An Introduction to
Algebraic Geometry and Statistical Learning Theory

Sumio Watanabe
Tokyo Institute of Technology

December 17, 2009

Abstract
This article introduces the book, “algebraic geometry and statistical learning theory.” A parametric model in statistics or a learning machine in information science is called singular if it is not identifiable or if its Fisher information matrix is not positive definite. Although a lot of statistical models and learning machines are singular, their statistical properties have been left unknown. In this book, an algebraic geometrical method is established on which we can construct new statistical theory for singular models.

Four main formulas are proved. Firstly, we show that any log likelihood function can be represented by a common standard form, based on resolution of singularities. Secondly, the asymptotic behavior of the Bayes marginal likelihood is derived by the zeta function theory. Thirdly, the asymptotic expansions of Bayes generalization and training errors are proved, which enable us to make a widely applicable information criterion for singular models. And lastly, the symmetry of the generalization and training errors in the maximum a posteriori method is proved.

In this book, algebraic geometry is explained for non-mathematicians, and the concrete, applicable, and nontrivial formulas are introduced. Also it is theoretically shown that, in singular models, Bayes estimation is more appropriate than one point estimation, even asymptotically.

1 Outline of the book
A parametric model in statistics or a learning machine in information science is called singular if the map from the parameter to the probability distribution is not one-to-one, or if its Fisher information matrix is not positive definite. A lot of statistical models are singular, for example, artificial neural networks, reduced rank regressions, normal mixtures, binomial mixtures, hidden Markov models, stochastic context-free grammars, Bayesian networks, and so on. In general, if a statistical model contains hierarchical structure, sub-module, or hidden variables, then it is singular.

If a statistical model is singular, then the log likelihood function can not be approximated by any quadratic form, resulting that the conventional statistical theory of regular statistical models does not hold. In fact, Cramer-Rao inequality has no meaning, asymptotic normality of the maximum likelihood estimator does not hold, and the Bayes a posteriori distribution can not be approximated by any normal distribution. Neither AIC corresponds to the asymptotic average generalization error nor BIC is equal to the asymptotic Bayes marginal likelihood. It has been difficult to study singular models, because there are so many types of singularities in their log likelihood functions.
In this book, we introduce an algebraic geometrical method by which new statistical theory can be constructed. Four main formulas are established.

- The log likelihood function can be given a common standard form using resolution of singularities, even applied to more complex models.
- The asymptotic behavior of the Bayes marginal likelihood is derived based on zeta function theory.
- A new method is derived to estimate the generalization error in Bayes estimation from training error.
- The generalization errors of maximum likelihood and a posteriori methods are clarified by empirical process theory on algebraic varieties.

In this book, algebraic geometry, zeta function theory, and empirical process theory are explained for non-mathematicians, which are useful to study statistical theory of singular statistics.

2 Main Results of the Book

2.1 Introduction

Let $p(x|w)$ be a probability density function of $x \in \mathbb{R}^N$ for a parameter $w \in W \subset \mathbb{R}^d$, which is called a statistical model or a learning machine. A statistical model $p(x|w)$ is called singular if the map $w \mapsto p(\cdot|w)$ is not one-to-one or if its Fisher information matrix contains the zero eigen value.

Let $X_1, X_2, \ldots, X_n$ be random variables which are independently subject to the probability distribution $q(x)dx$. The set

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

and $q(x)dx$ are called a set of random samples and the true distribution, respectively. In this book, we study the case when the set of true parameters

$$W_0 = \{w \in W; q(x) = p(x|w)\}$$

is not a single point but an analytic set with singularities. From the statistical point of view, this case is important in model selection, hyperparameter optimization, and hypothesis test for singular statistical models.

Example 1. In a regression model,

$$p(y|x, a, b, c, d) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} (y - a \sin(bx) - c \sin(dx))^2 \right),$$

and $q(y|x) = p(y|x, 0, 0, 0, 0)$, then

$$W_0 = \{ab + cd = 0, ab^3 + cd^3 = 0\}.$$ 

Therefore $W_0$ is an algebraic set with singularities.
2.2 Main Formula I

The Kullback-Leibler divergence $K(w)$ and the empirical one $K_n(w)$ are respectively defined by

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

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

Note that $K_n(w)$ is equal to the minus log likelihood ratio function. In this book we mainly study the case when $K(w)$ is a real analytic function. Then $W_0 = \{w \in W; K(w) = 0\}$ is a real analytic set. It is difficult to study the set $W_0$ because it contains complicated singularities in general. However, the fundamental theorem called resolution of singularities in algebraic geometry ensures the following fact. For an arbitrary $K(w) \geq 0$, there exist a $d$-dimensional manifold $M$ and a real analytic map $g : M \to W$ such that, in each local coordinate of $M$,

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

$$\varphi(g(u))|g'(u)| = u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d},$$

where $|g'(u)|$ is the Jacobian determinant of $w = g(u)$ and $k_1, k_2, \ldots, k_d, h_1, h_2, \ldots, h_d$ are non-negative integers. By using this fact, we can prove the first theorem.

**Theorem 1** For an arbitrary singular statistical model, the empirical Kullback-Leibler divergence is given by the common standard form,

$$K_n(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d} - \frac{1}{\sqrt{n}} \xi_n(u) u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d},$$

where $\xi_n(u)$ is an empirical process on $M$ which converges to a gaussian process $\xi(u)$ in law.

Theorem 1 is important because it enables us to study any singular models by using the common standard form. Note that a regular model is a very special example of singular models, hence Theorem 1 holds even for a regular statistical model.

2.3 Main Formula II

The generalized Bayes a posteriori distribution with $\beta > 0$ is defined by

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

where $\varphi(w)$ is an a priori distribution and $Z_n$ is the Bayes marginal likelihood. The minus log Bayes marginal likelihood is defined by

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

which is called the Bayes free energy or the Bayes description length in information theory. The random variable $F$ plays the central role in model evaluation, model selection, and hyperparameter optimization. Let

$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i)$$

be the empirical entropy. Based on Theorem 1, we can prove the following theorem.
Theorem 2  The asymptotic expansion

\[ F = \beta n S_n + \lambda \log n - (m - 1) \log \log n + O_p(1) \]

holds, where \( \lambda > 0 \) is a rational number and \( m \) is a natural number. Two constants \((-\lambda)\) and \( m \) are respectively equal to the maximum pole and its order of the zeta function that is defined by the analytic continuation of the holomorphic function

\[ \zeta(z) = \int K(w)^z \varphi(w) dw, \quad \text{Re}(z) > 0. \]

The concrete values of \( \lambda \) and \( m \) can be found by a finite algebraic calculation. The constant \( \lambda \) is called a log canonical threshold in algebraic geometry, which is equal to

\[ \lambda = \min \min_j \left( \frac{h_j + 1}{2k_j} \right), \]

where \( \min_{\alpha} \) shows the minimum value over all local coordinates of \( M \). If a statistical model is regular, then \( \lambda = d/2 \) and \( m = 1 \), whereas, if it is singular, \( \lambda \neq d/2 \) and \( m \neq 1 \) in general. For example, in the case eq.(1), \( \lambda = 2/3 \) and \( m = 1 \). Theorem 2 shows that BIC for a regular model should be improved in singular models.

2.4 Main Formula III

Let \( E_w[ \cdot ] \) be the average using the generalized Bayes a posteriori distribution \( p(w|D_n) \). The Bayes predictive distribution is defined by

\[ p^*(x) = E_w[p(x|w)]. \]

The important random variables in Bayes statistics are

\[ G = -\int q(x) \log p^*(x) dx, \]
\[ T = -\frac{1}{n} \sum_{i=1}^{n} \log p^*(X_i), \]

where \( G \) and \( T \) are the average and empirical log losses of Bayes estimation. In information science, \( G \) and \( T \) are called the generalization and training errors respectively. The entropy \( S \) is defined by

\[ S = -\int q(x) \log q(x) dx. \]

We introduce a new concept functional variance,

\[ V = \sum_{i}^{n} \{ E_w[(\log p(X_i|w))^2] - E_w[(\log p(X_i|w))]^2 \}. \]

Theorem 3  There exist two constants \( \lambda \) and \( \nu = \nu(\beta) \) such that the following asymptotic expansions hold,

\[ E[G] = S + \left( \frac{\lambda - \nu}{\beta} + \nu \right) \frac{1}{n} + o(\frac{1}{n}), \quad (2) \]
\[ E[T] = S + \left( \frac{\lambda - \nu}{\beta} - \nu \right) \frac{1}{n} + o(\frac{1}{n}), \quad (3) \]
\[ E[V] = \frac{2\nu}{\beta} + o(1), \quad (4) \]

where \( \lambda > 0 \) is the same constant as Theorem 2.
The constant $\nu$ is called a singular fluctuation. Unfortunately, it is still difficult to obtain the concrete value of $\nu$, however, it can be estimated from random samples by eq.(4). If a model is regular then $\lambda = \nu = d/2$. By eliminating $\lambda$ and $\nu$ from eq.(2), eq.(3), and eq.(4), we obtain the widely applicable information criterion (WAIC),

$$E[G] = E[T] + \frac{\beta}{n} E[V] + o\left(\frac{1}{n}\right).$$

Moreover, we can prove that this equation holds for arbitrary set of $(q(x), p(x|w), \varphi(w))$, even if the true distribution can not be realized by the parametric model, in other words, $W_0$ is the empty set. Therefore, the Bayes generalization error can be estimated from the Bayes training error and the functional variance without any information or any assumption about the true distribution.

### 2.5 Main Formula IV

The minus log likelihood function or the log loss function is defined by

$$L_n(w) = -\sum_{i=1}^{n} \log p(X_i|w) - a_n \log \varphi(w),$$

where $\{a_n \geq 0\}$ is a non-decreasing function that satisfies

$$\lim_{n \to \infty} a_n n^b = 0$$

for an arbitrary $b > 0$. Let $\hat{w}$ be the parameter that minimizes $L_n(w)$. We assume that $W$ is contained in a compact set. The generalization and training errors of the maximum a posteriori estimation are respectively defined by

$$G = -\int q(x) \log p(x|\hat{w})dx,$$

$$T = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i|\hat{w}).$$

**Theorem 4** In the maximum a posteriori estimation, the symmetry of the generalization and training errors holds,

$$\lim_{n \to \infty} E[n(G - S)] = -\lim_{n \to \infty} E[n(T - S)] = \frac{1}{4} E\left[\min_{g(u) \in W_0} |\xi(u)|^2\right],$$

(5)

where $W_{00}$ is a subset of $W_0$ that is determined by the sequence $\{a_n\}$.

Since the limit value of eq.(5) is given by the average of the maximum value of the random process, the generalization error of one point estimation is far larger than that by the Bayes estimation in general. Therefore, the Bayes estimation is more appropriate than the maximum likelihood method and the maximum a posteriori method in singular statistical models.

### 2.6 Regular and Singular

This book generalizes the regular statistical theory to singular statistical theory. The correspondence between regular theory and singular theory is given in Table.2.6. From the mathematical point of view, singular statistical theory is obtained by the functional methodology.
<table>
<thead>
<tr>
<th></th>
<th>regular</th>
<th>singular</th>
</tr>
</thead>
<tbody>
<tr>
<td>algebra</td>
<td>linear algebra</td>
<td>ring and ideal</td>
</tr>
<tr>
<td>geometry</td>
<td>differential geometry</td>
<td>algebraic geometry</td>
</tr>
<tr>
<td>analysis</td>
<td>real-valued function</td>
<td>function-valued func.</td>
</tr>
<tr>
<td>probability theory</td>
<td>central limit theorem</td>
<td>empirical process</td>
</tr>
<tr>
<td>Fisher inform. matrix</td>
<td>positive definite</td>
<td>semi-positive def.</td>
</tr>
<tr>
<td>Cramer-Rao inequality</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>asymptotic normality</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ML estimator</td>
<td>asymptotic efficient</td>
<td>not efficient</td>
</tr>
<tr>
<td>Bayes a posteriori</td>
<td>normal distribution</td>
<td>singular dist.</td>
</tr>
<tr>
<td>learning coefficient</td>
<td>(d/2)</td>
<td>(\lambda)</td>
</tr>
<tr>
<td>singular fluctuation</td>
<td>(d/2)</td>
<td>(\nu)</td>
</tr>
<tr>
<td>Bayes marginal</td>
<td>((d/2) \log n)</td>
<td>(\lambda \log n)</td>
</tr>
<tr>
<td>Information criterion</td>
<td>AIC</td>
<td>WAIC</td>
</tr>
<tr>
<td>examples</td>
<td>normal, binomial</td>
<td>mixtures</td>
</tr>
<tr>
<td></td>
<td>linear regression</td>
<td>neural networks</td>
</tr>
<tr>
<td></td>
<td>linear prediction</td>
<td>hidden Markov</td>
</tr>
</tbody>
</table>

Table 1: Correspondence between regular and singular

### 3 Book Information

#### 3.1 Author, publisher, and series

Book Name: Algebraic Geometry and Statistical Learning Theory  
Author: Sumio Watanabe  
Publisher: Cambridge University Press, United Kingdom, August, 2009, ISBN=9780521864671  
Series: Cambridge Monographs on Applied and Computational Mathematics  

#### 3.2 Short History

For regular statistical models, asymptotic theory was established by R.A. Fisher, 1925. Akaike information criterion (AIC), Bayes information criterion (BIC), and the minimum description length (MDL) were proposed by Akaike 1974, Schwarz 1978, and Rissanen 1984, respectively.

The concept of zeta function was firstly found by Gel’fand and Silov, 1954. Resolution of singularities was proved by H. Hironaka, 1964. Application of resolution theorem to the zeta function was found by M.F. Atiyah, 1970. Bernstein-Sato polynomial was found by Bernstein and Sato independently, 1972.

Algebraic geometrical structure in statistical learning theory was firstly found in 1999 [1, 2, 3], where asymptotic expansion of the Bayes marginal, Theorem 2 was derived. It was applied to many learning machines, for example, normal mixtures [5], naive Bayesian networks [7], and reduced rank regressions [8]. It was introduced in the lectures of algebraic statistics [13]. The case when the Kullback-Leibler distance between the true parameter and singularities is in proportion to \(1/n\) was studied [4]. It was proved that the exchange ratio of the exchange Monte
Calro method is determined by the log canonical threshold [12]. The variational Bayes or the mean field approximation of singular models were studied [9, 10, 11]. The common standard form of the likelihood ratio function, Theorem 1, and the symmetry of the generalization and training errors, Theorem 4, were shown [6]. The equation of state in statistical learning, Theorem 3, was found [15, 16], even if the true distribution is not realized by the statistical model [17].

References


