Singularities in Learning Theory

Sumio Watanabe
P&I Lab., Tokyo Institute of Technology,
Mailbox R2-5, 4259 Nagatsuda, Midori-ku,
Yokohama, 226-8503 Japan

May 10, 2006

Abstract

This paper shows the problems of singularities in learning theory. We introduce two important observables in learning theory, stochastic complexity and generalization error. A lot of learning machines used in information science have singularities in their parameter space, resulting that their mathematical properties have been left unknown. In this paper we show that the asymptotic behaviors of stochastic complexity and generalization error can be identified by the largest pole and its order of the zeta function of the learning machine.

1 Introduction

Let $\mathbf{R}^N$ and $\mathbf{R}^d$ be respectively N and d dimensional Euclidean spaces. Also let $(\Omega, \mathcal{B}, P)$ be a probability space, and $X_1, X_2, ..., X_n$ be random variables defined as measurable functions from Ω to $\mathbf{R}^N$. Assume that random variables $X_1, X_2, ..., X_n$ are independently subject to the probability distribution $q(x)dx$ where $q(x) > 0$ is a positive and measurable function on $x \in \mathbf{R}^N$ and dx is Lebesgue measure on $\mathbf{R}^N$. Let D_n denote the set of random variables,

$$D_n = \{X_1, X_2, ..., X_n\}.$$

The set D_n is referred to as the set of random samples and the probability distribution $q(x)dx$ is called as the true distribution.

A learning machine is defined by the conditional probability distribution $p(x|w)dx$, where $p(x|w)$ is a positive and measurable function of $x \in \mathbf{R}^N$ for a given $w \in \mathbf{R}^d$. Here w is called a parameter of the learning machine. Let $\varphi(w)dw$ be a probability distribution on $\mathbf{R}^d$. In learning theory we study two probability distributions. The a posteriori distribution of w is defined by

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

where $Z(D_n)$ be the normalized function defined by

$$Z(D_n) = \int \varphi(w) \prod_{i=1}^{n} p(X_i|w)\, dw.$$

The predictive distribution is defined by

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

which is the estimated probability distribution from the random samples. The main problem of learning theory can be expressed as the following.

Main Problem in Learning Theory. Establish mathematical foundation that is necessary to clarify the difference between two probability distributions $q(x)dx$ and $p(x|D_n)dx$.

Remarks. In information science, only the random variables D_n are given, whereas the true probability distribution $q(x)dx$ is unknown. Information scientist employs a learning machine $p(x|w)dx$ and estimates "the true distribution is approximately equal to the predictive distribution $p(x|D_n)dx$. Learning theory is needed to answer how precise the estimation is.

2 Stochastic Complexity and Generalization Error

The relative entropy from a probability distribution $p_1(x)dx$ to $p_2(x)dx$ is defined by

$$\int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx,$$

which is nonnegative and is equal to zero if and only if $p_1(x)dx = p_2(x)dx$ $(\forall x)$. The *generalization error* is defined as the relative entropy from the true distribution $q(x)dx$ to the predictive distribution $p(x|D_n)$,

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

The generalization error is a measurable function of the random variables D_n, hence it is also a random variable. It is expected that $G(D_n) \to 0$ when n tends to infinity. The goal of learning theory is to clarify the asymptotic behavior of the random variable $G(D_n)$.

We define the log loss function by

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

Then the *stochastic complexity* is given by

$$F(D_n) = -\log \int \exp(-\sum_{i=1}^{n} f(X_i, w)) \; \varphi(w) \; dw,$$

which is also a random variable. It is quite easy to show that

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

where $E[\cdot]$ shows the expectation value. This relation shows that there is a mathematical relation between the generalization error and the stochastic complexity.

Remark. The stochastic complexity is also called the random free energy. It is quite important in mathematical physics to clarify the asymptotic behavior of the stochastic complexity.

3 Mathematical Problem

The set of true parameters is defined by

$$W_0 = \{w \in \mathbf{R}^d; q(x) = p(x|w) \quad (\forall x)\}.$$

If the set of true parameters consists of one point w_0 and the Fisher information matrix

$$I(w_0) = \int \frac{\partial f(x,w_0)}{\partial w_i} \frac{\partial f(x,w_0)}{\partial w_j} p(x|w_0) dx$$

is positive definite, then it is easy to show the asymptotic behaviors of $G(D_n)$ and $F(D_n)$. It is well known in information theory, theoretical physics, and statistics that

$$\begin{aligned} n\, E[G(D_n)] &\rightarrow \frac{d}{2}, \\ E[F(D_n)] - \frac{d \log n}{2} &\rightarrow const., \end{aligned}$$

where this fact was proved between 1920-1940.

However, almost all learning machines used in modern information science do not have positive definite Fisher information matrices. Moreover, W_0 is not one point but an algebraic set or an analytic set in general.

Mathematical Problem in Learning Theory. Establish mathematical foundation on which we can construct learning theory when W_0 is an algebraic set or an analytic set.

Remark. If a learning machine estimates the structure of the true distribution or if it estimates the hidden variables of the true distribution, then it has singularities (degenerate Fisher information matrices) in its parameter space. Almost all learning machines such as neural networks, Bayesian networks, gaussian mixtures, mixtures of binomial distributions, spin systems, hidden Markov models, probabilistic contex-free grammars, and Boltzmann machines have singularities.

4 Results

The main results are described in this section.

Condition (A). The a priori distribution $\varphi(w)$ is given by the following form,

$$\varphi(w) = \begin{cases} \varphi_0(w) & (\varphi_1(w) > 0, ..., \varphi_k(w) > 0) \\ 0 & \text{(otherwise)} \end{cases}$$

where $\varphi_0(w), .., \varphi_k(w)$ are real analytic functions. The support of $\varphi(w)$ is compact.

Condition (B). The function

$$f(x, w) = \log(q(x)/p(x|w))$$

is an $L^s(U)$ valued analytic function of $w \in \mathbf{R}^d$ $(s \geq 2)$ where $W \subset \mathbf{R}^d$ is an open set which includes the support of $\varphi(w)$. Here $L^s(U)$ is the Banach space defined by

$$L^s(U) = \{g; \int_U |g(x)|^s q(x) dx < \infty\}.$$

These conditions are assumed in the following theorems. Under the condition (B), the function

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

is an analytic function of w.

The first theorem claims that the stochastic complexity has asymptotic expansion.

Theorem 1 *Assume two conditions (A) and (B) for $s \geq 4$. Then the stochastic complexity has the following asymptotic expansion,*

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

Here $(-\lambda_1)$ and m_1 are respectively the largest pole and its order of the meromorphic function $\zeta(z)$ $(z \in \mathbf{C})$ on the entire complex plane that is the analytic continuation of

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

The random variables $R(D_n), R$ satisfy

$$R(D_n) \to R \quad (\textit{convergence in law}),$$

and

$$\lim_{n\to\infty} E[R(D_n)] = E[R].$$

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

Condition (C). We define a function $M(x)$ by

$$M(x) = \sup_{w \in K*} |f(x,w)|.$$

where K^* is an open set of C^d which contains the support of φ. There exists a constant $\gamma > 0$ such that

$$E[M(X)^2 \exp(\gamma M(X)^2)] < \infty.$$

This condition is needed in the second theorem.

Theorem 2 *Assume that condition (A),(B) and (C) for $s \geq 4$. Then there exists a random variable G such that the convergence in law*

$$nG(D_n) \to G$$

holds and

$$nE[G(D_n)] \to E[G] = \lambda_1,$$

where λ_1 is equal to that in Theorem 1.

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

5 Algorithms

In information science, we need an algorithm to obtain the largest pole and its order of the zeta function. It is well known that, if we find the resolution of singularities of $H(w) = 0$, then λ_1 and m_1 can be immediately calculated. In fact, these coefficients of a lot of learning machines have been found by resolution of singularities.

6 Conclusion

This paper shortly introduced the problem of singularities in learning theory.

References

[1] Sumio Watanabe, "Daisu Kika to Gakushu Riron," Morikita Shuppan, in Japanese (2006).

[2] Sumio Watanabe, "Algebraic Geometry and Statistical Learning Theory," to be published in Cambridge University Press (In English).

[3] M.Aoyagi, S.Watanabe, "Stochastic complexities of reduced rank regression in Bayesian estimation," International Journal of Neural Networks, 18(7),pp.924-933,2005.

[4] S.Watanabe,"Algebraic analysis for nonidentifiable learning machines", Neural Computation, 13(4), pp.899-933, 2001.

[5] S.Watanabe, K.Fukumizu, K.Hagiwara, S.Amari, "Learning Theory of Singular Statistical Models," Trans. IEICE,J88-D2 (2), pp.159-169, 2005. (Survay Paper).

[6] K. Yamazaki, S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models", Neurocomputing, to appear.