Proceedings of the IASTED International Conference
COMPUTATIONAL INTELLIGENCE
November 20–22, 2006, San Francisco, CA, USA
ISBN Hardcopy: 0-88986-602-3 / CD: 0-88986-603-1

GENERALIZATION ERROR OF THREE LAYERED LEARNING MODEL IN BAYESIAN ESTIMATION

Miki Aoyagi
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku, R2-5
Yokohama, 226-8503, Japan
email: aoyagi.m.aa@m.titech.ac.jp

Sumio Watanabe
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku, R2-5
Yokohama, 226-8503, Japan
email: swatanab@pi.titech.ac.jp

ABSTRACT
In this paper, we obtain the asymptotic forms of the generalization errors for some three layered learning models in Bayesian estimation. The generalization error measures how precisely learning models can approximate true density functions which produce learning data. We use a recursive blowing up process for analyzing the Kullback function of the learning model. Then, we have the maximum pole of its zeta function which is defined by the integral of the Kullback function and an *a priori* probability density function. In [1, 2], it was proved that the maximum pole of the zeta function asymptotically gives the generalization error of the hierarchical learning model.

KEY WORDS
Generalization error, non-regular learning machines, Bayesian estimate, resolution of singularities, Kullback function, zeta function.

1 Introduction

Hierarchical learning models are such as a layered neural network, reduced rank regression, a normal mixture model and a Boltzmann machine. These are known as effective learning models for analyzing complicated data practically. Therefore the generalization errors (learning efficiency) of these models may be smaller than those of regular statistical models. However, their generalization errors cannot be analyzed by using classic theories of regular statistical models, for example, model selection methods AIC[3], TIC[4], HQ[5], NIC[6], BIC[7], MDL[8], since their Fisher matrix functions are singular. For most examples, only upper bounds of their generalization errors were calculated but not the exact values. These models are called non-regular.

There are usually considered to be direct and inverse problems. The direct problem would be to solve the generalization error with a known true density function. The inverse problem is to find proper learning models and learning algorithms to minimize the generalization error under the condition of an unknown true density function. The inverse problem is important for practical usage, but in order to tackle the inverse problem, first the direct problem has to be solved. So it is necessary and crucial to construct fundamental mathematical theories for solving the direct problem. Recently, it was proved that the maximum poles of the zeta functions for hierarchical learning models asymptotically give their generalization errors as follows [1, 2]. Let x be an input with a probability density function $q(x)$ and y an output. We assume that n training samples $\{x_i\}_{i=1}^n$ are randomly selected from $q(x)$ and that $\{y_i\}_{i=1}^n$ are obtained by the conditional probability density function $q(y|x)$. Let $(x^n, y^n) := \{(x_i, y_i)\}_{i=1}^n$. The aim of the learning system is to estimate the function $q(y|x)$ by using (x^n, y^n). Let us consider a learning model $p(y|x, w)$ which infers a probabilistic output y from a given input x, where w is a parameter. Fix an *a priori* probability density function $\psi(w)$ on the parameter set W. Then, the *a posteriori* probability density function $p(w|(x^n, y^n))$ is written by

$$p(w|(x^n, y^n)) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^{n} p(y_i|x_i, w),$$

where $Z_n = \int_W \psi(w)\prod_{i=1}^n p(y_i|x_i, w)\mathrm{d}w$. So the average inference $p(y|x, (x^n, y^n))$ of the Bayesian density function is given by

$$p(y|x, (x^n, y^n)) = \int p(y|x, w)p(w|(x^n, y^n))\mathrm{d}w.$$

Here we define a measure function between the true density function $q(y|x)$ and the predictive density function $p(y|x, (x^n, y^n))$:

$$K(q||p) = \int q(y|x)\log\frac{q(y|x)}{p(y|x, (x^n, y^n))}q(x)\mathrm{d}x\mathrm{d}y.$$

This function is always positive and satisfies $K(q||p) = 0$ if and only if $q(y|x) = p(y|x, (x^n, y^n))$. The expectation value of that function over training samples is called the generalization error. It clarifies how precisely $p(y|x, (x^n, y^n))$ can approximate $q(y|x)$. Assume that the true probability density function $q(y|x)$ is expressed by $q(y|x) = p(y|x, w^*)$, where w^* is constant. Let $G(n)$ be the generalization error :

$$G(n) = E_n\{\int p(y|x, w^*)\log\frac{p(y|x, w^*)}{p(y|x, (x^n, y^n))}q(x)\mathrm{d}x\mathrm{d}y\},$$

where $E_n\{\cdot\}$ is the expectation value over the set of n training samples. Define the zeta function $J(z)$ of a complex variable z for the learning model by

$$J(z) = \int K(w)^z\psi(w)\mathrm{d}w,$$

where $K(w)$ is the Kullback function:

$$K(w) = \int p(y|x,w^*) \log \frac{p(y|x,w^*)}{p(y|x,w)} q(x) \mathrm{d}x \mathrm{d}y.$$

Then, for the maximum pole $-\lambda$ of $J(z)$ and its order θ, we have

$$G(n) \cong \lambda/n - (\theta-1)/(n \log n) \text{ as } n \to \infty. \quad (1)$$

Therefore, our aim is to obtain λ and θ. The values λ and θ can be calculated by using a blowing up process.

In spite of those mathematical foundations, the main terms of most generalization errors are unknown by the following reasons. By Hironaka's Theorem [9], it is known that the desingularization of an arbitrary polynomial can be obtained by using a blowing up process. However the desingularization of any polynomial in general, although it is known as a finite process, is very difficult. Furthermore, most of the Kullback functions are degenerate (over $\mathbb{R}$) with respect to their Newton polyhedrons, singularities of the Kullback functions are not isolated, and the Kullback functions are not simple polynomials, i.e., they have parameters, for example, M, H and N of

$\sum_{n=1}^{H} \sum_{k=1}^{M} \sum_{j=1}^{N} (\sum_{i=1}^{H} a_{ki} b_{ij}^{2n-1})^2$,

whose learning model is one of three layered neural networks. Therefore, to obtain the desingularization of the Kullback functions is a new problem even in mathematics, since these singularities are very complicated and so most of them have not been investigated so far.

In this paper, we consider the zeta functions

$$\int \Psi = \int (\sum_{n=1}^{H} \sum_{k=1}^{M} \sum_{j=1}^{N} (\sum_{i=1}^{H} a_{ki} b_{ij}^{p(n-1)+1})^2)^z \prod_{k=1}^{M} \prod_{i=1}^{H'} \mathrm{d}a_{ki} \prod_{i=1}^{H'} \prod_{j=1}^{N} \mathrm{d}b_{ij}, \quad (2)$$

where M, N, H, H', p are natural numbers with $H' \leq H \leq 2H'$ and a_{ki}, b_{ij} for $i > H'$ are all constants. The three layered neural network with N input units, H' hidden units and M output units has the zeta function $\int \Psi$ with $p = 2$, if it is trained for estimating the true distribution with $H - H'$ hidden units. Also a normal mixture model has the zeta function $\int \Psi$ with $p = 1$, $M = 1$, $\sum_{j=1}^{H'} a_{1j} = 1$ and $\sum_{j=H'+1}^{H} a_{1j} = -1$ [10].

In this paper, we consider the case of $H' = 2$ and arbitrary M, N by using a recursive blowing up process.

We already have λ and θ if $M = N = 1$ and any H, H' in [11] and [12]. In the paper [13], we have clarified λ and θ of the reduced rank regression which is the three layered neural network with linear hidden units. This model is the case of

$$\int (\sum_{k=1}^{M} \sum_{j=1}^{N} (\sum_{i=1}^{H} a_{ki} b_{ij})^2)^z \prod_{k=1}^{M} \prod_{i=1}^{H'} \mathrm{d}a_{ki} \prod_{i=1}^{H'} \prod_{j=1}^{N} \mathrm{d}b_{ij}.$$

2 Resolution of singularities

In this section, we introduce Hironaka's Theorem [9] on the resolution of singularities and construction of blowing up. Blowing up is a main tool in the resolution of singularities of an algebraic variety.

Theorem[Hironaka [9]]

Let f be a real analytic function in a neighborhood of $w = (w_1, \cdots, w_d) \in \mathbf{R}^d$ with $f(w) = 0$. There exists an open set $V \ni w$, a real analytic manifold U and a proper analytic map μ from U to V such that
(1) $\mu : U - \mathcal{E} \to V - f^{-1}(0)$ is an isomorphism, where $\mathcal{E} = \mu^{-1}(f^{-1}(0))$,
(2) for each $u \in U$, there are local analytic coordinates $(u_1, \cdots, u_n)$ such that $f(\mu(u)) = \pm u_1^{s_1} u_2^{s_2} \cdots u_n^{s_n}$, where $s_1, \cdots, s_n$ are non-negative integers.

Next we explain blowing up along a manifold used in this paper. Define a manifold $\mathcal{M}$ by gluing k open sets $U_i \cong \mathbf{R}^d$, $i = 1, 2, \cdots, k (d \geq k)$ as follows.

Denote a coordinate of U_i by $(\xi_{1i}, \cdots, \xi_{di})$.

Define an equivalence relation

$(\xi_{1i}, \xi_{2i}, \cdots, \xi_{di}) \sim (\xi_{1j}, \xi_{2j}, \cdots, \xi_{dj})$

at $\xi_{ji} \neq 0$ and $\xi_{ij} \neq 0$, by $\xi_{ij} = 1/\xi_{ji}, \xi_{jj} = \xi_{ii}\xi_{ji}, \xi_{hj} = \xi_{hi}/\xi_{ji} (1 \leq h \leq k, h \neq i, j), \xi_{\ell j} = \xi_{\ell i} (k+1 \leq \ell \leq d)$, and set $\mathcal{M} = \coprod_{i=1}^{k} U_i / \sim$.

Also define $\pi : \mathcal{M} \to \mathbf{R}^d$ by

$$U_i \ni (\xi_{1i}, \cdots, \xi_{ni}); \mapsto (\xi_{ii}\xi_{1i}, \cdots, \xi_{ii}\xi_{i-1i}, \xi_{ii}, \xi_{ii}\xi_{i+1i}, \cdots, \xi_{ii}\xi_{ki}, \xi_{k+1i}, \cdots, \xi_{di}).$$

This map is well-defined and called blowing up along

$$X = \{(w_1, \cdots, w_k, w_{k+1}, \cdots, w_d) \in \mathbf{R}^d \mid w_1 = \cdots = w_k = 0\}.$$

The blowing map satisfies
(1) $\pi : \mathcal{M} \to \mathbf{R}^d$ is proper and
(2) $\pi : \mathcal{M} - \pi^{-1}(X) \to \mathbf{R}^d - X$ is isomorphic.

3 Main Results

Define the norm of a matrix $C = (c_{ij})$ by $||C|| = \sqrt{\sum_{i,j} |c_{ij}|^2}$.

For simplicity, we use the notation $\mathrm{d}a$ instead of $\prod_{i=1}^{M} \prod_{j=1}^{H} \mathrm{d}a_{ij}$ for $a = (a_{ij})$.

Main Theorem 1 *Let* $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ & \vdots \\ a_{M1} & a_{M2} \end{pmatrix}$, $B_j = \begin{pmatrix} b_{1j} & b_{1j}^{p+1} \\ b_{2j} & b_{2j}^{p+1} \end{pmatrix}$ *and* $B = (B_1, \cdots, B_N)$.

Set

$$\Psi = ||AB||^{2z} \mathrm{d}a \mathrm{d}b, \quad (3)$$

then the maximum pole $-\lambda$ of $\int_{||AB||<1} \Psi$ and its order θ are as follows.

(a)

(i) If $N \geq M+1$ *then* $\lambda = M$ *and* $\theta = 1$.

(ii) If $N = M$ *then* $\lambda = \frac{2N+p(M+N-1)}{2(p+1)}$ *and* $\theta = 1$.

(iii) If $N = M-1$ *then* $\lambda = N$ *and* $\theta = 2$.

(iv) If $N < M-1$ *then* $\lambda = N$ *and* $\theta = 1$.

(b) The case of $a_{k1} + a_{k2} = 1$.

(i) If $N > M+p$ *then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 1$.

(ii) If $N = M+p$ *then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 2$.

(iii) If $M-1 < N < M+p$ *then* $\lambda = \frac{2N+p(M+N-1)}{2(p+1)}$ *and* $\theta = 1$.

(iv) If $N = M-1$ *then* $\lambda = N$ *and* $\theta = 2$.

(v) If $N < M-1$ *then* $\lambda = N$ *and* $\theta = 1$.

Main Theorem 2 *Let* $A = \begin{pmatrix} a_{11} & a_{12} & a^*_{13} \\ a_{21} & a_{22} & a^*_{23} \\ & \vdots & \\ a_{M1} & a_{M2} & a^*_{M3} \end{pmatrix}$,

$B_j = \begin{pmatrix} b_{1j} & b_{1j}^{p+1} & b_{1j}^{2p+1} \\ b_{2j} & b_{2j}^{p+1} & b_{2j}^{2p+1} \\ b^*_{3j} & b^{*\,p+1}_{3j} & b^{*\,2p+1}_{3j} \end{pmatrix}$, $B = (B_1, \cdots, B_N)$

and a^*_{k3}, b^*_{3j} *for all* k, j *are constants. Suppose* $a^*_{k3} b^*_{3j} \neq 0$ *for some* k *and* j.

Set

$$\Psi = ||AB||^{2z} \mathrm{d}a\mathrm{d}b, \tag{4}$$

then the maximum pole $-\lambda$ *of* $\int_{||AB||<1} \Psi$ *and its order* θ *are as follows.*

Let N_1 *be the number of the set* $\{j \mid b^*_{3j} \neq 0\}$.

(a)

(i) If $N > M + \frac{p+1}{2}$ *then* $\lambda = \frac{2M+N}{2}$ *and* $\theta = 1$.

(ii) If $N_1 = 1$ *and* $N = M + \frac{p+1}{2}$ *(or* $N = M + \frac{p}{2}$*) then* $\lambda = \frac{2M+N}{2}$ *and* $\theta = 2$.

(iii) If $N_1 = 1$, $M+1 \leq N < M + \frac{p+1}{2}$ *and* p *is odd then* $\lambda = \frac{3M+3N-1+(p-1)(4M+2N-1)/2}{2(p+1)}$ *and* $\theta = 1$.

(iv) If $N_1 = 1$, $M+1 \leq N < M + \frac{p}{2}$ *and* p *is even then* $\lambda = \frac{3M+3N-1+(4M+2N-1)(p/2-1)}{2p}$ *and* $\theta = 1$.

(v) If $N_1 \geq 2$ *and* $M+1 < N$ *then* $\lambda = \frac{2M+N}{2}$ *and* $\theta = 1$.

(vi) If $N_1 \geq 2$ *and* $M+1 = N$ *then* $\lambda = \frac{2M+N}{2}$ *and* $\theta = 2$.

(vii) If $N = M$ *then* $\lambda = \frac{3M+3N-1}{4}$ *and* $\theta = 1$.

(viii) If $N = M-1$ *then* $\lambda = \frac{M+2N}{2}$ *and* $\theta = 2$.

(ix) If $N < M-1$ *then* $\lambda = \frac{M+2N}{2}$ *and* $\theta = 1$.

(b) The case of $a_{k1} + a_{k2} = 1$ *and* $a^*_{k3} = -1$.

(i) If $N > M + \frac{p+1}{2}$ *then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 1$.

(ii) If $N_1 = 1$ *and* $N = M + \frac{p+1}{2}$ *(or* $N = M + \frac{p}{2}$*) then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 2$.

(iii) If $N_1 = 1$, $M+1 \leq N < M + \frac{p+1}{2}$ *and* p *is odd then* $\lambda = \frac{M+3N-1+(p-1)(2M+2N-1)/2}{2(p+1)}$ *and* $\theta = 1$.

(iv) If $N_1 = 1$, $M+1 \leq N < M + \frac{p}{2}$ *and* p *is even then* $\lambda = \frac{M+3N-1+(2M+2N-1)(p/2-1)}{2p}$ *and* $\theta = 1$.

(v) If $N_1 \geq 2$ *and* $M+1 < N$ *then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 1$.

(vi) If $N_1 \geq 2$ *and* $M+1 = N$ *then* $\lambda = \frac{M+N}{2}$ *and* $\theta = 2$.

(vii) If $N = M$ *then* $\lambda = \frac{M+3N-1}{4}$ *and* $\theta = 1$.

(viii) If $N = M-1$ *then* $\lambda = N$ *and* $\theta = 2$.

(ix) If $N < M-1$ *then* $\lambda = N$ *and* $\theta = 1$.

Here we show two examples which have the zeta function $\int \Psi$ in (2).

Example 1

Consider the three layered neural network with N input units, H' hidden units and M output units which is trained for estimating the true distribution with $H-H'$ hidden units. Denote an input value by $x = (x_j) \in \mathbb{R}^N$ with a probability density function $q(x)$ which has a compact support $\tilde{W}$. Then an output value $y = (y_k) \in \mathbb{R}^M$ of the three layered neural network is given by $y_k = f_k(x, w)$+(noise), where $w = \{a_{ki}, b_{ij}; 1 \leq k \leq M, 1 \leq i \leq H', 1 \leq j \leq N\}$ and

$$f_k(x, w) = \sum_{i=1}^{H'} a_{ki} \tanh(\sum_{j=1}^{N} b_{ij} x_j).$$

Consider a statistical model

$$p(y|x, w) = \frac{1}{(2\pi)^{M/2}} \exp(-\frac{1}{2}||y - f(x, w)||^2).$$

Assume that the true distribution

$$p(y|x, w^*) = \frac{1}{(2\pi)^{M/2}} \exp(-\frac{1}{2}||y - f(x, w^*)||^2),$$

is included in the learning model, where $w^* = \{a^*_{ki}, b^*_{ij}; 1 \leq k \leq M, H'+1 \leq i \leq H, 1 \leq j \leq N\}$ and $f_k(x, w^*) = \sum_{i=H'+1}^{H} (-a^*_{ki}) \tanh(\sum_{j=1}^{N} b^*_{ij} x_j)$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w^*) > 0$. Then it has the zeta function $\int \Psi$ in (2) with $p = 2$.

This is proved by using a Taylor expansion together with Lemma 5 in [1].

Example 2

We consider the normal mixture model

$$p(x|w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{H'} a_{1i} \exp(-\frac{\sum_{j=1}^{N}(x_j - b_{ij})^2}{2}),$$

where $w = \{a_{1i}, b_{ij}; 1 \leq i \leq H', 1 \leq j \leq N\}$ and $\sum_{i=1}^{H'} a_{1i} = 1$. Set the true distribution by

$$p(x|w^*) = \frac{1}{(2\pi)^{N/2}} \sum_{i=H'+1}^{H} (-a^*_{1i}) \exp(-\frac{\sum_{j=1}^{N}(x_j - b^*_{ij})^2}{2}),$$

where $w^* = \{a^*_{1i}, b^*_{ij}; H'+1 \leq i \leq H, 1 \leq j \leq N\}$ and $\sum_{i=H'+1}^{H} a^*_{1i} = -1$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w^*) > 0$. Then it has the zeta function $\int \Psi$ in (2) with $p = 1$ and $M = 1$ [10].

4 Proof of Main Theorem 1

We need the following inductive statement $(*)$ of k for calculating poles by using a blowing up process.

$$(*)\qquad \Psi^* = \{v_{11}^{2k}(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2k}\left\|\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\Big\}^z v_{11}^{2N-1+(k-1)(M+N-1)}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1 = v_{11}^{p+1-k}b'_{21}$ and $B_j = b'_{2j}$, $j \geq 2$.

Construct blowing up of Ψ in (3) along the submanifold $\{b_{ij} = 0, i = 1, 2, 1 \leq j \leq N\}$.

Let $b_{11} = v_{11}$ and $b_{ij} = v_{11}b_{ij}$ for $(i,j) \neq (1,1)$.

Then we have

$$\Psi = \left\|\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ & \vdots \\ a_{M1} & a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^{2z} v_{11}^{2N-1}\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1 = \begin{pmatrix} v_{11} & v_{11}^{p+1} \\ v_{11}b_{21} & v_{11}^{p+1}b_{21}^{p+1}\end{pmatrix}$ and $B_j = \begin{pmatrix} v_{11}b_{1j} & v_{11}^{p+1}b_{1j}^{p+1} \\ v_{11}b_{2j} & v_{11}^{p+1}b_{2j}^{p+1}\end{pmatrix}$, $j \geq 2$.

By Lemmas 2 and 3 in [13], the maximum pole of $\int_W \Psi$ and its order are equal to those of $\int_W \Psi'$, where

$$\Psi' = \left\|\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ & \vdots \\ a_{M1} & a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^{2z} v_{11}^{2N-1}\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

$$B_1 = \begin{pmatrix} v_{11} & 0 \\ v_{11}b_{21} & v_{11}^{p+1}\begin{vmatrix} 1 & 1 \\ b_{21} & b_{21}^{p+1}\end{vmatrix}\end{pmatrix} \text{ and}$$

$$B_j = \begin{pmatrix} 0 & 0 \\ v_{11}\begin{vmatrix} 1 & b_{1j} \\ b_{21} & b_{2j}\end{vmatrix} & v_{11}^{p+1}\begin{vmatrix} 1 & b_{1j}^{p+1} \\ b_{21}^{p+1} & b_{2j}^{p+1}\end{vmatrix}\end{pmatrix},$$

$j \geq 2$.

We change the variables $d_{11}, \ldots, d_{M1}$ from $a_{11}, \ldots, a_{M1}$ by setting

$$\begin{pmatrix} d_{11} \\ d_{21} \\ \vdots \\ d_{M1}\end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ & \vdots \\ a_{M1} & a_{M2}\end{pmatrix}\begin{pmatrix} 1 \\ b_{21}\end{pmatrix}.$$

The case of $\mathbf{a_{i1} + a_{i2} = 1}$. If $a_{i1} + a_{i2} = 1$, then $d_{i1} = a_{i1}(1 - b_{21}) + b_{21}$. If $b_{21} = 1$, then $d_{i1} = 1$, so blowing up is completed.

We obtain

$$\Psi' = \{v_{11}^2(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^2\left\|\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z v_{11}^{2N-1}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where

$B_1 = v_{11}^p\begin{vmatrix} 1 & 1 \\ b_{21} & b_{21}^{p+1}\end{vmatrix}$ and

$B_j = \left(\begin{vmatrix} 1 & b_{1j} \\ b_{21} & b_{2j}\end{vmatrix} \quad v_{11}^p\begin{vmatrix} 1 & b_{1j}^{p+1} \\ b_{21}^{p+1} & b_{2j}^{p+1}\end{vmatrix}\right)$, $j \geq 2$.

Let $b'_{21} = \begin{vmatrix} 1 & 1 \\ b_{21} & b_{21}^{p+1}\end{vmatrix}$ and $b'_{2j} = \begin{vmatrix} 1 & b_{1j} \\ b_{21} & b_{2j}\end{vmatrix}$.

Then

$B_j = \begin{pmatrix} b'_{2j} & v_{11}^p((b'_{2j} + b_{21}b_{1j})^{p+1} - (b_{21}b_{1j})^{p+1})\end{pmatrix}$, $j \geq 2$.

Again by Lemmas 2 and 3 in [13], we set

$$\Psi'' = \{v_{11}^2(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^2\left\|\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|\}^{2z} v_{11}^{2N-1}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1 = v_{11}^p b'_{21}$ and $B_j = b'_{2j}$, $j \geq 2$.

We have the inductive statement $(*)$ of $k = 1$.

Construct blowing up of Ψ^* along the submanifold $\{v_{11} = b'_{22} = \cdots = b'_{2N} = d_{11} = \cdots = d_{M1} = 0\}$. Then we have (I), (II), (III) cases.

(I) Let $d_{11} = u_{11}$, $v_{11} = u_{11}v_{11}$, $b'_{2j} = u_{11}b'_{2j}$ for $j \geq 2$ and $d_{i1} = u_{11}d_{i1}$ for $i \geq 2$ in Ψ^*.

Then, we have

$$\Psi'^* = \{v_{11}^{2k}u_{11}^{2k+2}(1 + d_{21}^2 + \cdots + d_{M1}^2) + v_{11}^{2k}u_{11}^{2k+2}\left\|\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z \times v_{11}^{2N-1+(k-1)(M+N-1)}u_{11}^{2N-1+k(M+N-1)}\mathrm{d}d\mathrm{d}u\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1 = v_{11}^{p+1-k}u_{11}^{p-k}b'_{21}$ and $B_j = b'_{2j}$, $j \geq 2$.

We have poles $-\dfrac{2N + (k-1)(M+N-1)}{2k}$ and $-\dfrac{2N + k(M+N-1)}{2k+2}$.

(II) Let $b'_{22} = v_{22}$, $v_{11} = v_{22}v_{11}$, $b'_{2j} = v_{22}b'_{2j}$ for $j \geq 3$ and $d_{i1} = v_{22}d_{i1}$ for $i \geq 1$ in Ψ^*.

Then, we have

$$\Psi'^* = \{v_{11}^{2k}v_{22}^{2k+2}(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2k}v_{22}^{2k+2}\left\|\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z v_{11}^{2N-1+(k-1)(M+N-1)}v_{22}^{2N-1+k(M+N-1)}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1 = v_{11}^{p+1-k}v_{22}^{p-k}b'_{21}$, $B_2 = 1$ and $B_j = b'_{2j}$, $j \geq 3$.

Again by using Lemmas 2 and 3 in [13], the maximum pole of $\int_W \Psi'^*$ and its order are equal to those of

$\int_W \Psi''^*$, where

$$\Psi''^* = \{v_{11}^{2k}v_{22}^{2k+2}(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2k}v_{22}^{2k+2}\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}\right\|^2\}^z v_{11}^{2N-1+(k-1)(M+N-1)}v_{22}^{2N-1+k(M+N-1)}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b.$$

The case of $\mathbf{a_{i1}+a_{i2}=1}$. If $a_{i1}+a_{i2}=1$, then $a_{i2}=(1-v_{11}^{2k-2}v_{22}^{2k}d_{i1})/(1-b_{21})$. Therefore, it completes blowing up.

Construct blowing up of Ψ''^* along the submanifold $\{a_{12}=\cdots=a_{M2}=d_{11}=\cdots=d_{M1}=0\}$.

Let $d_{11}=u_{11}$, $d_{i1}=u_{11}d_{i1}$ for $i\geq 2$ and $a_{i2}=u_{11}a_{i2}$ for $i\geq 1$ in Ψ''^*.

Then

$$\Psi'''^* = \{v_{11}^{2k}v_{22}^{2k+2}u_{11}^2(1+d_{21}^2+\cdots+d_{M1}^2) + v_{11}^{2k}v_{22}^{2k+2}u_{11}^2\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}\right\|^2\}^z v_{11}^{2N-1+(k-1)(M+N-1)}v_{22}^{2N-1+k(M+N-1)}u_{11}^{2M-1}\mathrm{d}d\mathrm{d}u\mathrm{d}v\mathrm{d}a\mathrm{d}b.$$

We have poles $-\dfrac{2N+(k-1)(M+N-1)}{2k}$, $-\dfrac{2N+k(M+N-1)}{2k+2}$ and $-M$.

(III) Let $b'_{2j}=v_{11}b'_{2j}$ for $j\geq 2$ and $d_{i1}=v_{11}d_{i1}$ for $i\geq 1$ in Ψ^*.

Then, we have

$$\Psi'^* = \{v_{11}^{2k+2}(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2k+2}\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z v_{11}^{2N-1+k(M+N-1)}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1=v_{11}^{p-k}b'_{21}$ and $B_j=b'_{2j}$, $j\geq 2$.

Therefore we have the inductive statement $(*)$ of $k+1$.

We finish induction at $k=p+1$ and we have

$$\Psi^*_{p+1} = \{v_{11}^{2(p+1)}(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2(p+1)}\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z v_{11}^{2N-1+p(M+N-1)}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1=b'_{21}$ and $B_j=b'_{2j}$, $j\geq 2$.

Construct blowing up of Ψ^*_{p+1} along the submanifold $\{b_{21}=\cdots=b_{2N}=d_{11}=\cdots=d_{M1}=0\}$. Then we need (I'), (II') cases.

(I') Let $b'_{21}=v_{21}$, $b'_{2j}=v_{21}b'_{2j}$ for $j\geq 2$ and $d_{i1}=v_{21}d_{i1}$ for $i\geq 1$ in Ψ^*_{p+1}.

Then, we have

$$\Psi'^*_{p+1} = \{v_{11}^{2(p+1)}v_{21}^2(d_{11}^2+\cdots+d_{M1}^2) + v_{11}^{2(p+1)}v_{21}^2\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}\right\|^2\}^z v_{11}^{2N-1+p(M+N-1)}v_{21}^{M+N-1}\mathrm{d}d\mathrm{d}v\mathrm{d}a\mathrm{d}b.$$

The case of $\mathbf{a_{i1}+a_{i2}=1}$. Since $a_{i2}=(1-v_{11}^{2p}v_{21}^2d_{i1})/(1-b_{21})$, blowing up is finished.

Construct blowing up of Ψ'^*_{p+1} along the submanifold $\{a_{12}=\cdots=a_{M2}=d_{11}=\cdots=d_{M1}=0\}$.

Let $d_{11}=u_{11}$, $d_{i1}=u_{11}d_{i1}$ for $i\geq 2$ and $a_{i2}=u_{11}a_{i2}$ for $i\geq 1$ in Ψ'^*_{p+1}.

Then, we have

$$\Psi''^*_{p+1} = \{v_{11}^{2(p+1)}v_{21}^2u_{11}^2(1+d_{21}^2+\cdots+d_{M1}^2) + v_{11}^{2(p+1)}v_{21}^2u_{11}^2\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}\right\|^2\}^z v_{11}^{2N-1+p(M+N-1)}v_{21}^{M+N-1}u_{11}^{2M-1}\mathrm{d}d\mathrm{d}u\mathrm{d}v\mathrm{d}a\mathrm{d}b.$$

We have poles $-\dfrac{2N+p(M+N-1)}{2(p+1)}$, $-\dfrac{M+N}{2}$ and $-M$.

(II') Let $d_{11}=u_{11}$, $b'_{2j}=u_{11}b'_{2j}$ for $j\geq 1$ and $d_{i1}=u_{11}d_{i1}$ for $i\geq 2$ in Ψ^*_{p+1}.

Then we have

$$\Psi'^*_{p+1} = \{v_{11}^{2(p+1)}u_{11}^2(1+d_{21}^2+\cdots+d_{M1}^2) + v_{11}^{2(p+1)}u_{11}^2\left\|\begin{pmatrix} a_{12}\\ a_{22}\\ \vdots\\ a_{M2}\end{pmatrix}(B_1,\cdots,B_N)\right\|^2\}^z v_{11}^{2N-1+p(M+N-1)}u_{11}^{M+N-1}\mathrm{d}d\mathrm{d}u\mathrm{d}v\mathrm{d}a\mathrm{d}b,$$

where $B_1=b'_{21}$ and $B_j=b'_{2j}$, $j\geq 2$.

We have poles $-\dfrac{2N+p(M+N-1)}{2(p+1)}$ and $-\dfrac{M+N}{2}$.

From all the inductions, we have poles $-\dfrac{2N+(k-1)(M+N-1)}{2k}$ for $k=1,\cdots,p+1$, $-\dfrac{M+N}{2}$ and M.

If $N \geq M-1$ then $\frac{2N+(k-1)(M+N-1)}{2k} \geq \frac{2N+k(M+N-1)}{2(k+1)}$.

If $N - \frac{p}{p+2} \geq M$ then $\frac{2N+p(M+N-1)}{2(p+1)} \geq M$.

Therefore

(i) If $N \geq M+1$ then $\lambda = M$ and $\theta = 1$.

(ii) If $N = M$ then $\lambda = \frac{2N+p(M+N-1)}{2(p+1)}$ and $\theta = 1$.

(iii) If $N = M-1$ then $\lambda = N$ and $\theta = 2$.

(iv) If $N < M-1$ then $\lambda = N$ and $\theta = 1$.

The case of $\mathbf{a_{i1} + a_{i2} = 1}$. M does not appear.

As space is limited, the proof of Main Theorem 2 is omitted here. We also use a blowing up process, but need more complicated method, since we have the constant terms such as b_{3j}^*.

5 Conclusion

In this paper, we consider the asymptotic form of the generalization error in Bayesian estimation for the three layered learning models.

Let $p(y|x,w) = \frac{1}{(2\pi)^{M/2}} \exp(-\frac{1}{2}||y - f(x,w)||^2)$, where $f_k(x,w) = \sum_{i=1}^{2} a_{ki} \tanh(\sum_{j=1}^{N} b_{ij} x_j)$ for $1 \leq k \leq M$. This model is the three layered neural network with N input units, 2 hidden units and M output units. If the true distribution is zero, the generalization errors $G(n)$ in (1) is given by Main Theorem 1 (a) with $p = 2$. If it is trained for estimating the true distribution represented by the model with 1 hidden units, $G(n)$ is obtained by Main Theorem 2 (a).

The normal mixture model $p(x|w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{2} a_{1i} \exp(-\frac{\sum_{j=1}^{N}(x_j - b_{ij})^2}{2})$, with $a_{11} + a_{12} = 1$ has the generalization error $G(n)$ given by Main Theorem 2 (b) with $p = 1$ and $M = 1$, if a true distribution is $p(x|w^*) = \frac{1}{(2\pi)^{N/2}} \exp(-\frac{\sum_{j=1}^{N}(x_j - b_j^*)^2}{2})$. If $p = 1$ and $M = 1$, Main Theorem 1 (b) corresponds to the case of the true distribution $p(x|w^*) = \frac{1}{(2\pi)^{N/2}} \exp(-\frac{\sum_{j=1}^{N} x_j^2}{2})$. It is easily checked that Main Theorem 1 (b) and Main Theorem 2 (b) are the same results if $p = 1$, since their difference is only $(b_j^*) \neq 0$ or $(b_j^*) = 0$.

The applications of our result are as follows. From these results, we can construct mathematical foundation for analyzing and developing the precision of the MCMC method [14], [15]. Also we would compare these values to such as the generalization error of localized Bayes estimation [16].

We could see that the blowing up method in algebraic geometry can be effectively used for solving the problems in the learning theory. Our future purpose is to improve our method for applying arbitrary H and H'.

Acknowledgments

This research was supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 16700218.

References

[1] S. Watanabe, Algebraic analysis for nonidentifiable learning machines, *Neural Computation*, 13 (4), 2001, 899-933.

[2] S. Watanabe, Algebraic geometrical methods for hierarchical learning machines, *Neural Networks*, 14 (8) 2001, 1049-1060.

[3] H. Akaike, A new look at the statistical model identification, *IEEE Trans. on Automatic Control*, 19 1974, 716-723.

[4] K. Takeuchi, Distribution of an information statistic and the criterion for the optimal model, *Mathematical Science*, 153, 1976, 12-18.

[5] E. J. Hannan and B. G. Quinn, The determination of the order of an autoregression. *Journal of Royal Statistical Society, Series B*, 41, 1979, 190-195.

[6] N. J. Murata, S. G. Yoshizawa, and S. Amari, Network information criterion - determining the number of hidden units for an artificial neural network model, *IEEE Trans. on Neural Networks*, 5 (6), 1994, 865-872.

[7] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, 6 (2) 1978, 461-464.

[8] J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. on Information Theory*, 30 (4) 1984, 629-636.

[9] H. Hironaka, Resolution of Singularities of an algebraic variety over a field of characteristic zero, *Annals of Math*, 79 1964, 109-326.

[10] S. Watanabe, K. Yamazaki and M. Aoyagi, Kullback Information of Normal Mixture is not an Analytic Function, *Technical report of IEICE*, NC2004, 2004, 41-46.

[11] M. Aoyagi and S. Watanabe, Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network, *IEICE Trans.*, J88-D-II, 10, 2005, 2112-2124, (English version : *Systems and Computers in Japan*, John Wiley & Sons, Inc. (in press).

[12] M. Aoyagi, The zeta function of learning theory and generalization error of three layered neural perceptron, *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities*, 2006, (in press).

[13] M. Aoyagi and S. Watanabe, Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation, *Neural Networks*, 18, 2005, 924-933.

[14] N. Nakano, K. Takahashi and S. Watanabe, On the evaluation criterion of the MCMC Method in Singular Learning machines, *IEICE Trans.*, J88-D-II, 10, 2005, 2011-2020.

[15] K. Nagata and S. Watanabe, A proposal and effectiveness of the optimal approximation for Bayesian posterior distribution, *Workshop on Information-Based Induction Sciences*, 2005, 99-104.

[16] S. Takamatsu, S. Nakajima and S. Watanabe, Generalization Error of Localized Bayes Estimation in Reduced Rank Regression, *Workshop on Information-Based Induction Sciences*, 2005, 81-86.