Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity

Miki Aoyagi[1] **and Kenji Nagata**[2]

[1] Department of Mathematics, College of Science & Technology
Nihon University
1-8-14, Surugadai, Kanda, Chiyoda-ku, 101-8308, Japan,
aoyagi.miki@nihon-u.ac.jp.

[2]Graduate School of Frontier Science, The University of Tokyo
5-1-5, Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan, nagata@mns.k.u-tokyo.ac.jp.

Keywords: Generalization error, three layered neural networks, normal mixture models, non-regular learning machine, resolution of singularities, zeta function

Abstract

Recently, the term "algebraic statistics" arises from the study of probabilistic models and techniques for statistical inference using methods from algebra and geometry (Sturmfels, 2008). Our study is to consider the generalization error and stochastic complexity in learning theory by using the log canonical threshold in algebraic geometry. Such thresholds correspond to the main term of the generalization error in Bayesian estimation, which is called a learning coefficient (Watanabe, 2001a,b). The learning coefficient serves to measure the learning efficiencies in hierarchical learning models. In this paper, we consider learning coefficients for Vandermonde matrix type singularities, by using a new approach : focusing on the generators of the ideal which defines singularities. We give new tight bound values of learning coefficients for the Vandermonde matrix type singularities and the explicit values with certain conditions. By applying

our results, the learning coefficients of three layered neural networks and normal mixture models are shown.

1 Introduction

In this paper, we consider the generalization error and stochastic complexity in learning theory by using a log canonical threshold in algebraic geometry.

The log canonical threshold $\lambda_Z(Y, f)$ is analytically defined by

$$\lambda_Z(Y, f) = \sup\{c : |f|^{-c} \text{ is locally } L^2 \text{ near } Z\},$$

over $\mathbb{C}$ and

$$\lambda_Z(Y, f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ near } Z\},$$

over $\mathbb{R}$ for a nonzero holomorphic function f over $\mathbb{C}$ or an analytic function f over $\mathbb{R}$ on a smooth variety Y, where $Z \subset Y$ is a closed subscheme (Kollár, 1997; Mustata, 2002). It is known that if f is a polynomial or a convergent power series, $\lambda_0(\mathbb{C}^d, f)$ is the largest root of the Bernstein-Sato polynomial $b(s) \in \mathbb{C}[s]$ of f, where $b(s)f^s = Pf^{s+1}$ for a linear differential operator P (Bernstein, 1972; Bjőrk, 1979; Kashiwara, 1976). The log canonical threshold $\lambda_Z(Y, f)$ also corresponds to the largest pole of $\int_{\text{near } Z} |f|^{2\xi}\psi(w)dw$ over $\mathbb{C}$, ($\int_{\text{near } Z} |f|^{\xi}\psi(w)dw$ over $\mathbb{R}$) for a complex variable ξ, where $\psi(w)$ is a $C^\infty-$ function with a compact support and $\psi(w) \neq 0$ on Z.

Such thresholds serve to measure the learning efficiencies in hierarchical learning models, i.e., they correspond to the main terms of generalization errors in learning systems.

The purpose of the learning system is to estimate an unknown true density function which distributes data. Real data associated with genetic analysis, data mining, image or speech recognition, artificial intelligence, the control of a robot, time series prediction, and so on, are very complicated and usually not generated by a simple normal distribution. Hierarchical learning models such as the layered neural network, the Boltzmann machine, the reduced rank regression and the normal mixture model are known to be effective learning models for analyzing such data. They are, however, non-regular statistical models, which cannot be analyzed using the classic theories of regular statistical models (Hartigan, 1985; Sussmann, 1992; Hagiwara et al., 1993; Fukumizu, 1996).

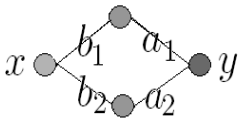


Figure 1: Simple three layered neural network : One input unit, one output unit and two hidden units. The output y is expressed by $y = a_1 \tanh(b_1 x) + a_2 \tanh(b_2 x) +$ (noise) where x is the input.

For example, consider a simple three layered neural network that has one input unit, one output unit and two hidden units (Fig. 1). The model is expressed by the probability form of one input $x \in \mathbb{R}$, one output $y \in \mathbb{R}$ with a parameter $w = (a_1, a_2, b_1, b_2) \in \mathbb{R}^4$:

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - a_1 \tanh(b_1 x) - a_2 \tanh(b_2 x))^2).$$

Assume that the true density function is $p(y|x, w_t^*)$ with $w_t^* = 0$. Then the true parameter set is $\{w = (a_1, a_2, b_1, b_2) \in \mathbb{R}^4 | p(y|x, w_t^*) = p(y|x, w)\} = \{w = (a_1, a_2, b_1, b_2) \in \mathbb{R}^4 | a_1 \tanh(b_1 x) + a_2 \tanh(b_2 x) = 0, \text{ for any } x\} = \{b_1 = \pm b_2, a_1 = \mp a_2\} \cup \{a_1 b_1 = a_2 b_2 = 0\}$. This set does not consist of only one point, resulting in a non-positive definite Fisher information matrix. Usually, the true parameter set of non-regular models is an analytic set with complicated singularities. Consequently, it is difficult to solve theoretical problems, such as clarifying generalization errors in learning theory.

The generalization error measures the difference between the true density function $q(z)$ and the predictive density function $p(z|(z)^n)$ obtained using n distributed training samples $(z)^n = (z_1, \ldots, z_n)$ of z from the true density function $q(z)$.

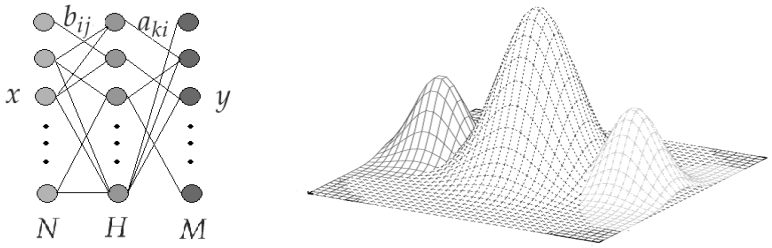
In the case of Figure 1, the notation z corresponds to (x, y), and we have $p(x, y|w) = p(y|x, w)q(x)$ with a probability density function $q(x)$ of an input value x.

We define it as the average Kullback distance between $q(z)$ and $p(z|(z)^n)$:

$$G(n) = E_n\{\int q(z) \log \frac{q(z)}{p(z|(z)^n)} dz\},$$

where E_n is the expectation value over n training samples. This function clarifies precisely how $p(z|(z)^n)$ can approximate $q(z)$. Thus, $G(n)$ is also called a learning curve or a learning efficiency. The classic model selection methods of regular statistical models such as AIC (Akaike, 1974), TIC (Takeuchi, 1976), HQ (Hannan & Quinn, 1979),

NIC (Murata et al., 1994), BIC (Schwarz, 1978), and MDL (Rissanen, 1984), cannot apply to the generalization error for non-regular models, since the true parameter set of regular models should be one point and its Fisher information matrix is positive definite. Therefore, it is important to construct a mathematical foundation for clarifying generalization errors of non-regular models. It is well known that Bayesian estimation is more appropriate than the maximum likelihood method when a learning machine is non-regular (Akaike, 1980; Mackay, 1992). We usually consider the generalization error in terms of a direct and an inverse problem. The direct problem involves solving the generalization error with a known true density function. The inverse problem is finding 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 solve it, we first need to solve the direct problem. In this paper, we consider the direct problem of Vandermonde matrix type singularities over the real field (Definition 4).



(a) (b)

Figure 2: (a) The three layered neural network : N input units, M output units and H hidden units. The output y_k is expressed by $y_k = \sum_{i=1}^{H} a_{ki} \tanh(\sum_{j=1}^{N} b_{ij} x_j) +$ (noise) where x_j is an input. (b) The normal mixture model with identity matrix variances: $p(z|w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{H} a_i \exp(-\frac{\sum_{j=1}^{N}(z_j - b_{ij})^2}{2})$, with $\sum_{i=1}^{H} a_i = 1$, $a_i \geq 0$, which has H peaks.

By focusing on the generators of the ideal which defines singularities, we firstly show that learning coefficients for the three layered neural network and for normal mixture models with identity matrix variances are obtained by the Vandermonde matrix type singularities (Theorem 8 and Theorem 9 Fig. 2). Next, we have (1) new tight bound values of learning coefficients for the Vandermonde matrix type singularities (Theorem

10), and (2) the explicit values under certain conditions (Theorem 11). The explicit values in Theorem 11 are equal to the bound values in Theorem 10. By applying these results, we have the learning coefficients for the three layered neural network and for the normal mixture models (Theorem 12, Theorem 13).

Learning coefficients for mixtures of binomial distributions are also obtained by Vandermonde matrix type singularities (Yamazaki et al. (2010)). The Vandermonde matrix type is a generic one in neural computation because almost all singularities in neural computation are classified into the Vandermonde matrix type.

We have already obtained learning coefficients for the three layered neural network with one input unit and one output unit (Aoyagi, 2005a, 2006). Learning coefficients in the case of the normal mixture models with dimension one have been obtained recently (Aoyagi, 2010a). We have also obtained the exact asymptotic forms of the generalization errors for the reduced rank regression (Aoyagi, 2005b) and for the restricted Boltzmann machine model (Aoyagi, 2010b). Rusakov & Geiger (2005) obtained them for Naive Bayesian networks.

This paper consists of five sections. In Section 2, we summarize the framework of Bayesian learning models. Section 3 describes our main results. To confirm our theoretical results, numerical experiments are shown in Section 4, and we give our conclusions in Section 5.

2 Generalization error and stochastic complexity in Bayesian estimation

In this paper, we consider the stochastic complexity and the generalization error in Bayesian estimation.

Let $q(z)$ be a true probability density function and $(z)^n := \{z_i\}_{i=1}^n$ be n training independent and identical samples from $q(z)$. Consider a learning model which is written by a probability form $p(z|w)$, where w is a parameter. The purpose of the learning system is to estimate $q(z)$ from $(z)^n$ by using $p(z|w)$.

Let $p(w|(z)^n)$ be the *a posteriori* probability density function:

$$p(w|(z)^n) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^{n} p(z_i|w),$$

where $\psi(w)$ is an *a priori* probability density function on the parameter set W and

$$Z_n = \int_W \psi(w) \prod_{i=1}^{n} p(z_i|w) \mathrm{d}w.$$

So the average inference $p(z|(z)^n)$ of the Bayesian density function is given by

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

which is the predictive density function.

Set

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

This function always has a nonnegative value and satisfies $K(q||p) = 0$ if and only if $q(z) = p(z|(z)^n)$.

The generalization error $G(n)$ is its expectation value E_n over n training samples:

$$G(n) = E_n\{ \int q(z) \log \frac{q(z)}{p(z|(z)^n)} \mathrm{d}z\}.$$

Let

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(z_i)}{p(z_i|w)}.$$

The average stochastic complexity or the free energy is defined by

$$F(n) = -E_n\{\log \int \exp(-nK_n(w)) \psi(w) \mathrm{d}w\}.$$

Then we have $G(n) = F(n+1) - F(n)$ for an arbitrary natural number n (Levin et al., 1990; Amari et al., 1992; Amari & Murata, 1993). $F(n)$ is known as the Bayesian criterion in Bayesian model selection (Schwarz, 1978), stochastic complexity in universal coding (Rissanen, 1986; Yamanishi, 1998), Akaike's Bayesian criterion in optimization of hyperparameters (Akaike, 1980) and evidence in neural network learning (Mackay, 1992). In addition, $F(n)$ is an important function for analyzing the generalization error.

It has recently been proved that the largest pole of a zeta function gives the generalization error of hierarchical learning models asymptotically (Watanabe, 2001a,b, 2010). We assume that the true density distribution $q(z)$ is included in the learning model, i.e., $q(z) = p(z|w_t^*)$ for $w_t^* \in W$, where W is the parameter space.

Define the zeta function $J(\xi)$ of a complex variable ξ for the learning model by

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

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

$$K(w) = \int p(z|w_t^*) \log \frac{p(z|w_t^*)}{p(z|w)} \mathrm{d}z.$$

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

$$F(n) = \lambda \log n - (\theta - 1) \log \log n + O(1), \tag{1}$$

where $O(1)$ is a bounded function of n, and

$$G(n) \cong \frac{\lambda}{n} - \frac{\theta - 1}{n \log n} \text{ as } n \to \infty. \tag{2}$$

Therefore, our aim in this paper is to obtain λ and θ.

To assist in achieving this aim, we use the desingularization in algebraic geometry (Watanabe, 2009; Fulton, 1993). In algebraic geometry, a learning coefficient corresponds to a log-canonical threshold. Many studies on it in algebraic geometry have usually been done on an algebraically closed field such as the complex field. One of the recent results is relating to arc spaces by Mustata (2002). Our study is over the real field and it is therefore, a new problem, even in mathematics, to obtain desingularizations of such Kullback functions.

3 Main Results

In this section, we show our main results.

3.1 Learning coefficients for Vandermonde matrix type singularities

In this paper, we denote by a^*, b^*, w^* constants, using the suffix $*$. Also for simplicity, we denote $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ instead of $w = \{a_{ki}, b_{ij}\}_{1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N}$, in this paper.

Define the norm of a matrix $C = (c_{ij})$ by $||C|| = \sqrt{\sum_{i,j} |c_{ij}|^2}$. Set $\mathbb{N}_{+0} = \mathbb{N} \cup \{0\}$.

Definition 1 *Let $\lambda_{w^*}(f)$ be the largest pole of $\int_U |f|^\xi \psi dw$ and $\theta_{w^*}(f)$ its order, where U is a sufficiently small neighborhood of w^*, f is a real analytic function in a neighborhood of w^* and ψ is a C^∞ function with compact support and $\psi(w^*) \neq 0$.*

We give below Lemma 2 as they are frequently used in this paper.

Lemma 2 (Aoyagi (2009, 2010a); Lin (2010)) *Let U be a neighborhood of $w^* \in \mathbb{R}^d$. Let $\mathcal{I}$ be the ideal generated by $f_1, \ldots, f_n$ which are analytic functions defined on U. Also let $\psi(w)$ be a C^∞ function on U with compact support. If $g_1, \ldots, g_m \in \mathcal{I}$, then $\lambda_{w^*}(f_1^2 + \cdots + f_n^2)$ is greater than $\lambda_{w^*}(g_1^2 + \cdots + g_m^2)$. In particular, if $g_1, \ldots, g_m$ generate the ideal $\mathcal{I}$ then*

$$\lambda_{w^*}(f_1^2 + \cdots + f_n^2) = \lambda_{w^*}(g_1^2 + \cdots + g_m^2).$$

(Proof)

The fact $g_1^2 + \cdots + g_m^2 \leq P(f_1^2 + \cdots + f_n^2)$ for $P >> 1$ yields this lemma.

Q.E.D.

Definition 3 *Fix $Q \in \mathbb{N}$. Define $[b_1^*, b_2^*, \cdots, b_N^*]_Q = \gamma_i(0, \cdots, 0, b_i^*, \cdots, b_N^*)$ if $b_1^* = \cdots = b_{i-1}^* = 0$, $b_i^* \neq 0$, and $\gamma_i = \begin{cases} 1 & \text{if } Q \text{ is odd,} \\ |b_i^*|/b_i^* & \text{if } Q \text{ is even.} \end{cases}$*

Definition 4 *Fix $Q \in \mathbb{N}$ and $m \in \mathbb{N}_{+0}$.*

$$\text{Let } A = \begin{pmatrix} a_{11} & \cdots & a_{1H} & a_{1,H+1}^* & \cdots & a_{1,H+r}^* \\ a_{21} & \cdots & a_{2H} & a_{2,H+1}^* & \cdots & a_{2,H+r}^* \\ & \vdots & & & \vdots & \\ a_{M1} & \cdots & a_{MH} & a_{M,H+1}^* & \cdots & a_{M,H+r}^* \end{pmatrix}, I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N,$$

$$B_I = (\prod_{j=1}^N b_{1j}^{\ell_j}, \prod_{j=1}^N b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^N b_{Hj}^{\ell_j}, \prod_{j=1}^N {b_{H+1,j}^*}^{\ell_j}, \cdots, \prod_{j=1}^N {b_{H+r,j}^*}^{\ell_j})^t,$$

and

$$B = (B_I)_{\ell_1 + \cdots + \ell_N = Qn+m, n \geq 0} = (B_{(m,0,\cdots,0)}, B_{(m-1,1,\cdots,0)}, \cdots, B_{(0,0,\cdots,m)}, B_{(m+Q,0,\cdots,0)}, \cdots)$$

(t denotes the transpose).

a_{ki} and b_{ij} $(1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N)$ are the variables in a neighborhood of a_{ki}^* and b_{ij}^*, where a_{ki}^* and b_{ij}^* are fixed constants.

Let $\mathcal{I}$ be the ideal generated by the elements of AB.

We call singularities of $\mathcal{I}$ Vandermonde matrix type singularities.

To simplify, we usually assume that

$$(a^*_{1,H+j}, a^*_{2,H+j}, \cdots, a^*_{M,H+j})^t \neq 0, (b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N}) \neq 0,$$

for $1 \leq j \leq r$ *and*

$$[b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N}]_Q \neq [b^*_{H+j',1}, b^*_{H+j',2}, \cdots, b^*_{H+j',N}]_Q,$$

for $j \neq j'$.

The Vandermonde matrix type singularities are degenerate with respect to their Newton polyhedrons (Fulton, 1993), their singularities are not isolated.

In general, singularities appeared in learning theory have such properties, and therefore, obtaining the log canonical thresholds is a still difficult problem.

Remark 1 *The ideal* $\mathcal{I}$ *in Definition 4 is also generated by the elements of* AB' *where*

$$B' = (B_I)_{\ell_1+\cdots+\ell_N = Qn+m, 0 \leq n \leq H+r-1}.$$

Example 5 *If* $m = N = M = r = 1, Q = H = 2$, *then we have* $A = \begin{pmatrix} a_{11} & a_{12} & a^*_{13} \end{pmatrix}$,

$$B' = \begin{pmatrix} b_{11} & b_{11}^3 & b_{11}^5 \\ b_{21} & b_{21}^3 & b_{21}^5 \\ b^*_{31} & {b^*_{31}}^3 & {b^*_{31}}^5 \end{pmatrix}.$$

These A, B' *are for the simple neural network in Figure 1:*

$$p(y|x,w) = \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}(y - a_{11}\tanh(b_{11}x) - a_{12}\tanh(b_{21}x))^2),$$

and the true distribution

$$p(y|x,w^*_t) = \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}(y + a^*_{13}\tanh(b^*_{31}x))^2).$$

Example 6 *If* $Q = m = M = r = 1, H = 2, N = 2$, *then we have* $A = \begin{pmatrix} a_{11} & a_{12} & a^*_{13} \end{pmatrix}$,

$$B' = \begin{pmatrix} b_{11} & b_{12} & b_{11}^2 & b_{11}b_{12} & b_{12}^2 & b_{11}^3 & b_{11}b_{12}^2 & b_{11}^2 b_{12} & b_{12}^3 \\ b_{21} & b_{22} & b_{21}^2 & b_{21}b_{22} & b_{22}^2 & b_{21}^3 & b_{21}b_{22}^2 & b_{21}^2 b_{22} & b_{22}^3 \\ b^*_{31} & b^*_{32} & {b^*_{31}}^2 & b^*_{31}b^*_{32} & {b^*_{32}}^2 & {b^*_{31}}^3 & b^*_{31}{b^*_{32}}^2 & {b^*_{31}}^2 b^*_{32} & {b^*_{32}}^3 \end{pmatrix}.$$

If $a^*_{13} = -1$, *these* A, B' *are for a normal mixture model with identity matrix variances*

$$p(z|w) = \frac{a_{11}}{2\pi}\exp(-\frac{(z_1 - b_{11})^2 + (z_2 - b_{12})^2}{2}) + \frac{a_{12}}{2\pi}\exp(-\frac{(z_1 - b_{21})^2 + (z_2 - b_{22})^2}{2}),$$

$\sum_{i=1}^{2} a_{1i} = 1$, $a_{1i} \geq 0$, *and the true distribution is*

$$p(z|w_t^*) = \frac{1}{2\pi}(-a_{13}^*) \exp(-\frac{(z_1 - b_{31}^*)^2 + (z_2 - b_{32}^*)^2}{2}), -a_{13}^* = 1.$$

In this paper, we denote

$$A_{M,H} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} \\ a_{21} & a_{22} & \cdots & a_{2H} \\ & & \vdots & \\ a_{M1} & a_{M2} & \cdots & a_{MH} \end{pmatrix}, B_{H,N,I} = \begin{pmatrix} \prod_{j=1}^{N} {b_{1j}}^{\ell_j} \\ \prod_{j=1}^{N} {b_{2j}}^{\ell_j} \\ \vdots \\ \prod_{j=1}^{N} {b_{Hj}}^{\ell_j} \end{pmatrix} and$$

$B_{H,N}^{(Q,m)} = (B_{H,N,I})_{\ell_1+\ldots+\ell_N=Qn+m, 0\leq n\leq H-1}$.

Also we denote

$$(A_{M,H}, \mathbf{a}^*) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a_{1,H+1}^* \\ a_{21} & a_{22} & \cdots & a_{2H} & a_{2,H+1}^* \\ & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a_{M,H+1}^* \end{pmatrix}, \begin{pmatrix} B_{H,N,I} \\ \mathbf{b}^* \end{pmatrix} = \begin{pmatrix} \prod_{j=1}^{N} {b_{1j}}^{\ell_j} \\ \prod_{j=1}^{N} {b_{2j}}^{\ell_j} \\ \vdots \\ \prod_{j=1}^{N} {b_{Hj}}^{\ell_j} \\ \prod_{j=1}^{N} {b_{H+1,j}^*}^{\ell_j} \end{pmatrix}$$

and $\begin{pmatrix} B_{H,N}^{(Q,m)} \\ \mathbf{b}^* \end{pmatrix} = \begin{pmatrix} B_{H,N,I} \\ \mathbf{b}^* \end{pmatrix}_{\ell_1+\ldots+\ell_N=Qn+m, 0\leq n\leq H}$, where $\mathbf{a}^* = \begin{pmatrix} a_{1,H+1}^* \\ \vdots \\ a_{M,H+1}^* \end{pmatrix}$

and $\mathbf{b}^* = (b_{H+1,1}^*, \cdots, b_{H+1,N}^*)$.

Theorem 7 (Aoyagi (2010a)) *Consider a sufficiently small neighborhood U of*

$$w^* = \{a_{ki}^*, b_{ij}^*\}_{1\leq i\leq H},$$

and variables $w = \{a_{ki}, b_{ij}\}_{1\leq i\leq H}$ in the set U.

Set $(b_{01}^{**}, b_{02}^{**}, \cdots, b_{0N}^{**}) = (0, \ldots, 0)$.

Let each $(b_{11}^{**}, b_{12}^{**}, \cdots, b_{1N}^{**})$, ..., $(b_{r'1}^{**}, b_{r'2}^{**}, \cdots, b_{r'N}^{**})$ be a different real vector in

$$[b_{i1}^*, b_{i2}^*, \cdots, b_{iN}^*]_Q \neq 0, \text{ for } i = 1, \ldots, H + r.$$

That is,

$$\{(b_{11}^{**}, \cdots, b_{1N}^{**}), \ldots, (b_{r'1}^{**}, \cdots, b_{r'N}^{**}) \; ; \; [b_{i1}^*, \cdots, b_{iN}^*]_Q \neq 0, i = 1, \ldots, H + r\}.$$

Then r' is uniquely determined and $r' \geq r$ by the assumption in Definition 4, and set $(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^{*}_{H+i,1}, \cdots, b^{*}_{H+i,N}]_Q$, for $1 \leq i \leq r$.

Assume that

$$[b^{*}_{i1}, \cdots, b^{*}_{iN}]_Q = \begin{cases} 0, & 1 \leq i \leq H_0 \\ (b^{**}_{11}, \cdots, b^{**}_{1N}), & H_0 + 1 \leq i \leq H_0 + H_1, \\ (b^{**}_{21}, \cdots, b^{**}_{2N}), & H_0 + H_1 + 1 \leq i \leq H_0 + H_1 + H_2, \\ \vdots & \\ (b^{**}_{r'1}, \cdots, b^{**}_{r'N}), & H_0 + \cdots + H_{r'-1} + 1 \leq i \leq H_0 + \cdots + H_{r'}, \end{cases}$$

and $H_0 + \cdots + H_{r'} = H$.

Then we have

$$\lambda_{w^*}(||AB||^2) = \lambda_{w^{(0)*}}(||A_{M,H_0} B^{(Q,m)}_{H_0,N}||^2)$$

$$+ \sum_{\alpha=1}^{r} \lambda_{w^{(\alpha)*}}(||(A_{M,H_\alpha}, \mathbf{a}^{(\alpha)*}) \begin{pmatrix} B^{(1,0)}_{H_\alpha,N} \\ \mathbf{b}^{(\alpha)*} \end{pmatrix} ||^2) + \sum_{\alpha=r+1}^{r'} \lambda_{w^{(\alpha)*}}(||A_{M,H_\alpha} B^{(1,0)}_{H_\alpha,N}||^2),$$

where $w^{(0)*} = \{a^{*}_{ki}, 0\}_{1 \leq i \leq H_0}$, $w^{(\alpha)*} = \{a^{*}_{k,H_0+\cdots+H_{\alpha-1}+i}, b^{**}_{\alpha j}\}_{1 \leq i \leq H_\alpha}$, $\mathbf{a}^{(\alpha)*} = \begin{pmatrix} a^{*}_{1,H+\alpha} \\ \vdots \\ a^{*}_{M,H+\alpha} \end{pmatrix}$

and $\mathbf{b}^{(\alpha)*} = (b^{**}_{\alpha 1}, \cdots, b^{**}_{\alpha N})$ *for* $\alpha \geq 1$.

Moreover,

$$\lambda_{w^*}(||AB||^2) = \frac{Mr'}{2} + \lambda_{w_1^{(0)*}}(||A_{M,H_0} B^{(Q,m)}_{H_0,N}||^2)$$

$$+ \sum_{\alpha=1}^{r} \lambda_{w_1^{(\alpha)*}}(||(A_{M,H_\alpha-1}, \mathbf{a}^{(\alpha)*}) B^{(1,1)}_{H_\alpha,N}||^2) + \sum_{\alpha=r+1}^{r'} \lambda_{w_1^{(\alpha)*}}(||A_{M,H_\alpha-1} B^{(1,1)}_{H_\alpha-1,N}||^2),$$

where $w_1^{(0)*} = \{a^{*}_{ki}, 0\}_{1 \leq i \leq H_0}$,

$w_1^{(\alpha)*} = \{a^{*}_{k,H_0+\cdots+H_{\alpha-1}+i}, 0\}_{2 \leq i \leq H_\alpha}$ *and* $\mathbf{a}^{(\alpha)*} = \begin{pmatrix} a^{*}_{1,H+\alpha} \\ \vdots \\ a^{*}_{M,H+\alpha} \end{pmatrix}$ *for* $\alpha \geq 1$.

Theorem 7 is used for the proofs of Theorems 8 and 9.

3.2 Three layered neural network

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 r hidden units.

Denote an input value by $x = (x_j) \in \mathbb{R}^N$ with a probability density function $q(x)$. 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 $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ and

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

Consider a statistical model

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

and $p(x, y|w) = p(y|x, w)q(x)$. Assume that the true distribution

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

is included in the learning model, where $w_t^* = \{a_{k,H+i}^*, b_{H+i,j}^*\}_{1 \leq i \leq r}$ and $f_k(x, w_t^*) = \sum_{i=1}^{r} (-a_{k,H+i}^*) \tanh(\sum_{j=1}^{N} b_{H+i,j}^* x_j)$.

Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$. We have

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

and the notaion (x, y) for the three layered neural network corresponds to z in Section 2.

Let $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a_{1,H+1}^* & \cdots & a_{1,H+r}^* \\ a_{21} & a_{22} & \cdots & a_{2H} & a_{2,H+1}^* & \cdots & a_{2,H+r}^* \\ & & \vdots & & & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a_{M,H+1}^* & \cdots & a_{M,H+r}^* \end{pmatrix}$, $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$,

$$B_I = (\prod_{j=1}^{N} b_{1j}^{\ell_j}, \prod_{j=1}^{N} b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^{N} b_{Hj}^{\ell_j}, \prod_{j=1}^{N} {b_{H+1,j}^*}^{\ell_j}, \cdots, \prod_{j=1}^{N} {b_{H+r,j}^*}^{\ell_j})^t,$$

and $B = (B_I)_{\ell_1 + \cdots + \ell_N = 2n-1, 1 \leq n \leq H+r}$. Set

$$\Psi = ||AB||^{2\xi} \mathrm{d}a\mathrm{d}b.$$

Theorem 8 *Consider a sufficiently small neighborhood U of $w^* = \{a_{ki}^*, b_{ij}^*\}_{1 \leq i \leq H}$, where $||AB|| = 0$ at w^*. Let $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ be in U.*

The learning coefficient for the three layered neural network is the largest pole of $\int_{||AB|| < 1} \Psi$, i.e., $\lambda_{w^*}(||AB||^2)$ with $Q = 2$ and $m = 1$.

This is proved by using a Taylor expansion $\tanh(x) = x + c_1 x^3 + c_2 x^5 + \cdots$ ($c_1, c_2, \ldots \in \mathbb{R}$) together with Lemma 5 in (Watanabe, 2001a).

3.3 Normal mixture model

We consider a normal mixture model with identity matrix variances

$$p(z|w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{H} a_i \exp(-\frac{\sum_{j=1}^{N}(z_j - b_{ij})^2}{2}),$$

where $w = \{a_i, b_{ij}\}_{1 \leq i \leq H}$ and $\sum_{i=1}^{H} a_i = 1$, $a_i \geq 0$.

Set the true distribution by

$$p(z|w_t^*) = \frac{1}{(2\pi)^{N/2}} \sum_{i=H+1}^{H+r} (-a_i^*) \exp(-\frac{\sum_{j=1}^{N}(z_j - b_{ij}^*)^2}{2}),$$

where $w_t^* = \{a_i^*, b_{ij}^*\}_{H+1 \leq i \leq H+r}$ and $\sum_{i=H+1}^{H+r} a_i^* = -1$, $a_i^* < 0$. (In order to simplify the followings, we use the values $a_i^* < 0$ not $a_i^* > 0$.)

Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$.

Let $A = (a_1, \cdots, a_H, a_{H+1}^*, \ldots, a_{H+r}^*)$, $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$,

$$B_I = (\prod_{j=1}^{N} b_{1j}^{\ell_j}, \prod_{j=1}^{N} b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^{N} b_{Hj}^{\ell_j}, \prod_{j=1}^{N} {b_{H+1,j}^*}^{\ell_j}, \cdots, \prod_{j=1}^{N} {b_{H+r,j}^*}^{\ell_j})^t,$$

and $B = (B_I)_{\ell_1 + \cdots + \ell_N = n, 1 \leq n \leq H+r}$ (t denotes the transpose).

Then the learning coefficient of the normal mixture model is the largest pole of

$$\int \Psi \;=\; \int_{||AB||^2 < 1} ||AB||^{2\xi} \prod_{i=1}^{H} \mathrm{d}a_i \prod_{i=1}^{H} \prod_{j=1}^{N} \mathrm{d}b_{ij}, \tag{3}$$

with $\sum_{j=1}^{H} a_j = 1$, $a_i \geq 0$ and $\sum_{j=H+1}^{H+r} a_j^* = -1$, $a_j^* < 0$ (Watanabe et al., 2004).

Note that we have the relations $\sum_{j=1}^{H} a_j = 1$, $a_i \geq 0$ and $\sum_{j=H+1}^{H+r} a_j^* = -1$, $a_j^* < 0$. We need to modify the function $||AB||^2$ for obtaining the largest pole of $\int \Psi$ by using Vandermonde matrix type singularities. The following theorem is available for such purpose.

Theorem 9 *Consider a sufficiently small neighborhood U of $w^* = \{a_i^*, b_{ij}^*\}_{1 \leq i \leq H}$, where $||AB|| = 0$ at w^*. Let $w = \{a_i, b_{ij}\}_{1 \leq i \leq H}$ be in U.*

Let each $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N})$, ..., $(b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ *be a different real vector in* $(b^{*}_{i1}, b^{*}_{i2}, \cdots, b^{*}_{iN})$ *for* $i = 1, \ldots, H + r$, *that is,*

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \cdots, b^{**}_{r'N})\ ;\ (b^{*}_{i1}, \cdots, b^{*}_{iN}), i = 1, \ldots, H\}.$$

Then r' *is uniquely determined and* $r' \geq r$ *by the assumption in Definition 4, and set* $(b^{**}_{i1}, \cdots, b^{**}_{iN}) = (b^{*}_{H+i,1}, \cdots, b^{*}_{H+i,N})$, *for* $1 \leq i \leq r$.

Assume that

$$(b^{*}_{i1}, \cdots, b^{*}_{iN}) = \begin{cases} (b^{**}_{11}, \cdots, b^{**}_{1N}), & 1 \leq i \leq H_1, \\ (b^{**}_{21}, \cdots, b^{**}_{2N}), & H_1 + 1 \leq i \leq H_1 + H_2, \\ \qquad \vdots & \\ (b^{**}_{r'1}, \cdots, b^{**}_{r'N}), & H_1 + \cdots + H_{r'-1} + 1 \leq i \leq H_1 + \cdots + H_{r'}, \end{cases}$$

and $H_1 + \cdots + H_{r'} = H$.

Then we have

$$\lambda_{w^*}(||AB||^2) = \sum_{\alpha=1}^{r'-1} \lambda_{w_1^{(\alpha)*}}\left(a_1^{(\alpha)2}\right)$$

$$+ \sum_{\alpha=1}^{r} \lambda_{w^{(\alpha)*}}(||(A_{1,H_\alpha - 1}, a^{*}_{H+\alpha}) B^{(1,1)}_{H_\alpha, N}||^2) + \sum_{\alpha=r+1}^{r'} \lambda_{w^{(\alpha)*}}(||A_{1,H_\alpha - 1} B^{(1,1)}_{H_\alpha - 1, N}||^2),$$

where $w_1^{(\alpha)*} = \begin{cases} a^{*}_{H_1+\cdots+H_{\alpha-1}+1} + \cdots + a^{*}_{H_1+\cdots+H_\alpha} + a^{*}_{H+\alpha}, & 1 \leq \alpha \leq r, \\ a^{*}_{H_1+\cdots+H_{\alpha-1}+1} + \cdots + a^{*}_{H_1+\cdots+H_\alpha}, & r+1 \leq \alpha \leq r'-1, \end{cases}$

$w^{(\alpha)*} = \{a_i^{(\alpha)*}, b_{ij}^{(\alpha)*}\}_{2 \leq i \leq H_\alpha} = \{a^{*}{}_{H_1+\cdots+H_{\alpha-1}+i}, 0\}_{2 \leq i \leq H_\alpha}$.

The proof for this theorem appears in Appendix A.

Theorem 9 is proved by using Theorem 7.

These Theorems 8 and 9 show that both learning coefficients for the three layered neural network and the normal mixture model are obtained by using Vandermonde matrix type singularities.

3.4 New bound values of learning coefficients

The next theorem gives new bound values of the largest pole for Vandermonde matrix type singularities. Let $\langle {k \atop l} \rangle = \dfrac{k!}{l!(k-l)!}$, for natural numbers k, l.

Theorem 10 *We use the same notations as in Theorem 7. We have the followings.*

Let $\mathrm{bound}_{\mathbf{i}} = \begin{cases} \frac{MH}{2}, & \text{if } mM \leq N-1, \text{ and } \mathbf{i}=1, \\ \frac{mM(H-1)+N}{2m}, & \text{if } mM \leq N-1, \text{ and } \mathbf{i}=2, \\ \frac{NH}{2m}, & \text{if } N \leq mM \leq m(N-1), \\ \frac{NH}{2m}, & \text{if } M \geq N, (N-1)(m-1) \geq 1, \\ \frac{2HN+Q(M(1+k)+(N-1)(2H-k-1))k}{4Qk+4m}, & \text{if } M \geq N, (N-1)(m-1)=0, \end{cases}$

for $\mathbf{i}=1,2$, *where* $k=\max\{i \in \mathbb{Z}; 2H \geq (Qi(i-1)+2mi)(M-N+1)\}$.

Also let $\mathrm{bound}_3 = \dfrac{NH+\sum_{i=0}^{k'-1} MQ(k'-i)\langle {}^{N+m+Qi-1}_{N-1} \rangle}{2m+2Qk'}$,

where $k'=\max\{i \in \mathbb{Z}; NH \geq M\sum_{i'=0}^{i-1}(m+Qi')\langle {}^{N+m+Qi'-1}_{N-1} \rangle\}$.

We have

$$\lambda_0(||A_{M,H}B_{H,N}^{(Q,m)}||^2) \leq \min\{\mathrm{bound}_1, \mathrm{bound}_3\},$$
$$\lambda_0(||(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q,m)}||^2) \leq \min\{\mathrm{bound}_2, \mathrm{bound}_3\}.$$

The proof appears in Appendix B.

Remark 2 *We have*

$$\min_{A_{M,H-1}^*} \lambda_{(A_{M,H-1}^*,0)}(||(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q,m)}||^2) = \lambda_0(||(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q,m)}||^2).$$

3.5 Exact values of learning coefficients

Theorem 11 Case 1 *Consider the case of* $H=1$. *We have*

$$\lambda_0(\|A_{M1}B_{1N}^{(Q,m)}\|^2) = \min\{\frac{M}{2}, \frac{N}{2m}\}, \text{ and its order } \theta_0(\|A_{M1}B_{1N}^{(Q,m)}\|^2) = \begin{cases} 1, & \text{if } mM \neq N, \\ 2, & \text{if } mM = N, \end{cases}$$

and

$$\lambda_0(\|\mathbf{a}^*B_{1N}^{(Q,m)}\|^2) = \frac{N}{2m}, \text{ and its order } \theta_0(\|A_{M1}B_{1N}^{(Q,m)}\|^2) = 1.$$

Case 2 *Consider the case of* $H=2$.

1. *Let* $\lambda = \lambda_0(\|A_{M2}B_{2N}^{(Q,m)}\|^2)$, *and* θ *its order.*

   Then we have

   (a) *If* $mM \leq N-1$ *then* $\lambda = M$ *and* $\theta = 1$.

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

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

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

(e) If $m = 2$, $N = 1, M = 1$, *then* $\lambda = \frac{1}{2}$ *and* $\theta = 2$.

(f) If $m = 2$, $N < mM$, $M > 1$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 1$.

(g) If $m \geq 2$, $N = mM$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 3$.

(h) If $m > 2$, $N < mM$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 1$.

2. *Let* $\lambda = \lambda_0(\|(A_{M1}, \mathbf{a}^*)B_{2N}^{(Q,m)}\|^2)$, *and* θ *its order.*

Then

(a) If $m \geq 2$, $mM \leq N - 1$ *then* $\lambda = \frac{mM+N}{2m}$ *and* $\theta = 1$.

(b) If $m = 1$, $N \geq M + Q + 1$ *then* $\lambda = \frac{N+M}{2}$ *and* $\theta = 1$.

(c) If $m = 1$, $N = M + Q$ *then* $\lambda = \frac{N+M}{2}$ *and* $\theta = 2$.

(d) If $m = 1$, $M + 1 \leq N \leq M + Q - 1$ *then* $\lambda = \frac{2N+Q(2N-1)}{2(Q+1)}$ *and* $\theta = 1$.

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

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

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

(h) If $m = 2$, $N = 1, M = 1$, *then* $\lambda = \frac{1}{2}$ *and* $\theta = 2$.

(i) If $m = 2$, $N < mM$, $M > 1$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 1$.

(j) If $m \geq 2$, $N = mM$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 2$.

(k) If $m > 2$, $N < mM$ *then* $\lambda = \frac{N}{m}$ *and* $\theta = 1$.

Its proof appears in Appendix C.

3.6 A learning coefficient for three layered neural network

By using Section 3.4 and 3.5, we have the followings.

Theorem 12 *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 r hidden units. Then their learning coefficients λ and θ in (1) and (2) are as follows.*

Let

$$\begin{aligned}\mathrm{bound}_0 &= \frac{Mr}{2} + \min_{H_0+\cdots+H_r=H}\{\frac{H_0N + (M(1+k_0) + (N-1)(2H_0 - k_0 - 1))k_0}{4k_0+2} \\ &\qquad + \sum_{\alpha=1}^{r} \frac{2H_\alpha N + (M(1+k_\alpha) + (N-1)(2H_\alpha - k_\alpha - 1))k_\alpha}{4k_\alpha + 4}\} \\ &= \frac{Mr}{2} + \frac{H(N-1)}{2} + \min\{\frac{r}{2} + \frac{H - r + (M-N+1)(k' + k'^2)}{4k'+2}, \\ &\qquad \frac{r-1}{2} + \frac{2(H-r+1) + (M-N+1)(k'' + k''^2)}{4k''+4}\},\end{aligned}$$

where $k_0 = \max\{i \in \mathbb{Z}; H_0 \geq i^2(M-N+1)\}$, $k_\alpha = \max\{i \in \mathbb{Z}; 2H_\alpha \geq (i^2+i)(M-N+1)\}$ *for* $\alpha \geq 1$, $k' = \max\{i \in \mathbb{Z}; H - r \geq i^2(M-N+1)\}$ *and* $k'' = \max\{i \in \mathbb{Z}; 2(H-r-1) \geq (i^2+i)(M-N+1)\}$.

Also let $\mathrm{bound}_3 = \dfrac{N(H-r) + \sum_{i=0}^{k'-1} 2M(k'-i)\langle {N+2i \atop N-1} \rangle}{2 + 4k'}$,
where $k' = \max\{i \in \mathbb{Z}; N(H-r) \geq M\sum_{i'=0}^{i-1}(1+2i')\langle {N+2i' \atop N-1} \rangle\}$.

If $M < N$, *then* $\lambda \leq \min\{\dfrac{MH+Nr}{2}, \dfrac{(M+N)r}{2} + \mathrm{bound}_3\}$.

If $M \geq N$, *then* $\lambda \leq \min\{\mathrm{bound}_0, \mathrm{bound}_3\}$.

Especially,

1. $H - r = 0$: $\lambda = r(\frac{M+N}{2}), \theta = 1$.

2. $H = 1, r = 0$: $\lambda = \min\{\frac{M}{2}, \frac{N}{2}\}, \theta = \begin{cases} 1, & \text{if } M \neq N, \\ 2, & \text{if } M = N. \end{cases}$

3. $H - r = 1, r \geq 1$:

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

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

   (c) If $N = M$ then $\lambda = (r-1)(\frac{M+N}{2}) + \frac{3M+3N-1}{4}$ and $\theta = 1$.

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

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

4. $H = 2, r = 0$:

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

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

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

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

Our results are tighter than Watanabe's bounds (Watanabe (2001c)).

Its proof appears in Appendix D.

Remark 3 *See Lemma 17 in Appendix D about*
$\min\{\frac{r}{2}+\frac{H-r+(M-N+1)(k'+k'^2)}{4k'+2}, \frac{r-1}{2}+\frac{2(H-r+1)+(M-N+1)(k''+k''^2)}{4k''+4}\}$,
in detail.

3.7 A learning coefficient for normal mixture model

By using section 3.4, 3.5 similarly and by the theorem of dimension one $(N=1)$ in the paper (Aoyagi (2010a)), we have the followings.

Theorem 13 *Consider normal mixture models with* H *peaks and the true distribution with* r *peaks. Then their learning coefficients* λ *and* θ *in (1) and (2) are as follows.*

Let $\mathrm{bound}_3 = \dfrac{N(H-r+1)+\sum_{i=0}^{k'-1}(k'-i)\langle {}^{N+i}_{N-1}\rangle}{2+2k'}$,
where $k' = \max\{i \in \mathbb{Z}; N(H-r+1) \geq \sum_{i'=0}^{i-1}(1+i')\langle {}^{N+i'}_{N-1}\rangle\}$.

If $1 < N$, *then* $\lambda \leq \min\{\dfrac{H-1+Nr}{2}, \dfrac{(N+1)(r-1)}{2} + \mathrm{bound}_3\}$.

If $N = 1$, *then*

$$\lambda = r-1+\frac{i+i^2+2(H-(r-1))}{4(i+1)}, \theta = \begin{cases} 1, & \text{if } i^2+i < 2(H-(r-1)), \\ 2, & \text{if } i^2+i = 2(H-(r-1)), \end{cases}$$

where $i = \max\{j \in \mathbb{Z}\ ;\ j^2+j \leq 2(H-(r-1))\}$.

Especially,

1. $H-r=0$ *:* $\lambda = \frac{r-1+rN}{2}, \theta = 1$.

2. $H-r=1$ *:*

   (a) If $N > 2$, $\lambda = \frac{r(N+1)}{2}$, $\theta = 1$.

   (b) If $N = 2$, $\lambda = \frac{3r}{2}$, $\theta = 2$.

   (c) If $N = 1$, $\lambda = \frac{3}{4} + r - 1$, $\theta = 1$.

4 Numerical Analysis

In order to confirm our theoretical results, we simulated Bayesian estimation of normal mixture models, and compared the theoretical value of the coefficient λ and its numerical value.

4.1 Numerical Setting

We use the same notations in Section 3.3.

We experimented on N dimensional training data, where $N = 1, \cdots, 4$. Set by $(r, H) = (1, 2), (2, 3)$, where r is the number of peaks for a true distribution and H is the one for a learner distribution. Consequently, we simulated the eight cases of Bayesian estimation of normal mixture models. Then, the theoretical value of their largest pole $-\lambda$ of $J(\xi)$ and its order θ can be calculated from Theorem 13, which are shown in Table 1.

	N	r	H	λ	θ		N	r	H	λ	θ
case 1	1	1	2	0.75	1	case 5	1	2	3	1.75	1
case 2	2	1	2	1.50	2	case 6	2	2	3	3.00	2
case 3	3	1	2	2.00	1	case 7	3	2	3	4.00	1
case 4	4	1	2	2.50	1	case 8	4	2	3	5.00	1

Table 1: The theoretical value of the largest poles $-\lambda$ of $J(\xi)$ and its order θ for our experimental conditions.

The number n of training data was set 200, 400, 600, 800 and 1000. The *a priori* distribution $\psi(w)$ was defined by the uniform distribution for the mixing ratio $-a_i^*$ and the N-dimensional normal distribution whose mean and variance are respectively 0.0 and 10.0 for each mean b_i^* of the peaks.

In Bayesian estimation, it is necessary for calculating the expectation over the *a posteriori* distribution. For this purpose, in our experiments, we used the exchange Monte Carlo (EMC) method for sampling from the *a posteriori* distribution. The exchange Monte Carlo (EMC) method, one of the Markov chain Monte Carlo methods, is found to be appropriate for sampling from the *a posteriori* distribution in non-regular

(a) true distribution (b) predictive distribution

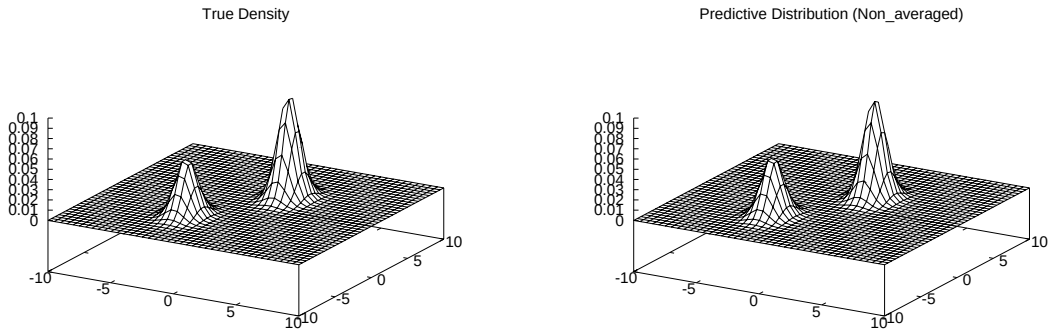


Figure 3: Examples of a true distribution and of the corresponding numerical predictive distribution in the case that $N = 2, r = 2, H = 3$.

statistical models such as neural networks and normal mixture models (Nagata, 2008a). The detailed setting of EMC method in this study was similarly set as the setting in the paper (Nagata, 2008c).

In this experiment, we generate the $T = 4000$ samples of parameter $\{w_t\}(t = 1, \cdots, T)$ by the EMC method, and calculate the predictive distribution by using these samples as follows,

$$p(x|(x)^n) \approx \frac{1}{T} \sum_{t=1}^{T} p(x|w_t).$$

Examples of a true distribution and of the corresponding numerical predictive distribution are shown in Figure 3 in the case that $N = 2, r = 2, H = 3$. In order to evaluate the experimental value of coefficient λ, we calculated the generalization error from the $n' = 10000$ test data $\{x'_i\}(i = 1, \cdots, n')$ as follows,

$$G(n) \approx E_n \left\{ \frac{1}{n'} \sum_{i=1}^{n'} \log \frac{q(x'_i)}{p(x'_i|(x)^n)} \right\}.$$

We can evaluate the validity of our results by comparing the asymptotic form of generalization error shown in Eq.(2) and the experimental one.

4.2 Numerical Results

Figure 4 shows the numerical results in the case that $r = 1$ and $H = 2$. In each figure, the horizontal axis is the number n of training data, and the vertical one the

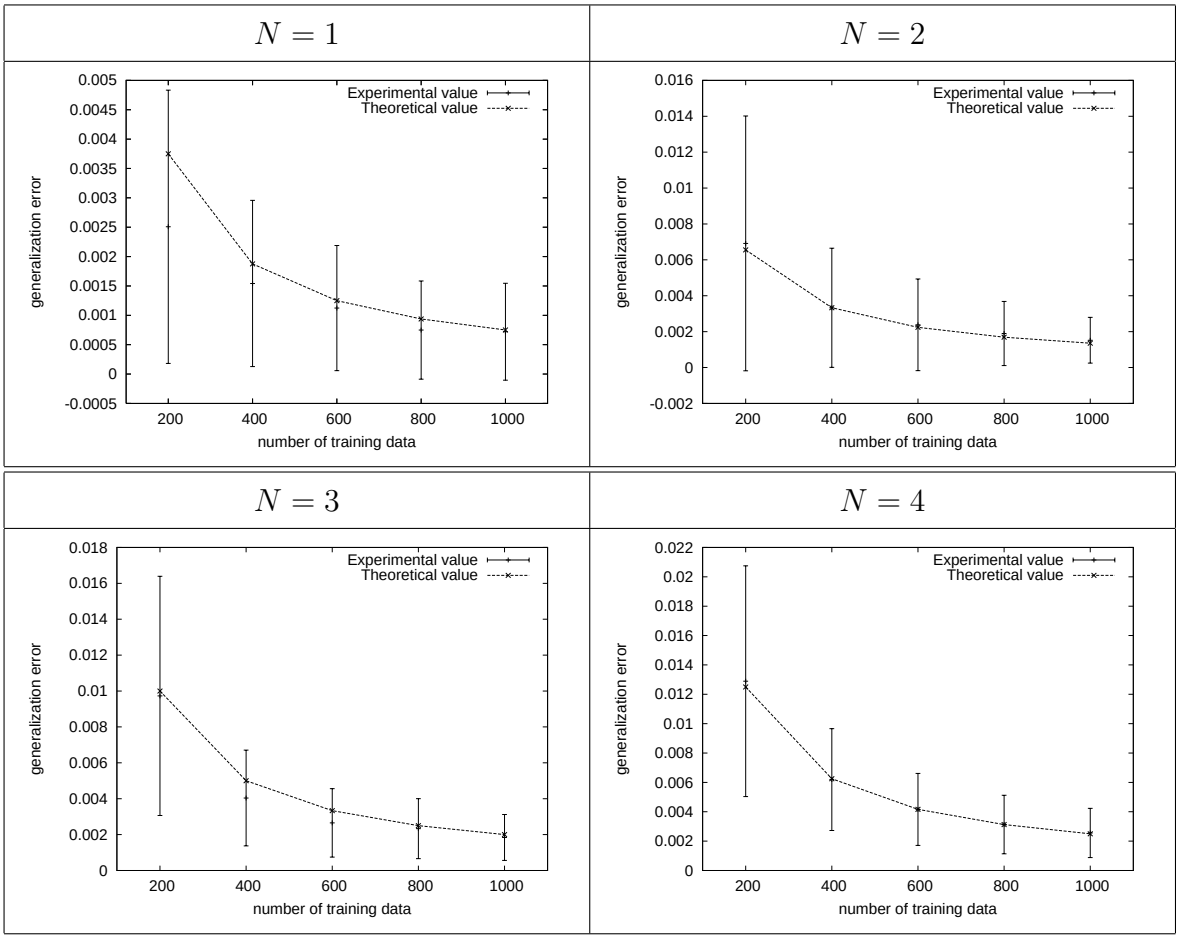


Figure 4: Comparison between the theoretical generalization error and the numerical one in the case that $r = 1$ and $H = 2$. Dashed lines in these figures indicate the theoretical value, and error bars "average $\pm$ standard deviation" of 100 numerical values.

value of generalization error. The dashed lines in these figures indicate the theoretical values of generalization errors, which is calculated from Eq.(2). The error bars indicate "average $\pm$ standard deviation" for all 100 sets of training data. In the same way, we also simulated in the case that $r = 2$ and $H = 3$. Figure 5 shows its numerical results.

According to these results, the experimental value of generalization error converges to the theoretical one as the number n of training data increases. Some differences between the theoretical results and the experimental results may be influenced by the lower term of the generalization error in Eq.(2).

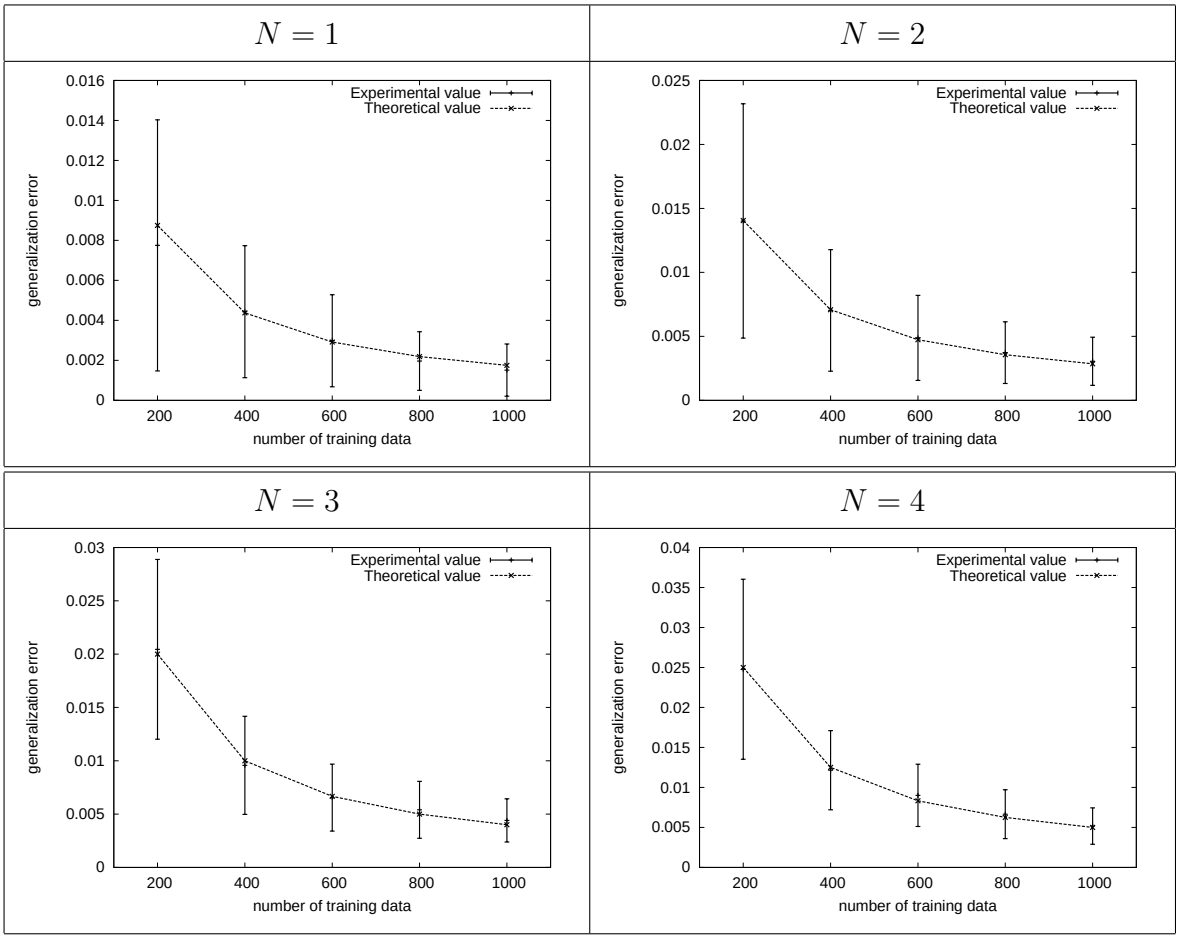


Figure 5: Comparison between the theoretical generalization error and the numerical one in the case that $r = 2$ and $H = 3$.

5 Conclusion

In this paper, we consider learning coefficients for Vandermonde matrix type singularities. Theorems 8 and 9 show that learning coefficients for three layered neural networks and for normal mixture models with identity matrix variances are obtained by the same type of singularities, i.e., Vandermonde matrix type singularities. Yamazaki et al. (2010) shows that learning coefficients for mixtures of binomial distributions are also obtained by Vandermonde matrix type singularities. Vandermonde matrix type is a generic one in neural computation, so these facts seem to imply that Vandermonde matrix type singularities are essential for learning theory.

We also show new tight bound values of learning coefficients for Vandermonde matrix type singularities (Theorem 10) and the explicit values in some conditions (Theorem

11).

By applying our results, we consider the learning coefficients of three layered neural networks with certain number of hidden units (Theorem 12), and normal mixture models. with certain number of peaks in Bayesian estimation (Theorem 13).

Numerical results in Section 4 confirm our theoretical results.

Our future research aims to improve our method to obtain the generalization error for both models, by using the bound values in Theorem 10.

We believe that extending our results would provide a mathematical foundation for the analysis of various multi-layered models.

This study involves applying techniques of algebraic geometry to learning theory and it seems that we can contribute to the development of both these fields in the future.

The application of our results is as follows. The results of this paper introduce a mathematical measure of preciseness for numerical calculations such as the Markov Chain Monte Carlo. In the paper (Nagata, 2008a), mathematical foundation for analyzing and developing the precision of the MCMC method is constructed by using the theoretical values of marginal likelihoods. Moreover, the paper (Nagata, 2008b) studied the setting of temperatures for the exchange MCMC method and proved the mathematical relation between the symmetrized Kullback function and the exchange ratio, from which an optimal setting of temperatures could be devised. Our theoretical results will be helpful in these numerical experiments.

Furthermore, these values have been compared with those of the generalization error of a localized Bayes estimation (Takamatsu et al., 2005).

Acknowledgement This research was supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 22540224 and Grant-in-aid for Basic Science Research of Nihon University.

Appendix A

The proof for Theorem 9 is as follows.

Let $A = (a_1, \ldots, a_H, a^*_{H+1}, \ldots, a^*_{H+r})$, $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$,

$$B_I = \left(\prod_{j=1}^N b_{1j}^{\ell_j}, \prod_{j=1}^N b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^N b_{Hj}^{\ell_j}, \prod_{j=1}^N {b^*_{H+1,j}}^{\ell_j}, \cdots, \prod_{j=1}^N {b^*_{H+r,j}}^{\ell_j} \right)^t,$$

and $B = (B_I)_{\ell_1+\cdots+\ell_N=n, 1\le n\le H+r}$.

Then we have

$$AB_I = (a_1, \ldots, a_{H-1}, a^*_{H+1}, \ldots, a^*_{H+r}) \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H-1,j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \prod_{j=1}^N {b^*_{H+1,j}}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N {b^*_{H+r,j}}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix},$$

by using $\sum_{i=1}^H a_i = 1$, $\sum_{i=H+1}^r a_i^* = -1$.

Let $B'_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H-1,j}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \prod_{j=1}^N {b^*_{H+1,j}}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N {b^*_{H+r,j}}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$ and $B' = (B'_I)_{\ell_1+\cdots+\ell_N=n, 1\le n\le H+r}$.

Since

$$\prod_{j=1}^N b_{kj}^{\ell_j} - \prod_{j=1}^N b_{Hj}^{\ell_j} = (\prod_{j=1}^{N-1} b_{kj}^{\ell_j}) b_{kN}^{\ell_N - 1}(b_{kN} - b_{HN}) + b_{HN}((\prod_{j=1}^{N-1} b_{kj}^{\ell_j}) b_{kN}^{\ell_N-1} - (\prod_{j=1}^{N-1} b_{Hj}^{\ell_j}) b_{HN}^{\ell_N-1}),$$

$$(\prod_{j=1}^{N-1} b_{kj}^{\ell_j}) b_{kN}^{\ell_N-1}(b_{kN} - b_{HN}) = (\prod_{j=1}^{N-1} b_{kj}^{\ell_j}) b_{kN}^{\ell_N-2}(b_{kN} - b_{HN})^2 + b_{HN}(\prod_{j=1}^{N-1} b_{kj}^{\ell_j}) b_{kN}^{\ell_N-2}(b_{kN} - b_{HN}),$$

and so on, we have a regular matrix R such that $B'R = (B''_I)_I$ where

$$B''_I = \begin{pmatrix} \prod_{j=1}^N (b_{1j} - b_{Hj})^{\ell_j} \\ \prod_{j=1}^N (b_{2j} - b_{Hj})^{\ell_j} \\ \vdots \\ \prod_{j=1}^N (b_{H-1,j} - b_{Hj})^{\ell_j} \\ \prod_{j=1}^N (b^*_{H+1,j} - b_{Hj})^{\ell_j} \\ \vdots \\ \prod_{j=1}^N (b^*_{H+r,j} - b_{Hj})^{\ell_j} \end{pmatrix}.$$

Set

$$
\begin{aligned}
&(a_1'^{(1)}, \ldots, a_{H_1}'^{(1)}) = (a_1, \ldots, a_{H_1}),\\
&(a_1'^{(2)}, \ldots, a_{H_2}'^{(2)}) = (a_{H_1+1}, \ldots, a_{H_1+H_2}),\\
&\qquad \vdots\\
&(a_1'^{(r')}, \ldots, a_{H_{r'}}'^{(r')}) = (a_{H_1+\cdots+H_{r'-1}+1}, \ldots, a_{H_1+\cdots+H_{r'}}).
\end{aligned}
$$

and

$$
\begin{aligned}
&(b_{1j}'^{(1)}, \ldots, b_{H_1 j}'^{(1)}) = (b_{1j} - b_{Hj}, \ldots, b_{H_1 j} - b_{Hj}),\\
&(b_{1j}'^{(2)}, \ldots, b_{H_2 j}'^{(2)}) = (b_{H_1+1,j} - b_{Hj}, \ldots, b_{H_1+H_2,j} - b_{Hj}),\\
&\qquad \vdots\\
&(b_{1j}'^{(r')}, \ldots, b_{H_{r'}-1,j}'^{(r')}) = (b_{H_1+\cdots+H_{r'-1}+1,j} - b_{Hj}, \ldots, b_{H_1+\cdots+H_{r'}-1,j} - b_{Hj}).
\end{aligned}
$$

for $1 \le j \le N$.

Let $A'^{(\alpha)} = \begin{cases} (a_1'^{(\alpha)}, a_2'^{(\alpha)}, \cdots, a_{H_\alpha}'^{(\alpha)}, a^*_{H+\alpha}), & \text{for } 1 \le \alpha \le r, \alpha \le r'-1 \\ (a_1'^{(\alpha)}, a_2'^{(\alpha)}, \cdots, a_{H_\alpha}'^{(\alpha)}, 0), & \text{for } H+1 \le \alpha \le r'-1, \end{cases}$

$A'^{(r')} = \begin{cases} (a_1'^{(r')}, a_2'^{(r')}, \cdots, a_{H_{r'}-1}'^{(r')}, a^*_{H+r'}), & \text{if } r' = r, \\ (a_1'^{(r')}, a_2'^{(r')}, \cdots, a_{H_{r'}-1}'^{(r')}, 0), & \text{if } r' > r, \end{cases}$ and

$B'^{(\alpha)} = (B_I'^{(\alpha)})_{\ell_1+\ell_2+\cdots+\ell_N = n, 1 \le n \le H_{r'}}$, where

$$
B_I'^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}'^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}'^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha,j}'^{(\alpha)\ \ell_j} \\ \prod_{j=1}^N (b^{**}_{\alpha j} - b_{Hj})^{\ell_j} \end{pmatrix}, 1 \le \alpha \le r'-1, B_I'^{(r')} = \begin{pmatrix} \prod_{j=1}^N b_{1j}'^{(r')\ell_j} \\ \prod_{j=1}^N b_{2j}'^{(r')\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_{r'}-1,j}'^{(r')\ \ell_j} \\ \prod_{j=1}^N (b^{**}_{r'j} - b_{Hj})^{\ell_j} \end{pmatrix}.
$$

By Theorem 7, we only need to consider the case $\sum_{\alpha=1}^r ||A'^{(\alpha)} B'^{(\alpha)}||^2$ instead of $||AB||^2$.

Since we assume that $(b^{**}_{\alpha 1}, \cdots, b^{**}_{\alpha N}) \ne (b^*_{H1}, \cdots, b^*_{HN}) = (b^{**}_{r'1}, \cdots, b^{**}_{r'N})$ for $1 \le \alpha \le r'-1$, we may set $b^{**}_{\alpha 1} - b_{H1} \ne 0$.

Since $B_I'^{(\alpha)}\frac{1}{(b_{\alpha 1}^{**}-b_{H1})^{\ell_1}} = \begin{pmatrix} \left(\frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{1j}'^{(\alpha)^{\ell_j}} \\ \left(\frac{b_{21}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{2j}'^{(\alpha)^{\ell_j}} \\ \vdots \\ \left(\frac{b_{H_\alpha 1}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{H_\alpha j}'^{(\alpha)\,\ell_j} \\ \prod_{j=2}^{N}(b_{\alpha j}^{**}-b_{Hj})^{\ell_j} \end{pmatrix}$, there exists a regular matrix R'' such that $B'^{(\alpha)}R'' = B''^{(\alpha)} = (\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, B_I''^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, n\in\mathbb{N}}$, where

$$B_I''^{(\alpha)} = \begin{pmatrix} 0 \\ \left(\frac{b_{21}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{2j}'^{(\alpha)^{\ell_j}} - \left(\frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{1j}'^{(\alpha)^{\ell_j}} \\ \vdots \\ \left(\frac{b_{H_\alpha 1}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{H_\alpha j}'^{(\alpha)\,\ell_j} - \left(\frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{1j}'^{(\alpha)^{\ell_j}} \\ \prod_{j=2}^{N}(b_{\alpha j}^{**}-b_{Hj})^{\ell_j} - \left(\frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N} b_{1j}'^{(\alpha)^{\ell_j}} \end{pmatrix}.$$

We have, therefore, a regular matrix R''' such that

$$B''^{(\alpha)}R''' = B'''^{(\alpha)} = (\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, B_I''^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, n\in\mathbb{N}},$$

where

$$B_I'''^{(\alpha)} = \begin{pmatrix} 0 \\ \left(\frac{b_{21}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}} - \frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N}(b_{2j}'^{(\alpha)} - b_{1j}'^{(\alpha)})^{\ell_j} \\ \vdots \\ \left(\frac{b_{H_\alpha 1}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}} - \frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N}(b_{H_\alpha j}'^{(\alpha)} - b_{1j}'^{(\alpha)})^{\ell_j} \\ \left(1 - \frac{b_{11}'^{(\alpha)}}{b_{\alpha 1}^{**}-b_{H1}}\right)^{\ell_1}\prod_{j=2}^{N}(b_{\alpha j}^{**} - b_{Hj} - b_{1j}'^{(\alpha)})^{\ell_j} \end{pmatrix}.$$

Set $a_1^{(\alpha)} = \begin{cases} a_1'^{(\alpha)} + a_2'^{(\alpha)} + \cdots + a_{H_\alpha}'^{(\alpha)} + a_i^*, & \text{for } 1 \le \alpha \le r, \alpha \le r'-1, \\ a_1'^{(\alpha)} + a_2'^{(\alpha)} + \cdots + a_{H_\alpha}'^{(\alpha)}, & \text{for } H+1 \le \alpha \le r'-1, \end{cases}$

$$A^{(\alpha)} = \begin{cases} (a_2^{(\alpha)}, a_3^{(\alpha)}, \cdots, a_{H_\alpha}^{(\alpha)}, a_{H+\alpha}^*) = (a_2^{\prime(\alpha)}, a_3^{\prime(\alpha)}, \cdots, a_{H_\alpha}^{\prime(\alpha)}, a_{H+\alpha}^*), \\ \qquad \text{for } 1 \leq \alpha \leq r, \alpha \leq r' - 1 \\ (a_2^{(\alpha)}, a_3^{(\alpha)}, \cdots, a_{H_\alpha}^{(\alpha)}) = (a_2^{\prime(\alpha)}, a_3^{\prime(\alpha)}, \cdots, a_{H_\alpha}^{\prime(\alpha)}), \\ \qquad \text{for } H+1 \leq \alpha \leq r' - 1, \end{cases}$$

$$A^{(r')} = \begin{cases} (a_2^{(r')}, a_3^{(r')}, \cdots, a_{H_{r'}}^{(r')}, a_{H+\alpha}^*) = (a_1^{\prime(r')}, a_2^{\prime(r')}, \cdots, a_{H_\alpha - 1}^{\prime(r')}, a_{H+\alpha}^*), & \text{if } r = r', \\ (a_2^{(r')}, a_3^{(r')}, \cdots, a_{H_{r'}}^{(r')}) = (a_1^{\prime(r')}, a_2^{\prime(r')}, \cdots, a_{H_\alpha - 1}^{\prime(r')}), & \text{if } r < r', \end{cases}$$

and $\begin{pmatrix} b_{11}^{(\alpha)} & b_{12}^{(\alpha)} & \cdots & b_{1N}^{(\alpha)} \\ & & \vdots & \\ b_{H_\alpha 1}^{(\alpha)} & b_{H_\alpha 2}^{(\alpha)} & \cdots & b_{H_\alpha N}^{(\alpha)} \end{pmatrix}$

$$= \begin{pmatrix} \frac{b_{21}^{\prime(\alpha)}}{b_{\alpha 1}^{**} - b_{H1}} - \frac{b_{11}^{\prime(\alpha)}}{b_{\alpha 1}^{**} - b_{H1}} & b_{22}^{\prime(\alpha)} - b_{12}^{\prime(\alpha)} & \cdots & b_{2N}^{\prime(\alpha)} - b_{1N}^{\prime(\alpha)} \\ & & \vdots & \\ \frac{b_{H_\alpha 1}^{\prime(\alpha)}}{b_{\alpha 1}^{**} - b_{H1}} - \frac{b_{11}^{\prime(\alpha)}}{b_{\alpha 1}^{**} - b_{H1}} & b_{H_\alpha 2}^{\prime(\alpha)} - b_{12}^{\prime(\alpha)} & \cdots & b_{H_\alpha N}^{\prime(\alpha)} - b_{1N}^{\prime(\alpha)} \\ 1 - \frac{b_{11}^{\prime(\alpha)}}{b_{\alpha 1}^{**} - b_{H1}} & b_{\alpha 2}^{**} - b_{H2} - b_{12}^{\prime(\alpha)} & \cdots & b_{\alpha N}^{**} - b_{HN} - b_{1N}^{\prime(\alpha)} \end{pmatrix}.$$

$$\begin{pmatrix} b_{11}^{(r')} & b_{12}^{(r')} & \cdots & b_{1N}^{(r')} \\ & & \vdots & \\ b_{H_{r'}1}^{(r')} & b_{H_{r'}2}^{(r')} & \cdots & b_{H_{r'}N}^{(r')} \end{pmatrix} = \begin{pmatrix} b_{11}^{\prime(r')} & b_{12}^{\prime(r')} & \cdots & b_{1N}^{\prime(r')} \\ & & \vdots & \\ b_{H_{r'}-1,1}^{\prime(r')} & b_{H_{r'}-1,2}^{\prime(r')} & \cdots & b_{H_{r'}-1,N}^{\prime(r')} \\ b_{r'1}^{**} - b_{H1} & b_{r'2}^{**} - b_{H2} & \cdots & b_{r'N}^{**} - b_{HN} \end{pmatrix}.$$

Then we have Theorem 9.

Q.E.D.

Lemma 14 *Let* $I = (\ell_1, \ldots, \ell_N) \in (\mathbb{N} \cup \{0\})^N$, $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$ *and*

$B = (B_I)_{\ell_1 + \cdots + \ell_N = Qn+m, 0 \leq n \leq H-1.}$.

Also let $B'_I = \begin{pmatrix} 0 \\ b_{21}^{\ell_1} \prod_{j=2}^N {b'_{2j}}^{\ell_j} \\ \vdots \\ b_{H1}^{\ell_1} \prod_{j=2}^N {b'_{Hj}}^{\ell_j} \end{pmatrix}$ *for* $\ell_2 + \cdots + \ell_N \neq 0$, $B'_{(\ell_1,0,\ldots,0)} = B_{(\ell_1,0,\ldots,0)}$

and $B' = (B'_I)_{\ell_1 + \cdots + \ell_N = Qn+m, 0 \leq n \leq H-1, \ell_2 + \cdots + \ell_N \neq 0}$.

Set $b'_{ki} = b_{ki} - b_{k1} b_{1i} / b_{11}$ *for* $b_{11} \neq 0$ *and* $k, i \geq 2$.

Then we have, for a regular matrix R, $BR = B'$.

(Proof)

We have $\begin{pmatrix} \prod_{j=1}^{N} b_{1j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^{N} b_{Hj}^{\ell_j} \end{pmatrix} = \begin{pmatrix} b_{11}^{\ell_1} \prod_{j=2}^{N} b_{1j}^{\ell_j} \\ b_{21}^{\ell_1} \prod_{j=2}^{N} b_{2j}^{\ell_j} \\ \vdots \\ b_{H1}^{\ell_1} \prod_{j=2}^{N} b_{Hj}^{\ell_j} \end{pmatrix} = \begin{pmatrix} b_{11}^{\ell_1} \prod_{j=2}^{N} b_{1j}^{\ell_j} \\ b_{21}^{\ell_1} \prod_{j=2}^{N} (b'_{2j} + b_{21}b_{1j})^{\ell_j} \\ \vdots \\ b_{H1}^{\ell_1} \prod_{j=2}^{N} (b'_{Hj} + b_{H1}b_{1j})^{\ell_j} \end{pmatrix}$.

Also we have $\begin{pmatrix} b_{11}^{\ell_1} \prod_{j=2}^{N} b_{1j}^{\ell_j} \\ b_{21}^{\ell_1} \prod_{j=2}^{N} (b_{21}b_{1j})^{\ell_j} \\ \vdots \\ b_{H1}^{\ell_1} \prod_{j=2}^{N} (b_{H1}b_{1j})^{\ell_j} \end{pmatrix} = \begin{pmatrix} b_{11}^{\ell_1} \prod_{j=2}^{N} b_{1j}^{\ell_j} \\ b_{21}^{\sum \ell_j} \prod_{j=2}^{N} b_{1j}^{\ell_j} \\ \vdots \\ b_{H1}^{\sum \ell_j} \prod_{j=2}^{N} b_{1j}^{\ell_j} \end{pmatrix}$ and

$$\begin{pmatrix} 0 \\ b_{21}^{\ell_1} \prod_{j=2}^{N} {b'_{2j}}^{\ell'_j} (b_{21}b_{1j})^{\ell_j - \ell'_j} \\ \vdots \\ b_{H1}^{\ell_1} \prod_{j=2}^{N} {b'_{Hj}}^{\ell'_j} (b_{H1}b_{1j})^{\ell_j - \ell'_j} \end{pmatrix} = \begin{pmatrix} 0 \\ b_{21}^{\ell_1 + \sum(\ell_j - \ell'_j)} \prod_{j=2}^{N} {b'_{2j}}^{\ell'_j} \prod_{j=2}^{N} b_{1j}^{\ell_j - \ell'_j} \\ \vdots \\ b_{H1}^{\ell_j + \sum(\ell_j - \ell'_j)} \prod_{j=2}^{N} {b'_{Hj}}^{\ell'_j} \prod_{j=2}^{N} b_{1j}^{\ell_j - \ell'_j} \end{pmatrix}.$$

Q.E.D.

Appendix B

In this section, we give the proof of Theorem 10

Assume that $H_0 = H$.

Let

$$\Psi = ||AB||^2, \tag{4}$$

$\phi = \mathrm{d}a\mathrm{d}b$, V be a sufficiently small neighborhood of 0 and $J(\xi) = \int_V \Psi^{\xi} \phi$.

Bound values : bound$_1$ and bound$_2$

By using a blowing up process together with an inductive method in algebraic geometry (Watanabe, 2009; Fulton, 1993), we show that we have the following functions (5) and (6) below.

Let

$$\phi = \prod_{i=1}^{H'} v_i^{T_i} \mathrm{d}v\mathrm{d}a\mathrm{d}b, \tag{5}$$

where

$$\begin{aligned} T_i &= mM(i-1)+(H-i+1)N+Q(iM+(H-i)(N-1)) \\ &\quad +Q((i+1)M+(H-i-1)(N-1))+\cdots+Q(H'M+(H-H')(N-1))-1 \\ &= mM(i-1)+(H-i+1)N \\ &\quad +Q(M(i+H')+(N-1)(2H-H'-i))(H'-i+1)/2-1, \end{aligned}$$

and

$$\begin{aligned} \Psi &= (v_1^{QH'+m}v_2^{Q(H'-1)+m}\cdots v_{H'}^{Q+m})^2||A_1||^2 \\ &\quad +\sum_{\ell_1=Qn+m,n\geq H'}(v_1^{\ell_1}v_2^{\ell_1-Q}\cdots v_{H'}^{\ell_1-(H'-1)Q})^2||A_2{f'}_{\ell_1,0,\cdots,0}||^2 \\ &\quad +\sum_{\substack{\ell_1+\cdots+\ell_N=Qn+m,\\ \ell_2+\cdots+\ell_N>0}}(v_1^{\ell_1+(QH'+1)(\ell_2+\cdots+\ell_N)}v_2^{\ell_1+(Q(H'-1)+1)(\ell_2+\cdots+\ell_N)}\cdots v_{H'}^{\ell_1+(Q+1)(\ell_2+\cdots+\ell_N)})^2 \\ &\quad \times||A_2f_{\ell_1,\ell_2,\cdots,\ell_N}||^2, \end{aligned} \tag{6}$$

where $A_1=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H'} \\ & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH'} \end{pmatrix}$, $A_2=\begin{pmatrix} a_{1,H'+1} & a_{1,H'+2} & \cdots & a_{1H} \\ & \vdots & & \\ a_{M,H'+1} & a_{M,H'+2} & \cdots & a_{MH} \end{pmatrix}$,

$$\begin{aligned} {f'}_{Qn+m,0,\cdots,0} &= \begin{pmatrix} b_{H'+1,1}^{m+Q(n-H')}((b_{H'+1,1}v_2\cdots v_{H'})^Q-1)((b_{H'+1,1}v_3\cdots v_{H'})^Q-1)\cdots((b_{H'+1,1})^Q-1) \\ \vdots \\ b_{H1}^{m+Q(n-H')}((b_{H1}v_2\cdots v_{H'})^Q-1)((b_{H1}v_3\cdots v_{H'})^Q-1)\cdots((b_{H1})^Q-1) \end{pmatrix} \end{aligned}$$

and

$$f_{\ell_1,\ell_2,\cdots,\ell_N}=\begin{pmatrix} \prod_{j=1}^N b_{H'+1,j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H,j}^{\ell_j} \end{pmatrix}.$$

Construct the blow-up of the function (4) along the submanifold $\{b_{ij}=0, 1\leq i\leq H, 1\leq j\leq N\}$. Let $b_{11}=v_1$, $b_{ij}=v_1b'_{ij}$, $(i,j)\neq(1,1)$.

Set $b''_{ij}=b'_{ij}-b'_{i1}b'_{1j}$ for $i\geq 2$ and $a'_{i1}=a_{i1}+a_{i2}b_{21}^m+a_{i3}b_{31}^m+\cdots+a_{iH}b_{H1}^m$ for $1\leq i\leq M$. By using Lemma 2 in Section 3.1 and setting $a_{i1}=a'_{i1}$, $b_{ij}=b''_{ij}$ again, we need to consider the functions

$$\phi = v_1^{NH-1}\mathrm{d}v\mathrm{d}a\mathrm{d}b, \tag{7}$$

and

$$\begin{aligned}\Psi &= (v_1^m)^2||A_1||^2 \\ &+ \sum_{\ell_1=Qn+m, n\geq 1} (v_1^{\ell_1})^2||A_2 {f'}_{\ell_1,0,\cdots,0}||^2 \\ &+ \sum_{\substack{\ell_1+\cdots+\ell_N=Qn+m,\\ \ell_2+\cdots+\ell_N>0}} (v_1^{\ell_1+\ell_2+\cdots+\ell_N})^2||A_2 f_{\ell_1,\ell_2,\cdots,\ell_N}||^2,\end{aligned} \tag{8}$$

where $A_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{M1} \end{pmatrix}$, $A_2 = \begin{pmatrix} a_{12} & a_{13} & \cdots & a_{1H} \\ & \vdots & & \\ a_{M2} & a_{M3} & \cdots & a_{MH} \end{pmatrix}$,

$f'_{Qn+m,0,\cdots,0} = \begin{pmatrix} b_{21}^{m+Q(n-1)}(b_{21}^Q - 1) \\ \vdots \\ b_{H1}^{m+Q(n-1)}(b_{H1}^Q - 1) \end{pmatrix}$ and $f_{\ell_1,\ell_2,\cdots,\ell_N} = \begin{pmatrix} \prod_{j=1}^N b_{2j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$.

We construct the blow-up of the above function (8) along the submanifold $\{v_1 = 0, a_{k1} = 0, b_{ij} = 0, 1 \leq k \leq M, 2 \leq i \leq H, 2 \leq j \leq N\}$ Q times. Let $a_{k1} = v_1^Q a'_{k1}$, $b_{ij} = v_1^Q b'_{ij}$, $1 \leq k \leq M, 2 \leq i \leq H, 2 \leq j \leq N$.

We have the $J(\xi)$'s poles $\frac{NH+p(M+(H-1)(N-1))}{2(m+p)}$ for $0 \leq p \leq Q$ and the functions Eqs. (5) and (6) with $H' = 1$, by setting $a_{i1} = a'_{i1}$, $b_{ij} = b'_{ij}$.

Assume Eqs. (5) and (6). Construct the blow-up of function (6) along the submanifold $\{b_{ij} = 0, H' + 1 \leq i \leq H, 1 \leq j \leq N\}$.

Let $b_{H'+1,1} = v_{H'+1}$ and $b_{ij} = v_{H'+1} b'_{ij}$ for $H' + 1 \leq i \leq H, 1 \leq j \leq N$, $(i, j) \neq (H' + 1, 1)$.

Set

$$\begin{aligned} & b''_{ij}((v_2 \cdots v_{H'+1})^Q - 1)((v_3 \cdots v_{H'+1})^Q - 1) \cdots ((v_{H'+1,1})^Q - 1) \\ = \; & b'_{ij} - b'_{H'+1,j} b'_{i1}((b_{i1} v_2 \cdots v_{H'+1})^Q - 1)((b_{i1} v_3 \cdots v_{H'+1})^Q - 1) \cdots ((b_{i1} v_{H'+1})^Q - 1) \end{aligned}$$

for $i \geq H' + 2$ and

$$\begin{aligned} & a'_{i,H'+1} = a_{i,H'+1}((b_{H'+1,1} v_2 \cdots v_{H'+1})^Q - 1)((b_{H'+1,1} v_3 \cdots v_{H'+1})^Q - 1) \cdots ((b_{H'+1,1})^Q - 1) \\ & + a_{i,H'+2} b_{H'+2,1}^m((b_{H'+2,1} v_2 \cdots v_{H'+1})^Q - 1)((b_{H'+2,1} v_3 \cdots v_{H'+1})^Q - 1) \cdots ((b_{H'+2,1})^Q - 1) \\ & + \cdots + a_{iH} b_{H1}^m((b_{H1} v_2 \cdots v_{H'+1})^Q - 1)((b_{H1} v_3 \cdots v_{H'+1})^Q - 1) \cdots ((b_{H1})^Q - 1) \end{aligned}$$

for $1 \leq i \leq M$. By using Lemma 2 and setting $a_{i1} = a'_{i1}$, $b_{ij} = b''_{ij}$ again, we need to

consider the functions

$$\phi = v_{H'+1}^{(H-H')N-1} \prod_{i=1}^{H'} v_i^{T_i} \mathrm{d}v\mathrm{d}a\mathrm{d}b, \tag{9}$$

where

$$\begin{aligned} T_i &= mM(i-1) + (H-i+1)N + Q(M(i+H') \\ &\quad +(N-1)(2H-H'-i))(H'-i+1)/2 - 1, \end{aligned}$$

for $1 \leq i \leq H'$ and

$$\begin{aligned} &\Psi = (v_1^{QH'+m} v_2^{Q(H'-1)+m} \cdots v_{H'}^{m+Q})^2 ||A_1||^2 \\ &+(v_1^{QH'+m} v_2^{Q(H'-1)+m} \cdots v_{H'}^{m+Q} v_{H'+1}^{m})^2 (a_{1,H'+1}^2 + \cdots + a_{M,H'+1}^2) \\ &+ \sum_{\ell_1 = Qn+m, n \geq H'+1} (v_1^{\ell_1} v_2^{\ell_1 - Q} \cdots v_{H'}^{\ell_1-(H'-1)Q} v_{H'+1}^{\ell_1 - H'Q})^2 ||A_2 {f'}_{\ell_1,0,\cdots,0}||^2 \\ &+ \sum_{\substack{\ell_1+\cdots+\ell_N = Qn+m, \\ \ell_2+\cdots+\ell_N > 0}} (v_1^{\ell_1+(QH'+1)(\ell_2+\cdots+\ell_N)} v_2^{\ell_1+(Q(H'-1)+1)(\ell_2+\cdots+\ell_N)} \cdots v_{H'+1}^{\ell_1+\ell_2+\cdots+\ell_N})^2 ||A_2 f_{\ell_1,\ell_2,\cdots,\ell_N}||^2, \end{aligned} \tag{10}$$

where $A_1 = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H'} \\ & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH'} \end{pmatrix}$, $A_2 = \begin{pmatrix} a_{1,H'+2} & a_{1,H'+3} & \cdots & a_{1H} \\ & \vdots & & \\ a_{M,H'+2} & a_{M,H'+3} & \cdots & a_{MH} \end{pmatrix}$,

$f'_{Qn+m,0,\cdots,0}$

$$= \begin{pmatrix} b_{H'+2,1}^{m+Q(n-H'-1)}((b_{H'+2,1} v_2 \cdots v_{H'} v_{H'+1})^Q - 1)((b_{H'+2,1} v_3 \cdots v_{H'} v_{H'+1})^Q - 1) \cdots ((b_{H'+2,1})^Q - 1) \\ \vdots \\ b_{H1}^{m+Q(n-H'-1)}((b_{H1} v_2 \cdots v_{H'})^Q - 1)((b_{H1} v_3 \cdots v_{H'})^Q - 1) \cdots ((b_{H1})^Q - 1) \end{pmatrix}$$

and

$$f_{\ell_1,\ell_2,\cdots,\ell_N} = \begin{pmatrix} \prod_{j=1}^N b_{H'+2,j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H,j}^{\ell_j} \end{pmatrix}.$$

We construct the blow-up of the above function along the submanifold $\{v_{H'+1} = 0, a_{ki'} = 0, 1 \leq k \leq M, 1 \leq i' \leq H', \}$, m times. By letting $a_{ki'} = a'_{ki'} v_{H'+1}$, we have the poles $\frac{iMH'+N(H-H')}{2i}$ for $1 \leq i \leq m$.

Fix $1 \leq p \leq H'+1$. We construct the blow-up of the above function along the submanifold $\{v_p = 0, a'_{ki'} = 0, b_{ij} = 0, 1 \leq k \leq M, 1 \leq i' \leq H'+1, H'+2 \leq i \leq H, 2 \leq j \leq N\}$ Q times. Let $a'_{ki'} = v_p^Q a''_{ki'}$, $b_{ij} = v_p^Q b'_{ij}$, $1 \leq k \leq M, 1 \leq i' \leq H'+1, H'+2 \leq i \leq H, 2 \leq j \leq N$.

We have the $J(\xi)$'s poles

$$\frac{mM(p-1)+(H-p+1)N+Q(M(p+H')+(N-1)(2H-H'-p))(H'-p+1)/2+p'(M(H'+1)+(N-1)(H-H'-1)}{2Q(H'-p+1)+2m+2p'}$$

for $1 \leq p \leq H'+1$, $0 \leq p' \leq Q$ and the functions Eqs. (5) and (6) with $H'+1$, by setting $a_{ki'} = a'_{ki'}$, $b_{ij} = b'_{ij}$.

If $a_{1H}, \cdots, a_{MH}$ are constants, then we have the $J(\xi)$'s poles $\frac{NH+p(M+(H-1)(N-1))}{2(m+p)}$ for $0 \leq p \leq Q$, $\frac{iMH'+N(H-H')}{2i}$ for $1 \leq i \leq m$, $1 \leq H' \leq H-1$, and

$$\frac{mM(p-1)+(H-p+1)N+Q(M(p+H')+(N-1)(2H-H'-p))(H'-p+1)/2+p'(M(H'+1)+(N-1)(H-H'-1)}{2Q(H'-p+1)+2m+2p'}$$

for $1 \leq p \leq H'+1$, $1 \leq H' \leq H-2$, $0 \leq p' \leq Q$.

Bound values : bound$_3$

We next show that we have the following function (11) below.

$$\Psi = \sum_{0 \leq n \leq H-1} v_1^{2(Qn+m)} ||A^{(n)}_{M,\langle {N+m+nQ-1 \atop N-1} \rangle}||^2 \tag{11}$$

where $A_{M,\langle {N+m+nQ-1 \atop N-1} \rangle} = \begin{pmatrix} a'_{1,k_{n-1}+1} & \cdots & a'_{1,k_n} \\ \vdots & & \vdots \\ a'_{M,k_{n-1}+1} & \cdots & a'_{M,k_n} \end{pmatrix}$, where $k_n = \sum_{i=0}^{n} \langle {N+m+iQ-1 \atop N-1} \rangle$.

Construct the blow-up of the function (4) along the submanifold $\{b_{ij} = 0, 1 \leq i \leq H, 1 \leq j \leq N\}$. Let $b_{11} = v_1$, $b_{ij} = v_1 b'_{ij}$, $(i,j) \neq (1,1)$.

Let $f_{i,I} = \prod_{j=1}^{N} {b'_{ij}}^{\ell_j}$ for $I = (\ell_1, \cdots, \ell_N)$. Number the elements in the set $\{I = (\ell_1, \cdots, \ell_N) : \ell_1 + \cdots + \ell_N\}$ from 1 to $\sum_{i=0}^{H-1} \langle {N+m+iQ-1 \atop N-1} \rangle$, and we denote $I^{(1)}, I^{(2)}, \cdots$. We can assume that $\ell_1^{(k)} + \cdots + \ell_N^{(k)} \leq \ell_1^{(k+1)} + \cdots + \ell_N^{(k+1)}$.

Choose b'^{*}_{ij} properly such that all

$$Y_{i,k} = \begin{vmatrix} f_{1I^{(1)}} & f_{1I^{(2)}} & \cdots & f_{1I^{(k)}} & f_{1I^{(k+1)}} & \\ f_{2I^{(1)}} & f_{2I^{(2)}} & \cdots & f_{2I^{(k)}} & f_{2I^{(k+1)}} & \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ f_{kI^{(1)}} & f_{kI^{(2)}} & \cdots & f_{kI^{(k)}} & f_{kI^{(k+1)}} & \\ f_{iI^{(1)}} & f_{iI^{(2)}} & \cdots & f_{iI^{(k)}} & f_{iI^{(k+1)}} & \end{vmatrix}$$

are not zero in a small neighborhood of $\{b^*_{ij}\}$.

Then by setting $a'_{i,k+1} = a_{i,k+1}Y_{k+1,k} + a_{i,k+2}Y_{k+2,k} + a_{i,k+3}Y_{k+3,k} + \cdots + a_{iH}Y_{H,k}$ for $1 \leq i \leq M$ and by using Lemma 2 in Section 3.1 and Lemma 15 below, we have Eq. (11).

We have the $J(\xi)$'s poles $\frac{T_{H'}}{2m+2Q(H'+1)}$ for $0 \leq H' \leq H$, where

$$\begin{aligned} T_{H'} &= NH + MQ(H'+1)\langle \begin{matrix} N+m-1 \\ N-1 \end{matrix} \rangle + MQH'\langle \begin{matrix} N+m+Q-1 \\ N-1 \end{matrix} \rangle \\ &\quad + \cdots + MQ\langle \begin{matrix} N+m+QH'-1 \\ N-1 \end{matrix} \rangle. \end{aligned}$$

Lemma 15 *We have*

$$\begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} & b_{1,i} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} & b_{2,i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} & b_{i-1,i} \\ b_{i,1} & b_{i,2} & \cdots & b_{i,i-1} & b_{i,i} \end{vmatrix} \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} & b_{1,j} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} & b_{2,j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} & b_{i-1,j} \\ b_{k,1} & b_{k,2} & \cdots & b_{k,i-1} & b_{k,j} \end{vmatrix}$$

$$- \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} & b_{1,j} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} & b_{2,j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} & b_{i-1,j} \\ b_{i,1} & b_{i,2} & \cdots & b_{i,i-1} & b_{i,j} \end{vmatrix} \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} & b_{1,i} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} & b_{2,i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} & b_{i-1,i} \\ b_{k,1} & b_{k,2} & \cdots & b_{k,i-1} & b_{k,i} \end{vmatrix},$$

$$= \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} \\ \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} \end{vmatrix} \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1,i-1} & b_{1,i} & b_{1,j} \\ b_{21} & b_{22} & \cdots & b_{2,i-1} & b_{2,i} & b_{2,j} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{i-1,1} & b_{i-1,2} & \cdots & b_{i-1,i-1} & b_{i-1,i} & b_{i-1,j} \\ b_{i,1} & b_{i,2} & \cdots & b_{i,i-1} & b_{i,i} & b_{i,j} \\ b_{k,1} & b_{k,2} & \cdots & b_{k,i-1} & b_{k,i} & b_{k,j} \end{vmatrix},$$

where $i < j, k$.

Appendix C

In this section, we obtain the largest pole λ of $\int_{||AB||^2<1} ||AB||^{2\xi}$ and its order θ for $H \leq 2$, where $||AB||^2 = 0$ defines Vandermonde matrix type singularities.

Case 1 For $A = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{M1} \end{pmatrix}$ and $B = (b_{11}^m, b_{12}^m, \ldots, b_{1N}^m)$, we have

$$\lambda = \min\{\frac{M}{2}, \frac{N}{2m}\}, \theta = \begin{cases} 1, & \text{if } mM \neq N, \\ 2, & \text{if } mM = N. \end{cases}$$

by constructing the blow-up along the submanifold $\{b_{1j} = 0, 1 \leq j \leq N\}$.

For $A = \begin{pmatrix} a_{11}^* \\ a_{21}^* \\ \vdots \\ a_{M1}^* \end{pmatrix}$ and $B = (b_{11}, b_{12}, \ldots, b_{1N})$, we have

$$\lambda = \frac{N}{2m}, \theta = 1.$$

by constructing the blow-up along the submanifold $\{b_{1j} = 0, 1 \leq j \leq N\}$.

Case 2 Let $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \\ a_{M1} & a_{M2} \end{pmatrix}$, $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \end{pmatrix}$ and $B = (B_I)_{\ell_1+\ldots+\ell_N=Qn+m, 0\leq n\leq 1}$.

Construct the blow-up along the submanifold $\{b_{ij} = 0, 1 \leq i \leq 2, 1 \leq j \leq N\}$. Let $b_{11} = v_1$ and $b_{ij} = v_1 b'_{ij}$ for $(i,j) \neq (1,1)$. Set $b''_{2i} = b'_{2i} - b_{21}b_{1i}$ for $i \geq 2$ and $a'_{k1} = a_{k1} + a_{k2}b_{21}^m$ for $k \geq 1$.

By Lemmas 2 and 14, we need to consider

$$v_1^{2m} \| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} \| \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2} \end{pmatrix} \begin{pmatrix} v_1^Q b_{21}^m (b_{21}^Q - 1) & \prod_{j=1}^N b_{2j}''^{l_j} \end{pmatrix}_{l_1+\cdots+l_N=m, l_2+\cdots+l_N>0} \|^2.$$

Again set $b_{2i} = b''_{2i}$. Construct the blow-up along the submanifold $\{b_{2j} = 0, 1 \leq j \leq N\}$.

(I) Let $b_{21} = v_2$ and $b_{2j} = v_2 b'_{2j}$ for $j \geq 2$, and we need to consider

$$v_1^{2m} \| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} v_2^{2m} \| \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2} \end{pmatrix} \begin{pmatrix} v_1^Q (v_2^Q - 1) & b'_{22} & \dots & b'_{2N} \end{pmatrix} \|^2.$$

(II) Let $b_{22} = v_2$ and $b_{2j} = v_2 b''_{2j}$ for $j \neq 2$, and we need to consider

$$v_1^{2m} \| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} v_2^{2m} \| \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{M2} \end{pmatrix} \|^2.$$

We, therefore, have the poles

$$-\frac{2N + (k-1)(M+N-1)}{2(k-1)+2m} (k = 1, \dots, Q+1), -\frac{Mk+N}{2k} (k = 1, \dots, m), -M,$$

1. If $mM \leq N - 1$ then $\lambda = M$ and $\theta = 1$.
2. If $m = 1$, $M = N$, then $\lambda = \frac{2N+Q(2N-1)}{2(Q+1)}$ and $\theta = 1$.
3. If $m = 1$, $N = M - 1$ then $\lambda = N$ and $\theta = 2$.
4. If $m = 1$, $N < M - 1$ then $\lambda = N$ and $\theta = 1$.
5. If $m = 2$, $N = 1, M = 1$, then $\lambda = \frac{1}{2}$ and $\theta = 2$.
6. If $m = 2$, $N < mM$, $M > 1$ then $\lambda = \frac{N}{m}$ and $\theta = 1$.
7. If $m \geq 2$ $N = mM$ then $\lambda = \frac{N}{m}$ and $\theta = 3$.
8. If $m > 2$ $N < mM$ then $\lambda = \frac{N}{m}$ and $\theta = 1$.

Let $A = \begin{pmatrix} a_{11} & a^*_{12} \\ a_{21} & a^*_{22} \\ \vdots & \\ a_{M1} & a^*_{M2} \end{pmatrix}$, $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \end{pmatrix}$ and $B = (B_I)_{\ell_1 + \dots + \ell_N = Qn+m, 0 \leq n \leq 1}$.

Let all $b^*_{ij} = 0$. Construct the blow-up along the submanifold $\{b_{ij} = 0, 1 \le i \le 2, 1 \le j \le N\}$. Let $b_{11} = v_1$ and $b_{ij} = v_1 b'_{ij}$ for $(i,j) \ne (1,1)$. Set $b''_{2i} = b'_{2i} - b_{21}b_{1i}$ for $i \ge 2$ and $a'_{k1} = a_{k1} + a_{k2}b^m_{21}$ for $k \ge 1$. By Lemmas 2 and 14, we need to consider

$$v_1^{2m}\| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} \| \begin{pmatrix} a^*_{12} \\ a^*_{22} \\ \vdots \\ a^*_{M2} \end{pmatrix} \begin{pmatrix} v_1^Q b^m_{21}(b^Q_{21} - 1) & \prod_{j=1}^N b''^{l_j}_{2j} \end{pmatrix}_{l_1+\cdots+l_N=m, l_2+\cdots+l_N>0} \|^2.$$

Again set $b_{2i} = b''_{2i}$. Construct the blow-up along the submanifold $\{b_{2j} = 0, 1 \le j \le N\}$.

(I) Let $b_{21} = v_2$ and $b_{2j} = v_2 b'_{2j}$ for $j \ge 2$, and we need to consider

$$v_1^{2m}\| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} v_2^{2m} \| \begin{pmatrix} a^*_{12} \\ a^*_{22} \\ \vdots \\ a^*_{M2} \end{pmatrix} \begin{pmatrix} v_1^Q(v_2^Q - 1) & b'_{22} & \dots & b'_{2N} \end{pmatrix} \|^2.$$

(II) Let $b_{22} = v_2$ and $b_{2j} = v_2 b'_{2j}$ for $j \ne 2$, and we need to consider

$$v_1^{2m}\| \begin{pmatrix} a'_{11} \\ a'_{21} \\ \vdots \\ a'_{M1} \end{pmatrix} \|^2 + v_1^{2m} v_2^{2m} \| \begin{pmatrix} a^*_{12} \\ a^*_{22} \\ \vdots \\ a^*_{M2} \end{pmatrix} \|^2.$$

We, therefore, have the poles

$$-\frac{2N + (k-1)(M+N-1)}{2(k-1)+2m}(k = 1, \dots, Q+1), -\frac{Mk+N}{2k}(k = 1, \dots, m),$$

and

1. If $m \ge 2$, $mM \le N-1$ then $\lambda = \frac{mM+N}{2m}$ and $\theta = 1$.
2. If $m = 1$, $N \ge M+Q+1$ then $\lambda = \frac{N+M}{2}$ and $\theta = 1$.
3. If $m = 1$, $N = M+Q$ then $\lambda = \frac{N+M}{2}$ and $\theta = 2$.
4. If $m = 1$, $M+1 \le N \le M+Q-1$ then $\lambda = \frac{2N+Q(2N-1)}{2(Q+1)}$ and $\theta = 1$.

5. If $m = 1$, $N = M$ then $\lambda = \frac{2N+Q(2N-1)}{2(Q+1)}$ and $\theta = 1$.

6. If $m = 1$, $N = M - 1$ then $\lambda = N$ and $\theta = 2$.

7. If $m = 1$, $N < M - 1$ then $\lambda = N$ and $\theta = 1$.

8. If $m = 2$, $N = 1, M = 1$, then $\lambda = \frac{1}{2}$ and $\theta = 2$.

9. If $m = 2$, $N < mM$, $M > 1$ then $\lambda = \frac{N}{m}$ and $\theta = 1$.

10. If $m \geq 2$ $N = mM$ then $\lambda = \frac{N}{m}$ and $\theta = 2$.

11. If $m > 2$ $N < mM$ then $\lambda = \frac{N}{m}$ and $\theta = 1$.

Appendix D

Set

$$\lambda_0(H) = \frac{(k_0^2 + k_0)(M - N + 1) + H}{4k_0 + 2},$$

where $k_0 = \max\{i \in \mathbb{Z} \mid (M - N + 1)i^2 \leq H\}$, and

$$\lambda_1(H) = \frac{(k_1 + k_1^2)(M - N + 1) + 2H}{4(k_1 + 1)},$$

where $k_1 = \max\{i \in \mathbb{Z} \mid (M - N + 1)(i^2 + i) \leq 2H\}$.

We have

$$\lambda_0(H) = \min_{i \geq 0} \frac{(i + i^2)(M - N + 1) + H}{4i + 2}, \lambda_1(H) = \min_{i \geq 0} \frac{(i + i^2)(M - N + 1) + 2H}{4(i + 1)}.$$

Lemma 16 *(1)* $\sum_{\alpha=1}^{r} \lambda_1(H_\alpha) \geq \frac{r-1}{2} + \lambda_1(\sum_{\alpha=1}^{r} H_\alpha - 1)$.

(2) $\lambda_0(H_0) + \lambda_1(H_1) \geq \min\{\lambda_0(H_0 + H_1 - 1) + \frac{1}{2}, \lambda_1(H_0 + H_1)\}$.

(Proof)

(1) Let $H', H'' \geq 2$. Since for some $k_1(k_1+1) \leq 2H'/(M-N+1), k_2(k_2+1) \leq 2H''/(M-N+1)$, we have

$$\begin{aligned}
& \lambda_1(H') + \lambda_1(H'') - \lambda_1(H'+H''-1) - \lambda_1(1) \\
\geq\ & \frac{k_1(M-N+1)}{4} + \frac{H'}{2(k_1+1)} + \frac{k_2(M-N+1)}{4} + \frac{H''}{2(k_2+1)} \\
& -\frac{(k_1+k_2)(M-N+1)}{4} - \frac{H'+H''-1}{2(k_1+k_2+1)} - \frac{1}{2} \\
=\ & \frac{H'(k_2+1)k_2 + H''(k_1+1)k_1 - (k_1+1)(k_2+1)(k_1+k_2)}{2(k_1+1)(k_2+1)(k_1+k_2+1)(M-N+1)} \\
\geq\ & \frac{(k_1+1)(k_2+1)k_2 + (k_2+1)(k_1+1)k_1 - (k_1+1)(k_2+1)(k_1+k_2)}{2(k_1+1)(k_2+1)(k_1+k_2+1)(M-N+1)} = 0
\end{aligned}$$

Therefore, we have

$$\sum_{\alpha=1}^{r} \lambda_1(H_\alpha) \geq (r-1)\lambda_1(1) + \lambda_1(\sum_{\alpha=1}^{r} H_\alpha - 1).$$

(2) Let $H' \geq 0$, $H'' \geq 1$. We have

$$\begin{aligned}
& \frac{(k_0^2+k_0)(M-N+1)+H'}{4k_0+2} + \frac{(k_1^2+k_1)(M-N+1)+2H''}{2k_1+2} \\
& -\frac{(k_0^2+k_0)(M-N+1)+(H'+1)}{4k_0+2} + \frac{(k_1^2+k_1)(M-N+1)+2(H''-1)}{2k_1+2} \\
=\ & \frac{1}{2(k_1+1)} - \frac{1}{4k_0+2},
\end{aligned}$$

Therefore,

(i) if $\lambda_0(H') + \lambda_1(H'') = \frac{(k_0^2+k_0)(M-N+1)+H'}{4k_0+2} + \frac{(k_1^2+k_1)(M-N+1)+2H''}{2k_1+2}$ and $2k_0+1 \geq k_1+1$ then inductively

$$\lambda_0(H') + \lambda_1(H'') \geq \lambda_0(H'+1) + \lambda_1(H''-1) \geq \lambda_0(H'+H''-1) + \lambda_1(1).$$

(ii) If $\lambda_0(H'+1) + \lambda_1(H''-1) = \frac{(k_0^2+k_0)(M-N+1)+H'+1}{4k_0+2} + \frac{(k_1^2+k_1)(M-N+1)+2(H''-1)}{2k_1+2}$ and $2k_0+1 < k_1+1$ then inductively

$$\lambda_0(H'+1) + \lambda_1(H''-1) > \lambda_0(H') + \lambda_1(H'') > \lambda_1(H'+H'').$$

From (i) and (ii), we have

$$\lambda_0(H_0) + \lambda_1(H_1) \ \geq\ \min\{\lambda_0(H_0+H_1-1) + \frac{1}{2}, \lambda_1(H_0+H_1)\}. \qquad (12)$$

Q.E.D.

Therefore, by some computations, we have

$$\begin{aligned}
&\min_{H_0+\cdots+H_r=H, H_1\geq 1,\cdots,H_r\geq 1}\{\frac{H_0N+(M(1+k_0)+(N-1)(2H_0-k_0-1))k_0}{4k_0+2}\\
&\qquad\qquad +\sum_{\alpha=1}^{r}\frac{2H_\alpha N+(M(1+k_\alpha)+(N-1)(2H_\alpha-k_\alpha-1))k_\alpha}{4k_\alpha+4}\}\\
=\ &\min_{H_0+\cdots+H_r=H, H_1\geq 1,\cdots,H_r\geq 1}\{\frac{H_0(N-1)}{2}+\frac{H_0+k_0(M-N+1)(1+k_0)}{4k_0+2}\\
&\qquad\qquad +\sum_{\alpha=1}^{r}\{\frac{H_\alpha(N-1)}{2}+\frac{2H_\alpha+(M-N+1)(k_\alpha+k_\alpha^2)}{4k_\alpha+4}\}\}\\
=\ &\frac{H(N-1)}{2}+\min_{H_0+\cdots+H_r=H, H_1\geq 1,\cdots,H_r\geq 1}\{\lambda_0(H_0)+\sum_{\alpha=1}^{r}\lambda_1(H_\alpha)\}\\
=\ &\frac{H(N-1)}{2}+\min\{\lambda_0(H-r)+\frac{r}{2},\lambda_1(H-r+1)+\frac{r-1}{2}\}
\end{aligned}$$

where $k_0 = \max\{i \in \mathbb{Z}; H_0 \geq i^2(M-N+1)\}$ and $k_\alpha = \max\{i \in \mathbb{Z}; 2H_\alpha \geq (i^2+i)(M-N+1)\}$ for $\alpha \geq 1$.

Lemma 17 *We have the followings.*

- *If* $M-N+1=1$, *then*

  $\lambda_0(H-1)+\frac{1}{2} > \lambda_1(H)$ *for* $2 \leq H \leq 9$.

  $\lambda_0(H-1)+\frac{1}{2} = \lambda_1(H)$ *for* $H = 10$.

  $\lambda_0(H-1)+\frac{1}{2} < \lambda_1(H)$ *for* $H > 10$.

- *If* $M-N+1=2$, *then*

  $\lambda_0(H-1)+\frac{1}{2} = \lambda_1(H)$ *for* $H = 2$.

  $\lambda_0(H-1)+\frac{1}{2} > \lambda_1(H)$ *for* $3 \leq H \leq 5$.

  $\lambda_0(H-1)+\frac{1}{2} = \lambda_1(H)$ *for* $6 \leq H \leq 9$.

  $\lambda_0(H-1)+\frac{1}{2} < \lambda_1(H)$ *for* $H > 9$.

- *If* $M-N+1 \geq 3$, *then*

  $\lambda_0(H-1)+\frac{1}{2} = \lambda_1(H)$ *for* $H \leq M-N+1$.

  $\lambda_0(H-1)+\frac{1}{2} > \lambda_1(H)$ *for* $M-N+1 < H \leq M-N+4$.

  $\lambda_0(H-1)+\frac{1}{2} = \lambda_1(H)$ *for* $H = M-N+5$.

  $\lambda_0(H-1)+\frac{1}{2} < \lambda_1(H)$ *for* $H > M-N+5$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, *3* 716 – 723.

Akaike, H. (1980). Likelihood and Bayes procedure. *Bayesian Statistics*, Valencia, Spain, University Press, 143 – 166.

Amari S. and Murata N. (1993). Statistical theory of learning curves under entropic loss. *Neural Computation*, 140 – 153.

Amari S., Fujita N. and Shinomoto S. (1992). Four Types of Learning Curves, *Neural Computation*, **4**, 608 – 618.

Aoyagi M. (2006) The zeta function of learning theory and generalization error of three layered neural perceptron, *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities*, **1501**, 153 – 167.

Aoyagi M. (2009) Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network, *International Journal of Pure and Applied Mathematics*, **52**-2, 177 – 204.

Aoyagi M. (2010a) A Bayesian Learning Coefficient of Generalization Error and Vandermonde Matrix-Type Singularities, *Communications in Statistics - Theory and Methods*, **39**-15, 2667 – 2687.

Aoyagi M. (2010b) Stochastic Complexity and Generalization Error of a Restricted Boltzmann Machine in Bayesian Estimation, *Journal of Machine Learning Research*, **11**-Apr, 1243 – 1272.

Aoyagi M. and Watanabe S. (2005a) Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network, *IEICE Trans. J88-D-II*, **10**, 2112 – 2124.

Aoyagi M. and Watanabe S. (2005b) Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation, *Neural Networks*, **18**, 924 – 933.

Bernstein I. N. (1972) The analytic continuation of generalized functions with respect to a parameter, *Functional Analysis Applications*, **6**, 26 – 40.

Bjőrk J. E. (1979) Rings of differential operators, *Amsterdam: North-Holland*,

Fukumizu K. (1996) A regularity condition of the information matrix of a multilayer perceptron network, *Neural Networks*, **9**-5, 871 – 879.

Fulton W. (1993) Introduction to toric varieties, Annals of Mathematics Studies, *Princeton University Press*.

Hagiwara K., Toda N. and Usui S.(1993) On the problem of applying AIC to determine the structure of a layered feed-forward neural network, *Proceedings of IJCNN Nagoya Japan*, **3**, 2263 – 2266.

Hannan E. J. and Quinn B. G. (1979) The determination of the order of an autoregression, *Journal of Royal Statistical Society, Series B*, **41**, 190 – 195.

Hartigan J. A.(1985) A failure of likelihood asymptotics for normal mixtures, *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, **2**, 807 – 810.

Kashiwara M. (1976) B-functions and holonomic systems, *Inventions Math.*, **38**, 33 – 53.

Kollár J. (1997) Singularities of pairs, *Algebraic geometry-Santa Cruz 1995, Proc. Sympos. Pure Math., Amedsr. Math. Soc., Providence, RI*, **62**, 221 – 287.

Levin E., Tishby N. and Solla S. A. (1990) A statistical approaches to learning and generalization in layered neural networks, *Proceedings of IEEE*, **78**-10, 1568 – 1674.

Lin S. (2010) Asymptotic approximation of marginal likelihood integrals, *arXiv:1003.5338*.

Mackay D. J. (1992) Bayesian interpolation, *Neural Computation*, **4**-2, 415 – 447.

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

Mustata M. (2002) Singularities of pairs via jet schemes, *J. Amer. Math. Soc.*, **15**, 599 – 615.

Nagata K. and Watanabe S. (2008a) Exchange Monte Carlo Sampling from Bayesian Posterior for Singular Learning Machines, *IEEE Transactions on Neural Networks*, **19**-7, 1253 – 1266.

Nagata K. and Watanabe S. (2008b) Asymptotic Behavior of Exchange Ratio in Exchange Monte Carlo Method, *International Journal of Neural Networks*, **21**-7, 980 – 988.

Nagata K. and Watanabe S. (2008c) Design of Exchange Monte Carlo Method for Bayesian Learning in Normal Mixture Models, *Proceedings of International Conference on Neural Information Processing* (to appear)

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

Rissanen J. (1986) Stochastic complexity and modeling, *Annals of Statistics*, **14** 1080 – 1100.

Rusakov D. and Geiger D. Rissanen J. (2005) Asymptotic Model Selection for Naive Bayesian Networks, *Journal of Machine Learning Research*, **6** 1 – 35.

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

Sturmfels B. (2008) Open problems in algebraic statistics, *in Emerging Applications of Algebraic Geometry, (editors M. Putinar and S. Sullivant), I.M.A. Volumes in Mathematics and its Applications*, **149** 351 – 364.

Sussmann H. J. (1992) Uniqueness of the weights for minimal feed-forward nets with a given input-output map, *Neural Networks*, **5** 589 – 593.

Takamatsu S., Nakajima S. & Watanabe S.(2005) Generalization Error of Localized Bayes Estimation in Reduced Rank Regression, *Workshop on Information-Based Induction Sciences*, 81 – 86.

Takeuchi K.(1976) Distribution of an information statistic and the criterion for the optimal model, *Mathematical Science*, **153** 12 – 18.

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

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

Watanabe S.(2001c) Learning efficiency of redundant neural networks in Bayesian estimation, *IEEE Transactions on Neural Networks*, **12**-6 1475 – 1486.

Watanabe S.(2009) Algebraic Geometry and Statistical Learning Theory, *Cambridge Monographs on Applied and Computational Mathematics*, **25**

Watanabe S.(2010) Equations of states in singular statistical estimation, *Neural Networks*, **23**-1 20 – 34.

Watanabe S., Yamazaki K. and Aoyagi M.(2004) Kullback Information of Normal Mixture is not an Analytic Function, *Technical report of IEICE, NC2004*, 41 – 46.

Yamazaki K., Aoyagi M. Watanabe S.(2010) Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram, *Neural Networks*, **23**-1 35 – 43.

Yamanishi K.(1998) A decision-theoretic extension of stochastic complexity and its applications to learning, *IEEE Trans. on Information Theory*, **44**-4 1424 – 1439.