

# Equations of States in Singular Statistical Estimation

Sumio Watanabe  
Precision and Intelligence Laboratory  
Tokyo Institute of Technology  
4259 Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan  
E-mail: swatanab@pi.titech.ac.jp

October 22, 2018

## Abstract

Learning machines that have hierarchical structures or hidden variables are singular statistical models because they are nonidentifiable and their Fisher information matrices are singular. In singular statistical models, neither does the Bayes *a posteriori* distribution converge to the normal distribution nor does the maximum likelihood estimator satisfy asymptotic normality. This is the main reason that it has been difficult to predict their generalization performance from trained states. In this paper, we study four errors, (1) the Bayes generalization error, (2) the Bayes training error, (3) the Gibbs generalization error, and (4) the Gibbs training error, and prove that there are universal mathematical relations among these errors. The formulas proved in this paper are equations of states in statistical estimation because they hold for any true distribution, any parametric model, and any *a priori* distribution. Also we show that the Bayes and Gibbs generalization errors can be estimated by Bayes and Gibbs training errors, and we propose widely applicable information criteria that can be applied to both regular and singular statistical models.

## 1 Introduction

Recently, many learning machines are being used in information processing systems. For example, layered neural networks, normal mixtures, binomial mixtures, Bayes networks, Boltzmann machines, reduced rank regressions, hidden Markov models, and stochastic context-free grammars are being employed in pattern recognition, time series prediction, robotic control, human modeling, and biostatistics. Although their generalization performances determine the accuracy of the information systems, it has been difficult to estimate generalization errors based on training errors, because such learning machines are singular statistical models.

A parametric model is called regular if the mapping from the parameter to the probability distribution is one-to-one and if its Fisher information matrix is always positive definite. If a statistical model is regular, then the Bayes *a posteriori* distribution converges to the normal distribution, and the maximum likelihood estimator satisfies asymptotic normality. Based on such properties, the relation between the generalization error and the training error was clarified, on which some information criteria were proposed.

On the other hand, if the mapping from the parameter to the probability distribution

is not one-to-one or if the Fisher information matrix is singular, then the parametric model is called singular. In general, if a learning machine has hierarchical structure or hidden variables, then it is singular. Therefore, almost all learning machines are singular. For singular learning machines, the log likelihood function can not be approximated by any quadratic form of the parameter, with the result that the conventional relationship between generalization errors and training errors does not hold either for the maximum likelihood method [6] [5][7] or Bayes estimation [12]. Singularities strongly affect generalization performances [15] and learning dynamics [1]. Therefore, in order to establish the mathematical foundation of singular learning theory, it is necessary to construct the formulas which hold even in singular learning machines.

Recently, we proved [13][15] that the generalization error in Bayes estimation is asymptotically equal to  $\lambda/n$ , where  $\lambda > 0$  is the rational number determined by the zeta function of a learning machine and  $n$  is the number of training samples. In regular statistical models,  $\lambda = d/2$ , where  $d$  is the dimension of the parameter space, whereas in singular statistical models,  $\lambda$  depends strongly on the learning machine, the true distribution, and the *a priori* probability distribution. In practical applications, the true distribution is often unknown, hence it has been difficult to estimate the generalization error from the training error. To estimate the generalization error when we do not have any information about the true distribution, we need a general formula which holds independently of singularities.

In this paper, we study four errors, (1) the Bayes generalization error  $B_g$ , (2) the Bayes training error  $B_t$ , (3) the Gibbs generalization error  $G_g$ , and (4) the Gibbs training error  $G_t$ , and prove the formulas

$$\begin{aligned} E[B_g] - E[B_t] &= 2\beta(E[G_t] - E[B_t]) + o\left(\frac{1}{n}\right), \\ E[G_g] - E[G_t] &= 2\beta(E[G_t] - E[B_t]) + o\left(\frac{1}{n}\right), \end{aligned}$$

where  $E[\cdot]$  denotes the expectation value and  $0 < \beta < \infty$  is the inverse temperature of the *a posteriori* distribution. These equations assert that the increased error from training to generalization is in proportion to the difference between the Bayes and Gibbs training errors. It should be emphasized that these formulas hold for any true distribution, any learning machine, any *a priori* probability distribution, and any singularities, therefore they reflect the universal laws of statistical estimation. Also, based on the formula, we propose widely applicable information criteria (WAIC) which can be applied to both regular and singular learning machines. In other words, we can apply WAIC without any knowledge about the true distribution.

This paper consists of six parts. In Section 2, we describe the main results of this paper. In Section 3, we propose widely applicable information criteria and show how to apply them to statistical estimation. In Section 4, we prove the main results in the mathematically rigorous way. In Sections 5 and 6, we discuss and conclude of this paper. The proofs of lemmas are quite technical hence they are presented in Appendix.

## 2 Main Results

Let  $(\Omega, \mathcal{B}, P)$  be a probability space, and  $X : \Omega \rightarrow \mathbf{R}^N$  be a random variable whose probability distribution is  $q(x)dx$ . Here  $\mathbf{R}^N$  denotes the  $N$  dimensional Euclidean space. We assume that the random variables  $X_1, X_2, \dots, X_n$  are independently subject to the same probability distribution as  $X$ . In learning theory,  $q(x)dx$  is called the true distribution and  $D_n = \{X_1, X_2, \dots, X_n\}$  is a set of training samples. A learning machine is defined by a parametric probability density function  $p(x|w)$  of  $x \in \mathbf{R}^N$  for a given parameter  $w \in W \subset \mathbf{R}^d$ , where  $W$  is a set of parameters. An *a priori* probability density function  $\varphi(w)$  is defined on  $W$ . The Bayes *a posteriori* probability density  $p(w|D_n)$  for a given set of training samples  $D_n$  is defined by

$$p(w|D_n) = \frac{1}{C_n} \varphi(w) \left( \prod_{i=1}^n p(X_i|w) \right)^\beta,$$

where  $\beta > 0$  is the inverse temperature and  $C_n > 0$  is the normalizing constant. The expectation value with respect to this probability distribution is denoted by  $E_w[\cdot]$ . Also  $E_{D_n}[\cdot]$  and  $E_X[\cdot]$  denote respectively the expectation values over  $D_n$  and  $X$ . We sometimes omit  $D_n$  and simply use  $E[\cdot]$ . We study the four errors, defined below.

(1) Bayes generalization error,

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

(2) Bayes training error,

$$B_t = \frac{1}{n} \sum_{j=1}^n \log \frac{q(X_j)}{E_w[p(X_j|w)]}.$$

(3) Gibbs generalization error,

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

(4) Gibbs training error,

$$G_t = E_w \left[ \frac{1}{n} \sum_{j=1}^n \log \frac{q(X_j)}{p(X_j|w)} \right].$$

These four errors are measurable functions of  $D_n$ , hence they are also random variables.

**Remark.** The Bayes generalization error is equal to the Kullback-Leibler distance from the true distribution  $q(x)$  to the Bayes predictive distribution  $E_w[p(x|w)]$ . The Gibbs generalization error is equal to the average of the Kullback-Leibler distance from the true distribution to the Gibbs estimation. They show the accuracy of Bayes and Gibbs estimations, it is important for statistical learning machines to be able to estimate them from random samples.

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

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

In this paper, we mainly study the singular case, that is to say, the situation when the set of true parameters  $\{w \in W; q(x) = p(x|w)\}$  consists of more than one point and the Fisher information matrix is not positive definite. We assume the following three conditions.

**(A.1)** Assume that the set of parameters  $W$  is a compact set which is the closure of an open set in  $\mathbf{R}^d$ . The set  $W$  is defined by

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

where  $\pi_1(w), \dots, \pi_k(w)$  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  $C^\infty$ -class function and  $\varphi_1(w) \geq 0$  is an analytic function.

**(A.2)** Let  $s \geq 6$  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 the 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 for  $M(x) \equiv \sup_{w \in W^*} |f(X, w)|$ ,

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

and there exists  $t > 0$  such that, for  $Q(x) \equiv \sup_{K(w) \leq t} p(x|w)$

$$\int M(x)^2 Q(x) dx < \infty.$$

**Remark.** These assumptions are needed for the mathematical reasons.

(1) These conditions allow for the case that the set of true parameters  $W_0 = \{w \in$

$W; q(x) = p(x|w)\}$  is not a single point but an algebraic set or an analytic set with singularities. In general, the Fisher information matrix has zero eigenvalues. On the other hand, in conventional statistical learning theory, it is assumed that  $W_0$  consists of one point and the Fisher information matrix is positive definite. On the assumptions of this paper, we can not use any result of conventional statistical learning theory.

(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 a singularity in general. For 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 other hand, if  $|w| = \infty$  is not a singularity,  $\mathbf{R}^d \cup \{|w| = \infty\}$  can be understood as a compact set and the same theorems established in this paper hold.

(3) The condition that  $\pi_1(w), \dots, \pi_k(w)$  and  $\varphi_1(w)$  are analytic functions is necessary because if one of them is a  $C^\infty$  class function, there exists a pathological example. In fact, if  $\varphi_1(w) = \exp(-1/\|w\|^2)$  in a neighborhood of the origin and the set of true parameters is the origin, then the four errors may not be in proportion to  $1/n$ .

(4) The condition  $s \geq 6$  is needed to ensure the existence of the asymptotic expansion of the Bayes generalization error in our proof. (See the proof of Theorem 1.)

(5) Some non-analytic statistical models can be made analytic. For example, in a simple mixture model  $p(x|a) = ap_1(x) + (1-a)p_2(x)$  for some probability densities  $p_1(x)$  and  $p_2(x)$ , the log density ratio function  $f(x, a)$  is not analytic at  $a = 0$ , but it can be made analytic by the representation  $p(x|\theta) = \alpha^2 p_1(x) + \beta^2 p_2(x)$ , on the manifold  $\theta \in \{\alpha^2 + \beta^2 = 1\}$ . As is shown in the proofs, if  $W$  is contained in an analytic manifold, then the same theorems hold as stated in this paper.

(6) Note that

$$\int M(x)^6 q(x) dx < \infty. \quad (1)$$

Based on assumptions (A.1), (A.2), and (A.3), we prove the following results.

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

$$nB_g \rightarrow B_g^*, \quad nB_t \rightarrow B_t^*, \quad nG_g \rightarrow G_g^*, \quad nG_t \rightarrow G_t^*.$$

(2) *As  $n \rightarrow \infty$ , the following convergence in probability holds,*

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

(3) The expectation values of the four errors converge as follows,

$$E[nB_g] \rightarrow E[B_g^*], \quad E[nB_t] \rightarrow E[B_t^*],$$

$$E[nG_g] \rightarrow E[G_g^*], \quad E[nG_t] \rightarrow E[G_t^*].$$

For the proof of this theorem, see Section 4. the following Theorem is the main result of this paper.

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

$$E[B_g^*] - E[B_t^*] = 2\beta(E[G_t^*] - E[B_t^*]), \quad (2)$$

$$E[G_g^*] - E[G_t^*] = 2\beta(E[G_t^*] - E[B_t^*]). \quad (3)$$

**Remark.** (1) Theorem 2 asserts that the increases of errors from training to prediction are in proportion to the difference between the Bayes and Gibbs training errors. We refer to 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. It is proved that the equations of states hold even if the true distribution is not contained in the parametric model [22].

(2) Although the equations of states hold universally, the four errors themselves depend strongly on a true distribution, a learning machine, an *a priori* distribution, and singularities.

(3) Theorem 2 also asserts a conservation law, namely, the difference between the Bayes error and the Gibbs error is invariant between training and generalization,

$$E[G_g^*] - E[B_g^*] = E[G_t^*] - E[B_t^*]. \quad (4)$$

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

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

$$\begin{pmatrix} E[B_g^*] \\ E[G_g^*] \end{pmatrix} = \begin{pmatrix} 1 - 2\beta & 2\beta \\ -2\beta & 1 + 2\beta \end{pmatrix} \begin{pmatrix} E[B_t^*] \\ E[G_t^*] \end{pmatrix}. \quad (5)$$

**Remark.** (1) From eq.(5), 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) Since the set of eigenvalues of the linear transform in eq.(5) is  $\{1\}$ , and the dimension

of the linear invariant subspace is one, there is no conservation law other than eq.(4).

(3) A statistical model is called *regular* if the set of true parameters  $W_0 = \{w \in W; q(x) = p(x|w)\}$  consists of a single point and if the 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

$$\begin{aligned} E[B_g^*] &= \frac{d}{2}, & E[G_g^*] &= \left(1 + \frac{1}{\beta}\right) \frac{d}{2}, \\ E[B_t^*] &= -\frac{d}{2}, & E[G_t^*] &= \left(-1 + \frac{1}{\beta}\right) \frac{d}{2}, \end{aligned}$$

which is a special case of Theorem 2.

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

$$\zeta(z) = \int_W K(w)^z \varphi(w) dw. \quad (6)$$

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][17]). They are denoted as follows,

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

The order of each pole  $\lambda_k$  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** *As  $n \rightarrow \infty$ , the convergence in probability*

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

*holds. Therefore*

$$E[G_g^*] + E[G_t^*] = \frac{2\lambda}{\beta}. \quad (7)$$

Also the following corollary holds.

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

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

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

From these theorems and corollaries, if one knows the true distribution, one can predict the Bayes and Gibbs generalization errors from the Bayes and Gibbs training errors with probability one, as  $n$  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. Lastly, by Theorems 2 and 3, the following corollary is immediately proved.

**Corollary 3** *Let  $\nu = \nu(\beta) = \beta(E[G_t^*] - E[B_t^*])$ . Then*

$$\begin{aligned} E[B_g^*] &= \frac{\lambda - \nu}{\beta} + \nu, \\ E[B_t^*] &= \frac{\lambda - \nu}{\beta} - \nu, \\ E[G_g^*] &= \frac{\lambda}{\beta} + \nu, \\ E[G_t^*] &= \frac{\lambda}{\beta} - \nu. \end{aligned}$$

*Therefore Bayes learning is asymptotically determined by  $\lambda$  and  $\nu$ .*

In general  $\nu(\beta)$  depend on  $\beta > 0$ . In regular statistical models,  $\lambda = \nu = d/2$  for arbitrary  $\beta > 0$ , whereas in singular learning machines, they are different in general. Corollary 2 was firstly discovered in [13][15]. Since the constant  $\lambda$  depends strongly 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 [16], normal mixtures [24], reduced rank regressions [2], Boltzmann machines [25], and hidden Markov models [26]. Also the behavior of  $\lambda$  was analyzed for the case when Jeffreys' prior is employed as an *a priori* distribution [14], and in the case when the distance of the true distribution from the singularity is in proportion to  $1/\sqrt{n}$  [18].

### 3 Widely Applicable Information Criteria

The main purpose of this paper is to prove the theorems above. However, in order to illustrate the importance of the results of this paper, we propose widely applicable information criteria and introduce an experiment. Experimental analysis of practical applications is a topic for future study.

### 3.1 Basic Concepts

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

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

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 combining the entropy of the true distribution with Corollary 1,

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

we obtain the equations,

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

Let us define widely applicable information criteria (WAIC) by

$$\begin{aligned} \text{WAIC}_1 &= BL_t + 2\beta (GL_t - BL_t), \\ \text{WAIC}_2 &= GL_t + 2\beta (GL_t - BL_t). \end{aligned}$$

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

$$\begin{aligned} E[BL_g] &= E[\text{WAIC}_1] + o\left(\frac{1}{n}\right), \\ E[GL_g] &= E[\text{WAIC}_2] + o\left(\frac{1}{n}\right). \end{aligned}$$

Therefore,  $\text{WAIC}_1$  and  $\text{WAIC}_2$  provide indices for model evaluation.

**Remark.** If a model is regular and the true distribution is contained in the parametric model, then  $\lambda = d/2$  and

$$2\beta(E[G_t^*] - E[B_t^*]) = d \tag{8}$$

hold. It is proved in [22] that, even if a model  $p(x|w)$  does not contain the true distribution  $q(x)$ , the equations of states hold if the Hessian matrix of the Kullback-Leibler distance is positive definite at the unique optimal parameter  $w^*$  that minimizes the Kullback-Leibler distance from  $q(x)$  to  $p(x|w)$ . In such a case,

$$2\beta(E[G_t^*] - E[B_t^*]) = \text{tr}(IJ^{-1}), \quad (9)$$

where  $I$  and  $J$  are  $d \times d$  matrices defined by

$$\begin{aligned} I_{ij} &= \int \partial_i f(x, w^*) \partial_j f(x, w^*) q(x) dx, \\ J_{ij} &= - \int \partial_i \partial_j f(x, w^*) q(x) dx. \end{aligned}$$

Here we used a notation,  $\partial_i = (\partial/\partial w_i)$ . Moreover, as  $n \rightarrow \infty$  convergence in probability

$$2\beta(G_t^* - B_t^*) \rightarrow \text{tr}(IJ^{-1}) \quad (10)$$

holds. If  $\beta \rightarrow \infty$ , both the Bayes and Gibbs estimations result in the maximum likelihood method. Therefore, for regular statistical models, WAIC has asymptotically the same variance as AIC. In other words, WAIC can be understood as information criteria of generalized from AIC. For singular learning machines, neither eq.(8) nor (9) holds, for example,  $J^{-1}$  does not exist, whereas WAIC gives the accurate generalization error.

**Remark.** In Bayes estimation, the marginal likelihood or the stochastic complexity

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

is often used in model selection and hyperparameter optimization. We clarified its behavior for singular learning machines in [15]. In regular statistical models,  $F$  is asymptotically equal to BIC, however, in singular models, it is not equal to BIC even asymptotically. Note that  $F$  does not correspond to the generalization error, hence the optimal model for the minimizing  $F$  does not minimize the generalization error in general. The Bayes and Gibbs generalization errors are important because they correspond directly to the Kullback-Leibler distance from the true distribution to the estimated one. In this paper, we make mathematically new information criteria which correspond to the generalization error. Even for regular statistical models, there is much research and discussion which compares AIC with BIC. It is a topic for future study to compare the marginal likelihood and the equations of states from the viewpoint of statistical methodology.

**Remark.** In conventional Bayes estimation, the inverse temperature  $\beta = 1$  is used. Hence WAIC for  $\beta = 1$  is most important. On the other hand, WAIC for general  $\beta$  shows the

$H$	Theory	$E[B_g]$	$\sigma[B_g]$	$E[\text{WAIC}_1]$	$\sigma[\text{WAIC}_1]$
1		6.215318	0.034043	6.214185	0.230465
2		3.013187	0.118109	2.993593	0.225722
3	0.027000	0.028422	0.007393	0.025139	0.006886
4	0.030000	0.030830	0.007678	0.027207	0.008176
5	0.032000	0.033030	0.008418	0.030152	0.008728
6	0.034000	0.034978	0.008832	0.031382	0.009778

Table 1: Experimental Results

effect of the inverse temperature on the generalization and training errors. Moreover, in applications, one may use  $\beta$  as a hyperparameter. In such a case, it can be optimized by the minimization of WAIC.

### 3.2 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

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

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  $\text{rank}(B_0A_0) = H_0$ . The algebraic variety of the true parameters is defined by  $K(A, B) = 0$ , where

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

has complicated singularities. We conducted experiments for the case that  $N_1 = N_2 = 6$ ,  $H_0 = 3$ ,  $\beta = 1$ ,  $n = 500$ , 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  $H = 1, 2, \dots, 6$  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. The expectation values  $B_g$  and  $\text{WAIC}_1$  were obtained by averaging over 25 trials, that is to say, 25 sets of training samples were independently taken from the true distribution. In Table.1, theoretical values of  $E[B_g]$  for  $\beta = 1$  were obtained from [2]. Learning machines with  $H = 1, 2$  do not contain the true distribution, hence theoretical values do not exist. The two values  $E[B_g]$  and  $\sigma[B_g]$  are the experimental average and standard deviation of the Bayes generalization error, respectively. The two values  $E[\text{WAIC}_1]$  and  $\sigma[\text{WAIC}_1]$  are the experimental average and standard deviation of  $\text{WAIC}_1$ , respectively. The experimental results show that the average behavior of the Bayes generalization error could be estimated by that of  $\text{WAIC}_1$ . However, the standard

deviations of the WAIC<sub>1</sub> and the Bayes generalization error are not small. Note that, even in regular statistical models, the standard deviations of the generalization error and AIC are also not small.

## 4 Singular Learning Theory

In this section, we shall prove the main theorems. Proofs of the lemmas are rather technical, hence they are given in Appendix.

### 4.1 Outline of the Proof

We prove the main theorems by the following procedure.

- (1) Firstly we show that only the neighborhoods of the true parameters essentially affect the four errors.
- (2) By using resolution of singularities, the set of parameters can be understood as the image of an analytic map from a manifold, on which all singularities of the true parameters are of normal crossing type.
- (3) We prove that the four errors converges in law to functionals of a tight gaussian process on the set of true parameters in the manifold.
- (4) Expectations of the four errors converge to those of functionals of the tight gaussian process.
- (5) The relations between the four errors are derived by partial integration of the gaussian process.

### 4.2 Basic Properties

By using the log density ratio function  $f(x, w)$ , we define the empirical Kullback-Leibler distance by

$$K_n(w) = \frac{1}{n} \sum_{i=1}^n f(X_i, w).$$

For a given constant  $a > 0$ , we define an expectation value restricted to the set  $\{w \in W; K(w) \leq a\}$  by

$$E_w[f(w)|_{K(w) \leq a}] = \frac{\int_{K(w) \leq a} f(w) e^{-\beta n K_n(w)} \varphi(w) dw}{\int_{K(w) \leq a} e^{-\beta n K_n(w)} \varphi(w) dw}.$$

We define four errors respectively by

$$B_g(a) = E_X \left[ -\log E_w [e^{-f(X, w)} |_{K(w) \leq a}] \right],$$

$$\begin{aligned}
B_t(a) &= \frac{1}{n} \sum_{j=1}^n -\log E_w[e^{-f(X_j, w)} |_{K(w) \leq a}], \\
G_g(a) &= E_w[K(w) |_{K(w) \leq a}], \\
G_t(a) &= E_w[K_n(w) |_{K(w) \leq a}].
\end{aligned}$$

Since  $W$  is compact and  $K(w)$  is an analytic function,  $\overline{K} = \sup_{w \in W} K(w)$  is finite. Then,  $B_g(\overline{K}) = B_g$ ,  $B_t(\overline{K}) = B_t$ ,  $G_g(\overline{K}) = G_g$ , and  $G_t(\overline{K}) = G_t$ . Also we define  $\eta_n(w)$  for  $w$  such that  $K(w) > 0$  by

$$\eta_n(w) = \frac{K(w) - K_n(w)}{\sqrt{K(w)}}, \quad (11)$$

and

$$H_t(a) = \sup_{0 < K(w) \leq a} |\eta_n(w)|^2.$$

$H_t(\overline{K})$  is denoted by  $H_t$ .

**Lemma 1** *For an arbitrary  $a > 0$ , the following inequalities hold.*

$$\begin{aligned}
B_t(a) &\leq G_t(a) \leq \frac{3}{2}G_g(a) + \frac{1}{2}H_t(a), \\
0 &\leq B_g(a) \leq G_g(a), \\
-\frac{1}{4}H_t(a) &\leq G_t(a).
\end{aligned}$$

For the proof of this lemma, see Section 7. In particular, by putting  $a = \overline{K}$ , we have

$$\begin{aligned}
B_t &\leq G_t \leq \frac{3}{2}G_g + \frac{1}{2}H_t, \\
0 &\leq B_g \leq G_g, \\
-\frac{1}{4}H_t &\leq G_t.
\end{aligned}$$

**Remark.** A sequence of random variables  $\{R_n\}$  is called asymptotically uniformly integrable (AUI) if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} E[I_M(R_n)] = 0,$$

where

$$I_M(x) = \begin{cases} 0 & (|x| < M) \\ |x| & (|x| \geq M) \end{cases}.$$

The following properties are well known [23].

- (1) If the convergence in law  $R_n \rightarrow R$  holds and  $R_n$  is AUI, then  $E[R_n] \rightarrow E[R]$ .
- (2) If  $R_n$  is AUI and if a random variable  $S_n$  satisfies  $|S_n| \leq R_n$ , then  $S_n$  is also AUI.
- (3) If there exist  $p > 0$  and  $C > 0$  such that  $E[|R_n|^p] < C$ , then  $R_n^q$  ( $0 < q < p$ ) is AUI.

By Lemma 1, if  $nH_t(a)$ ,  $nG_g(a)$ , and  $nB_t(a)$  are AUI, then  $nB_g(a)$  and  $nG_t(a)$  are AUI.

**Lemma 2** (1) *There exists a constant  $C_H > 0$  such that*

$$E[(nH_t)^3] = C_H < \infty.$$

(2) *For an arbitrary  $\alpha > 0$ ,*

$$Pr(nH_t > n^\alpha) \leq \frac{C_H}{n^{3\alpha}}. \quad (12)$$

For the proof of this lemma, see Section 7. Lemma 2 shows that  $nH_t$  is asymptotically uniformly integrable.

**Lemma 3** (1) *The four errors  $nB_g$ ,  $nB_t$ ,  $nG_g$ , and  $nG_t$  are all asymptotically uniformly integrable.*

(2) *For an arbitrary  $\epsilon > 0$ , following convergences in probability hold*

$$\begin{aligned} n(B_g - B_g(\epsilon)) &\rightarrow 0, \\ n(B_t - B_t(\epsilon)) &\rightarrow 0, \\ n(G_g - G_g(\epsilon)) &\rightarrow 0, \\ n(G_t - G_t(\epsilon)) &\rightarrow 0. \end{aligned}$$

For the proof of this lemma, see Section 7. Based on this Lemma,  $B_g(\epsilon)$ ,  $B_t(\epsilon)$ ,  $G_g(\epsilon)$ , and  $G_t(\epsilon)$  are referred to as the major parts of the four errors.

### 4.3 Resolution of Singularities

By Lemma3, the main region in the parameter set to be studied is

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

for a sufficiently small  $\epsilon > 0$ . By applying Hironaka's resolution theorem 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  $U_\alpha$  is a local coordinate and a proper analytic map  $g : U_\alpha \rightarrow W_\epsilon$ , expressed as  $w = g(u)$ , such that in each  $U_\alpha$ , the 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 the Jacobian determinant,  $k = (k_1, k_2, \dots, k_d)$  and  $h = (h_1, h_2, \dots, h_d)$  are sets of nonnegative integers, and  $b(u) > 0$  is a  $C^\infty$  class function. Note that  $g(u)$ ,  $k$ , and  $h$  depend on the local coordinate  $U_\alpha$ , however, to keep notation simple, we omit  $\alpha$  that

identifies the local coordinate. By applying partitions 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  $C^\infty$  class function, without loss of generality. 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 [15]. Since  $W_\epsilon$  is compact and  $g$  is a proper map,  $g^{-1}(W_\epsilon)$  is also compact. For our purpose, we need only the compact subset  $g^{-1}(W_\epsilon)$  in  $\mathcal{M}$ . Therefore, hereinafter we use the notation  $\mathcal{M}$  for  $g^{-1}(W_\epsilon)$ , 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 4** *There exists an  $L^s(q)$  valued analytic function  $\mathcal{M} \ni u \mapsto a(x, u) \in L^s(q)$  such that*

$$f(x, g(u)) = a(x, u) u^k, \tag{13}$$

$$E_X[a(X, u)] = u^k, \tag{14}$$

$$K(g(u)) = 0 \Rightarrow E_X[a(X, u)^2] = 2, \tag{15}$$

$$E_X[\|a(X)\|^s] < \infty. \tag{16}$$

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, with the result that the log density ratio function is also trivial. For the proof of this lemma, see Section 7. We define  $\|a(X)\| = \sup_{u \in \mathcal{M}} |a(X, u)|$ .

#### 4.4 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  $a^*(x, u) = E_X[a(X, u)] - a(x, u)$ . Note that  $|\xi_n(u)| = |\eta_n(g(u))|$ , where  $\eta_n(w)$  in eq.(11) is ill-defined on  $K(w) = 0$  on  $W$ , but  $\xi_n(u)$  is well-defined on  $K(g(u)) = 0$  on  $\mathcal{M}$ . In other words, resolution of singularities ensures  $\eta_n$  is well-defined. We have the following Lemma.

**Lemma 5** *The empirical process satisfies*

$$\begin{aligned} E[\|\xi_n\|^6] &< Const. < \infty \\ E[\|\nabla\xi_n\|^6] &< Const. < \infty \end{aligned}$$

where *Const.* does not depend on  $n$ , and  $\|\nabla\xi_n\| = \sum_{j=1}^d \|\partial_j\xi_n\|$ .

Let the Banach space of 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 normed space. It was proved in [19] 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 the covariance matrix of finite points. In a singular learning machine, the Fisher information matrix is singular, however,  $E_X[a(X, u)a(X, v)]$  can be understood as a generalized version of the Fisher information matrix.

Let  $\xi(u)$  be an arbitrary differentiable function. We define the average of  $f(u)$  over  $\mathcal{M}$  for the given function  $\xi(u)$  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 sum over all coordinates of  $\mathcal{M}$ ,  $\sigma$  is a constant which satisfies  $0 \leq \sigma \leq 1$ , and

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

**Lemma 6** *Assume that  $k_1 > 0$ . For an arbitrary analytic function  $\xi(u)$ ,*

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

where  $\partial_1 = (\partial/\partial u_1)$ , and  $c_1, c_2, c_3 > 0$  are constants which are determined by  $k_1, h_1, \beta$ , and  $\|\psi\| \|1/\psi\|$ .

Note that, by Lemma 6,  $G_g(\epsilon)$  is asymptotically uniformly integrable. For the proof of this Lemma, see Section 7.

Since  $w = g(u)$ , we rewrite the major parts of four errors by using the empirical process  $\xi_n(u)$ ,

$$B_g(\epsilon) = E_X[-\log E_u^0[e^{-a(X,u)u^k}|\xi_n]], \quad (17)$$

$$B_t(\epsilon) = \frac{1}{n} \sum_{j=1}^n -\log E_u^0[e^{-a(X_j,u)u^k}|\xi_n], \quad (18)$$

$$G_g(\epsilon) = E_u^0[u^{2k}|\xi_n], \quad (19)$$

$$G_t(\epsilon) = E_u^0[u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u)|\xi_n]. \quad (20)$$

In each local coordinate  $[0, 1]^d$ , without loss of generality, we can assume that there exists  $r$  such that

$$u = (x, y) \in \mathbf{R}^r \times \mathbf{R}^{r'},$$

where  $r' = d - r$ , multi-indices  $k = (k, k')$  and  $h = (h, h')$  satisfy

$$\frac{h_1 + 1}{2k_1} = \dots = \frac{h_r + 1}{2k_r} = \lambda_\alpha < \frac{h'_1 + 1}{2k'_1} \leq \dots,$$

where  $(-\lambda_\alpha)$  and  $r$  are respectively equal to the largest pole and its order of the meromorphic function that is given by the analytic continuation of

$$\int_{[0,1]^d} u^{2kz+h} du.$$

We define the multi-index  $\mu = (\mu_1, \dots, \mu_{r'}) \in \mathbf{R}^{r'}$  by

$$\mu_i = h'_i - 2k'_i \lambda_\alpha.$$

Then

$$\mu_i > h'_i - 2k'_i \left( \frac{h'_i + 1}{2k'_i} \right) = -1,$$

hence  $y^\mu$  is integrable in  $[0, 1]^{r'}$ . Both  $\lambda_\alpha$  and  $r$  depend on the local coordinate. Let  $\lambda$  be the smallest  $\lambda_\alpha$ , and  $m$  be the largest  $r$  among the coordinates for which  $\lambda = \lambda_\alpha$ . Then  $(-\lambda)$  and  $m$  are respectively equal to the largest pole and its order of the zeta function of eq.(6). Let  $\alpha^*$  be the index of the set of all coordinates that satisfy  $\lambda_\alpha = \lambda$  and  $r = m$ . As is shown by the following lemma, only the coordinates  $U_{\alpha^*}$  affect the four errors. Let  $\sum_{\alpha^*}$  denote the sum over all such coordinates.

For a given function  $f(u)$ , we adopt the notation  $f_0(y) = f(0, y)$ . For example,  $a_0(X, y) = a(X, 0, y)$ ,  $\xi_0(y) = \xi(0, y)$ , and  $\psi_0(y) = \psi(0, y)$ . The expectation value 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  $\int dy$  denotes  $\int_{[0,1]^{r'}} dy$  and

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

Then we have the following lemma.

**Lemma 7** *Let  $p \geq 0$  be a constant. There exists  $c_1 > 0$  such that, for an arbitrary  $C^1$ -class function  $f(u)$  and analytic function  $\xi(u)$ , the following inequality holds,*

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

where  $\|\nabla f\| = \sum_j \|\partial_j f\|$ .

We define four functionals of a given function  $\xi(u)$  by

$$B_g^*(\xi) \equiv \frac{1}{2} E_X [ E_{y,t}[a_0(X, y) t^{1/2} |\xi|^2] ], \quad (21)$$

$$B_t^*(\xi) \equiv G_t^*(\xi) - G_g^*(\xi) + B_g^*(\xi), \quad (22)$$

$$G_g^*(\xi) \equiv E_{y,t}[t|\xi], \quad (23)$$

$$G_t^*(\xi) \equiv E_{y,t}[t - t^{1/2} \xi_0(y)|\xi]. \quad (24)$$

Note that these four functionals do not depend on  $n$ . From the definition, we can prove the following lemma.

**Lemma 8** *For an arbitrary real measurable function  $\xi(u)$ ,*

$$G_g^*(\xi) + G_t^*(\xi) = \frac{2\lambda}{\beta}.$$

#### 4.5 Proof of Theorem 1

Firstly we show that the following convergences in probability hold.

$$nB_g(\epsilon) - B_g^*(\xi_n) \rightarrow 0, \quad (25)$$

$$nB_t(\epsilon) - B_t^*(\xi_n) \rightarrow 0, \quad (26)$$

$$nG_g(\epsilon) - G_g^*(\xi_n) \rightarrow 0, \quad (27)$$

$$nG_t(\epsilon) - G_t^*(\xi_n) \rightarrow 0. \quad (28)$$

Based on eq.(19) and eq.(23), we obtain eq.(27) by Lemma 7. Also based on eq.(20) and eq.(24), we obtain eq.(28) by Lemma 7. To prove eq.(25), 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

$$\begin{aligned} nB_g(\epsilon) &= nE_u^0[u^{2k} | \xi_n] - \frac{n}{2} E_X E_u^0[a(X, u)^2 u^{2k} | \xi_n] \\ &\quad + \frac{n}{2} E_X E_u^0[a(X, u)u^k | \xi_n]^2 + \frac{1}{6} nb_g^{(3)}(\sigma^*), \end{aligned} \quad (29)$$

where we have used  $E_X[a(X, u)] = u^k$ . The first term on the right hand side of eq.(29) is  $nG_g(\epsilon)$ . By Lemma 7, we can prove the convergence in probability

$$\begin{aligned} &\left| nE_X E_u^0[a(X, u)^2 u^{2k} | \xi_n] - E_X E_{y,t}[a_0(X, y)^2 t | \xi_n] \right| \\ &\leq \frac{c_1}{\log n} e^{4\beta \|\xi_n\|^2} E_X [ \beta \|\nabla \xi_n\| \|a(X)^2\| + \|\nabla a(X)^2\| + \|a(X)^2\| ] \rightarrow 0 \end{aligned} \quad (30)$$

holds. The proof of eq.(30) is as follows. Two empirical processes  $\xi_n(u)$  and  $\partial \xi(u)$  respectively converge in law to  $\xi(u)$  and  $\partial \xi(u)$  in the Banach space with the sup norm  $\|\cdot\|$ . Therefore, their continuous functionals  $\|\xi_n\|$ ,  $\|\partial \xi_n\|$ , and  $e^{4\beta \|\xi_n\|^2}$  also converge in law. Note that  $1/\log n$  goes to zero. In general, if a sequence of random variables converges to zero in law, then it converges to zero in probability, hence we obtain the convergence in probability eq.(30). In the following proofs, we use the same method.

Since  $E_X[a_0(X, y)] = 2$ , the sum of the first two terms of the right hand side of eq.(29) converges to zero in probability. For the third term, by using the notation  $E_X[a(X, u)a(X, v)] = \rho(u, v)$ ,  $\rho_0(u, y) = \rho(u, (0, y))$ , and  $\rho_{00}(y', y) = \rho((0, y'), (0, y))$ , and applying Lemma 7,

$$\begin{aligned} &\left| nE_X E_u^0[a(X, u)u^k | \xi_n]^2 - E_{y,t}[a_0(X, y)t^{1/2} | \xi_n]^2 \right| \\ &\leq \left| \sqrt{n} E_u^0 \left[ u^k (\sqrt{n} E_v^0 [\rho(u, v)v^k] - E_{y,t}[\rho_0(u, y)t^{1/2}]) \right] \right| \\ &\quad + \left| E_{y,t} \left[ t^{1/2} (\sqrt{n} E_u^0 [\rho_0(u, y)u^k] - E_{y',t'}[\rho_{00}(y', y)(t't)^{1/2}]) \right] \right| \\ &\leq \frac{c_1 \sqrt{n}}{\log n} E_u^0[u^k] e^{4\beta \|\xi_n\|^2} (\beta \|\nabla \xi_n\| \|\rho\| + \|\nabla \rho\| + \|\rho\|) \\ &\quad + \frac{c_1}{\log n} e^{4\beta \|\xi_n\|^2} (\beta \|\nabla \xi_n\| \|\rho\| + \|\nabla \rho\| + \|\rho\|), \end{aligned} \quad (31)$$

where ' $|\xi_n$ ' is omitted to keep the notation simple. The equation (31) converges to zero in probability by Lemma 6. Therefore the difference between the third term and  $B_g^*(\xi_n)$  converges to zero in probability. For the last term, we have

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

By applying Lemma 6,

$$|nb_g^{(3)}(\sigma^*)| \leq \frac{6c_2}{n^{1/2}} E_X \left[ \|a(X)\|^3 \{1 + \|\xi_n\|^3 + \|\partial\xi_n\|^3 + \|a(X)\|^{3/2} + \|\partial a(X)\|^{3/2}\} \right], \quad (32)$$

which shows that  $nb_g^{(3)}(\sigma^*)$  converges to zero in probability. Hence eq.(25) is proved. Let us prove eq.(26). 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

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

Then by applying Lemma 6,  $nb_t^{(3)}(\sigma^*)$  converges to zero in probability in the same way as for eq.(32). By the same methods as used with eq.(30) and eq.(31), replacing respectively  $E_X[\|a(X)^2\|]$  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

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

hold, with the result that the convergence in probability

$$nB_t(\epsilon) - nG_t(\epsilon) + nG_g(\epsilon) - nB_g(\epsilon) \rightarrow 0. \quad (33)$$

holds. Therefore eq.(26) is obtained. By combining eq.(25)-eq.(28) with Lemma 3 (2), the following convergences in probability hold,

$$nB_g - B_g^*(\xi_n) \rightarrow 0, \quad (34)$$

$$nB_t - B_t^*(\xi_n) \rightarrow 0, \quad (35)$$

$$nG_g - G_g^*(\xi_n) \rightarrow 0, \quad (36)$$

$$nG_t - G_t^*(\xi_n) \rightarrow 0. \quad (37)$$

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  $\xi_n \rightarrow \xi$ , the convergences in law

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

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

#### 4.6 Proof of Theorem 2

Let  $\{(x_i, g_i); i = 1, 2, \dots, N\}$  be a set of independent random variables which are subject to the probability distribution

$$q(x) = \frac{e^{-g^2/2}}{\sqrt{2\pi}}.$$

A tight gaussian process is defined by

$$\zeta_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n a(x_i, u) g_i.$$

Then, in the same way as the convergence in law  $\xi_n(u) \rightarrow \xi(u)$  was proved, the convergence in law  $\zeta_n(u) \rightarrow \xi(u)$  can be proved, because  $\zeta_n(u)$  has the same expectation and covariance.

$$E[\zeta_n(u)] = 0, \tag{38}$$

$$E[\zeta_n(u)\zeta_n(v)] = E_X[a(X, u)a(X, v)]. \tag{39}$$

In other words, both  $\zeta_n(u)$  and  $\xi_n(u)$  converge in law to the same random process  $\xi(u)$ . Moreover, we can prove that  $\zeta_n(u)$  satisfies  $E[\|\zeta_n\|^s] < \infty$  ( $s \geq 6$ ) in the same way. Therefore we can prove equations of a gaussian random process  $\xi(u)$  by using the convergence in law  $\zeta_n(u) \rightarrow \xi(u)$ . Since  $g_i$  is subject to the standard normal distribution,

$$E[g_i F(g_i)] = E\left[\frac{\partial}{\partial g_i} F(g_i)\right] \tag{40}$$

holds for a differentiable function of  $F(x)$  which satisfies  $|F(x)|/|x|^k, |F(x)'|/|x|^k \rightarrow 0$  ( $|x| \rightarrow \infty$ ) for some  $k > 0$ .

Let us prove Theorem 2. We use the notation,

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

where  $u = (x, y)$ . Also we define the expectation value of  $f(u, t)$  for a given function  $\xi(u)$ ,

$$\langle f(u, t) \rangle_\xi = \frac{\int du^* \int_0^\infty dt f(u, t) t^{\lambda-1} e^{-\beta t + \xi(u)\beta\sqrt{t}}}{\int du^* \int_0^\infty dt t^{\lambda-1} e^{-\beta t + \xi(u)\beta\sqrt{t}}}.$$

Note that Lemma 8 is equivalent to

$$\langle 2t \rangle_\xi - \langle \sqrt{t} \xi(u) \rangle_\xi = \frac{2\lambda}{\beta}.$$

By this equation and  $|\sqrt{t}\xi(u)| \leq (t + \xi(u)^2)/2$ ,

$$\langle t \rangle_\xi \leq \frac{4\lambda}{3\beta} + \frac{\langle \xi(u)^2 \rangle_\xi}{3}, \quad (41)$$

$$\langle |\sqrt{t}\xi(u)| \rangle_\xi \leq \frac{2\lambda}{3\beta} + \frac{2\langle \xi(u)^2 \rangle_\xi}{3}, \quad (42)$$

hold for an arbitrary function  $\xi(u)$ . Note that  $\langle \xi(u)^2 \rangle_\xi \leq \|\xi\|^2$ , because  $\|\cdot\|$  is the sup norm. The expectations of  $B_g^*$ ,  $G_g^*$ , and  $G_t^*$  can be written by

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

where  $A$  is a constant defined by

$$A \equiv E[\frac{\int du^* \xi(u) Y'(\xi(u))}{Z(\xi)}].$$

We introduce  $A_n$  by using  $\zeta_n(u)$ ,

$$\begin{aligned} A_n &= E[\frac{\int du^* \zeta_n(u) Y'(\zeta_n(u))}{Z(\zeta_n)}] \\ &= \beta E[\langle \zeta_n \sqrt{t} \rangle_{\zeta_n}] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \beta E[g_i \langle a(x_i, u) \sqrt{t} \rangle_{\zeta_n}]. \end{aligned}$$

Then by eq.(42),  $\langle \zeta_n(u) \sqrt{t} \rangle_{\zeta_n}$  is asymptotically uniformly integrable, hence  $A_n \rightarrow A$  ( $n \rightarrow \infty$ ). On the other hand, we define

$$\begin{aligned} B_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \beta E[\frac{\partial}{\partial g_i} \langle a(x_i, u) \sqrt{t} g_i \rangle_{\zeta_n}] \\ &= E[\int du^* \{ \frac{1}{\sqrt{n}} \sum_{i=1}^n a(x_i, u) \frac{\partial}{\partial g_i} \} \frac{Y'(\zeta_n(u))}{Z(\zeta_n)}]. \end{aligned}$$

Then by using

$$\begin{aligned} \frac{\partial}{\partial g_i} \left( \frac{Y'(\zeta_n(u))}{Z(\zeta_n)} \right) &= \frac{Y''(\zeta_n(u)) a(x_i, u)}{\sqrt{n} Z(\zeta_n)} \\ &\quad - \frac{Y'(\zeta_n(u))}{\sqrt{n} Z(\zeta_n)^2} \int dv^* Y'(\zeta_n(v)) a(x_i, v), \end{aligned}$$

we have

$$\beta \frac{\partial}{\partial g_i} \langle a(x_i, u) \sqrt{t} \rangle_{\zeta_n} = \frac{\beta^2}{\sqrt{n}} \left( \langle a(x_i, u)^2 t \rangle_{\zeta_n} - \langle a(x_i, u) \sqrt{t} \rangle_{\zeta_n}^2 \right).$$

Hence

$$B_n = E[\frac{\beta^2}{n} \sum_{i=1}^n \left( \langle a(x_i, u)^2 t \rangle_{\zeta_n} - \langle a(x_i, u) \sqrt{t} \rangle_{\zeta_n}^2 \right)].$$

Also,

$$\zeta_n(u)^2 \leq \frac{1}{n} \left( \sum_{k=1}^n a(x_k, u)^2 \right) \left( \sum_{k=1}^n g_k^2 \right). \quad (43)$$

From eq.(41), eq.(42), and eq.(43), both  $\langle a(x_i, u)\sqrt{t} \rangle_{\zeta_n}$  and  $(\partial/\partial g_i)\langle a(x_i, u)\sqrt{t} \rangle_{\zeta_n}$  are bounded by a finite sum of quadratic forms of  $g_i$ . Hence by eq.(40),  $A_n = B_n$ . Lastly, since  $\langle (1/n) \sum_{i=1}^n a(x_i, u)^2 t \rangle_{\zeta_n}$  and  $\langle (1/\sqrt{n}) \sum_{i=1}^n a(x_i, u)\sqrt{t} \rangle_{\zeta_n}^2$  are asymptotically uniformly integrable by eq.(41), eq.(42), we obtain  $B_n \rightarrow B$ , where

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

Here we have used  $E_X[a(X, u)^2] = 2$  for  $K(g(u)) = 0$  by Lemma 4. Since  $A_n = B_n$ ,  $A_n \rightarrow A$ , and  $B_n \rightarrow B$ , we have  $A = B$ . Therefore

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

which completes Theorem 2. (Q.E.D.)

#### 4.7 Proof of Theorem 3

From Lemma 8, it follows that

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

Then by Theorem 1 and Lemma 3, we obtain Theorem 3. (Q.E.D.)

## 5 Discussion

In this section, we discuss the theorems in this paper.

Firstly, Theorem 1 was derived from definitions of the four errors. As is shown in the proof,

$$B_t = G_t - \hat{G}_g + \hat{B}_g + o_p\left(\frac{1}{n}\right),$$

where  $o_p(1/n)$  is a random variable whose order is smaller than  $1/n$  and

$$\begin{aligned} \hat{G}_g &= \frac{1}{2n} \sum_{j=1}^n E_w \left[ \left( \log \frac{q(X_j)}{p(X_j|w)} \right)^2 \right], \\ \hat{B}_g &= \frac{1}{2n} \sum_{j=1}^n E_w \left[ \log \frac{q(X_j)}{p(X_j|w)} \right]^2. \end{aligned}$$

Here convergences in probability  $n(\hat{G}_g - G_g) \rightarrow 0$  and  $n(\hat{B}_g - B_g) \rightarrow 0$  hold. We need the information about the true distribution to calculate both  $\hat{G}_g$  and  $\hat{B}_g$ , however, we do

not need it to calculate

$$V \equiv 2(\hat{G}_g - \hat{B}_g) = \frac{1}{n} \sum_{j=1}^n E_w[(\log p(X_j|w))^2] - \frac{1}{n} \sum_{j=1}^n E_w[\log p(X_j|w)]^2.$$

The random variable  $V$  is the variance of the *a posteriori* distribution. By using  $V$ ,  $WAIC_1$  and  $WAIC_2$  can be replaced by

$$WAIC_1 = BL_t + \beta V,$$

$$WAIC_2 = GL_t + \beta V.$$

The third criterion  $WAIC_3$

$$WAIC_3 = BL_t - GL_t + \hat{G}_g - \hat{B}_g$$

can be used as an index to examine how precisely the asymptotic theory holds. In other words, the value  $|WAIC_3|$  is the error of the asymptotic theory.

Secondly, let us study Theorem 2. This theorem is essentially derived from the fact that the empirical process  $\xi_n(u)$  converges to the tight gaussian process  $\xi(u)$  and that the partial integration formula

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

holds for  $\xi(u)$ .

Thirdly, Theorem 3 is proved by the property of the integral

$$Z_\lambda(\beta|a) = \sum_{\alpha^*} \int du^* \int_0^\infty dt t^{\lambda-1} e^{-\beta t + a\beta\sqrt{t}}.$$

That is to say, Theorems 2 and 3 are essentially proved by partial integration.

Fourthly, in this paper, we proved three results eqs.(2), (3), and (7). The two relations of eq.(2) and eq.(3) hold universally, independently of singularities, whereas the third relation of eq.(7) depends strongly on singularities. To determine the values of the four errors, one more relation is needed. However, it seems that there is no such relation. Hence in order to determine the four errors, we may have to evaluate at least one of the four errors. For example

$$E[G_t] = \frac{\partial}{\partial \beta} E\left[-\log Z_\lambda(\beta|\xi(u))\right].$$

It is conjectured that this value is determined by the generalized Fisher information matrix  $E_X[a(X, u)a(X, v)]$  on the set of true parameters  $\mathcal{M}_0$ . To investigate this problem in a mathematically rigorous way is a problem for future study.

Fifthly, we assumed that the log density ration function  $f(x, w)$  is an  $L^s(q)$ -valued analytic function. Even if  $f(x, )$  is not analytic, if  $f(x, ) = u^k a(x, )$  holds and  $a(x, )$

satisfies some assumptions proved in Lemmas, then the theorem holds. However, if  $f(x, \cdot)$  is not analytic, then there is examples in which  $f(x, \cdot) = u^k a(x, \cdot)$  does not hold and it is not easy to judge whether  $f(x, \cdot) = u^k a(x, \cdot)$  holds or not. It is the future study the equations of states in this paper in the more weak conditions.

Lastly, let us compare the result of this paper with the asymptotic theory of regular statistical models. In regular statistical models, the set of true parameters consists of just one point,  $W_0 = \{w_0\}$ . By the transform  $w = g_0(u) = w_0 + I(w_0)^{1/2}u$ , where  $I(w)$  is the Fisher information matrix,

$$\begin{aligned} K(g_0(u)) &\cong \frac{1}{2}|u|^2, \\ K_n(g_0(u)) &\cong \frac{1}{2}|u|^2 - \frac{\xi_n}{\sqrt{n}} \cdot u, \end{aligned}$$

where  $I(w_0)$  is Fisher information matrix and  $\xi_n = (\xi_n(1), \xi_n(2), \dots, \xi_n(d))$  is defined by

$$\xi_n(k) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial u_k} \log p(X_i | g_0(u)) \Big|_{u=0}.$$

Here each  $\xi_n(k)$  converges in law to the standard normal distribution. Statistical learning theory for regular models is based on the convergence in law  $\xi_n \rightarrow \xi$ , whereas that for singular models, it is based on the fact that  $\xi_n(u) \rightarrow \xi(u)$ .

## 6 Conclusion

Based on singular learning theory, we established the equations of states in learning, and proposed widely applicable information criteria.

### Acknowledgment

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

## 7 Appendix

### 7.1 Proof of Lemma 1

Since  $B_g(a)$  is the Kullback-Leibler distance from  $q(x)$  to  $E_w[p(x|w)|_{K(w) \leq a}]$ ,  $B_g(a) \geq 0$ . Using Jensen's inequality,

$$E_w[e^{-f(x,w)} |_{K(w) \leq a}] \geq e^{-E_w[f(x,w) |_{K(w) \leq a}]} \quad (\forall x),$$

we have  $B_g(a) \leq G_g(a)$  and  $B_t(a) \leq G_t(a)$ . If  $0 < K(w) \leq a$ ,

$$K_n(w) = K(w) - \sqrt{K(w)} \eta_n(w)$$

$$\begin{aligned}
&\geq (\sqrt{K(w)} - \frac{\eta_n(w)}{2})^2 - \frac{\eta_n(w)^2}{4} \\
&\geq -\frac{1}{4}H_t(a).
\end{aligned}$$

Hence  $-H_t(a)/4 \leq G_t(a)$ . Also we have

$$K_n(w) \leq \frac{3}{2}K(w) + \frac{1}{2}\eta_n(w)^2. \quad (44)$$

Therefore  $G_t(a) \leq \frac{3}{2}G_g(a) + \frac{1}{2}H_t(a)$ . (Q.E.D.)

## 7.2 Proof of Lemma 2

(1) For any  $\epsilon > 0$  and  $a > 0$ , by the definition of  $\eta_n(w)$ ,

$$\sqrt{n} \eta_n(w) = \frac{1}{\sqrt{K(w)}} \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^n (E_X[f(X, w)] - f(X_j, w))$$

is an empirical process and  $f(x, w)$  is an analytic function of  $w$ , hence

$$E[\sup_{\epsilon < K(w) < a} |\sqrt{n}\eta_n|^6] < \text{const.}$$

[23][19][20]. It is proven in Lemma 5 that  $E[(nH_t(\epsilon))^3]$  also satisfies the same inequality.

(2) Let the random variable  $S$  be defined by

$$S = \begin{cases} 1 & (\text{if } nH_t > n^\alpha) \\ 0 & (\text{otherwise}) \end{cases}.$$

Then  $E[S] = Pr(nH_t > n^\alpha)$  and

$$C_H = E[(nH_t)^3] \geq E[(nH_t)^3 S] \geq E[S] n^{3\alpha},$$

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

## 7.3 Proof of Lemma 3

We use the notation,

$$\begin{aligned}
S_1(f(w)) &= \int_{K(w) \geq \epsilon} f(w) e^{-n\beta K_n(w)} \varphi(w) dw, \\
S_0(f(w)) &= \int_{K(w) < \epsilon} f(w) e^{-n\beta K_n(w)} \varphi(w) dw.
\end{aligned}$$

By using the inequality,

$$\frac{1}{2}K(w) - \frac{1}{2}\eta_n(w)^2 \leq K_n(w) \leq \frac{3}{2}K(w) + \frac{1}{2}\eta_n(w)^2,$$

we have inequalities for arbitrary  $f(w), g(w) > 0$ ,

$$\begin{aligned}
S_1(f(w)) &\leq (\sup_w f(w)) e^{-n\beta\epsilon/2} \exp(\frac{\beta}{2}nH_t), \\
S_0(g(w)) &\geq c_0 (\inf_w g(w)) n^{-\lambda} \exp(-\frac{\beta}{2}nH_t),
\end{aligned}$$

where  $(-\lambda)$  is the largest pole of  $\zeta(z)$  and  $c_0 > 0$  is a constant which satisfies the inequality [15]

$$\int_{K(w) < \epsilon} \exp\left(-\frac{3\beta n}{2}K(w)\right)\varphi_1(w)dw \geq \frac{c_0}{n^\lambda}.$$

Hence

$$\frac{S_1(f(w))}{S_0(g(w))} \leq \frac{\sup_w f(w)}{\inf_w g(w)} s(n),$$

where

$$s(n) = \frac{n^\lambda}{c_0} e^{-n\beta\epsilon/2+n\beta H_t}.$$

Then

$$|\log s(n)| \leq n\beta\epsilon/2 + n\beta H_t + \lambda \log n + |\log c_0|.$$

By using the function  $M(x) \geq 0$  used in eq.(1), we define  $M_n$  by

$$M_n \equiv \frac{1}{n} \sum_{j=1}^n M(X_j).$$

Then

$$E[M_n^3] \leq E[(\sum(M(X_j)/n)^3)] \leq E[(\sum M(X_j)^3/n)] = E_X[M(X)^3] < \infty.$$

(1) Firstly, we study Bayes generalization error.

$$\begin{aligned} n(B_g - B_g(\epsilon)) &= nE_X\left[-\log \frac{E_w[e^{-f(X,w)}]}{E_w[e^{-f(X,w)}|_{K(w) \leq \epsilon}]}\right] \\ &= nE_X\left[-\log\left(1 + \frac{S_1(e^{-f(X,w)})}{S_0(e^{-f(X,w)})}\right) + \log\left(1 + \frac{S_1(1)}{S_0(1)}\right)\right]. \end{aligned}$$

Therefore

$$\begin{aligned} n|B_g - B_g(\epsilon)| &\leq nE_X\left[\log\left(1 + \frac{S_1(e^{-f(X,w)})}{S_0(e^{-f(X,w)})}\right) + \log\left(1 + \frac{S_1(1)}{S_0(1)}\right)\right] \\ &\leq nE_X\left[\log(1 + s(n) e^{2\sup_w |f(X,w)|}) + \log(1 + s(n))\right] \\ &\leq nE_X\left[\log(1 + s(n) e^{2M(X)})\right] + ns(n). \end{aligned}$$

The second term converges to zero in probability because of Lemma 2. Let  $f_1(n)$  be the first term,

$$f_1(n) = nE_X[\log(1 + s(n) e^{2M(X)})].$$

Let us define

$$\Theta_1(x) = \begin{cases} 1 & (2M(x) > n\beta\epsilon/4) \\ 0 & (2M(x) \leq n\beta\epsilon/4) \end{cases}. \quad (45)$$

Then by using  $\log(1+x) \leq x$  and  $\log(1+e^x) \leq |x| + 1$ ,

$$\begin{aligned} f_1(n) &= nE_X[(1 - \Theta_1(X)) \log(1 + s(n) e^{2M(X)})] \\ &\quad + nE_X[\Theta_1(X) \log(1 + s(n) e^{2M(X)})] \\ &\leq ns(n) \exp(n\beta\epsilon/4) \\ &\quad + nE_X[\Theta_1(X)(2M(X) + |\log s(n)| + 1)], \end{aligned}$$

which converges to zero in probability because, from the inequality eq.(1),

$$\begin{aligned} E_X[\Theta_1(X)M(X)] &\leq \left(\frac{4}{n\beta\epsilon}\right)^5 E[M(X)^6], \\ E_X[\Theta_1(X)] &\leq \left(\frac{4}{n\beta\epsilon}\right)^6 E[M(X)^6]. \end{aligned}$$

It follows that  $n(B_g - B_g(\epsilon)) \rightarrow 0$ . Secondly, we prove the convergence in probability  $n(B_t - B_t(\epsilon)) \rightarrow 0$ .

$$\begin{aligned} n|B_t - B_t(\epsilon)| &\leq \sum_{j=1}^n \{\log(1 + s(n) e^{2\sup|f(X_j, w)|}) + \log(1 + s(n))\} \\ &\leq \sum_{j=1}^n \log(1 + s(n)e^{2M(X_j)}) + n \log(1 + s(n)) \equiv L_n \end{aligned} \quad (46)$$

where eq.(46) is the definition of  $L_n$ . To prove the convergence in probability  $L_n \rightarrow 0$ , it is sufficient to prove convergence in mean  $E[L_n] \rightarrow 0$ . Let the random variable  $\Theta_2$  be

$$\Theta_2 = \begin{cases} 1 & (nH_t > n\beta\epsilon/4) \\ 0 & (nH_t \leq n\beta\epsilon/4) \end{cases} \quad (47)$$

Then

$$\begin{aligned} E[L_n] &= E[L_n(1 - \Theta_2)] + E[L_n\Theta_2] \\ &\leq nE_X[\log(1 + (n^\lambda/c_0) e^{2M(X) - n\beta\epsilon/4})] \\ &\quad + n^{\lambda+1} \exp(-n\beta\epsilon/4)/c_0 \\ &\quad + E[\Theta_2 n(2M_n + |\log s(n)| + 1)] \\ &\quad + E[\Theta_2 n(|\log s(n)| + 1)] \end{aligned}$$

The first term goes to zero can be proved in the same way as  $f_1(n) \rightarrow 0$ . The second term goes to zero as a real sequence. Both the third and fourth terms go to zero because

$$\begin{aligned} E[\Theta_2 n M_n] &\leq nPr(nH_t > n)^{1/2} E[M_n^2]^{1/2}, \\ E[n\Theta_2(n\beta\epsilon)] &= n^2\beta\epsilon Pr(nH_t > n\beta\epsilon/4), \\ E[n\Theta_2(nH_t)] &\leq nPr(nH_t > n\beta\epsilon/4)^{1/2} E[(nH_t)^2]^{1/2}, \end{aligned}$$

and by using Lemma 2. Thus we obtain  $n(B_t - B_t(\epsilon)) \rightarrow 0$ . Thirdly, the Gibbs generalization error can be estimated as

$$\begin{aligned} n|G_g - G_g(\epsilon)| &\leq \left| n \frac{S_0(K(w)) + S_1(K(w))}{S_0(1) + S_1(1)} - \frac{nS_0(K(w))}{S_0(1)} \right| \\ &\leq \frac{nS_1(K(w))}{S_0(1)} + \frac{nS_0(K(w))S_1(1)}{S_0(1)^2} \\ &\leq 2n \bar{K} s(n), \end{aligned} \quad (48)$$

which converges to zero in probability. Lastly, in the same way, the Gibbs training error satisfies

$$\begin{aligned} n|G_t - G_t(\epsilon)| &\leq 2n s(n) \sup_w |K_n(w)| \\ &\leq 2n s(n) M_n \end{aligned}$$

which converges to zero in probability.

(2) Firstly, from Lemma 2,  $nH_t$  is AUI. Secondly, let us prove  $nB_t$  is AUI. Let  $L_n$  be the term in eq.(46). Then

$$|nB_t| \leq |nB_t(\epsilon)| + L_n.$$

Moreover, by employing a function,

$$b(s) = -\frac{1}{n} \sum_{j=1}^n \log E_w[e^{-sf(X_j, w)}],$$

there exists  $0 < s^* < 1$  such that

$$nB_t = nb(1) = \sum_{j=1}^n \frac{E_w[f(X_j, w)e^{-s^*f(X_j, w)}]}{E_w[e^{-s^*f(X_j, w)}]}.$$

Hence

$$|nB_t| \leq \sum_{j=1}^n \sup_w |f(X_j, w)| \leq nM_n$$

Therefore

$$|nB_t| \leq |nB_t(\epsilon)| + B^*,$$

where

$$B^* \equiv \begin{cases} nM_n & (nH_t > \epsilon\beta n/4) \\ L_n & (nH_t \leq \epsilon\beta n/4) \end{cases}.$$

By summing the above equations,

$$E[|nB_t|^{3/2}] \leq E[2|nB_t(\epsilon)|^{3/2}] + E[2(B^*)^{3/2}].$$

In Lemma 5, we prove that  $E[|nB_t(\epsilon)|^{3/2}] < \infty$ . By Lemma 2 (2) with  $\delta$  such that  $n^\delta = \epsilon\beta n/4$ , we have  $P(H_t > \epsilon\beta/4) \leq C'_H/n^3$ , hence

$$\begin{aligned} E[(B^*)^{3/2}] &\leq E[\Theta_2(B^*)^{3/2}] + E[(1 - \Theta_2)(B^*)^{3/2}] \\ &\leq E[(nM_n)^3]^{1/2} E[\Theta_2]^{1/2} \\ &\quad + E[(1 - \Theta_2)(L_n)^{3/2}] < \infty. \end{aligned}$$

The first term is finite because  $E[\Theta_2] = Pr(nH_t > n\beta\epsilon/4)$ . Finiteness of the second term can be proved in the same way as proving that  $E[(1 - \Theta_2)L_n] \rightarrow 0$ . Hence  $|nB_t|$  is AUI. Lastly, we show that  $nG_g$  is AUI. From eq.(48),

$$0 \leq nG_g \leq nG_g(\epsilon) + 2n s(n) \bar{K}.$$

Moreover, always  $nG_g \leq n\bar{K}$ , by definition. Therefore

$$nG_g \leq nG_g(\epsilon) + K^*$$

where

$$\begin{aligned} K^* &\equiv \begin{cases} n\bar{K} & (nH_t > n^{2/3}) \\ \bar{K} n s(n) & (nH_t \leq n^{2/3}) \end{cases} \\ &\leq \begin{cases} n\bar{K} & (nH_t > n^{2/3}) \\ \bar{K} e^{-n\beta\epsilon/3} & (nH_t \leq n^{2/3}) \end{cases}. \end{aligned}$$

Then

$$0 \leq E[(nG_g)^{3/2}] \leq E[2(nG_g(\epsilon))^{3/2}] + E[2(K^*)^{3/2}].$$

It is proven in Lemma 6 that  $E[(nG_g(\epsilon))^{3/2}] < \infty$ . By Lemma 2 with  $\delta = 2/3$ , we have  $P(nH_t > n^{2/3}) \leq C_H/n^2$ , hence

$$E[(K^*)^{3/2}] \leq n^{3/2} \bar{K}^{3/2} \frac{C_H}{n^2} + \bar{K} e^{-n\beta\epsilon/2} < \infty.$$

Hence  $nG_g$  is AUI. Since  $E[(nH_t)^3] < \infty$ ,  $E[(nB_t)^{3/2}] < \infty$ , and  $E[(nG_g)^{3/2}] < \infty$  all four errors are also AUI by Lemma 1. (Q.E.D.)

#### 7.4 Proof of Lemma 4

By the definition of the Kullback-Leibler distance and  $f(x, g(u)) = \log(q(x)/p(x|g(u)))$ , for arbitrary  $u \in \mathcal{M}$ ,

$$\begin{aligned} K(g(u)) &= \int f(x, g(u))q(x)dx \\ &= \int (e^{-f(x, g(u))} + f(x, g(u)) - 1)q(x)dx \\ &= \int \frac{f(x, g(u))^2}{2} e^{-t^* f(x, g(u))} q(x)dx, \end{aligned}$$

where  $0 < t^* < 1$ . Let  $U'$  be a neighborhood of  $u = 0$ . For arbitrary  $L > 0$  the set  $D_L$  is defined by

$$D_L \equiv \{x \in \mathbf{R}^N; \sup_{u \in U'} |f(x, g(u))| \leq L\}.$$

Then for any  $u \in U'$ ,

$$u^{2k} \geq \int_{D_L} \frac{f(x, g(u))^2}{2} e^{-L} q(x)dx,$$

with the result that, for any  $u^k \neq 0$  ( $u \in U'$ ),

$$1 \geq e^{-L} \int_{D_L} \frac{f(x, g(u))^2}{2u^{2k}} q(x)dx. \quad (49)$$

Since  $f(x, g(u))$  is an  $L^s(q)$ -valued real analytic function, it is given by an absolutely convergent power series,

$$\begin{aligned} f(x, g(u)) &= \sum_{\alpha} a_{\alpha}(x)u^{\alpha} \\ &= a(x, u)u^k + b(x, u)u^k, \end{aligned}$$

where

$$\begin{aligned} a(x, u) &= \sum_{\alpha \geq k} a_\alpha(x) u^{\alpha-k}, \\ b(x, u) &= \sum_{\alpha < k} a_\alpha(x) u^{\alpha-k}, \end{aligned}$$

and  $\sum_{\alpha \geq k}$  denotes the sum over indices that satisfy

$$\alpha_i \geq k_i \quad (i = 1, 2, \dots, d) \quad (50)$$

and  $\sum_{\alpha < k}$  denotes the sum over indices that do not satisfy eq.(50). Here  $a(x, u)$  is an  $L^s(q)$ -valued real analytic function. From eq.(49), for an arbitrary  $u^k \neq 0$  ( $u \in U'$ ),

$$\begin{aligned} 1 &\geq e^{-L} \int_{D_L} (a(x, u) + b(x, u))^2 q(x) dx \\ &\geq \frac{e^{-L}}{2} \int_{D_L} b(x, u)^2 q(x) dx - e^{-L} \int_{D_L} a(x, u)^2 q(x) dx. \end{aligned}$$

Here  $|a(x, u)|$  is a bounded function of  $u \in U'$ . If  $b(x, u) \equiv 0$  does not hold, then  $|b(x, u)| \rightarrow \infty$  ( $u \rightarrow 0$ ), hence we can choose  $u$  and  $D_L$  so that the above inequality does not hold. Therefore, we have  $b(x, u) \equiv 0$ , which shows eq.(13). From

$$u^{2k} = \int f(x, g(u)) q(x) dx = \int a(x, u) u^k q(x) dx,$$

we obtain eq.(14). To prove eq.(15), it is sufficient to prove  $E_X[a(X, u)^2] = 2$  when  $K(g(u)) = 0$ . Let the Taylor expansion of  $f(x, g(u))$  be

$$f(x, g(u)) = \sum_{\alpha} a_\alpha(x) u^\alpha.$$

Then

$$|a_\alpha(x)| \leq \frac{M(x)}{R^\alpha} \quad (51)$$

where  $R$  is the associated convergence radii and

$$a(x, u) = \sum_{\alpha \geq k} a_\alpha(x) u^{\alpha-k}.$$

Hence

$$\begin{aligned} |a(x, u)| &\leq \sum_{\alpha \geq k} \frac{M(x)}{R^\alpha} r^{\alpha-k} \\ &= c_1 \frac{M(x)}{R^k}, \end{aligned}$$

where  $c_1 > 0$  is a constant. For arbitrary  $u$  ( $u^k \neq 0$ ),

$$1 = \int \frac{a(x, u)^2}{2} e^{-t^* a(x, u) u^k} q(x) dx,$$

where  $0 < t^* < 1$ . Put

$$S(x, u) = \frac{a(x, u)^2}{2} e^{-t^* a(x, u) u^k} q(x).$$

Then

$$\begin{aligned} S(x, u) &\leq c_1 \frac{M(x)^2}{R^{2k}} \max\{1, e^{-a(x, u) u^k}\} q(x) \\ &= c_1 \frac{M(x)^2}{R^{2k}} \max\{q(x), p(x|w)\} \\ &\leq c_1 \frac{M(x)^2}{R^{2k}} Q(x). \end{aligned}$$

By the fundamental condition (A.3),  $M(x)^2 Q(x)$  is an integrable function, hence  $S(x, u)$  is bounded by the integrable function. By using Lebesgue's convergence theorem, as  $u^k \rightarrow 0$ , we obtain

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

for any  $u$  that satisfies  $u^{2k} = 0$ , which proves eq.(15). Lastly, since  $f(x, u)$  is an  $L^s(q)$  valued analytic function,  $a(x, u)$  is also an  $L^s(q)$  valued analytic function. Moreover, eq.(51) shows eq.(16). (Q.E.D.)

## 7.5 Proof of Lemma 5

The proof is given in [19] and Theorem 39 in [20].

## 7.6 Proof of Lemma 6

Let  $u = (u_1, u_2, \dots, u_d)$ . Since at least one of non-negative integers  $k_1, \dots, k_d$  is not equal to zero, we can assume  $k_1 \geq 1$  without loss of generality. Put  $g(u) = u_2^{k_2} \dots u_d^{k_d}$  and  $h(u) = u_2^{h_2} \dots u_d^{h_d}$ . Then  $u^k = u_1^{k_1} g(u)$ ,  $u^h = u_1^{h_1} h(u)$ , where either  $g(u)$  or  $h(u)$  do not depend on  $u_1$ . We adopt the notation,

$$\begin{aligned} N_p &= \sum_{\alpha} \int_{[0,1]^d} u_1^{p k_1 + h_1} g(u)^p h(u) e^{-\beta n u^{2k} + f(u)} du, \\ f(u) &= \beta \sqrt{n} u^k \xi(u) + \sigma u^k a(X, u), \end{aligned}$$

By the definition and  $c_1 = \|\psi\|/\|1/\psi\|$ ,

$$E_u^\sigma[u^{2k}|\xi] \leq c_1 \frac{N_2}{N_0}.$$

By applying partial integration to  $N_2$ ,

$$\begin{aligned} N_2 &= - \sum_{\alpha} \int_{[0,1]^d} \frac{h(u)}{2\beta n k_1} u_1^{h_1+1} e^{f(u)} \partial_1(e^{-\beta n u^{2k}}) du \\ &\leq \sum_{\alpha} \int_{[0,1]^d} \frac{h(u)}{2\beta n k_1} \partial_1(u_1^{h_1+1} e^{f(u)}) e^{-\beta n u^{2k}} du \\ &= \sum_{\alpha} \int_{[0,1]^d} \frac{u_1^{h_1} h(u)}{2\beta n k_1} e^{-\beta n u^{2k} + f(u)} (h_1 + 1 + u_1 \partial_1 f(u)) du. \end{aligned}$$

From the definition of  $f(u)$

$$\begin{aligned} u_1 \partial_1 f(u) &= \beta \sqrt{n} (k_1 u^k \xi(u) + u^k \partial_1 \xi(u)) \\ &\quad + \sigma k_1 u^k a(X, u) + \sigma u^k \partial_1 a(X, u). \end{aligned}$$

By using inequalities

$$\begin{aligned} |\sqrt{n} u^k \xi(u)| &\leq \frac{1}{2} (nu^{2k} + \xi(u)^2), \\ |\sqrt{n} u^k \partial_1 \xi(u)| &\leq \frac{1}{2} (nu^{2k} + (\partial_1 \xi(u))^2), \end{aligned}$$

and  $|u^k| \leq 1$ ,

$$|u_1 \partial_1 f(u)| \leq \frac{\beta}{2} \{k_1 (nu^{2k} + \|\xi\|^2) + nu^{2k} + \|\partial_1 \xi\|^2\} + k_1 \sigma \|a\| + \sigma \|\partial_1 a\|.$$

Hence

$$\frac{N_2}{N_0} \leq \frac{1}{2nk_1} \left\{ \frac{n(k_1 + 1)}{2} \frac{N_2}{N_0} + h_1 + 1 + k_1 \|\xi\|^2 + \|\partial_1 \xi\|^2 + \frac{k_1 \sigma \|a\| + \sigma \|\partial_1 a\|}{\beta} \right\},$$

with the result that

$$z_1 \frac{N_2}{N_0} \leq \frac{1}{2nk_1} \left\{ h_1 + 1 + k_1 \|\xi\|^2 + \|\partial_1 \xi\|^2 + \frac{k_1 \sigma \|a\| + \sigma \|\partial_1 a\|}{\beta} \right\},$$

where  $z_1 = (3k_1 - 1)/(4k_1)$ , which shows the first half of the lemma. Let us prove the latter half. Firstly,

$$E_u^\sigma [u^{3k} |\xi|] \leq c_3 \frac{N_3}{N_0}.$$

In the same way as for the first half, by applying partial integration, we have

$$N_3 \leq \sum_\alpha \int_{[0,1]^d} \frac{u^h u^k}{2\beta n k_1} e^{-\beta n u^{2k} + f(u)} (h_1 + k_1 + 1 + u_1 \partial_1 f(u)) du.$$

Therefore, we obtain

$$\begin{aligned} \frac{N_3}{N_0} &\leq \frac{1}{2\beta k_1 n} \left\{ \frac{N_3}{N_0} \frac{n\beta(k_1 + 1)}{2} + \frac{N_1}{N_0} \right. \\ &\quad \left. \times (k_1 + h_1 + 1 + \frac{\beta k_1}{2} \|\xi\|^2 + \frac{\beta}{2} \|\partial_1 \xi\|^2 + k_1 \sigma \|a\| + \sigma \|\partial_1 a\|) \right\}. \end{aligned}$$

Therefore

$$z_1 \frac{N_3}{N_0} \leq \frac{1}{2\beta k_1 n} \frac{N_1}{N_0} (k_1 + h_1 + 1 + \frac{\beta k_1}{2} \|\xi\|^2 + \frac{\beta}{2} \|\partial_1 \xi\|^2 + k_1 \sigma \|a\| + \sigma \|\partial_1 a\|).$$

By using Cauchy-Schwarz inequality, that is to say,  $N_1/N_0 \leq (N_2/N_0)^{1/2}$ , and and by applying the result of the first half and Hölder's inequality,

$$\begin{aligned} \frac{N_3}{N_0} &\leq \frac{1}{2\beta k_1 n} \left( \frac{N_2}{N_0} \right)^{1/2} \left\{ k_1 + h_1 + 1 + \frac{\beta k_1}{2} \|\xi\|^2 + \frac{\beta}{2} \|\partial_1 \xi\|^2 + k_1 \sigma \|a\| + \sigma \|\partial_1 a\| \right\} \\ &\leq \frac{C}{n^{3/2}} \{1 + \|\xi\|^2 + \|\partial_1 \xi\|^2 + \sigma \|a\| + \sigma \|\partial_1 a\|\}^{3/2}, \end{aligned}$$

where  $C > 0$  is a constant which is determined by  $k_1, h_1$ , and  $\beta$ . In general,

$$\left(\frac{1}{5} \sum_{k=1}^5 |a_k|^2\right)^{3/2} \leq \frac{1}{5} \sum_{k=1}^5 |a_k|^3,$$

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

## 7.7 Proof of Lemma 7

For given functions  $\xi(u)$  and  $g(u)$ , we define

$$\begin{aligned} A^p(\xi, g) &\equiv \sum_{\alpha} \int_{[0,1]^r} dx \int_{[0,1]^{r'}} dy (x^{2k} y^{2k'})^p x^h y^{h'} g(x, y) \\ &\times e^{-n\beta x^{2k} y^{2k'} + \sqrt{n}\beta x^k y^{k'} \xi(x, y)}. \end{aligned} \quad (52)$$

Then

$$E_u^0[u^{2pk} f(u) | \xi] = \frac{A^p(\xi, f\psi)}{A^0(\xi, \psi)}.$$

It is rewritten as

$$A^p(\xi, g) = \sum_{\alpha} \int_0^{\infty} dt \int dx dy \delta(t - nx^{2k} y^{2k'}) x^h y^{h'} g(x, y) \frac{t^p}{n^p} e^{-\beta t - \beta \sqrt{t} \xi(x, y)}.$$

To analyze  $\delta(\cdot)$  function, we need the fact that, for  $\text{Re}(z) > 0$ ,

$$\begin{aligned} \int_{[0,1]^r} (a x^{2k})^z x^h dx &= a^z \prod_{j=1}^r \int_0^1 x_j^{2k_j z + h_j} dx_j \\ &= \frac{a^z}{2^r k_1 \cdots k_r (z + \lambda_{\alpha})^r}. \end{aligned}$$

By applying the inverse Mellin transform to this equation, we have

$$\int_{[0,1]^r} \delta(t - ax^{2k}) x^h dx = \begin{cases} c_0 \frac{t^{\lambda_{\alpha}-1}}{a^{\lambda_{\alpha}}} (\log \frac{a}{t})^{r-1} & (0 < t < a) \\ 0 & (\text{otherwise}) \end{cases}$$

where  $c_0 = 1/(2^r (r-1)! k_1 \cdots k_r)$ . If  $g_0(y) = g(0, y)$  and  $\xi_0(y) = \xi(0, y)$  then

$$A^p(\xi_0, g_0) = \sum_{\alpha} \int_0^{\infty} dt \int_{t < ny^{2k'} < n} dy c_0 \frac{y^{\mu} t^{p+\lambda_{\alpha}-1}}{n^{p+\lambda_{\alpha}}} e^{-\beta t - \beta \sqrt{t} \xi_0(y)} \left(\log \frac{ny^{2k'}}{t}\right)^{r-1} g_0(y). \quad (53)$$

where the region ' $t < ny^{2k'} < n$ ' denotes the set  $\{y \in [0, 1]^s; t < ny^{2k'} < n\}$ . Then by using eq.(53),

$$\begin{aligned} |A^p(\xi, g)| &\leq c_0 \|g\| e^{-\beta \|\xi\|^2/2} \int_0^{\infty} dt \int_{[0,1]^r} dy \frac{y^{\mu} t^{p+\lambda_{\alpha}-1}}{n^{p+\lambda_{\alpha}}} \left|\log \frac{ny^{2k'}}{t}\right|^{r-1} e^{-\beta t/2} \\ &\leq c_1 \|g\| e^{-\beta \|\xi\|^2/2} \frac{(\log n)^{r-1}}{n^{p+\lambda_{\alpha}}}, \end{aligned} \quad (54)$$

where  $c_1 > 0$  is a constant. In the same way,

$$|A^p(\xi, g)| \geq c'_1 \min |g| e^{-3\beta \|\xi\|^2/2} \frac{(\log n)^{r-1}}{n^{p+\lambda_{\alpha}}}. \quad (55)$$

Let  $\lambda$  be the smallest value in  $\{\lambda_\alpha; \alpha\}$ . Then  $(-\lambda)$  is equal to the largest pole of  $\zeta(z)$ . The coordinate  $U_\alpha$  whose  $\lambda_\alpha$  is equal to the smallest one  $\lambda_\alpha = \lambda$  and whose  $r$  is equal to the largest one  $r = m$  is denoted by  $U_{\alpha^*}$ . The sum  $\sum_{\alpha^*}$  denotes the sum restricted to such coordinates. Let  $A_*^p(\xi, g)$  be the sum of  $A^p(\xi, g)$  restricted in this way, in other words,  $\sum_\alpha$  is replaced by  $\sum_{\alpha^*}$  in eq.(52). Also we define  $C_*^p(\xi, g) = A_*^p(\xi, g) - A_*^p(\xi_0, g_0)$ . There exists  $x^* \in [0, 1]^r$  such that

$$e^{-\beta\sqrt{t}\xi(x,y)}g(x,y) - e^{-\beta\sqrt{t}\xi(0,y)}g(0,y) = \sum_{j=1}^r x_j \{ \partial_j g(x^*, y) - \beta\sqrt{t}g\partial_j \xi(x^*, y) \} e^{-\beta\sqrt{t}\xi(x^*, y)}$$

Hence

$$|C_*^p(\xi, g)| \leq c_2(\|\nabla g\| + \beta\|g\|\|\nabla \xi\|) e^{-\beta\|\xi\|^2/2} \frac{(\log n)^{m-2}}{n^{p+\lambda}}. \quad (56)$$

By expanding eq.(53), we have

$$\begin{aligned} A_*^p(\xi_0, g_0) &= \sum_{k=1}^m A_*^{pk}(\xi_0, g_0) \\ A_*^{pk}(\xi_0, g_0) &= \sum_{\alpha^*} \frac{(\log n)^{k-1}}{n^{p+\lambda}} \binom{m-1}{k-1} \int_0^\infty dt \int_{t < ny^{2k'} < n} dy \\ &\quad \times c_0 y^\mu (\log \frac{y^{2k'}}{t})^{m-k} g_0(y) t^{p+\lambda-1} e^{-t+\sqrt{t}\xi_0(y)}. \end{aligned}$$

The largest order term among them is  $A_*^{pm}(\xi_0, g_0)$ . We define  $B_*^{pm}(\xi_0, g_0)$  from  $A_*^{pm}$  by replacing the integral region of  $y$ ,

$$\begin{aligned} B_*^{pm}(\xi_0, g_0) &= \sum_{\alpha^*} \frac{(\log n)^{m-1}}{n^{p+\lambda}} \int_0^\infty dt \int_{[0,1]^r} dy \\ &\quad \times c_0 y^\mu g_0(y) t^{p+\lambda-1} e^{-\beta t + \beta\sqrt{t}\xi_0(y)}. \end{aligned}$$

The difference between  $A_*^{pm}(\xi_0, g_0)$  and  $B_*^{pm}(\xi_0, g_0)$  is smaller than  $\|g\|e^{-\|\xi\|^2/2}/n^{p+\lambda}$ , and

$$|A_*^{pk}(\xi_0, g_0)| \leq c_3 \|g\| e^{-\beta\|\xi\|^2/2} \frac{(\log n)^{k-1}}{n^{p+\lambda}} \quad (1 \leq k \leq m), \quad (57)$$

$$|B_*^{pm}(\xi_0, g_0)| \geq c_3' \|g\| e^{-3\beta\|\xi\|^2/2} \frac{(\log n)^{r-1}}{n^{p+\lambda}}. \quad (58)$$

By the definition,

$$D \equiv E_u^0[u^{2pk} f(u)|\xi] - E_{y,t}[t^p f(0, y)|\xi] = \frac{A^p(\xi, f\psi)}{A^0(\xi, \psi)} - \frac{B_*^{pm}(\xi_0, f_0\psi_0)}{B_*^{0m}(\xi_0, \psi_0)}.$$

Then using eqs.(54)-(58),

$$\begin{aligned} R^p(\xi, g) &\equiv A^p(\xi, g) - B_*^{pm}(\xi_0, g_0) \\ &= A_o^p(\xi, g) + C_*^p(\xi, g) + \sum_{k=1}^m A_*^{pk}(\xi_0, g_0) - B_*^{pm}(\xi_0, g_0), \end{aligned}$$

where  $A_o^p(\xi, g) = A^p(\xi, g) - A_*^p(\xi, g)$  is the sum over  $\alpha$  that are not  $\alpha^*$ . Therefore

$$|R^p(\xi, g)| \leq \frac{c_4}{n^{p+\lambda}} e^{-\beta\|\xi\|^2} (\|g\| + \beta\|g\|\|\nabla \xi\| + \|\nabla g\|)$$

Thus

$$\begin{aligned}
|D| &\leq \frac{|R^p(\xi, f\psi)|}{A^0(\xi, \psi)} + \frac{|R^0(\xi, \psi)| |B_*^{pm}(\xi_0, f_0\psi_0)|}{A^0(\xi, \psi) B_*^{0m}(\xi_0, \psi_0)} \\
&\leq \|\psi\| \left\| \frac{1}{\psi} \right\| \frac{c_5}{n^p \log n} e^{4\beta\|\xi\|^2} (\|g\| + \beta\|g\| \|\nabla\xi\| + \|\nabla g\|)
\end{aligned}$$

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

## 7.8 Proof of Lemma 8

By using partial integration, for an arbitrary  $a \in R$ ,

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

Hence

$$\int_0^\infty dt (2t - \sqrt{t}a - \frac{2\lambda}{\beta}) t^{\lambda-1} e^{-\beta t + \beta \sqrt{t}a} = 0. \tag{59}$$

which shows Lemma 8. (Q.E.D.)

## 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 Berkeley 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] K.Nagata, S.Watanabe, "The Exchange Monte Carlo Method for Bayesian Learning in Singular Learning Machines," *Proc. of WCCI2006*,(Canada, Cancouver), 2006.
- [12] 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.
- [13] 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.
- [14] S. Watanabe, gAlgebraic information geometry for learning machines with singularities,h *Advances in Neural Information Processing Systems*, (Denver, USA), pp.329-336. 2001.
- [15] S.Watanabe, "Algebraic analysis for nonidentifiable learning machines," *Neural Computation*, Vol.13, No.4, pp.899-933, 2001.
- [16] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," *Neural Networks*, Vol.14, No.8,pp.1049-1060, 2001.
- [17] S. Watanabe, "Learning efficiency of redundant neural networks in Bayesian estimation," *IEEE Transactions on Neural Networks*, Vol.12, No.6, 1475-1486, 2001.
- [18] 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.
- [19] S.Watanabe,"Algebraic geometry of singular learning machines and symmetry of generalization and training errors," *Neurocomputing*, Vol.67,pp.198-213,2005.
- [20] S. Watanabe,"Algebraic geometry and learning theory,h Morikita publishing, 2006.

- [21] S.Watanabe, "Generalization and training errors in Bayes and Gibbs estimations in singular learning machines," IEICE technical report in neuro computing, IEICE-NC, December, 2007.
- [22] S. Watanabe, "On a relation between a limit theorem in learning theory and singular fluctuation," *IEICE Technical Report*, No.NC2008-111, pp.45-50, 2009.
- [23] A. W. van der Vaart, Jon A. Wellner, "Weak Convergence and Empirical Processes," Springer,1996.
- [24] 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.
- [25] 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.
- [26] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models", *Neurocomputing*, Vol.69, pp.62-84, 2005.