数理解析研究所講究録
第1694巻 2010年 1-19

Log canonical threshold and singularities in learning theory

Miki Aoyagi

ARISH, Nihon University,
Nihon University Kaikan Daini Bekkan, 12-5, Goban-cho, Chiyoda-ku,
Tokyo 102-8251, Japan. Email: aoyagi.miki@nihon-u.ac.jp

Abstract

In this paper, we show new bounds of the log canonical threshold for Vandermonde matrix type singularities and summarize our recent results for the log canonical thresholds of singularities in learning models.

Keywords log canonical threshold, resolution of singularities, generalization error, hierarchical learning models,

1 Introduction

Recently, the term "algebraic statistics" arises from the study of probabilistic models and techniques for statistical inference using methods from algebra and geometry (Sturmfels [29]). Our study is to consider the generalization error and the stochastic complexity in learning theory by using the log canonical threshold in real and complex analysis and algebraic geometry.

The log canonical threshold $c_Z(Y,f)$ in algebraic geometry is analytically defined by

$$c_Z(Y,f) = \sup\{c : |f|^{-c} \text{ is locally } L^2 \text{ on } U\},$$

over the complex field and

$$c_Z(Y,f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ on } U\},$$

over the real field for a nonzero regular function f on a smooth variety Y, where $Z \subset Y$ is a closed subscheme and U a neighborhood of Z (Kollár [19], Mustata [22]).

It is also known that $c_0(\mathbb{C}^d, f)$ 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 [12], Björk[13], Kashiwara[18]). Let $\zeta(z) = \int_U |f(w)|^{2z}\psi dw \wedge d\overline{w}$ over the complex field and $\zeta(z) = \int_U |f(w)|^z\psi dw$ over the real field, where ψ is a C^∞ function with compact support. Atiyah proved that $\zeta(z)$ is a meromorphic function on $\mathbb{C}$, and its poles are negative rational numbers, by using resolution of singularities [11]. The largest pole of $\zeta(z)$ corresponds to $-c_Z(Y,f)$, if $\overline{supp}(\psi) \supset Z$. We denote by $\theta_Z(Y,f)$ the order of the

largest pole in this paper. For simple example, we have $c_0(\mathbb{C}, z^m) = 1/m$, $\theta_0(\mathbb{C}, z^m) = 1$ and $c_0(\mathbb{R}, x^m) = 1/m$, $\theta_0(\mathbb{R}, x^m) = 1$.

We have many differences between the real field and the complex field, for example, log canonical thresholds over the complex field are less than 1, while those over the real field are not necessarily less than 1. In algebraic geometry and algebraic analysis, these studies are usually done over an algebraically closed field (Kollár [19], Mustata [22]). We cannot apply results over an algebraically closed field to our cases over the real field, directly.

The theoretical study of hierarchical learning models has been rapidly developed in recent years. The data analyzed by such learning models are associated with image or speech recognition, artificial intelligence, the control of a robot, genetic analysis, data mining, time series prediction, and so on. They are very complicated and usually not generated by a simple normal distribution, as they are influenced by many factors. Hierarchical learning models such as the normal mixture model, the Boltzmann machine, layered neural network and reduced rank regression may be known to be effective learning models. They, however, likewise have complicated, i.e., non-regular statistical structures, which cannot be analyzed using the classic theories of regular statistical models. (Hartigan [17], Sussmann [30], Hagiwara, Toda, & Usui [16], Fukumizu [14]). The theoretical study has therefore been started to construct a mathematical foundation for non-regular statistical models.

Watanabe proved that the largest pole of a zeta function for a non-regular statistical model gives the main term of the generalization error of hierarchical learning models in Bayesian estimation (Watanabe [32], [33]). The generalization error of a learning model is a difference between a true density function and a predictive density function obtained using distributed training samples. It is one of the most important topic in learning theory. The largest pole of a zeta function for a learning model, which is called a Bayesian learning coefficient, corresponds to the log canonical threshold.

In this paper, we show new bounds of the log canonical thresholds for Vandermonde matrix type singularities. Vandermonde matrix type singularities have been recognized to be related to Bayesian learning coefficients for the three layered neural network (Aoyagi & Watanabe [8], Aoyagi [4], [5]), normal mixture models (Watanabe, Yamazaki & Aoyagi [36], Aoyagi [6]), and the mixtures of binomial distribution (Yamazaki, Aoyagi & Watanabe [38]). These facts seem to imply that the singularities are essential for learning theory. We also overview our recent results of singularities for the restricted Boltzmann machine (Aoyagi [7]) and the reduced rank regression (Aoyagi & Watanabe [9]). Such singularities are degenerate with respect to their Newton polyhedrons and non-isolation of their singularities (Fulton [15]). In several papers, only upper bounds of these values were reported before (Watanabe [31], Watanabe & Watanabe [35], Yamazaki & Watanabe [39], [40], Nishiyama & Watanabe [25]). Rusakov and Geiger [27] considered them for Naive Bayesian networks.

Such our results were used for analyzing and developing the precision of the Markov Chain Monte Carlo (Nagata & Watanabe, [23]) and for studying the setting of temperatures for the exchange MCMC method (Nagata & Watanabe [24]).

2 Main Result

In this paper, we denote by a^*, b^* constants and denote by a^* if the variable a is in a sufficiently small neighborhood of a^*.

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

2.1 Vandermonde matrix type singularities

Definition 1 *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 2 *Fix $Q \in \mathbb{N}$ and $m \in \mathbb{N}_{+0}$.*

Let $MH + HN$ variables $w = \left\{ \begin{pmatrix} a_{11} & \cdots & a_{1H} \\ a_{21} & \cdots & a_{2H} \\ & \vdots & \\ a_{M1} & \cdots & a_{MH} \end{pmatrix}, \begin{pmatrix} b_{11} & \cdots & b_{1N} \\ b_{21} & \cdots & b_{2N} \\ & \vdots & \\ b_{H1} & \cdots & b_{HN} \end{pmatrix} \right\}$ and

$rM + rN$ *constants* $w_t^* = \left\{ \begin{pmatrix} a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & \vdots & \\ a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix}, \begin{pmatrix} b^*_{H+1,1} & \cdots & b^*_{H+1,N} \\ b^*_{H+2,1} & \cdots & b^*_{H+2,N} \\ & \vdots & \\ b^*_{H+r,1} & \cdots & b^*_{H+r,N} \end{pmatrix} \right\}$.

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 $B = (B_I)_{\ell_1+\cdots+\ell_N = Qn+m, 0 \leq n \leq H+r-1}$ (t denotes the transpose), where A is an $M \times (H+r)$ dimensional matrix and B is an $(H+r) \times \sum_{n=0}^{H+r-1} \frac{(Qn+m+N-1)!(Qn+m)!}{(N-1)!}$ dimensional matrix.

We call singularities of $||AB||^2 = 0$ 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 $j \neq j'$.

Let w, w_t^*, A and B be as in Definition 2. Let w be in a sufficiently small neighborhood of

$$w^* = \left\{ \begin{pmatrix} a_{11}^* & \cdots & a_{1H}^* \\ a_{21}^* & \cdots & a_{2H}^* \\ & \vdots & \\ a_{M1}^* & \cdots & a_{MH}^* \end{pmatrix}, \begin{pmatrix} b_{11}^* & \cdots & b_{1N}^* \\ b_{21}^* & \cdots & b_{2N}^* \\ & \vdots & \\ b_{H1}^* & \cdots & b_{HN}^* \end{pmatrix} \right\}.$$

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

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

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

Then $r' \geq r$ and set $(b_{i1}^{**}, \cdots, b_{iN}^{**}) = [b_{H+i,1}^*, \cdots, b_{H+i,N}^*]_Q$, for $1 \leq i \leq r$.
It is natural to assume that

$$\left.\begin{matrix} [b_{11}^*, \cdots, b_{1N}^*]_Q \\ \vdots \\ [b_{H_0 1}^*, \cdots, b_{H_0 N}^*]_Q \end{matrix}\right\} = 0,$$

$$\left.\begin{matrix} [b_{H_0+1,1}^*, \cdots, b_{H_0+1,N}^*]_Q \\ \vdots \\ [b_{H_0+H_1,1}^*, \cdots, b_{H_0+H_1,N}^*]_Q \end{matrix}\right\} = (b_{11}^{**}, \cdots, b_{1N}^{**}),$$

$$\left.\begin{matrix} [b_{H_0+H_1+1,1}^*, \cdots, b_{H_0+H_1+1,N}^*]_Q \\ \vdots \\ [b_{H_0+H_1+H_2,1}^*, \cdots, b_{H_0+H_1+H_2,N}^*]_Q \end{matrix}\right\} = (b_{21}^{**}, \cdots, b_{2N}^{**}),$$

$$\vdots$$

$$\left.\begin{matrix} [b_{H_0+\cdots+H_{r'-1}+1,1}^*, \cdots, b_{H_0+\cdots+H_{r'-1}+1,N}^*]_Q \\ \vdots \\ [b_{H_0+\cdots+H_{r'-1}+H_{r'},1}^*, \cdots, b_{H_0+\cdots+H_{r'-1}+H_{r'},N}^*]_Q \end{matrix}\right\} = (b_{r'1}^{**}, \cdots, b_{r'N}^{**}).$$

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

Theorem 1 *(Aoyagi [5])*
We have

$$c_{w^*}(||AB||^2) = \sum_{\alpha=0}^{r'} c_{w^{(\alpha)*}}(||A^{(\alpha)}B^{(\alpha)}||^2),$$

where $w^{(\alpha)*} = \{a_{ki}^{(\alpha)*}, b_{ij}^{(\alpha)*}\} = \{a_{k,H_0+\cdots+H_{\alpha-1}+i}^*, b_{\alpha j}^{**}\}_{1 \leq k \leq M, 1 \leq i \leq H_\alpha, 1 \leq j \leq N}$,
$I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$,

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} \\ & & \vdots & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} \end{pmatrix}, B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha j}^{(\alpha)\ell_j} \end{pmatrix}, \text{ for } \alpha = 0, r+1 \leq \alpha \leq r',$$

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} & a_{1,H+\alpha}^{*} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} & a_{2,H+\alpha}^{*} \\ & & \vdots & & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} & a_{M,H+\alpha}^{*} \end{pmatrix}, B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^{N} b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^{N} b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^{N} b_{H_\alpha j}^{(\alpha)\ \ell_j} \\ \prod_{j=1}^{N} b_{\alpha j}^{**\ell_j} \end{pmatrix}, \text{ for } 1 \leq \alpha \leq r,$$

$B^{(0)} = (B_I^{(0)})_{\ell_1+\cdots+\ell_N=Qn+m, 0\leq n\leq H_0-1}$, $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, 0\leq n\leq H_\alpha-1}$ *for* $r+1 \leq \alpha \leq r'$ *and* $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, 0\leq n\leq H_\alpha}$ *for* $1 \leq \alpha \leq r$.

$B^{(0)}$, $B^{(\alpha)}(1 \leq \alpha \leq r)$ *and* $B^{(\alpha)}(r+1 \leq \alpha \leq r')$ *are* $H_0 \times \sum_{n=0}^{H_0-1} \frac{(Qn+m+N-1)!(Qn+m)!}{(N-1)!}$, $(H_\alpha+1) \times \sum_{n=0}^{H_\alpha} \frac{(n+N-1)!n!}{(N-1)!}$ *and* $H_\alpha \times \sum_{n=0}^{H_\alpha-1} \frac{(n+N-1)!n!}{(N-1)!}$ *dimensional matrices, respectively.*

Theorem 1 shows certain orthogonality conditions of the log canonical threshold of Vandermonde matrix type singularities. Usually, r corresponds to the number of elements in a true distribution. It means that the Bayesian learning coefficient related with such singularities is the sum of each for the small model with respect to each element of a true distribution.

Theorem 2 *(Aoyagi [5]) We use the same notations as in Theorem 1. If* $N = 1$, *we have*

$$\begin{aligned} c_{w^*}(||AB||^2) &= \frac{MQk_0(k_0+1)+2H_0}{4(m+k_0Q)} \\ &+ \frac{Mr'}{2} + \sum_{\alpha=1}^{r} \frac{Mk_\alpha(k_\alpha+1)+2H_\alpha}{4(1+k_\alpha)} + \sum_{\alpha'=r+1}^{r'} \frac{Mk_{\alpha'}(k_{\alpha'}+1)+2(H_{\alpha'}-1)}{4(1+k_{\alpha'})}, \end{aligned}$$

where

$$\begin{aligned} k_0 &= \max\{i \in \mathbb{Z}; 2H_0 \geq M(i(i-1)Q+2mi)\}, \\ k_\alpha &= \max\{i \in \mathbb{Z}; 2H_\alpha \geq M(i^2+i)\} \text{ for } 1 \leq \alpha \leq r, \\ k_{\alpha'} &= \max\{i \in \mathbb{Z}; 2(H_{\alpha'}-1) \geq M(i^2+i)\} \text{ for } r+1 \leq \alpha' \leq r'. \end{aligned}$$

and

$$\theta_{w^*}(||AB||^2) = \#\Theta + 1$$

where

$$\begin{aligned} \Theta = \{k_0, k_\alpha, k_{\alpha'} \ : \ & 2H_0 = M(k_0(k_0-1)Q+2mk_0), 2H_\alpha = M(k_\alpha^2+k_\alpha), 1 \leq \alpha \leq r, \\ & 2(H_{\alpha'}-1) = M(k_{\alpha'}^2+\alpha'), r+1 \leq \alpha' \leq r'\}. \end{aligned}$$

The next theorem gives new bounds for the log canonical threshold of Vandermonde matrix type singularities.

Theorem 3 *We use the same notations as in Theorem 1. We have the followings.*

(1)

$$c_{w^{(0)*}}(||A^{(0)}B^{(0)}||^2) \leq \begin{cases} \frac{MH_0}{2}, & \text{if } mM \leq N-1, \\ \frac{NH_0}{2m}, & \text{if } N \leq mM \leq m(N-1), \\ \frac{NH_0}{2m}, & \text{if } M \geq N, (N-1)(m-1) \geq 1, \\ \frac{2H_0N + Q(M(1+k_0) + (N-1)(2H_0 - k_0 - 1))k_0}{4Qk_0 + 4m}, & \\ \text{if } M \geq N, (N-1)(m-1) = 0, & \end{cases}$$

where $k_0 = \max\{i \in \mathbb{Z}; 2H_0 \geq (Qi(i-1) + 2mi)(M - N + 1)\}$.

(2) *For* $1 \leq \alpha \leq r$,

$$c_{w^{(\alpha)*}}(||A^{(\alpha)}B^{(\alpha)}||^2) \leq \begin{cases} \frac{MH_\alpha + N}{2}, & \text{if } M \leq N-1, \\ \frac{M}{2} + \frac{2H_\alpha N + (M(1+k_\alpha) + (N-1)(2H_\alpha - k_\alpha - 1))k_\alpha}{4k_\alpha + 4}, & \\ \text{if } M \geq N, & \end{cases}$$

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

(3) *For* $r+1 \leq \alpha' \leq r'$,

$$c_{w^{(\alpha')*}}(||A^{(\alpha')}B^{(\alpha')}||^2) \leq \begin{cases} \frac{MH_{\alpha'}}{2}, & \text{if } M \leq N-1, \\ \frac{M}{2} + \frac{2(H_{\alpha'} - 1)N + (M(1+k_{\alpha'}) + (N-1)(2H_{\alpha'} - 3 - k_{\alpha'}))k_{\alpha'}}{4k_{\alpha'} + 4}, & \\ \text{if } M \geq N, & \end{cases}$$

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

(Proof)

Assume that $H_0 = H$.

Let

$$\Psi = ||AB||^2, \tag{1}$$

$\phi = dadb$, V is a sufficiently small neighborhood of 0 and $J(z) = \int_V \Psi^z \phi$.

By using a blowing up process together with an inductive method, we show that we have the following functions (2) and (3) below.

Let

$$\phi = \prod_{i=1}^{H'} v_i^{T_i} dv dadb, \tag{2}$$

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

$$\begin{aligned}\Psi &= (v_1^{QH'+m}v_2^{Q(H'-1)+m}\cdots v_{H'}^{Q+m})^2||A_1||^2 \qquad (3)\\ &+ \sum_{\ell_1=Qn+m,n\geq H'}(v_1^{\ell_1}v_2^{\ell_1-Q}\cdots v_{H'}^{\ell_1-(H'-1)Q})^2||A_2f'_{\ell_1,0,\cdots,0}||^2\\ &+ \sum_{\substack{\ell_1+\cdots+\ell_N=Qn+m,\\ \ell_2+\cdots+\ell_N>0}}(v_1^{\ell_1+(QH'+1)(\ell_2+\cdots+\ell_N)}v_2^{\ell_1+(Q(H'-1)+1)(\ell_2+\cdots+\ell_N)}\cdots v_{H'}^{\ell_1+(Q+1)(\ell_2+\cdots+\ell_N)})^2\\ &\qquad\times||A_2f_{\ell_1,\ell_2,\cdots,\ell_N}||^2,\end{aligned}$$

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

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

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 (1) 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 1 in Section 2.4 and setting $a_{i1}=a'_{i1}$, $b_{ij}=b''_{ij}$ again, we need to consider the functions

$$\phi = v_1^{NH-1}\mathrm{d}v\mathrm{d}a\mathrm{d}b, \qquad (4)$$

and

$$\begin{aligned}\Psi &= (v_1^m)^2||A_1||^2 \qquad (5)\\ &+ \sum_{\ell_1=Qn+m,n\geq 1}(v_1^{\ell_1})^2||A_2f'_{\ell_1,0,\cdots,0}||^2\\ &+ \sum_{\substack{\ell_1+\cdots+\ell_N=Qn+m,\\ \ell_2+\cdots+\ell_N>0}}(v_1^{\ell_1+\ell_2+\cdots+\ell_N})^2||A_2f_{\ell_1,\ell_2,\cdots,\ell_N}||^2,\end{aligned}$$

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 (5) 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 $a_{k1}=v_1^Qa'_{k1}$, $b_{ij}=v_1^Qb'_{ij}$, $1\leq k\leq M, 2\leq i\leq H, 2\leq j\leq N$.

We have the $J(z)$'s poles $\frac{NH+p(M+(H-1)(N-1))}{2(m+p)}$ for $0 \leq p \leq Q$ and the functions Eqs. (2) and (3) with $H'=1$, by setting $a_{i1}=a'_{i1}$, $b_{ij}=b'_{ij}$.

Assume Eqs. (2) and (3). Construct the blow-up of function (3) 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

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

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

$$\phi = v_{H'+1}^{(H-H')N-1}\prod_{i=1}^{H'}v_i^{T_i}\mathrm{d}v\mathrm{d}a\mathrm{d}b, \tag{6}$$

where

$$T_i = mM(i-1)+(H-i+1)N+Q(M(i+H')+(N-1)(2H-H'-i))(H'-i+1)/2-1,$$

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

$$\begin{aligned} \Psi &= (v_1^{QH'+m}v_2^{Q(H'-1)+m}\cdots v_{H'}^{m+Q})^2||A_1||^2 \\ &+(v_1^{QH'+m}v_2^{Q(H'-1)+m}\cdots v_{H'}^{m+Q}v_{H'+1}^{m})^2(a^2_{1,H'+1}+\cdots+a^2_{M,H'+1}) \\ &+\sum_{\ell_1=Qn+m, n\geq H'+1}(v_1^{\ell_1}v_2^{\ell_1-Q}\cdots v_{H'}^{\ell_1-(H'-1)Q}v_{H'+1}^{\ell_1-H'Q})^2||A_2f'_{\ell_1,0,\cdots,0}||^2 \\ &+\sum_{\substack{\ell_1+\cdots+\ell_N=Qn+m,\\ \ell_2+\cdots+\ell_N>0}}(v_1^{\ell_1+(QH'+1)(\ell_2+\cdots+\ell_N)}v_2^{\ell_1+(Q(H'-1)+1)(\ell_2+\cdots+\ell_N)}\cdots v_{H'+1}^{\ell_1+\ell_2+\cdots+\ell_N})^2||A_2f_{\ell_1,\ell_2,\cdots,\ell_N}||^2, \end{aligned} \tag{7}$$

where $A_1=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H'} \\ & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH'} \end{pmatrix}$, $A_2=\begin{pmatrix} a_{1,H'+2} & a_{1,H'+3} & \cdots & a_{1H} \\ & \vdots & & \\ a_{M,H'+2} & a_{M,H'+3} & \cdots & a_{MH} \end{pmatrix}$, $f'_{Qn+m,0,\cdots,0}=$

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

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 $J(z)$'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. (2) and (3) with $H'+1$, by setting $a_{ki'} = a'_{ki'}$, $b_{ij} = b'_{ij}$.

Q.E.D.

Conjecture The bound values in Theorem 3 are the exact ones.

2.2 Restricted Boltzmann machine

In this section, we show our results for the restricted Boltzmann machine.

Let $2 \leq L \in \mathbb{N}$ and $K \in \mathbb{N}$.

From now on, for simplicity, we denote

$$\{\{n\}\} = \begin{cases} 0, & \text{if } n = 0 \mod 2, \\ 1, & \text{if } n = 1 \mod 2, \end{cases} \quad \{\{(n_1, \cdots, n_m)\}\} = (\{\{n_1\}\}, \cdots, \{\{n_m\}\}).$$

Let $D = (d_{ij})$ be an $L \times K$ matrix with $|d_{ij}| < 1$.

Denote $D^J = \prod_{i=1}^L \prod_{j=1}^K d_{ij}^{J_{ij}}$, where $J = (J_{ij})$ is an $L \times K$ matrix with $J_{ij} \in \{0,1\}$.

Set $\mathcal{I} = \{I = (I_i) \in \{0,1\}^L | \{\{\sum_{i=1}^L I_i\}\} = 0\}$, and $D^I = \sum_{\substack{J:\{\{\sum_{i=1}^L J_{ij}\}\}=0 \\ \{\{\sum_{j=1}^K J_{ij}\}\}=I_i}} D^J$ for $I \in \mathcal{I}$.

Let

$$\Psi_B = \sum_{I \in \mathcal{I}} \left(\frac{D^I}{D^0} - \frac{D^{*I}}{D^{*0}}\right)^2. \tag{8}$$

From the eigenvalue analysis method, we obtain the following theorem.

Theorem 4 *(Aoyagi & Watanabe [10], Aoyagi [7])*

Case 1 *If* $L = 2$ *then* $c_{D^*}(\Psi_B) = 1/2$ *and its order* $\theta_{D^*}(\Psi_B) = \begin{cases} 2, & \text{if } K = 1, D^* = 0 \\ 1, & \text{otherwise } K \geq 2. \end{cases}$

Case 2 *If* $L = 3$ *then* $c_{D^*}(\Psi_B) = \begin{cases} 3/4, & \text{if } K = 1, D^* = 0 \\ 1/2, & \text{if } K = 1, D^* \neq 0, \prod_{i=1}^3 D_{i1}^* = 0 \\ 3/2, & \text{if } K = 1, \prod_{i=1}^3 D_{i1}^* \neq 0 \\ 3/2, & \text{if } K \geq 2, \end{cases}$

$$\text{and its order } \theta_{D^*}(\Psi_B) = \begin{cases} 3, & \text{if } K = 2, D^* = 0, \\ 2, & \text{if } K = 2, D^* \neq 0, b^*_{i_0 j} = b^*_{i_1 j} = 0 \text{ for } 1 \leq j \leq K, \\ 2, & \text{if } K = 2, b^*_{i_0 j_0} b^*_{i_1 j_0} \neq 0, b_{i_2 j_0} = b^*_{ij} = 0 \\ & \text{for } 1 \leq i \leq 3, 1 \leq j \leq K, j \neq j_0, \\ 1, & \text{otherwise}, \end{cases}$$

where $i_0, i_1, i_2 \in \{1, 2, 3\}$ *are different from each other and* $1 \leq j_0 \leq K$.

For its proof, we use the eigenvalues and the eigenvectors of the matrix $C_j = (c_j^{I,I'})$ where $d_j^I = \prod_{i=1}^L d_{ij}^{I_i}$, and $c_j^{I,I'} = d_j^{I''}$ with $\{\{I' + I''\}\} = I$, for $I, I', I'' \in \mathcal{I}$.

We obtain $c_{D^*}(\Psi_B)$ and its order $\theta_{D^*}(\Psi_B)$ for $L > K$ using a recursive blowing up.

Theorem 5 *(Aoyagi [7]) Assume that* $L > K$ *and* $D^* = 0$. *Then we have* $c_{D^*}(\Psi_B) = \frac{LK}{4}$ *and its order* $\theta_{D^*}(\Psi_B) = \begin{cases} 1, & \text{if } L > K + 1, \\ L, & \text{if } L = K + 1. \end{cases}$

We also bound values of $c_{D^*}(\Psi_B)$ for other cases.

Theorem 6 *(Aoyagi [7])*

Let $(d_{1j}, d_{2j}, \cdots, d_{Lj}) \neq 0$ *for* $j = 1, \ldots, K_0$ *and* $(d_{1j}, d_{2j}, \cdots, d_{Lj}) = 0$ *for* $j = K_0 + 1, \ldots, K$ *in* V, *where* V *is a sufficiently small neighborhood of* D^*.

Then we have

$$\frac{L(K-K_0)}{4} \leq c_{D^*}(\Psi_B) \leq \frac{L(K-K_0)}{4} + \frac{LK_0}{2}, \qquad \text{if } L > K - K_0$$

$$\frac{L(L-1)}{4} + \frac{LK_0}{2} \leq c_{D^*}(\Psi_B) \leq \frac{2K_0 + (L-1)(L-2)}{4} + \frac{LK_0}{2} \left(< \frac{LK_0}{2} + \frac{L(K-K_0)}{4} \right), \qquad \text{if } L \leq K - K_0.$$

2.3 Reduced rank regression

Let

$$\{w = (A_R, B_R) \mid A_R \text{ is an } H_R \times N_R \text{ matrix}, B_R \text{ is an } M_R \times H_R \text{ matrix}\},$$

be the set of parameters.

Let $\Psi_R = ||A_R B_R - A_R^* B_R^*||^2$

Theorem 7 *(Aoyagi & Watanabe [8])*

The log canonical threshold $c_{w^*}(\Psi_R)$ *and its order* $\theta_{w^*}(\Psi_R)$ *are given as the followings: Let* r *be the rank of* $A_R^* B_R^*$.

Case 1 *Let* $N_R + r \leq M_R + H_R$, $M_R + r \leq N_R + H_R$ *and* $H_R + r \leq M_R + N_R$.

(a) If $M_R + H_R + N_R + r$ *is even, then* $\theta_{w^*}(\Psi_R) = 1$ *and*

$$c_{w^*}(\Psi_R) = \frac{-(H_R + r)^2 - M_R^2 - N_R^2 + 2(H_R + r)M_R + 2(H_R + r)N_R + 2M_R N_R}{8}.$$

(b) If $M_R + H_R + N_R + r$ *is odd, then* $\theta_{w^*}(\Psi_R) = 2$ *and*

$$c_{w^*}(\Psi_R) = \frac{-(H_R + r)^2 - M_R^2 - N_R^2 + 2(H_R + r)M_R + 2(H_R + r)N_R + 2M_R N_R + 1}{8}.$$

Case 2 *Let* $M_R+H_R < N_R+r$. *Then* $\theta_{w^*}(\Psi_R) = 1$ *and* $c_{w^*}(\Psi_R) = \dfrac{H_RM_R - H_Rr + N_Rr}{2}$.

Case 3 *Let* $N_R+H_R < M_R+r$. *Then* $\theta_{w^*}(\Psi_R) = 1$ *and* $c_{w^*}(\Psi_R) = \dfrac{H_RN_R - H_Rr + M_Rr}{2}$.

Case 4 *Let* $M_R + N_R < H_R + r$. *Then* $\theta_{w^*}(\Psi_R) = 1$ *and* $c_{w^*}(\Psi_R) = \dfrac{M_RN_R}{2}$.

2.4 Remarks

The following remarks are useful for our proofs.

Remark 1 *If a regular function* $f(x) \neq 0$, $x \in \mathbb{R}^d$ *is non-degenerate with respect to its Newton polyhedron* Γ_+ *and if* $c = \min\{c' \geq 0 : c'\mathbf{e} \in \Gamma_+\} > 1$ *then we have* $c_0(f) = 1/c$ *and* $\theta_0(f) = \min\{d, \theta'\}$, *where* $\mathbf{e} = (1, \ldots, 1)^t$ *and* θ' *is the number of faces* $T \ni c\mathbf{e}$ *with dimension* $d-1$ *of* Γ_+ *[15].*

Remark 2 *Let*

$$f_1 = u_1^{s_{11}}u_2^{s_{12}}\cdots u_d^{s_{1d}}, f_2 = u_1^{s_{21}}u_2^{s_{22}}\cdots u_d^{s_{2d}}, \cdots, f_p = u_1^{s_{p1}}u_2^{s_{p2}}\cdots u_d^{s_{pd}}, g = u_1^{t_1}u_2^{t_2}\cdots u_d^{t_d}\mathrm{d}u$$

and Γ_+ *be the Newton diagram of* $f_1^2 + \cdots + f_p^2$.

Let $c = \min\{c' \geq 0 : c'(\mathbf{t}+\mathbf{e}) \in \Gamma_+\}$ *and* $\theta = \min\{d, \theta'\}$, *where* $\mathbf{e} = (1, \ldots, 1)^t$, $\mathbf{t} = (t_1, \ldots, t_d)^t$ *and* θ' *is the number of faces* $T \ni c(\mathbf{t}+\mathbf{e})$ *with dimension* $d-1$ *of* Γ_+.

Then, the largest pole of $\int$ *near* $0(f_1^2 + \cdots + f_p^2)^z g$ *is* $1/c$ *and its order is* θ. *In this case, the condition* $c > 1$ *is not necessary.*

Corollary 1 *Let* $f_\alpha(x_1^{(\alpha)}, \ldots, x_{d_\alpha}^{(\alpha)}) \geq 0$ *be a regular function and* $c_{w_\alpha^*}(f_\alpha) = c_\alpha$, $\theta_{w_\alpha^*}(f_\alpha) = \theta_\alpha$, *for* $\alpha = 1, \ldots, r$.

Then for $f(x_1^{(1)}, \ldots, x_{d_1}^{(1)}, \ldots, x_1^{(r)}, \ldots, x_{d_r}^{(r)}) = \sum_{\alpha=1}^r f_\alpha$ *and* $w^* = (w_1^*, \ldots, w_r^*)$, *we have* $c_{w^*}(f) = \sum_{\alpha=1}^r c_\alpha$, $\theta_{w^*}(f) = \sum_{\alpha=1}^r(\theta_\alpha - 1) + 1$.

(Proof)

By blowing ups at w_α^*, we may set

$$f_\alpha^z \mathrm{d}x^{(\alpha)} = ({u_1^{(\alpha)}}^{2s_1^{(\alpha)}} {u_2^{(\alpha)}}^{2s_2^{(\alpha)}} \cdots {u_{d_\alpha}^{(\alpha)}}^{2s_{d_\alpha}^{(\alpha)}})^z {u_1^{(\alpha)}}^{t_1^{(\alpha)}} {u_2^{(\alpha)}}^{t_2^{(\alpha)}} \cdots {u_{d_\alpha}^{(\alpha)}}^{t_{d_\alpha}^{(\alpha)}} \mathrm{d}u^{(\alpha)}$$

on one of local analytic coordinate systems and

$$c_\alpha = \frac{t_1^{(\alpha)}+1}{2s_1^{(\alpha)}} = \cdots = \frac{t_{\theta_\alpha}^{(\alpha)}+1}{2s_{\theta_\alpha}^{(\alpha)}} < \frac{t_i^{(\alpha)}+1}{2s_i^{(\alpha)}}, \text{ for } i \geq \theta_\alpha + 1.$$

Let $d = \sum_{\alpha=1}^r d_\alpha$ and

$$L = (\mathbf{l}_1, \cdots, \mathbf{l}_d) = \begin{pmatrix} l_{11}^{(1)} & l_{12}^{(1)} & \cdots & l_{1d}^{(1)} \\ \vdots & \vdots & \cdots & \vdots \\ l_{d_1 1}^{(1)} & l_{d_1 2}^{(1)} & \cdots & l_{d_1 d}^{(1)} \\ \vdots & \vdots & \cdots & \vdots \\ l_{11}^{(r)} & l_{12}^{(r)} & \cdots & l_{1d}^{(r)} \\ \vdots & \vdots & \cdots & \vdots \\ l_{d_r 1}^{(r)} & l_{d_r 2}^{(r)} & \cdots & l_{d_r d}^{(r)} \end{pmatrix}, l_{ij}^{(\alpha)} \in \mathbb{N}.$$

Set the mapping by

$$u =^L u' = (u_1'^{l_{11}^{(1)}} u_2'^{l_{12}^{(1)}} \cdots u_d'^{l_{1d}^{(1)}}, u_1'^{l_{21}^{(1)}} u_2'^{l_{22}^{(1)}} \cdots u_d'^{l_{2d}^{(1)}}, \cdots, u_1'^{l_{d_r1}^{(r)}} u_2'^{l_{d_r2}^{(r)}} \cdots u_d'^{l_{d_rd}^{(r)}}),$$

for $u' = (u_1', \cdots, u_d')$.

Then we have

$$f^z \prod_{\alpha=1}^r \mathrm{d}x^{(\alpha)} = (\sum_{\alpha=1}^r u_1^{(\alpha)^{2s_1^{(\alpha)}}} u_2^{(\alpha)^{2s_2^{(\alpha)}}} \cdots u_{d_\alpha}^{(\alpha)^{2s_{d_\alpha}^{(\alpha)}}})^z \prod_{\alpha=1}^r u_1^{(\alpha)^{t_1^{(\alpha)}}} u_2^{(\alpha)^{t_2^{(\alpha)}}} \cdots u_{d_\alpha}^{(\alpha)^{t_{d_\alpha}^{(\alpha)}}} \mathrm{d}u^{(\alpha)}$$

$$= (\sum_{\alpha=1}^r u_1'^{2\sum_{i=1}^{d_\alpha} s_i^{(\alpha)} l_{i1}^{(\alpha)}} \cdots u_d'^{2\sum_{i=1}^{d_\alpha} s_i^{(\alpha)} l_{id}^{(\alpha)}})^z u_1'^{\sum_{\alpha=1}^r \sum_{i=1}^{d_\alpha} (t_i^{(\alpha)}+1) l_{i1}^{(\alpha)} - 1} \cdots u_d'^{\sum_{\alpha=1}^r \sum_{i=1}^{d_\alpha} (t_i^{(\alpha)}+1) l_{id}^{(\alpha)} - 1} \mathrm{d}u',$$

on a local coordinate system u'.

If L is related with a face $\sigma(L)$ with dimension d of a refinement of the fan defined by the Newton diagram of $\sum_{\alpha=1}^r u_1^{(\alpha)^{2s_1^{(\alpha)}}} u_2^{(\alpha)^{2s_2^{(\alpha)}}} \cdots u_{d_\alpha}^{(\alpha)^{2s_{d_\alpha}^{(\alpha)}}}$, then there exists α_0 such that $\sum_{i=1}^{d_{\alpha_0}} s_i^{(\alpha_0)} l_{ij}^{(\alpha_0)} \leq \sum_{i=1}^{d_\alpha} s_i^{(\alpha)} l_{ij}^{(\alpha)}$, for $\alpha = 1, \ldots, r$ and $j = 1, \ldots, d$. Therefore, we have poles

$$\lambda_j := \frac{\sum_{\alpha=1}^r \sum_{i=1}^{d_\alpha} (t_i^{(\alpha)} + 1) l_{ij}^{(\alpha)}}{2 \sum_{i=1}^{d_{\alpha_0}} s_i^{(\alpha_0)} l_{ij}^{(\alpha_0)}}, j = 1, \ldots, d,$$

on a local coordinate system u'.

We have

$$\lambda_j \geq \sum_{\alpha=1}^r \frac{\sum_{i=1}^{d_\alpha} (t_i^{(\alpha)} + 1) l_{ij}^{(\alpha)}}{2 \sum_{i=1}^{d_\alpha} s_i^{(\alpha)} l_{ij}^{(\alpha)}} \geq \sum_{\alpha=1}^r c_\alpha,$$

and $\lambda_j = \sum_{\alpha=1}^r c_\alpha$, if and only if

$$(a) \quad l_{ij}^{(\alpha)} = 0, i \geq \theta_\alpha + 1, 1 \leq \alpha \leq r, \quad (b) \quad \sum_{i=1}^{d_1} s_i^{(1)} l_{ij}^{(1)} = \cdots = \sum_{i=1}^{d_r} s_i^{(r)} l_{ij}^{(r)}.$$

We can choose $\sum_{\alpha=1}^r \theta_\alpha - (r-1)$ independent vectors $\mathbf{l}_j$ satisfing (a) and (b), and this fact completes the proof.

Q.E.D.

Lemma 1 *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. If $g_1, \ldots, g_m \in I$, then $c_{w^*}(f_1^2 + \cdots + f_n^2)$ is greater than $c_{w^*}(g_1^2 + \cdots + g_m^2)$. In particular, if $g_1, \ldots, g_m$ generate the ideal $\mathcal{I}$ then*

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

3 Bayesian learning theory

In this section, we overview Bayesian learning theory, especially the stochastic complexity and the generalization error.

It is well known that Bayesian estimation is more appropriate than the maximum likelihood method when a learning machine is non-regular (Akaike [1], Mackay [21]).

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

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

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

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

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

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

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

which is the predictive density function.

Set

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

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

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

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

Let

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

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, Tishby & Solla [20], Amari, Fujita & Shinomoto [2], Amari & Murata [3]). $F(n)$ is known as the Bayesian criterion in Bayesian model selection (Schwarz [28]), stochastic complexity in universal coding (Rissanen [26], Yamanishi [37], Akaike's Bayesian criterion in optimization of hyperparameters (Akaike [1]) and evidence in neural network learning (Mackay [21]). 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 [32], [33]). We assume that the true density distribution $q(x)$ is included in the learning model, i.e., $q(x) = p(x|w_t^*)$ for $w_t^* \in W$, where W is the parameter space.

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:

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

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

$$F(n) = \lambda \log n - (\theta - 1) \log \log n + O(1), \tag{9}$$

where $O(1)$ is a bounded function of n, and if $G(n)$ has an asymptotic expansion,

$$G(n) \cong \frac{\lambda}{n} - \frac{\theta - 1}{n \log n} \text{ as } n \to \infty. \tag{10}$$

Therefore, our aim is to obtain λ and θ.

Note that for $Z = \{w : K(w) = 0\}$, $\lambda = c_Z(W, K(w)) = \sup\{c : |K|^{-c}$ is locally L^1 near $Z\}$, which is the log canonical threshold of $K(w)$ and its order $\theta = \theta_Z(W, K(w))$.

(a) 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 $q(x)$ 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

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

is included in the learning model, where $w_t^* = \{a_{ki}^*, b_{ij}^*; 1 \leq k \leq M, H + 1 \leq i \leq H + r, 1 \leq j \leq N\}$ and $f_k(x, w_t^*) = \sum_{i=H+1}^{H+r} (-a_{ki}^*) \tanh(\sum_{j=1}^{N} b_{ij}^* x_j)$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$.

Then λ and θ for the model corresponts the logcanonical threshold $c_{w^*}(||AB||^2)$ and its order with $Q = 2$ and $m = 1$, where A and B are defined in Definition 2.

(b) The normal mixture model with H peaks which is trained for estimating the true distribution with r peaks :

Consider a 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

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

where $w_t^* = \{a_{1i}^*, b_{ij}^*; H+1 \leq i \leq H+r, 1 \leq j \leq N\}$ and $\sum_{i=H+1}^{H+r} a_{1i}^* = -1$. Suppose that an *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with a compact support W where $\psi(w_t^*) > 0$.

Then λ and θ for the model corresponts the logcanonical threshold $c_{w^*}(||AB||^2)$ and its order with $Q = 1$, $M = 1$ and $m = 1$, where A and B are defined in Definition 2.

(a) and (b) as above show that λ in Eqs. (9), (10) for three layered neural networks and for normal mixture models are obtained by the same type of singularities, i.e., Vandermonde matrix type singularities. The paper [38], moreover, shows that λ for mixtures of binomial distributions is also obtained by Vandermonde matrix type singularities. These facts seem to imply that Vandermonde matrix type singularities are essential for learning theory.

(c) The restricted Boltzmann machine with L binary observable units and K binary hidden units y:

Set

$$p(x, y|a) = \frac{\exp(\sum_{i=1}^{L}\sum_{j=1}^{K} a_{ij}x_iy_j)}{Z(a)},$$

where

$$Z(a) = \sum_{x_i=\pm 1, y_j=\pm 1,} \exp(\sum_{i=1}^{L}\sum_{j=1}^{K} a_{ij}x_iy_j),$$

$x = (x_i) \in \{1, -1\}^L$ and $y = (y_j) \in \{1, -1\}^K$.

Consider a restricted Boltzmann machine

$$p(x|a) = \sum_{y_j=\pm 1} p(x,y|a) = \frac{\prod_{j=1}^K (\prod_{i=1}^L \exp(a_{ij}x_i) + \prod_{i=1}^L \exp(-a_{ij}x_i))}{Z(a)}$$

$$= \frac{\prod_{j=1}^K \prod_{i=1}^K \cosh(a_{ij})}{Z(a)}$$

$$\times \prod_{j=1}^K (2 \sum_{0\le p\le L/2} \sum_{i_1<\cdots<i_{2p}} x_{i_1}x_{i_2}\cdots x_{i_{2p}} \tanh(a_{i_1 j}) \tanh(a_{i_2 j}) \cdots \tanh(a_{i_{2p} j})).$$

Let $D = (d_{ij}) = (\tanh(a_{ij}))$. Denote $D^J = \prod_{i=1}^L \prod_{j=1}^K b_{ij}^{J_{ij}}$ and $x^J = \prod_{i=1}^L x_i^{\sum_{j=1}^K J_{ij}}$, where $J = (J_{ij})$ is an $L \times K$ matrix with $J_{ij} \in \{0,1\}$.

Then we have

$$p(x|a) = \frac{2^K \prod_{j=1}^K \prod_{i=1}^L \cosh(a_{ij})}{Z(a)} \sum_{J:\{\{\sum_{i=1}^L J_{ij}\}\}=0 \text{ for all } j} D^J x^J.$$

Let

$$Z(b) = \frac{Z(a)}{2^K \prod_{j=1}^K \prod_{i=1}^K \cosh(a_{ij})}.$$

Set $\mathcal{I} = \{I = (I_i) \in \{0,1\}^L | \{\{\sum_{i=1}^L I_i\}\} = 0\}$, and $D^I = \sum_{\substack{J:\{\{\sum_{i=1}^L J_{ij}\}\}=0 \\ \{\{\sum_{j=1}^K J_{ij}\}\}=I_i}} D^J$ for $I \in \mathcal{I}$.

Then we have

$$p(x|a) = \frac{1}{Z(b)} \sum_{I\in\mathcal{I}} D^I x^I$$

and $Z(b) = 2^K D^0$.

Assume that the true distribution is $p(x|a^*)$ with $a^* = (a_{ij}^*)$ and set $D^* = d^* = (d_{ij}^*) = (\tanh(a_{ij}^*))$.

Then λ and θ for the model corresponts the logcanonical threshold $c_{D^*}(\sum_{I\in\mathcal{I}}(\frac{D^I}{D^0} - \frac{D^{*I}}{D^{*0}})^2)$ and its order appeared in Section 2.2.

Remark 3 *Rusakov and Geiger [27] obtained λ and θ for the following class of Naive Bayesian networks with two hidden states and binary features:*

$$p(x|e,f,t) = t\prod_{i=1}^L e_i^{(1+x_i)/2}(1-e_i)^{(1-x_i)/2} + (1-t)\prod_{i=1}^K f_i^{(1+x_i)/2}(1-f_i)^{(1-x_i)/2}.$$

where $x \in \{1,-1\}^L$, $e = \{e_i\}_{i=1}^L \in \mathbb{R}^L$, $f = \{f_i\}_{i=1}^L \in \mathbb{R}^L$ and $0 \le t \le 1$. Our models with one hidden unit ($K = 1$) are obtained by setting $t = 1/2$, $\tanh(a_i) = 2e_i - 1$ and $f_i = -e_i$. The relation $f_i = -e_i$ creates a parameter space different from that of our models.

(d) The reduced rank regression model with M_R input units, N_R output units and H_R hidden units:

Let

$$\{w = (A_R, B_R) \mid A_R \text{ is an } H_R \times N_R \text{ matrix}, B_R \text{ is an } M_R \times H_R \text{ matrix}\},$$

be the set of parameters.

Denote the input value by x and the output value y of the reduced rank regression model, which is given by

$$y = A_R B_R x + (\text{noise}).$$

Consider the statistical model

$$p(y|x,w) = \frac{1}{(\sqrt{2\pi})_R^N} \exp(-\frac{1}{2}(y - A_R B_R x)^2).$$

Then λ and θ for the model corresponts the logcanonical threshold $c_{(A_R^*, B_R^*)}(||A_R B_R - A_R^* B_R^*||^2)$ and its order appeared in Section 2.3.

References

[1] Akaike, H.: Likelihood and Bayes procedure. Bayesian Statistics (Bernald J.M. eds.) University Press, Valencia, Spain (1980) 143–166

[2] Amari, S., Fujita, N., Shinomoto, S.: Four Types of Learning Curves. Neural Computation 4-4 (1992) 608–618

[3] Amari, S., Murata, N.: Statistical theory of learning curves under entropic loss. Neural Computation **5** (1993) 140–153

[4] Aoyagi, M.: The zeta function of learning theory and generalization error of three layered neural perceptron. RIMS Kokyuroku, Recent Topics on Real and Complex Singularities (2006) No. 1501, pp.153-167.

[5] Aoyagi, M.: Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network in Bayesian estimation, International Journal of Pure and Applied Mathematics (2009) vol. 52, No. 2, 177-204.

[6] Aoyagi, M.: Bayesian learning coefficient of generalization error and Vandermonde matrix type singularities, Communications in Statistics - Theory and Methods (2009) (to appear)

[7] Aoyagi, M.: Stochastic Complexity and Generalization Error of a Restricted Boltzmann Machine in Bayesian Estimation (preprint).

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

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

[10] Aoyagi, M., Watanabe, S.: Resolution of Singularities and Stochastic Complexity of Complete Bipartite Graph-Type Spin Model in Bayesian Estimation, Modeling Decisions for Artificial Intelligence, Lecture Notes in Computer Science, Springer (2007) No. 4617, pp. 443 - 454.

[11] Atiyah, M. F., Resolution of singularities and division of distributions. Comm. Pure and Appl. Math., 13, (1970) 145–150.

[12] Bernstein, I. N.: The analytic continuation of generalized functions with respect to a parameter. Functional Anal. Appl., **6** (1972) 26–40

[13] Björk, J. E.: Rings of differential operators. Amsterdam: North-Holland (1979)

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

[15] Fulton, W.: Introduction to toric varieties. Annals of Mathematics Studies Princeton University Press (1993) p131

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

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

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

[19] Kollár, J.: Singularities of pairs, Algebraic geometry-Santa Cruz 1995, Proc. Sympos. Pure Math., **62**, Amer. Math. Soc., Providence, RI, (1997 221–287

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

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

[22] Mustata, M.: Singularities of pairs via jet schemes, J. Amer. Math. Soc. **15** (2002), 599-615.

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

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

[25] Nishiyama, Y., Watanabe, S.: Asymptotic Behavior of Free Energy of General Boltzmann Machines in Mean Field Approximation. Technical report of IEICE NC2006-38 (2006) 1–6

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

[27] Rusakov, D., Geiger, D.: Asymptotic Model Selection for Naive Bayesian Networks. Journal of Machine Learning Research 6 (2005) 1–35

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

[29] Sturmfels, B.: 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**, Springer, New York, (2008) 351-364

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

[31] Watanabe, S.,"On the generalization error by a layered statistical model with Bayesian estimation," IEICE Trans., J81-A, pp. 1442-1452, 1998, (English version : Elect. and Comm. in Japan., John Wiley and Sons, 83 (6), pp.95-106, 2000.

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

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

[34] Watanabe, S., Hagiwara, K., Akaho, S., Motomura, Y., Fukumizu, K., Okada M., Aoyagi, M.: Theory and Application of Learning System. Morikita (2005) p. 195 (Japanese)

[35] WATANABE, K. and WATANABE, S. (2003). Upper Bounds of Bayesian Generalization Errors in Reduced Rank Regression. *IEICE Trans.* **J86-A** (3) 278-287 (In Japanese).

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

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

[38] Yamazaki, K., Aoyagi, M., Watanabe, S.: Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram, Neural Networks (to appear).

[39] Yamazaki, Y., Watanabe, S.: Singularities in mixture models and upper bounds of stochastic complexity. International Journal of Neural Networks, **16** (2003), 1029–1038

[40] Yamazaki, K., Watanabe, S.: Singularities in Complete Bipartite Graph-Type Boltzmann Machines and Upper Bounds of Stochastic Complexities. IEEE Trans. on Neural Networks **16**-2 (2005) 312–324