A Formula of Equations of States in Singular Learning Machines

Sumio Watanabe, *Senior Member, IEEE*

***Abstract*— Almost all learning machines used in computational intelligence are not regular but singular statistical models, because they are nonidentifiable and their Fisher information matrices are singular. In singular learning machines, neither the Bayes a posteriori distribution converges to the normal distribution nor the maximum likelihood estimator satisfies the asymptotic normality, resulting that it has been difficult to estimate generalization performances. In this paper, we establish a formula of equations of states which holds among Bayes and Gibbs generalization and training errors, and show that two generalization errors can be estimated from two training errors. The equations of states proved in this paper hold for any true distribution, any learning machine, and a priori distribution, and any singularities, hence they define widely applicable information criteria.**

I. Introduction

A. Background

A lot of learning machines used in computer science and brain informatics are not regular but singular statistical models. A model is called regular if the mapping from the parameter to the probability distribution is one-to-one and if Fisher information matrix is always positive definite. Almost all learning machines employed in computational intelligence have hierarchical structures or hidden variables, hence they are singular learning machines. In regular statistical models, Bayes a posteriori distribution converges to the normal distribution and the maximum likelihood estimator satisfies asymptotic normality. Whereas, in singular learning machines, Bayes a posteriori distribution converges to the singular distribution [17] and the maximum likelihood estimator diverges to infinity [6], [5], [7]. Singularities in the parameter space strongly affect learning dynamics [1]. These are universal phenomena in singular learning machines, which prevent us from analyzing their generalization performances.

Recently, we established an algebraic geometrical method for singular learning machines [18], [19], [20], [21], [22]. Based on resolution of singularities, we proved that Beyes marginal likelihood is determined by the largest pole of zeta function. It was shown by these results that the generalization errors of singular learning machines depend on singularities. In practical applications, the true distribution is unknown, hence it has been difficult to estimate the singularities of the true set of parameters.

Sumio Watanabe is with PI Lab. in Tokyo Institute of Technology. Sumio Watanabe, Mailbox R2-5, 4259 Nagatuda, Midori-ku, Yokohama, 226-8503, Japan

This work was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

B. Equations of States in Learning

In this paper, we derive a formula which holds among four errors, Bayes generalization B_g, Bayes training B_t, Gibbs generalization G_g, and Gibbs training G_t. The equations of states proved in this paper are

$$\begin{aligned} E[B_g] - E[B_t] &= E[G_g] - E[G_t] \\ &= 2\beta(E[G_t] - E[B_t]). \end{aligned}$$

The formula holds for arbitrary true distribution, arbitrary learning machine, arbitrary a priori distribution, and arbitrary singularities. Hence we can estimate Bayes and Gibbs generalization errors from Bayes and Gibbs training errors. In other words, we can construct widely applicable information criteria which can be used in both regular and singular learning machines.

C. Examples of Singular Learning Machines

The following learning machines are not regular but singular statistical models. (1) layered neural networks, (2) radial basis functions, (3) normal mixtures, (4) binomial mixtures, (5) reduced rank regressions, (6) Boltzmann machines, (7) Bayes networks, and (8) hidden Markov models. Almost all learning machines are singular [26]. In these learning machines, if a learning machine is smaller compared with the true distribution, then the set of true parameter is an analytic set with singularities. In singular learning machine, AIC does not correspond to the average prediction error, and BIC does not equal to the asymptotic evidence. Hence in order to establish a mathematical foundation for model selection and hypothesis testing, we need singular learning theory.

II. Main Results

Let $q(x)$ be a probability density function on N dimensional Euclidean space $\mathbf{R}^N$ and X be a random variable which is subject to $q(x)$. Also Let $X_1, X_2, ..., X_n$ be random variables which are independently subject to the same probability distribution as X. In learning theory, $q(x)$ and $X_1, X_2, ..., X_n$ are respectively called the true distribution and a set of training samples.

A learning machine is defined by a probability distribution $p(x|w)$ of $x \in \mathbf{R}^n$ for a given parameter $w \in \mathbf{R}^d$. A probability density function $\varphi(w)$ is also defined on $\mathbf{R}^d$, which is called an a priori distribution. The a posteriori distribution $\rho(w)$ is defined by

$$\rho(w) = \frac{1}{Z}\,\varphi(w)\,\Big(\prod_{i=1}^{n} p(X_i|w)\Big)^{\beta},$$

where $\beta > 0$ is an inverse temperature. Let $E_w[\cdot]$ be the expectation value by this probability distribution. We define

four errors.
(1) Bayes generalization error.

$$B_g = E_X[\log \frac{q(X)}{E_w p(X|w)}].$$

(2) Bayes training error.

$$B_t = \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{E_w p(X_i|w)}.$$

(3) Gibbs generalization error.

$$G_g = E_w E_X[\log \frac{q(X)}{p(X|w)}].$$

(4) Gibbs training error.

$$G_t = E_w \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)}.$$

We need the mathematical assumptions which ensure the theorems. Let us define a log density ratio function by

$$f(x,w) = \log \frac{q(x)}{p(x|w)}.$$

In this paper, we assume the following three conditions.
(A.1) Assume that the set of parameters W is a a compact set which is a closure of an open set in $\mathbf{R}^d$. The set W is defined by

$$W = \{w \in \mathbf{R}^d; \pi_1(w) \geq 0, \cdots \pi_k(w) \geq 0\},$$

where $\pi_1(w), \cdots, \pi_k(w)$ are analytic functions, and the *a priori* probability density $\varphi(w)$ is given by $\varphi(w) = \varphi_0(w)\varphi_1(w)$ where $\varphi_0(w) > 0$ is a C^∞-class function and $\varphi_1(w) \geq 0$ is an analytic function.
(A.2) Let $s \geq 9/2$ be a constant, and $L^s(q)$ be the complex Banach space defined by

$$L^s(q) = \{f(x) \ ; \ \int |f(x)|^s q(x)dx < \infty\}.$$

Assume that there exists an open set $W' \subset \mathbf{C}^d$ which contains W such that a function $W' \ni w \mapsto f(\cdot, w)$ is an $L^s(q)$ valued analytic function.
(A.3) Let $W_0 = \{w \in W \ ; \ q(x) = p(x|w)\}$ be the set of true parameters. The set W_0 is not the empty set and there exists an open set $W^* \subset \mathbf{C}^d$ which contains W such that

$$E_X[\sup_{w \in W*} |f(X,w)|^s)] < \infty.$$

Remark. These assumptions are need by the mathematical reasons.
(1) These conditions allow the case that the set of true parameters $W_0 = \{w \in W; q(x) = p(x|w)\}$ is not one point but an algebraic set or an analytic set with singularities, and that Fisher information matrix has zero eigen values.
(2) The condition that W is compact is necessary because, even if the log density ratio function is an analytic function of the parameter, $|w| = \infty$ is singularity in general. By this reason, If W is not compact and W_0 contains $|w| = \infty$, the maximum likelihood estimator does not exist in general. In fact, if $x = (x_1, x_2)$, $w = (a, b)$, and $f(x,w) = (x_2 - a\sin(bx_1))^2/2$, and W_0 contains $\{a = 0\}$, then the maximum likelihood estimator never exists. On the ther hand, if $|w| = \infty$ is not singularity, $\mathbf{R}^d \cup \{|w| = \infty\}$ can be understood as a compact set and the same theorems as this paper hold.

Based on the assumptions (A.1),(A.2), and (A.3), we show Theorems.

Theorem *1:* (1) There exist random variables B_g^*, B_t^*, G_g^*, and G_t^* such that, when $n \to \infty$, the following convergences in law hold.

$$nB_g \to B_g^*, \qquad nB_t \to B_t^*,$$
$$nG_g \to G_g^*, \qquad nG_t \to G_t^*.$$

(2) When $n \to \infty$, the convergence in probability holds,

$$n(B_g - B_t - G_g + G_t) \to 0.$$

(3) Expectation values of four errors converge,

$$E[nB_g] \to E[B_g^*], \qquad E[nB_t] \to E[B_t^*],$$
$$E[nG_g] \to E[G_g^*], \qquad E[nG_t] \to E[G_t^*].$$

For the proof of this theorem, see section IV. The following Theorem is the main formula of this paper.

Theorem *2:* (**Equations of States in Statistical Estimation**). The following equations hold.

$$\begin{aligned} E[B_g^*] - E[B_t^*] &= E[G_g^*] - E[G_t^*] \\ &= 2\beta(E[G_t^*] - E[B_t^*]). \end{aligned} \tag{1}$$

Remark. (1) Theorem 2 shows the increases of errors from training to prediction are in proportion to the difference between Bayes and Gibbs training. We call Theorem 2 as **Equations of States in Statistical Estimation**, because they hold for any true distribution, any learning machine, any *a priori* distribution, and any singularities.
(2) Although the equations of states hold universally, the four errors themselves strongly depend on a true distribution, a learning machine, an *a priori* distribution, and singularities.
(3) Theorem 2 also shows the conservation law that the difference from the Bayes error to Gibbs error is invariant between training and generalization,

$$E[G_g^*] - E[B_g^*] = E[G_t^*] - E[B_t^*]. \tag{2}$$

As is shown in Theorem 1, this conservation law holds not only as expectations, but also as random variables, when the number of training samples tends to infinity.

Corollary *1:* The two generalization errors can be estimated by the two training errors,

$$\begin{pmatrix} E[B_g^*] \\ E[G_g^*] \end{pmatrix} = \begin{pmatrix} 1-2\beta & 2\beta \\ -2\beta & 1+2\beta \end{pmatrix} \begin{pmatrix} E[B_t^*] \\ E[G_t^*] \end{pmatrix}. \tag{3}$$

Remark. (1) From eq.(3), it follows that

$$\begin{pmatrix} E[G_t^*] \\ E[B_t^*] \end{pmatrix} = \begin{pmatrix} 1-2\beta & 2\beta \\ -2\beta & 1+2\beta \end{pmatrix} \begin{pmatrix} E[G_g^*] \\ E[B_g^*] \end{pmatrix},$$

which shows that there is a symmetry between generalization errors and training errors.

(2) A statistical model is called *regular* if the set of true parameters $W_0 = \{w \in W; q(x) = p(x|w)\}$ consists of one point and if Fisher Information matrix is always positive definite. Note that a regular model is a very special example of singular learning machines. For a regular statistical model, we have

$$E[B_g^*] = \frac{d}{2}, \qquad E[G_g^*] = (1 + \frac{1}{\beta})\frac{d}{2},$$
$$E[B_t^*] = -\frac{d}{2}, \qquad E[G_t^*] = (-1 + \frac{1}{\beta})\frac{d}{2},$$

which is a special case of Theorem 2.

Theorem 2 reveals the universal relations among four errors. It holds even if the set of true parameters has complex singularities. However, its fact simultaneously shows that we can extract no information of singularities directly from Theorem 2. Theorem 3 shows that four errors have important information about singularities. The Kullback-Leibler information is

$$K(w) = E_X[f(X, w)] = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

The *zeta function* of a learning machine is defined by

$$\zeta(z) = \int_W K(w)^z \ \varphi(w) \ dw. \tag{4}$$

The zeta function is a holomorphic function of a complex variable z in the region $Re(z) > 0$, which can be analytically continued to a meromorphic function on the entire complex plane. Its poles are all real, negative, and rational numbers (for the proof, see [4][9][22]). They are referred to as

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

The order of each pole λ_k is denoted by m_k. We simply use notations $\lambda = \lambda_1$ and $m = m_1$ for the largest pole and its order respectively.

Theorem *3:* (1) When $n \to \infty$, the convergence in probability

$$nG_g + nG_t - \frac{2\lambda}{\beta} \to 0$$

holds. Therefore

$$E[G_g^*] + E[G_t^*] = \frac{2\lambda}{\beta}. \tag{5}$$

Also the following corollary holds.

Corollary *2:* The following convergence in probability holds,

$$nB_g - nB_t + 2nG_t - \frac{2\lambda}{\beta} \to 0.$$

In particular, if $\beta = 1$, $E[B_g^*] = \lambda$.

By this theorem and corollary, if one knows the true distribution, one can predict Bayes and Gibbs generalization errors from Bayes and Gibbs training errors with probability one, when n tends to infinity. In practical applications, we seldom know the true distribution, however, this fact is useful in computer simulation research of learning theory and statistics. Corollary 2 was firstly discovered in [18][20]. Since the constant λ strongly depends on the true distribution, the learning machine, and the *a priori* distribution, it characterizes the properties of learning machines. The values of several models have been studied in neural networks [21], normal mixtures [30], reduced rank regressions [2], Boltzmann machines [31], hidden Markov models [32]. Also the behavior of λ was analyzed in the case when Jeffreys' prior is employed as an *a priori* distribution [19], and in the case when the distance of the true distribution from the singularity is in proportion to $1/\sqrt{n}$ [23].

III. Widely Applicable Information Criteria

The main purpose of this paper is to prove the theorems. However, in order to illustrate the reason why the results of this paper are important, we propose widely information criteria and introduce an experiment. Experimental analysis in practical applications is a future study.

A. Basic Concepts

Based on Corollary 1, we establish new information criteria which can be used in both regular and singular learning machines. Let us define Bayes generalization loss, Bayes training loss, Gibbs generalization loss, and Gibbs training loss by

$$\begin{aligned} BL_g &= E_X[-\log E_w[p(X|w)]], \\ BL_t &= \frac{1}{n}\sum_{j=1}^{n} -\log E_w[p(X_j|w)], \\ GL_g &= E_w E_X[-\log p(X|w)], \\ GL_t &= E_w[\frac{1}{n}\sum_{j=1}^{n} -\log p(X_j|w)]. \end{aligned}$$

These losses are random variables. Both training losses BL_t and GL_t can be numerically calculated based on training samples D_n and a learning machine $p(x|w)$ without any knowledge of the true density function $q(x)$. By adding the entropy of the true distribution to Corollary 1

$$S = -\int q(x) \log q(x) dx = E\Big[\frac{1}{n}\sum_{i=1}^{n} -\log q(X_i)\Big],$$

we obtain the equations,

$$\begin{aligned} E[BL_g] &= 2\beta E[GL_t] + (1 - 2\beta)E[BL_t] + o(\frac{1}{n}), \\ E[GL_g] &= (1 + 2\beta)E[GL_t] - 2\beta E[BL_t] + o(\frac{1}{n}). \end{aligned}$$

Let us define widely applicable information criteria (WAIC) by

$$\begin{aligned} WAIC_1 &= 2\beta \ GL_t + (1 - 2\beta) \ BL_t, \\ WAIC_2 &= (1 + 2\beta) \ GL_t - 2\beta \ BL_t. \end{aligned}$$

Then expectations of two criteria respectively equal to the Bayes and Gibbs generalization losses,

$$E[BL_g] = E[WAIC_1] + o(\frac{1}{n}),$$
$$E[GL_g] = E[WAIC_2] + o(\frac{1}{n}).$$

Therefore, $WAIC_1$ and $WAIC_2$ give the indices for model evaluation.

Remark. If a model is regular,

$$2\beta(E[G_t^*] - E[B_t^*]) = d, \tag{6}$$

holds. When $\beta \to \infty$, both Bayes and Gibbs estimations result in the maximum likelihood method, hence eq.(1) gives the well-known information criterion AIC. In other words, WAIC can be understood as generalized information criteria of AIC for singular learning machines. In singular learning machines, eq.(6) does not hold, however, the symmetry of generalization and training errors also holds when the parameter set is compact [24].

B. Experiments

We studied reduced rank regressions. The input and output vector is $x = (x_1, x_2) \in \mathbf{R}^{N_1} \times \mathbf{R}^{N_2}$ and the parameter is $w = (A, B)$ where A and B are respectively $N_1 \times H$ and $H \times N_2$ matrices. The learning machine is

$$p(x|w) = q(x_1)\frac{1}{(2\pi\sigma^2)^{N_2/2}} \exp(-\frac{1}{2\sigma^2}\|x_2 - BAx_1\|^2).$$

Since $q(x_1)$ has no parameter, it is not estimated. The true distribution is determined by matrices A_0 and B_0 such that $\mathrm{rank}(B_0A_0) = H_0$. The algebraic variety of the true parameters $K(A, B) = 0$, where

$$K(A, B) \propto \|BA - B_0A_0\|^2,$$

has complicated singularities. We conducted experiments in a case when $N_1 = N_2 = 6$, $H_0 = 3$, $\beta = 1$, $n = 1000$, and $\sigma = 0.1$. The *a priori* distribution was $p(A, B) \propto \exp(-2.0 \cdot 10^{-5}(\|A\|^2 + \|B\|^2))$. Reduced rank regressions with hidden units $H = 1, 2, .., 6$ were employed. The *a posteriori* distribution was numerically approximated by the Metropolis method, where initial 5000 steps were omitted and 2000 parameters were collected after every 200 steps. Expectation values B_g, $WAIC_1$, E_g, $WAIC_2$ were averaged by 20 trials, that is to say, 20 sets of training samples were independently taken from the true distribution. Theoretical values of $E[B_g]$ for $\beta = 1$ were given in [2]. The results in table.1 show why the results of this paper are important.

H	Theory	B_g	$WAIC_1$	G_g	$WAIC_2$
1		1.677973	1.668674	1.688998	1.679448
2		0.826303	0.797272	0.853480	0.823484
3	0.013500	0.012696	0.014204	0.026243	0.027413
4	0.015000	0.014491	0.015340	0.030163	0.030554
5	0.016000	0.015600	0.016078	0.031975	0.032011
6	0.017000	0.016481	0.016687	0.033334	0.033048

TABLE I
EXPERIMENTAL REULTS

IV. SINGULAR LEARNING THEORY

In this paper, we show outline of theorems. For the mathematically rigorous proof, see [28].

A. Resolution of Singularities

The main region in the parameter set to be studied is

$$W_\epsilon = \{w \in W \ ; \ K(w) \le \epsilon\}$$

for a sufficiently small $\epsilon > 0$. The set $W \setminus W_\epsilon$ does not affect the asymptotic behaviors [28]. By applying the Hironaka desingularitzation theorem [8] to $K(w)(\epsilon - K(w))\varphi_1(w)\pi_1(w)\cdots\pi_k(w)$, there exist a manifold $\mathcal{M} = \cup_\alpha U_\alpha$ where U_α is a local coordinate and a proper analytic map $g : U_\alpha \to W_\epsilon$, written by $w = g(u)$, such that in each U_α, functions $K(w)$, $(\epsilon - K(w))$, $\varphi_1(w)$, $\pi_1(w)$, $\cdots$, and $\pi_k(w)$ are all normal crossing. That is to say,

$$K(g(u)) = u^{2k} = \prod_{j=1}^{d} u_j^{2k_j},$$

and

$$\varphi(g(u))|g'(u)| = b(u)|u^h| = b(u)|\prod_{j=1}^{d} u_j^{h_j}|,$$

where $|g'(u)|$ is Jacobian, $h = (k_1, k_2, ..., k_d)$ and $k = (h_1, h_2, .., h_d)$ are sets of nonnegative integers, and $b(u) > 0$ is a C^∞ class function. Note that $g(u)$, k, and h depend on the local coordinate U_α, however, for simple notations, we omit α that identifies the local coordinate. By applying the partition of unity to $\mathcal{M}$, we can assume that $g^{-1}(W)$ is the union of coordinates $[0, 1]^d$ and that

$$\varphi(g(u))|g'(u)| = u^h \ \psi(u),$$

where $\psi(u) > 0$ is a C^∞ class function. Existence of such a manifold $\mathcal{M}$ and an analytic map $w = g(u)$ is well known in algebraic geometry [10], algebraic analysis[4], [9], and learning theory [20]. Since W is compact and g is a proper map, $g^{-1}(W)$ is also compact. For our purpose, we need only the compact subset $g^{-1}(W)$ in $\mathcal{M}$. Therefore, hereafter we use notation $g^{-1}(W) = \mathcal{M}$, which is a compact subset of the manifold. The set of true parameters is denoted by $W_0 = \{w \in W \ ; \ K(w) = 0\}$ and $\mathcal{M}_0 = \{u \in \mathcal{M} \ ; \ K(g(u)) = 0\}$.

Let us define the supremum norm by

$$\|f\| = \sup_{u \in \mathcal{M}} |f(u)|.$$

Then we have a standard form of the log density ratio function.

Lemma *1:* There exists an $L^s(q)$ valued analytic function $\mathcal{M} \ni u \mapsto a(x,u) \in L^s(q)$ such that

$$\begin{aligned} f(x,g(u)) &= a(x,u)\ u^k, \\ E_X[a(X,u)] &= u^k, \\ K(g(u))=0 &\Rightarrow E_X[a(X,u)^2]=2, \\ E_X[\|a(X)\|^s] &< \infty. \end{aligned}$$

This lemma shows that, if there are only normal crossing singularities in the parameter set, the ideal generated by the set of true parameters is trivial, resulting that the log density ratio function is also trivial. For the proof of this lemma, see [28]. We define $\|a(X)\| = \sup_{u\in\mathcal{M}} |a(X,u)|$.

B. Empirical Processes

An empirical process $\xi_n(u)$ is defined by

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a^*(X_i,u)$$

where $a^*(x,u) = E_X[a(X,u)] - a(x,u)$. Then the empirical process satisfies the following Lemma.

Lemma *2:* The empirical process satisfies

$$\begin{aligned} E[\|\xi_n\|^4] &< Const. < \infty \\ E[\|\nabla\xi_n\|^4] &< Const. < \infty \end{aligned}$$

where $Const.$ does not depend on n, and $\|\nabla\xi_n\| = \sum_{j=1}^{d} \|\partial_j \xi_n\|$.
Let the Banach space of the uniformly bounded and continuous functions on $\mathcal{M}$ be

$$B(\mathcal{M}) = \{f(u)\ ;\ \|f\| < \infty\}.$$

Since $\mathcal{M}$ is compact, $B(\mathcal{M})$ is a separable norm space. It was proved in [24] that the empirical process $\xi_n(u)$ defined on $B(\mathcal{M})$ weakly converges to the tight gaussian process $\xi(u)$ that satisfies

$$\begin{aligned} E_\xi[\xi(u)] &= 0, \\ E_\xi[\xi(u)\xi(v)] &= E_X[a^*(X,u)a^*(X,v)]. \end{aligned}$$

If $u, v \in \mathcal{M}_0$,

$$E_X[a^*(X,u)a^*(X,v)] = E_X[a(X,u)a(X,v)].$$

It is well known that a tight gaussian process is uniquely determined by its expectation and covariance matrix of finite points. In a singular learning machine, Fisher information matrix is singular, however, $E_X[a(X,u)a(X,v)]$ can be understood as the generalized concept of the Fisher information matrix.

Let $\xi(u)$ be an arbitrary differentiable function. We define the average $f(u)$ over $\mathcal{M}$ for $\xi(u)$ by

$$E_u^\sigma[f(u)|\xi] = \frac{\sum_\alpha \int_{[0,1]^d} f(u)\ Z(u,\xi)\ du}{\sum_\alpha \int_{[0,1]^d} Z(u,\xi)\ du},$$

where $\sum_\alpha$ is the summation over all coordinates of $\mathcal{M}$, $0 \le \sigma \le 1$, and

$$Z(u,\xi) = u^h\ \psi(u)\ e^{-\beta n u^{2k} + \beta\sqrt{n} u^k \xi(u) + \sigma u^k a(X,u)}.$$

Lemma *3:* Assume that $k_1 > 0$. For an arbitrary analytic function $\xi(u)$,

$$\begin{aligned} E_u^\sigma[u^{2k}|\xi] &\le \frac{c_1}{n}\{1 + \|\xi\|^2 + \|\partial_1\xi\|^2 \\ &\quad + \sigma\|a(X)\| + \sigma\|\partial_1 a(X)\|\}, \\ E_u^\sigma[u^{3k}|\xi] &\le \frac{c_2}{n^{3/2}}\{1 + \|\xi\|^3 + \|\partial_1\xi\|^3 \\ &\quad + (\sigma\|a(X)\|)^{3/2} + (\sigma\|\partial_1 a(X)\|)^{3/2}\}, \end{aligned}$$

where $\partial_1 = (\partial/\partial u_1)$, $c_1 = c_3(k_1+1)/\beta + 1/2$, $c_2 = c_3 5((k_1+h_1+1)/\beta+1)^{3/2}$, and $c_3 = \|\psi\|\|1/\psi\|$.
Note that, by Lemma 3, $G_g(\epsilon)$ is asymptotically uniformly integrable. For the proof of this Lemma, see [28].

Since $w = g(u)$, we rewrite the major parts of four errors as

$$B_g(\epsilon) = E_X[-\log E_u^0[e^{-a(X,u)u^k}|\xi_n]], \quad (7)$$

$$B_t(\epsilon) = \frac{1}{n}\sum_{j=1}^{n} -\log E_u^0[e^{-a(X_j,u)u^k}|\xi_n], \quad (8)$$

$$G_g(\epsilon) = E_u^0[u^{2k}|\xi_n], \quad (9)$$

$$G_t(\epsilon) = E_u^0[u^{2k} - \frac{1}{\sqrt{n}}\ u^k \xi_n(u)|\xi_n]. \quad (10)$$

Without loss of generality, in each local coordinate, we can assume $u = (x,y)$ $x \in \mathbf{R}^r$, $y \in \mathbf{R}^{r'}$, $(r' = d - r)$, $k = (k,k')$, $h = (h,h')$, and

$$\frac{h_1+1}{2k_1} = \cdots = \frac{h_r+1}{2k_r} = \lambda_\alpha < \frac{h_1'+1}{2k_1'} \le \cdots.$$

We define $\mu = h' - 2k'\lambda_\alpha \in \mathbf{R}^{r'}$, then

$$\mu_i > h_i' - 2k_i'\frac{h_i'+1}{2k_i'} = -1,$$

hence y^μ is integrable in $[0,1]^{r'}$. Both λ_α and r depend on the local coordinate. Let λ be the smallest λ_α, and m be the largest r among the coordinates in which $\lambda = \lambda_\alpha$. Then $(-\lambda)$ and m are respectively equal to the largest pole and its order of the zeta function of eq.(4). Let α^* be the index of the set of coordinates which satisfy $\lambda_\alpha = \lambda$ and $r = m$. As is shown in the following lemma, only coordinates U_{α^*} affect four errors. Let $\sum_{\alpha^*}$ be the sum of such coordinates.

For a given function $f(u)$, we use a notation $f_0(y) = f(0,y)$. Also $a_0(X,y) = a(X,0,y)$. The expectation for a given function $\xi(u)$ is defined by

$$E_{y,t}[f(y,t)|\xi] = \frac{\sum_{\alpha^*}\int_0^\infty dt \int dy\ f(y,t)\ Z_0(y,t,\xi)}{\sum_{\alpha^*}\int_0^\infty dt \int dy\ Z_0(y,t,\xi)}$$

where $\int dy$ shows $\int_{[0,1]^{r'}} dy$ and

$$Z_0(y,t,\xi) = y^\mu\ t^{\lambda-1} e^{-\beta t + \beta\sqrt{t}\ \xi_0(y)} \psi_0(y).$$

The following is the last lemma.

Lemma *4:* Let $p \geq 0$ be a constant. There exists $c_1 > 0$ such that, for arbitrary C^1-class function $f(u)$ and analytic function $\xi(u)$, the following inequality holds,

$$\left| n^p \ E_u^0[u^{2pk} f(u)|\xi] - \ E_{y,t}[t^p f_0(y)|\xi] \right| \leq \frac{c_1}{\log n} \exp(4\beta\|\xi\|^2)\{\beta\|\nabla\xi\|\|f\| + \|\nabla f\| + \|f\|\}$$

where $\|\nabla f\| = \sum_j \|\partial_j f\|$.
We define four functionals of a given function $\xi(u)$ by

$$\begin{aligned} B_g^*(\xi) &\equiv \frac{1}{2} E_X[\ E_{y,t}[a_0(X,y)t^{1/2}|\xi]^2\ ], & (11)\\ B_t^*(\xi) &\equiv G_t^*(\xi) - G_g^*(\xi) + B_g^*(\xi), & (12)\\ G_g^*(\xi) &\equiv E_{y,t}[t|\xi], & (13)\\ G_t^*(\xi) &\equiv E_{y,t}[t - t^{1/2}\xi_0(y)|\xi]. & (14) \end{aligned}$$

Note that these four functionals do not depend on n.

C. Proof of Theorem 1

Firstly we show the following convergences in probability hold.

$$\begin{aligned} nB_g(\epsilon) - B_g^*(\xi_n) &\to 0, & (15)\\ nB_t(\epsilon) - B_t^*(\xi_n) &\to 0, & (16)\\ nG_g(\epsilon) - G_g^*(\xi_n) &\to 0, & (17)\\ nG_t(\epsilon) - G_t^*(\xi_n) &\to 0. & (18) \end{aligned}$$

Based on eq.(9) and (13), eq.(17) is obtained by Lemma 4. Also based on eq.(10) and (14), eq.(18) is obtained by Lemma 4. To prove eq.(15), we define

$$b_g(\sigma) \equiv E_X\left[-\log E_u^0[e^{-\sigma a(X,u)u^k}|\xi_n]\ \right],$$

then, it follows that $nB_g(\epsilon) = nb_g(1)$ and there exists $0 < \sigma^* < 1$ such that

$$\begin{aligned} nB_g(\epsilon) &= nE_u^0[u^{2k}|\xi_n] \\ &\quad -\frac{n}{2}E_X E_u^0[a(X,u)^2u^{2k}|\xi_n] \\ &\quad +\frac{n}{2}E_X E_u^0[a(X,u)u^k|\xi_n]^2 \\ &\quad +\frac{1}{6}nb_g^{(3)}(\sigma^*), \end{aligned} \quad (19)$$

where we used $E_X[a(X,u)] = u^k$. The first term in the right hand side of eq.(19) is $nG_g(\epsilon)$. By Lemma 4, the convergences in probability

$$\begin{aligned} &\left| nE_X E_u^0[a(X,u)^2u^{2k}|\xi_n] - E_X E_{y,t}[a_0(X,y)^2 t|\xi_n] \right| \\ \leq\ & \frac{c_1}{\log n} e^{4\beta\|\xi_n\|^2} E_X[\ \beta\|\nabla\xi_n\|\|a(X)\| \\ & +\|\nabla a(X)\| + \|a(X)\|\ ] \to 0 \end{aligned} \quad (20)$$

holds. Since $E_X[a_0(X,y)] = 2$, the sum of the first two terms of the right hand side of eq.(19) converges to zero in probability. For the third term, by using the notations $E_X[a(X,u)a(X,v)] = \rho(u,v)$, $\rho_0(u,y) = \rho(u,(0,y))$, and $\rho_{00}(y',y) = \rho((0,y'),(0,y))$, and applying Lemma 4,

$$\begin{aligned} &|nE_X E_u^0[a(X,u)u^k|\xi_n]^2 - E_{y,t}[a_0(X,y)t^{1/2}|\xi_n]^2| \\ \leq\ & \left| \sqrt{n}E_u^0\left[u^k(\sqrt{n}E_v^0[\rho(u,v)v^k] - E_{y,t}[\rho_0(u,y)t^{1/2}])\right]\right| \\ & +\left| E_{y,t}\left[t^{1/2}(\sqrt{n}E_u^0[\rho_0(u,y)u^k]\right.\right. \\ & \left.\left. -E_{y',t'}[\rho_{00}(y',y)(t't)^{1/2}])\right]\right| \\ \leq\ & \frac{c_1\sqrt{n}}{\log n} E_u^0[u^k]\ e^{4\beta\|\xi_n\|^2}(\beta\|\nabla\xi_n\|\|\rho\| + \|\nabla\rho\| + \|\rho\|) \\ & +\frac{c_1}{\log n} e^{4\beta\|\xi_n\|^2}(\beta\|\nabla\xi_n\|\|\rho\| + \|\nabla\rho\| + \|\rho\|), \end{aligned} \quad (21)$$

where '$|\xi_n$' is omitted for simple notation. The equation (21) converges to zero in probability by Lemma 3. Therefore the difference between the third term and $B_g^*(\xi_n)$ converges to zero in probability. For the last term, we have

$$\begin{aligned} |nb^{(3)}(\sigma^*)| &= \left| E_X\left\{ E_u^{\sigma^*}[a(X,u)^3u^{3k}|\xi_n] \right.\right. \\ &\quad +2E_u^{\sigma^*}[a(X,u)|\xi_n]^3 \\ &\quad -3E_u^{\sigma^*}[a(X,u)^2u^{2k}|\xi_n] \\ &\quad \left.\left. \times E_u^{\sigma^*}[a(X,u)u|\xi_n]\right\}\right| \\ &\leq 6nE_X\left[\|a(X)\|^3\ E_u^{\sigma^*}[u^{3k}|\xi_n]\right]. \end{aligned}$$

By applying Lemma 3,

$$\begin{aligned} |nb_g^{(3)}(\sigma^*)| &\leq \frac{6c_2}{n^{1/2}} E_X\left[\|a(X)\|^3\ \{1 + \|\xi_n\|^3 + \|\partial\xi_n\|^3 \right. \\ &\quad \left. +\|a(X)\|^{3/2} + \|\partial a(X)\|^{3/2}\}\right], \end{aligned} \quad (22)$$

which shows $nb_g^{(3)}(\sigma^*)$ converges to zero in probability. Hence eq.(15) is proved. Let us prove eq.(16). By defining

$$b_t(\sigma) = \frac{1}{n}\sum_{j=1}^n -\log E_u^0[e^{-\sigma a(X_j,u)u^k}|\xi_n],$$

it follows that $nB_t(\epsilon) = nb_t(1)$ and there exists $0 < \sigma^* < 1$ such that

$$\begin{aligned} nB_t(\epsilon) &= nG_t(\epsilon) - \frac{1}{2}\sum_{j=1}^n E_u^0[a(X_j,u)^2u^{2k}|\xi_n] \\ &\quad +\frac{1}{2}\sum_{j=1}^n E_u^0[a(X_j,u)u^k|\xi_n]^2 + \frac{1}{6}nb_t^{(3)}(\sigma^*), \end{aligned}$$

Then by applying Lemma 3, $nb_t^{(3)}(\sigma^*)$ converges to zero in probability by the same way as eq.(22). By the same methods as eq.(20) and eq.(21) replacing respectivey $E_X[\|a(X)^2\|]$ and $\rho(u,v)$ with $(1/n)\sum_j \|a(X_j)^2\|$ and $\rho_n = (1/n)\sum_j a(X_j,u)a(X_j,v)$, convergences in probability

$$\begin{aligned} \frac{1}{2}\sum_{j=1}^n E_u^0[a(X_j,u)^2u^{2k}|\xi_n] - G_g^*(\xi_n) &\to 0 \\ \frac{1}{2}\sum_{j=1}^n E_u^0[a(X_j,u)u^k|\xi_n]^2 - B_g^*(\xi_n) &\to 0 \end{aligned}$$

hold, resulting that the convergence in probability

$$nB_t(\epsilon) - nG_t(\epsilon) + nG_g(\epsilon) - nB_g(\epsilon) \to 0. \quad (23)$$

holds. Therefore eq.(16) is obtained. By combining eq.(15)-eq.(18), convergences in probability holds,

$$\begin{aligned} nB_g - B_g^*(\xi_n) &\to 0, & (24)\\ nB_t - B_t^*(\xi_n) &\to 0, & (25)\\ nG_g - G_g^*(\xi_n) &\to 0, & (26)\\ nG_t - G_t^*(\xi_n) &\to 0. & (27) \end{aligned}$$

Four functionals $B_g^*(\xi)$, $B_t^*(\xi)$, $G_g^*(\xi)$, and $G_t^*(\xi)$ are continuous functions of $\xi \in B(\mathcal{M})$. From the convergence in law of the empirical process $\xi_n \to \xi$, convergences in law

$$\begin{aligned} B_g^*(\xi_n) \to B_g^*(\xi), \quad & B_t^*(\xi_n) \to B_t^*(\xi),\\ G_g^*(\xi_n) \to G_g^*(\xi), \quad & G_t^*(\xi_n) \to G_t^*(\xi), \end{aligned}$$

are derived. Therefore Theorem 1 (1) and (2) are obtained. Theorem 1 (3) is shown in [28] (Q.E.D.)

D. Proof of Theorem 2

Before proving the theorem, we prepare a property of gaussian process. Let $\{g_i\}_{i=1}^{\infty}$ be independent gaussian random variables on $\mathbf{R}$ which satisfy $E[g_i] = 0,\ E[g_i g_j] = \delta_{ij}$. For such random variables,

$$E[g_i F(g_i)] = E[\frac{\partial}{\partial g_i} F(g_i)]$$

holds for a differentiable function of $F(\cdot)$. Since $L^2(q)$ is a separable Hilbert space, there exists a complete orthonormal system $\{e_k(x)\}_{k=1}^{\infty}$. By defining

$$b_k(u) = \int a(x,u)\ e_k(u)\ q(x)\ dx,$$

we have a relations

$$\begin{aligned} a(x,u) &= \sum_{k=1}^{\infty} b_k(u) e_k(x),\\ E_X[a(X,u)a(X,v)] &= \sum_{k=1}^{\infty} b_k(u) b_k(v). \end{aligned}$$

The tight gaussian process defined by

$$\xi^*(u) = \sum_{k=1}^{\infty} b_k(u)\ g_k$$

has the same expectations and covariance matrices as $\xi(u)$, therefore it is subject to the same probability distribution as $\xi(u)$. Thus we can identify $\xi(u) = \xi^*(u)$ in calculation of expectation values.

$$E[\xi(u)\xi(v)] = \sum_{i=1}^{\infty} b_i(u) b_i(v). \quad (28)$$

If $u, v \in \mathcal{M}_0$, then

$$E[\xi(u)\xi(v)] = E_X[a(X,u)a(X,v)]. \quad (29)$$

Let us prove Theorem 2. We use notations,

$$\begin{aligned} Y(a) &= \int_0^{\infty} dt\ t^{\lambda-1}\ e^{-\beta t + a\beta\sqrt{t}},\\ \int du^* &= \sum_{\alpha^*} \int dx\ dy\ \delta(x)\ y^{\mu},\\ Z(\xi_n) &= \int du^*\ Y(\xi(u)), \end{aligned}$$

where $u = (x, y)$ are introduced. Then

$$\begin{aligned} 2E[B_g^*] &= \frac{1}{\beta^2} E[E_X[\Big(\frac{\int du^* a(X,u) Y'(\xi(u))}{Z(\xi)}\Big)^2]]\\ E[G_g^*] &= \frac{1}{\beta^2} E[\frac{\int du^* Y''(\xi(u))}{Z(\xi)}]\\ E[G_t^*] &= \frac{1}{\beta^2} E[\frac{\int du^* Y''(\xi(u))}{Z(\xi)}] - \frac{A}{\beta} \end{aligned}$$

where A is a constant defined by

$$\begin{aligned} A &\equiv E[\frac{\int du^*\ \xi(u) Y'(\xi(u))}{Z(\xi)}]\\ &= E[\int du^*\ \{\sum_{i=1}^{\infty} b_i(u) g_i\} \frac{Y'(\xi(u))}{Z(\xi)}]\\ &= E[\int du^* \{\sum_{i=1}^{\infty} b_i(u) \frac{\partial}{\partial g_i}\} \frac{Y'(\xi(u))}{Z(\xi)}]. \end{aligned}$$

Then by usining

$$\begin{aligned} \frac{\partial}{\partial g_i} \frac{Y'(\xi(u))}{Z(\xi)} &= \frac{Y''(\xi(u)) b_i(u)}{Z(\xi)}\\ &\quad - \frac{Y'(\xi(u))}{Z(\xi)^2} \int dv^*\ Y'(\xi(v)) b_i(v), \end{aligned}$$

we obtain

$$\begin{aligned} A &= E[\int du^*\ \frac{Y''(\xi(u)) \sum_{i=1}^{\infty} b_i(u)^2}{Z(\xi)}]\\ &\quad - E[\int du^* \int dv^* \frac{Y'(\xi(u)) Y'(\xi(v)) \sum_i b_i(u) b_i(v)}{Z(\xi)^2}]\\ &= E[\int du^* \frac{2Y''(\xi(u))}{Z(\xi)}]\\ &\quad - E[\int du^* \int dv^* \frac{Y'(\xi(u)) Y'(\xi(v)) E[\xi(u)\xi(v)]}{Z(\xi)^2}], \end{aligned}$$

where we used eq.(28). By applying eq.(29),

$$\begin{aligned} A &= 2E[\frac{\int du^* Y''(\xi(u))}{Z(\xi)}]\\ &\quad - EE_X[\Big(\frac{\int du^*\ a(X,u) Y'(\xi(u))}{Z(\xi)}\Big)^2]\\ &= 2\beta^2 E[G_g^*] - 2\beta^2 E[B_g^*]. \end{aligned}$$

Since

$$A = \beta(E[G_g^*] - E[G_t^*]),$$

we obtain the Theorem 2. (Q.E.D.)

E. Proof of Theorem 3

By using the partial integration, for an arbitrary a,

$$\int_0^\infty e^{-\beta t}\, 2t^\lambda\, e^{\beta a\sqrt{t}}\, dt = \frac{1}{\beta}\int_0^\infty e^{-\beta t}\frac{\partial}{\partial t}\Big(2t^\lambda\, e^{\beta a\sqrt{t}}\Big)\, dt.$$

Hence

$$\int_0^\infty dt\, (2t - \sqrt{t}a - \frac{2\lambda}{\beta})\, t^{\lambda-1}\, e^{-\beta t+\beta\sqrt{t}a} = 0.$$

It follows that

$$G_g^*(\xi_n) + G_t^*(\xi_n) - \frac{2\lambda}{\beta} = 0,$$

which is the conclusion of Theorem 3. (Q.E.D.)

V. Discussion

We define a constant $\nu \geq 0$ by

$$\nu = \frac{1}{2\beta}E\Big[\frac{\int du^*\xi(u)Y'(\xi(u))}{Z(\xi)}\Big].$$

Note that ν is a constant for n, but it depends on β. Then it follows that

$$E[B_g] = (\frac{\lambda-\nu}{\beta}+\nu)\frac{1}{n} + o(\frac{1}{n}), \quad (30)$$

$$E[B_t] = (\frac{\lambda-\nu}{\beta}-\nu)\frac{1}{n} + o(\frac{1}{n}), \quad (31)$$

$$E[G_g] = (\frac{\lambda}{\beta}+\nu)\frac{1}{n} + o(\frac{1}{n}), \quad (32)$$

$$E[G_t] = (\frac{\lambda}{\beta}-\nu)\frac{1}{n} + o(\frac{1}{n}). \quad (33)$$

Here $\lambda > 0$ is a constant which does not depend on β. If a learning machine is regular then $\lambda = \nu = d/2$, where d is the dimension of the parameter space. As is proven in this paper, by measuring both Bayes and Gibbs training errors, we can estimate λ and ν using equations of states. It is a future study to clarify the function $\nu = \nu(\beta)$. This value has important information of singularities.

VI. Conclusion

Based on singular learning theory, we established the equations of states in learning which hold for any true distribution, any learning machine, and a priori distribution, and any singularities. Using the equations of states, we proposed widely applicable information criteria, which can be used in both regular and singular learning machines.

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,"Stochastic complexities of reduced rank regression in Bayesian estimation," Neural Networks, Vol.18,No.7, pp.924-933, 2005.

[3] 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.

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

[10] J. Kollár,"Lectures on Resolution of Singularities," Princeton University Press, (Princeton), 2007.

[11] Kenji Nagata and Sumio Watanabe,"Exchange Monte Carlo Sampling from Bayesian Posterior for Singular Learning Machines", to appear in IEEE Transactions on Neural Networks.

[12] Kenji Nagata, Sumio Watanabe, "Asymptotic Behavior of Exchange Ratio in Exchange Monte Carlo Method", International Journal of Neural Networks, to appear.

[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] S. Watanabe,"Algebraic geometry and learning theory," Morikita publishing, 2006.

[26] S. Watanabe,"Almost all learning machines are singular," Proc. of international Symposimum on IEEE FOCI 2007.

[27] S.Watanabe,"Generalization and training errors in Bayes and Gibbs estimations in singular learning machines," IEICE technical report (NC2007-75), Vol.2007-12, pp.25-30, December, 2007.

[28] S.Watanabe,"Equations of States in Singular Statistical Estimation," arXiv:0712.653, 5th December. 2007.

[29] A. W. van der Vaart, Jon A. Wellner, "Weak Convergence and Empirical Processes," Springer,1996.

[30] 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.

[31] 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.

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