned} \tag{6}$$

より,

$$K_T(A,B)=\int\sum_{i=1}^{N}\{\sum_{j=1}^{H_0}\sum_{k=1}^{M}(b_{ij}a_{jk}-b_{ij}^*a_{jk}^*)x_k\}^2\frac{q(\mathbf{x})}{2\sigma^2}d\mathbf{x} \tag{7}$$

$$K_Z(A,B)=\int\sum_{i=1}^{N}\{\sum_{j=H_0+1}^{H}\sum_{k=1}^{M}b_{ij}a_{jk}x_k\}^2\frac{q(\mathbf{x})}{2\sigma^2}d\mathbf{x} \tag{8}$$

とすると, 式 (5), (6), (7), (8) より,

$$K(A,B)\leqq 2K_T(A,B)+2K_Z(A,B) \tag{9}$$

である. $K_T(A,B)$ と $K_Z(A,B)$ とは共通の変数を含まないので, $K(A,B), K_T(A,B), K_Z(A,B)$ のそれぞれから定義されるゼータ関数の極のうち最も原点に近いものをそれぞれ, $-\lambda, -\lambda_T, -\lambda_Z$ とすると,

$$\lambda \leqq \lambda_T+\lambda_Z \tag{10}$$

が成り立つので, λ_T, λ_Z について考える.

まず, λ_T について考える. 式 (7) より, $K_T(A,B)$ は, 学習モデルが中間ユニットの数 H_0 個で, 真のモデルと同じモデルで学習を行った場合のカルバック情報量である. このとき, $K_T(A,B)=0$ を満たすパラメータ A,B の集合は, NH_0+MH_0 次元のパラメータ空間の中で H_0^2 次元の部分多様体をなす(注 1)ので,

$$\lambda_T=\frac{1}{2}(N+M-H_0)H_0 \tag{11}$$

である.

次に λ_Z を考える. 式 (8) より, $K_Z(A,B)$ は真のモデルが 0 のとき, 中間ユニット $H-H_0$ 個の学習モデルで学習を行った場合のカルバック情報量である. そこでまず一般的に, 真のモデルが 0 のとき, 中間ユニットの数 H の学習モデルで学習を行ったときのカルバック情報量 $K_0(A,B)$ のゼータ関数の極を求める. カルバック情報量 $K_0(A,B)$ は不等式

$$\begin{aligned} K_0(A,B) &= \frac{1}{\sigma^2}\int\sum_{i=1}^{N}\{\sum_{j=1}^{H}\sum_{k=1}^{M}b_{ij}a_{jk}x_k\}^2q(\mathbf{x})d\mathbf{x} \\ &= \frac{1}{\sigma^2}\int\sum_{i=1}^{N}\{\sum_{k=1}^{M}x_k\sum_{j=1}^{H}b_{ij}a_{jk}\}^2q(\mathbf{x})d\mathbf{x} \\ &\leqq \frac{M}{\sigma^2}\int\sum_{i=1}^{N}\sum_{k=1}^{M}x_k^2\{\sum_{j=1}^{H}b_{ij}a_{jk}\}^2q(\mathbf{x})d\mathbf{x} \\ &\leqq \frac{\tilde{M}}{\sigma^2}\sum_{i=1}^{N}\sum_{k=1}^{M}\{\sum_{j=1}^{H}b_{ij}a_{jk}\}^2 \end{aligned} \tag{12}$$

を満たす(注 2). 事前分布は原点の近傍で 0 にならないと仮定しているので, ゼータ関数の定義

$$J(z)=\int\{\sum_{i=1}^{N}\sum_{k=1}^{M}\{\sum_{j=1}^{H}b_{ij}a_{jk}\}^2\}^z\varphi(A,B)dAdB$$

において, $\varphi(A,B)$ はすべてのパラメータが区間 $[-1,1]$ にあるときだけ値をもつ一様分布としても一般性を失わない.

$$a_{11}=a, a_{12}=aa'_{12}, a_{13}=aa'_{13}, \cdots, a_{HM}=aa'_{HM}$$

(注 1) : $BA=BG^{-1}GA$ のように $(H_0\times H_0)$ の行列 $G(\det G\neq 0)$ を作用させても $K_T(A,B)$ は変わらないため $\{A,B : K_T(A,B)=0\}$ は H_0^2 次元の部分多様体をなす.

(注 2) : ここで $\tilde{M}=M\times\max_{1\leqq k\leqq M}\{\int x_k^2q(\mathbf{x})d\mathbf{x}\}$ は存在するとしている.

$b_{11}=b, b_{12}=bb'_{12}, b_{13}=bb'_{13}, \cdots, b_{NH}=bb'_{NH}$

と変数変換して，定数倍を省略すると，$J(z)$ は，

$$\int a^{2z}b^{2z}\{K'_0(A',B')\}^z a^{MH-1}b^{NH-1}dadbdA'dB'$$

$$=\frac{1}{2z+MH}\frac{1}{2z+NH}\int\{K'_0(A',B')\}^z dA'dB'$$

という項を含むことがわかる．ここで，

$$K'_0(A',B')=\sum_{i=1}^{N}\sum_{k=1}^{M}\{\sum_{j=1}^{H}b'_{ij}a'_{jk}\}^2$$

(ただし $a'_{11}=b'_{11}=1$) である．よって，この $J(z)$ は $z=-\frac{MH}{2}, -\frac{NH}{2}$ を極として持つ．このことから，$K_0(A,B)$ のゼータ関数の極のうち最も原点に近いものを $-\lambda_0$ とすると，

$$\lambda_0 \leqq \min(M,N)\frac{H}{2} \tag{13}$$

が成り立つ．

また，H 以下の自然数 h に対して，

$$K_h(A,B)=\sum_{i=1}^{N}\sum_{k=1}^{M}\{\sum_{j=1}^{h}b_{ij}a_{jk}\}^2$$

とすると，(12) 式より，

$$K_0(A,B)\leqq\frac{\tilde{M}}{\sigma^2}K_H(A,B)$$

であり，

$$\begin{aligned}K_H(A,B)&=\sum_{i=1}^{N}\sum_{k=1}^{M}\{\sum_{j=1}^{H}b_{ij}a_{jk}\}^2\\&\leqq 2\sum_{i=1}^{N}\sum_{k=1}^{M}(b_{i,H-1}a_{H-1,k}+b_{iH}a_{Hk})^2\\&\quad+2\sum_{i=1}^{N}\sum_{k=1}^{M}(\sum_{j=1}^{H-2}b_{ij}a_{jk})^2\\&=2K_2(A,B)+2K_{H-2}(A,B)\end{aligned}\tag{14}$$

が $H\geqq 3$ のとき，成り立つ．これより，$1\leqq h\leqq H$ に対して，$K_h(A,B)$ のゼータ関数の最も原点に近い極を $-\lambda_h$ とすると，

$$\lambda_0\leqq\lambda_H\leqq\lambda_2+\lambda_{H-2}$$

が成立するので，くりかえして用いることにより

$$\lambda_0\leqq\lambda_H\leqq\begin{cases}\frac{H}{2}\lambda_2 & (H\text{ が偶数のとき})\\ \frac{H-1}{2}\lambda_2+\lambda_1 & (H\text{ が奇数のとき})\end{cases}\tag{15}$$

が得られる．まず λ_2 について考える．$K_2(A,B)$ のゼータ関数

$$J_2(z)=\int\{\sum_{i=1}^{N}\sum_{k=1}^{M}(b_{i1}a_{1k}+b_{i2}a_{2k})^2\}^z\varphi(A,B)dAdB$$

において，$b_{i1}=b_{i2}b'_{i1}$，$a_{2k}=a_{1k}a'_{2k}$ の変数変換を考え，

$$K'_2(A',B')=\sum_{i=1}^{N}\sum_{k=1}^{M}b_{i2}^2a_{1k}^2(b'_{i1}+a'_{2k})^2\tag{16}$$

とすると，$J_2(z)$ は，

$$\int K'_2(A',B')^z b_{12}\cdots b_{N2}a_{11}\cdots a_{1M}dA'dB'\tag{17}$$

という項を持つ．ここで，

$$b'_{i1}+a'_{21}=p_i\quad(1\leqq i\leqq N)$$

$$b'_{11}+a'_{2k}=q_k\quad(2\leqq k\leqq M)$$

とおくと，

$$\begin{aligned}K'_2(A',B')&=\sum_{i=1}^{N}a_{11}^2b_{i2}^2p_i^2+\sum_{k=2}^{M}a_{1k}^2b_{12}^2q_k^2\\&\quad+\sum_{i=2}^{N}\sum_{k=2}^{M}a_{1k}^2b_{i2}^2(p_i+q_k-p_1)^2\end{aligned}$$

となる．ここで，

$$p_1=p, p_i=pp'_i\ (2\leqq i\leqq N)$$

$$q_k=pq'_k\ (2\leqq k\leqq M)$$

の変数変換を考えると，

$$\begin{aligned}K''_2(A'',B'')&=\sum_{i=1}^{N}a_{11}^2b_{i2}^2p_i'^2+\sum_{k=2}^{M}a_{1k}^2b_{12}^2q_k'^2\\&\quad+\sum_{i=2}^{N}\sum_{k=2}^{M}a_{1k}^2b_{i2}^2(p'_i+q'_k-p'_1)^2\end{aligned}\tag{18}$$

$(p'_1=1)$ としたとき，(17) は，

$$\int p^{2z+M+N-2}\{K''_2(A'',B'')\}^z dpdA''dB''$$

$$=\frac{1}{2z+M+N-1}\int\{K''_2(A'',B'')\}^z dA''dB''$$

という項を含む．よって，

$$\lambda_2\leqq\frac{M+N-1}{2}\tag{19}$$

である． λ_1 は (13) 式において $H=1$ の場合を考えて，

$$\lambda_1 \leqq \frac{\min(M,N)}{2} \tag{20}$$

(15)，(19)，(20) より，

$$\lambda_0 \leqq \lambda_H \leqq \begin{cases} \frac{M+N-1}{2}\frac{H}{2} \\ \quad (H \text{ が偶数のとき}) \\ \frac{M+N-1}{2}\frac{H-1}{2}+\frac{\min(M,N)}{2} \\ \quad (H \text{ が奇数のとき}) \end{cases} \tag{21}$$

が成り立つ．

$$\begin{cases} \frac{M+N-1}{2} < \min(M,N) & (M=N \text{ のとき}) \\ \frac{M+N-1}{2} \geqq \min(M,N) & (M \neq N \text{ のとき}) \end{cases}$$

であることを考えて，λ_Z は (13)，(21) 式の H を $H-H_0$ にすることで，

$$\lambda_Z \leqq \begin{cases} \frac{2M-1}{2}\frac{H-H_0}{2} \\ \quad (M{=}N \text{ かつ } H{-}H_0 \text{ が偶数のとき}) \\ \frac{2M-1}{2}\frac{H-H_0}{2}+\frac{1}{4} \\ \quad (M{=}N \text{ かつ } H{-}H_0 \text{ が奇数のとき}) \\ \min(M,N)\frac{H-H_0}{2} \\ \quad (\text{それ以外のとき}) \end{cases} \tag{22}$$

を満たすことがわかる．

式 (10)，(11)，(22) より，

$$\lambda \leqq \begin{cases} \frac{1}{4}\{(2M-1)H+(2M-2H_0+1)H_0\} \\ \quad (M{=}N \text{ かつ } H{-}H_0 \text{ が偶数のとき}) \\ \frac{1}{4}\{(2M-1)H+(2M-2H_0+1)H_0+1\} \\ \quad (M{=}N \text{ かつ } H{-}H_0 \text{ が奇数のとき}) \\ \frac{1}{2}\{\min(M,N)H+\max(M,N)H_0-H_0^2\} \\ \quad (\text{それ以外のとき}) \end{cases} \tag{23}$$

が成り立つ．

5. 実　　験

ここでは，真のモデルが 0 のときのベイズ汎化誤差を数値実験により計算する．

5.1 実験方法

パラメータの事前分布を $\varphi(A,B)$ とし，独立な n 個の入力 $\mathbf{x} \in R^M$ と出力 $\mathbf{y} \in R^N$ の組 $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$ を学習データとして用い学習を行うとき，学習誤差 $E(A,B)$ を

$$E(A,B) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|\mathbf{y}_i - BA\mathbf{x}_i\|^2 - \log\varphi(A,B) \tag{24}$$

とする．このとき，ベイズ事後分布 $p_n(A,B)$ と，予測分布 $p_n(\mathbf{y}|\mathbf{x})$ は，

$$p_n(A,B) = \frac{1}{Z}\varphi(A,B)\prod_{i=1}^{n}p(\mathbf{y}_i|\mathbf{x}_i,A,B) = \frac{1}{Z'}\exp\{-E(A,B)\} \tag{25}$$

$$p_n(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x},A,B)p_n(A,B)dAdB \tag{26}$$

である (Z，Z' は正規化定数)．

$E(A,B)$ に関する最急降下法に，平均 0 分散 $2/\tau$ の正規分布に従う白色雑音 dZ_τ/dt を加えることで得られる確率微分方程式

$$\frac{dW}{dt} = -\frac{\partial E(A,B)}{\partial W} + \frac{dZ_\tau}{dt} \tag{27}$$

(確率変数 W は A,B 中の全てのパラメータ) は，定数 τ が十分小さいとき，$t\to\infty$ で，W の確率密度関数がある密度関数 $q(w)$ に収束するとすると，

$$q(w) = \frac{1}{Z}\exp\{-E(w)\}$$

となることが知られている．よって，今回の実験は (25) 式のベイズ事後分布 $p_n(A,B)$ を，学習誤差 $E(A,B)$ に関する最急降下法に正規雑音を加えることで実現する．すなわち，$W(k)$ で時刻 k におけるパラメータを表すとして，(27) 式を差分化した

$$W(k+1) = W(k) - \tau\frac{\partial E(A,B)}{\partial W(k)} + Z_\tau \tag{28}$$

によりパラメータの更新を行う．ここで，Z_τ は平均 0 分散 2τ の正規雑音である．そして，十分な繰り返しの後，一定間隔でパラメータを記憶していくことで，ベイズ事後分布に従うパラメータのセットを得ることができる．これらを $\{(A_k,B_k)\}_{k=1}^r$ で表す．これらを用いて，(26) 式のベイズ予測分布は，

$$p_n(\mathbf{y}|\mathbf{x}) \cong \frac{1}{r}\sum_{k=1}^{r}p(\mathbf{y}|\mathbf{x},A_k,B_k)$$

により構成できるので，s 個のテストサンプル $\{(\mathbf{x}_l,\mathbf{y}_l)\}_{l=1}^s$ について予測を行う．このとき (4) 式のベイズ汎化誤差を，

$$G(n) \cong E\Big[\frac{1}{s}\sum_{l=1}^{s}\{\log\frac{q(\mathbf{y}_l|\mathbf{x}_l,O,O)}{p_n(\mathbf{y_l}|\mathbf{x}_l)}\}\Big].$$

により計算する．

H	理論値	$\tau = 10^{-4}$	$\tau = 10^{-5}$
1	0.025	0.0232 ± 0.0030	0.0291 ± 0.0036
2	0.045	0.0459 ± 0.0040	0.0535 ± 0.0042
3	0.070	0.0674 ± 0.0053	0.0779 ± 0.0060
4	0.090	0.0833 ± 0.0056	0.0955 ± 0.0063
5	0.115	0.0975 ± 0.0060	0.1097 ± 0.0064

表1　真のモデルが0のときのランク縮小写像のベイズ汎化誤差

5.2 実験結果

真のモデルが0のとき，入力数5，出力数5のランク縮小写像で，中間ユニット数 $H = 1, 2, \cdots, 5$ の場合について，ベイズ汎化誤差を計算する．学習例数 $n = 100$ で (28) 式に従って1000000回学習を繰り返すときの，900000〜1000000回目の間を100等分し，その時点のパラメータを記憶し，100セットのパラメータを用意する．パラメータの事前分布 $\varphi(A, B)$ を

$$\varphi(A, B) \propto \exp(-\frac{\|A\|^2 + \|B\|^2}{2})$$

となるようにとる(注3)．このとき (28) 式における $\frac{\partial E(A,B)}{\partial W(k)}$ の項は，

$$\frac{\partial E(A, B)}{\partial A(k)} = A(k) - \frac{1}{\sigma^2}\sum_{i=1}^{100} B(k)^T(\mathbf{y}_i - B(k)A(k)\mathbf{x}_i)\mathbf{x}_i^T$$

$$\frac{\partial E(A, B)}{\partial B(k)} = B(k) - \frac{1}{\sigma^2}\sum_{i=1}^{100}(\mathbf{y}_i - B(k)A(k)\mathbf{x}_i)(A(k)\mathbf{x}_i)^T$$

である(注4)．単位時間 τ は $\tau = 10^{-4}, 10^{-5}$ の2通りについて計算した．用意したパラメータを用い，10000個のテストサンプルについて予測を行ったときのベイズ汎化誤差の計算結果を表1に示す．値は学習データの出方について100回の繰り返しの平均をとったときの，(平均値) ± 2(標準偏差)$/\sqrt{100}$ を示してある．この区間に真のベイズ汎化誤差の値が含まれる確率は約95% である．

$\tau = 10^{-4}$ のときは，ほぼ得られた上限の範囲内に入っていると言える．$\tau = 10^{-5}$ のときは，$\tau = 10^{-4}$ のときに比べ全ての H の場合で大きな値となっている．

(注3)：任意の行列 $A = [a_{ij}]$ に対して，$\|A\|^2 = \sum_{ij} a_{ij}^2$ とする．

(注4)：任意の関数 f，行列 $A = [a_{ij}]$ に対して，$\frac{\partial f}{\partial A} = [\frac{\partial f}{\partial a_{ij}}]$ とする．

6. 考　　察

入力数 M，中間ユニット数 H，出力数 N のランク縮小写像と同じパラメータ数を持つ正則モデルでベイズ推測を行った場合汎化誤差係数は「パラメータ数/2」$= (MH + NH)/2$ となるのに対し，得られた上限はこれよりも遥かに小さな値である．また，ランク縮小写像で最尤推定を行なったときの汎化誤差係数は，「パラメータ数/2」よりも大きくなることが知られている．これより，ランク縮小写像の学習にはベイズ法が有効であると言える．また，最急降下法に正規雑音を加えて事後確率を実現したとき，同じ繰り返しでも，単位時間 τ の値によってはベイズ事後確率が正しく構成されない場合があると考えられる．数値実験の結果，$\tau = 10^{-5}$ のときに，上限値を越えてしまっていたのはこのためであろう．事後確率が正しく構成されない理由として，繰り返し回数や取り出すパラメータの数が少ないということが考えられる．τ の値が小さくなるにつれて1回のパラメータの更新におけるパラメータの変化量は小さくなる．このために取り出されるパラメータセットがパラメータ空間内で偏った部分のみから取り出され，事後確率が正しく構成されなかったと考えられる．この方法で事後確率を正しく構成するには，パラメータ空間の広い範囲からパラメータが取り出されるように，τ の値に応じて，十分な繰り返しとパラメータ数が必要であると考えられる．ベイズ事後確率の構成には，MCMCを用いるなど，他にいくつかの方法が考えられるが，得られるパラメータの系列が事後確率に従っているかどうか判定することができないため，それらの方法との比較も困難である．これは今後の課題である．

7. 結　　論

代数幾何的な手法を用いて，ランク縮小写像のベイズ汎化誤差の上限を与え，証明を行った．その結果，ランク縮小写像の学習にはベイズ法が有効であるこ

とがわかった．今後の課題として，ベイズ事後確率を実現する手法を考える必要があり，そのためには，得られるパラメータの系列が事後確率に従っているかどうか判定する規準を与えることが必要である．

本研究は部分的に科学研究費補助金 12680370 の援助を受けた．

文　　献

[1] H. Akaike, “Likelihood and Bayes procedure,” *Bayesian Statistics*, (Bernald J.M. eds.) University Press, Valencia, Spain, pp.143-166, 1980.

[2] S. Amari, N.Fujita and S. Shinomoto, “Four Types of Learning Curves,” *Neural Computation*, Vol.4, No.4, pp.608-618,1992.

[3] Baldi,P.F., Hornik,K. (1995) Learning in linear neural networks: a survey. *IEEE Trans. on Neural Networks*,Vol.6,pp.837-858.

[4] Dacunha-Castelle, D., & Gassiat, E. (1997). Testing in locally conic models, and application to mixture models. *Probability and Statistics*, 1, 285-317.

[5] Fukumizu, K. (1999) Generalization error of linear neural networks in unidentifiable cases. *Lecture Notes on Computer Science*, 1720, 51-62, Springer.

[6] J.A.Hartigan, “A Failure of likelihood asymptotics for normal mixtures,” *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, Vol.2, 807-810, 1985.

[7] E.Levin, N. Tishby and S.A.Solla, “A statistical approaches to learning and generalization in layered neural networks. *Proc. of IEEE*, Vol.78, No.10, pp.1568-1674, 1990.

[8] D.J. Mackay, “Bayesian interpolation,” *Neural Computation*, Vol.4, No.2, pp.415-447, 1992.

[9] G.Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, Vol.6, No.2, 461-464, 1978.

[10] Watanabe, S. (1997) On the essential difference between neural networks and regular statistical models. *Proceedings of 2nd International Conference on Computational Intelligence and Neuroscience*, 2, 149-152.

[11] S.Watanabe, “Algebraic analysis for singular statistical estimation,” *Lecture Notes on Computer Science*, Vol.1720, pp.39-50, Springer,1999.

[12] S.Watanabe,“Algebraic analysis for nonidentifiable learning machines,” *Neural Computation*, Vol.13, No.4, pp.899-933,2001.

[13] S.Watanabe,“Algebraic information geometry for learning machines with singularitiess,” *Advances in Neural Information Processing Systems*, Vol.13, pp.329-336. 2001.

[14] S. Watanabe, ”Algebraic geometrical methods for hierarchical learning machines,” Neural Networks, Vol.14, No.8,pp.1049-1060, 2001.

[15] S.Watanabe and K.Fukumizu, “Probabilistic design of layered neural networks based on their unified framework,” *IEEE Trans. on Neural Networks*, Vol.6, No.3, pp.691-702,1995.

[16] K.Yamanishi, “A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. on Information Theory*, Vol.44, No.4, pp.1424-1439, 1998.

[17] K.Yamazaki, S.Watanabe,“ ,”A probabilistic algorithm to calculate the learning curves of hierarchical learning machines with singularities,” to appear in Trans. on IEICE, D-II, 2002.