The maximum pole of the zeta function for learning theory

Miki AOYAGI and Sumio WATANABE

Abstract

Layered neural network is non-regular learning machine. Therefore, there are many difficult problems to solve. Recently it was proved that the maximum pole of the zeta function asymptotically gives the stochastic complexity of non-regular learning machine in Bayesian estimation, where the zeta function for the learning theory is the integral of the Kullback distance and a certain a priori probability density function. For several non-regular learning machines, upper bounds of the main term in the asymptotic form of the stochastic complexity were obtained. The exact values have been left unknown, because of their computational complexities. In this paper, we introduce a new computational technique, and compute explicitly the main term in the asymptotic form of the stochastic complexity in the case of a three layered neural network.

1 Introduction

Hierarchical learning machines such as reduced rank regression, multi-layer perceptrons, normal mixtures and Boltzmann machines, are important research topics, and have useful applications in many fields. These learning models are called *non-regular* (*non-identifiable*) statistical models. A few mathematical theories for such learning machines are known. So it is necessary and crucial to construct fundamental mathematical theories.

The main topic of this paper is about the zeta function which is the integral of the *Kullback* function and a certain a priori probability density function. Recently, Watanabe[7, 8] proved that the maximum pole of the zeta function asymptotically gives the stochastic complexity of non-regular

learning machine. Furthermore, he showed that the poles of the zeta function can be calculated by using desingularization. By Hironaka's Theorem[5], it is known that the desingularization of an arbitrary polynomial can be obtained by using the blowing-up process. However the desingularization of any polynomial in general, although it is known as a finite process, is very difficult.

In order to calculate the maximum pole of the zeta function, first we obtain the desingularization of the *Kullback* function.

The main problems in obtaining the desingularization are that

- most of the *Kullback* functions are degenerate (over $\mathbf{R}$) with respect to their Newton polyhedrons,
- the *Kullback* functions have parameters, for example, p of Equation (1),
- singular points are not isolated.

We note that there are many classical results to calculate the maximum pole of the zeta function using the desingularization of a plane curve in the dimension two. Also there have been many investigations for the case of the prehomogenious vector space, which corresponds a special case. The *Kullback* function do not occur in the prehomogenious vector space.

In this paper, we give the desingularization of the zeta function for a layered neural network by using a recursive blowing-up and obtain the exact maximum pole.

The applications of our result from the viewpoint of Learning theory are as follows. First, using our result, we can discuss the model selection method for Bayesian estimation. Second, we can analyze and develop the precision of the MCMC method. By the MCMC method, the estimated value of a marginal likelihood had been calculated for the hyper-parameter estimation and the model selection method of complex learning models. We formulated the theoretical value of a marginal likelihood which is given in this paper. Then we can compare the calculated value and the theoretical value.

2 Poles of the zeta function for a three neural network

In this section, we show how to obtain the poles of the zeta function of a learning model in the case of a three neural network.

Consider a three-layer neural perceptron with one input unit, p hidden units, and one output unit.

Let $w = (a_1, \cdots, a_p, b_1, \cdots, b_p) \in \mathbb{R}^p$ be a parameter. Denote the input value by x.

Then the statistical model of the three layered neural network is

$$p(y|x,w) = \frac{1}{(\sqrt{2\pi})^N} \exp(-\frac{1}{2}(y - f(x,w))^2),$$

where $f(x,w) = \sum_{m=1}^{p} a_k \tanh(b_k x)$.

Assume that the true parameter w is $w = 0$ and that the *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with compact support W where $\psi(0) > 0$.

Then the zeta function of this model is written as follows [1].

$$J(z) = \int_W \{\sum_{n=1}^{p} (\sum_{m=1}^{p} a_m b_m^{2n-1})^2\}^z \prod_{m=1}^{p} \mathrm{d}a_m \mathrm{d}b_m. \tag{1}$$

This is obtained by using the Tayler expansion of hyperbolic tangent.

Main Theorem

Let $-\lambda$ be the maximum pole of $\int_W \Psi$ and m its order.

Set $i_0 = \max\{i \mid i^2 \leq p\}$. Then

$$\lambda = \frac{i_0 + i_0^2 + p}{4i_0 + 2}, \quad \ell = \begin{cases} 2 & (i_0^2 = p), \\ 1 & (i_0^2 < p). \end{cases}$$

Consider the following differential form

$$\Psi = \{\sum_{n=1}^{p} (\sum_{m=1}^{p} a_m b_m^{2n-1})^2\}^z \prod_{m=1}^{p} \mathrm{d}a_m \mathrm{d}b_m.$$

Put the auxiliary function $f_{n,l}$ by

$$f_{n,l}(x_1, \cdots, x_l) = \sum_{j_1 + \cdots + j_l = n-l} x_1^{2j_1} \cdots x^{2j_l} > 0.$$

This function satisfies

$$\begin{aligned} &f_{n,l}(x_1, \cdots, x_{l-1}, y_l) - f_{n,l}(x_1, \cdots, x_{l-1}, z_l) \\ &= ((y_l)^2 - (z_l)^2) f_{n,l+1}(x_1, \cdots, x_{l-1}, z_l, y_l). \end{aligned}$$

Let

$$c_i = \sum_{m=i}^{p} a_m b_m (b_m^2 - b_1^2)(b_m^2 - b_2^2) \cdots (b_m^2 - b_{i-1}^2).$$

Then we have

$$\begin{aligned} \Psi &= \{\sum_{n=1}^{p} (f_{n,1}(b_1)c_1 + f_{n,2}(b_1,b_2)c_2 + f_{n,3}(b_1,b_2,b_3)c_3 + \cdots \\ &\qquad + f_{n,n}(b_1,\ldots,b_n)c_n)^2\}^z \prod_{m=1}^{p} \mathrm{d}a_m \mathrm{d}b_m. \end{aligned}$$

Proof of Main Theorem : Part 1

Let $J, J^{(\alpha)}, J_m^{(\alpha)}$ be elements in $\mathbb{R}^\alpha$. Denote $J^{(\alpha)} = (J^{(\alpha')}, *)$ by $J^{(\alpha)} > J^{(\alpha')} (\alpha > \alpha')$ and $J^{(\alpha)} = (0, \cdots, 0)$ by $J^{(\alpha)} = 0^{(\alpha)}$ or $J^{(\alpha)} = 0$. Set $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$.

We need to calculate poles of the following function by using the blowing-up process together with an inductive method of k, K, α.

Inductive statement

Set $s(J) = \#\{m; k \le m \le p, J_m^{(\alpha)} = J\}$, $s(i, J) = \#\{m; k \le m \le i-1, J_m^{(\alpha)} = J\}$, for $J \in \mathbb{R}^\alpha$, where # implies the number of elements.

(a) $K \ge k$,

(b) $\Psi = \{v_1^{t_1} v_2^{t_2} v_3^{t_3} \cdots v_{k-1}^{t_{k-1}} \Big(d_1^2 + (d_1 f_{2,1} + d_2 f_{2,2})^2 + \cdots + (d_1 f_{K-1,1} + \cdots$

$+ d_{K-1} f_{K-1,K-1})^2 + \sum_{n=K}^{p} (d_1 f_{n,1} + \cdots + d_{K-1} f_{n,K-1} + \sum_{i=K}^{p} f_{n,i} c_i)^2 \Big)\}^z$

$\prod_{m=1}^{k-1} v_m^{q_m} \prod_{m=1}^{K-1} \mathrm{d}d_m \prod_{m=K}^{p} \mathrm{d}a_m \prod_{m=1}^{k-1} \mathrm{d}v_m \prod_{m=k}^{p} \mathrm{d}b_m$.

Here, $t_l, q_m \in \mathbb{Z}_+$. Also, there exist $RJ^{(\alpha)} \subset \mathbb{R}^\alpha$, $t(i, J, l) \in \mathbb{Z}_+$ and functions $g(i, m) \ne 0$, $(K \le i \le p, 1 \le l \le k-1, i \le m \le p)$ such that

$$c_i = v_1^{t(i,0,1)} v_2^{t(i,0,2)} \cdots v_{k-1}^{t(i,0,k-1)} \sum_{\substack{i \le m \le p \\ J_m^{(\alpha)} = 0}} g(i,m) a_m b_m \prod_{\substack{k \le i' < i \\ J_{i'}^{(\alpha)} = 0}} (b_m^2 - b_{i'}^2)$$

$$+ \sum_{J \in RJ^{(\alpha)}} v_1^{t(i,J,1)} v_2^{t(i,J,2)} \cdots v_{k-1}^{t(i,J,k-1)} \sum_{\substack{i \le m \le p \\ J_m^{(\alpha)} = J}} g(i,m) a_m b_m \prod_{\substack{k \le i' < i \\ J_{i'}^{(\alpha)} = J}} (b_m - b_{i'})$$

$$+\sum_{J\notin RJ^{(\alpha)},J\neq 0} v_1^{t(i,J,1)}v_2^{t(i,J,2)}\cdots v_{k-1}^{t(i,J,k-1)}\sum_{\substack{i\le m\le p\\ J_m^{(\alpha)}=J}} g(i,m)a_m\prod_{\substack{k\le i'<i\\ J_{i'}^{(\alpha)}=J}}(b_m-b_{i'}).$$

(c) $J_{i'}^{(\alpha)}\neq J_i^{(\alpha)}$ for $k\le i'<i<K$ and $J_i^{(\alpha)}\notin RJ^{(\alpha)}\cup\{0\}$ for $k\le i<K$.

(d) Let $\tilde{t}(i,J,l):=t_l/2+t(i,J,l)$, where $J\in\mathbb{R}^\alpha$, $K\le i\le p, 1\le l\le k-1$. There exist $D_{J^{(\mu)},l}\in\mathbb{Z}_+$ such that

$$\begin{aligned}\tilde{t}(i,J,l)=&\sum_{J>0^{(\mu)}} D_{0^{(\mu)},l}(2s(i,0^{(\mu)})+1)\\ &+\sum_{\substack{J>J^{(\mu)}\\ J^{(\mu)}\in RJ^{(\mu)}}} D_{J^{(\mu)},l}(s(i,J^{(\mu)})+1)+\sum_{\substack{J>J^{(\mu)}\\ J^{(\mu)}\notin RJ^{(\mu)},J^{(\mu)}\neq 0}} D_{J^{(\mu)},l}s(i,J^{(\mu)}).\end{aligned}$$

(e) There exist $g_l\in\mathbb{Z}_+$, $\eta_{k',l}^{(\xi)}\in\mathbb{Z}_+$ $(2\le k'\le K-1, 1\le\xi\le g_l, 1\le l\le k-1)$ such that

$$\begin{aligned}&\frac{t_l}{2}=\sum_{\xi=1}^{g_l}(1+\eta_{2,l}^{(\xi)}+\cdots+\eta_{K-1,l}^{(\xi)}),\\ &0\le\eta_{2,l}^{(\xi)}\le 2, 0\le\eta_{2,l}^{(\xi)}+\eta_{3,l}^{(\xi)}\le 4,\\ &\vdots\\ &0\le\eta_{2,l}^{(\xi)}+\eta_{3,l}^{(\xi)}+\cdots+\eta_{K-1,l}^{(\xi)}\le 2(K-2).\end{aligned}$$

(f) Let $\varphi_l^{(\xi)}:=p+2\eta_{2,l}^{(\xi)}+\cdots+(K-1)\eta_{K-1,l}^{(\xi)}$. There exist $\phi_l\in\mathbb{Z}_+(1\le l\le k-1)$ such that $g_l\le\sum_{J_m^{(\alpha)}>J^{(\mu)}} D_{J^{(\mu)},l}$ and

$$q_l+1 \;=\; \sum_{\xi=1}^{g_l}\varphi_l^{(\xi)}+\phi_l+\sum_{m=k}^{p}(-g_l+\sum_{J_m^{(\alpha)}>J^{(\mu)}} D_{J^{(\mu)},l}).$$

The end of inductive statement

Statements (d), (e) and (f) are needed when we compare poles.
If $J_m^{(\alpha)}=0$ for all m,α, then $\alpha=k-1$ and

(a') $k=K$.

(b') $c_i = v_1^{t(i,0,1)} v_2^{t(i,0,2)} \cdots v_{k-1}^{t(i,0,k-1)} \sum_{i \le m \le p} a_m b_m \prod_{k \le i' < i} (b_m^2 - b_{i'}^2)$, for $k \le i \le p$.

(d') $D_{0^{(l-1)},l} = 1$, others 0.
$\tilde{t}(i, 0^{(k-1)}, l) = D_{0^{(l-1)},l}(2(i-l)+1) = 2(i-l)+1$.

(e') $\frac{t_l}{2} = 1 + \eta_{2,l}^{(1)} + \cdots + \eta_{k-1,l}^{(1)}$,
$\eta_{2,l}^{(1)} = 0, \ldots, \eta_{l-1,l}^{(1)} = 0, \eta_{l,l}^{(1)} = 2, \ldots, \eta_{k-2,l}^{(1)} = 2$,
$0 \le \eta_{k-1,l}^{(1)} \le 2, t(k, 0^{(k-1)}, l) + \eta_{k-1,l}^{(1)} = 2$.

(f') Set $\varphi_l^{(1)} := p + 2\eta_{2,l}^{(1)} + \cdots + (k-1)\eta_{k-1,l}^{(1)}$. Then $q_l + 1 = \varphi_l^{(1)}$.

The proof of Part 1 will appear in [1].

Proof of Main Theorem : Part 2

To obatain the maximum pole, we need the following four lemmas.

Lemma 2.1 *The case of $J_m^{(\alpha)} = 0$ for all m and α, then we obtain the following poles.*

$$\begin{array}{cc}
-\frac{p}{2}, & \\
-\frac{p+k}{4}, & -\frac{p+2k}{6}, \\
-\frac{p+2k+k+1}{8}, & -\frac{p+2k+2(k+1)}{10}, \\
\vdots & \\
-\frac{p+(i-1)(2k-2+i)+k+i-1}{4i}, & -\frac{p+(i-1)(2k-2+i)+2(k+i-1)}{4i+2}, \\
\vdots & \\
-\frac{(p-k-1)(k-2+p)+p-1}{4(p-k)}, & -\frac{p+(p-k-1)(k-2+p)+2(p-2)}{4(p-k)+2}.
\end{array}$$

Proof This can be proved by using the proof of Part 1.

Lemma 2.2 *If $a_m, b_m > 0$, $m \in \mathbb{N}$, then $\frac{\sum a_m}{\sum b_m} \ge \min\{\frac{a_m}{b_m}\}$.*

Lemma 2.3 *Let $k \in \mathbb{N}$.*

Assume $0 \le \eta_2 \le 2, \cdots, 0 \le \eta_2 + \eta_3 + \cdots + \eta_{k-1} \le 2(K-2)$, where $\eta_{k'} \in \mathbb{Z}_+$ and $2 \le k' \le K-1$.

Set

$$t := 1 + \eta_2 + \cdots + \eta_{K-1},$$
$$\varphi := p + 2\eta_2 + \cdots + (K-1)\eta_{K-1},$$

and

$$t = 2i + m, \quad i \in \mathbb{N}, \quad m = 0 \text{ or } 1.$$

Then

$$\frac{\varphi}{2t} > \frac{p + i^2 + im}{4i + 2m} = \frac{p + 1 + 1 + 2 + 2 + \cdots + (i-1) + (i-1) + i + im}{2t}.$$

Remark $-\frac{p+i^2+im}{4i+2m}$ stated in Lemma 2.3 are equal to poles where $k = 1$ in Lemma 2.1.

Lemma 2.4 *The maximum pole among ones obtained in Part 1 is one of poles in Lemma 2.1.*

Proof This proof will also appear in [1].

Therfore, the maximum pole is the maximum one among

$$-\frac{p}{2}, -\frac{p+1}{4}, -\frac{p+2}{6},$$
$$-\frac{p+4}{8}, -\frac{p+6}{10},$$
$$\vdots$$
$$-\frac{p+i^2}{4i}, -\frac{p+i^2+i}{4i+2},$$
$$\vdots$$
$$-\frac{p+(p-1)^2}{4(p-1)}, -\frac{p+(p-1)^2+(p-1)}{4(p-1)+2}.$$

So Main Theorem follows.

Acknowledgments

This research was supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 16700218.

References

[1] Aoyagi, M. and Watanabe, S. Resolution of Singularities and the Generalization Error with Bayesian Estimation for Layered Neural Network, preprint.

[2] Aoyagi, M. and Watanabe, S. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation, Newral Network, (to appear).

[3] Atiyah, M. F. (1970). Resolution of singularities and division of distributions, *Comm. Pure and Appl. Math,* **13** 145-150.

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

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

[6] Sato, M. and Shintani, T. (1974). On zeta functions associated with prehomogeneous vector space. *Annals of Math.* **100** 131-170.

[7] Watanabe, S. (2001a). Algebraic analysis for nonidentifiable learning machines. *Neural Computation.* **13** (4) 899-933.

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

Miki AOYAGI
Department of Mathematics
Sophia University
7-1 Kioi-cho, Chiyoda-ku,
Tokyo, 102-8554 JAPAN

Sumio WATANABE
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku
Yokohama, 226-8503 JAPAN