Analytic Equivalence of Bayes a Posteriori Distributions

Takeshi Matsuda and Sumio Watanabe

Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology
matsuken@cs.pi.titech.ac.jp

Abstract. A lot of learning machines which have hidden variables or hierarchical structures are singular statistical models. They have singular Fisher information matrices and different learning performance from regular statistical models. In this paper, we prove mathematically that the learning coefficient is determined by the analytic equivalence class of Kullback information, and show experimentally that the stochastic complexity by the MCMC method is also given by the equivalence class.

1 Introduction

Learning machines such as layered neural networks, normal mixtures, hidden Markov models, Boltzmann machines, Bayes networks and stochastic context-free grammars are not regular statistical models, because their Fisher information matrices are not positive definite. These learning machines are called *singular statistical models* because they are not subject to the conventional statistical theory of regular statistical models. In fact, neither the distribution of the maximum likelihood estimator nor the Bayes a posteriori distribution converges to the normal distribution, even when the number of training samples goes to infinity.

Recently, it was proved that the generalization performance of a singular learning machine in Bayes estimation is determined by the algebraic geometrical structure of the learning machine [5]. The generalization error G, which is defined as the expectation value of the Kullback information from the true distribution to the Bayes predictive distribution, is equal to

$$G = \frac{\lambda}{n} + o(\frac{1}{n}),$$

where n is the number of training samples and λ is the learning coefficient. The constant $(-\lambda)$ is equal to the largest pole of the zeta function of a learning machine,

$$\zeta(z) = \int H(w)^z \varphi(w) dw \quad (z \in \mathbf{C}),$$

where $H(w)$ is the Kullback information from the true distribution to the learning machine with the parameter w and $\varphi(w)$ is the Bayes a priori distribution.

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 113–121, 2006.

The learning coefficients of some learning machines, for example, a three-layer perceptron and a reduced rank regression, have been obtained by using resolution of singularities [2,3], which clarified that the generalization errors of singular learning machines are smaller than those of regular statistical models, if Bayes estimation is employed in learning.

In this paper, we introduce the concept of analytic equivalence between the Kullback informations, and show the following three facts.

(1) We prove that, if two learning machines are analytically equivalent, then they have the same learning coefficient.
(2) For the case when the Kullback information is defined on the two-dimensional Euclidean space, we derive the concrete learning coefficient of a given equivalence class.
(3) We show experimentally that the the stochastic complexity obtained by the Markov chain Monte Carlo method is also determined by the analytic equivalence class.

In regular statistical models, the asymptotic behavior of a learning machine is completely determined by the Fisher information matrix, whereas in singular learning machines, it is determined by the analytic equivalence class of the Kullback information.

2 Statistical Framework of Machine Learning

In this section, we summarize the well known statistical framework of Bayes estimation.

2.1 Bayes Learning

Let $q(x)$ be a probability density function called as the true distribution which is defined on the N-dimensional Euclidean space, $\mathbf{R}^N$. A set of random variables

$$X^n = (X_1, X_2, ..., X_n)$$

consists of training samples which are independently taken from the probability distribution $q(x)dx$. The integer n is referred to as the number of training samples. A learning machine is represented by a conditional probability density function $p(x|w)$ where w is a d-dimensional parameter. When an a priori probability density function $\varphi(w)$ is given on $\mathbf{R}^d$, the Bayes a posteriori distribution is defined by

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

where and $Z(X^n)$ is the normalizing constant. The Bayes predictive distribution is also defined by

$$p(x|X^n) = \int p(x|w)\ p(w|X^n)\ dw,$$

which is the estimated probability density function on $\mathbf{R}^N$ by Bayes learning. The Generalization error $G(n)$ is measured by the average Kullback information from the true distribution $q(x)dx$ to the predictive distribution $p(x|X^n)dx$,

$$G(n) = E\Big[\int q(x) \log \frac{q(x)}{p(x|X^n)} dx\Big],$$

where $E[\cdot]$ denotes the expectation value overall sets of X^n. Also we define the stochastic complexity by

$$F(n) = E\Big[-\log Z(X^n)\Big] + n \int q(x) \log q(x) dx.$$

It is easy to show that

$$G(n) = F(n+1) - F(n)$$

holds for an arbitrary natural number n. The stochastic complexity indicates how appropriate the set $p(x|w)$ and $\varphi(w)$ is for a given training sample set X^n.

2.2 Asymptotic Theory

In learning theory, it is important to clarify the asymptotic behaviors of $G(n)$ and $F(n)$. The relation between algebraic geometry of the Kullback information and singular learning machines was clarified, and the following theorem was proved.

Theorem 1. *When n tends to infinity, the generalization error and the stochastic complexity are respectively given by*

$$G(n) = \frac{\lambda}{n} + o(\frac{1}{n}),$$
$$F(n) = \lambda \log n - (m-1) \log \log n + O(1),$$

where $(-\lambda)$ and m are respectively equal to the largest pole and its order of the zeta function,

$$\zeta(z) = \int H(w)^z \, \varphi(w) \, dw.$$

Here $H(w)$ is the Kullback information

$$H(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

Proof. The proof is given in [5].

This theorem shows that the learning coefficient is determined by the Kullback information $H(w)$ and the a priori distribution $\varphi(w)$.

3 Analytic Equivalence of Kullback Information

In this section we introduce the concept of analytic equivalence and prove that, if the Kullback informations are analytically equivalent, then they have the same learning coefficient.

Definition 1. *Let U and V be open sets in $\mathbf{R}^d$ whose closures are compact. Two real analytic functions $K(w)$ on U and $H(w)$ on V are said to be analytically equivalent if there exists a bijective analytic map $g : V \to U$*

$$H(w) = K(g(w)) \quad (w \in V)$$

and the Jacobian $|g'(w)|$ satisfies the condition that $\epsilon < |g'(w)| < C$ in V for some $\epsilon, C > 0$.

Then by the definition of the analytic equivalence, the following theorem holds.

Theorem 2. *Assume that two real analytic functions $K(w)$ on U and $H(w)$ on V are analytically equivalent. Then two zeta functions*

$$\zeta_1(z) = \int_U K(w)^z dw$$
$$\zeta_2(z) = \int_V H(w)^z dw$$

have the same largest pole.

Proof. It is well known that $\zeta_1(z)$ and $\zeta_2(z)$ are meromorphic functions and all poles of them are negative and real numbers [4]. From the definition, it follows that

$$\begin{aligned}\zeta_2(z) &= \int_U H(g(w))^z |g'(w)| dw \\ &= \int_U K(w)^z |g'(w)| dw\end{aligned}$$

Let $(-\lambda)$ be the largest pole of $\zeta_1(z)$. When z is real and $z > -\lambda$

$$\epsilon |\zeta_1(z)| < |\zeta_2(z)| < C|\zeta_1(z)|$$

This inequality shows that the largest poles should coincide.

Note that two zeta functions do not have the same second largest pole in general.

Definition 2. *Let $v = (v_1, \ldots, v_n)$ be a set of nonnegative integers. For a monomial $x^u = x_1^{u_1} \cdots x_n^{u_n}$, we define the weighted degree $ord_w(x^u)$ with the weight v by*

$$ord_w(x^u) =< v, u >= v_1 u_1 + \ldots + v_n u_n.$$

A polynomial is said to be guasi-homogeneous if it is a linear combination of the monomials which have the same weighted degree with some weight.

Definition 3. *An analytic function f is said to have an algebraic isolated singularity at O, if the dimension of a real vector space*

$$M(f) = R[[x_1, \cdots, x_n]]/ < \frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n} >$$

is finite.

The following theorem shows a sufficient condition of the analytic equivalence.

Theorem 3. *Let f be an analytic function*

$$f = f_d + f_{d+1} + f_{d+2} + \cdots, f_d \neq 0$$

where f_d is a quasi-homegeneous polynomial of degree q with weight $\mathtt{v}$. If the weighted degree of $\mathtt{x}^{u_i}$ exceeds d, $c_1, \ldots, c_s$ are constants, then f and $f_d + c_1\mathtt{x}^{u_1} + \cdots + c_s\mathtt{x}^{u_s}$ are analytically equivalent.

Proof. For the proof of this theorem, see [1].

4 Two Dimensional Parameter Space

In the previous section, we have shown that the learning coefficient is determined by the analytic equivalence class. In this section we give the concrete learning coefficients for a given analytic function on two-dimensional space.

Theorem 4. *Let f be an analytic function given by*

$$f(x, y) = \Sigma_{k+l=4} a_{kl} x^k y^l + \Sigma_{2 \leq i+j \leq 3} b_{ij} x^i y^j,$$

where a_{kl} and b_{ij} are the real number coefficients of $x^k y^l$, $x^i y^j$, respectively. We consider the zeta function such that

$$\zeta(z) = \int f^{sz} dxdy.$$

Then largest pole of the zeta function is as follows.

$$\lambda = \begin{cases} \frac{2}{as} & (k' > i', l' = j') or (\Sigma b_{ij} x^i y^j \text{is a symmetric expression.}) \\ \frac{2n+1}{(4n+l')s} & (4 - a < j', k' > i', l' < j') \\ \frac{2}{(4-a)s} & (4 - a \geq j', k' > i', l' > j') \end{cases}$$

where k' is the value of the minimum k which satisfies $a_{kl} \neq 0$, l' is the value of the minimum l which satisfies $a_{kl} \neq 0$, i' is the value of the minimum i which satisfies $b_{ij} \neq 0$, j' is the value of the minimum j which satisfies $b_{ij} \neq 0$.

Proof. Let X be the curve $f(x, y) = 0$. We put $x = x$, $y = xy$ on X_{11}. Then we have $f_{11}^{sz} =$

$$x^{asz+1} y^{sl'z} (\Sigma_{k+l=4} a_{kl} x^{4-a} y^{l-l'} + \Sigma_{2 \leq i+j \leq 3} b_{ij} x^{i+j-a} y^{j-l'})^{sz}.$$

Similarly, we put $x = xy$, $y = y$ on X_{12}. Then we obtain $f_{12}^{sz} =$

$$x^{ai'z}y^{asz+1}(\Sigma_{k+l=4}a_{kl}x^{k-i'}y^{k+l-a} + 1 + \Sigma_{2\leq i+j\leq 3, i\neq i'}b_{ij}x^{i-i'}y^{i+j-a})^{sz}.$$

where $k' > i', l' < j'$.
Hence,

$$\zeta(z) = \int_{X_{11}} f_{11}^{sz}dxdy + \int_{X_{12}} f_{12}^{sz}dxdy.$$

Here, $\Sigma_{k+l=4}a_{kl}x^{4-a}y^{l-l'} + \Sigma_{2\leq i+j\leq 3}b_{ij}x^{i+j-a}y^{j-l'}$ and $x^{4-a} + y^{j'-l'}$ are analytic equvalence. Therefore, we have to consider only about

$$\int_{X_{11}} x^{asz+1}y^{sl'z}(x^{4-a} + y^{j'-l'})^{sz}dxdy.$$

When continuing resolution of singularity, the zeta function $\zeta(z)$ is as follows.

$$\zeta(z) = \int_{X_{n+1,1}} x^{n(4sz+2)+sl'z}y^{4sz+2+sl'z}(1+\cdots)^{sz}dxdy + \int_{X_{n+1,2}} x^{asz+1}y^{n(4sz+2)+sl'z}(x^{k+l-a} + y^{j'-l'-n(k+l-a)})^{sz}dxdy,$$

where $(j' - l') = n(k + l - a)$. Hence, we obtain

$$\lambda = \frac{2n+1}{(4n+l')s}.$$

In the other case, too, it is possible to prove in the same way. Also, if replacing x and y, we can get the value of λ in all cases.

5 Stochastic Complexity by MCMC Method

In the previous section, we have shown that the learning coefficients are determined by the analytic equivalence classes. In this section, by comparing the theoretical results with the numerical results by the Markov chain Monte Carlo method, we show that the stochastic complexities in real applications are also determined by the equivalence classes. Let us study the function,

$$F(n) = -\log \int \exp(-nf(x,y))\varphi(x,y)dxdy.$$

From the theoretical point of view, it has the asymptotic expansion,

$$F(n) = \lambda \log n - (m-1)\log\log n.$$

In the real applications of Bayes estimation, $F(n)$ is numerically calculated by

$$F(n) = \int_0^1 E_t[nf(x,y))]dt$$

where $E_t[\cdot]$ shows the expectation value over the probability distribution,

$$E_t[nf(x,y)] = \int nf(x,y)p_t(x,y)dxdy,$$

where

$$p_t(x,y) \propto \exp(-ntf(x,y))\varphi(x,y).$$

The random samples subject to $p_t(x,y)$ can be generated by the MCMC method. The following tables show the experimental results of $F(n)$ for $n = 10000$.

analytic function	$F(n)$
$y^2 + x^3$	1.900087
$y^2 + x^3 + y^3$	1.902980
$y^2 + x^3 + xy^3$	1.965838
$y^2 + x^3 + x^4$	2.012928
$y^2 + x^3 + y^3 + xy^3 + x^4$	1.862806
$y^2 + x^3 + y^5$	1.930222
$y^2 + x^3 + x^{10}$	1.910901
$y^2 + x^3 + y^5x^{10}$	1.909953
$y^2 + x^3 + x^{10} + y^5$	1.914395
$y^2 + x^3 + x^{10} + y^5 + y^5x^{10}$	1.909564
$y^2 + x^3 + x^{100}y^{100}$	1.890016
$y^2 + x^3 + x^{100}y^{100} + x^{100} + y^{100}$	1.895358

analytic function	$F(n)$
$y^3 + x^5$	1.115474
$y^3 + x^5 + x^2y^2$	1.162772
$y^3 + x^5 + x^{10}y^{10}$	1.115357
$y^3 + x^5 + x^{15}y^{10}$	1.115979
$y^3 + x^5 + x^{10}y^{15}$	1.114232
$y^3 + x^5 + x^{20}y^{20}$	1.115570
$y^3 + x^5 + x^{100}y^{100}$	1.103289
$y^3 + x^5 + x^{100}y^{100} + x^{100} + y^{100}$	1.101518

analytic function	$F(n)$
$y^5 + x^7$	0.563414
$y^5 + x^7 + x^{10}y^2$	0.555947
$y^5 + x^7 + x^2y^{10}$	0.561365
$y^5 + x^7 + x^{10}y^{10}$	0.562976
$y^5 + x^7 + x^{10}y^{15}$	0.559782
$y^5 + x^7 + x^{15}y^{10}$	0.552454
$y^5 + x^7 + x^{100}y^{100}$	0.566214
$y^5 + x^7 + x^{100}y^{100} + x^{100} + y^{100}$	0.564594

These results show that the numerically calculated stochastic complexities are determined by the analytic equivalence classes.

6 Discussion

In this paper, we have studied the relation between the learning coefficients and analytic equivalence classes. Let us discuss the results from three viewpoints.

Firstly, from the mathematical point of view, the result of this paper is devoted to the case of isolated singularities. In almost all learning machines, their singularities are not isolated, however, there is no simple criterion that can judge the analytic equivalence for non-isolated singularities. To construct the mathematical criterion of the analytic equivalence class in singular learning machines is the problem for the future study.

Secondly, from the statistical point of view, our result is a generalization of Fisher's asymptotic statistics. If two analytic functions have nondenerate Hesse matrices, then they are analytically equivalent. This is the reason why the learning coefficients of regular statistical models are determined by the dimensions of the parameter spaces. In singular learning machines, even if the learning machines have the same-dimensional parameter spaces, they have different learning coefficients in general. Consequently, the concept of the analytic equivalence class is a generalization of the Fisher information matrix.

And lastly, from the learning theoretical point of view, our result shows how the stochastic complexities are determined in the real world applications. The stochastic complexity is important in Bayes learning, which is applied to model selection and hyperpramater optimization. However, it is well known that it requires huge computational costs to calculate the stochastic compelxity. We expect that a new efficient algorithm based on the concept of analytic equivalence class.

7 Conclusion

We proved mathematically that the learning coefficients are determined by the analytic equivalence class of the Kullback information, and showed experimentally that the practical stochastic complexities are also determined by the analytic equivalence class. To construct the mathematical method which enables us to calculate the learning coefficients for the higher dimensional Kullback information is the problem for the future study.

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

References

1. V.I.Arnol'd, "Normal forms of functions in neighbourhoods of degenerate critical points," Russian Mathematical Surveys, Vol.29,No.2,pp.10-50,1974
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," Communications of Pure and Applied Mathematics, 13, 145-150, 1970.
5. S.Watanabe,"Algebraic Analysis for Non-identifiable Learning Machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001