International Journal of Pure and Applied Mathematics
Volume 52 No. 2 2009, 177-204

LOG CANONICAL THRESHOLD OF VANDERMONDE MATRIX TYPE SINGULARITIES AND GENERALIZATION ERROR OF A THREE-LAYERED NEURAL NETWORK IN BAYESIAN ESTIMATION

Miki Aoyagi
Advanced Research Institute for the Sciences and Humanities
Nihon University
Nihon University Kaikan Daini Bekkan
12-5, Goban-cho, Chiyoda-ku, Tokyo 102-8251, JAPAN
e-mail: aoyagi.miki@nihon-u.ac.jp

Abstract: The log canonical threshold of Vandermonde matrix type singularities over the real number field serves to measure the learning efficiencies in hierarchical learning models. Imposing certain orthogonality conditions for such singularities, explicit computational results for the log canonical thresholds are given. In applying such results to a three-layered neural network, we are able to clarify its generalization error and its stochastic complexity found useful in learning theory.

AMS Subject Classification: 32S10, 14Q15, 62D05, 62M20, 62M45
Key Words: log canonical threshold, zeta function, resolution of singularities, generalization error, layered neural networks

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 [24]). Our study is to consider the generalization error and the stochastic complexity in learning theory by using the log canonical threshold in algebraic geometry.

Received: March 2, 2009

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

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

over $\mathbb{C}$ and

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

over $\mathbb{R}$ for a nonzero regular function f on a smooth variety Y, where $Z \subset Y$ is a closed subscheme (Kollár [17], Mustata [20]). It has been established 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 [9], Björk [10], Kashiwara [16]). The log canonical threshold $c_Z(Y,f)$ 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 and $\psi(w) \neq 0$ on Z.

In this paper, we consider the log canonical threshold of Vandermonde matrix type singularities over the real number field (Definition 3). We have recently proved that such thresholds serve to measure the learning efficiencies in hierarchical learning models, i.e., they correspond to the main term of the generalization error for hierarchical learning models (Aoyagi [4], Aoyagi et al [6], Watanabe [29], Yamazaki et al [31]).

Hierarchical learning models such as layered neural networks, reduced rank regressions, normal mixture models and Boltzmann machines are known as effective learning models to analyze complicated data influenced by many factors. The theoretical study of hierarchical learning models has seen a rapid development in recent years, after these models were recognized as not having been analyzed using the classic theories of regular statistical models due to the fact that they have singular Fisher metrics (see Hartigan [14], Sussmann [25], Hagiwara et al [13], Fukumizu [11]). These models are called non-regular statistical models.

Watanabe proved that the largest pole of a zeta function for the hierarchical learning model gives asymptotically the main term of the generalization error of the model (Watanabe [26], [27]). Clarifying these generalization errors is one of the important topics in learning theory. We have shown that the log canonical threshold of Vandermonde matrix type singularities gives the main terms of the generalization error for three-layered neural networks, normal mixture models and mixtures of binomial distributions (Aoyagi [4], Aoyagi et al [6], Watanabe [29], Yamazaki et al [31]). The Vandermonde matrix type singularities are degenerate with respect to their Newton polyhedrons (Fulton [12]), and are not isolated.

In general, singularities in learning theory have similar properties, and therefore, obtaining the largest pole of the zeta functions in their application to learning theory is still a difficult problem. Moreover, our study is over the real number field and not the complex number field. In algebraic geometry and algebraic analysis, such studies are usually done over an algebraically closed field (Kollár [17], Mustata [20]). There are many differences between the real number field and the complex number field; for example, log canonical thresholds over the complex number field are less than 1, while those over the real number field are not necessarily so. Therefore, we cannot directly apply results established over an algebraically closed field to the present application.

In this paper, we begin by developing certain orthogonality conditions for Vandermonde matrix type singularities (Theorem 1). These conditions mean that the learning model "learns" a true distribution independently of each element (Section 3). Theorem 2 gives explicit computational results for the log canonical thresholds under certain conditions. In applying such results, we consider the generalization error and the stochastic complexity of three-layered neural networks (Theorem 4).

In [7], we obtained learning efficiencies for the reduced rank regression which is a three-layered neural network with linear hidden units. Rusakov and Geiger [22] obtained these efficiencies for Naive Bayesian networks. In a recent paper [8], we have also obtained these for one-dimensional normal mixture models.

The rest of the paper consists of two sections and Appendices. In Section 2, we present our main results concerning the Vandermonde matrix type singularities. In Section 3, we summarize the framework of Bayesian learning models and our result for three-layered neural networks, while the Appendices detail supplementary material of theorems left out for the sake of brevity and continuity in the main text.

2. Vandermonde Matrix Type Singularities

In this paper, we denote constants by a^*, b^*, etc.

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 $\mathbb{N}_{+0} = \mathbb{N} \cup \{0\}$, where $\mathbb{N}$ is the set of all natural numbers.

Definition 1. Denote

$$c_Z(f) = c_Z(\mathbb{R}^d, f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ near } Z\} \text{ over } \mathbb{R}$$

and by $\theta_Z(f)$ its order, i.e., the order of the largest pole of $\int_{\text{near } Z} |f|^z \mathrm{d}x$, for a nonzero regular function f on $\mathbb{R}^d$, where $Z \subset \mathbb{R}^d$ is a closed subscheme.

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 $Q \in \mathbb{N}$ and $m \in \mathbb{N}_{+0}$. Let

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

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

and $B = (B_I)_{\ell_1+\cdots+\ell_N = Qn+m, 0 \le n \le H+r-1}$ (t denotes the transpose).

We call singularities of $||AB||^2 = 0$ Vandermonde matrix type singularities.

To simplify matters, 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 $j \neq j'$.

From here on, we set A and B as in Definition 3.

Remark 1. By the ascending chain condition, we have $\langle AB \rangle = \langle AB' \rangle$ where $B' = (B_I)_{\ell_1+\cdots+\ell_N = Qn+m, 0 \le n \le H'}$ and $H' \ge H + r - 1$.

Example 1. If $N = 1$, $m = 0$, and $r = 0$, we have $A = \begin{pmatrix} a_{11} & \cdots & a_{1H} \\ a_{21} & \cdots & a_{2H} \\ & \vdots & \\ a_{M1} & \cdots & a_{MH} \end{pmatrix}$

and $B = \begin{pmatrix} 1 & b_{11}^Q & b_{11}^{2Q} & \cdots & b_{11}^{Q(H-1)} \\ 1 & b_{21}^Q & b_{21}^{2Q} & \cdots & b_{21}^{Q(H-1)} \\ & & \vdots & & \\ 1 & b_{H1}^Q & b_{H1}^{2Q} & \cdots & b_{H1}^{Q(H-1)} \end{pmatrix}$.

(The matrix B with $Q = 1$ as above is usually called a Vandermonde matrix.)

Example 2. If $N = 3$, $m = Q = 1$ and $r = H = 1$, we have $A = \begin{pmatrix} a_{11} & a^*_{12} \\ a_{21} & a^*_{22} \\ \vdots & \vdots \\ a_{M1} & a^*_{M,2} \end{pmatrix}$ and

$$B = \begin{pmatrix} b_{11} & b_{11}^2 & b_{12} & b_{12}^2 & b_{13} & b_{13}^2 & b_{11}b_{12} & b_{11}b_{13} & b_{12}b_{13} \\ b^*_{21} & {b^*_{21}}^2 & b_{22} & {b^*_{22}}^2 & b_{23} & {b^*_{23}}^2 & b^*_{21}b^*_{22} & b^*_{21}b^*_{23} & b^*_{22}b^*_{23} \end{pmatrix}.$$

Theorem 1. *Consider a sufficiently small neighborhood U_{w^*} of*

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

and $w = \{a_{ki}, b_{ij}\}_{1\le k\le M, 1\le i\le H, 1\le j\le N} \in U_{w^*}$.

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

Let each $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ be a different real vector in $[b^*_{i1}, b^*_{i2}, \cdots, b^*_{iN}]_Q \neq 0$, 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 $r' \ge r$ and set $(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^*_{H+i,1}, \cdots, b^*_{H+i,N}]_Q$, for $1 \le i \le r$.

Assume that

$$\left.\begin{array}{c} [b^*_{11}, \cdots, b^*_{1N}]_Q \\ \vdots \\ [b^*_{H_01}, \cdots, b^*_{H_0N}]_Q \end{array}\right\} = 0,$$

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

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

$$\vdots$$

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

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

Then we have

$$c_{w^*}(||AB||^2) = \sum_{\alpha=0}^{r'} c_{w^{(\alpha)*}}(||A^{(\alpha)}B^{(\alpha)}||^2),$$
$$\theta_{w^*}(||AB||^2) = (\sum_{\alpha=0}^{r'} \theta_{w^{(\alpha)*}}(||A^{(\alpha)}B^{(\alpha)}||^2) - 1) + 1,$$

where

$$w^{(\alpha)^*} = \{a_{ki}^{(\alpha)^*}, b_{ij}^{(\alpha)^*}\} = \{a^*_{k,H_0+\cdots+H_{\alpha-1}+i}, b^{**}_{\alpha j}\}_{1\le k\le M, 1\le i\le H_\alpha, 1\le j\le N},$$

$$I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N,$$

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

for $\alpha = 0, r+1 \le \alpha \le r'$,

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

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

Proof. Set

$$\begin{cases} (a_{i1}^{(0)}, \ldots, a_{iH_0}^{(0)}) = (a_{i1}, \ldots, a_{iH_0}), \\ (a_{i1}^{(1)}, \ldots, a_{iH_1}^{(1)}) = (a_{i,H_0+1}, \ldots, a_{i,H_0+H_1}), \\ \quad\vdots \\ (a_{i1}^{(r')}, \ldots, a_{iH_{r'}}^{(r')}) = (a_{i,H_0+\cdots+H_{r'-1}+1}, \ldots, a_{i,H_0+\cdots+H_{r'}}), \end{cases}$$

for $1 \le i \le M$, and

$$\begin{cases} (b_{1j}^{(0)}, \ldots, b_{H_0 j}^{(0)}) = (b_{1j}, \ldots, b_{H_0 j}), \\ (b_{1j}^{(1)}, \ldots, b_{H_1 j}^{(1)}) = (b_{H_0+1,j}, \ldots, b_{H_0+H_1,j}), \\ \quad \vdots \\ (b_{1j}^{(r')}, \ldots, b_{H_{r'} j}^{(r')}) = (b_{H_0+\cdots+H_{r'-1}+1,j}, \ldots, b_{H_0+\cdots+H_{r'},j}), \end{cases}$$

for $1 \le j \le N$.

For $\gamma_i(b_{i1}^{(\alpha)}, \cdots, b_{iN}^{(\alpha)}) = [b_{i1}^{(\alpha)}, \cdots, b_{iN}^{(\alpha)}]_Q$, we again set $a_{ki}^{(\alpha)}$ by $a_{ki}^{(\alpha)}/(\gamma_i)^m$ and $b_{ij}^{(\alpha)}$ by $b_{ij}^{(\alpha)}\gamma_i$, $1 \le j \le N$ and $1 \le k \le M$.

The main arguments of the proof appear in the appendices. By applying Lemma 5 in Appendix A and Corollary 1 in Appendix B, we have then established the validity of this theorem. □

In learning theory, r corresponds to the number of elements of a true distribution.

This theorem shows that the Bayesian learning coefficient associated with such singularities is the sum of each such coefficient for the small model with respect to each element of a true distribution (cf. Section 3).

Theorem 2. *We use the same notations as in Theorem 1. If $N = 1$, we have*

$$c_{w^*}(||AB||^2) = \frac{MQk_0(k_0+1) + 2H_0}{4(m+k_0Q)} + \frac{Mr'}{2}$$
$$+ \sum_{\alpha=1}^{r} \frac{Mk_\alpha(k_\alpha+1) + 2H_\alpha}{4(1+k_\alpha)} + \sum_{\alpha=r+1}^{r'} \frac{Mk'_\alpha(k'_\alpha+1) + 2(H_\alpha - 1)}{4(1+k'_\alpha)},$$
$$\theta_{w^*}(||AB||^2) = 1 + \#\Theta,$$

where

$$k_0 = \max\{i \in \mathbb{Z}; 2H_0 \ge M(i(i-1)Q + 2mi)\},$$
$$k_\alpha = \max\{i \in \mathbb{Z}; 2H_\alpha \ge M(i^2 + i)\}, \text{ for } 1 \le \alpha \le r,$$
$$k'_\alpha = \max\{i \in \mathbb{Z}; 2(H_\alpha - 1) \ge M(i^2 + i)\}, \text{ for } r+1 \le \alpha \le r',$$

and

$$\Theta = \{k_0, k_\alpha, k'_\alpha \;; \quad \begin{aligned} &2H_0 = M(k_0(k_0-1)Q + 2mk_0), \\ &2H_\alpha = M(k_\alpha^2 + k_\alpha), \text{ for } 1 \le \alpha \le r, \\ &2(H_\alpha - 1) = M(k'^2_\alpha + k'_\alpha), \text{ for } r+1 \le \alpha \le r'\}. \end{aligned}$$

For the proof of Theorem 2, we use Theorem 1 and a similar method to be found in [6], [4], where we used recursive blow-ups and toric resolution. The proof is very complicated because we see all branches of the recursive blow-ups at every non-isolated singularity.

Recently, we have also obtained explicit values of $c_{w^*}(||AB||^2)$ for general natural numbers N and M but for $H \leq 2$, see [5].

Our long term intention is to obtain the log canonical thresholds of Vandermonde matrix type singularities in general.

3. Learning Theorem

In this section, we present an overview of learning theory focusing especially on stochastic complexity and generalization error in Bayesian estimation.

A learning system consists of data, a learning model and a learning algorithm. The purpose of such a system is to estimate an unknown true density function from data distributed by the true density function. The data in learning theory are usually very complicated and not generated by a simple normal distribution. For example, such data are associated with image or speech recognition, artificial intelligence, the control of a robot, genetic analysis, data mining, time series prediction. Learning models to analyze such data should likewise have complicated structures. Hierarchical learning models such as the layered neural network model, the Boltzmann machine, the reduced rank regression model and the normal mixture model are known as effective learning models. These models are called non-regular statistical models and cannot be analyzed using the classic theories of regular statistical models (Hartigan [14], Sussmann [25], Hagiwara et al [13], Fukumizu [11]). The theoretical study has therefore been started to construct a mathematical foundation for non-regular statistical models.

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 topics in learning theory. The largest pole of a zeta function for a learning model, which is called a learning coefficient, gives the main term of the generalization error.

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 is always 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

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

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 et al [18], Amari et al [2], [3]). $F(n)$ is known as the Bayesian criterion in Bayesian model selection (Schwarz [23]), stochastic complexity in universal coding (see Rissanen [21], see Yamanishi [30]), Akaike's Bayesian criterion in optimization of hyperparameters (Akaike [1]) and evidence in neural network learning (Mackay [19]). Therefore, $F(n)$ 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 [26], [27]). 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.

Theorem 3. (see Watanabe [26], [27]) *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

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

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

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

To prove the above theorem, Watanabe used the function

$$v(t) = \int \delta(t - K(w)) \varphi(w) \mathrm{d}w = \frac{\partial}{\partial t} \int_{K(w)<t} \varphi(w) \mathrm{d}w,$$

which satisfies $\int v(t) f(t) dt = \int f(K(w)) \psi(w) dw$ for any analytic function $f(t)$. The Laplace transform of $v(t)$ is

$$Z(n) = \int \exp(-nK(w)) \varphi(w) dw,$$

and the Mellin transform of $v(t)$ is

$$\zeta(z) = \int K(w)^z \varphi(w) dw = \int t^z v(t) dt.$$

The key point of the proof is that by using poles of $\zeta(z)$ and the inverse Mellin transform of $\zeta(z)$, he obtained the asymptotic expansion of $v(t)$, and then the asymptotic expansion of $Z(n)$. The analysis of the difference between $-\log Z(n)$ and $F(n)$ completes the proof.

In learning theory, λ is, therefore, an essential value, which corresponds to the log canonical threshold of $K(w)$.

We here show the following two hierarchical learning models such that the log canonical thresholds of Vandermonde matrix type singularities are equal to their λ.

(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 $x^{(1)} = (x_j^{(1)}) \in \mathbb{R}^N$ with a probability density function $q(x)$ which has a compact support $\tilde{W}$. Then an output value $x^{(2)} = (x_k^{(2)}) \in \mathbb{R}^M$ of the three-layered neural network is given by $x_k^{(2)} = f_k(x^{(1)}, w) +$ (noise), where $w = \{a_{ki}, b_{ij}; 1 \le k \le M, 1 \le i \le H, 1 \le j \le N\}$ and

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

Consider a statistical model

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

Assume that the true distribution

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

is included in the learning model, where $w_t^* = \{a_{ki}^*, b_{ij}^*; 1 \le k \le M, H+1 \le i \le H+r, 1 \le j \le N\}$ and

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

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 the model has the zeta function $\int_W ||AB||^{2z} dw$ with $Q = 2$ and $m = 1$, where A and B are defined in Definition 3.

The Taylor expansion $\tanh x = \sum_{i=1}^{\infty} \alpha_i x^{2(i-1)+1}$, with $\alpha_i \neq 0$ at 0 together with Lemma 5 in Watanabe [26] proves this fact.

Remark 2. Let $\sigma(x) = \sum_{i=1}^{\infty} \alpha_i x^{Q(i-1)+1}$ and $\alpha_i \neq 0$. The maximum pole of

$$\int_W (\int_{\tilde{W}} (\sum_{m=1}^{p} a_m^{(w)} \sigma(b_m^{(w)} x) - \sum_{m=1}^{p} a_m^* \sigma(b_m^* x))^2 q(x) \mathrm{d}x)^z \psi(w) \mathrm{d}w,$$

and its order are the same as in Main Theorem 1.

(b) The normal mixture model with H peaks which is trained for estimating the true distribution with r peaks (Watanabe et al [29]):

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 $w = \{a_{1i}, b_{ij}; 1 \le i \le H, 1 \le j \le N\}$ 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 $w_t^* = \{a_{1i}^*, b_{ij}^*; H+1 \le i \le H+r, 1 \le j \le N\}$ 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 the model has the zeta function $\int_W ||AB||^{2z} dw$ with $Q = 1$, $M = 1$ and $m = 1$, where A and B are defined in Definition 3.

Statements (a) and (b) as given above show that for three-layered neural networks and for normal mixture models λ in Theorem 3 are obtained by the same type of singularities, i.e., Vandermonde matrix type singularities. Moreover, the paper of Yamazaki et al [31] shows that for mixtures of binomial distributions λ is also obtained by Vandermonde matrix type singularities. These results seem to imply that Vandermonde matrix type singularities are essential in resolving current analytical difficulties arising in learning theory.

Theorem 4. *We use the same notations in statement (a).*

For the three-layered neural network with one input unit, the maximum pole $-\lambda$ and its order θ in (1) and (2) are obtained by

$$\lambda = \min_{\tilde{w} \in W^*} c_{\tilde{w}}(||AB||^2)$$

with its order θ, where $Q = 2$, $m = 1$ and $W^* = \{\tilde{w} \in \mathbb{R}^d \mid f(x^{(1)}, \tilde{w}) = f(x^{(1)}, w_t^*)$ for any $x^{(1)}\}$.

More precisely,

— For $H - r + 1 \le \begin{cases} 10, & M = 1, \\ 5, & M = 2, \\ 4 + M, & M \ge 3, \end{cases}$

we have $\lambda = (r-1)\frac{M+1}{2} + \frac{M}{2} + \frac{M(k_1^2 + k_1) + 2(H - r + 1)}{4(k_1 + 1)}$,

$$\theta = \begin{cases} 1, & \text{if } M(k_1^2 + k_1) < 2(H - r + 1), \\ 2, & \text{if } M(k_1^2 + k_1) = 2(H - r + 1), \end{cases}$$

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

— For $H - r + 1 > \begin{cases} 10, & M = 1, \\ 5, & M = 2, \\ 4 + M, & M \ge 3, \end{cases}$

we have $\lambda = r\frac{M+1}{2} + \frac{M(k_0^2 + k_0) + H - r}{4k_0 + 2}$,

$$\theta = \begin{cases} 1, & \text{if } Mk_0^2 < H - r, \\ 2, & \text{if } Mk_0^2 = H - r, \end{cases}$$

where $k_0 = \max\{i \in \mathbb{Z} \mid Mk_0^2 \leq H - r\}$.

Its proof is obtained by setting $Q = 2$ and $m = 1$ in Theorem 2 and the following Lemma 1.

Lemma 1. *Set*

$$\lambda_0(Q, H_0) = \frac{MQ({k_0}^2 + k_0) + 2H_0}{4(m + k_0 Q)},$$

where $k = \max\{i \in \mathbb{Z} \mid M(Q(i^2 - i) + 2mi) \leq 2H_0\}$, *and*

$$\lambda_1(H_1) = \frac{M}{2} + \frac{M(k_1 + k_1^2) + 2H_1}{4(1 + k_1)} = \frac{M((k_1 + 1) + (k_1 + 1)^2) + 2H_1}{4(1 + k_1)},$$

where $k_1 = \max\{i \in \mathbb{Z} \mid M(k_1^2 + k_1) \leq 2H_1\}$.

We have

1. $(r - 1)\lambda_1(1) + \lambda_1(\sum_{i=1}^{r} H_i - r + 1) \leq \sum_{\tau_1=1}^{r} \lambda_1(H_i)$.

2. $\lambda_0(Q, H_0 + \sum_{i=r+1}^{r'} H_i) \leq \lambda_0(Q, H_0) + \sum_{i=r+1}^{r'} \lambda_1(H_i - 1)$.

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

4. If $m \geq 2$, *then*

$$\lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1) \leq \lambda_1(H_0 + H_1).$$

5. If $m = 0, 1$ *and* $Q = 1$, *then*

$$\lambda_1(H_0 + H_1) \leq \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1).$$

6. Let $m = 1$ *and* $Q \geq 2$.

If $1 \leq H_0 + H_1 \leq M$, *then* $\lambda_1(H_0 + H_1) = \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$.

There exists $\tilde{H} > M$ *such that if* $M + 1 \leq H_0 + H_1 \leq \tilde{H}$ *then* $\lambda_1(H_0 + H_1) \geq \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$, *and if* $\tilde{H} + 1 \leq H_0 + H_1$ *then* $\lambda_1(H_0 + H_1) < \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$.

7. Let $m = 0$ *and* $Q \geq 2$.

If $1 \leq H_0 + H_1 \leq M - 1$, *then* $\lambda_1(H_0 + H_1) > \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$.

There exists $\tilde{H} > M$ *such that if* $M \leq H_0 + H_1 \leq \tilde{H}$ *then* $\lambda_1(H_0 + H_1) \geq \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$, *and if* $\tilde{H} + 1 \leq H_0 + H_1$ *then* $\lambda_1(H_0 + H_1) < \lambda_1(1) + \lambda_0(Q, H_0 + H_1 - 1)$.

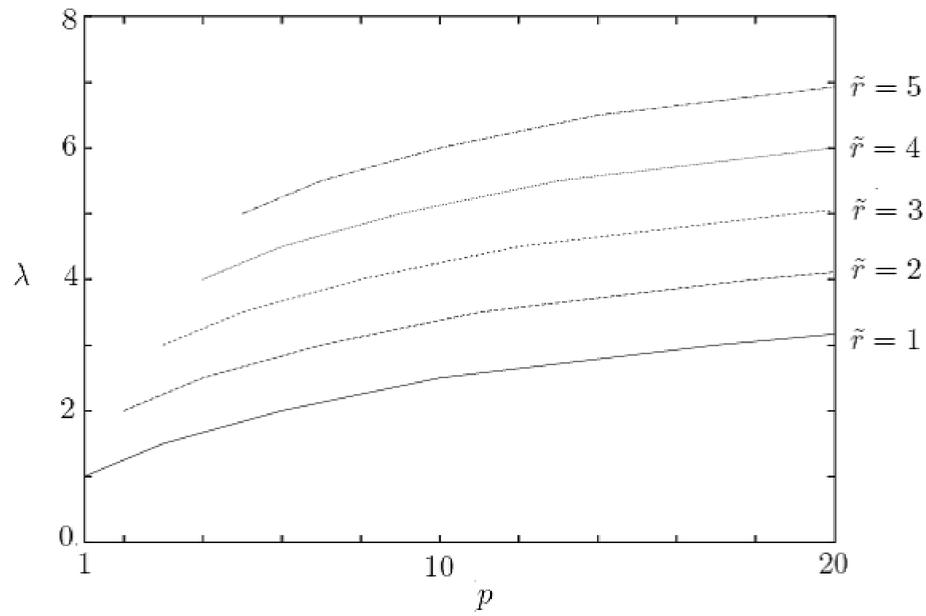

Figure 1: The curves of λ when $r = 1, 2, 3, 4, 5$. x-axis is H and y-axis is λ.

As space is limited, we omit its proof here.

Example 3. Assume that $M = 1$ and a true distribution is given by

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

and a learning model by

$$p(x^{(2)}|x^{(1)}, w) = \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}||x^{(2)} - \sum_{i=1}^{H} a_i \tanh(b_i x^{(1)})||^2).$$

If $H_0 + H_1 + H_2 = H$ and $b_1^* = \cdots = b_{H_0}^* = 0, b_{H_0+1}^* = \cdots = b_{H_0+H_1}^* = 1, a_{H_0+1}^* + \cdots + a_{H_0+H_1}^* = \frac{1}{2}, b_{H_0+H_1+1}^* = \cdots = b_{H_0+H_1+H_2}^* = 2, a_{H_0+H_1+1}^* + \cdots + a_{H_0+H_1+H_2}^* = \frac{1}{2}$, then we have $p(x^{(2)}|x^{(1)}, w_t^*) = p(x^{(2)}|x^{(1)}, w^*)$.

The above theorem shows that for $H - 2 + 1 \leq 10$, we have $\lambda = \frac{3}{2} + \frac{k_1^2 + k_1 + 2(H-1)}{4(k_1+1)}$, $\theta = \begin{cases} 1, & \text{if } k_1^2 + k_1 < 2(H-1), \\ 2, & \text{if } k_1^2 + k_1 = 2(H-1), \end{cases}$

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

For $H - 1 > 10$, we have $\lambda = 2 + \frac{k_0^2 + k_0 + H - 2}{4k_0 + 2}$,

$$\theta = \begin{cases} 1, & \text{if } k_0^2 < H-2, \\ 2, & \text{if } k_0^2 = H-2, \end{cases} \quad \text{where } k_0 = \max\{i \in \mathbb{Z} \mid k_0^2 \leq H-2\}.$$

Figure 1 shows the curves of λ when $M = 1$ and $r = 1, 2, 3, 4, 5$.

References

[1] H. Akaike, *Likelihood and Bayes Procedure. Bayesian Statistics* (Ed. J.M. Bernald), University Press, Valencia, Spain (1980), 143-166.

[2] S. Amari, N. Fujita, S. Shinomoto, Four types of learning curves, *Neural Computation*, **4**, No. 4 (1992), 608-618.

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

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

[5] M. Aoyagi, K. Nagata, Learning coefficient of generalization error of three layered neural networks and normal mixture models in Bayesian estimation, *Preprint*.

[6] M. Aoyagi, S. Watanabe, Resolution of singularities and the generalization error with Bayesian estimation for Layered neural network, *IEICE Trans*, J88-D-II, **10** (2005), 2112-2124; English version: *Systems and Computers in Japan*, John Wiley and Sons, In Press.

[7] M. Aoyagi, S. Watanabe, Stochastic complexities of reduced rank regression in Bayesian estimation, *Neural Networks*, **18** (2005), 924-933.

[8] M. Aoyagi, Learning coefficient of generalization error of normal mixture model with dimension one in Bayesian estimation, *Preprint*.

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

[10] J.E. Björk, *Rings of Differential Operators*, Amsterdam, North-Holland (1979).

[11] K. Fukumizu, A regularity condition of the information matrix of a multi-layer perceptron network, *Neural Networks*, **9**, No. 5 (1996), 871-879.

[12] W. Fulton, *Introduction to Toric Varieties*, Annals of Mathematics Studies Princeton University Press (1993), 131.

[13] K. Hagiwara, N. Toda, S. Usui, 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.

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

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

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

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

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

[19] D.J. Mackay, Bayesian interpolation, *Neural Computation*, **4**, No. 2 (1992), 415-447.

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

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

[22] D. Rusakov, D. Geiger, Asymptotic model selection for Naive Bayesian networks, *Journal of Machine Learning Research*, **6** (2005), 1-35.

[23] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, **6**, No. 2 (1978), 461-464.

[24] B. Sturmfels, Open problems in algebraic statistics, In: *Emerging Applications of Algebraic Geometry* (Ed-s: M. Putinar, S. Sullivant), I.M.A., Volumes in Mathematics and its Applications, **149**, Springer, New York (2008), 351-364.

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

[26] S. Watanabe, Algebraic analysis for nonidentifiable learning machines, *Neural Computation*, **13**, No. 4 (2001), 899-933.

[27] S. Watanabe, Algebraic geometrical methods for hierarchical learning machines, *Neural Networks*, **14**, No. 8 (2001), 1049-1060.

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

[29] S. Watanabe, K. Yamazaki, M. Aoyagi, Kullback information of normal mixture is not an analytic function, *Technical Report of IEICE*, **NC2004** (2004), 41-46.

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

[31] K. Yamazaki, M. Aoyagi, S. Watanabe, Asymptotic analysis of Bayesian generalization error with Newton diagram, *Preprint*.

Appendix A: Proof of Theorem 1

Lemma 2. *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*

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

Lemma 3. *Let*

$$B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ & \vdots & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix} \text{ and } \mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}.$$

Consider a sufficiently small neighborhood U of $\{b_i^*\}_{1\le i\le H}$ and $\{b_i\}_{1\le i\le H}\in U$. Let $b_i^*=\gamma_i|b_i^*|$. Set $\mathbf{b}''_{ij}=\begin{cases}\gamma_i^m\prod_{|b_k^*|=|b_i^*|,1\le k\le j-1}(b_k/\gamma_k-b_i/\gamma_i), & \text{if } b_i^*\neq 0,\\ b_i^m\prod_{b_k^*=0,1\le k\le j-1}(b_k^Q-b_i^Q), & \text{if } b_i^*=0,\end{cases}$ *for $1\le j\le i$ and $\mathbf{b}''_j=(0,\cdots,0,\mathbf{b}''_{jj},\cdots,\mathbf{b}''_{Hj})^t$, for $1\le j\le H$.*

Then there exists a regular matrix R such that

$$B'R=\left(\ \mathbf{b}''_1,\mathbf{b}''_2,\ldots,\mathbf{b}''_H\ \right).$$

Proof. We only need to prove that the vector space generated by $\mathbf{b}''_1,\mathbf{b}''_2,\ldots,$ $\mathbf{b}''_H$ is equal to that generated by $\mathbf{b}'_1,\mathbf{b}'_2,\ldots,\mathbf{b}'_H$.

Some computation shows that the vector space generated by $(b_1^m,\cdots,b_H^m)^t$,
$(0,b_2^m(b_1^Q-b_2^Q),\cdots,b_H^m(b_1^Q-b_H^Q))^t$,
$(0,0,b_3^m(b_1^Q-b_3^Q)(b_2^Q-b_3^Q),\cdots,b_H^m(b_1^Q-b_H^Q)(b_2^Q-b_H^Q))^t$,
$\vdots$
$(0,\cdots,0,b_1^m(b_1^Q-b_H^Q)\cdots(b_{H-1}^Q-b_H^Q))^t$
is equal to that generated by $\mathbf{b}'_1,\mathbf{b}'_2,\ldots,\mathbf{b}'_H$. Therefore, we may set
$\mathbf{b}'_1=(b_1^m,\cdots,b_H^m)$,
$\mathbf{b}'_2=(0,b_2^m(b_1^Q-b_2^Q),\cdots,b_H^m(b_1^Q-b_H^Q))^t$,
$\vdots$
$\mathbf{b}'_H=(0,\cdots,0,b_1^m(b_1^Q-b_H^Q)\cdots(b_{H-1}^Q-b_H^Q))^t$.

We use an induction.

From now on, denote by $\langle\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_H\rangle$ the vector space generated by vectors $\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_H$.

It is easy to check that $\langle\mathbf{b}'_1,\mathbf{b}'_2,\ldots,\mathbf{b}'_H\rangle=\langle\mathbf{b}'_1,\mathbf{b}'_2,\ldots,\mathbf{b}'_{H-1},\mathbf{b}''_H\rangle$.

Let $g_{j,j}(x),g_{j+1,j}(x),\ldots,g_{H,j}(x)$ be polynomials of x, $b_{j-1},\ldots,b_1$ such that $g_{j',j}(x\gamma_{j'})=g_{j'',j}(x\gamma_{j''})$ if $|b_{j'}^*|=|b_{j''}^*|\neq 0$ and $g_{j',j}(x)-g_{j'',j}(x')$ can be devided by $x^Q-x'^Q$ if $b_{j'}^*=b_{j''}^*=0$.

Assume that $(0,\cdots,0,g_{j,j}(b_j)\mathbf{b}''_{jj},\cdots,g_{H,j}(b_H)\mathbf{b}''_{Hj})^t$ is an element of $\langle\mathbf{b}''_j,$ $\ldots,\mathbf{b}''_H\rangle$ and that

$$\langle\mathbf{b}'_1,\cdots,\mathbf{b}'_H\rangle=\langle\mathbf{b}'_1,\cdots,\mathbf{b}'_{j-1},\mathbf{b}''_j,\cdots,\mathbf{b}''_H\rangle.$$

Since $\mathbf{b}'_{j-1} =$

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_{j-1}^m(b_1^Q - b_{j-1}^Q)\cdots(b_{j-2}^Q - b_{j-1}^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q)\cdots(b_{j-2}^Q - b_H^Q) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j-1,j-1}(b_{j-1})\mathbf{b}''_{j-1,j-1} \\ \vdots \\ g_{H,j-1}(b_H)\mathbf{b}''_{H,j-1} \end{pmatrix},$$

where

$$g_{j-1,j-1}(b_{j-1}) \neq 0, \ldots, g_{H,j-1}(b_H) \neq 0,$$

$g_{j',j-1}(\gamma_{j'}x) = g_{j'',j-1}(\gamma_{j''}x)$ if $|b^*_{j'}| = |b^*_{j''}| \neq 0$ and $g_{j',j-1}(x) - g_{j'',j-1}(x')$ can be divided by $x'^Q - x^Q$ if $b^*_{j'} = b^*_{j''} = 0$, we have

$$\mathbf{b}'_{j-1} = \mathbf{b}''_{j-1}g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (g_{j,j-1}(b_j) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{j,j-1} \\ \vdots \\ (g_{H,j-1}(b_H) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{H,j-1} \end{pmatrix}$$

$$= \mathbf{b}''_{j-1}g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{j,j} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{H,j} \end{pmatrix},$$

where $\begin{cases} g_{k,j}(b_k) = g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}), \\ \quad \text{if } |b^*_k| \neq |b^*_{j-1}|, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b_{j-1}/\gamma_{j-1} - b_k/\gamma_k), \\ \quad \text{if } |b^*_k| = |b^*_{j-1}| \neq 0, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b^Q_{j-1} - b^Q_k) \\ \quad \text{if } b^*_k = b^*_{j-1} = 0. \end{cases}$

By the inductive assumption, $(0, \cdots, 0, g_{j,j}(b_j)\mathbf{b}''_{j,j}, \cdots, g_{H,j}(b_H)\mathbf{b}''_{H,j})^t$ is an element of the vector space generated by $\mathbf{b}''_j, \cdots, \mathbf{b}''_H$.

Therefore, $\langle \mathbf{b}'_1, \cdots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \cdots, \mathbf{b}'_{j-1}, \mathbf{b}''_j, \cdots, \mathbf{b}''_H \rangle$
$= \langle \mathbf{b}'_1, \cdots, \mathbf{b}'_{j-2}, \mathbf{b}''_{j-1}, \mathbf{b}''_j, \cdots, \mathbf{b}''_H \rangle$. □

Lemma 4. *Let*

$$B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ \vdots & & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix} \text{ and } \mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}.$$

Consider a sufficiently small neighborhood U *of* $\{b_i^*\}_{1\le i\le H}$ *and* $\{b_i\}_{1\le i\le H} \in U$. *Let* $b_i^* = \gamma_i|b_i^*|$. *Let each* $|b_1^{**}|, \ldots, |b_r^{**}|$ *be a different real number in* $\{|b_i^*| \,;\, |b_i^*| \ne 0\}$:

$$\{|b_1^{**}|, \ldots, |b_r^{**}| \,;\, |b_i^{**}| \ne |b_j^{**}|, i \ne j\} = \{|b_i^*| \,;\, |b_i^*| \ne 0\}.$$

Also set $b_0^{**} = 0$.

Assume that $b_1^* = \cdots = b_{H_0}^* = b_0^{**}$, $|b_{H_0+1}^*| = \cdots = |b_{H_0+H_1}^*| = |b_1^{**}|$, $\ldots$, $|b_{H_0+\cdots+H_{r-1}+1}^*| = \cdots = |b_{H_0+\cdots+H_r}^*| = |b_r^{**}|$.

Set

$$\begin{aligned} (b_1^{(0)}, \ldots, b_{H_0}^{(0)}) &= (b_1, \ldots, b_{H_0}), \\ (b_1^{(1)}, \ldots, b_{H_1}^{(1)}) &= (b_{H_0+1}, \ldots, b_{H_0+H_1}), \\ &\vdots \\ (b_1^{(r)}, \ldots, b_{H_r}^{(r)}) &= (b_{H_0+\cdots+H_{r-1}+1}, \ldots, b_{H_0+\cdots+H_r}). \end{aligned}$$

Let ${b_i^{(\alpha)}}^* = \gamma_i^{(\alpha)}|{b_i^{(\alpha)}}^*|$.

Then there exists a regular matrix R *such that*

$$B'R = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix},$$

where $B^{(0)} = \begin{pmatrix} {b_1^{(0)}}^m & {b_1^{(0)}}^{Q+m} & \cdots & {b_1^{(0)}}^{Q(H_0-1)+m} \\ \vdots & \vdots & & \vdots \\ {b_{H_0}^{(0)}}^m & {b_{H_0}^{(0)}}^{Q+m} & \cdots & {b_{H_0}^{(0)}}^{Q(H_0-1)+m} \end{pmatrix}$ *and*

$$B^{(\alpha)} = \begin{pmatrix} {\gamma_1^{(\alpha)}}^m & {\gamma_1^{(\alpha)}}^m b_1^{(\alpha)}/\gamma_1^{(\alpha)} & {\gamma_1^{(\alpha)}}^m (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^2 & \cdots & {\gamma_1^{(\alpha)}}^m (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^{H_\alpha-1} \\ \vdots & \vdots & & & \vdots \\ {\gamma_{H_\alpha}^{(\alpha)}}^m & {\gamma_{H_\alpha}^{(\alpha)}}^m b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)} & {\gamma_{H_\alpha}^{(\alpha)}}^m (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^2 & \cdots & {\gamma_{H_\alpha}^{(\alpha)}}^m (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^{H_\alpha-1} \end{pmatrix}$$

for $1 \le \alpha \le r$.

Proof. Set

$$\mathbf{b''}^{(0)}_1 = \begin{pmatrix} b_1^{(0)m} \\ b_2^{(0)m} \\ \vdots \\ b_{H_0}^{(0)m} \end{pmatrix} \text{ and } \mathbf{b''}^{(0)}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_j^{(0)m} \prod_{1\le k\le j-1}(b_k^{(0)Q} - b_j^{(0)Q}) \\ \vdots \\ b_{H_0}^{(0)m} \prod_{1\le k\le j-1}(b_k^{(0)Q} - b_{H_0}^{(0)Q}) \end{pmatrix} \text{ for } j \ge 2.$$

Also set , $\mathbf{b''}^{(\alpha)}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \gamma_j^{(\alpha)m} \prod_{1\le k\le j-1}(b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_j^{(\alpha)}/\gamma_j^{(\alpha)}) \\ \vdots \\ \gamma_{H_\alpha}^{(\alpha)m} \prod_{1\le k\le j-1}(b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_H^{(\alpha)}/\gamma_H^{(\alpha)}) \end{pmatrix}$ for $1 \le \alpha \le r$, $2 \le j \le i$.

Then, by Lemma 3, there exists a regular matrix R such that $B'R =$

$$\begin{pmatrix} \mathbf{b''}^{(0)}_1 & \mathbf{b''}^{(0)}_2 & \cdots & \mathbf{b''}^{(0)}_{H_0} & 0 & & & \cdots & & & & 0 \\ \mathbf{b''}^{(1)}_1 & \mathbf{b''}^{(1)}_1 & \cdots & \mathbf{b''}^{(1)}_1 & \mathbf{b''}^{(1)}_1 & \mathbf{b''}^{(1)}_2 & \cdots & \mathbf{b''}^{(1)}_{H_1} & 0 & & \cdots & 0 \\ & & \vdots & & & & \vdots & & & & & \\ \mathbf{b''}^{(r)}_1 & \mathbf{b''}^{(r)}_1 & \cdots & \mathbf{b''}^{(r)}_1 & \mathbf{b''}^{(r)}_1 & \mathbf{b''}^{(r)}_1 & \cdots & \mathbf{b''}^{(r)}_1 & \cdots & \mathbf{b''}^{(r)}_1 & \cdots & \mathbf{b''}^{(r)}_{H_r} \end{pmatrix}.$$

Therefore, we have $B'RR' =$

$$\begin{pmatrix} \mathbf{b''}^{(0)}_1 & \mathbf{b''}^{(0)}_2 & \cdots & \mathbf{b''}^{(0)}_{H_0} & 0 & & & \cdots & & & & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{b''}^{(1)}_1 & \mathbf{b''}^{(1)}_2 & \cdots & \mathbf{b''}^{(1)}_{H_1} & 0 & & \cdots & 0 \\ & & \vdots & & & & \vdots & & & & & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & \mathbf{b''}^{(r)}_1 & \cdots & \mathbf{b''}^{(r)}_{H_r} \end{pmatrix},$$

for some regular matrix R'.

By applying Lemma 3 to $B^{(\alpha)}$, we have the proof. □

Lemma 5. *Let* $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$ *and* $B = (B_I)_{\ell_1+\ldots+\ell_N = Q(n-1)+m, n\in\mathbb{N}}$.

Consider a sufficiently small neighborhood U' *of*

$$\{b^*_{ij}\}_{1\le i\le H, 1\le j\le N} \text{ and } \{b_{ij}\}_{1\le i\le H, 1\le j\le N} \in U'.$$

Let each $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N})$, ..., $(b^{**}_{r1}, b^{**}_{r2}, \cdots, b^{**}_{rN})$ *be a different real vector in*

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

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r1}, \cdots, b^{**}_{rN})\} = \{[b^*_{i1}, \cdots, b^*_{iN}]_Q \neq 0; i = 1, \ldots, H\}.$$

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

Assume that

$$\left.\begin{array}{c} [b^*_{11}, \cdots, b^*_{1N}]_Q \\ \vdots \\ [b^*_{H_0 1}, \cdots, b^*_{H_0 N}]_Q \end{array}\right\} = 0,$$

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

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

$$\vdots$$

$$\left.\begin{array}{c} [b^*_{H_0+\cdots+H_{r-1}+1,1}, \cdots, b^*_{H_0+\cdots+H_{r'-1}+1,N}]_Q \\ \vdots \\ [b^*_{H_0+\cdots+H_{r-1}+H_r,1}, \cdots, b^*_{H_0+\cdots+H_{r-1}+H_r,N}]_Q \end{array}\right\} = (b^{**}_{r1}, \cdots, b^{**}_{rN}).$$

and $H_0 + \cdots + H_r = H$.

Set

$$\begin{array}{l} (b^{(0)}_{1j}, \ldots, b^{(0)}_{H_0 j}) = (b_{1j}, \ldots, b_{H_0 j}), \\ (b^{(1)}_{1j}, \ldots, b^{(1)}_{H_1 j}) = (b_{H_0+1,j}, \ldots, b_{H_0+H_1,j}), \\ \quad\vdots \\ (b^{(r)}_{1j}, \ldots, b^{(r)}_{H_r j}) = (b_{H_0+\cdots+H_{r-1}+1,j}, \ldots, b_{H_0+\cdots+H_r,j}), \end{array}$$

for $1 \leq j \leq N$.

Let $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$, $B^{(\alpha)}_I = \begin{pmatrix} {\gamma^{(\alpha)}_1}^{m-|I|} \prod_{j=1}^N {b^{(\alpha)}_{1j}}^{\ell_j} \\ {\gamma^{(\alpha)}_2}^{m-|I|} \prod_{j=1}^N {b^{(\alpha)}_{2j}}^{\ell_j} \\ \vdots \\ {\gamma^{(\alpha)}_{H_\alpha}}^{m-|I|} \prod_{j=1}^N {b^{(\alpha)}_{H_\alpha j}}^{\ell_j} \end{pmatrix}$

and $B^{(0)} = (B^{(0)}_I)_{\ell_1+\ldots+\ell_N=m+Q(n-1), n\in\mathbb{N}}$, $B^{(\alpha)} = (B^{(\alpha)}_I)_{\ell_1+\ldots+\ell_N=n, n\in\mathbb{N}_{+0}}$ *for*

$1 \leq \alpha \leq r$, *where*

$$\gamma_i^{(\alpha)}(b_{i1}^{(\alpha)^*}, \cdots, b_{iN}^{(\alpha)^*}) = [b_{i1}^{(\alpha)^*}, \cdots, b_{iN}^{(\alpha)^*}]_Q.$$

Then there exists a regular matrix R such that

$$BR = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix}.$$

Proof. The key point of the proof is to use $\begin{pmatrix} \prod_{j=1}^N {b_{1j}}^{\ell_j} \\ \prod_{j=1}^N {b_{2j}}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N {b_{Hj}}^{\ell_j} \end{pmatrix} =$

$$\begin{pmatrix} {b_{11}}^{\ell'_1} \prod_{j=2}^N {b_{1j}}^{\ell_j} & 0 & \cdots & 0 \\ 0 & {b_{21}}^{\ell'_1} \prod_{j=2}^N {b_{2j}}^{\ell_j} & \cdots & 0 \\ \vdots & \ddots & & 0 \\ 0 & 0 & \cdots & {b_{H1}}^{\ell'_1} \prod_{j=2}^N {b_{Hj}}^{\ell_j} \end{pmatrix} \begin{pmatrix} {b_{11}}^{\ell_1 - \ell'_1} \\ {b_{21}}^{\ell_1 - \ell'_1} \\ \vdots \\ {b_{H1}}^{\ell_1 - \ell'_1} \end{pmatrix},$$

and Lemma 4. □

Appendix B: Toric Variety

Here we introduce toric varieties (Fulton [11, 28]). Most of the Kullback functions are degenerate (over $\mathbb{R}$) with respect to their Newton polyhedrons. So we cannot directly obtain desingularization using toric varieties. We can however, use the idea partially for obtaining the maximum pole.

Set $\mathbb{R}_+ = \{r \in \mathbb{R} \mid r \geq 0\}$.

Definition 4. (Convex Rational Polyhedral Cone) A convex polyhedral cone σ is a cone generated by a finite number of vectors $\mathbf{a}_j$ $(j = 1, \ldots, i)$ in $\mathbb{R}^d$: $\sigma = \mathbb{R}_+\mathbf{a}_1 + \cdots + \mathbb{R}_+\mathbf{a}_i = \{r_1\mathbf{a}_1 + \cdots + r_i\mathbf{a}_i \in \mathbb{R}^d \mid r_1 \geq 0, \cdots, r_i \geq 0\}$.

A strongly convex rational polyhedral cone σ is a cone which is generated by vectors $\mathbf{a}_j$ $(i = 1, \ldots, i)$ in $\mathbb{Z}^d$ ("rational"), and contains no line through the origin ("strong").

Definition 5. (Dual of a Set) The dual $\sigma^\vee$ of any set σ is defined by $\sigma^\vee = \{\mathbf{u} \in \mathbb{R}^d \mid \langle \mathbf{u}, \mathbf{v} \rangle \geq 0 \text{ for all } \mathbf{v} \in \sigma\}$.

If σ is a convex polyhedral cone, then $\sigma^\vee$ is also a convex polyhedral cone and $\sigma^\vee \cap \mathbb{Z}^d$ is a finitely generated semigroup (Fulton [12]).

Definition 6. (Face of a Cone) A face $\sigma_{\mathbf{u}}$ of a convex polyhedral cone σ is $\sigma_{\mathbf{u}} = \sigma \cap \{\mathbf{u}\}^\perp = \{\mathbf{v} \in \sigma \mid \langle \mathbf{u}, \mathbf{v} \rangle = 0\}$ for some $\mathbf{u} \in \sigma^\vee$.

Definition 7. (Fan) A fan Δ is a collection of strongly convex rational polyhedral cones, satisfying the following conditions: every face of a cone in Δ is also a cone in Δ, and the intersection of two cones in Δ is a face of each.

Suppose that $\mathbf{a}_1, \ldots, \mathbf{a}_i \in \mathbb{Z}^d$ of a cone $\sigma = \mathbb{R}_+\mathbf{a}_1 + \cdots + \mathbb{R}_+\mathbf{a}_i$, are the first points in $\mathbb{Z}^d$ along the edges of σ. Then σ is called non-singular if $\mathbf{a}_1, \ldots, \mathbf{a}_i$ is a part of a basis of $\mathbb{Z}^d$.

Also a fan Δ is called non-singular if every cone in Δ is non-singular.

Definition 8. (Toric Variety) For a fan Δ and a cone $\sigma \in \Delta$, consider a group ring

$$R(\sigma) = \bigoplus_{\mathbf{u} \in \sigma^\vee \cap \mathbb{Z}^d} \mathbb{R}x^u = \{ \sum_{\mathbf{u} \in \sigma^\vee \cap \mathbb{Z}^d} c_{\mathbf{u}} x^{\mathbf{u}} \text{ finite sum } \mid c_{\mathbf{u}} \in \mathbb{R}\},$$

where $x^{\mathbf{u}}$ is a basis, as $\mathbf{u}$ varies over $\mathbf{u} \in \sigma^\vee$ with multiplication $x^{\mathbf{u}} x^{\mathbf{u}'} = x^{\mathbf{u}+\mathbf{u}'}$. Let

$$\begin{aligned} U_\sigma &= \mathrm{Hom}(R(\sigma), \mathbb{R}) \\ &= \{P : R(\sigma) \to \mathbb{R} \mid \text{ ring homomorphism with } P(1) = 1\}. \end{aligned}$$

The toric variety $X(\Delta)$ is defined by taking the disjoint union of U_σ, $\sigma \in \Delta$, and gluing U_σ to U_τ by the identification at $U_{\sigma\cap\tau}$. For $\sigma, \tau \in \Delta$, $U_{\sigma\cap\tau}$ is identified as a principal open subvariety of U_σ and U_τ.

A fan Δ is non-singular, then $X(\Delta)$ is a non-singular manifold (Fulton [12]).

Definition 9. (Refinement of a Fan) A fan Δ' is called a refinement of a fan Δ, if there exists $\sigma \in \Delta$ such that $\sigma' \subset \sigma$ for any $\sigma' \in \Delta'$, and if $\cup_{\sigma\in\Delta}\sigma = \cup_{\sigma'\in\Delta'}\sigma'$.

Definition 10. (Rational Convex Polytope) A convex polytope is defined as $\Gamma = \cap_{i=1}^m \{\mathbf{u} \in \mathbb{R}^d \mid \langle \mathbf{u}, \mathbf{v}_i \rangle \geq \rho_i\}$, for some $\mathbf{v}_i \in \mathbb{R}^d$ and $\rho_i \in R$, which is the convex hull of a finite set of points.

If $\mathbf{v}_i \in \mathbb{Z}^d$ and $\rho_i \in \mathbb{Z}$ then the convex polytope is called rational.

A face $\Gamma(\mathbf{v})$ of Γ for $\mathbf{v} \in \mathbb{Z}^d$, is the intersection with a supporting affine hyperplane: $\Gamma(\mathbf{v}) = \{\mathbf{u} \in \Gamma \mid \langle \mathbf{u}, \mathbf{v} \rangle = \min_{\mathbf{u}'\in\Gamma} \langle \mathbf{u}', \mathbf{v} \rangle\}$.

Theorem 5. (see Fulton [12]) *Let Γ be a rational convex polytope. Define a cone σ_F by $\sigma_F = \{\mathbf{v} \in \mathbb{R}^d \mid \Gamma(\mathbf{v}) \supset F\}$, for a face F of Γ. Then $\Delta =$*

$\{\sigma_F \mid F$ *is a face of* $\Gamma\}$ *is a fan.*

Theorem 6. (see Fulton [12]) *For any fan* Δ*, there is a refinement* Δ' *of* Δ *so that* Δ' *is non-singular.*

Then the morphism map from $X(\Delta')$ *to* $X(\Delta)$ *induced by the natural map* $U_{\sigma'} \to U_\sigma$ *for* $\sigma' \subset \sigma$*, is a resolution of singularities.*

For $i = 1, \ldots, d$, set $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{Z}^d$, whose i-th element is 1, and $V = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$. (T denotes the transpose).

Let

$$L = (\mathbf{l}_1, \cdots, \mathbf{l}_d) = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1d} \\ l_{21} & l_{22} & \cdots & l_{2d} \\ \vdots & \vdots & \cdots & \vdots \\ l_{d1} & l_{d2} & \cdots & l_{dd} \end{pmatrix}.$$

Define ${}^L y = (y_1^{l_{11}} y_2^{l_{12}} \cdots y_d^{l_{1d}}, y_1^{l_{21}} y_2^{l_{22}} \cdots y_d^{l_{2d}}, \cdots, y_1^{l_{d1}} y_2^{l_{d2}} \cdots y_d^{l_{dd}})$, for $y = (y_1, \cdots, y_d)$.

Fix

$$\tilde{\Delta} = \{\sigma \mid \sigma = \sum_{i=1}^{m} \mathbb{R}_+ \mathbf{v}_i, \mathbf{v}_i \in V, 1 \leq i \leq m \leq d\} \cup \{0\}, \tag{3}$$

in this paper. $\cup_{\sigma \in \tilde{\Delta}} \sigma$ is the first quadrant.

Then the toric variety $X(\tilde{\Delta})$ is identified as $\mathbb{R}^d$ by the map

$$U_{\sum_{i=1}^d \mathbb{R}^d \mathbf{e}_i} \tilde{\to} \mathbb{R}^d; \quad P \mapsto (y_1, \ldots, y_d) := (P(x^{e_1}), \ldots, P(x^{e_d})).$$

Let Δ be a non-singular fan and a refinement of $\tilde{\Delta}$ in (3).

The toric variety $X(\Delta)$ is constructed as follows.

For a d-dimensional $\sigma = \sum_{i=1}^d \mathbb{R}_+ \mathbf{a}_i \in \Delta$, where the set of $\mathbf{a}_1, \ldots, \mathbf{a}_d$ is a basis of $\mathbb{Z}^d$, we have $U_\sigma \cong \mathbb{R}^d$; $U_\sigma \ni P \mapsto (y_1, \ldots, y_d) := (P(x^{\mathbf{a}_1}), \ldots, P(x^{\mathbf{a}_d})) \in \mathbb{R}^d$.

For d-dimensional $\sigma = \sum_{i=1}^d \mathbb{R}_+ \mathbf{a}_i \in \Delta$ and $\tau = \sum_{i=1}^d \mathbb{R}_+ \mathbf{b}_i \in \Delta$, assume $\mathbf{a}_{s_i}, \mathbf{b}_{t_i} \notin \sigma \cap \tau, i = 1, \cdots, m_0$.

Take the coordinate systems of U_σ and U_τ by y^σ and y^τ, respectively.

The identification on $U_{\sigma \cap \tau}$ is

$$y^\sigma \sim y^\tau \iff {}^{A_\tau^{-1} A_\sigma} y^\sigma = y^\tau, y^\sigma_{s_i} \neq 0, y^\tau_{t_i} \neq 0, i = 1, \cdots, m_0,$$

where $A_\sigma = (\mathbf{a}_1, \cdots, \mathbf{a}_d)$ and $A_\tau = (\mathbf{b}_1, \cdots, \mathbf{b}_d)$.

Then $X(\Delta)$ is $\amalg_{\dim \sigma = d} U_\sigma / \sim$.

The map π from $X(\Delta)$ to $X(\tilde{\Delta}) \cong \mathbb{R}^d$ is

$$\pi_\sigma : y^\sigma = (y_1, \cdots, y_d) \in U_\sigma \mapsto {}^{A_\sigma} y^\sigma \in \mathbb{R}^d.$$

Lemma 6. *Let $L = (\mathbf{l}_1, \ldots, \mathbf{l}_d)$ be any regular $d \times d$ matrix, where d dimensional vectors $\mathbf{l}_i$ are in $\mathbb{Z}_+^d$.*

Set $\sigma_L = \sum_{i=1}^d \mathbb{R}_+ \mathbf{l}_i$.

Then there is a refinement fan Δ of $\tilde{\Delta}$ in (3) such that $\sigma_L \in \Delta$.

Proof. Set $\mathbf{e} = (1, \ldots, 1)^T$ and $\rho_i = \langle \mathbf{e}, \mathbf{l}_i \rangle$ for $i = 1, \ldots, d$. Let $\Gamma = \cap_{i=1}^d \{\mathbf{u} \in \mathbb{R}^d \mid \langle \mathbf{u}, \mathbf{l}_i \rangle \geq \rho_i\} \cap_{i=1}^d \{\mathbf{u} \in \mathbb{R}^d \mid \langle \mathbf{u}, \mathbf{e}_i \rangle \geq 0\}$. Then by Theorem 5, $\Delta = \{\sigma_F \mid F \text{ is a face of } \Gamma\}$ is a fan where $\sigma_F = \{\mathbf{v} \in \mathbb{R}^d \mid \Gamma(\mathbf{v}) \supset F\}$. Let $F = \Gamma(\mathbf{l}_1 + \cdots + \mathbf{l}_d) = \{\mathbf{u} \in \Gamma \mid \langle \mathbf{u}, \mathbf{l}_1 + \cdots + \mathbf{l}_d \rangle = \min_{\mathbf{u}' \in \Gamma} \langle \mathbf{u}', \mathbf{l}_1 + \cdots + \mathbf{l}_d \rangle\} = \cap_{i=1}^d \{\mathbf{u} \in \Gamma \mid \langle \mathbf{u}, \mathbf{l}_i \rangle = \rho_i\}$. Since L is regular, $F = \{\mathbf{e}\}$. We will show that $\sigma_L = \sigma_F \in \Delta$. The fact $\mathbf{e} \in \Gamma(\mathbf{l}_i)$ yields $\sigma_L \subset \sigma_F$. Suppose $\mathbf{v} \in \sigma_F \setminus \sigma_L$ and $\mathbf{v} = r_1 \mathbf{l}_1 + \cdots + r_d \mathbf{l}_d$ for $r_i \in \mathbb{R}$. Then some r_i are minus. Assume that $r_{i_1} < 0$. Let $\mathbf{u}_1$ be a vector satisfying $\langle \mathbf{u}_1, \mathbf{l}_i \rangle = 0$ for $i \neq i_1$ and $\langle \mathbf{u}_1, \mathbf{l}_{i_1} \rangle = 1$. For a large number I, we have $\mathbf{e} + \mathbf{u}_1/I \in \Gamma$ and $\langle \mathbf{e}, \mathbf{v} \rangle = \sum_{i=1}^d r_i \rho_i > \langle \mathbf{e} + \mathbf{u}_1/I, \mathbf{v} \rangle = \sum_{i=1}^d r_i \rho_i + r_{i_1}/I$. This is a contradiction to $\mathbf{v} \in \sigma_F$, i.e., $\mathbf{e} \in \Gamma(\mathbf{v})$. Therefore $\sigma_F = \sigma_L$. Finally we show that Δ is a refinement of $\tilde{\Delta}$. Let F be any face of Γ. If $\sigma_F \not\subset \sum_{i=1}^d \mathbb{R}_+ \mathbf{e}_i$, then there is a vector $\mathbf{v} = (v_1, \ldots, v_d)^T \in \sigma_F$ with some $v_{i_0} < 0$. For any large number I, $\mathbf{e} + I\mathbf{e}_{i_0} \in \Gamma$ and $\langle \mathbf{e} + I\mathbf{e}_{i_0}, \mathbf{v} \rangle \to -\infty$ as $I \to \infty$. This is a contradiction to $\langle \mathbf{u}, \mathbf{v} \rangle = \min_{\mathbf{u}' \in \Gamma} \langle \mathbf{u}', \mathbf{v} \rangle$ for any $\mathbf{u} \in F$. Therefore $\sigma_F \subset \sum_{i=1}^d \mathbb{R}_+ \mathbf{e}_i$. Since $\min_{\mathbf{u}' \in \Gamma} \langle \mathbf{u}', \mathbf{e}_i \rangle \geq 0$, we have $\Gamma(\mathbf{e}_i) = \{\mathbf{u} \in \Gamma \mid \langle \mathbf{u}, \mathbf{e}_i \rangle = \min_{\mathbf{u}' \in \Gamma} \langle \mathbf{u}', \mathbf{e}_i \rangle\} \neq \phi$. Therefore $\sigma_{\Gamma(\mathbf{e}_i)} \supset \mathbb{R}_+ \mathbf{e}_i$. That is, $\cup_F \sigma_F = \sum_{i=1}^d \mathbb{R}_+ \mathbf{e}_i$. □

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_+ (Fulton [12]).

Remark 3. Let

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

$g = u_1^{t_1} u_2^{t_2} \cdots u_d^{t_d} du$ 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 $\int_{\text{near } 0} (f_1^2 + \cdots + f_p^2)^z g$ is $1/c$ and its order is θ. In this case, the condition $c > 1$ is not necessary.

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

Then for $f(x_1^{(1)}, \dots, x_{d_1}^{(1)}, \dots, x_1^{(r)}, \dots, x_{d_r}^{(r)}) = \sum_{\alpha=1}^r f_\alpha$ and $w^* = (w_1^*, \dots, w_r^*)$, 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

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

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

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

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

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

We can choose $\sum_{\alpha=1}^{r} \theta_\alpha - (r-1)$ independent vectors $\mathbf{l}_j$ satisfing (a) and (b) by using Lemma 6, and this fact completes the proof. □