2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Weighted Blowups of Kullback Information and Application to Multinomial Distributions

Takeshi Matsuda† and Sumio Watanabe‡

†Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology,
Mailbox R2-5, 4259 Nagatsuta-chou, Midori-Ku, Yokohama, Kanagawa, 226-8503,JAPAN
‡Precision and Intelligence Laboratory, Tokyo Institute of Technology,
Mailbox R2-5, 4259 Nagatsuta-chou, Midori-Ku, Yokohama, Kanagawa, 226-8503,JAPAN
Email: matsuken@cs.pi.titech.ac.jp, swatanabe@cs.pi.titech.ac.jp

Abstract—Singular learning machines such as mixture models, neural networks and Bayesian networks are used in many fields of information engineering. However, they are not subject to the conventional statistical theory of regular statistical models, because their Fisher information matrices are degenerate. Recently, the generalization performance of singular learning machines was clarified based on resolution of singularities. In this paper, we propose a new method to compute learning coefficients using weighted blowups and show its effectiveness by application to the mixture of multinomial distributions.

1. Introduction

Learning machines such as layered neural networks, mixture models, hidden Markov models, and Bayes networks are being used in many fields of information engineering. However, mathematical analysis of these learning machines is generally not easy because their parameter spaces include singularities. Above learning machines are called singular learning machines, because their Fisher information matrices are not positive definite and they are not subject to the conventional statistical theory of regular statistical models. In [5], it was shown that the generalization performance of singular learning machines is determined by the zeta function of the Kullback-Leibler information from the true distribution $q(x)$ to a learning machine $p(x|\omega)$,

$$H(\omega) = \int q(x) \log \frac{q(x)}{p(x|\omega)} dx, \tag{1}$$

where $q(x)$ and $p(x|\omega)$ are probability density functions on n dimensional Euclidian space and ω is a d dimensional parameter. The zeta function of a learning machine is defined by

$$\zeta(z) = \int H(\omega)^z \varphi(\omega) d\omega, \tag{2}$$

where $\varphi(\omega)$ is an *a priori* probability density function contained in C_0^∞ class. It is well known that $\zeta(z)$ is a holomorphic function in $Re(z) > 0$ and that it can be analytically continued to the meromorphic function on the entire complex plane [3]. All poles of the zeta function are negative and rational numbers. Let $(-\lambda)$ be the largest pole of the zeta function $\zeta(z)$. Then the average Bayes generalization error is given by

$$G(N) = \frac{\lambda}{N} + o(\frac{1}{N}). \tag{3}$$

The constant λ determines the learning curve, hence it is called a learning coefficient. To evaluate how appropriate a learning machine is for a given set of training samples, the learning coefficient plays an important role, for example, it is the theoretical base for the optimal model selection and statistical hypothesis test.

The learning coefficients of a three-layer perceptron and a reduced rank regression were obtained by using recursive blowups [1], [2]. It was proposed that the toric modification is useful if Newton diagram is non-degenerate [6]. However, it is still difficult to find a complete resolution map in several singular learning machines. In this paper, we propose a new method to find the desingularization of learning machines using weighted blowups, and show its effectiveness by application to the mixture of multinomial distributions.

2. Bayes Learning

In this section, we overview the statistical framework of Bayes learning and the asymptotic theory of the generalization error. Assume that training samples $X^N = \{X_1, X_2, ..., X_N\}$ are independently taken from the true distribution $q(x)$ on $\mathbf{R}^n$.

1. Prepare a learning machine $p(x|\omega)$ which is defined by a conditional probability density function of $x \in \mathbf{R}^n$ for a given parameter $\omega \in \mathbf{R}^d$ and an *a priori* distribution $\varphi(\omega)$.

2. Define a Bayes *a posteriori* distribution

$$p(\omega|X^N) = \frac{1}{Z(X^N)} \varphi(\omega) \prod_{i=1}^{N} p(X_i|\omega),$$

where

$$Z(X^N) = \int d\omega\varphi(\omega)\prod_{i=1}^{N} p(x_i|\omega)$$

is the normalizing constant.

3. Bayes predictive distribution is defined by

$$p(x|X^N) = \int p(x|\omega)p(\omega|X^N)d\omega.$$

The average Bayes generalization error is defined by the average Kullback-Leibler information from the true distribution to the Bayes predictive distribution,

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

As eq.(3), if N is sufficiently large, then $G(N)$ is equal to λ/N where λ is the learning coefficient, where $(-\lambda)$ is the largest pole of the zeta function, eq.(2). Hironaka's resolution theorem ensures that there exists a set of a manifold U and a proper analytic map $g : U \to \mathbf{R}^d$ such that the Kullback-Leibler information in eq.(1) can be made to be normal crossing, in other words,

$$H(g(u)) = u_1^{2k_1}u_2^{2k_2}\cdots u_d^{2k_d}$$

on every local coordinate of U, where $k_1, k_2, ..., k_d$ are nonnegative integers. Once we find the manifold U and a resolution map g, then we can easily find the largest pole of the zeta function because

$$\zeta(z) = \int H(g(u))^z\varphi(g(u))|g'(u)|du$$

where $|g'(u)|$ is the Jacobian determinant of the map g. Therefore, if we find a resolution map, then we obtain the learning coefficient.

3. Proposed Method

In general, it is not easy to find desingularization for a given learning machine. In this paper we propose a new method of employing the weighted blowup [4] to find a resolution map and show the effectiveness by an application to the mixture of multinomial distributions.

Firstly, let us consider the monomial $\omega^a = \omega_1^{a_1}\omega_2^{a_2}\cdots\omega_d^{a_d}$. Let $s = (s_1, s_2, \cdots, s_d)$ be a set of nonnegative integers. For a given set s and ω^a, we define the weighted degree v_{ω^a} with the weight s by

$$v_{\omega^a} = a_1s_1 + \cdots + a_ds_d.$$

Example 1 *For the monomial $\omega_1^2\omega_2^3$ and $s = (2, 1)$, the weighted degree $v_{\omega_1^2\omega_2^3}$ is*

$$v_{\omega_1^2\omega_2^3} = 2\cdot 2 + 3\cdot 1 = 7.$$

Secondly, a polynomial is said to be quasi-homogeneous if it is expressed by a linear combination of monomials which have the same weighted degree with some weight.

Example 2 *We give an example of a quasi-homogeneous polynomial. Let $H = \omega_1^2 + \omega_2^3 + \omega_3^5$ and $s = (15, 10, 6)$. Then*

$$\begin{aligned} v_{\omega_1^2} &= 2\cdot 15 + 0\cdot 10 + 0\cdot 6 = 30,\\ v_{\omega_2^3} &= 0\cdot 15 + 3\cdot 10 + 0\cdot 6 = 30,\\ v_{\omega_3^5} &= 0\cdot 15 + 0\cdot 10 + 5\cdot 6 = 30. \end{aligned}$$

Therefore, H is a quasi-homogeneous polynomial of weighted degree 30 with weight s.

Weighted Blowup. Weighted blowup is defined as follows. Let H be an analytic function

$$H = H_l + H_{l+1} + H_{l+2} + \cdots,$$

where $H_l \neq 0$ is a quasi-homogeneous polynomial of weighted degree l with weight $s = (s_1, \cdots, s_d)$. Let M be an open set of d-dimensional Euclidian space and $(\omega_1, \cdots, \omega_d)$ be a coordinate of M. Also let N_i be an open subset of d-dimensional Euclidian space and $(\omega_{1i}, \cdots, \omega_{di})$ be a coordinate of N_i $(i = 1, 2, .., d)$. We define the mapping $g_i : N_i \to M$ as follows:

$$\begin{cases} \omega_1 = \omega_{ii}^{\frac{s_1}{s_i}}\omega_{1i},\\ \omega_2 = \omega_{ii}^{\frac{s_2}{s_i}}\omega_{2i},\\ \cdots\\ \omega_i = \omega_{ii},\\ \cdots\\ \omega_d = \omega_{ii}^{\frac{s_d}{s_i}}\omega_{di}. \end{cases} \tag{4}$$

Then $N = N_1 \cup \cdots \cup N_d$ is a nonsingular manifold by glueing of coordinates using eq.(4) for $i = 1, \cdots, d$. The set $\{N, g_i\}$ is called the weighted blowup with weight $s = (s_1, \cdots, s_d)$.

Example 3 *Let $H(\omega) = \omega_1^2 + \omega_2^3 + \omega_3^5 + \omega_1^2\omega_2$. Then*

$$H(\omega) = H_{30} + H_{40},$$

where $H_{30} = \omega_1^2 + \omega_2^3 + \omega_3^5$ and $H_{40} = \omega_1^2\omega_2$.

$$\begin{aligned} H(g_1(\omega_1)) &= \omega_{11}^{30}(1 + \omega_{12}^3 + \omega_{13}^5 + \epsilon_1)\\ H(g_2(\omega_2)) &= \omega_{22}^{30}(\omega_{21}^2 + 1 + \omega_{23}^5 + \epsilon_2)\\ H(g_3(\omega_3)) &= \omega_{33}^{30}(\omega_{31}^2 + \omega_{32}^3 + 1 + \epsilon_3) \end{aligned}$$

where ϵ_i $(i = 1, 2, 3)$ is the monomials of smaller order, which gives resolution of singularities of H.

If a resolution of singularities of the Kullback-Leibler information is found by weighted blow-up, then we can calculate the learning coefficient.

Example 4 *Let H be the same function as example 3. The largest pole of $\zeta(z) = \int H(\omega)^z \varphi(\omega) d\omega$ can be obtained by*

$$\zeta(z) = \int H(g_1(\omega_1))^z \omega_{11}^{30} d\omega$$

$$+ \int H(g_2(\omega_2))^z \omega_{22}^{30} d\omega$$

$$+ \int H(g_3(\omega_3))^z \omega_{33}^{30} d\omega,$$

resulting that the learning coefficient is $\frac{31}{30}$.

4. Application to Mixture of Multinomial Distributions

Mixtures of multinomial distributions are used in information systems in a clustering problem of medical data and natural language processing. In this section, we show the effectiveness of the proposed method by an application to a mixture of multinomial distributions. A multinomial distribution, in which m trials are judged into n categories, is defined by a probability distribution on the set $\{(x_1, x_2, .., x_n); \sum_{j=1}^n x_j = m\}$,

$$\frac{m!}{x_1!x_2! \cdots x_n!} \prod_{k=1}^n p_k^{x_k}, \qquad (5)$$

where $p_1 + \cdots + p_n = 1$. A mixture of l multinomial distributions is defined by

$$p(x_1, \cdots, x_n|\omega) = \frac{m!}{x_1!x_2! \cdots x_n!} \Big\{ \sum_{i=1}^l a_i \prod_{k=1}^n p_{ik}^{x_k} \Big\}, \quad (6)$$

where l is a natural number, $p_{i1} + \cdots + p_{in} = 1$ $(i = 1, 2, ..., l)$, and ω is a parameter defined by

$$\omega = (\{a_1, a_2, \cdots, a_l\}, \{p_{i1}, p_{i2}, \cdots, p_{in}\}_{i=1}^l)$$

which satisfies

$$\sum_{j=1}^l a_j = 1, \qquad a_j \geq 0,$$

$$\sum_{i=1}^n p_{ji} = 1, \qquad p_{ji} \geq 0.$$

A mixture of multinomial distributions is not a regular statistical model but a singular one. We derive the learning coefficient of a mixture of trinomial distributions made of two components by applying a weighted blowup.

Theorem 1 *Assume that a learning machine made of two components and the true distribution made of one component are represented respectively by*

$$\begin{aligned} p(x_1, x_2|\omega) &= \frac{m!}{x_1!x_2!x_3!} \Big\{ a p_1^{x_1} p_2^{x_2} p_{12}^{x_3} \\ &\quad + (1-a) p_3^{x_1} p_4^{x_2} p_{34}^{x_3} \Big\}, \\ q(x_1, x_2|\omega) &= \frac{m!}{x_1!x_2!x_3!} q_1^{x_1} q_2^{x_2} q_{12}^{x_3}, \end{aligned}$$

where $p_{12} = 1 - p_1 - p_2$, $p_{34} = 1 - p_3 - p_4$, $q_{12} = 1 - q_1 - q_2$, and $x_3 = m - x_1 - x_2$. Then the average Bayes generalization error is given by

$$G(N) = \frac{3}{2N} + o(\frac{1}{N}).$$

where N is the number of training samples.

Proof 1 *The Kullback information is given by*

$$H(\omega) = \sum_{x_1+x_2=0}^{x_1+x_2=m} q(x_1, x_2|\omega) \log \frac{q(x_1, x_2|\omega)}{p(x_1, x_2|\omega)}. \qquad (7)$$

Put $b = p_1 - q_1$, $c = p_2 - q_2$, $d = p_3 - q_1$ and $e = p_4 - q_2$, where $-1 < b < 1$, $-1 < c < 1$, $-1 < d < 1$ and $-1 < e < 1$. By using ideal theory, we can prove that $H(\omega)$ is analytically equivalent to the following polynomial,

$$\begin{aligned} H(\omega) &= (ab + (1-a)d)^2 + (ac + (1-a)e)^2 \\ &\quad + (abc + (1-a)de)^2 + (ab^2 + (1-a)d^2)^2 \\ &\quad + (ac^2 + (1-a)e^2)^2. \end{aligned}$$

We can assume that $a \neq 0$ or $a \neq 1$ hence Jacobian determinant of the transform $d_1 = ab + (1-a)d$ and $e_1 = ac + (1-a)e$ is not equal to zero. Therefore the Kullback information is equivalent to

$$\begin{aligned} H(\omega') &= d_1^2 + e_1^2 + (ab^2 + \frac{(d_1 - ab)^2}{1-a})^2 \\ &\quad + (ac^2 + \frac{(e_1 - ac)^2}{1-a})^2 \\ &\quad + (abc + \frac{(d_1 - ab)(e_1 - ac)}{1-a})^2. \end{aligned}$$

Let us apply the proposed method. The weight $(1, 1, 1, 3, 3)$ is associated with variables (a, b, c, d_1, e_1), by which the weighted blowup is defined by

$$\begin{cases} a = a_1 \\ b = a_1 b_1 \\ c = a_1 c_1 \\ d_1 = a_1^3 d_{11} \\ e_1 = a_1^3 e_{11} \end{cases} \qquad (8)$$

$$\begin{cases} a = b_2 a_2 \\ b = b_2 \\ c = b_2 c_2 \\ d_1 = b_2^3 d_{12} \\ e_1 = b_2^3 e_{12} \end{cases} \qquad (9)$$

$$\begin{cases} a = c_3 a_3 \\ b = c_3 b_3 \\ c = c_3 \\ d_1 = c_3^3 d_{13} \\ e_1 = c_3^3 e_{13} \end{cases} \quad (10)$$

$$\begin{cases} a = d_{14} a_4 \\ b = d_{14} b_4 \\ c = d_{14} c_4 \\ d_1 = d_{14}^3 \\ e_1 = d_{14}^3 e_{14} \end{cases} \quad (11)$$

$$\begin{cases} a = e_{15} a_5 \\ b = e_{15} b_5 \\ c = e_{15} c_5 \\ d_1 = e_{15}^3 d_{15} \\ e_1 = e_{15}^3 \end{cases} \quad (12)$$

From the transformation eq.(8), we have the local zeta function

$$\int a_1^{6z+8}\{d_{11}^2 + e_{11}^2 + (b_1^2 + a_1\frac{(a_1^2 d_{11} - b_1)^2}{1-a_1})^2$$
$$+(c_1^2 + a_1\frac{(a_1^2 e_{11} - c_1)^2}{1-a_1})^2$$
$$+(b_1 c_1 + a_1\frac{(a_1^2 d_{11} - b_1)(a_1^2 e_{11} - c_1)}{1-a_1})^2\}^z d\omega'. \quad (13)$$

Hence, we get a pole $-\frac{3}{2}$*. By the symmetry, it is sufficient to study the transform eq.(9) and eq.(11). From the transform eq.(9), we obtain*

$$\int b_2^{6z+8}\{d_{12}^2 + e_{12}^2 + (a_2 + b_2\frac{(b_2 d_{12} - a_2)^2}{1-a_2 b_2})^2$$
$$+(a_2 c_2^2 + b_2\frac{(b_2 e_{12} - a_2 c_2)^2}{1-a_2 b_2})^2$$
$$+(a_2 c_2 + b_2\frac{(b_2 d_{12} - a_2)(b_2 e_{12} - a_2 c_2)}{1-a_2 b_2})^2\}^z d\omega'. \quad (14)$$

And from eq.(11)

$$\int d_{14}^{6z+8}\{1 + e_{14}^2 + (a_4 b_4^2 + d_{14}\frac{(d_{14} - a_4 b_4)^2}{1-a_4 d_{14}})^2$$
$$+(a_4 c_4^2 + d_{14}\frac{(d_{14} e_{14} - a_4 c_4)}{1-a_4 d_{14}})^2$$
$$+(a_4 b_4 c_4 + d_{14}\frac{(d_{14} - a_4 b_4)(d_{14} e_{14} - a_4 b_4)}{1-a_4 d_{14}})^2. \quad (15)$$

For eq.(13), the weighted blowup with weight $(1, 1, 2, 2)$ *is associated with the variables* $(b_1, c_1, d_{11}, e_{11})$*, resulting that the zeta function has pole* $-\frac{3}{2}$*. Using blow up at the origin for eq.(14), we get pole* $-\frac{3}{2}$*. Therefore, we obtain* $\lambda = \frac{3}{2}$*. Hence, the Bayes generalization error is given by*

$$G(N) = \frac{3}{2N} + o(\frac{1}{N}).$$

5. Conclusion

In this paper we proposed a new method to calculate the learning coefficient by using weighted blowup, and showed its effectiveness by an application to a mixture of multinomial distributions. It is a study for the future to extend the proposed method to general multinomial distributions.

Acknowledgments

This work was supported by the Ministry of Education, Science, Sports, and Culture in Japan, Grant-in-aid for scientific research 18079007.

References

[1] M.Aoyagi, S, Watanabe "Stochastic complexities of reduced rank regression in Bayesian estimation," *Neural Networks.*, vol.18, No.7, pp.924–933, 2005.

[2] M. Aoyagi, S, Watanabe "Resolution of singularities and generalization error with Bayesian estimation for layered neural network," *IEICE Trans.*, J88-D-II, No.10, pp.2112–2124, 2005.

[3] M. Atiyah, "Resolution of singularities and division of distributions," *Communications of Pure and Applied Mathematics.*, vol.13, pp.145–150, 1970.

[4] T. Matsuda, S, Watanabe "On a Singular Point to Contribute to a Learning Coefficient and Weighted Resolution of Singularities," *Proceedings of Internatinal Conference of Artificial Neural Networks (ICANN).*, Part1, LNCS 4131, pp.11–18, 2007.

[5] S, Watanabe "Algebraic Analysis for Non-identifiable Learning Machines," *Neural Computation.*, vol.13, No.4, pp.899–933, 2001.

[6] K. Yamazaki, S. Watanabe, "Newton Diagram and Stochastic Complexity in Mixture of Binomial Distribution," *Algorithmic Learning Theory (ALT04)*, Padova, Italy, pp.105–110, 2004.