NNK
Title:
Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram

Authors:
Keisuke Yamazaki, Miki Aoyagi and Sumio Watanabe

Affiliations:
Precision and Intelligence Laboratory, Tokyo Institute of Technology

Corresponding Author:
Keisuke Yamazaki

Postal Mail:
Keisuke Yamazaki
Watanabe Laboratory
Precision and Intelligence Laboratory
Tokyo Institute of Technology
R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

Phone: +81-45-924-5018
Fax: +81-45-924-5018
E-mail: k-yam@pi.titech.ac.jp

Asymptotic Analysis of Bayesian Generalization Error with Newton Diagram

Keisuke Yamazaki[†], Miki Aoyagi[‡] and Sumio Watanabe[†]

[†]Precision and Intelligence Laboratory,
Tokyo Institute of Technology

R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan

[‡]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.

Abstract

Statistical learning machines that have singularities in the parameter space, such as hidden Markov models, Bayesian networks, and neural networks, are widely used in the field of information engineering. Singularities in the parameter space determine the accuracy of estimation in the Bayesian scenario. The Newton diagram in algebraic geometry is recognized as an effective method by which to investigate a singularity. The present paper proposes a new technique to plug the diagram in the Bayesian analysis. The proposed technique allows the generalization error to be clarified and provides a foundation for efficient model selection. We apply the proposed technique to mixtures of binomial distributions.

Keywords: Bayes generalization error, Statistical singular models, Newton diagram

1 Introduction

One of the tasks in the statistical machine learning is to investigate the generalization performance. In particular, in parametric models, there have been a number of theoretical studies on generalization under various conditions. Identifiability is an important condition in order to guarantee that the model is statistically regular in the asymptotic situation. However, almost all practical machines, e.g., mixture models, hidden Markov models, and Bayesian networks, are unidentifiable (we present a formal definition in Section 2.1).

The importance of the study of unidentifiable models has been reported (Hartigan, 1985; Amari & Ozeki, 2001). In some models, such as mixture models, the maximum likelihood estimator often diverges. Dacunha-Castelle and Gassiat (Dacunha-Castelle & Gassiat, 1997) proposed that the asymptotic behavior of the log likelihood ratio of the maximum likelihood method can be analyzed based on the theory of empirical processes by choosing a locally conic parameterization. Hagiwara (Hagiwara, 2002) has shown that the maximum likelihood method makes training errors very small but conjectured that this method also makes generalization errors very large. These results imply that the Bayes estimation also requires a novel method to analyze the unidentifiable models.

Therefore, the algebraic geometrical analysis was developed for the Bayesian scenario (Watanabe, 2001). This analysis allows the asymptotic form of the generalization error to be derived in many unidentifiable models (Aoyagi & Watanabe, 2005; Rusakov & Geiger, 2005; Yamazaki & Watanabe, 2003a; Yamazaki & Watanabe, 2003b; Yamazaki & Watanabe, 2005b; Yamazaki & Watanabe, 2005a). Using these theoretical forms of the error, a number of approximation methods, such as the variational and empirical Bayes methods, are evaluated (Watanabe & Watanabe, 2006; Nakajima & Watanabe, 2007). Moreover, the coefficient of the leading term in the form includes the information of the distribution generating data and can be used for model selection (Yamazaki et al., 2006). Therefore, it is necessary to reveal the generalization error of unidentifiable models for both theoretical and practical interests.

The algebraic geometrical approach reveals that the properties of singularities in the parameter space are essential in order to determine the accuracy of Bayes estimation. In algebraic geometry, singularities have been investigated for a long time, and it has been proven that the properties can be clarified by iterative blow-up (Hironaka, 1964). The former study (Aoyagi & Watanabe, 2004) follows this method in order to reveal the asymptotic error of neural networks. However, the transform could be very complicated, and the iterative method is not straightforward when the parameter space is high dimensional. The Newton diagram is another tool that can be used to find the transform (Fulton, 1993; Ewald, 1996). It is known that the blowing-up method is always available for the singularities, to which the Newton diagram method is applicable. These two methods are actually equivalent in the sense of the final result of transform. However, the processes to find it are totally different from each other. Compared with the iterative method, the Newton diagram is geometrically intuitive and has a smaller computational cost. Therefore, in the present paper, we propose a new approach to utilize the diagram as a sophisticated analysis for unidentifiable models and demonstrate the efficiency of this method by applying it to mixtures of binomial distributions. Actually, the mixture of binomial distributions is a practical model used for the gene analysis and the mutational spectrum analysis (Glazko et al., 1998). However, it is not straightforward to confirm whether

our asymptotic result is still valid in some real-world situation. Then, this paper focuses on the comparison between the iterative and proposed methods.

The remainder of the present paper is organized as follows. Section 2 summarizes the Bayes estimation and briefly introduces the algebraic geometrical method. Section 3 describes the Newton diagram through an example. The formal definitions are presented in the Appendices. In Section 4, we consider the application of the proposed method to mixture models. Finally, the discussion and conclusions are presented in Section 5.

2 Bayes Estimation and Singularities

In this section, we show that the Bayes generalization error given by the average Kullback divergence is determined by a zeta function. This is the relationship between the statistical learning theory and algebraic geometry.

2.1 Bayesian Learning on Unidentifiable Models

Let $X^n = (X_1, X_2, \cdots, X_n)$, such that $X_i \in R^M$ is a set of training samples that are independently and identically distributed. The number of training samples is n. The training samples and testing samples are taken from the true probability distribution $q(x)$. Let $p(x|w)$ and $\varphi(w)$ be a learning machine and a prior, respectively. Then, the a posteriori probability distribution is defined by

$$p(w|X^n) = \frac{1}{Z_0(X^n)}\ \varphi(w)\ \prod_{i=1}^{n} p(X_i|w), \tag{1}$$

where $Z_0(X^n)$ is a normalizing constant. The Bayesian predictive distribution $p(x|X^n)$ is given by

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw. \tag{2}$$

The generalization error $G(n)$ is the average Kullback information from the true distribution to the Bayesian predictive distribution,

$$G(n) = E_{X^n}\Big[\int q(x)\log\frac{q(x)}{p(x|X^n)}dx\Big], \tag{3}$$

where $E_{X^n}[\cdot]$ is the expectation over all samples. The error is a decreasing function with respect to the sample size n. The asymptotic form is given by

$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n\log n} + o\big(\frac{1}{n\log n}\big), \tag{4}$$

when the learning machine $p(x|w)$ can attain the true model $q(x)$, i.e., when there exist the true parameters, which are defined by $\{w^*|p(x|w^*) = q(x)\}$. The existence of w^* causes the error to go to zero as the sample size increases. Eq. (4) shows that the posterior is constructed around the true parameters and that distributed points from the posterior make the error converge to zero. The terms in the asymptotic form show the speed of convergence. In identifiable models, $\lambda = d/2$ and $m = 1$, where d is the number of parameters. The coefficient λ is strongly connected to the score of the model selection (Levin et al., 1989; Yamanishi, 1998). For example, $\lambda = d/2$ in identifiable models appears in the penalty term of BIC (Schwarz, 1978) or MDL (Rissanen, 1986).

Here, let us formally define the identifiability.

Definition 1 (identifiability) *If $p(x|w_1) \neq p(x|w_2)$ for $w_1 \neq w_2$, the model $p(x|w)$ is identifiable. Otherwise, it is unidentifiable.*

Our model $p(x|w)$ now includes unidentifiable cases. The true parameters construct a set of parameters. If the set consists of only a point, the model is identifiable and the posterior asymptotically converges to a Gaussian distribution, the mean of which is the point. It is not a special case that the true model $q(x)$ is expressed as a set of parameters $\{w^*\}$. For example, let $q(x) = N(x; 0, 1)$ and $p(x|w) = aN(x; b, 1) + (1 - a)N(x; c, 1)$, where $x \in R^1$, $N(x; \mu, \sigma^2)$ is a Gaussian distribution with the mean μ and the variance σ^2, and $w = \{a, b, c\}$, such that $0 \leq a \leq 1$. The true parameters are expressed as the set $\{a = 1, b = 0\} \cup \{a = 0, c = 0\} \cup \{b = c = 0\}$. This means that the model is unidentifiable if the size of the learning model is larger than that of the true model. Such a situation often occurs in practice.

2.2 Singularities and the Generalization Error

We next explain how to calculate the coefficients λ and m in Eq.(4). We define the Kullback divergence from the true distribution $q(x)$ to the learning model $p(x|w)$ by

$$H(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx, \tag{5}$$

and assume that it is an analytic function. Watanabe (Watanabe, 2001) proved that λ and m are defined by the largest pole of a zeta function

$$J(z) = \int H(w)^z \varphi(w) dw, \tag{6}$$

where z is a complex variable. This zeta function is known to have poles on only real negative rational points. Let the largest pole and its order be $-\lambda$ and m, respectively. Then, these factors determine the Bayes generalization error.

Example 1 *Let the Kullback divergence be defined by*

$$H(w) = w_1^2 w_2^2 w_3^4 w_4^6, \tag{7}$$

where $w = \{w_1, w_2, w_3, w_4\}$ *are the parameters of some learning model. For simplicity, a prior is the uniform distribution* $[-1, 1]$*, for each* w_i*. The zeta function is then given by*

$$J(z) \propto \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} (w_1^2 w_2^2 w_3^4 w_4^6)^z dw_1 dw_2 dw_3 dw_4. \tag{8}$$

It is easy to integrate the function and find the poles,

$$J(z) \propto \frac{1}{(2z+1)^2}\frac{1}{4z+1}\frac{1}{6z+1} f(z), \tag{9}$$

where $f(z)$ *is a holomorphic function. Note that the holomorphic function does not affect the poles. Thus, the largest pole* $z = -\lambda = -1/6$ *and its order* $m = 1$ *are derived. The generalization error is asymptotically calculated as*

$$G(n) = \frac{1}{6n} + o\big(\frac{1}{n \log n}\big). \tag{10}$$

As you may notice, the integration can be computed if $H(w)$ takes the form

$$H(w) = f_c(w) \prod_{i=1}^{d} w_i^{2\alpha_i}, \tag{11}$$

where $\alpha_i \in Z$ and $f_c(w)$ is positive in the neighborhood of $w = 0$. In the present paper, this form is referred to as the *product form*. If Eq.(5) has a polynomial form such as $H(w) = (w_1 w_2 + w_3 w_4)^2$, there is a mapping $w = g(u)$, where $H(g(u))$ has the product form with respect to u (Hironaka, 1964). It is proved that the mapping will be found by the iteration of blow-ups, which is a special coordinate transform. The final mapping $g(u)$ is referred to as the *resolution of singularities*.

In the statistical learning, $H(w)$ is not in product form and finding the resolution of singularities is still complicated (Aoyagi & Watanabe, 2005). Under certain conditions, the Newton diagram allows $g(u)$ to be found in a very systematic and geometric manner, as will be described in the following section.

3 Newton Diagram and Resolution of Singularities

In this section, we introduce the Newton diagram and its relation to the resolution of singularities. According to this method, we can rewrite the polynomial form of $H(w)$ in product form

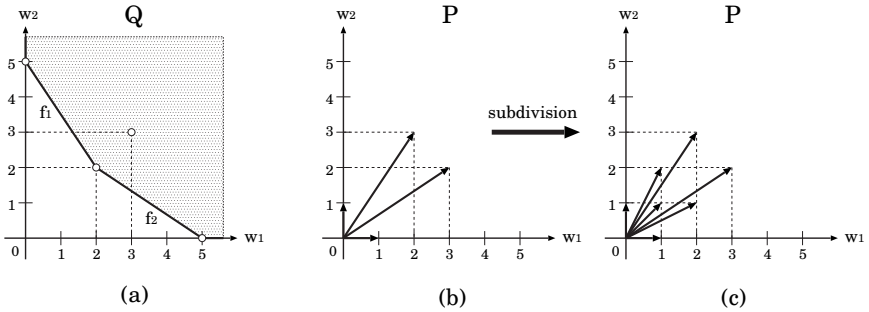

Figure 1: (a) Newton diagram, (b) fan, and (c) subdivided fan

$H(g(u))$ and find the largest pole of the zeta function. This method will be explained with a simple intuitive example. Refer to the appendix for the formal notations.

Assume that $H(w)$ is defined as

$$H(w) = w_1^5 + w_1^3 w_2^3 + w_1^2 w_2^2 + w_2^5, \tag{12}$$

where $w = (w_1, w_2)$. Let us consider the space of the exponent part, where (x, y) corresponds to (w_1^x, w_2^y). In this space, Eq. (12) has four points $v = (5,0), (3,3), (2,2), (0,5)$. The *Newton diagram* is the shaded area in Fig.1 (a) constructed by these points. (See the formal definition in the appendix.) The *face* is the border of the diagram. Note that, when the diagram is d dimensional, there are $d-1, d-2, \ldots, 1, 0$ dimensional faces as the border. In this example, there are four faces, f_1=$[(0,5),(2,2)]$, f_2=$[(2,2),(5,0)]$, $[(0,5),(0,\infty)]$, and $[(5,0),(\infty,0)]$. (The zero-dimensional faces are abbreviated.)

Let us define the following functions:

$$f_{\gamma 1}(w) = w_2^5 + w_1^2 w_2^2, \tag{13}$$

$$f_{\gamma 2}(w) = w_1^5 + w_1^2 w_2^2, \tag{14}$$

which consist of the compact faces f_1 and f_2, respectively. The function $H(w)$ is *non-degenerative* iff $\forall i$

$$\left\{(w_1, w_2) | \frac{\partial f_{\gamma i}(w)}{\partial w_1} = \frac{\partial f_{\gamma i}(w)}{\partial w_2} = 0\right\} \subset \{(w_1, w_2) | w_1 w_2 = 0\}. \tag{15}$$

This is the condition for applying the Newton diagram. We can easily confirm that the function (12) is non-degenerate. According to the definition,

$$\begin{cases} \frac{\partial f_{\gamma 1}(w)}{\partial w_1} & = 2w_1 w_2^2 = 0 \\ \frac{\partial f_{\gamma 1}(w)}{\partial w_2} & = 5w_2^4 + 2w_1^2 w_2 = 0. \end{cases} \tag{16}$$

The solution space is $w_2 = 0$, which is included in $\{w_1 w_2 = 0\}$. Evaluating $\partial f_{\gamma 2}(w)/\partial w_1 = \partial f_{\gamma 2}(w)/\partial w_2 = 0$, we can derive the solution space $w_1 = 0 \subset \{w_1 w_2 = 0\}$.

The *fan* is defined by orthogonal vectors of the face. Consider a parallelogram that consists of two arbitrary vectors of the fan. If the parallelogram includes a point of the lattice, add the vector from the origin to the point. The *subdivided fan* is defined by adding the vectors until there is no point in the parallelograms. The fan and the subdivided fan of the Newton diagram are shown by (b) and (c), respectively, in Fig. 1. For example, considering the parallelogram spanned by $(1, 0)$ and $(3, 2)$, we can find the inner point $(2, 1)$ of the lattice. In the similar way, we can find $(1, 2)$ and $(1, 1)$, too. Note that the subdivided fan is regarded as base vectors in the lattice space, which allows the fan to include neither $(2, 2)$ nor $(3, 1)$. For higher dimensional cases, see (Fulton, 1993).

Choose two vectors, such as $(3, 2)$ and $(1, 1)$, from the subdivided fan and define the matrix

$$A = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix}$$

with combining them.

If $\det A = \pm 1$, the mapping according to the matrix such that

$$\begin{cases} w_1 = u_1^3 u_2^1 \\ w_2 = u_1^2 u_2^1 \end{cases} \tag{17}$$

provides the product form (11),

$$H(g(u)) = u_1^{10} u_2^4 (u_1^5 u_2 + u_1^5 u_2^2 + 1 + u_2). \tag{18}$$

We can find a resolution of singularities at the origin considering all sets of vectors such that $\det A = \pm 1$. Each A provides local coordinates.

In order to find the largest pole of the zeta function determined by Eq. (6), it is necessary to find an efficient vector of the subdivided fan in the Newton diagram. The largest pole depends on the ratio between the Jacobian $|g'(u)|$ and the power of the common factor in $H(g(u))$. If a vector of the subdivided fan is $(a_1, \ldots, a_j, \ldots, a_d)$ and $H(g(u))$ has the common factor u_j^β, a pole is $z = -\alpha/\beta$, where $\alpha = \sum_i a_j$. For example, $\alpha = 3 + 2 = 5$ for u_1 in the above mapping. It is easy to find $\beta = 10$ in (18). Then a pole is calculated by $z = -5/10 = -1/2$. Finding the largest pole requires to check all possible variables u_j with all possible matrices A.

Here, we summarize the above-mentioned procedure as an algorithm.

Algorithm 1 (Resolution of singularities with the Newton diagram) _

1. *Draw the Newton diagram of $H(w)$.*

2. *Find the faces.*

3. *Verify that $H(w)$ is non-degenerative.*

4. *Find and subdivide the fan.*

5. *Select d vectors and define the mapping A such that* $\det A = \pm 1$.

6. *Find the largest pole based on all possible As.*

Note that d is the dimension of the parameter space. Because of the subdivision of the fan, we can always find the matrix A in *5*. Moreover, an algorithm and software can be used to find the faces, fans, and A automatically (Aranson, 2001).

If a learning machine has degenerative $H(w)$, we cannot apply this method directly. Therefore, we propose a method by which to change the function into a non-degenerative function. This method is described in the next section.

4 Application to the Mixture Model

In this section, let us apply the proposed method to a mixture of binomial distributions and analyze the generalization error. Mixture models are commonly used in a number of engineering fields and are representative singular models. The binomial mixture model has a simple structure of the parameter space because each component consists of a parameter. Therefore, as the first step for developing the proposed method, we focus on the model in order to apply the proposed method. It is a challenging future study to extend the result to other mixture models.

First, we formulate the model and derive the Kullback divergence. Since the divergence is degenerative, the Newton diagram is not directly applicable. Therefore, we next propose a method by which to change the coordinates of the parameter space, which makes the diagram usable.

4.1 Mixture of Binomial Distributions

A mixture of binomial distributions is formulated by

$$p(x=k|w) = \binom{N}{k} \left\{ \sum_{i=1}^{K} a_i p_i^k (1-p_i)^{N-k} \right\}, \tag{19}$$

where N, K are integers such that $2K \geq N$, $k = 0, 1, \cdots, N$, $(N\ k)^T$ is the number of combinations of N elements taken k at a time, and

$$w = (\{a_i\}_{i=1}^{K-1}, \{p_i\}_{i=1}^{K}) \tag{20}$$

is a parameter such that $0 < p_i \leq 1/2$, $a_i \geq 0$, and

$$a_K = 1 - \sum_{i=1}^{K-1} a_i. \tag{21}$$

A binomial distribution is defined by

$$\binom{N}{k} \bar{p}^k (1-\bar{p})^{N-k}, \tag{22}$$

where $1 < \bar{p} \leq 1/2$. Thus, the mixture (19) has K components.

4.2 Kullback Divergence of the Binomial Mixture Model

We use the following notation:

$$H(w) \gtrless K(w), \tag{23}$$

where there are positive constants C_1, C_2 such that

$$C_1 K(w) \leq H(w) \leq C_2 K(w) \tag{24}$$

in the neighborhood of $H(w) = 0$. Assume that the true distribution, which generates data, consists of K_0 components,

$$q(x=k) = \binom{N}{k} \left\{ \sum_{i=1}^{K_0} a_i^* p_i^{*k} (1-p_i^*)^{N-k} \right\}, \tag{25}$$

where $0 < p_i^* \leq 1/2$ are constants, $p_1^* < p_2^* < \cdots < p_{K_0}^*$, $a_i^* > 0$ and $a_{K_0}^* = 1 - \sum_{i=1}^{K_0-1} a_i^*$. The algebraic geometrical method requires the Kullback divergence, which is integral with respect to x (Eq.(5) and Example 1). When x is discrete, the Kullback divergence has a convenient property:

Lemma 1 *For a discrete domain X, it holds that*

$$H(w) = \sum_{x \in X} q(x) \log \frac{q(x)}{p(x|w)} \tag{26}$$

$$\gtrless \sum_{x \in X} \big\{ p(x|w) - q(x) \big\}^2. \tag{27}$$

Proof: Let us define a function of $y \in R^1$ as

$$S(y) = \log y + y - 1 \tag{28}$$

for $y > 0$. It is easy to confirm that there are constants C_{y1} and C_{y2} such that

$$C_{y1}(y-1)^2 \leq S(y) \leq C_{y2}(y-1)^2 \tag{29}$$

in the neighborhood of $y = 1$. Therefore,

$$H(w) = \sum_{x \in X} q(x) S(p(x|w)/q(x)) \tag{30}$$

$$\geqq \sum_{x \in X} q(x) \Big\{ \frac{p(x|w)}{q(x)} - 1 \Big\}^2 \tag{31}$$

$$\geqq \sum_{x \in X} \big\{ p(x|w) - q(x) \big\}^2. \tag{32}$$

(End of Proof)

As for the Kullback divergence of the binomial mixture, we can prove the following theorem:

Theorem 1 *If the learning machine is given by (19) and the true distribution is given by (25), the Kullback information (5) satisfies*

$$H(w) \geqq \sum_{k=1}^{N} \left\{ \sum_{j=1}^{K} a_j p_j^k - \sum_{j=1}^{K_0} a_j^* p_j^{*k} \right\}^2 . \tag{33}$$

Recall that a constant factor in the Kullback divergence does not affect the largest pole of the zeta function. Therefore, the generalization error is determined by the zeta function of the expression in the right-hand side, which no longer contains the term x.

Proof: Using Lemma 1, we can express the Kullback divergence of Eq. (5) as

$$H(w) \geqq \sum_{k=0}^{N} \left[p(x = k|w) - q(x = k) \right]^2 . \tag{34}$$

According to a property of $\geqq$, it holds that

$$H(w) = f(w)^2 + \{ f(w) + cg(w) \}^2 \tag{35}$$

$$\geqq f(w)^2 + g(w)^2, \tag{36}$$

where c is a constant. Starting with the case $k = N$, we can recursively apply this relation;

$$H(w) \geqq \sum_{k=0}^{N} \left\{ \sum_{j=1}^{K} a_j (1-p_j)^{N-k} p_j^k - \sum_{j=1}^{K_0} a_j^* (1-p_j^*)^{N-k} p_j^{*k} \right\}^2 \tag{37}$$

$$= \sum_{k=0}^{N-1} \left\{ \sum_{j=1}^{K} a_j (1-p_j)^{N-k} p_j^k - \sum_{j=1}^{K_0} a_j^* (1-p_j^*)^{N-k} p_j^{*k} \right\}^2 + \left\{ \sum_{j=1}^{K} a_j p_j^N - \sum_{j=1}^{K_0} a_j^* p_j^{*N} \right\}^2 \tag{38}$$

$$\geqq \sum_{k=0}^{N-2} \left\{ \sum_{j=1}^{K} a_j (1-p_j)^{N-k} p_j^k - \sum_{j=1}^{K_0} a_j^* (1-p_j^*)^{N-k} p_j^{*k} \right\}^2 + \left\{ \sum_{j=1}^{K} a_j p_j^{N-1} - \sum_{j=1}^{K_0} a_j^* p_j^{*N-1} \right\}^2 + \left\{ \sum_{j=1}^{K} a_j p_j^N - \sum_{j=1}^{K_0} a_j^* p_j^{*N} \right\}^2 \tag{39}$$

$$\geqq \sum_{k=1}^{N} \left\{ \sum_{j=1}^{K} a_j p_j^k - \sum_{j=1}^{K_0} a_j^* p_j^{*k} \right\}^2 . \tag{40}$$

(End of Proof)

4.3 Analysis of Generalization Error

We now apply the method with the Newton diagram to the mixture model. The following is the main theorem of the present paper.

Theorem 2 *Assume that the learning machine (19) and the true model (25) are $K+1$ and K component binomial mixtures, respectively. For a sufficiently large sample size n, the generalization error satisfies the following equation:*

$$G(n) = \left(K - \frac{1}{4} \right) \frac{1}{n} + o\left(\frac{1}{n \log n} \right). \tag{41}$$

Proof: Here, we present an outline of the proof. The full proof is given in the appendix. After changing the coordinates of the parameter space, we find that the faces consist of

$$(a_1 + a_2)^2, \tag{42}$$

$$\{a_j^2\} \qquad (j = 3, \cdots, K), \tag{43}$$

$$\{p_j^2\} \qquad (j = 2, \cdots, K+1). \tag{44}$$

This means that there exists a face described by

$$f_\gamma(w) = (a_1 + a_2)^2. \tag{45}$$

It is easy to find that

$$\{a_1 = -a_2\} \subset \left\{ \frac{\partial f_\gamma(w)}{\partial a_1} = \frac{\partial f_\gamma(w)}{\partial a_2} = 0 \right\} \not\subset \{a_1 a_2 = 0\}. \tag{46}$$

The term $(a_1 + a_2)^2$ makes the Kullback divergence degenerate, which implies that the Newton diagram is not applicable to the current condition. Therefore, we consider a transform to change the coordinates because the definition of the non-degenerate function depends on the coordinates.

We next propose a method by which to find the non-degenerate coordinates. Since the factor of the difficulty is the term $(a_1 + a_2)^2$, define a mapping

$$a_1' = a_1 + a_2. \tag{47}$$

Based on this mapping, the terms of faces will also change. Then, check whether the function is still degenerate in the new coordinates. The faces in the new coordinates consist of

$$(a_1 p_1 + a_1^* p_2)^2, \tag{48}$$

$$\{a_j^2\} \qquad (j = 3, \cdots, K), \tag{49}$$

$$\{p_j^2\} \qquad (j = 3, \cdots, K+1). \tag{50}$$

Now, the term $(a_1 p_1 + a_1^* p_2)^2$ makes the Kullback divergence degenerative. As before, define the mapping by

$$p_2' = a_1 p_1 + (a_2 - a_1 + a_1^*) p_2. \tag{51}$$

Finally, we obtain the faces constructed by

$$a_1^2 p_1^4, \tag{52}$$

$$\{a_j^2\} \qquad (j = 2, \cdots, K), \tag{53}$$

$$\{p_j^2\} \qquad (j = 2, \cdots, K+1). \tag{54}$$

It is easy to confirm that the function is non-degenerate with the coordinates. After subdividing the fan and finding the mapping A, we obtain the result.
(End of Proof)

When the Kullback divergence is non-degenerate, the terms in the faces are of the lowest order, i.e., the divergence is expressed as

$$H(w) \gtrless a_1^2 p_1^4 + \sum_{j=2}^{K} a_j^2 + \sum_{j=2}^{K+1} p_j^2 + \text{higher order terms}. \tag{55}$$

The method with the Newton diagram systematically prunes the unnecessary higher order terms for the resolution of singularities as we ignored the term $w_1^3 w_2^3$ of Eq. (12) in Section 3.

The following algorithm summarizes the method by which to find degenerate coordinates.

Algorithm 2 (Finding non-degenerate coordinates) _

1. *If the function $H(w)$ includes the term*

$$(w_1 + g_1(w\backslash w_1) + g_2(w\backslash w_1))^m \tag{56}$$

and the face, which makes the function degenerate, consists of

$$f_\gamma(w) = (w_1 + g_1(w\backslash w_1))^m, \tag{57}$$

where $m > 1$ is a natural number and g_1, g_2 are polynomials with respect to $w\backslash w_1$, then define a mapping

$$w_1' = w_1 + g_1(w\backslash w_1) + g_2(w\backslash w_1), \tag{58}$$

and repeat 1 with the new coordinates $\{w_1', w\backslash w_1\}$.

2. *Otherwise, end the algorithm.*

Note that Eq.(57) implies that g_2 includes only terms of order higher than g_1. If the condition in *1* is not satisfied, then the function could be still degenerate. In this case, we need another solution. Even though this constraint may seem strong, the binomial mixture model does not violate this constraint.

5 Discussion and Conclusions

First, let us point out the following two advantages of the Newton diagram to show the utility: (a) higher order terms in the diagram do not affect the generalization error and (b) the method by which to find the resolution of singularities is systematic. The former advantage means that the Newton diagram selects the essential terms, as mentioned in the previous section. The latter one requires a more detailed explanation. In the conventional (iterative) method, we repeatedly use a unit mapping formed by $w_1 = w_1' w_2$. For example, this unit mapping changes Eq. (12) into

$$H(w) = w_2^5 (w_1'^5 w_2 + w_1'^3 w_2^2 + w_1'^2 + w_2). \tag{59}$$

This procedure is referred to as blowing-up and can be interpreted as eliminating the common factor. Continuing this blowing-up, we obtained Eq. (18). As seen in this example, the

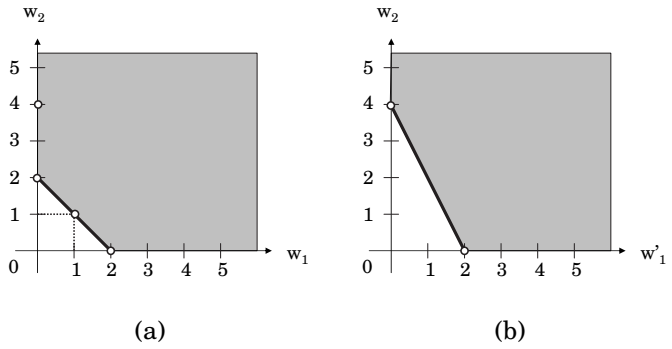

Figure 2: (a) $(w_1 + w_2)^2 + w_2^4$, (b) $w_1'^2 + w_2^4$

possibility of selecting the pair of parameters is not limited, and we need to search $w_2 = w_1 w_2'$ in to complete the coordinates. Furthermore, there is no method to evaluate the maximum time of iteration, although there is a guarantee that this algorithm stops (Hironaka, 1964). This means that the iterative method has an unknown computational cost in the sense of both the width and depth of the search tree. Compared to this difficulty, the proposed method has the limited cost of searching because each A determines local coordinates. Note that we can evaluate that the worst cost is at most k^d, where k and d are the number of vectors in subdivided fan and the dimension of parameters, respectively.

Second, we need to clarify the condition for which the proposed method can be applied. The only condition that need be considered is whether the function $H(w)$ is degenerate. A typical case to violate the condition is that the function includes the following term:

$$f_\gamma(w) \equiv (w_1 + w_2)^2 \tag{60}$$

as one of the faces. As in the proof of Theorem 2, the zero points of the partial differential with respect to w_1 or w_2 are not in the area $w_1 w_2 = 0$. More formally, we have

$$\{\frac{\partial f_\gamma}{\partial w_i} = 0\} \not\subset \{w_1 w_2 = 0\} \tag{61}$$

for $i = 1, 2$. This face is depicted in Fig.2(a). Intuitively, the function is non-degenerate when there is another point, which is not a vertex, on a face. According to Algorithm 2, new coordinates can be found by the affine mapping $w'1 = w_1 + w_2$ shown in Fig. 2 (b). Note that the point $(0, 4)$ constructs a face in new coordinates, though it was in the diagram before. The unnecessary terms are decided *after* the coordinates are fixed. Algorithm 2 cannot be applied to the case in which there is no affine mapping, i.e., the lowest order of the face is greater than

one. For example, if the face contains

$$f_\gamma(w) \equiv (w_1 w_2 + w_3 w_4)^2, \tag{62}$$

we cannot define any affine mapping to change into monomial. This type of term exists when the model has more redundant components $K > K_0 + 1$. Then, we can combine the proposed method with the iterative method. More precisely, blowing-up can be applied locally to such a term, and then Algorithm 2 will be available. In this case, we still ignore unnecessary higher order terms. It is an important and challenging future task to elucidate the types of unidentifiable models to which the proposed method can be applied.

A Formal Definitions of the Newton Diagram

Let the Taylor expansion of an analytic function $H(w)$ be

$$H(w) = \sum_v c_v w^v, \tag{63}$$

where $w = (w_1, \cdots, w_d) \in R^d$, $v = (v_1, \cdots, v_d) \in Q \subset Z^d$ and c_v is a constant. We use the notation that

$$w^v \equiv w_1^{v_1} w_2^{v_2} \cdots w_d^{v_d}. \tag{64}$$

Definition 2 (Newton diagram) *The convex hull of the subset*

$$\{v + v'; c_v \neq 0, v' \in R_+^d\} \tag{65}$$

is referred to as the Newton diagram $\Gamma_+(H)$.

For a given constant vector $a \in Z^d$, we define $l(a)$ by

$$l(a) \equiv \min\{\langle v, a\rangle; v \in \Gamma_+(H)\}, \tag{66}$$

where $\langle , \rangle$ is the inner product, $\langle v, a\rangle = \sum_{i=1}^d a_i v_i$.

Definition 3 (Face) *A face of* $\Gamma_+(H)$ *is defined by*

$$\gamma(a) \equiv \{v \in \Gamma_+(H); \langle v, a\rangle = l(a)\}. \tag{67}$$

Intuitively, the face is the border of the Newton diagram. Depending on a face γ, a polynomial f_γ is defined by

$$f_\gamma(w) \equiv \sum_{v \in \gamma} c_v w^v. \tag{68}$$

Definition 4 (Non-degenerate function) *The function $H(w)$ is said to be non-degenerate if and only if*

$$\left\{ w \in R^d; \frac{\partial f_\gamma}{\partial w_1}(w) = \cdots = \frac{\partial f_\gamma}{\partial w_d}(w) = 0 \right\} \subset \{ w_1 \cdots w_d = 0 \} \tag{69}$$

for an arbitrary compact face γ of $\Gamma_+(H)$. Otherwise, $H(w)$ is said to be degenerate.

Consider the dual space $P \subset Z^d$ of Q. The fan and the subdivided fan are defined on P.

Definition 5 (Fan) *A fan of Newton diagram Δ is a collection of convex polyhedral cones. More precisely, for a face γ, a cone is defined by*

$$\sigma_\gamma = \{ u \in P | \Gamma(u) \supset \gamma \}, \tag{70}$$

where

$$\Gamma(u) = \{ v \in \Gamma_+(H) | \langle v, u \rangle = \min_{v' \in \Gamma_+(H)} \langle v', u \rangle \}. \tag{71}$$

Then,

$$\Delta = \{ \sigma_\gamma | \gamma \text{ is a face of } \Gamma_+(H) \} \tag{72}$$

is a fan.

The subdivided fan can be defined in a similar manner. However, for simplicity, we omit the formal description. Intuitively, the construction is the same as the two-dimensional example (Section 3). Higher dimensional cases with formal definitions are written in (Fulton, 1993).

For a matrix

$$A = \begin{pmatrix} a_1^1 & \cdots & a_1^d \\ \vdots & \ddots & \vdots \\ a_d^1 & \cdots & a_d^d \end{pmatrix}, \tag{73}$$

and $w, u \in R^d$, we define

$$w = u^A \Leftrightarrow \begin{cases} w_1 = u_1^{a_1^1} \cdots u_d^{a_1^d} \\ \vdots \\ w_d = u_1^{a_d^1} \cdots u_d^{a_d^d} \end{cases}, \tag{74}$$

where a set, $a_1 = (a_1^1, \cdots, a_d^1), \cdots, a_d = (a_1^d, \cdots, a_d^d)$ is a part of the subdivided fan. Using the above definitions, the theorem that relates the Newton diagram to a resolution of singularities is known (Ewald, 1996), (Fulton, 1993).

Theorem 3 (Toric modification) *The map $\pi(u) : w = u^A, where \det A = \pm 1$, is called a real toric modification. The map $\pi^{-1}(U_0) \to W$ for the neighborhood $U_0 \subset R^d$ of the origin is a resolution of singularities if the function $H(w)$ is non-degenerate.*

B Proof of Theorem 2

According to Theorem 1, the Kullback information satisfies

$$H(w) \gtrless \sum_{k=1}^{N} \left\{ \sum_{j=1}^{K+1} a_j p_j^k - \sum_{j=1}^{K} a_j^* p_j^{*k} \right\}^2 . \tag{75}$$

We define the map $\Theta_1 : w \to w_1$, such that

$$a'_j \equiv a_j - a^*_{j-1} \qquad (j = 2, \cdots, K+1), \tag{76}$$
$$p'_j \equiv p_j - p^*_{j-1} \qquad (j = 2, \cdots, K+1), \tag{77}$$
$$a'_1 \equiv a_1, \tag{78}$$
$$p'_1 \equiv p_1 - p^*_1, \tag{79}$$

in order to shift the singularities to the origin. We write the variables a'_j and p'_j as a_j and p_j, respectively, in order to avoid the complicated notation. Then,

$$\begin{aligned} H(\Theta_1^{-1}(w_1)) \gtrless \mathcal{H}(w_1) \equiv \sum_{k=1}^{N} \Bigg\{ & c_2(k)(a_1 + a_2) + \sum_{j=3}^{K} c_j(k) a_j \\ & + c_1(k)(a_1 p_1 + (a_2 + a_1^*) p_2) \\ & + d_1(k)(a_1 p_1^2 + (a_2 + a_1^*) p_2^2) \\ & + \sum_{j=2}^{K+1} d_j(k) p_j + f_r(k, w_1) \Bigg\}^2 , \end{aligned} \tag{80}$$

where

$$c_2(k) = p_1^{*k} - p_K^{*k}, \tag{81}$$
$$c_j(k) = p_{j-1}^{*k} - p_K^{*k} \qquad (j = 3, \cdots, K), \tag{82}$$
$$c_1(k) = k p_1^{*k-1}, \tag{83}$$
$$d_1(k) = \binom{k}{2} p_1^{*k-2} \qquad (d_1(1) = 0), \tag{84}$$
$$d_j(k) = a^*_{j-1} k p_{j-1}^{*k-1} \qquad (j = 2, \cdots, K), \tag{85}$$
$$d_{K+1}(k) = k \left(1 - \sum_{j=1}^{K-1} a_j^* \right) p_K^{*k-1} \tag{86}$$

and $f_r(k, w)$ is the sum of the remaining terms in $\mathcal{H}(w_1)$.

Here, we show the vectors

$$v_1(k) = (c_1(k), c_2(k), \ldots, c_K(k), d_1(k), d_3(k), \ldots, d_{K+1}(k)) \tag{87}$$

for $k = 1, \ldots, 2K$ are linearly independent. Because of the condition $N \geq 2K$, the vectors are definable. It is sufficient to prove that the matrix

$$V_1 = \begin{bmatrix} v_1(1) \\ v_1(2) \\ \cdots \\ v_1(2K) \end{bmatrix} \tag{88}$$

satisfies that $\det V_1 \neq 0$. It is easy to confirm that

$$\det V_1 \propto \prod_{j=2}^{K} (p_1^* - p_j^*)^6 \prod_{1<i<j} (p_i^* - p_j^*)^4, \tag{89}$$

which proves $\det V_1 \neq 0$ since $p_1^* < p_2^* < \cdots < p_K^*$. Now, we define the vectors

$$v_2(k) = (c_2(k), \ldots, c_K(k), d_1(k), \ldots, d_{K+1}(k)) \tag{90}$$

for $k = 1, \ldots, 2K$. Similarly, we can define V_2 and show

$$\det V_2 \propto \prod_{j=2}^{K} (p_1^* - p_j^*)^6 \prod_{1<i<j} (p_i^* - p_j^*)^4. \tag{91}$$

So, these vectors are also linearly independent.

In order to find the faces, we rewrite the function as

$$\mathcal{H}(w_1) = \sum_{k=1}^{N} \{h_1(w_1, k)^2\} + 2 \sum_{k=1}^{N} \{h_1(w_1, k) h_2(w_1, k)\} + \sum_{k=1}^{N} \{h_2(w_1, k)^2\}, \tag{92}$$

where

$$h_1(w_1, k) = c_2(k)(a_1 + a_2) + \sum_{j=3}^{K} c_j(k) a_j + \sum_{j=2}^{K+1} d_j(k) p_j, \tag{93}$$

$$h_2(w_1, k) = c_1(k)(a_1 p_1 + (a_2 + a_1^*) p_2) + d_1(k)(a_1 p_1^2 + (a_2 + a_1^*) p_2^2) + f_r(k, w). \tag{94}$$

The terms in $h_2(w_1, k)$ are of higher order than those of $h_1(w_1, k)$ in the sense of w_1. Since the vectors

$$\{(c_2(k), \cdots, c_K(k), d_2(k), \cdots, d_{K+1}(k))\}_{k=1}^{N} \tag{95}$$

are linearly independent with respect to k, the following holds:

$$\sum_{k=1}^{N} \{h_1(w_1, k)^2\} \gtrless \sum_{k=1}^{N} c_2(k)^2 (a_1 + a_2)^2 + \sum_{k=1}^{N} \sum_{j=3}^{K} c_j(k)^2 a_j^2 + \sum_{k=1}^{N} \sum_{j=2}^{K+1} d_j(k)^2 p_j^2. \tag{96}$$

Then, the faces in the Newton diagram of $\mathcal{H}(w_1)$ consist of the following terms:

$$(a_1 + a_2)^2, \tag{97}$$

$$\{a_j^2\} \qquad (j = 3, \cdots, K), \tag{98}$$

$$\{p_j^2\} \qquad (j = 2, \cdots, K+1). \tag{99}$$

Because of the term $(a_1+a_2)^2$, the function $\mathcal{H}(w_1)$ is degenerate. We define the map $\Theta_2 : w_1 \to w_2$,

$$a_2' \equiv a_1 + a_2, \tag{100}$$

$$a_j' \equiv a_j \qquad (j = 1, 3, 4, \cdots, K), \tag{101}$$

$$p_j' \equiv p_j \qquad (j = 1, 2, \cdots, K+1). \tag{102}$$

For simplicity, we write the variables a_j' and p_j' as a_j and p_j, respectively. Then, $\mathcal{H}(w_1)$ is rewritten as

$$\mathcal{H}(\Theta_2^{-1}(w_2)) = \sum_{k=1}^{N} \{h_1^{'}(w_2, k)^2\} + 2\sum_{k=1}^{N} \{h_1'(w_2, k) h_2'(w_2, k)\} + \sum_{k=1}^{N} \{h_2^{'}(w_2, k)^2\}, \tag{103}$$

where

$$h_1'(w_2, k) = c_2(k)a_2 + \sum_{j=3}^{K} c_j(k)a_j + \sum_{j=3}^{K+1} d_j(k)p_j + c_1(k)(a_1 p_1 + (a_2 - a_1 + a_1^*)p_2), \tag{104}$$

$$h_2'(w_2, k) = d_1(k)(a_1 p_1^2 + (a_2 - a_1 + a_1^*)p_2^2) + f_r'(k, w_2), \tag{105}$$

and $f_r'(k, w_2)$ is the sum of the remaining terms in $\mathcal{H}(\Theta_2^{-1}(w_2))$. The terms in $h_2'(w_2, k)$ are of higher order than those of $h_1'(w_2, k)$ in the sense of w_2. According to the linear independency of the vectors

$$\{(c_1(k), \cdots, c_K(k), d_3(k), \cdots, d_{K+1}(k))\}_{k=1}^{N}, \tag{106}$$

the following holds:

$$\begin{aligned} \sum_{k=1}^{N} \{h_1^{'}(w_2, k)^2\} \geqslant & \sum_{k=1}^{N} c_2(k)^2 a_2^2 + \sum_{k=1}^{N}\sum_{j=3}^{K} c_j(k)^2 a_j^2 + \sum_{k=1}^{N}\sum_{j=3}^{K+1} d_j(k)^2 p_j^2 \\ & + \sum_{k=1}^{N} c_1(k)^2 (a_1 p_1 + (a_2 - a_1 + a_1^*)p_2)^2. \end{aligned} \tag{107}$$

Then, the faces in the Newton diagram of $\mathcal{H}(\Theta_2^{-1}(w_2))$ consist of the following terms:

$$(a_1 p_1 + a_1^* p_2)^2, \tag{108}$$

$$\{a_j^2\} \qquad (j = 2, \cdots, K), \tag{109}$$

$$\{p_j^2\} \qquad (j = 3, \cdots, K+1). \tag{110}$$

Because of the term $(a_1p_1+a_1^*p_2)^2$, the function $\mathcal{H}(\Theta_2^{-1}(w_2))$ is still degenerate. Thus, we define the map $\Theta_3 : w_2 \to w_3$,

$$a_j' \equiv a_j \qquad (j = 1, \cdots, K), \tag{111}$$

$$p_2' \equiv a_1p_1 + (a_2 - a_1 + a_1^*)p_2, \tag{112}$$

$$p_j' \equiv p_j \qquad (j = 1, 3, 4, \cdots, K+1). \tag{113}$$

Again, we write the variables a_j' and p_j' as a_j and p_j, respectively. Then, $\mathcal{H}(\Theta_2^{-1}(w_2))$ is rewritten as

$$\mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(w_3)) = \sum_{k=1}^{N}\{h_1^{''}(w_3)^2\} + 2\sum_{k=1}^{N}\{h_1''(w_3)h_2''(w_3)\} + \sum_{k=1}^{N}\{h_2^{''}(w_3)^2\}, \tag{114}$$

where

$$\begin{aligned} h_1''(w_3,k) =& c_2(k)a_2 + \sum_{j=3}^{K} c_j(k)a_j + \sum_{j=3}^{K+1} d_j(k)p_j + c_1(k)p_2 \\ &+ d_1(k)\left(a_1p_1^2 + \frac{(p_2 - a_1p_1)^2}{a_2 - a_1 + a_1^*}\right), \end{aligned} \tag{115}$$

$$h_2''(w_3,k) = f_r''(k,w_3), \tag{116}$$

and $f_r''(k,w_3)$ is the sum of the remaining terms in $\mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(w_3))$. The terms in $h_2''(w_3,k)$ are of higher order than those of $h_1''(w_3,k)$ in the sense of w_3. According to the linear independency of the vectors

$$\{(c_1(k), \cdots, c_K(k), d_1(k), d_3(k), \cdots, d_{K+1}(k))\}_{k=1}^{N}, \tag{117}$$

the following holds:

$$\begin{aligned} \sum_{k=1}^{N}\{h_1^{''}(w_3,k)^2\} \geqslant& \sum_{k=1}^{N} c_2(k)^2a_2^2 + \sum_{k=1}^{N}\sum_{j=3}^{K} c_j(k)^2a_j^2 + \sum_{k=1}^{N}\sum_{j=3}^{K+1} d_j(k)^2p_j^2 \\ &+ \sum_{k=1}^{N} c_1(k)^2p_2^2 + \sum_{k=1}^{N} d_1(k)^2\left(a_1p_1^2 + \frac{(p_2 - a_1p_1)^2}{a_2 - a_1 + a_1^*}\right)^2. \end{aligned} \tag{118}$$

Then, the faces in the Newton diagram of $\mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(w_3))$ consist of the following terms:

$$a_1^2p_1^4, \tag{119}$$

$$\{a_j^2\} \qquad (j = 2, \cdots, K), \tag{120}$$

$$\{p_j^2\} \qquad (j = 2, \cdots, K+1). \tag{121}$$

Finally, the function $\mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(w_3))$ is non-degenerate. Based on the subdivided fan, we can find the mapping A defined by

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 2 & 0 & \cdots & 1 \end{pmatrix}. \tag{122}$$

This matrix has a column vector $(0, 2, \cdots, 2, 1, 2, \cdots, 2)^T$ corresponding to the exponent part of $(a_1, a_2, \cdots, a_K, p_1, p_2, \cdots, p_{K+1})$ in the unit matrix. Thus, a resolution of singularities $g(u)$, where

$$u = (u_1, u_2, \cdots, u_K, v_1, v_2, \cdots, v_{K+1}) \tag{123}$$

is derived as

$$a_1 \equiv u_1, \tag{124}$$

$$a_j \equiv u_j v_1^2 \qquad (j = 2, \cdots, K), \tag{125}$$

$$p_1 \equiv v_1, \tag{126}$$

$$p_j \equiv v_j v_1^2 \qquad (j = 2, \cdots, K+1). \tag{127}$$

The Jacobian of $g(u)$ is

$$|g(u)| = v_1^{4K-2}, \tag{128}$$

and the common factor of $\mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(w_3))$ is v_1^4. Therefore, the largest pole λ' of

$$\int \mathcal{H}(\Theta_2^{-1}\Theta_3^{-1}(g(u)))^z |g(u)| du \tag{129}$$

is

$$\lambda' = K - \frac{1}{4}. \tag{130}$$

Since the Jacobians of maps $\Theta_1, \Theta_2, \Theta_3$ are not equal to zero, the largest pole of (6) is the same as λ', which is the coefficient of the generalization error.

References

Amari, S., & Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons. *IEICE Trans*, *E84-A 1*, 31–38.

Aoyagi, M., & Watanabe, S. (2004). The generalization error of reduced rank regression in bayesian estimation. *Proc. of ISITA* (pp. 1068–1073).

Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 924–933.

Aranson, A. B. (2001). Computation and applications of the newton polyhedrons. *Mathematics and Computers in Simulation*, *57*, 155–160.

Dacunha-Castelle, D., & Gassiat, E. (1997). Testing in locally conic models and application to mixture models. *Probability and Statistics*, *1*, 285–317.

Ewald, G. (1996). *Combinatorial convexity and algebraic geometry*, vol. 168 of *Graduate texts in mathematics*. Springer-Verlag.

Fulton, W. (1993). *Introduction to toric varieties*, vol. 131 of *Annals of Mathematics Studies*. Princeton University Press.

Glazko, G. B., Milanesi, L., & Rogozin, I. B. (1998). The subclass approach for mutational spectrum analysis: Application of the sem algorithm. *Journal of Theor. Biology*, *192*, 475–487.

Hagiwara, K. (2002). On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, *14*, 1979–2002.

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

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, *79*, 109–3262.

Levin, E., Tishby, N., & Solla, S. A. (1989). A statistical approach to learning and generalization in layered neural networks. *COLT '89: Proceedings of the second annual workshop on Computational learning theory* (pp. 245–260). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Nakajima, S., & Watanabe, S. (2007). Variational bayes solution of linear neural networks and its generalization performance. *Neural Comput.*, *19*, 1112–1153.

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

Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive bayesian networks. *The Journal of Machine Learning Research*, *6*, 1–35.

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

Watanabe, K., & Watanabe, S. (2006). Stochastic complexities of gaussian mixtures in variational bayesian approximation. *J. Mach. Learn. Res.*, *7*, 625–644.

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

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

Yamazaki, K., Nagata, K., Watanabe, S., & Müller, K.-R. (2006). A model selection method based on bound of learning coefficient. *Proc. of International Conference on Artificial Neural Networks* (pp. 371–380).

Yamazaki, K., & Watanabe, S. (2003a). Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks*, *16*, 1029–1038.

Yamazaki, K., & Watanabe, S. (2003b). Stochastic complexity of bayesian networks. *Proc. of UAI* (pp. 592–599).

Yamazaki, K., & Watanabe, S. (2005a). Algebraic geometry and stochastic complexity of hidden markov models. *Neurocomputing*, *69*, 62–84.

Yamazaki, K., & Watanabe, S. (2005b). Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities. *IEEE Trans. on Neural Networks*, *16*, 312–324.