Algebraic Geometry of Singular Learning Machines and Symmetry of Generalization and Training Errors

Sumio Watanabe
P&I Lab., Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan
Tel:+81-45-924-5017
Fax:+81-45-924-5018
E-mail: swatanab@pi.titech.ac.jp

October 26, 2004

Key Words: Algebraic geometry, Singular learning machines, resolution of singularities, likelihood function, generalization error

Algebraic Geometry of Singular Learning Machines
and Symmetry of Generalization and Training Errors

Sumio Watanabe, Tokyo Institute of Technology

Abstract

A lot of hierarchical learning machines such as neural networks and normal mixtures are singular learning machines. In such a learning machine, the likelihood function can not be approximated by any quadratic form, resulting that the conventional statistical theory does not hold. This paper proves the symmetrical property of the generalization and training errors based on the algebraic geometrical method. Firstly, a new parameterization is introduced by applying the resolution of singularities. Secondly, the asymptotic behavior of the likelihood function is clarified based on the empirical process theory. Lastly, the asymptotic forms of the generalization and training errors are derived. The result will be a mathematical foundation of model selection and hypothesis testing in singular learning machines.

1 Introduction

Almost hierarchical learning machines used in information science such as layered neural networks, normal mixtures, Boltzmann machines, Bayesian networks, hidden Markov Models, and reduced rank approximations are not regular statistical models. Such learning machines are nonidentifiable and singular (Watanabe,2001a), in other words, the mapping from parameters to probability distributions is not one-to-one and the set of parameters whose Kullback informations are equal to zero has singularities. For example, if a parameter w of a three layered neural network with H hidden units represents a network with H_0 hidden units ($H_0 < H$), then the set of parameters which realize the same inference from inputs to outputs as w is not one point but an analytic set with singularities. Here a set is called an analytic set if and it consists of all zero points of an analytic function. In this case such parameters are defined as zero points of Kullback information from the true distribution to the learning machine. The Fisher information matrices $I(w)$ of such parameters are degenerate, $\det I(w) = 0$, hence asymptotic normality of the maximum likelihood estimator does not hold. These are the reasons why neither the model selection method nor the hypothesis testing method has been established in singular learning machines.

It should be emphasized that the difficulty caused by singularities is not a special problem but a universal one. If a learning machine is made to have a layered or symmetrical structure, then there are a lot of singularities in the parameter space. When we compare several singular learning machines which almost attain the true distribution, then we have to clarify the effect of singularities in learning and generalization.

In a mixture of normal distributions, the failure of the testing hypothesis method of regular statistical models was pointed out in statistics (Hartigan,1985). In a three-layered neural network, the information criterion AIC does not correspond to the generalization error (Hagiwara, 2002). Also it was shown that the Bayes a posteriori distribution converges to a quite different one from the normal distribution and that the information criterion BIC is not equal to the Bayes marginal likelihood even if the number of training samples is sufficiently large (Watanabe, 1995). A phenomenon caused by singularities was studied from the viewpoint of information geometry (Amari, Park, & Ozeki,2002).

The effect of singularities in learning has been left unknown because it has been difficult to analyze them mathematically. In the neighborhood of singularities, neither the likelihood function nor Kullback information can be approximated by any quadratic form of parameters. The limiting distribution of the maximum likelihood estimator is still unknown, and the Bayes a posteriori distribution does not concentrate on one parameter. Hence the conventional statistical theory of regular models does not hold.

However, recently, a new method based on algebraic geometry has been developed, which enables us to rigorously analyze the asymptotic expansion of the Bayes marginal likelihood and the Bayes generalization error in general singular learning machines (Watanabe, 1999; Watanabe, 2001a;Watanabe, 2001b; Yamazaki & Watanabe,2003). It was shown that the resolution of singularities plays the central role in construction of mathematical foundation for learning machines with singularities.

In this paper, we study the maximum likelihood method and the maximum a posteriori method, and show that the algebraic geometrical method also gives a

sufficient condition for the symmetrical behavior of the generalization and training errors. It is already shown that the training error of the maximum likelihood method in a singular learning machine has quite different asymptotics from that in a regular statistical model, based on a locally conic parameterization (Dacunha-Castelle & Gaissart, 1997). However, it is still unknown whether arbitrary singular learning machine has the locally conic parameterization or not. Moreover, the relation between the generalization and training errors is still left unknown.

In this paper, we show that the resolution of singularities gives the general asymptotic behavior of the log-likelihood function, upon which the symmetrical property of the generalization and training errors can be proved. In the second section, the definitions of the generalization and training errors are introduced in the maximum likelihood method and the maximum a posteriori method. In the third section, the main theorem and the assumptions are described. The fourth section is devoted to an introduction of resolution of singularities and some theorems which are proved based on resolution of singularities. In the fifth section, the empirical process theory which is needed in nonidentifiable models is explained. In the sixth section, we give the proof of the main theorem using results in the fourth, fifth, and sixth sections. Discussions and conclusions are respectively given in the seventh and eighth sections.

2 Generalization and Training Errors

Let $p(x|w)$ represent a learning machine which is defined as a conditional probability distribution of $x \in \mathbf{R}^N$ for a given parameter $w \in \mathbf{R}^d$. The set of parameters is denoted by $W \subset \mathbf{R}^d$. Assume that the training samples $X_1, X_2, .., X_n$ are random variables independently taken from the true distribution $p(x|w_0)$, where w_0 is referred to as the true parameter. The log-likelihood function is defined by

$$L(w) = -\sum_{i=1}^{n} \log p(X_i|w) + \lambda_n f(w),$$

where $f(w) = 0$ if we adopt the maximum likelihood method, or $f(w) = -\log \pi(w)$ and $\lambda_n \equiv 1$ if we do the maximum a posteriori method using an a priori distribution $\pi(w)$. If λ_n is chosen as an increasing function of n, then $L(w)$ is called a loss function of a generalized a posteriori method. The λ_n is called the weight of the a

priori distribution. We assume that the a priori distribution is positive $\pi(w) > 0$ for arbitrary $w \in W$.

The estimated parameter $\hat{w}$ is defined by

$$\hat{w} = \arg\min_{w \in W} L(w).$$

Note that $\hat{w}$ is a random variable because it depends on training samples. The Kullback information $K(w)$ and the log-likelihood ratio $R(w)$ are respectively defined by

$$\begin{aligned} K(w) &= \int p(x|w_0) \log \frac{p(x|w_0)}{p(x|w)} dx, \\ R(w) &= \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i|w_0)}{p(X_i|w)}. \end{aligned}$$

The generalization error $E_g(n)$ and the training error $E_t(n)$ are respectively defined by

$$\begin{aligned} E_g(n) &= E_{X^n}[K(\hat{w})], \\ E_t(n) &= E_{X^n}[R(\hat{w})], \end{aligned}$$

where $E_{X^n}[\cdot]$ shows the expectation value over all sets of training samples, $X^n = (X_1, X_2, .., X_n)$.

Definition. It is said that the symmetry of the generalization and training errors holds, if and only if there exists a constant $\mu \geq 0$ such that

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

when n tends to infinity. The constant μ is referred to as the learning coefficient.

It is well known that, if the learning machine $p(x|w)$ is a regular statistical model, then the symmetry holds with $\mu = d/2$, where d is the number of parameters. This

fact is proven based on the property that the Fisher information matrix $I(w_0)$ is positive definite, which ensures the asymptotic normality of the maximum likelihood estimator. When the set of the true parameter is not one point or the Fisher information matrix is not positive definite, it has been left unknown whether the symmetry of both errors holds or not. In Bayes estimation, it is shown that the symmetry does not hold in general (Watanabe & Amari, 2003). This paper gives a sufficient condition upon which the symmetry holds in singular learning machines in the maximum likelihood method or the maximum a posteriori method.

Note that, in a singular learning machine, the set of true parameters

$$W_0 = \{w \in W \ ; \ K(w) = 0\}$$

is not one point but an analytic set with singularities. In the neighborhood of singularities, there is no coordinate which makes W_0 be a manifold.

Example. If a three-layer perceptron,

$$p(y|x,a,b,c,d) = \frac{1}{(2\pi)^{1/2}} \exp\Big[-\frac{1}{2}(y - a\tanh(bx) + c\tanh(dx))^2 \Big]$$

is trained using samples taken from the true distribution $p(y|x,0,0,0,0)q(x)$, where $q(x)$ is some probability distribution of x, then

$$K(a,b,c,d) = \frac{1}{2}\int (a\tanh(bx) - c\tanh(dx))^2 \ q(x) \ dx,$$

resulting that

$$W_0 = \{(a,b,c,d); ab + cd = 0, ab^3 + cd^3 = 0\}.$$

In this case, W_0 is defined by polynomials, which is called an algebraic variety. The origin $(0,0,0,0)$ is a singularity of W_0.

3 Main Theorem

We assume two conditions, (C.1) and (C.2).

(C.1) The set of parameters W is a compact set in $\mathbf{R}^d$ which is a closure of a nonempty open set. The set of true parameters

$$W_0 \equiv \{w \in W; K(w) = 0\}$$

is not an empty set.

(C.2) The log-likelihood ratio function

$$h(x, w) \equiv \log \frac{p(x|w_0)}{p(x|w)}$$

is an $L^2(w_0)$-valued analytic function of $w \in W$. Here $L^2(w_0)$ is the real Hilbert space defined by the square integrable functions,

$$L^2(w_0) = \{f(x) \text{ measurable } ; \int f(x)^2 p(x|w_0) dx < \infty\}.$$

The function

$$W \ni w \mapsto h(x, w) \in L^2(w_0)$$

is called an $L^2(w_0)$-valued analytic function if and only if the Taylor expansion of $h(x, w)$ among an arbitrary $w' \in W$ absolutely converges in some convergence radii by the norm of $L^2(w_0)$.

Since we assume W is a compact set and the likelihood function $L(w)$ is a continuous function of w, there exists a parameter $\hat{w}$ which minimizes $L(w)$. The following is the main theorem of this paper.

Theorem 1 *Let $p(x|w)$ be a learning machine which satisfies conditions (C.1) and (C.2). Assume that the weight of the a priori distribution satisfies*

$$\lim_{n \to \infty} \frac{\lambda_n}{\sqrt{n}} = 0.$$

Then, the symmetry of the generalization and training errors holds.

Remark. The first condition (C.1) assumes that the true distribution is contained in the learning machine. In order to make a model selection or hypothesis testing algorithm, this is a natural assumption. The second condition (C.2) assumes that the likelihood ratio function is an analytic function of the parameter. A lot of natural learning machines such as layered neural networks satisfy these conditions.

4 Resolution of Singularities

In the proof of the main theorem, the following theorem in algebraic geometry plays a central role.

Theorem 2 *(Resolution of Singularities) Let $K(w) \geq 0$ be a real analytic function defined in a neighborhood of $0 \in R^d$. Then there exists an open set $V \subset \mathbf{R}^d$, a d-dimensional real analytic manifold U and a proper analytic map $g : U \to V$ such that*
(1) $g : U \setminus U_0 \to V \setminus V_0$ is an invertible analytic function, where $V_0 = K^{-1}(0)$, $U_0 = g^{-1}(V_0)$.
(2) For each $P \in U$, there are local analytic coordinates $(u_1, u_2, .., u_d)$ centered at P so that, locally near P, we have

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

where $k_i \geq 0$ is nonnegative integers.

Remark. This is one of the most important theorems in algebraic geometry proven by (Hironaka, 1964). This theorem claims that any singularities in $K(w) = 0$ can be transformed into normal crossing ones $K(g(u)) = 0$ by using analytic mapping g. The map g is called *proper* if and only if the inverse $g^{-1}(C)$ of any compact set C is also compact. Based on this theorem, we can restrict singularities as normal crossing ones. However, the parameter space should be a manifold because the resolution of singularities is realized by blowing-ups. Moreover, the manifold is not orientable in general. The nonnegative integers $(k_1, k_2, .., k_d)$ depend on the local coordinate.

The application of the theorem to Schwarz distribution theory was proposed by (Atiyah, 1970). This theorem was also used in the proof of the rational proof of

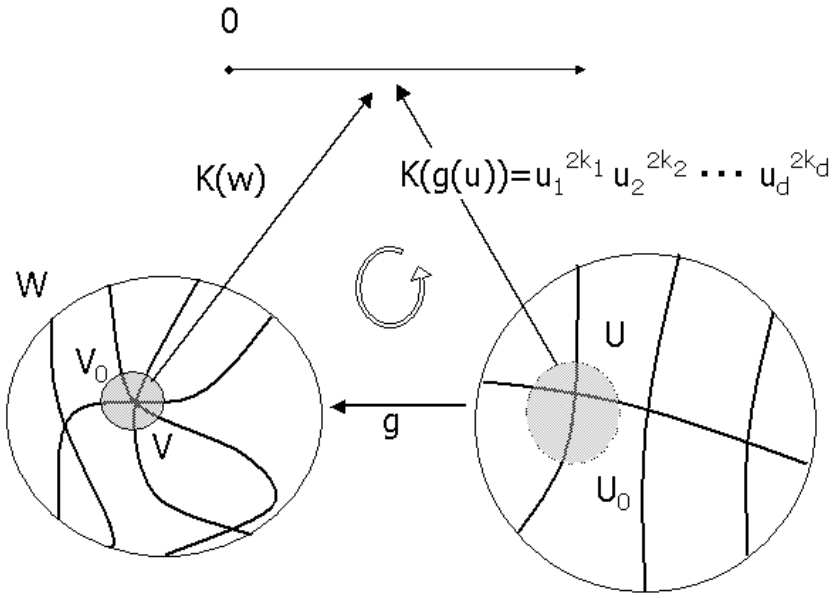

Figure 1: Resolution of Singularities

the b-function (Kashiwara, 1976). In learning theory, it was firstly clarified that the combination of this theorem and the zeta function provides the asymptotic expansion of the Bayes marginal likelihood (Watanabe, 1999; Watanabe,2001a). This paper shows that this theorem also gives an important theory to the maximum likelihood method and the maximum a posteriori method.

Hereafter, we apply the resolution theorem to the Kullback information $K(w)$, resulting that we obtain the real analytic function g and the local coordinates $(u_1, u_2, .., u_d)$.

Theorem 3 *Assume two conditions, (C.1) and (C.2). Then there exists an $L^2(w_0)$-valued analytic function $a(x,u)$ such that locally*

$$h(x, g(u)) = a(x,u)\ u_1^{k_1}\ u_2^{k_2}\ \cdots\ u_d^{k_d}, \tag{1}$$

and

$$\int a(x,u)\, p(x|w_0)\, dx = u_1^{k_1}\, u_2^{k_2}\, \cdots\, u_d^{k_d}. \tag{2}$$

(Proof) For a given $M > 0$, we define a subset $X_M \subset \mathbf{R}^N$ by

$$X_M = \{x \in \mathbf{R}^N; \max_{w \in W} |h(x,w)| \leq M\}.$$

There exists a constant $d_M > 0$ such that

$$|y| \leq M \Longrightarrow e^{-y} + y - 1 \geq d_M\, y^2.$$

Then for arbitrary $x \in X_M$,

$$e^{-h(x,w)} + h(x,w) - 1 \geq d_M\, h(x,w)^2.$$

It follows that

$$\begin{aligned} K(w) &= \int h(x,w)\, p(x|w_0)\, dx \\ &= \int \{e^{-h(x,w)} + h(x,w) - 1\}\, p(x|w_0)\, dx \\ &\geq \int_{X_M} \{e^{-h(x,w)} + h(x,w) - 1\}\, p(x|w_0)\, dx \\ &\geq d_M \int_{X_M} h(x,w)^2\, p(x|w_0)\, dx. \end{aligned}$$

If a parameter w is contained in a neighborhood of W_0, then by using $w = g(u)$, we obtain

$$u_1^{2k_1}\, u_2^{2k_2}\, \cdots u_d^{2k_d} \geq d_M \int_{X_M} h(x,g(u))^2\, p(x|w_0)\, dx.$$

Since $h(x,g(u))$ is an $L^2(w_0)$-valued analytic function, by using the orthonormal system $\{e_j(x)\}$ in the Hilbert space $L^2(w_0)$, we have

$$in the discussion, h(x,g(u)) = \sum_{j=1}^{\infty} e_j(x) f_j(u). \tag{3}$$

Here $f_j(u)$ is given by

$$f_j(u) = \int h(x,g(u)) e_j(u) p(x|w_0) dx,$$

resulting that it is an analytic function of u. Thus

$$u_1^{2k_1}\, u_2^{2k_2}\, \cdots u_d^{2k_d} \geq d_M \sum_{j=1}^{\infty} f_j(u)^2 \geq d_M\, f_j(u)^2.$$

Since $f_j(u)$ is an analytic function of u, there exists an analytic function $f_j^*(u)$ such that

$$f_j(u) = f_j^*(u) u_1^{k_1}\ u_2^{k_2} \cdots u_d^{k_d}.$$

By using the equation (3), it follows that

$$h(x, g(u)) = u_1^{k_1}\ u_2^{k_2} \cdots u_d^{k_d} \sum_{j=1}^{\infty} f_j^*(u) e_j(x),$$

whose convergence is ensured by the convergence of the equation (3). Let $a(x, u)$ be a function

$$a(x, u) = \sum_{j=1}^{\infty} f_j^*(u) e_j(x).$$

The log-likelihood ratio function $h(x, g(u))$ is an $L^2(w_0)$-valued analytic function, also is $a(x, u)$, which proves equation (1). The equation (2) is proved by the definition of $K(w)$. (Q.E.D.)

Remark. Even if two analytic function $\alpha(u), \beta(u)$ of u satisfy

$$0 \leq \alpha(u) \leq \beta(u),$$

the function $\alpha(u)/\beta(u)$ is not an analytic function in general. However, if $\beta(u)$ is normal crossing,

$$\beta(u) = u_1^{2k_1} \cdots u_d^{2k_d},$$

then $\alpha(u)/\beta(u)$ is an analytic function. That is to say, we need the normal crossing property in the proof of the foregoing theorem.

By the resolution of singularities, the Kullback information is locally represented by

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

Let us define a function $k(u)$ by

$$k(u) \equiv u_1^{k_1}\ u_2^{k_2}\ \cdots\ u_d^{k_d},$$

then, based on Theorem 3,

$$\begin{aligned} K(g(u)) &= k(u)^2, \\ h(x, g(u)) &= k(u) a(x, u). \end{aligned}$$

Theorem 4 *For arbitrary u, the function $a(x,u)$ has the following property.*

$$\int a(x,u)^2\ p(x|w_0)\ dx = 2.$$

(Proof) For $0 \leq t \leq 1$, we define an $L^2(w_0)$-valued analytic function $F(t,x,u)$

$$F(t,x,u) \equiv e^{-th(x,w)} + th(x,w) - 1.$$

Then

$$F(0,x,u) = \frac{\partial}{\partial t}F(0,x,u) = 0.$$

There exists t^* (U0 $\leq t^* \leq 1$) such that

$$\begin{aligned} F(1,x,u) &= \frac{1}{2}\frac{\partial^2}{\partial t^2}F''(t^*,x,u), \\ &= \frac{h(x,w)^2}{2}e^{-t^*h(x,w)}. \end{aligned}$$

Hence

$$\begin{aligned} K(g(u)) &= \int \{e^{-h(x,g(u))} + h(x,g(u)) - 1\}\ p(x|w_0)\ dx \\ &= \frac{1}{2}\int h(x,g(u))^2\ e^{-t^*h(x,g(u))}\ p(x|w_0)\ dx. \end{aligned}$$

Dividing this equation by $u_1^{2k_1}\cdots u_d^{2k_d}$ and $u_j \to 0$, we obtain

$$1 = \frac{1}{2}\int a(x,u)^2\ p(x|w_0)\ dx,$$

which completes the Theorem. (Q.E.D.)

5 Likelihood Function and Empirical Process

By using resolution of singularities, $R(w)$ is locally represented by

$$\begin{aligned} R(g(u)) &= \sum_{i=1}^{n} k(u)a(X_i,u) \\ &= n\ k(u)^2 + \sqrt{n}\ k(u)\ \psi_n(u), \end{aligned}$$

where

$$\psi_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{a(X_i,u) - k(u)\},$$

is an empirical process. Since $a(x,u)$ is an $L^2(w_0)$-valued analytic function and W is compact, $\psi_n(w)$ weakly converges to the gaussian process $\psi(w)$.

Remark. Since

$$\int a(x,u)\, p(x|w_0)\, dx = k(u),$$

it follows that $E_{X^n}[\psi_n(u)] \equiv 0$. Moreover,

$$\int a(x,u)^2\, p(x|w_0)\, dx = 2.$$

Hence, $\psi_n(x,u)$ converges to a normal distribution for every fixed u by the central limit theorem. Let $B(W)$ be the Banach space of all bounded function on W, which is complete by the norm,

$$\|\varphi\| = \sup_{w\in W} |\varphi(w)|.$$

In $B(W)$, the family of measurable sets is defined as the smallest σ-algebra that contains all open set in $B(W)$. Then $B(W)$ is a measurable space. The gaussian process $\psi(u)$ is uniquely defined by the expectation and covariance

$$\begin{aligned} E_\psi[\psi(u)] &= 0, \\ E_\psi[\psi(u_1)\psi(u_2)] &= E_X[(a(X,u_1)-k(u_1))(a(X,u_2)-k(u_2))]. \end{aligned}$$

The empirical process $\psi_n(u)$ and the gaussian process $\psi(u)$ are random variable on $B(W)$ and naturally define probability measures on $B(W)$. The weak convergence $\psi_n(u) \to \psi(u)$ is defined by

$$\lim_{n\to\infty} E_{X^n}[F(\psi_n)] = E_\psi[F(\psi)]$$

for arbitrary bounded and continuous map

$$F : B(W) \ni \varphi \mapsto F(\varphi) \in \mathbf{R}^1.$$

For the definition and the proof of the weak convergence, see (Aad W.van der Vaart & Jon Wellner, 1996). The inequality

$$E_{X^n}\|\psi_n(w)\|^2 < \infty, \tag{4}$$

can also be proved directly (Watanabe, 2001a). We obtained the following theorem.

Theorem 5 *There exists a parameterization $u = (u_1, u_2, .., u_d)$ in which the log-likelihood ratio function $R(w)$ can locally be represented by*

$$R(g(u)) = nk(u)^2 + \sqrt{n}k(u)\psi_n(u),$$

where

$$k(u) = u_1^{k_1}\ u_2^{k_2}\ \cdots\ u_d^{k_2},$$

and the weak convergence

$$\psi_n(u) \to \psi(u)$$

holds.

In this paper, we apply this theorem to the proof of the main theorem. This theorem might be useful in the more general learning theory, for example, hypothesis testing, and so on.

6 Proof of the Main Theorem

Let us prove the main theorem, Theorem 1, based on the results in the sections 4, 5, and 6. Instead of the parametrization $u \in \mathbf{R}^d$, let us introduce a new coordinate (t, s), where $t = k(u)$ and $s \in \mathbf{R}^{d-1}$. If one of $k_1, k_2, .., k_d$ is odd, then there exists a constant $\epsilon > 0$ such that t takes all value in

$$-\epsilon < t < \epsilon.$$

If otherwise then

$$0 \leq t < \epsilon.$$

Example. For example, if

$$k(u) = u_1^{k_1} u_2^{k_2} u_3^{k_3}$$

then one can choose a coordinate,

$$\begin{aligned} t &= k(u), \\ s_1 &= k_2 u_2^2 - k_1 u_1^2, \\ s_2 &= k_3 u_3^2 - k_1 u_1^2, \end{aligned}$$

in which the plane $t = const.$ is orthogonal to the line $s_1 = const., s_2 = const.$

Let us define a loss function $L_0(w)$ by

$$L_0(w) = R(w) + \lambda_n f(w).$$

Then minimization of $L_0(w)$ is equivalent to minimization of $L(w)$. By using Theorem 5, The function $L_0(w)$ is represented by

$$L_0(t, s) = nt^2 + \sqrt{n}\ t\ \psi_n(t, s) + \lambda_n f(t, s).$$

There exists t^* $(0 \leq |t^*| \leq |t|)$ such that

$$\begin{aligned} L_0(t, s) &= n\ t^2 + \sqrt{n}\ t\ \psi_n(0, s) + \lambda_n f(0, s) + F(t, s), \\ F(t, s) &= \sqrt{n} t^2 \partial_t \psi_n(t^*, s) + \lambda_n t \partial_t f(t^*, s). \end{aligned}$$

Firstly we consider the case

$$-\epsilon < t < \epsilon.$$

By putting $t = T/\sqrt{n}$,

$$L_0(\frac{T}{\sqrt{n}}, s) = T^2 + T\psi_n(0, s) + \lambda_n f(0, s) + F'(T, s),$$

where

$$F'(T, s) = \frac{T^2}{\sqrt{n}} \partial_t \psi_n(t^*, s) + \frac{\lambda_n T}{\sqrt{n}} \partial_t f(t^*, s).$$

Hence if $L_0(T/\sqrt{n}, s)$ is minimized at $(\hat{T}, \hat{s})$, then

$$\hat{T} = -\frac{\psi_n(0, \hat{s})}{2} + o_p(1).$$

Equivalently, if $L_0(t, s)$ is minimized at $(\hat{t}, \hat{s})$, then

$$\hat{t} = -\frac{\psi_n(0, \hat{s})}{2\sqrt{n}} + o_p(\frac{1}{\sqrt{n}}), \tag{5}$$

where $o_p(\frac{1}{\sqrt{n}})$ shows a random variable which satisfies the convergence in probability

$$\sqrt{n}\ o_p(\frac{1}{\sqrt{n}}) \to 0,$$

when n tends to zero. This convergence is easily shown by the fact that the closure of U is compact. (Since g is a proper mapping and W is compact, the closure of U is also compact). Equation (5) shows that

$$R(\hat{w}) = -\frac{|\psi_n(0,\hat{s})|^2}{4n} + o_p(\frac{1}{n}) \quad (6)$$

$$K(\hat{w}) = \frac{|\psi_n(0,\hat{s})|^2}{4n} + o_p(\frac{1}{n}). \quad (7)$$

Secondly, let us consider the case

$$0 \leq t < \epsilon.$$

If there exists s' such that

$$\psi_n(0,s') < 0,$$

then the same relations as equations (6) and (7) holds. If

$$\psi_n(0,s') \geq 0$$

for all s', then in the coordinate $(u_1, u_d, ..., u_d)$,

$$R(\hat{w}) = o_p(\frac{1}{n}) \quad (8)$$

$$K(\hat{w}) = o_p(\frac{1}{n}). \quad (9)$$

The equations (6),(7),(8),(9) indicates that the symmetry holds in both cases.

$$E_{X^n}[\sup_{w\in W} |\psi_n(w)|^2] < \infty$$

ensures that the learning coefficient is finite, which completes the main theorem. (Q.E.D.)

Remark. The set W_0 is not one point. The optimal parameter $\hat{w}$ may be distributed on a union of many local neighborhoods of W_0. The optimal $\hat{s}$ is determined by

$$\hat{s} = \arg\min_s \left[-\frac{|\psi_n(0,s)|^2}{4} + \lambda_n f(0,s). \right].$$

The place of $(\hat{t},\hat{s})$ strongly depends on λ_n, $\pi(w)$, and training samples. The main point of the proof is that the symmetry of generalization and training errors holds without respect to the place where $\hat{w}$ is distributed.

7 Discussion

In Bayesian estimation, it is known that the symmetry of generalization and training errors holds in regular learning machines, whereas it does not hold in singular learning machines. This paper firstly shows the symmetry holds in both the maximum likelihood method and the maximum a posteriori method.

7.1 Noncompact Case

In this paper, we study the case that the set of parameters W is compact. If W is not compact, then the maximum likelihood estimator may not exist, or even if it exists, the equation (4) does not hold in general (Hartigan, 1985). Theoretically speaking, it is important to clarify such cases. Based on the result of this paper, the phenomenon in the noncompact case might be conjectured as follows. The training error is smaller than the compact case, whereas the generalization error is larger than the compact case.

On the other hand, it has been proven that the generalization error of the Bayes estimation is far smaller than that of the maximum likelihood method or the maximum a posteriori method (Watanabe,2001a;Watanabe, 2001b;Watanabe & Amari, 2003). In singular learning machines, the a posteriori distribution does not converge to the normal distribution, resulting that one point estimator can not be the sufficient statistic even in the asymptotic case. It is the future study to clarify the relation between the maximum likelihood method and the Bayes estimation, and to construct the efficient sets of estimators.

7.2 Learning Coefficient

Let us consider the learning coefficient. In this subsection, we assume that, in all local coordinates U, there exists an odd number in $k_1, k_2, ..., k_d$.

Fisrtly, let us study the case

$$\limsup_{n\to\infty} \lambda_n \leq const.$$

This case includes the maximum likelihood method and the maximum a positerior

method. Then the learning coefficient μ in the symmetry is equal to

$$\mu = E_{\psi}[|\psi(\hat{u})|^2],$$

where

$$\hat{u} = \arg\min_{u \in U_0}\{-\frac{|\psi(u)|^2}{4} + \lambda_n f(g(u))\},$$

where $\psi(u)$ on U_0 is the gaussian process whose expectation vaule and covariance matrix are respectively given by zero and

$$E_{\psi}[\psi(u_1)\psi(u_2)] = E_X[a(X,u_1)a(X,u_2)].$$

Secondly, we study the case

$$\lim_{n\to\infty} \lambda_n = \infty$$

and

$$\lim_{n\to\infty} \lambda_n/\sqrt{n} = 0.$$

We define the set U_{00} of all parameters which attain the miminum of $f(g(u))$,

$$U_{00} = \{u \in U_0; f(g(u)) = \min_{u' \in U_0} f(g(u'))\}.$$

Since U and g is made by blowing-up, $g(u)$ is not a one-to-one mapping, resulting that U_{00} does not consist of one point. The learning coefficient is given by

$$\mu = E_{\psi}[\max_{u \in U_{00}} |\psi(u)|^2].$$

If the function f is defined on not W but U such that U_{00} consists of one point, then the learning coefficient is $\mu = 1/2$. However, $\mu > 1/2$ in general. Although the coefficient μ is determined by the maximum value of the gaussian distribution on the analytic set U_0, it can not be represented by any simple number in general. If $w \in W \setminus W_0$, then

$$a(X, g^{-1}(w)) = \frac{h(x, g^{-1}(w))}{\sqrt{K(g^{-1}(w))}},$$

which is not well defined if $w \in W_0$. Hence it is difficult to numerically construct $a(X, g^{-1}(w))$ in the parameter space W. In order to numerically estimate μ, we need to construct the gaussian process on U_{00}. The resolution process can be realized by using blowing-ups. The resolution results can be seen (Watanabe,2001a;

Rusakov&Geiger, 2002; Yamazaki &Watanabe,2003). For testing hypothesis in some concrete learning machine, the advanced algebraic geometrical method will be needed.

7.3 Construction of Learning Algorithms

In this paper, we have shown the symmetry of the generalization and training errors. In singular learning machines, the smaller training error does not always make the smaller generalization error. It is difficult to find the optimal parameter which minimizes the training error. However, even if we find the optimal parameter, it does not always result in the precise prediction. When we construct the training algorithm of heirarchical learning machines, we have to pay some attention to this point.

8 Conclusion

In this paper, we study the maximum likelihood method and the maximum a posteriori method, and show a sufficiently condition of the symmetry of the generalization and training errors. The resolution of singularities enables us to analyze the singular learning machine in a local coordinate where every singularities are normal crossing one. The symmetry holds under the natural conditions even in singular learning machines.

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

References

Amari,S., Park,H., and Ozeki,T. (2002) Geometrical Singularities in the Neuromanifold of Multilayer Perceptrons. *Advances in Neural Information Processing Systems*, 14, to appear.

Atiyah, M. F. (1970). Resolution of singularities and division of distributions. Communications of Pure and Applied Mathematics, 13, 145-150.

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

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. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 2, 807-810.

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. Annals of Mathematics, 79, 109-326.

Kashiwara, M. (1976). B-functions and holonomic systems. Inventiones Mathematicae, 38, 33-53.

Rusakov, D. & Geiger, D. (2002) Asymptotic model selection for naive Bayesian networks,*Proceedings of UAI'02*, to appear.

van der Vaart, A.W,&Weller,J (1996) Weak Convergence and Empirical Processes. Springer.

Watanabe,S.(1995) A generalized Bayesian framework for neural networks with singular Fisher information matrices.*International Symposium on Nonlinear Theory and Its Applications*,2,207-210.

Watanabe, S. (1999b). Algebraic analysis for singular statistical estimation. Lecture Notes in Computer Science, 1720, 39-50.

Watanabe, S. (2000). Algebraic analysis for non-regular learning machines. Advances in Neural Information Processing, 12, 356-362.

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

Watanabe, S. (2001b). Training and generalization errors of the hierarchical learning machines with algebraic singularities. IEICE Transactions, J84A (1), 99-108.

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

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