Stochastic Complexity and Newton Diagram

Keisuke YAMAZAKI[†], Miki AOYAGI[‡] and Sumio WATANABE[†]

[†] Precision and Intelligence Laboratory,
Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku, Yokohama,
226-8503 Japan, Mailbox R2-5
E-mail: {k-yam, swatanab}@pi.titech.ac.jp

[‡] Department of Mathematics,
Sophia University,
E-mail: miki-a@sophia.ac.jp

Abstract

Many singular learning machines such as neural networks and mixture models are used in the information engineering field. In spite of their wide range applications, their mathematical foundation of analysis is not yet constructed because of the singularities in the parameter space. In recent years, we developed the algebraic geometrical method that shows the relation between the efficiency in Bayesian estimation and the singularities. In this paper, we propose a new mathematical method to analyze singular learning machines based on the Newton diagram and toric deformation. Using the proposed method, we obtain the exact value of the asymptotic stochastic complexity, which is a criterion of the model selection, in a mixture of binomial distributions.

1. Introduction

In the information engineering field, many kinds of learning machines such as neural networks, mixture models and Bayesian networks are being used. In spite of the wide-range applications and technical learning algorithms, their mathematical properties are not yet clarified.

All learning models belong to either category, *identifiable* or *non-identifiable*. A learning model is generally represented as a probability density function $p(x|w)$, where w is a parameter. When a machine learns from sample data, its parameter is optimized. Thus, the parameter determines the probability distribution of the model. If the mapping from the parameter w to $p(x|w)$ is one-to-one, the model is *identifiable*, otherwise, *non-identifiable*.

There are many difficulties in analyzing the non-identifiable model using the conventional method. If the learning model attains the true distribution, the parameter space contains the true parameter(s). In non-identifiable models, the set of true parameters is not one point but an analytic set in the parameter space. Because the set includes many singularities, the Fisher information matrices are not positive definite. Thus, the log likelihood cannot be approximated by any quadratic form of the parameters in the neighborhood of singularities [1], [11]. That is one of the reasons why the non-identifiable model cannot be clarified. We refer to this model as a singular model.

In Bayesian estimation, the stochastic complexity [6], which is equal to the free energy or the minus marginal likelihood, is very important. Using this observable, we can select the optimal size of the model and derive its generalization error. It is well known that the stochastic complexity is equivalent to BIC in identifiable (statistical regular) models [7]. However, it does not hold in singular models.

In recent years, we have proven that the singularities in the parameter space strongly relate to the efficiency of the Bayesian estimation based on algebraic geometry. This relation reveals that the stochastic complexity is determined by the zeta function of the Kullback information from the true distribution to the learning model and of an a priori distribution. The analysis of the stochastic complexity results in finding the largest pole of the zeta function. Using this method, we have clarified the upper bounds of the stochastic complexities in concrete models, such as multi-layered perceptrons, mixture models, Bayesian networks and hidden Markov models [12], [13], [14]. Though we are actually able to analyze these singular models, it is not easy to find the largest pole. To find the largest pole is equivalent to find a resolution of singularities of the Kullback information according to the algebraic geometrical method [8]. However, if the Kullback information satisfies a certain non-degenerate condition (Definition 3 in Section 3), we can systemat-

This work was partially supported by the Ministry of Education, Science, Sports, and Culture in Japan, Grant-in-aid for scientific research 15500130, 2004.

ically derive a desingularization based on the Newton diagram [2], [3]. The problem is that almost singular models do not satisfy the condition. It seemed impossible to apply the method of the Newton diagram to these models by choosing an appropriate variable. In this paper, we propose an algorithm to make the Kullback information satisfy the condition and apply the algorithm to a mixture of binomial distribution, and reveal the stochastic complexity. The method with the Newton diagram shows advantages to analyze singular learning machines.

2. Bayesian Learning, Stochastic Complexity and Algebraic Geometry

In this section, we introduce the standard framework of Bayesian estimation. They are well known in statistical learning theory.

Let $X^n = (X_1, X_2, \cdots, X_n)$ be a set of training samples that are independent and identical, where n is the number of training samples. These and the testing samples are taken from the true probability distribution $q(x)$. Let $p(x|w)$ be a learning model. The a priori probability distribution $\varphi(w)$ is given on the set of parameters W. The a posteriori probability distribution is defined by

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

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

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw.$$

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

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

where E_{X^n} stands for the expectation value over all sets of training samples. We define another normalizing constant $Z(X^n)$,

$$\begin{aligned} Z(X^n) &= \int \exp(-nH_n(w))\varphi(w)dw, \\ H_n(w) &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(X_i)}{p(X_i|w)}. \end{aligned}$$

The stochastic complexity is defined by

$$F(X^n) = -\log Z(X^n).$$

We can select the optimal model and hyperparameters by minimizing $-\log Z_0(X^n)$. This is equivalent to minimizing the stochastic complexity, since

$$\begin{aligned} -\log Z_0(X^n) &= -\log Z(X^n) + S(X^n), \\ S(X^n) &= -\sum_{i=1}^{n}\log q(X_i), \end{aligned}$$

where the empirical entropy $S(X^n)$ is independent of the learners. The average stochastic complexity $F(n)$ is defined by

$$F(n) = -E_{X^n}\Big[\log Z(X^n)\Big]. \quad (1)$$

The relation is well known [5], [10],

$$G(n) = F(n+1) - F(n).$$

Thus, it is very important to clarify $F(n)$.

We define the Kullback information from the true distribution $q(x)$ to the learner $p(x|w)$ by

$$H(w) = \int q(x)\log\frac{q(x)}{p(x|w)}dx, \quad (2)$$

and the zeta function by

$$J(z) = \int H(w)^z\varphi(w)dw. \quad (3)$$

It is known that this function has real, negative and rational poles. Using a resolution of singularities $g(\cdot)$ in the algebraic geometrical method [8], [9], we can represent $H(w)$ as

$$H(g(u)) = u_1^{a_1}u_2^{a_2}\cdots u_d^{a_d}. \quad (4)$$

Then, we can find the largest pole $-\lambda$ and the order m by integrating $J(z)$. The asymptotic expansion of the stochastic complexity is described as

$$F(n) = \lambda\log n - (m-1)\log\log n + o(1).$$

The generalization error can be rewritten as

$$G(n) = \lambda/n + o(1/n).$$

3. Newton Diagram and Resolution of Singularities

In this section, we introduce the Newton diagram and the relation to a resolution of singularities.

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

$$H(w) = \sum_{v} c_v w^v,$$

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

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

Definition 1 *The convex hull of the subset*

$$\{v + v'; c_v \neq 0, v' \in R^d_+\}$$

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

Let us define a constant vector $a \in Z^d$ and

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

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

Definition 2 *A face of* $\Gamma_+(H)$ *is defined by*

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

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

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

Definition 3 *The function* $H(w)$ *is said to be non-degenerate if and only if*

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

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

Consider the dual space $P \subset Z^d$ of Q.

Definition 4 *The orthogonal vectors to faces are called a fan.*

Consider the parallelogram constituted by arbitrary two vectors. If it includes a point of P, add the vector from the origin to the point.

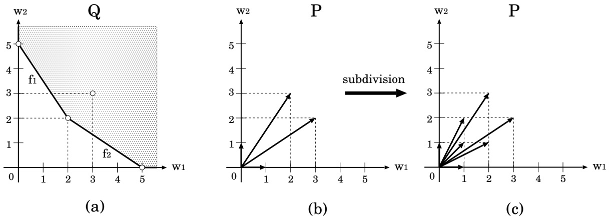
Definition 5 *Subdivision of* $\Gamma_+(H)$ *is defined by adding the vectors until there is no point in the parallelograms.*

For a matrix

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

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

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



(a) (b) (c)

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

where a set of $a_1 = (a_1^1, \cdots, a_d^1), \cdots, a_d = (a_1^d, \cdots, a_d^d)$ is a partof the subdivided fan. By using the above definitions, Theorem 1 is known which relates the Newton diagram to a resolution of singularities [3].

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

Theorem 1 claims that a resolution of singularities is determined by a set of all $\pi(u)$. In other words, the space W consists of the union of local coordinatates constituted by each A.

Example 1 *Assume* $H(w)$ *is defined by*

$$H(w) = w_1^5 + w_1^3 w_2^3 + w_1^2 w_2^2 + w_2^5,$$

where $w = (w_1, w_2)$, $v = (5, 0), (3, 3), (2, 2), (0, 5)$ *and* $c_v = 1$. *Then, the Newton diagram is depicted by the shaded area (Fig. 1 (a)) and it has four faces,* f_1, f_2, $[(0, 5), (0, \infty)]$, $[(5, 0), (\infty, 0)]$. *(Figure 1 (a)). It is easy to show the function* $H(w)$ *is non-degenerate. Then, the fan and the subdivided one of the Newton diagram are respectively depicted by Fig. 1 (b) and (c). We can select two vectors,* $(3, 2), (1, 1)$ *from the subdivided fan. The map* $g(u_1, u_2)$ *is defined by*

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

This gives an intended expression (4) of $H(w)$,

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

The above example shows that a resolution map is found which makes the function $H(w)$ rewritten as the expression (4) based on the Newton diagram. In order to find the largest pole of the zeta function determined by the equation (3), the problem comes down to find the efficient vectors 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_j and $H(g(u))$ has the common factor u_j^β, the vector settles a pole $-\alpha/\beta$, where $\alpha = \sum_i a_i^j$.

4. Main Results

In this section, let us introduce the algorithm and apply it to a mixture of binomial distributions.

4.1. Proposed Algorithm

We introduce an algorithm to change a degenerate Kullback information into non-degenerate. At first, we assume the following condition,

(A1) Assume that the Kullback information (2) is degenerate because the Newton diagram has a face such that

$$(\zeta + h(w))^2,$$

where ζ is one of parameters w and $h(w)$ is a polynomial of w.

Let $H(w)$ include the term,

$$(\zeta + h(w) + h'(w))^2,$$

where $h'(w)$ is a higher order polynomial of w than $h(w)$. This means the terms of $h'(w)$ are inside of the Newton diagram since $\zeta, h(w)$ constitute the face.

Next, the following is the proposed algorithm of this paper.

Algorithm 1 *(Step 1) Define the map* $\Pi : w \to w'$,

$$\zeta' \equiv \zeta + h(w) + h'(w).$$

(Step 2) According to the map Π, *redraw the Newton diagram of* $H(\Pi^{-1}(w'))$.

(Step 3) If $H(\Pi^{-1}(w'))$ *is degenerate and satisfies the condition (A1), regard* w' *as* w *and go to (Step 1). Otherwise, return* $H(\Pi^{-1}(w'))$.

Using this algorithm, we can reveal the stochastic complexity of some learning machines.

4.2. Mixture of Binomial Distributions

A mixture of binomial distributions is defined by

$$p(x = k|w) = \left(\begin{array}{c} N \\ k \end{array} \right) \left\{ \sum_{i=1}^{K+1} a_i p_i^k (1 - p_i)^{N-k} \right\}, \quad (5)$$

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

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

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

$$a_{K+1} = 1 - \sum_{i=1}^K a_i.$$

A binomial distribution is defined by

$$\left(\begin{array}{c} N \\ k \end{array} \right) \bar{p}^k (1 - \bar{p})^{N-k},$$

where $1 < \bar{p} \le 1/2$. Thus, the mixture (5) has $K + 1$ components.

This learning machine is used for the gene analysis and the mutational spectrum analysis [4].

4.3. Application of Algorithm

Assume that a learning machine consists of two components, and the true distribution consists of one component,

$$\begin{aligned} & p(x = k|w) \\ & \quad = \binom{N}{k} \left\{ a p_1^k (1 - p_1)^{N-k} + (1 - a) p_2^k (1 - p_2)^{N-k} \right\} \quad (6) \end{aligned}$$

$$q(x = k) \quad = \quad \left(\begin{array}{c} N \\ k \end{array} \right) p^{*k} (1 - p^*)^{N-k}, \quad (7)$$

where $0 < p^* \le 1/2, 0 < a^* < 1$ are constants. Based on the proposed algorithm and the method of the Newton diagram, we prove the following theorem.

Theorem 2 *If the learning machine is given by the equation (6) and the true distribution is given by the equation (7), then, for a sufficiently large natural number* n, *the stochastic complexity satisfies the equation,*

$$F(n) = \frac{3}{4} \log n + C,$$

where C *is a constant independent of* n.

(Proof of Theorem 2) By using some analytic evaluations, it follows that

$$C_1 \mathcal{H}(w_1) \le H(\Theta_1^{-1}(w_1)) \le C_2 \mathcal{H}(w_1),$$

where C_1, C_2 are positive constants independent of w_1, the map $\Theta_1 : w \to w_1$,

$$\begin{aligned} a &\equiv a', \\ p_1' &\equiv p_1 - p^*, \\ p_2' &\equiv p_2 - p^*, \end{aligned}$$

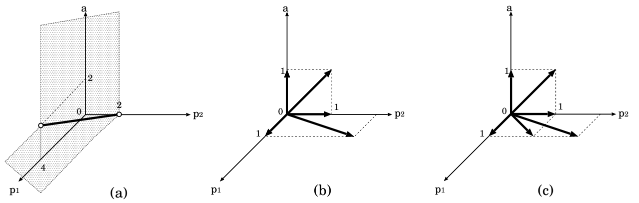


Figure 2: (a) the Newton diagram, (b) the fans, (c) the subdivided fans

and

$$\begin{aligned}\mathcal{H}(w_1) \quad \equiv \quad & (ap_1 + (1-a)p_2)^2 \\ & +(ap_1^2 + (1-a)p_2^2)^2 + (\text{higher order terms}).\end{aligned}$$

We write the variables, a' and p'_j as a and p_j respectively to avoid the complicated notation. It is known the largest pole in (3) is equal to that in the zeta function of $\mathcal{H}(w_1)$. In $\mathcal{H}(w_1)$, the faces of the Newton diagram consist of the term,

$$(ap_1 + p_2)^2,$$

Because of this, the function $\mathcal{H}(w_1)$ is degenerate. Regarding $p_2, ap_1, -ap_2$ as $\zeta, h(w), h'(w)$ in Algorithm 1, respectively, we apply it to this model. We define the map $\Pi : w_1 \to w'_1$ as

$$p'_2 \equiv ap_1 + (1-a)p_2.$$

The term $(ap_1 + (1-a)p_2)^2$ is rewritten as p_2^2 (We abbreviate $'$). Then, $\mathcal{H}(\Pi^{-1}(w'_1))$ is rewritten as

$$\begin{aligned}\mathcal{H}(\Pi^{-1}(w'_1)) \quad \equiv \quad & p_2^2 + (ap_1^2 + (1-a)\frac{(p_2 - ap_1)^2}{1-a})^2 \\ & +(\text{higher order terms}).\end{aligned}$$

At this time, the faces consist of the terms,

$$\begin{gathered} p_2^2, \\ a^2 p_1^4 \end{gathered}$$

The function is non-degenerate. The Newton diagram, the fans and the subdivided fans are depicted by Fig. 2 (a)-(c). We can select a set of vectors in the subdivided fan, $(0,1,2), (1,0,0), (1,0,1)$. According to them, a resolution of singularities $g(u)$, where

$$u = (u_1, v_1, v_2)$$

is derived as

$$\begin{aligned} a &\equiv v_1 v_2, \\ p_1 &\equiv u_1, \\ p_2 &\equiv u_1^2 v_2. \end{aligned}$$

The Jacobian of $g(u)$ is

$$|g(u)| = |u_1|^2 |v_2|,$$

and

$$\mathcal{H}(\Pi^{-1}(g(u))) = u_1^4 u_3^2 (c_1 + c_2 v_1^2 v_2^2 + \text{higher order terms}),$$

where c_1, c_2 are positive constants. Therefore, a pole in $\mathcal{H}(w_1)$ is $\lambda = 3/4$. As a matter of fact, it is the largest one by comparing all poles determined by other sets of vectors in the same way. **(End of Proof)**

5. Discussion & Conclusion

First, let us summarize two advantages of the method to use the Newton diagram. (a) *We are able to ignore higher-order terms which are inside the Newton diagram.* (b) *A resolution of singularities is systematically found by using the fans.* In the example in Section 3, we focused not the whole terms but particular terms in $H(w)$ to find a resolution of singularities. For instance, the term $w_1^3 w_2^3$ in Example 1 does not affect the resolution of singularities. This means we need to consider only the terms in the faces of the Kullback information. In general, it is not easy to derive a resolution of singularities from all terms in the Kullback information. Moreover, the number of resolutions exponentially grows when the dimension of the parameter increases. This is one of the reasons to use a partial resolution (blow-up) in the previous studies [12], [13], [14]. Based on these advantages, we revealed not an upper bound but the exact value of λ, the coefficient of the leading term in the stochastic complexity. It is expected the method with the Newton diagram can be applied to other singular models. For example, we obtained an upper bound of general mixture models [12], [13]. If the components are binomial distributions, the upper bound is $\lambda \leq 1$. Our result shows that it can be tighter on some conditions.

Second, let us talk about the degenerate Kullback information. As above stated, the method with the Newton diagram is useful to analyze learning machines with singularities. However, it assumes that the Kullback information is non-degenerate. In fact, most Kullback informations of the non-regular models are degenerate. It depends on the coordinates of the parameter whether a function is degenerate or not. In this paper, we proposed the algorithm that makes the Kullback information non-degenerate. We defined the map Π that changes some terms in the faces into a square of one variable since they make the function degenerate. The algorithm stops when the Kullback information is non-degenerate or a face does not satisfy the condition (A1).

The problem is the latter case since the Kullback information is still degenerate and we cannot apply the Newton diagram method. Then, the face includes a square of a quadratic form, such that

$$(\zeta_1\zeta_2 + \zeta_3\zeta_4 + \cdots)^2.$$

It is impossible to change this expression into a square of one variable such that $\zeta_1^{'2}$ with one-to-one mapping. This means that there are two kinds of degenerate functions, the case holding (A1) and the above case. Which function class can satisfy (A1) is still open to discuss, and it would construct a new algorithm to apply the method of the Newton diagram without (A1). They are our future studies.

Last, let us discuss the model selection problem. In statistical regular models, the stochastic complexity is equivalent to the well known BIC [7]. However, it is not in non-regular models. The dimension of the parameter w is $d = 3$. So, the coefficient λ is smaller than $d/2$. Furthermore, by using our result, the conventional approximations to calculate the stochastic complexity such as Markov Chain Monte Carlo and Variational Bayes methods can be evaluated.

References

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

[2] A. B. Aranson, "Computation and applications of the Newton polyhedrons," *Math. and Computers in Simulation*, 57, pp. 155-160, 2001.

[3] W. Fulton, "Introduction to toric varieties," Annals of Mathematics Studies, Vol. 131, Princeton University Press, 1993.

[4] G. B. Glazko, L. Milanesi, and I. B. Rogozin, "The subclass approach for mutational spectrum analysis: Application of the SEM algorithm," *Journal of Theor. Biology*, 192, pp. 475-487, 1998.

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

[6] J. Rissanen, "Stochastic complexity and modeling," *Annals of Statistics*, 14, pp. 1080-1100, 1986.

[7] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 6 (2), pp. 461-464, 1978.

[8] S. Watanabe, "Algebraic analysis for non-identifiable learning machines," *Neural Computation*, 13 (4), pp. 899-933, 2001.

[9] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," *Neural Networks*, 14 (8), pp. 1049-1060, 2001.

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

[11] K. Yamazaki and S. Watanabe,"A probabilistic algorithm to calculate the learning curves of hierarchical learning machines with singularities," *Trans. on IEICE*, J85-D-2(3), pp. 363-372, 2002.

[12] K. Yamazaki and S. Watanabe,"Singularities in mixture models and upper bounds of stochastic complexity," *International Journal of Neural Networks*, 16(2003), pp.1029-1038, 2003.

[13] K. Yamazaki and S. Watanabe, "Stochastic complexity of Bayesian networks," in *Proceedings of 19th Conference on Uncertainty in Artificial Intelligence*, pp. 592-599.

[14] K. Yamazaki and S. Watanabe, "Stochastic complexities of hidden Markov models," in *Proceedings of 13th Conference on Neural Networks for Signal Processing*, pp. 179-188.

[15] K. Yamazaki, M. Aoyagi and S. Watanabe, "Algorithm with Newton diagram for analyzing stochastic complexity," in *Technical report of IEICE*, to appear.