Entropy **2013**, *15*, 1-x; doi:10.3390/——

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Consideration on Singularities in Learning Theory and the Learning Coefficient

Miki Aoyagi

Department of Mathematics, College of Science & Technology, Nihon University, 1-8-14, Surugadai, Kanda, Chiyoda-ku, Tokyo 101-8308, Japan; E-Mail: aoyagi.miki@math.cst.nihon-u.ac.jp; Tel.: +81-3-5386-5878.

Received: 21 June 2013; in revised form: 29 August 2013 / Accepted: 30 August 2013 / Published: xx

---

Abstract: We consider the learning coefficients in learning theory and give two new methods for obtaining these coefficients in a homogeneous case: a method for finding a deepest singular point and a method to add variables. In application to Vandermonde matrix-type singularities, we show that these methods are effective. The learning coefficient of the generalization error in Bayesian estimation serves to measure the learning efficiency in singular learning models. Mathematically, the learning coefficient corresponds to a real log canonical threshold of singularities for the Kullback functions (relative entropy) in learning theory.

Keywords: learning coefficient; Kullback function (relative entropy); singular learning machine; resolution of singularities

---

1. Introduction

The purpose of a learning system is to estimate an unknown true density function (a probability model) that generates the data. Real data associated with, for example, genetic analysis, data mining, image or speech recognition, artificial intelligence, the control of a robot and time series prediction are very complicated and usually are not generated by a simple normal distribution. In Bayesian estimation, we set a learning model that is written in probabilistic form with parameters, and our goal is to estimate the true density function by *a predictive function* constructed with the learning model and such data. Therefore, the learning model should be abundant enough to capture the true density function's structure. 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. These are, however, singular learning models, which cannot be analyzed using the classic theory of regular statistical models, because singular learning models have a singular Fisher metric that is not always approximated by any quadratic form [1–4]. Therefore, it is difficult to analyze *their generalization errors*, which indicate how precisely the predictive function approximates the true density function.

In recent studies, Watanabe showed using algebraic geometry that the generalization and training errors are subject to a universal law and defined the model selection method "widely applicable information criterion" (WAIC) as a generalized Akaike information criterion (AIC) [5–9]. WAIC can even be applied to singular learning models, whereas AIC cannot. Using the WAIC, we can estimate the generalization errors from the training errors without any knowledge of the true probability density functions. The generalization errors relate to the generalization losses via the entropy of the true distribution. Thus, we can select a suitable model from among several statistical models by this method.

Computing the WAIC requires the values of the learning coefficient and the singular fluctuation, which are both birational invariants. Mathematically, the learning coefficient is the log canonical threshold (Definition 1) of the Kullback function (relative entropy), and the singular fluctuation is known as a statistically generalized log canonical threshold, which is obtained theoretically from the learning coefficient (Equation (1) in Section 2). These values can be obtained by Hironaka's Theorem (Appendix A). However, it is still difficult to obtain these within learning theory for several reasons, such as degeneration with respect to their Newton polyhedra and non-isolation of their singularities [10]. Moreover, in algebraic geometry and algebraic analysis, these studies are usually done over an algebraically closed field [11,12]; many differences exist for real and complex fields. For example, log canonical thresholds over the complex field are less than one, whereas those over the real field are not necessarily so. We, therefore, cannot apply results over an algebraically closed field to our current situation directly (Appendix B). One of the bottlenecks in learning theory is to obtain the learning coefficients and the singular fluctuation.

In this paper, we consider the learning coefficient of "Vandermonde matrices-type singularities" in statistical learning theory. The reason why we contribute only to such singularities is that the Vandermonde matrix type is generic and essential in learning theory. These log canonical thresholds give the learning coefficients of normal mixture models, three-layered neural networks and mixtures of binomial distributions, which are widely used as effective learning models (Sections 3.1 and 3.2 and [13]). Moreover, we prove Theorem 2 (the method for finding a deepest deepest singular point) and Theorem 3 (the method to add variables), which are very beneficial to obtain the log canonical threshold for the homogeneous case. Theorem 2 indicates the best point of singularities that gives the log canonical threshold. Therefore, this theorem is useful for the reduction of the number of blowup processes. Theorem 3 improves our recursive blowup method by simplifying coordinate system changes with added variables. These two theorems enable us to obtain a new bound for the log canonical thresholds of Vandermonde matrix-type singularities in Theorem 5. These bounds are much tighter than those in [14].

In the past few years, we have obtained the learning coefficients for reduced rank regression [15], for the three-layered neural network with one input unit and one output unit [16,17], and for the normal mixture models with a dimension of one [18]. The paper [14] derived bounds on the learning coefficients

for the Vandermonde matrix-type singularities and explicit values under some conditions. The learning coefficients for the restricted Boltzmann machine [19] have also been considered recently. Ref [20–22], respectively, obtained these for naive Bayesian networks and for directed tree models with hidden variables. These results give partial answers for the learning coefficients.

The rest of the paper is in three sections. Section 2 summarizes the framework of Bayesian learning models. In Section 3, we demonstrate our main theorems and consider the log canonical threshold of Vandermonde matrix-type singularities (Definition 3). We finish with our conclusions in Section 4.

2. Learning Coefficients and Singular Fluctuations

In this section, we present the theory of learning coefficients and singular fluctuations. Let $q(x)$ be a true probability density function of variables, $x \in \mathbb{R}^N$, and let $x^n := \{x_i\}_{i=1}^n$ be n training samples selected from $q(x)$ independently and identically. Consider a learning model that is written in probabilistic form as $p(x|w)$, where $w \in W \subset \mathbb{R}^d$ is a parameter. The purpose of the learning system is to estimate $q(x)$ from x^n using $p(x|w)$. Let $\psi(w)$ be an *a priori* probability density function on the parameter set, W, and $p(w|x^n)$ 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:

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

Let us define for the inverse temperature, β:

$$E_w[f(w)] = \frac{\int dw f(x)\psi(w)\prod_{i=1}^{n} p(x_i|w)^\beta}{\int dw \psi(w)\prod_{i=1}^{n} p(x_i|w)^\beta}$$

We usually set $\beta = 1$.

We then have a predictive density function, $p(x|X^n) = E_w[p(x|w)]$, which is the average inference of the Bayesian density function.

We next introduce the Kullback function, $K(q||p)$, and the empirical Kullback function, $K_n(q||p)$, for density functions $p(x), q(x)$:

$$K(q||p) = \sum_x q(x)\log\frac{q(x)}{p(x)}$$

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

The function, $K(p||q)$, always has a non-negative value and satisfies $K(q||p) = 0$, if and only if $q(x) = p(x)$.

The Bayesian generalization error, B_g, Bayesian training error, B_t, Gibbs generalization error, G_g, and Gibbs training error, G_t, are defined as follows:

$$B_g = K(q(x)||E_w[p(x|w)]))$$

$$B_t = K_n(q(x)|E_w[p(x_i|w)])$$

$$G_g = E_w[K(q(x)||p(x|w))]$$

and

$$G_t = E_w[K_n(q(x)||p(x|w))]$$

The most important of these is the Bayesian generalization error. This error describes how precisely the predictive function approximates the true density function.

Watanabe [6,7,23] proved the following four relations:

$$E[B_g] = \frac{\lambda + \nu\beta - \nu}{n\beta} + o(\frac{1}{n})$$

$$E[B_t] = \frac{\lambda - \nu\beta - \nu}{n\beta} + o(\frac{1}{n})$$

$$E[G_g] = \frac{\lambda + \nu\beta}{n\beta} + o(\frac{1}{n})$$

$$E[G_t] = \frac{\lambda - \nu\beta}{n\beta} + o(\frac{1}{n})$$

Thu,s we have:

$$E[B_g] = E[B_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

and

$$E[G_g] = E[G_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

Eliminating the expectation of the true probability density function from the above four errors and setting:

$$BL_g = -\sum_x q(x) \log E_w[p(x|w)]$$

$$BL_t = -\frac{1}{n}\sum_{i=1}^{n} \log E_w[p(x_i|w)]$$

$$GL_g = -E_w[\sum_x q(x) \log p(x|w)]$$

$$GL_t = -E_w[\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|w)]$$

we then have:

$$E[BL_g] = E[BL_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

and

$$E[GL_g] = E[GL_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

These two equations constitute the WAIC and show that we can estimate the Bayesian and Gibbs generalization errors from the Bayesian and Gibbs training errors without any knowledge of the true probability density functions. Training errors are calculated from training samples, x_i, using a learning

model, p. In real applications or experiments, we usually do not know the true distribution, but only the values of the training errors. Our purpose is to estimate the true distribution from the training samples, showing that these relations are effective. We can select a suitable model from among several statistical models by observing these values.

Let λ denote a learning coefficient and ν a singular fluctuation, both of which are birational invariants. Mathematically, λ is equal to the log canonical threshold introduced in Definition 1 and Appendix B. For regular models, $\lambda = \nu = d/2$ holds, where d is the dimension of the parameter space.

The difference between the Bayesian and Gibbs training errors converges to ν/n:

$$n\beta(E[G_t] - E[B_t]) \to \nu$$

These relations were shown using the resolution of singularities and the Schwarz distribution.

From the learning coefficient, λ, and its order, θ, the value, ν, is obtained theoretically as follows. Let $\xi(u)$ be an empirical process defined on the manifold obtained by a resolution of singularities, and $\sum_{u^*}$ denote the sum of local coordinates that attain the minimum λ and the maximum θ. We then have:

$$\nu = \frac{1}{2} E_\xi \frac{\int_0^\infty dt \sum_{u^*} \int du \xi(u) t^{\lambda - 1/2} e^{-\beta t + \beta\sqrt{t}\xi(u)}}{\int_0^\infty dt \sum_{u^*} \int du t^{\lambda - 1/2} e^{-\beta t + \beta\sqrt{t}\xi(u)}} \tag{1}$$

$\xi(u)$ is a random variable of a Gaussian process with mean zero and variance two. Our purpose in this paper is to obtain λ.

To assist in achieving this aim, we use the desingularization approach from algebraic geometry (*cf.* Appendix A). It is a new problem in algebraic geometry to obtain the desingularization of the Kullback functions, because the singularities of these functions are very complicated, and as such, most of these have not yet been investigated.

3. Main Theorems and Vandermonde Matrix-Type Singularities

We denote constants, such as a^*, b^* and w^*, by the suffix $*$. Additionally, for simplicity, we use the notation: $w = \{a_{ki}, b_{ij}\}_{1 \le i \le H}$ instead of: $w = \{a_{ki}, b_{ij}\}_{1 \le k \le M, 1 \le i \le H, 1 \le j \le N}$, because we always have $1 \le k \le M$ and $1 \le j \le N$ 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 *For a real analytic function, f, in a neighborhood, U, of w^* and a C^∞ function ψ with a compact support, let $\lambda_{w^*}(f, \psi)$ be the largest pole of $\int_U |f|^z \psi dw$ and $\theta_{w^*}(f, \psi)$ be its order. If $\psi(w^*) \neq 0$, then we denote $\lambda_{w^*}(f) = \lambda_{w^*}(f, \psi)$ and $\theta_{w^*}(f) = \theta_{w^*}(f, \psi)$, because the log canonical threshold and its order are independent of ψ.*

Definition 2 *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 3 *Fix $Q \in \mathbb{N}$.*

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

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

(the superscript, t, denotes matrix transposition).

a_{ki} *and* b_{ij} $(1 \le k \le M, 1 \le i \le H, 1 \le j \le N)$ *are 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 $1 \le j \le r$ *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 $j \neq j'$*.*

Example 1 *If* $N = Q = 1$ *and* $r = 0$*, then we have:* $B = \begin{pmatrix} b_{11} & b_{11}^2 & \cdots & b_{11}^H \\ b_{21} & b_{21}^2 & \cdots & b_{21}^H \\ & & \vdots & \\ b_{H1} & b_{H1}^2 & \cdots & b_{H1}^H \end{pmatrix}$.

This matrix is a Vandermonde matrix.

Example 2 *If* $Q = 1$*,* $M = 1$*,* $H = 2, N = 2$ *and* $r = 1$*, then we have:* $A = \begin{pmatrix} a_{11} & a_{12} & a^*_{1,3} \end{pmatrix}$ *and* $B = \begin{pmatrix} b_{11} & b_{12} & b_{11}^2 & b_{11}b_{12} & b_{12}^2 & b_{11}^3 & b_{11}b_{12}^2 & b_{11}^2b_{12} & b_{12}^3 \\ b_{21} & b_{22} & b_{21}^2 & b_{21}b_{22} & b_{22}^2 & b_{21}^3 & b_{21}b_{22}^2 & b_{21}^2b_{22} & b_{22}^3 \\ b^*_{31} & b^*_{32} & {b^*_{31}}^2 & b^*_{31}b^*_{32} & {b^*_{32}}^2 & {b^*_{31}}^3 & b^*_{31}{b^*_{32}}^2 & {b^*_{31}}^2b^*_{32} & {b^*_{32}}^3 \end{pmatrix}$.

In this paper, we denote:

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

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

Furthermore, we denote: $\mathbf{a}^* = \begin{pmatrix} a^*_{1,H+1} \\ \vdots \\ a^*_{M,H+1} \end{pmatrix}$ and

$$(A_{M,H}, \mathbf{a}^*) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a^*_{1,H+1} \\ a_{21} & a_{22} & \cdots & a_{2H} & a^*_{2,H+1} \\ & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a^*_{M,H+1} \end{pmatrix}$$

Theorem 1 ([18]) *Consider a sufficiently small neighborhood, U, of*

$$w^* = \{a^*_{ki}, b^*_{ij}\}_{1 \le i \le H}$$

and variables, $w = \{a_{ki}, b_{ij}\}_{1 \le i \le H}$, 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 3. 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 $H_0 + \cdots + H_{r'} = H$.

We then have:

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

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

where: $w_1^{(0)*} = \{a^*_{k,i}, 0\}_{1 \le i \le H_0}$,

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

Theorem 2 (Method for finding a deepest singular point) *Let* $f_1(w_1, \ldots, w_d)$, ..., $f_m(w_1, \ldots, w_d)$ *be homogeneous functions of* $w_1, \cdots, w_j$ $(j \leq d)$ *with the degree,* n_i, *of* $w_1, \cdots, w_j$. *Furthermore, let ψ be a C^∞ function, such that* $\psi_{(0,\cdots,0,w^*_{j+1},\cdots,w^*_d)} \geq \psi_{(w^*_1,\cdots,w^*_d)}$ *and ψ_w is homogeneous of* $w_1, \cdots, w_j$ *in a small neighborhood of* $(0, \cdots, 0, w^*_{j+1}, \cdots, w^*_d)$.

Then, we have:

$$\lambda_{(0,\cdots,0,w^*_{j+1},\cdots,w^*_d)}(f_1^2+\cdots+f_m^2,\psi) \leq \lambda_{(w^*_1,\cdots,w^*_j,w^*_{j+1},\cdots,w^*_d)}(f_1^2+\cdots+f_m^2,\psi)$$

(Proof)

Let d be the degree of $w_1,\cdots,w_j$ for ψ in a neighborhood of $(0,\cdots,0,w^*_{j+1},\cdots,w^*_d)$. Let us construct the blowup of f_1, ..., f_m along the submanifold, $\{v=0,w_i=0,1\leq i\leq j\}$. Let $w_i=vw'_i$ for $1\leq i\leq j$. We have: $v^{n_i}f_i(w'_1,\ldots,w'_d)$ and $(f_1(w)^2+f_2(w)^2+\cdots+f_m(w)^2)^z\psi dwdv=(v^{2n_1}f_1^2(w')+v^{2n_2}f_2^2(w')+\cdots+v^{2n_m}f_m(w')^2)^z\psi(w')v^{d+j}dw'dv$. Because: $v^{2n_i}f_i^2(w'_1,\ldots,w'_d)\leq f_i^2(w'_1,\ldots,w'_d)$ for $|v|<1$, we have: $v^{2n_1}f_1^2+\cdots+v^{2n_m}f_m^2\leq f_1^2+\cdots+f_m^2$, and, hence, by Lemma 1 in Appendix C:

$$\lambda_{(0,\cdots,0,w^*_{j+1},\cdots,w^*_d)}(f_1^2+\cdots+f_m^2,\psi)\leq \min\{d+j+1,\lambda_{w^*}(f_1^2+\cdots+f_m^2,\psi)\}$$

Furthermore, we consider the construction of the blowup of: f_1, ..., f_m along the submanifold: $\{w_i=0,1\leq i\leq j\}$, for which we have

$$\lambda_{(0,\cdots,0,w^*_{j+1},\cdots,w^*_d)}(f_1^2+\cdots+f_m^2,\psi)\leq d+j$$

Q.E.D.

In general, it is not true that:

$$\lambda_{w_0}(f_1^2+\cdots+f_m^2,\psi)\leq\lambda_{w^*}(f_1^2+\cdots+f_m^2,\psi)$$

even if $w_0\in\mathbb{R}^d$ satisfies:

$$f_i(w_0)=\frac{\partial f_i}{\partial w_j}(w_0)=0,1\leq i\leq m,1\leq j\leq d$$

Example 3 *Let* $f_1=x(x-1)^2$, $f_2=(y^2+(x-1)^2)((y-1)^6+x)$ *and* $f_3=(z^2+(x-1)^2)((z-1)^6+x)$. *Then, we have:* $f_1=f_2=f_3=\frac{\partial f_1}{\partial x}=\frac{\partial f_2}{\partial y}=\frac{\partial f_2}{\partial x}=\frac{\partial f_3}{\partial z}=\frac{\partial f_3}{\partial x}=0$ *if and only if* $x=1,y=0,z=0$.

In this case, we have $\lambda_{(1,0,0)}(f_1^2+f_2^2+f_3^2)=1/4+1/4+1/4>\lambda_{(0,1,1)}(f_1^2+f_2^2+f_3^2)=1/2+1/12+1/12$.

Theorem 3 (Method to add variables) *Let* $f_1(w_1,\ldots,w_d)$, ..., $f_m(w_1,\ldots,w_d)$ *be homogeneous functions of* $w_1,\cdots,w_d$ *of the degree,* n_i, *in* $w_1,\cdots,w_d$. *Set:* $f'_1(w_2,\ldots,w_d)=f_1(1,w_2,\ldots,w_d)$, ..., $f'_m(w_2,\ldots,w_d)=f_m(1,w_2,\ldots,w_d)$. *If* $w^*_1\neq 0$, *then we have:*

$$\lambda_{(w^*_1,\cdots,w^*_d)}(f_1^2+\cdots+f_m^2)=\lambda_{(w^*_2/w^*_1,\cdots,w^*_d/w^*_1)}(f_1'^2+\cdots+f_m'^2)$$

(Proof) Set $w'_2=w_2/w_1,\ldots,w'_d=w_d/w_1$. Then, we have:

$$f_i(w_1,w_2,\ldots,w_d)=w_1^{n_i}f_i(1,w'_2,\ldots,w'_d)=w_1^{n_i}f'_i$$

Since $w_1^{n_i}\neq 0$ on a small neighborhood of w^*_1, there exist positive real numbers, C,C', such that:

$$C(f_1^2+\cdots+f_m^2)\leq f_1'^2+\cdots+f_m'^2\leq C'(f_1^2+\cdots+f_m^2)$$

This completes the proof by Lemma 1 in Appendix C. Q.E.D.

Remark 1

The above theorem shows that we can set nonzero constants as variables to obtain the same log canonical threshold. However, in general, this is not true.

(1) Consider the function $f(x,y) = (y-1+x^2)^2$. We have $\lambda_{f=0}(f) = 1/2$, whereas $\lambda_{f(x,1)=0}(f(x,1)) = 1/4$.
(2) Consider the function $f(x_1,x_2,x_3,x_4,x_5,y) = (x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+y-1)^2$. We have $\lambda_{f=0}(f) = 1/2$, whereas $\lambda_0(f(x_1,x_2,x_3,x_4,x_5,1)) = 5/4$.

The second example shows that the following theorem over the complex field is not true over the real field.

Theorem 4 *[11]*

Let $f(x_1,\cdots,x_d,y)$ be a **holomorphic** *function near zero, and for a hyperplane H, let $g = f|_{y=0}$ (or $g = f_H$) denote the restriction of f to $y=0$ (or H). Then, $\lambda_{g=0}(g) \leq \lambda_{f=0}(f)$.*

Define: $\langle {k \atop l} \rangle = \dfrac{k!}{l!(k-l)!}$,

Theorem 5 *We use the same notation as in Theorem 1. Let:*

$$\text{bound}_1 = \min\{\frac{(H-i+1)N + d_i(s) + d_i'(s) + d_i''(s)}{2(count(i,s,k(s))-1)Q+2} : 1 \leq i \leq s, 1 \leq k(1), \cdots, k(s) \leq N, 1 \leq s \leq H\}$$

where: $count(i,s,j) = \#\{i_1 : i \leq i_1 \leq s, k(i_1) = j\}$, $C(i,s) = \#\{count(i,s,j) = 0, 1 \leq j \leq N\}$,

$$\begin{aligned}
d_i(s) &= (N-1)Q\sum_{s_1=i}^{s}(count(i,s_1,k(s_1))-1) \\
d_i'(s) &= M(i-1)\{(count(i,s,k(s))-1)Q+1\} \\
&\quad +QM \sum_{\substack{s_1=i, \\ count(i,s,k(s))>count(i,s_1,k(s_1))}}^{s-1} (count(i,s,k(s)) - count(i,s_1,k(s_1))), \\
d_i''(s) &= \begin{cases}
0, & \text{if } count(i,s,k(s)) = 1, \\
(H-s)\{C(i,s)Q + (N-1)Q(count(i,s,k(s))-2)\}, & \\
\quad \text{if } count(i,s,k(s)) \geq 2, N-1 \leq M, & \\
(H-s)\{C(i,s)Q + MQ(count(i,s,k(s))-2)\}, & \\
\quad \text{if } count(i,s,k(s)) \geq 2, C(i,s) \leq M < N-1, & \\
(H-s)\{MQ(count(i,s,k(s))-1)\}, & \\
\quad \text{if } count(i,s,k(s)) \geq 2, M \leq C(i,s). &
\end{cases}
\end{aligned}$$

Furthermore, let: $\text{bound}_2 = \dfrac{NH + \sum_{i=0}^{k'-1} MQ(k'-i)\langle {N+Qi \atop N-1} \rangle}{2+2Qk'}$,
where: $k' = \max\{i \in \mathbb{Z}; NH \geq M\sum_{i'=0}^{i-1}(1+Qi')\langle {N+Qi' \atop N-1} \rangle\}$, *and let:*

$$\text{bound}_3 = \frac{MH}{2}$$

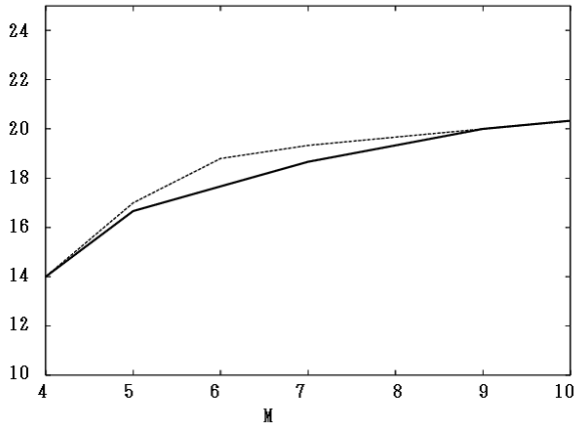
We have

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

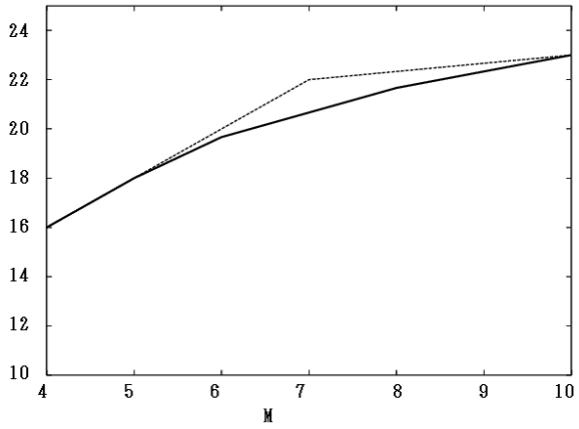
The proof appears in Appendix C.

Figures 1a–d show the values of new bounds, $\min\{\text{bound}_1, \text{bound}_2, \text{bound}_3\}$, for (a) $H = 7, N = 6$, (b) $H = 8, N = 6$, (c) $H = 7, N = 7$ and (d) $H = 8, N = 7$ with $Q = 2$, respectively. We compare these values with those obtained by the past work in [14]. In the figures, the horizontal axis is the number, M, and the vertical one, the value of such bounds. The dashed lines indicate the bounds obtained by the past work. These figures show that new bounds are not greater than old ones.

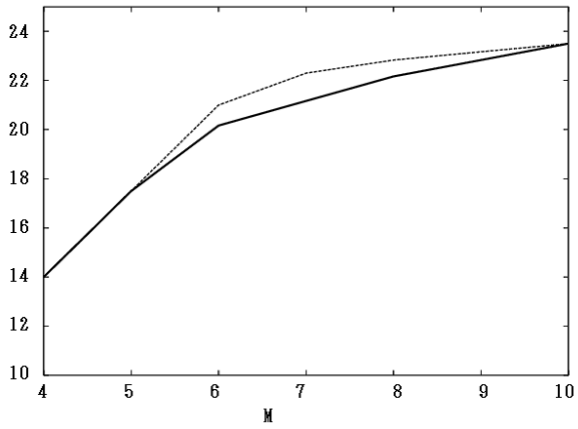
Figure 1. The values of new bounds, $\min\{\text{bound}_1, \text{bound}_2, \text{bound}_3\}$, for (**a**) $H = 7, N = 6$; (**b**) $H = 8, N = 6$; (**c**) $H = 7, N = 7$ and (**d**) $H = 8, N = 7$ with $Q = 2$, compared with the bounds obtained by the past work in [14].



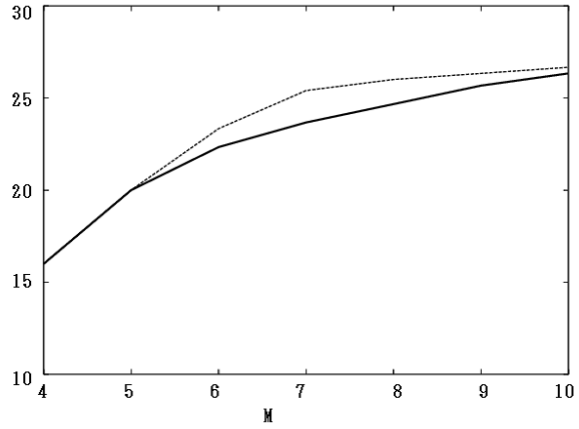
(**a**) $H = 7, N = 6$



(**b**) $H = 8, N = 6$



(**c**) $H = 7, N = 7$



(**d**) $H = 8, N = 7$

In paper [24], we had exact values for $N = 1$:

$$\lambda_0(||A_{M,H}B^{(Q)}_{H,1}||^2) = \frac{MQk(k+1) + 2H}{4(1+kQ)}$$

where: $k = \max\{i \in \mathbb{Z} : 2H \geq M(i(i-1)Q + 2i)\}$, and we had:

$$\theta = \begin{cases} 1, & \text{if } 2H > M(k(k-1)Q + 2k) \\ 2, & \text{if } 2H = M(k(k-1)Q + 2k) \end{cases}$$

We had other exact values when H is small on paper [14]. Both sets of exact values are the bounded values in Theorem 5.

3.1. A Learning Coefficient for a Three-Layered Neural Network

Consider the three-layered neural network with N input units, H hidden units and M output units, which are trained for estimating the true distribution with r hidden units. Their learning coefficients, λ, are as follows [14,24]:

$\lambda = \frac{Mr}{2} + \min\{\lambda_0(||A_{M,H_0}B^{(2)}_{H_0,N}||^2) + \sum_{\alpha=1}^{r}\lambda_0(||(A_{M,H_\alpha-1}, \mathbf{a}^*)B^{(1)}_{H_\alpha,N}||^2) : H_0+H_1+\cdots+H_r = H, \mathbf{a}^* \neq 0\}$.

3.2. A Learning Coefficient for a Normal Mixture Model

Consider normal mixture models with H peaks and the true distribution with r peaks. Then, their learning coefficients, λ, are as follows [14,18]:

$\lambda = \frac{r-1}{2} + \min\{\sum_{\alpha=1}^{r}\lambda_0(||(A_{M,H_\alpha-1}, \mathbf{a}^*)B^{(1)}_{H_\alpha,N}||^2) : \sum_{\alpha=1}^{r} H_\alpha = H, \mathbf{a}^* \neq 0\}$

In particular, we have for $N = 1$:

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

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

4. Conclusions

In this paper, we prove two theorems, Theorem 2 (the method for finding a deepest singular point) and Theorem 3 (the method to add variables) for obtaining learning coefficients in a homogeneous case. By applying these methods to Vandermonde matrix-type singularities and using the inclusion of ideals and recursive blowup from algebraic geometry, we found new bounds on learning coefficients for Vandermonde matrix-type singularities. These bounds are much tighter than those in [14]. Our future research aim is to improve our methods and to obtain exact values for the general machine model.

The learning coefficients from our recent results have been used very effectively by Drton [25,26] for model selection, using a method called "singular Bayesian information criterion (sBIC)", which can be applied to singular models, where the assumptions supporting the use of the standard BIC do not hold. Our theoretical results introduce a mathematical measure of precision to numerical calculations, such as Markov chain Monte Carlo (MCMC). Nagata and Watanabe [27,28] gave a mathematical foundation for analyzing and developing the precision of the MCMC method using our theoretical values of marginal likelihoods.

Acknowledgments

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

Conflicts of Interest

The authors declare no conflict of interest.

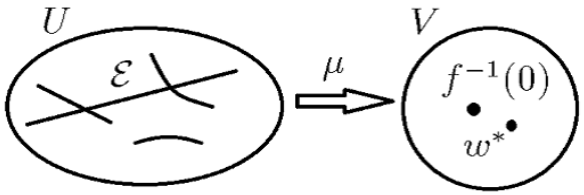
Appendix A

We introduce Hironaka's Theorem on desingularization.

Theorem 6 *[Desingularization, Hironaka (1964), (Figure A1)]*

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

Figure A1. Hironaka's Theorem: diagram of desingularization, μ, of f: $\mathcal{E}$ maps to $f^{-1}(0)$. $U - \mathcal{E}$ is isomorphic to $V - f^{-1}(0)$ by μ, where V is a small neighborhood of w^* with $f(w^*) = 0$.



Appendix B

The learning coefficient is the log canonical threshold of the Kullback function (relative entropy). In this section, we explain its difference for real and complex fields. Let f be a nonzero holomorphic function over $\mathbb{C}$ or an analytic function over $\mathbb{R}$ on a smooth variety, Y, and let $Z \subset Y$ be a closed subscheme. The log canonical threshold, $\lambda_Z(Y, f)$, is defined analytically as:

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

over $\mathbb{C}$, and:

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

over $\mathbb{R}$ [11,12]. It is known that if f is a polynomial or a convergent power series, then $\lambda_0(\mathbb{C}^d, f)$ 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 [29–31]. The log canonical threshold, $\lambda_Z(Y, f)$, also corresponds to the largest pole of $\int_{\text{near } Z} |f|^{2z}\psi(w)dw$ over $\mathbb{C}$, ($\int_{\text{near } Z} |f|^{z}\psi(w)dw$ over $\mathbb{R}$), where $\psi(w)$ is a $C^\infty-$ function with a compact support, such that $\psi(w) \neq 0$ on Z.

Appendix C

Using the blowup process and the method to add variables together with the inductive method for s, we demonstrate Theorem 5

We give below Lemma 1, as it is frequently used in the proofs.

Lemma 1 ([18,24,32]) *Let U be a neighborhood of $w^* \in \mathbb{R}^d$. Let I be the ideal generated by $f_1, \ldots, f_n$, which are analytic functions defined on U.*
(1) If $g_1^2 + \ldots + g_m^2 \leq f_1^2 + \cdots + f_n^2$, then $\lambda_{w^*}(g_1^2 + \cdots + g_m^2) \leq \lambda_{w^*}(f_1^2 + \cdots + f_n^2)$.
(2) If $g_1, \ldots, g_m \in I$, then $\lambda_{w^*}(g_1^2 + \cdots + g_m^2) \leq \lambda_{w^*}(f_1^2 + \cdots + f_n^2)$. In particular, if $g_1, \ldots, g_m$ generate the ideal I, then $\lambda_{w^*}(f_1^2 + \cdots + f_n^2) = \lambda_{w^*}(g_1^2 + \cdots + g_m^2)$.

The following lemma is also used in the proofs.

Lemma 2 ([19]) *Let I, J be the ideals generated by $f_1(w), \ldots, f_n(w)$ and $g_1(w'), \ldots, g_m(w')$, respectively. If w and w' are different variables, then*

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

Step 1

Let us consider the following procedure from $s = 1$ to $s = H$, and the generators of the ideal:

$$\mathcal{J} = \left\langle \begin{pmatrix} a_{i_0 1} & \cdots & a_{i_0 H} \end{pmatrix} \begin{pmatrix} b_{11}^{\ell_1} \cdots b_{1N}^{\ell_N} \\ b_{21}^{\ell_1} \cdots b_{2N}^{\ell_N} \\ \vdots \\ b_{H1}^{\ell_1} \cdots b_{HN}^{\ell_N} \end{pmatrix} : 1 \leq i_0 \leq M, \sum_{i=1}^{N} \ell_i = nQ + 1, n \geq 0 \right\rangle.$$

By constructing the blowup repeatedly and choosing one branch of the blowup process, we show the following (i)∼(v) in this subsection:

(i) $k(1), \cdots, k(s-1) \in \{1, 2, \cdots, N\}$,
(ii) $count(i_1, i_2, j) = \#\{k(i) = j \mid i_1 \leq i \leq i_2\}$ for $1 \leq j \leq N$,
(iii) $b_{ij} = \begin{cases} v_1 v_2 \cdots v_i b'_{ij}, & i < s-1, \\ v_1 v_2 \cdots v_{s-1} b'_{ij}, & i \geq s-1, \end{cases}$ and $b'_{ik(i)} = 1$,
(iv) The Jacobian is:

$$\prod_{i=1}^{H} \prod_{j=1}^{N} db_{ij} = \prod_{i=1}^{s-1} v_i^{(H-i+1)N-1+d_i} dv_i \prod_{i=1}^{s-1} \prod_{j=1}^{N} db'_{ij}$$

and

$$d_i = (N-1)Q \sum_{s_1=i}^{s-1} (count(i, s_1, k(s_1)) - 1),$$

(v)

$$\mathcal{J} = \left\langle a_{i_0 i_1} b_{i_1 k(i_1)} \prod_{k(i)=k(i_1),1\le i\le i_1-1} (b^Q_{i_1,k(i_1)} - b^Q_{i,k(i_1)}) \ : \ 1 \le i_0 \le M, 1 \le i_1 \le s-1 \right\rangle \quad (2)$$

$$+ \left\langle \begin{pmatrix} a_{i_0 s} & \cdots & a_{i_0 H} \end{pmatrix} \begin{pmatrix} b^{nQ+1}_{sj} \prod_{k(i)=j,1\le i\le s-1}(b^Q_{sj} - b^Q_{ij}) \\ \vdots \\ b^{nQ+1}_{Hj} \prod_{k(i)=j,1\le i\le s-1}(b^Q_{Hj} - b^Q_{ij}) \end{pmatrix} \ : \ 1 \le i_0 \le M, j = 1, \cdots, N, n \ge 0 \right\rangle$$

$$+ \left\langle \begin{pmatrix} a_{i_0 s} & \cdots & a_{i_0 H} \end{pmatrix} \begin{pmatrix} b^{\ell_1}_{s1} \cdots b^{\ell_N}_{sN} \\ \vdots \\ b^{\ell_1}_{H1} \cdots b^{\ell_N}_{HN} \end{pmatrix} : 1 \le i_0 \le M, \sum_{i=1}^{N} \ell_i = nQ+1, n \ge 1, \forall \ell_j < Qn+1 \right\rangle.$$

By Theorem 3, we can set $b'_{ik(i)}$ as a variable.

Now, we show the above by the inductive method.

Define $1 \le k(s) \le N$. Construct the blowup along $\{b'_{ij} = 0, s \le i \le H, 1 \le j \le N\}$. Set $b'_{ij} = v_s b''_{ij}$ for $s \le i \le H, 1 \le j \le N$, and set $b''_{sk(s)} = 1$.

By constructing the blowup along $\{v_i = b''_{sj} = 0, 1 \le j \le N, j \ne k(s)\}$ repeatedly, and by choosing one branch of the blowup process, set $b''_{sj} = b'''_{sj} \prod_{i=1}^{s-1} v_i^{d'_i}$ for $j \ne k(s)$, where $d'_i = (count(i, s, k(s)) - 1)Q$.

Consider a sufficiently small neighborhood of $\{v_i = 0\}$ using Theorem 2.

Set $f_{ik(s)} = b_{ik(s)} \prod_{k(i_1)=k(s),1\le i_1\le s-1}(b^Q_{ik(s)} - b^Q_{i_1 k(s)})$ for $i \ge s$, $\tilde{b}'_{ij} = b''_{ij} - b''_{sj}\frac{f_{ik(s)}}{f_{sk(s)}}$ and $\tilde{b}_{ij} = \tilde{b}'_{ij} \prod_{i_1=1}^{s} v_{i_1} = b_{ij} - b_{sj}\frac{f_{ik(s)}}{f_{sk(s)}}$ for $i \ge s+1, j \ne k(s)$.

We then have:

$$\begin{pmatrix} b^{\ell_1}_{s1} \cdots b^{\ell_N}_{sN} \\ \vdots \\ b^{\ell_1}_{H1} \cdots b^{\ell_N}_{HN} \end{pmatrix}$$

$$= \begin{pmatrix} \prod_{j=1}^{N} b^{\ell_j}_{sj} \\ b^{\ell_{k(s)}}_{s+1,k(s)} \prod_{j=1,j\ne k(s)}^{N} (\tilde{b}_{s+1,j} + b_{sj}\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\ell_j} \\ \vdots \\ b^{\ell_{k(s)}}_{H,k(s)} \prod_{j=1}^{N} (\tilde{b}_{Hj} + b_{sj}\frac{f_{Hk(s)}}{f_{sk(s)}})^{\ell_j} \end{pmatrix}$$

which is an element of the vector ideal:

$$\left\langle \begin{pmatrix} \prod_{j=1}^{N} b_{sj}^{\ell_j} \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{s+1,k(s)}^{\ell_{k(s)}} (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \\ \vdots \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{Hk(s)}^{\ell_{k(s)}} (\frac{f_{Hk(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \end{pmatrix} \right\rangle$$

$$+ \sum_{\substack{\ell'_{k(s)}=\ell_{k(s)},\\ 0\leq \ell'_j\leq \ell_j, \exists \ell'_j\neq \ell_j}} \left\langle \begin{pmatrix} 0 \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell'_j} b_{s+1,k(s)}^{\ell_{k(s)}} \prod_{j=1,j\neq k(s)}^{N} \tilde{b}_{s+1,j}^{\ell_j-\ell'_j} (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j} \\ \vdots \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell'_j} b_{Hk(s)}^{\ell_{k(s)}} \prod_{j=1,j\neq k(s)}^{N} \tilde{b}_{Hj}^{\ell_j-\ell'_j} (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j} \end{pmatrix} \right\rangle$$

Furthermore, we have: $\begin{pmatrix} \prod_{j=1}^{N} b_{sj}^{\ell_j} \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{s+1,k(s)}^{\ell_{k(s)}} (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \\ \vdots \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{Hk(s)}^{\ell_{k(s)}} (\frac{f_{Hk(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \end{pmatrix}$ is an element of:

$$\left\langle \frac{\prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}}{f_{sk(s)}^{\sum_{j=1,j\neq k(s)}^{N}\ell_j}} \begin{pmatrix} b_{s,k(s)}^{nQ+1} \prod_{k(i_1)=k(s),1\leq i_1\leq s-1} (b_{sk(s)}^{Q} - b_{i_1k(s)}^{Q}) \\ b_{s+1,k(s)}^{nQ+1} \prod_{k(i_1)=k(s),1\leq i_1\leq s-1} (b_{s+1k(s)}^{Q} - b_{i_1k(s)}^{Q}) \\ \vdots \\ b_{Hk(s)}^{nQ+1} \prod_{k(i_1)=k(s),1\leq i_1\leq s-1} (b_{Hk(s)}^{Q} - b_{i_1k(s)}^{Q}) \end{pmatrix} \ : \ n\geq 0 \right\rangle$$

Since $b''_{sj} = b'''_{sj} \prod_{i=1}^{s-1} v_i^{d'_i}$ for $j \neq k(s)$, where $d'_i = (count(i,s,k(s))-1)Q$, we have $b_{sj} = b'''_{sj} \prod_{i=1}^{s-1} v_i^{d'_i+1}$ and $\frac{\prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}}{f_{sk(s)}^{\sum_{j=1,j\neq k(s)}^{N}\ell_j}}$ is finite. That is, we have:

$$\begin{pmatrix} \prod_{j=1}^{N} b_{sj}^{\ell_j} \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{s+1,k(s)}^{\ell_{k(s)}} (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \\ \vdots \\ \prod_{j=1,j\neq k(s)}^{N} b_{sj}^{\ell_j}\, b_{Hk(s)}^{\ell_{k(s)}} (\frac{f_{Hk(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell_j} \end{pmatrix} \in \left\langle \begin{pmatrix} b_{s,k(s)}^{nQ} f_{sk(s)} \\ b_{s+1,k(s)}^{nQ} f_{s+1,k(s)} \\ \vdots \\ b_{Hk(s)}^{nQ} f_{Hk(s)} \end{pmatrix} \ : \ n\geq 0 \right\rangle$$

If we assume that for α:

$$\mathcal{J}_\alpha = \left\langle \begin{pmatrix} b_{sk(s)}^{nQ} f_{sk(s)} \\ \vdots \\ b_{Hk(s)}^{nQ} f_{Hk(s)} \end{pmatrix}, n\geq 0 \right\rangle + \left\langle \begin{pmatrix} b_{s1}^{\ell_1} \cdots b_{sN}^{\ell_N} \\ \vdots \\ b_{H1}^{\ell_1} \cdots b_{HN}^{\ell_N} \end{pmatrix} : \sum_{j=1}^{N} \ell_j = nQ+1, n\geq 1, \sum_{j=1,j\neq k(s)}^{N} \ell_j \leq \alpha \right\rangle$$

$$= \left\langle \begin{pmatrix} b_{sk(s)}^{nQ} f_{sk(s)} \\ \vdots \\ b_{Hk(s)}^{nQ} f_{Hk(s)} \end{pmatrix}, n\geq 0 \right\rangle + \left\langle \begin{pmatrix} 0 \\ b_{s+1,k(s)} \prod_{j=1,j\neq k(s)}^{N} \tilde{b}_{s+1,j}^{\ell_j} \\ \vdots \\ b_{Hk(s)} \prod_{j=1,j\neq k(s)}^{N} \tilde{b}_{Hj}^{\ell_j} \end{pmatrix} : \sum_{j=1}^{N} \ell_j = nQ+1, n\geq 1, \sum_{j=1,j\neq k(s)}^{N} \ell_j \leq \alpha \right\rangle$$

we have for $\sum_{j=1,j\neq k(s)}^{N}\ell_j=\alpha+1$:

$$\left\langle\begin{pmatrix} b_{s1}^{\ell_1}\cdots b_{sN}^{\ell_N}\\ \vdots\\ b_{H1}^{\ell_1}\cdots b_{HN}^{\ell_N}\end{pmatrix}\right\rangle+\mathcal{J}_\alpha=\left\langle\begin{pmatrix} 0\\ b_{s+1,k(s)}\prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{s+1,j}^{\ell_j}\\ \vdots\\ b_{Hk(s)}\prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{Hj}^{\ell_j}\end{pmatrix}\right\rangle+\mathcal{J}_\alpha$$

since for $\exists\ell'_j\neq 0$, we have:

$$\begin{pmatrix} 0\\ \prod_{j=1,j\neq k(s)}^{N}b_{sj}^{\ell'_j}b_{s+1,k(s)}^{\ell_{k(s)}}\ \prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{s+1,j}^{\ell_j-\ell'_j}\ (\frac{f_{s+1,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j}\\ \vdots\\ \prod_{j=1,j\neq k(s)}^{N}b_{sj}^{\ell'_j}b_{Hk(s)}^{\ell_{k(s)}}\ \prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{Hj}^{\ell_j-\ell'_j}\ (\frac{f_{H,k(s)}}{f_{sk(s)}})^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j}\end{pmatrix}$$

$$=\frac{\prod_{j=1,j\neq k(s)}^{N}b_{sj}^{\ell'_j}}{f_{sk(s)}^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j}}\begin{pmatrix} 0\\ b_{s+1,k(s)}^{\ell_{k(s)}}\ \prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{s+1,j}^{\ell_j-\ell'_j}\ f_{s+1,k(s)}^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j}\\ \vdots\\ b_{Hk(s)}^{\ell_{k(s)}}\ \prod_{j=1,j\neq k(s)}^{N}\tilde{b}_{Hj}^{\ell_j-\ell'_j}\ f_{H,k(s)}^{\sum_{j=1,j\neq k(s)}^{N}\ell'_j}\end{pmatrix}\in\mathcal{J}_\alpha$$

Therefore, by setting:

$$a'_{i_0s}b_{sk(s)}\prod_{k(i)=k(s),1\leq i\leq s-1}(b_{s,k(s)}^{Q}-b_{i,k(s)}^{Q})=\sum_{i_1=s}^{H}a_{i_0i_1}b_{i_1k(s)}\prod_{k(i)=k(s),1\leq i\leq s-1}(b_{i_1k(s)}^{Q}-b_{i,k(s)}^{Q})$$

for $1\leq i_0\leq M$, and by setting $b_{ij}=\tilde{b}_{ij}$ again, we have:

$$\mathcal{J}=\left\langle a_{i_0i_1}b_{i_1k(i_1)}\prod_{k(i)=k(i_1),1\leq i\leq i_1-1}(b_{i_1,k(i_1)}^{Q}-b_{i,k(i_1)}^{Q})\ :\ 1\leq i_0\leq M,1\leq i_1\leq s-1\right\rangle$$

$$+\left\langle a'_{i_0s}b_{sk(s)}\prod_{k(i)=k(s),1\leq i\leq s-1}(b_{s,k(s)}^{Q}-b_{i,k(s)}^{Q})\ :\ 1\leq i_0\leq M\right\rangle$$

$$+\left\langle\begin{pmatrix} a_{i_0s+1}&\cdots&a_{i_0H}\end{pmatrix}\begin{pmatrix} b_{s+1j}\prod_{k(i)=j,1\leq i\leq s}(b_{s+1j}^{Q}-b_{ij}^{Q})\\ \vdots\\ b_{Hj}\prod_{k(i)=j,1\leq i\leq s}(b_{Hj}^{Q}-b_{ij}^{Q})\end{pmatrix}\ :\ j=1,\cdots,N\right\rangle$$

$$+\left\langle\begin{pmatrix} a_{i_0s+1}&\cdots&a_{i_0H}\end{pmatrix}\begin{pmatrix} b_{s+1,1}^{\ell_1}\cdots b_{s+1,N}^{\ell_N}\\ \vdots\\ b_{H1}^{\ell_1}\cdots b_{HN}^{\ell_N}\end{pmatrix}:1\leq i_0\leq M,\sum_{i=1}^{N}\ell_i=nQ+1,n\geq 1,\forall\ell_j<Qn+1\right\rangle$$

with (i)$\sim$(iv).

Step 2

By Step 1, we need to consider the ideal:

$$\left\langle a_{i_0,i_1}\prod_{i=1}^{i_1}v_i^{(count(i,i_1,k(i_1))-1)Q+1}\ :\ 1\le i_0\le M, i_1\le s\right\rangle$$

$$+\left\langle \begin{pmatrix} a_{i_0s+1} & \cdots & a_{i_0H}\end{pmatrix}\prod_{i=1}^{s}v_i^{count(i,s,j)Q+1}\begin{pmatrix} b'_{s+1j}\\ \vdots\\ b'_{Hj}\end{pmatrix}\ :\ j=1,\cdots,N\right\rangle$$

$$+\left\langle \begin{pmatrix} a_{i_0s+1} & \cdots & a_{i_0H}\end{pmatrix}\prod_{i=1}^{s}v_i^{Q+1}\begin{pmatrix} b'^{\ell_1}_{s+1,1}\cdots b'^{\ell_N}_{s+1,N}\\ \vdots\\ b'^{\ell_1}_{H1}\cdots b'^{\ell_N}_{HN}\end{pmatrix}:1\le i_0\le M,\sum_{i=1}^{N}\ell_i=Q+1,\forall\ell_j<Q+1\right\rangle$$

with Jacobian:

$$\prod_{i=1}^{H}v_i^{(H-i+1)N-1+d_i(s)}dv_i\prod_{k=1}^{M}\prod_{i=1}^{H}da_{ki}\prod_{i=1}^{H}\prod_{j=1}^{N}db'_{ij}$$

where:

$$d_i(s)=(N-1)Q\sum_{s_1=i}^{s}(count(i,s_1,k(s_1))-1)$$

We have: $\lambda_0(||A_{M,H}B^{(Q)}_{H,N}||^2)\le$

$$\min\{\frac{(H-i+1)N+d_i(s)+d'_i(s)+d''_{i,i_{s+1},\cdots,i_H}(s)}{2(count(i,s,k(s))-1)Q+2}\ :\ 1\le i\le s, i_\alpha\ge 0, 1\le s\le H\}$$

where:

$$\begin{aligned}
d'_i(s) &= M(i-1)\{(count(i,s,k(s))-1)Q+1\}\\
&\quad +QM\sum_{\substack{s_1=i,\\ count(i,s,k(s))>count(i,s_1,k(s_1))}}^{s-1}(count(i,s,k(s))-count(i,s_1,k(s_1)))\\
d''_{i,i_{s+1},\cdots,i_H}(s) &= \sum_{\alpha=s+1}^{H}\{Mi_\alpha\\
&\quad +\sum_{\substack{j=1,j\neq k(s)\\ count(i,s,j)=0}}\max\{Q(count(i,s,k(s))-1)-i_\alpha,0\}\\
&\quad +\sum_{\substack{j=1,j\neq k(s)\\ count(i,s,j)\ge 1, count(i,s,k(s))\ge 2}}^{N}\max\{Q(count(i,s,k(s))-2)-i_\alpha,0\}\}
\end{aligned}$$

Set $C(i,s) = \#\{count(i,s,j) = 0, 1 \leq j \leq N\}$. Then, we have:

$$
\begin{aligned}
d_i''(s) &= \min\{d''_{i,i_{s+1},\cdots,i_H}, i_\alpha \geq 0\} \\
&= \begin{cases}
0, \qquad \text{if } count(i,s,k(s)) = 1 \\
(H-s)\{C(i,s)Q + (N-1)Q(count(i,s,k(s))-2)\} \\
\qquad \text{if } count(i,s,k(s)) \geq 2, N-1 \leq M \\
(H-s)\{C(i,s)Q + MQ(count(i,s,k(s))-2)\} \\
\qquad \text{if } count(i,s,k(s)) \geq 2, C(i,s) \leq M < N-1 \\
(H-s)\{MQ(count(i,s,k(s))-1)\} \\
\qquad \text{if } count(i,s,k(s)) \geq 2, M \leq C(i,s)
\end{cases}
\end{aligned}
$$

By the above equation, we have bound_1. By [19], we have bound_2 and bound_3, thus completing the proof.

References

1. Hartigan, J.A. A Failure of Likelihood Ratio Asymptotics for Normal Mixtures. In Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer, California, CA, USA, 1985; Volume 2, pp. 807–810.
2. Sussmann, H.J. Uniqueness of the weights for minimal feed-forward nets with a given input-output map. *Neural Netw.* **1992**, *5*, 589–593.
3. Hagiwara, K.; Toda, N.; Usui, S. On the problem of applying AIC to determine the structure of a layered feed-forward neural network. In Proceedings of the IJCNN Nagoya Japan, Nagoya Congress Center, Japan, 25–29 October 1993; Volume 3, pp. 2263–2266.
4. Fukumizu, K. A regularity condition of the information matrix of a multilayer perceptron network. *Neural Netw.* **1996**, *9*, 871–879.
5. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
6. Watanabe, S. Algebraic analysis for nonidentifiable learning machines. *Neural Comput.* **2001**, *13*, 899–933.
7. Watanabe, S. Algebraic geometrical methods for hierarchical learning machines. *Neural Netw.* **2001**, *14*, 1049–1060.
8. Watanabe, S. Algebraic geometry of learning machines with singularities and their prior distributions. *J. Jpn. Soc. Artif. Intell.* **2001**, *16*, 308–315.
9. Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: New York, NY, USA, 2009; Volume 25.
10. Fulton, W. *Introduction to Toric Varieties, Annals of Mathematics Studies*; Princeton University Press: Princeton, NJ, USA, 1993.
11. Kollár, J. Singularities of Pairs. In *Algebraic Geometry-Santa Cruz 1995, Series Proceedings of Symposia in Pure Mathematics*, 9–29 July 1995; American Mathematical Society: Providence, RI, USA, 1997,; Volume 62, pp. 221–287.
12. Mustata, M. Singularities of pairs via jet schemes. *J. Am. Math. Soc.* **2002**, *15*, 599–615.

13. Yamazaki, K.; Aoyagi, M.; Watanabe, S. Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Netw.* **2010**, *23*, 35–43.
14. Aoyagi, M.; Nagata, K. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity. *Neural Comput.* **2012**, *24*, 1569–1610.
15. Aoyagi, M.; Watanabe, S. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Netw.* **2005**, *18*, 924–933.
16. Aoyagi, M.; Watanabe, S. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II* **2005**, *10*, 2112–2124.
17. Aoyagi, M. The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku Recent Top. Real Complex Singul.* **2006**, *1501*, 153–167.
18. Aoyagi, M. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Commun. Stat. Theory Methods* **2010**, *39*, 2667–2687.
19. Aoyagi, M. Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *J. Algebr. Stat.* **2013**, in press.
20. Rusakov, D.; Geiger, D. Asymptotic Model Selection for Naive Bayesian Networks. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, Alberta, Canada, 1–4 August 2002; pp. 438–445.
21. Rusakov, D.; Geiger, D. Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.* **2005**, *6*, 1–35.
22. Zwiernik, P. An asymptotic behavior of the marginal likelihood for general Markov models. *J. Mach. Learn. Res.* **2011**, *12*, 3283–3310.
23. Watanabe, S. Equations of states in singular statistical estimation. *Neural Netw.* **2010**, *23*, 20–34.
24. Aoyagi, M. Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network. *Int. J. Pure Appl. Math.* **2009**, *52*, 177–204.
25. Drton, M. Conference Lecture: Reduced Rank Regression. Workshop on Singular Learning Theory, AIM 2011. Available online: http://math.berkeley.edu/~critch/slt2011/ (accessed on 16 December 2011)
26. Drton, M. Conference Lecture: Bayesian Information Criterion for Singular Models. Algebraic Statistics 2012 in the Alleghenies at The Pennsylvania State University. Available online: http://jasonmorton.com/aspsu2012/ (accessed on 15 June 2012).
27. Nagata, K.; Watanabe, S. Exchange Monte Carlo Sampling from Bayesian posterior for singular learning machines. *IEEE Trans. Neural Netw.* **2008**, *19*, 1253–1266.
28. Nagata, K.; Watanabe, S. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Int. J. Neural Netw.* **2008**, *21*, 980–988.
29. Bernstein, I.N. The analytic continuation of generalized functions with respect to a parameter. *Funct. Anal. Appl.* **1972**, *6*, 26–40.
30. Bjőrk, J.E. *Rings of Differential Operators*; North-Holland: Amsterdam, The Netherlands, 1979.
31. Kashiwara, M. B-functions and holonomic systems. *Invent. Math.* **1976**, *38*, 33–53.

32. Lin, S. Asymptotic approximation of marginal likelihood integrals. **2010**, arXiv:1003.5338v2.