Almost All Learning Machines are Singular

(Invited Paper in FOCI2007)

Sumio WATANABE
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuda, Midori-ku, Yokohama 226-8503
Email: swatanab@pi.titech.ac.jp

***Abstract*— A learning machine is called singular if its Fisher information matrix is singular.**

Almost all learning machines used in information processing are singular, for example, layered neural networks, normal mixtures, binomial mixtures, Bayes networks, hidden Markov models, Boltzmann machines, stochastic context-free grammars, and reduced rank regressions are singular.

In singular learning machines, the likelihood function can not be approximated by any quadratic form of the parameter. Moreover, neither the distribution of the maximum likelihood estimator nor the Bayes a posteriori distribution converges to the normal distribution, even if the number of training samples tends to infinity. Therefore, the conventional statistical learning theory does not hold in singular learning machines.

This paper establishes the new mathematical foundation for singular learning machines. We propose that, by using resolution of singularities, the likelihood function can be represented as the standard form, by which we can prove the asymptotic behavior of the generalization errors of the maximum likelihood method and the Bayes estimation. The result will be a base on which training algorithms of singular learning machines are devised and optimized.

I. Introduction

A learning machine is called regular if its Fisher information matrix is well-defined and positive definite. It is called *singular* if its Fisher information matrix is singular, in other words, it has zero eigenvalues.

A lot of learning machines used in practical applications are singular. For example, layered neural networks, normal mixtures, mixture of binomial distributions, Bayes networks, hidden Markov models, stochastic context-free grammars, Boltzmann machines, and reduced rank regressions are singular learning machines. In singular learning machines, the conventional statistical learning theory does not hold. The likelihood function can not be approximated by any quadratic form, resulting that neither the distribution of the maximum likelihood estimator nor the Bayes a posteriori distribution converges to a normal distribution [6] [5] [17]. Consequently, AIC, BIC, or MDL, respectively does not correspond to the average prediction error, the Bayes marginal likelihood, or the minimum description length. Since singular learning machines play a central role in information processing systems, it is important to construct the new mathematical foundation, on which statistical learning theory of them can be built.

Let $p(y|x,w)$ represent a learning machine which infers a probabilistic output y from an input x using a parameter w. Then

$$\{p(y|x,w); w \in W\}$$

is the set of all conditional probabilities, where W is the set of parameters. We can introduce a geometry on this set by using Kullback-Leibler distance, whose metric is given by the Fisher information matrix. In a regular statistical model, W can be understood as a smooth differential manifold, whereas, in a singular learning machine, it is an algebraic variety or an analytic set with singularities. Therefore, in order to construct statistical learning theory of singular models, we need a new mathematical foundation such as algebraic geometry and algebraic analysis. In this paper, we show that resolution of singularities gives the standard form of the likelihood function of a singular learning machine, by which the generalization errors of the maximum likelihood method and Bayes estimation are clarified.

II. Framework of Statistical Learning

A. Learning Machine and Training Data

Let $p(y|x,w)$ be a conditional probability distribution of an output y for a given input x and a parameter w. Here $p(y|x,w)$ is referred to as a learning machine. We assume that the set of all parameters W is a subset of the d dimensional Euclidean space. Let

$$D_n = \{(x_i, y_i); i = 1, 2, \ldots, n\}$$

be a set of training samples which are independently taken from the true and unknown distribution

$$q(x,y) = q(x)q(y|x).$$

By using the sample data set D_n, the learning machine estimates the true conditional probability $q(y|x)$ and obtains an trained inference $p^*(y|x)$. For example, in the maximum likelihood learning, the trained inference is given by

$$p^*(y|x) = p(y|x,\hat{w}),$$

where $\hat{w}$ is the maximum likelihood estimator (MLE). In the Bayes learning, the trained inference is given by

$$p^*(y|x) = \int p(y|x,w)\, p(w|D_n)\, dw,$$

where $p(w|D_n)$ is the Bayes a posteriori distribution. The generalization error of the trained inference $p^*(y|x)$ is defined by

the Kullback information from the true conditional probability to the estimated conditional probability,

$$G = \int\int q(y|x)q(x)\log\frac{q(y|x)}{p^*(y|x)}dxdy.$$

Note that $G \geq 0$ is a random variable because it is a function of the sample set D_n. Its expectation value $E[G]$ is called the average generalization error. The smaller generalization error means that the learning is more appropriate.

B. Regular and Singular Learning Machines

The Fisher information matrix $I(w) = \{I_{ij}(w)\}$ is defined by

$$I_{ij}(w) = \int\int L_i(w,x,y)L_j(w,x,y)p(y|x,w)q(x)dxdy,$$

where

$$L_i(w,x,y) = \frac{\partial}{\partial w_i}\log p(y|x,w).$$

Note that Fisher information matrix is positive semi-definite in both regular and singular models, in other words, its eigen values are all nonnegative.

Example. If a learning machine is a regression model using a parametric function $f(x,w)$

$$p(y|x,w) \propto \exp\Big(-\frac{1}{2}(y-f(x,w))^2\Big),$$

then the Fisher information matrix is equal to

$$I_{ij}(w) = \int \partial_i f(x,w)\partial_j f(x,w)p(y|x,w)q(x)dx,$$

where $\partial_i = (\partial/\partial w_i)$. The Fisher information matrix is positive definite if and only if $\{\partial_i f(x,w)\}$ is linearly independent.

A learning machine is called regular if $I(w)$ is positive definite for an arbitrary w. If otherwise, then the learning machine is called singular. In a singular learning machine, there exists a parameter w such that $\det I(w) = 0$. Such a parameter is called a singularity of the Fisher information matrix. At a singularity of the Fisher information matrix, the likelihood function can not be approximated by any quadratic form of the parameter.

Remark. If a learning machine has a singularity of the Fisher information matrix, the set of such points

$$S = \{w \in W; \det I(w) = 0\}$$

is not the empty set. In general, the set S is not one point, and it contains singularities in itself (singularities of the analytic set S).

Example. Let us study a function with a parameter $w = \{a_k, b_k\}$,

$$f(x,w) = \sum_{k=1}^{K}\Big\{a_k\sin(x) + b_k\cos(x)\Big\} \quad (0 \leq x \leq 2\pi),$$

then Fisher information matrix is always positive definite. Hence the regression model which employs Fourier series is regular. However, if a function with a parameter $w = \{a_k, b_k\}$,

$$f(x,w) = \sum_{k=1}^{K} a_k\tanh(b_k x) \quad (0 \leq x \leq 1) \tag{1}$$

is used in a regression model, then the set

$$\{\frac{\partial}{\partial a_k}f(x,w), \frac{\partial}{\partial b_k}f(x,w)\}$$

is not linearly independent if at least one of a_k or b_k is equal to zero. Hence the regression model is singular. It was shown [22] that the set of all singularities of the Fisher information matrix of this machine

$$S = \{w; \det I(w) = 0\}$$

is an algebraic variety which is determined by the ideal generated by

$$\sum_{k=1}^{K} a_k b_k^{2h-1} = 0 \quad (h = 1, 2, .., K).$$

This algebraic variety S contains singularities.

Remark. In singular learning machines, a parameter which corresponds to a smaller model is a singularity of the Fisher information matrix in the larger model, whereas, in regular statistical models, it is not.

Example. The normal distribution

$$p(x|a) \propto \exp(-\frac{||x-a||^2}{2})$$

is regular, whereas its mixture

$$p(x|a,b,c) \propto a\ \exp(-\frac{||x-b||^2}{2}) + (1-a)\ \exp(-\frac{||x-c||^2}{2})$$

is singular. In general, if a learning machine has a layered structure or a hidden variable, then it is singular, in general. Hence almost all learning machines in neural information processing is singular.

Remark. Some researchers claim that the conventional statistical learning theory even holds in a singular learning machine in the case when the optimal parameter for function approximation is uniquely determined. It is not true. Even if the optimal parameter in function approximation is unique, singularities affects the learning process by the bias-variance problem. It should be emphasized that singularities of the Fisher information matrix has the larger bias but the smaller variance than the ordinary points. In statistical engineering process such as model selection and hypothesis testing, we have to study the balance of the bias and the variance of the parameter. It depends on the true distribution, the learning machine, and the number of training samples [21][23] [13].

III. Mathematical Foundation

A. Standard Form of Singular Likelihood

In singular learning machines, the conventional statistical learning theory does not hold, hence we need a new mathematical foundation on which the likelihood function is appropriately treated. We propose the following theorem is the basic one for singular learning machines, which is called *resolution of singularities*, or *resolution theorem* in algebraic geometry.

Resolution Theorem. (Hironaka,1964). Let $H(w)$ be an analytic function on an open set W in $\mathbf{R}^d$, which satisfies $H(w) \geq 0$ ($w \in W$) and $H(w_0) = 0$ for some $w_0 \in W$. Then there exist both a d dimensional manifold U and an analytic function $g : U \to W$, such that, for an arbitrary coordinate in $u \in U$,

$$\begin{aligned} H(g(u)) &= u_1^{2k_1} \cdots u_d^{2k_d}, \\ |g'(u)| &= b(u) u_1^{h_1} \cdots u_d^{h_d}, \end{aligned}$$

where $k_1, \ldots, k_d, h_1, \ldots, h_d$ are nonnegative integers, $|g'(u)|$ is the Jacobian of the map $w = g(u)$, and $b(u) > 0$ is an analytic function.

Remark. This theorem is the well-known basic theorem in algebraic geometry proved by Hironaka [8], on which Atiyah and Kashiwara respectively made the foundation of distribution theory [4] and algebraic analysis [11]. It was firstly proposed in [18][20] that this theorem is essential to statistical learning theory. This theorem claims that the function $H(g(u))$ can be made as the direct product of $u_1, u_2, \ldots, u_d$, which is said to be *normal crossing*. The manifold U is not orientable in general. The analytic function $w = g(u)$ may not be invertible at u such that $H(g(u)) = 0$, however, invertible at u such that $H(g(u)) \neq 0$. For a given function $H(w)$, both the manifold U and analytic map $w = g(u)$ can be algorithmically found by using recursive blowing-ups or toric modification.

Let us define the empirical and average Kullback informations respectively by

$$\begin{aligned} K_n(w) &= \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(y_i|x_i)}{p(y_i|x_i, w)}, \\ K(w) &= \int \int q(x) q(y|x) \log \frac{q(y|x)}{p(y|x, w)} dx dy. \end{aligned}$$

For simplicity, we assume in this paper that there exists a parameter w_0 such that $q(y|x) = p(y|x, w_0)$, in other words, the true distribution is contained in the learning machine. Even if the true is not contained in a learning machine, singularities strongly affect learning process [22][23]. If there exists a parameter w_0 such that $K(w_0) = 0$, then $K(w)$ satisfies the assumptions of resolution theorem, hence we can apply the resolution theorem to $K(w)$. It is immediately shown that there exist both a manifold U and an analytic function $g : U \to W$ such that

$$\begin{aligned} K(g(u)) &= A(u)^2 \\ A(u) &= u_1^{k_1} \cdots u_d^{k_d}. \end{aligned}$$

By using this fact, the empirical Kullback information can be written as

$$K_n(g(u)) = nA(u)^2 + \sqrt{n} A(u) \xi_n(u), \tag{2}$$

where

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i, y_i, u).$$

Here the function $\psi(x, y, u)$ is defined by

$$\psi(x, y, u) = \frac{1}{A(u)} \Big(K(g(u)) - \log \frac{q(y|x)}{p(y|x, g(u))} \Big).$$

We propose that the equation (2) is the standard representation of the likelihood function of a singular learning machine.

Remark. The function $\psi(x, y, u)$ can be written as $\psi(x, y, g^{-1}(w))$ if $K(w) \neq 0$. However, it is ill-defined as the function of w at $K(w) = 0$ in general. On the other hand, we can prove $\psi(x, y, u)$ is an analytic function of u even when $K(g(u)) = 0$. The fact that $\psi(x, y, u)$ is well-defined function of u can be proved by the normal crossing property of $A(u)$.

The random process $\xi_n(u)$ is an *empirical process*, which satisfies

$$E[|\xi_n(u)|^2] = 2 \quad (\text{if} \quad K(g(u)) = 0).$$

Moreover, $\xi_n(u)$ weakly converges to the tight gaussian process $\xi(u)$ when n tends to infinity. Here the gaussian process $\xi(u)$ is uniquely identified by its average and covariance

$$\begin{aligned} E[\xi(u)] &= 0 \quad (\forall u), \\ E[\xi(u_1)\xi(u_2)] &= E[\psi(X, Y, u_1)\psi(X, Y, u_2)] \quad (\forall u_1, u_2). \end{aligned}$$

Then the empirical Kullback information can be written as[24],

$$K_n(g(u)) = \Big(\sqrt{n} A(u) + \frac{\xi(u)}{2} \Big)^2 - \frac{1}{4} |\xi(u)|^2 \tag{3}$$

on the manifold U.

B. Example

Let us study a learning machine

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2} (y - a\sigma(bx) - cx)^2 \Big),$$

where $\sigma(x) = x + x^2$ and $w = (a, b, c)$. We define the set of parameters by

$$W = \{w; |a| \leq 1, |b| \leq 1, |c| \leq 1\}.$$

Assume that the true distribution is

$$q(y|x) q(x) = p(y|x, 0, 0, 0)\, q(x),$$

where $q(x)$ is the standard normal distribution (average 0, variance 1). Then the Kullback information is

$$K(w) = \frac{1}{2} \Big((ab + c)^2 + 3a^2 b^4 \Big).$$

Note that $K(w) = 0$ defines an algebraic variety,

$$ab = c = 0.$$

The function $K(w)$ is not normal crossing, however, we can make it normal crossing by the recursive blowing-ups. Let us introduce four sets

$$\begin{aligned} W_1 &= \{w; |a| \leq |c|\}, \\ W_2 &= \{w; |a| \geq |c|, |ab| \leq |ab + c|\}, \\ W_3 &= \{w; |a| \geq |c|, |ab + c| \leq |ab^2|\}, \\ W_4 &= \{w; |a| \geq |c|, |ab^2| \leq |ab + c| \leq |ab|\}. \end{aligned}$$

It follows that $W = \cup_i W_i$. Let U_i $(i = 1, 2, 3, 4)$ be sets contained in $\mathbf{R}^3$,

$$U_i = \{(a_i, b_i, c_i)\}.$$

A function $g_i : U_i \to W$ $(i = 1, 2, 3, 4)$ is defined on each U_i such that

$$\begin{aligned} a = a_1 c_1, &\quad b = b_1, &\quad c = c_1, \\ a = a_2, &\quad b = b_2 c_2, &\quad c = a_2(1 - b_2)c_2, \\ a = a_3, &\quad b = b_3, &\quad c = a_3 b_3(b_3 c_3 - 1), \\ a = a_4, &\quad b = b_4 c_4, &\quad c = a_4 b_4 c_4(c_4 - 1). \end{aligned}$$

Then U_i can be understood as a local coordinate whose union defines a manifold $U = \cup_i U_i$, where two points $u_i \in U_i$ and $u_j \in U_j$ are identified as a one point in U if and only if $g_i(u_i) = g_j(u_j)$. In each coordinate, the Kullback information is given by

$$\begin{aligned} 2K(g(u)) &= c_1^2\{(a_1 b_1 + 1)^2 + 3a_1^2 b_1^4\}, \\ &= a_2^2 c_2^2(1 + 3b_2^4 c_2^2), \\ &= a_3^2 b_3^4(c_3^2 + 3), \\ &= a_4^2 b_4^2 c_4^4(1 + 3b_4^2). \end{aligned}$$

It is easy to see that the empirical process $\xi_n(u)$ is well-defined in each coordinate.

IV. Maximum Likelihood and Maximum A Posteriori

In the maximum likelihood method and the maximum a posteriori method, the loss function

$$L(w) = -\sum_{i=1}^{n} \log p(y_i|x_i, w) - a_n \log \varphi(w)$$

is minimized, where $\varphi(w)$ is some a priori probabilty distribution on W and $\{a_n \geq 0\}$ is a real sequence. The parameters that minimize $L(w)$ with $a_n = 0$ and $a_n = 1$ are respectively called the maximum likelihood estimator and the maximum a posteriori estimator. Sometimes the other estimators are used with the other conditions on a_n. Let $\hat{w}$ be the parameter that minimizes $L(w)$. Then the average training error $E_t(n)$ and the average generalization error $E_g(n)$ are respectively defined by

$$\begin{aligned} E_t(n) &= E[K_n(\hat{w})], \\ E_g(n) &= E[K(\hat{w})], \end{aligned}$$

where $E[\ \ ]$ shows the expectation value over all sets of training samples. Even in singular learning machines, it was proven in [24] that, if W is contained in a compact set, the symmetry of two errors hold,

$$\begin{aligned} E_t(n) &= -\frac{\mu}{n} + o(\frac{1}{n}), \\ E_g(n) &= \frac{\mu}{n} + o(\frac{1}{n}), \end{aligned}$$

where μ is a constant. In regular statistical models, it is well known that $\mu = d/2$. However, in singular learning machines, μ is not equal to $d/2$. In general, $\mu > d/2$.

Let us assume that, in each local coordinate, at least one of $k_1, ..., k_d$ is an odd number. When

$$\lim_{n\to\infty} a_n = \alpha,$$

then

$$\mu = \frac{1}{4} E[\psi(\hat{u})^2].$$

The estimator $\hat{u}$ is defined by

$$\hat{u} = \arg \inf_{u \in U_0} \Big(-\frac{\xi(u)^2}{4} + \alpha\varphi(g(u))\Big),$$

where U_0 is the set of parameters which satisfy $K(g(u)) = 0$. Therefore, if $a_n = 0$, then

$$\mu = \frac{1}{4} E[\sup_{u \in U_0} \xi(u)^2].$$

If $\lim_n a_n = \infty$ and $\lim_n a_n/\sqrt{n} = 0$, then

$$\mu = \frac{1}{4} E[\sup_{u \in U_{00}} \xi(u)^2],$$

where U_{00} is the set of parameters which maximize $\varphi(g(u))$ in U_0.

Remark. If W is not a compact set, MLE often does not exist. Even if MLE exists, the generalization error becomes larger than μ/n. The MLE in singular a singular learning machine has a quite stronger over-fitting property than that of a regular statistical models. The maximum likelihood method is more inappropriate for singular learning machines than regular statistical models. On the other hand, as we show in the following section, the average generalization error of Bayes estimation is far smaller than that of the maximum likelihood method in singular learning machines.

V. Bayes Estimation

A. Theoretical Results

In Bayes learning, the free energy or the marginal likelihood is given by

$$F(D_n) = -\log \int \prod_{i=1}^{n} p(y_i|x_i, w)\ \varphi(w)\ dw,$$

where $\varphi(w)$ is an a priori distribution on W. It is easy to show the average generalization error of Bayes is equal to the increase of the marginal likelihood

$$E[G] = E[F(D_{n+1})] - E[F(D_n)] - nS,$$

where S is the entropy of the true distribution,

$$S = -\int\int q(x)q(y|x)\log q(y|x)dxdy.$$

The zeta function of a learning machine [20] is important in Bayes theory,

$$\zeta(z) = \int K(w)^z \varphi(w)\, dw. \quad (Re(z) > 0).$$

The function $\zeta(z)$, which is a holomorphic function in $Re(z) > 0$, can be analytically continued to the meromorphic function on the entire complex plane. The analytic continuation is ensured by again the resolution theorem [4] or the existence of b-function [11]. It is also proved that the poles of the zeta function are all real and negative integers,

$$0 > -\lambda_1 > -\lambda_2 > -\lambda_3 > \cdots.$$

Let m_1 be the order of the largest pole $(-\lambda_1)$. Then it was proved that the free energy has the asymptotic expansion [18][20]

$$F(D_n) = nS_n + \lambda_1 \log n - (m_1 - 1)\log\log n + R(D_n),$$

where S_n is the empirical entropy of the true distribution

$$S_n = -\frac{1}{n}\sum_{i=1}^{n} \log q(y_i|x_i),$$

which does not depend on the learning machine, and $R(D_n)$ is a random variable which weakly converges to the random variable represented by the random process $\xi(u)$. By using this result, we obtain that, if the Bayes generalization error has asymptotic expansion, then it is

$$E[G] = \frac{\lambda_1}{n} + o(\frac{1}{n}).$$

If the learning machine is regular, $\lambda_1 = d/2$ and $m_1 = 1$. If it is singular and $\varphi(w) > 0$ at some singularity of the Fisher information matrix, then $\lambda_1 < d/2$ and $m_1 \geq 1$. Note that the Jeffreys' prior defined by

$$\varphi(w) \propto \sqrt{\det I(w)},$$

is equal to zero at the singularity of the Fisher information matrix. If the Jeffreys' prior is employed, then $\lambda_1 \geq d/2$ [19].

By using the resolution theorem again, the largest pole and its order can be exactly obtained. The largest pole $(-\lambda_1)$ of

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

is given by

$$\lambda_1 = \min_j \frac{h_j + 1}{2k_j}, \tag{4}$$

and its order m_1 is equal to the number of j which attains the minimum in equation (4).

Complete resolution of singularities was given in a layered neural network [2] and a reduced rank regression [3]. If a learning machine given by equation (1) is trained so as to learn the true distribution $a_1 = a_2 = \cdots = 0$, then

$$\lambda_1 = \frac{[\sqrt{H}]^2 + [\sqrt{H}] + H}{4[\sqrt{H}] + 2},$$

where $[\sqrt{H}]$ is the largest integer that is not larger than $\sqrt{H}$. It is still an open problem to find the complete resolution of some learning machines. However, if a partial resolution of singularities is found, then the upper bound of λ_1 is simultaneously obtained. Partial resolutions were found in general three-layer perceptrons [21], gaussian mixtures [25], Boltzmann machines[26], and hidden Markov models[27].

B. Variational Bayes

In the variational Bayes (VB) learning, the a posteriori distribution $p(w|D_n)$ is approximated by the VB posterior distribution

$$r(w) = r_1(w_1)r_2(w_2)\cdots r_d(w_d).$$

The probability distributions $r_1(w_1), \cdots, r_d(w_d)$ are determined by minimization of the Kullback information

$$\int r(w)\log\frac{r(w)}{p(w|D_n)}dw.$$

In singular learning machines, the posterior distribution does not converge to any normal distribution, resulting that the VB estimation is different from the true Bayes a posteriori distribution even if the number of training samples goes to infinity. The VB free energy defined by

$$F_{vb}(D_n) = F(D_n) + \int r(w)\log\frac{r(w)}{p(w|D_n)}dw$$

has the different asymptotic form than the true Bayes marginal likelihood. The asymptotic forms of the VB free energy were clarified in normal mixtures [15], in general mixtures models [16], hidden Markov models [9], stochastic context-free grammars [10], and Boltzmann machines [14]. Unfortunately, in VB learning, the average generalization error is not equal to the increase of the free energy, hence the generalization error of VB learning is still an open problem. It is shown that the behavior of the VB generalization error is quite different from the VB free energy in reduced rank regression [13].

C. Markov Chain Monte Calro

In singular learning machines, the true Bayes a posteriori distribution has a quite complex form in the neighborhood of singularities. It is often difficult for both the Markov chain Monte Carlo and VB to approximate such probability distribution. Recently, it was shown that an improved method, the exchange Monte Carlo, is appropriate for constructing the a posteriori distribution in singular learning machines, resulting in the smaller generalization errors [12]. The theoretical values of the free energy gives a good index which measures preciseness of the approximation of the a posteriori distributions.

VI. Conclusion

A lot of learning machines used in practical applications are singular, to which the conventional statistical learning theory can not be applied.

This paper introduces the standard form of the likelihood function in singular learning machines, by which the asymptotic behaviors of the training and generalization errors are clarified.

Nowadays, new training algorithms such as the Variational Bayes and the exchange Monte Calro method are being developed. Based on the theoretical results of singular learning machines, we can develop new information tools which measures the speed and preciseness of new training algorithms.

Acknowledgment

This work was supported by the Ministry of Education, Science, Sports, and Culture in Japan, Grand-in-aid for scientific research 15500310 and 18079007.

References

[1] S.-i. Amari, H. Park, and T. Ozeki, "Singularities Affect Dynamics of Learning in Neuromanifolds," Neural Comput., 18(5),pp.1007 - 1065,2006.

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

[3] M.Aoyagi, S.Watanabe,"Stochastic complexities of reduced rank regression in Bayesian estimation," Neural Networks, Vol.18,No.7, pp.924-933, 2005.

[4] M.F.Atiyah,"Resolution of singularities and division of distributions," Comm. Pure Appl. Math., Vol.13,pp.145-150,1970.

[5] K. Hagiwara, "On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario, " Neural Comput., Vol.14,Vol.8, pp.1979 - 2002, 2002.

[6] J.A.Hartigan,"A failure of likelihood asymptotics for normal mixture," Proc. of Barkeley Conf. in honor of Jerzy Neyman and Jack Keifer, Vol.2, pp.807-810,1985.

[7] T. Hayasaka, M. Kitahara, and S. Usui, "On the Asymptotic Distribution of the Least-Squares Estimators in Unidentifiable Models," Neural Comput., Vol.16 ,No.1, pp.99 - 114, 2004.

[8] H. Hironaka, "Resolution of singularities of an algebraic variety over a field of characteristic zero," Ann. of Math., Vol.79, 109-326,1964.

[9] T.Hosino, K.Watanabe,S.Watanabe,"Stochastic complexity of Hidden Markov Models on the Variational Bayesian Learning," to appear in IEICE Transactions (D-II).

[10] T.Hosino, K.Watanabe, S.Watanabe,"Free Energy of Stochastic Context Free Grammar on Variational Bayes. " pp.407-416, Proc. of ICONIP,2006.

[11] M. Kashiwara, "B-functions and holonomic systems," Inventions Math., 38, 33-53.1976.

[12] K.Nagata, S.Watanabe, "The Exchange Monte Carlo Method for Bayesian Learning in Singular Learning Machines," Proc. of WCCI2006,(Canada, Cancouver), 2006.

[13] S. Nakajima, S. Watanabe, "Variational Bayes Solution of Linear Neural Networks and its Generalization Performance," Neural Computation, to appear.

[14] Y. Nishiyama, S.Watanabe, "Asymptotic Behavior of Stochastic Complexity of Complete Bipartite Graph-type Boltzmann Machines," Proc. of ICONIP2006, (Chaina,HongKong) to appear, 2006.

[15] K.Watanabe, S.Watanabe, "Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation," Journal of Machine Learning Research, Vol.7, (Apr),pp. 625–644, 2006.

[16] K.Watanabe, S.Watanabe,"Stochastic complexities of general mixture models in variational Bayesian leaning," Neural Networks, to appear.

[17] S.Watanabe, "Generalized Bayesian framework for neural networks with singular Fisher information matrices," Proc. of International Symposium on Nonlinear Theory and Its applications, (Las Vegas), pp.207-210, 1995.

[18] S.Watanabe, "Algebraic analysis for singular statistical estimation," Proc. of International Journal of Algorithmic Learning Theory, Lecture Notes on Computer Sciences, 1720, pp.39-50, 1999.

[19] S. Watanabe, "Algebraic information geometry for learning machines with singularities," Advances in Neural Information Processing Systems, (Denver, USA), pp.329-336. 2001.

[20] S.Watanabe, "Algebraic Analysis for Nonidentifiable Learning Machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001.

[21] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," Neural Networks, Vol.14, No.8,pp.1049-1060, 2001.

[22] S. Watanabe, "Learning efficiency of redundant neural networks in Bayesian estimation," IEEE Transactions on Neural Networks, Vol.12, No.6, 1475-1486,2001.

[23] S.Watanabe, S.-I.Amari,"Learning coefficients of layered models when the true distribution mismatches the singularities", Neural Computation, Vol.15,No.5,1013-1033, 2003.

[24] S.Watanabe,"Algebraic geometry of singular learning machines and symmetry of generalization and training errors," Neurocomputing, Vol.67,pp.198-213,2005.

[25] K.Yamazaki, S.Watanabe,"Singularities in mixture models and upper bounds of stochastic complexity." International Journal of Neural Networks, Vol.16, No.7, pp.1029-1038,2003.

[26] K.Yamazaki, S.Watanabe," Singularities in Complete bipartite graph-type Boltzmann machines and upper bounds of stochastic complexities", IEEE Trans. on Neural Networks, Vol. 16 (2), pp.312-324, 2005.

[27] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models", Neurocomputing, Vol.69, pp.62-84, 2005.