1
Introduction

In this book, we study a system which perceives the real world. Such a system has to estimate an information source by observation. If the information source is a probability distribution, then the estimation process is called statistical learning, and the system is said to be a statistical model or a learning machine.

A lot of statistical models have hierarchical layers, hidden variables, a collection of modules, or grammatical structures. Such models are nonidentifiable and contain singularities in their parameter spaces. In fact, the map from a parameter to a statistical model is not one-to-one, and the Fisher information matrix is not positive definite. Such statistical models are called singular. It has been difficult to examine the learning process of singular models, because there has been no mathematical theory for such models.

In this book, we establish a mathematical foundation which enables us to understand the learning process of singular models. This chapter gives an overview of the book before a rigorous mathematical foundation is developed.

1.1 Basic concepts in statistical learning

To describe what statistical learning is, we need some basic concepts in probability theory. For the reader who is unfamiliar with probability theory, Section 1.6 summarizes the key results.

1.1.1 Random samples

Let N be a natural number and $\mathbb{R}^N$ be the N-dimensional real Euclidean space. We study a case when information data are represented by vectors in $\mathbb{R}^N$.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

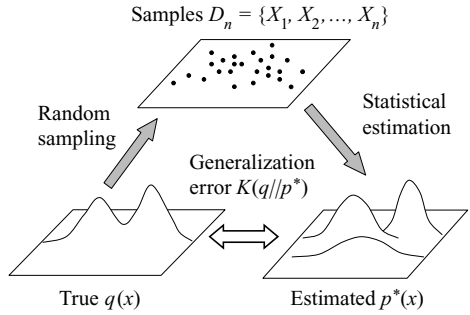


Fig. 1.1. Statistical learning

Firstly, using Figure 1.1, let us explain what statistical learning is. Let $(\Omega, \mathcal{B}, P)$ be a probability space and $X : \Omega \to \mathbb{R}^N$ be a random variable which is subject to a probability distribution $q(x)dx$. Here $q(x)$ is a probability density function and dx is the Lebesgue measure on $\mathbb{R}^N$.

We assume that random variables $X_1, X_2, \ldots, X_n$ are independently subject to the same probability distribution as X, where n is a natural number. In statistical learning theory, $q(x)$ is called a true probability density function, and random variables $X_1, X_2, \ldots, X_n$ are random samples, examples, random data, or training samples. The probability density function of the set of independent random samples is given by

$$q(x_1)q(x_2)\cdots q(x_n).$$

In practical applications, we obtain their realizations by observations. The natural number n is said to be the number of random samples. The set of random samples or random data is denoted by

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

The main purpose of statistical learning is to construct a method to estimate the true probability density function $q(x)$ from the set D_n.

In this book, we study a method which employs a parametric probability density function. A conditional probability density function $p(x|w)$ of $x \in \mathbb{R}^N$ for a given parameter $w \in W$ is called a learning machine or a statistical model, where W is the set of all parameters. Sometimes the notation $p(x|w) = p_w(x)$ is used. Then $w \mapsto p_w$ gives a map from the parameter to the probability density function. We mainly study the case when W is a subset of the d-dimensional real Euclidean space $\mathbb{R}^d$ or a d-dimensional real analytic manifold. An *a priori* probability density function $\varphi(w)$ is defined on W. We assume that, for any $w \in W$, the support of the probability distribution $p(x|w)$ is equal to that of

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

$q(x)$ and does not depend on w. That is to say, for any $w \in W$,

$$\overline{\{x \in \mathbb{R}^N; p(x|w) > 0\}} = \overline{\{x \in \mathbb{R}^N; q(x) > 0\}},$$

where $\overline{S}$ is the closure of a set S in $\mathbb{R}^N$.

In statistical learning or statistical inference, the main study concerns a method to produce a probability density function $p^*(x)$ on $\mathbb{R}^N$ based on the set of random samples D_n, using the parametric model $p(x|w)$. Such a function

$$D_n \mapsto p^*(x)$$

is called a statistical estimation method or a learning algorithm. Note that there are a lot of statistical estimation methods and learning algorithms. The probability density function $p^*(x)$, which depends on the set of random variables D_n, is referred to as the estimated or trained probability density function. Generally, it is expected that the estimated probability density function $p^*(x)$ is a good approximation of the true density function $q(x)$, and that it becomes better as the number of random samples increases.

1.1.2 Kullback–Leibler distance

In order to compare two probability density functions, we need a quantitative value which shows the difference between two probability density functions.

Definition 1.1 (Kullback–Leibler distance) For given probability density functions $q(x)$, $p(x) > 0$ on an open set $A \subset \mathbb{R}^N$, the Kullback–Leibler distance or relative entropy is defined by

$$K(q\|p) = \int_A q(x) \log \frac{q(x)}{p(x)} dx.$$

If the integral is not finite, $K(q\|p)$ is defined as $K(q\|p) = \infty$.

Theorem 1.1 *Assume that $q(x)$, $p(x) > 0$ are continuous probability density functions on an open set A. Then the following hold.*
(1) For arbitrary $q(x)$, $p(x)$, $K(q\|p) \geq 0$.
(2) $K(q\|p) = 0$ if and only if $q(x) = p(x)$ for any $x \in A$.

Proof of Theorem 1.1 Let us introduce a real function

$$S(t) = -\log t + t - 1 \quad (0 < t < \infty).$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Then $S(t) \geq 0$, and $S(t) = 0$ if and only if $t = 1$. Since $\int q(x)dx = \int p(x)dx = 1$,

$$K(q\|p) = \int_A q(x)\, S\Big(\frac{p(x)}{q(x)}\Big)\, dx,$$

which shows (1). Assume $K(q\|p) = 0$. Since $S(p(x)/q(x))$ is a nonnegative and continuous function of x, $S(p(x)/q(x)) = 0$ for any $x \in A$, which is equivalent to $p(x) = q(x)$. □

Remark 1.1 The Kullback–Leibler distance is called the relative entropy in physics. In information theory and statistics, the Kullback–Leibler distance $K(q\|p)$ represents the loss of the system $p(x)$ for the information source $q(x)$. The fact that $K(q\|p)$ is not symmetric for $q(x)$ and $p(x)$ may originate from the difference of their roles. Historically, relative entropy was first defined by Boltzmann and Gibbs in statistical physics in the nineteenth century. In the twentieth century it was found that relative entropy plays a central role in information theory and statistical estimation.

We can measure the difference between the true density function $q(x)$ and the estimated one $p^*(x)$ by the Kullback–Leibler distance:

$$K(q\|p^*) = \int q(x) \log \frac{q(x)}{p^*(x)} dx.$$

In statistical learning theory, $K(q\|p^*)$ is called the generalization error of the method of statistical estimation $D_n \mapsto p^*$. In general, $K(q\|p^*)$ is a measurable function of the set of random samples D_n, hence it is also a real-valued random variable. The training error is defined by

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

which is also a random variable. One of the main purposes of statistical learning theory is to clarify the probability distributions of the generalization and training errors for a given method of statistical estimation. The expectation values $E[K(q\|p^*)]$ and $E[K_n(q\|p^*)]$ are respectively called the mean generalization error and the training error. If the mean generalization error is smaller, the statistical estimation method is more appropriate. The other purpose of statistical learning theory is to establish a mathematical relation between the generalization error and the training error. If the generalization error can be estimated from the training error, we can select the suitable model or hyperparameter among several statistical possible models.

Definition 1.2 (Likelihood function) For a given set of random samples D_n and a statistical model $p(x|w)$, the likelihood function $L_n(w)$ of $w \in W \subset \mathbb{R}^d$ is defined by

$$L_n(w) = \prod_{i=1}^{n} p(X_i|w).$$

If $p(x|w) = q(x)$, then $L_n(w)$ is equal to the probability density function of D_n.

Definition 1.3 (Log likelihood ratio function) For a given true distribution $q(x)$ and a parametric model $p(x|w)$, the log density ratio function $f(x, w)$, the Kullback–Leibler distance $K(w)$, and the log likelihood ratio function $K_n(w)$ are respectively defined by

$$f(x, w) = \log \frac{q(x)}{p(x|w)}, \tag{1.1}$$

$$K(w) = \int q(x) f(x, w) dx, \tag{1.2}$$

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w), \tag{1.3}$$

where $K_n(w)$ is sometimes referred to as an empirical Kullback–Leibler distance.

From the definition,

$$E[f(X, w)] = E[K_n(w)] = K(w).$$

By using the empirical entropy

$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i), \tag{1.4}$$

the likelihood function satisfies

$$-\frac{1}{n} \log L_n(w) = K_n(w) + S_n. \tag{1.5}$$

The empirical entropy S_n does not depend on the parameter w, hence maximization of the likelihood function $L_n(w)$ is equivalent to minimization of $K_n(w)$.

Remark 1.2 If a function $S(t)$ satisfies $S''(t) > 0$ and $S(1) = 0$, then

$$\int_A q(x)\, S\Big(\frac{p(x)}{q(x)}\Big)\, dx$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

has the same property as the Kullback–Leibler distance in Theorem 1.1. For example, using $S(t) = (1 - t^a)/a$ for a given $a, 0 < a < 1$, a generalized distance is defined by

$$K^{(a)}(q\|p) = \int q(x)\Big(\frac{1 - (p(x)/q(x))^a}{a}\Big)\, dx.$$

For example, if $a = 1/2$, Hellinger's distance is derived,

$$K^{(1/2)}(q\|p) = \int (\sqrt{q(x)} - \sqrt{p(x)})^2 dx.$$

In general Jensen's inequality claims that, for any measurable function $F(x)$,

$$\int q(x)S(F(x))dx \geq S\Big(\int q(x)F(x)dx\Big),$$

where the equality holds if and only if $F(x)$ is a constant function on $q(x) > 0$. Hence $K^{(a)}(q\|p) \geq 0$ and $K^{(a)}(q\|p) = 0$ if and only if $q(x) = p(x)$ for all x. Hence $K^{(a)}(q\|p)$ indicates a difference of $p(x)$ from $q(x)$. The Kullback–Leibler distance is formally obtained by $a \to +0$. For arbitrary probability density functions $q(x)$, $p(x)$, the Kullback–Leibler distance satisfies $K(q\|p) \geq K^{(a)}(q\|p)$, because

$$K(q\|p) - K^{(a)}(q\|p) = \int q(x)\Big(\frac{af(x,w) + e^{-af(x,w)} - 1}{a}\Big)dx \geq 0.$$

Moreover, if $K(q\|p) \neq 0$ then

$$\lim_{a\to+0} \frac{K_a(q\|p)}{K(q\|p)} = 1.$$

Therefore, from the learning theory of $K(q\|p)$, we can construct a learning theory of $K^{(a)}(q\|p)$.

Remark 1.3 If $E[K(w)] < \infty$ then, by the law of large numbers, the convergence in probability

$$K_n(w) \to K(w)$$

holds for each $w \in W$. Furthermore, if $E[K(w)^2] < \infty$ then, by the central limit theorem,

$$\sqrt{n}(K_n(w) - K(w))$$

converges in law to the normal distribution, for each $w \in W$. Therefore, for each w, the convergence in probability

$$-\frac{1}{n}\log L_n(w) \to K(w) - S$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

holds, where S is the entropy of the true distribution $q(x)$,

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

It might seem that minimization of $K_n(w)$ is equivalent to minimization of $K(w)$. If these two minimization problems were equivalent, then maximization of $L_n(w)$ would be the best method in statistical estimation. However, minimization and expectation cannot be commutative.

$$E[\min_w K_n(w)] \neq \min_w E[K_n(w)] = \min_w K(w). \tag{1.6}$$

Hence maximization of $L_n(w)$ does not mean minimization of $K(w)$. This is the basic reason why statistical learning does not result in a simple optimization problem. To clarify the difference between $K(w)$ and $K_n(w)$, we have to study the meaning of the convergence $K_n(w) \to K(w)$ in a functional space. There are many nonequivalent functional topologies. For example, sup-norm, L^p-norm, weak topology of Hilbert space L^2, Schwartz distribution topology, and so on. It strongly depends on the topology of the function space whether the convergence $K_n(w) \to K(w)$ holds or not. The Bayes estimation corresponds to the Schwartz distribution topology, whereas the maximum likelihood or *a posteriori* method corresponds to the sup-norm. This difference strongly affects the learning results in singular models.

1.1.3 Fisher information matrix

Definition 1.4 (Fisher information matrix) For a given statistical model or a learning machine $p(x|w)$, where $x \in \mathbb{R}^N$ and $w \in \mathbb{R}^d$, the Fisher information matrix

$$I(w) = \{I_{jk}(w)\} \quad (1 \leq j, k \leq d)$$

is defined by

$$I_{jk}(w) = \int \Big(\frac{\partial}{\partial w_j} \log p(x|w)\Big) \Big(\frac{\partial}{\partial w_k} \log p(x|w)\Big)\, p(x|w)\, dx$$

if the integral is finite.

By the definition, the Fisher information matrix is always symmetric and positive semi-definite. It is not positive definite in general. In some statistics textbooks, it is assumed that the Fisher information matrix is positive definite, and that the Cramer–Rao inequality is proven; however, there are a lot of statistical models and learning machines in which Fisher information matrices

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

have zero eigenvalue. The Fisher information matrix is positive definite if and only if

$$\left\{\frac{\partial}{\partial w_j}\log p(x|w)\right\}_{j=1}^{d}$$

is linearly independent as a function of x on the support of $p(x|w)$. Since

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

the Fisher information matrix is positive definite if and only if

$$\left\{\frac{\partial}{\partial w_j} f(x, w)\right\}_{j=1}^{d}$$

is linearly independent as a function of x. By using $\int p(x|w)dx = 1$ for an arbitrary w, it is easy to show that

$$I_{jk}(w) = -\int \left(\frac{\partial^2}{\partial w_j \partial w_k}\log p(x|w)\right) p(x|w)\, dx.$$

If $q(x) = p(x|w_0)$, then

$$I_{jk}(w_0) = \frac{\partial^2}{\partial w_j \partial w_k} K(w_0).$$

Therefore, the Fisher information matrix is equal to the Hessian matrix of the Kullback–Leibler distance at the true parameter.

Remark 1.4 If the Fisher information matrix is positive definite in a neighborhood of the true parameter w_0, $K(w) > 0$ $(w \neq w_0)$ holds, and the Kullback–Leibler distance can be approximated by the positive definite quadratic form, then

$$K(w) \approx \tfrac{1}{2}(w - w_0) \cdot I(w_0)(w - w_0),$$

where $u \cdot v$ shows the inner product of two vectors u, v. If the Fisher information matrix is not positive definite, then $K(w)$ cannot be approximated by any quadratic form in general. This book establishes the mathematical foundation for the case when the Fisher information matrix is not positive definite.

Remark 1.5 (Cramer–Rao inequality) Assume that random samples $\{X_i; i = 1, 2, \ldots, n\}$ are taken from the probability density function $\prod_{i=1}^{n} p(x_i|w)$, where $w = (w_1, w_2, \ldots, w_d) \in \mathbb{R}^d$. A function from random samples to the parameter space

$$\{u_j(x_1, x_2, \ldots, x_n); j = 1, 2, \ldots, d\} \in \mathbb{R}^d$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

is called an unbiased estimator if it satisfies

$$E[u_j(X_1, X_2, \ldots, X_n) - w_j] \equiv \int (u_j(x_1, x_2, \ldots, x_n) - w_j) \times \prod_{i=1}^{n} p(x_i|w) dx_i = 0$$

for arbitrary $w \in \mathbb{R}^d$. Under certain are conditions which ensure that $\int dx_j$ and $(\partial/\partial w_k)$ are commutative for arbitrary j, k,

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_k} E[u_j(X_1, X_2, \ldots, X_n) - w_j] \\ &= E[(u_j - w_j) \sum_{i=1}^{n} \frac{\partial}{\partial w_k} \log p(X_i, w)] - \delta_{jk}. \end{aligned}$$

Therefore,

$$\delta_{jk} = E[(u_j - w_j) \sum_{i=1}^{n} \frac{\partial}{\partial w_k} \log p(X_i, w)].$$

For arbitrary d-dimensional vectors $\mathbf{a} = (a_j)$, $\mathbf{b} = (b_k)$,

$$(\mathbf{a} \cdot \mathbf{b}) = E\Big[\Big(\sum_{j=1}^{d} a_j(u_j - w_j)\Big)\Big(\sum_{k=1}^{d}\sum_{i=1}^{d} b_k \frac{\partial}{\partial w_k} \log p(X_i, w)\Big)\Big].$$

By applying the Cauchy–Schwarz inequality

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq n\ (\mathbf{a} \cdot V\mathbf{a})(\mathbf{b} \cdot I(w)\mathbf{b}), \tag{1.7}$$

where $V = (V_{jk})$ is the covariance matrix of $u - w$,

$$V_{jk} = E[(u_j - w_j)(u_k - w_k)]$$

and $I(w)$ is the Fisher information matrix. If $I(w)$ is positive definite, by putting

$$\mathbf{a} = I(w)^{1/2}\mathbf{c},$$
$$\mathbf{b} = I(w)^{-1/2}\mathbf{c},$$

it follows that

$$\|\mathbf{c}\|^2 \leq n\ (\mathbf{c} \cdot I(w)^{1/2} V I(w)^{1/2}\mathbf{c})$$

holds for arbitrary vector $\mathbf{c}$, hence

$$V \geq \frac{I(w)^{-1}}{n}. \tag{1.8}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

This relation, the Cramer–Rao inequality, shows that the covariance matrix of any unbiased estimator cannot be made smaller than the inverse of the Fisher information matrix. If $I(w)$ has zero eigenvalue and $\mathbf{b}$ is an eigenvector for zero eigenvalue, eq.(1.7) shows that either V is not a finite matrix or no unbiased estimator exists. For statistical models which have a degenerate Fisher information matrix, we have no effective unbiased estimator in general.

1.2 Statistical models and learning machines

1.2.1 Singular models

Definition 1.5 (Identifiablity) A statistical model or a learning machine $p(x|w)$ ($x \in \mathbb{R}^N$, $w \in W \subset \mathbb{R}^d$) is called identifiable if the map

$$W \ni w \mapsto p(\ |w)$$

is one-to-one, in other words,

$$p(x|w_1) = p(x|w_2) \quad (\forall x \in \mathbb{R}^d) \Longrightarrow w_1 = w_2.$$

A model which is not identifiable is called nonidentifiable or unidentifiable.

Definition 1.6 (Positive definite metric) A statistical model or a learning machine $p(x|w)$ ($x \in \mathbb{R}^N$, $w \in W \subset \mathbb{R}^d$) is said to have a positive definite metric if its Fisher information matrix $I(w)$ is positive definite for arbitrary $w \in \mathbf{W}$. If a statistical model does not have a positive definite metric, it is said to have a degenerate metric.

Definition 1.7 (Singular statistical models) Assume that the support of the statistical model $p(x|w)$ is independent of w. A statistical model $p(x|w)$ is said to be regular if it is identifiable and has a positive definite metric. If a statistical model is not regular, then it is called strictly singular. The set of singular statistical models consists of both regular and strictly singular models.

Mathematically speaking, identifiability is neither a necessary nor a sufficient condition of positive definiteness of the Fisher information matrix. In fact, if $p(x|a)$ ($x, a \in \mathbb{R}^1$) is a regular statistical model, then $p(x|a^3)$ is identifiable but has a degenerate Fisher information matrix. Also $p(x|a^2)$ ($|a| > 1$) has a nondegenerate Fisher information matrix but is nonidentifiable. These are trivial examples in which an appropriate transform or restriction of a parameter makes models regular.

However, a lot of statistical models and learning machines used in information science have simultaneously nonidentifiability and a degenerate metric.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Moreover, they contain a lot of singularities which cannot be made regular by any transform or restriction.

In this book, we mainly study singular statistical models or singular learning machines. The following statistical models are singular statistical models.

(1) Layered neural networks
(2) Radial basis functions
(3) Normal mixtures
(4) Binomial and multinomial mixtures
(5) Mixtures of statistical models
(6) Reduced rank regressions
(7) Boltzmann machines
(8) Bayes networks
(9) Hidden Markov models
(10) Stochastic context-free grammar

These models play the central role of information processing systems in artificial intelligence, pattern recognition, robotic control, time series prediction, and bioinformatics. They determine the preciseness of the application systems. Singular models are characterized by the following features.

(1) They are made by superposition of parametric functions.
(2) They have hierarchical structures.
(3) They contain hidden variables.
(4) They consist of several information processing modules.
(5) They are designed to obtain hidden knowledge from random samples.
(6) They estimate the probabilistic grammars.

In singular statistical models, the knowledge or grammar to be discovered corresponds to singularities in general. Figure 1.2 shows an example of the correspondence between parameters and probability distributions in normal mixtures.

Remark 1.6 (Equivalence relation) The condition that $p(x|w_1) = p(x|w_2)$ for arbitrary x does not mean

$$\frac{\partial^k p(x|w_1)}{\partial w_1^k} = \frac{\partial^k p(x|w_2)}{\partial w_2^k} \quad (k = 1, 2, 3, \ldots).$$

Even if $p(x|w_1) \approx p(x|w_2)$, their derivatives are very different in general. The preciseness of statistical estimation is determined by the derivative of $p(x|w)$, hence results of statistical estimations are very different if $p(x|w_1) \approx p(x|w_2)$.

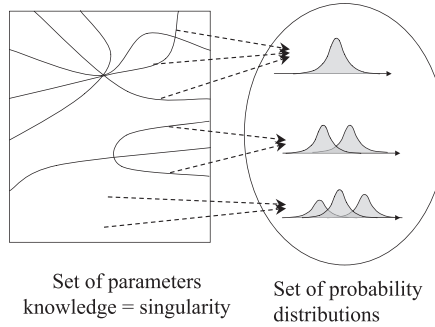


Fig. 1.2. Map from parameter to probability distribution

One can introduce an equivalence relation $\sim$ into the set of parameters W,

$$w_1 \sim w_2 \Longleftrightarrow p(x|w_1) = p(x|w_2) \quad (\forall x).$$

Then the map $(W/_{\sim}) \ni w \to p_w$ is one-to-one. However, the quotient set $(W/_{\sim})$ is neither the Euclidean space nor a manifold. Therefore, it is still difficult to construct statistical learning theory on $(W/_{\sim})$. In this book, we show that there is a birational map in algebraic geometry which enables us to establish singular learning theory.

Remark 1.7 (No asymptotic normality) If a model is regular, then the Bayes *a posteriori* distribution can be approximated by the normal distribution

$$\frac{1}{Z_n} \exp\Big(-\frac{n}{2}(w - w_0)I(w_0)(w - w_0)\Big),$$

where w_0 is the unique parameter such that $q(x) = p(x|w_0)$. Also the maximum likelihood estimator and the maximum *a posteriori* estimator are asymptotically subject to the normal distribution. Such a property is called asymptotic normality. However, singular statistical models do not have such a property, with the result that almost all statistical theories using asymptotic normality do not hold in singular statistical models.

Remark 1.8 (True generic condition) In a lot of statistical models and learning machines, the set of parameters at which the Fisher information matrices are degenerate

$$W_{(0)} = \{w \in W \ ; \ \det(I(w)) = 0\}$$

is a measure zero subset in $\mathbb{R}^d$. Hence one might suppose that, in generic cases, the true parameter w_0 is seldom contained in $W_{(0)}$, and that the learning theory assuming $\det(I(w_0)) > 0$ may be sufficient in practical applications. However,

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

this consideration is wrong. On the contrary, in general cases, we have to optimize a statistical model or a learning machine by comparing several probable models and hyperparameters. In such cases, we always examine models under the condition that the optimal parameter lies in a neighborhood of $W_{(0)}$. Especially in model selection, hyperparameter optimization, or hypothesis testing, we need the theoretical results of the case $w_0 \in W_{(0)}$ because we have to determine whether $w_0 \in W_{(0)}$ or not. Therefore, the superficial generic condition $\det(I(w)) > 0$ does not have true generality.

Remark 1.9 (Singular theory contains regular theory) Statistical theory of regular models needs identifiability and a nondegenerate Fisher information matrix. In this book, singular learning theory is established on the assumption that neither identifiability nor a positive definite Fisher information matrix is necessary. Of course, even if a model is regular, the singular learning theory holds. In other words, a regular model is understood as a very special example to which singular learning theory can be applied. From the mathematical point of view, singular learning theory contains regular learning theory as a very special part. For example, the concepts AIC (Akaike's information criterion) and BIC (Bayes information criterion) in regular statistical theory are completely generalized in this book.

1.2.2 Density estimation

Let us introduce some examples of regular and singular statistical models.

Example 1.1 (Regular model) A parametric probability density function of $(x, y) \in \mathbb{R}^2$ for a given parameter $w = (a, b) \in \mathbb{R}^2$ defined by

$$p(x, y|a, b) = \frac{1}{2\pi} \exp\Big(-\frac{(x-a)^2 + (y-b)^2}{2}\Big)$$

is a regular statistical model, where the set of parameters is $W = \{(a, b) \in \mathbb{R}^2\}$. This is a two-dimensional normal distribution. For given random samples (X_i, Y_i), the likelihood function is

$$L_n(a, b) = \frac{1}{(2\pi)^n} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\{(X_i - a)^2 + (Y_i - b)^2\}\Big).$$

If the true distribution is given by (a_0, b_0),

$$q(x, y) = p(x, y|a_0, b_0),$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

the log likelihood ratio function is

$$K_n(a,b) = \frac{a^2 - a_0^2 + b^2 - b_0^2}{2} - (a-a_0)\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) - (b-b_0)\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big).$$

The Kullback–Leibler distance is

$$K(a,b) = \tfrac{1}{2}\{(a-a_0)^2 + (b-b_0)^2\}.$$

Note that $K(a,b) = 0$ if and only if $a = a_0$ and $b = b_0$. The Fisher information matrix

$$I(a,b) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is positive definite for an arbitrary (a,b).

Example 1.2 (Singular model) Let us introduce another parametric probability density function of $x \in \mathbb{R}^1$ defined by

$$p(x|a,b) = \frac{1}{\sqrt{2\pi}}\Big\{(1-a)\,e^{-\frac{x^2}{2}} + a\,e^{-\frac{(x-b)^2}{2}}\Big\}.$$

The set of parameters is

$$W = \{w = (a,b); 0 \le a \le 1,\ -\infty < b < \infty\}.$$

This model is called a normal mixture. If the true distribution is given by

$$q(x) = p(x|a_0, b_0),$$

then the log likelihood ratio function is

$$K_n(a,b) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{1 + a_0\,(\exp(b_0 X_i - b_0^2/2) - 1)}{1 + a\,(\exp(bX_i - b^2/2) - 1)}\Big)$$

and

$$K(a,b) = \int \log\Big(\frac{1 + a_0\,(\exp(b_0 x - b_0^2/2) - 1)}{1 + a\,(\exp(bx - b^2/2) - 1)}\Big)\, q(x)\,dx.$$

If $a_0 b_0 \neq 0$, then $K(a,b) = 0$ is equivalent to $a = a_0$ and $b = b_0$. In such cases, the Fisher information matrix $I(a_0, b_0)$ is positive definite. However, if $a_0 b_0 = 0$, then $K(a,b) = 0$ is equivalent to $ab = 0$, and the Fisher information matrix $I(a_0, b_0) = 0$. The function $K(a,b)$ can be expanded as

$$K(a,b) = \tfrac{1}{2}a^2 b^2 + \cdots,$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

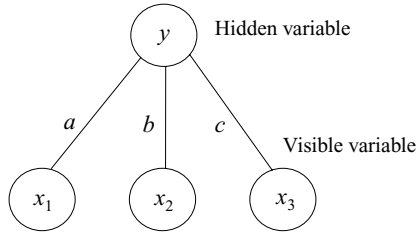


Fig. 1.3. Bayesian network with hidden unit

which shows that $K(a, b)$ cannot be approximated by any quadratic form. If we make a model selection algorithm or a hypothesis test procedure for this model, then we have to study the case $K(a, b) \cong 1/\sqrt{n}$ where n is the number of random samples. Hence we have to evaluate the effect of the singularity in the set $ab = 0$.

Example 1.3 (Bayesian network with a hidden unit) Let X_1, X_2, X_3 and Y are random variables which take values $\{-1, 1\}$. The Bayesian network shown in Figure 1.3 is defined by the probability distribution of $X = (X_1, X_2, X_3)$ and Y,

$$p(x, y|w) = \frac{1}{Z(a, b, c)} \exp(ax_1y + bx_2y + cx_3y),$$

where $w = (a, b, c)$ and $Z(a, b, c)$ is a normalizing constant. Let X be a set of visible units and Y a hidden unit. The probability distribution of X is given by the marginal distribution,

$$\begin{aligned} p(x|w) &= \frac{1}{Z(a, b, c)} \sum_{y=\pm 1} \exp(ax_1y + bx_2y + cx_3y) \\ &= \frac{1}{2\ Z(a, b, c)} \cosh(ax_1 + bx_2 + cx_3). \end{aligned}$$

By using $\tanh(ax_i) = \tanh(a)x_i$ for $x_i = \pm 1$, and

$$\begin{aligned} \cosh(u + v) &= \cosh(u)\cosh(v) + \sinh(u)\sinh(v), \\ \sinh(u + v) &= \sinh(u)\cosh(v) + \cosh(u)\sinh(v), \end{aligned}$$

we have

$$p(x|w) = \tfrac{1}{8}\{1 + t(a)t(b)x_1x_2 + t(b)t(c)x_2x_3 + t(c)t(a)x_3x_1\},$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $t(a) = \tanh(a)$. Assume that the true distribution is given by $q(x) = p(x|0, 0, 0) = 1/8$. Then the Kullback–Leibler distance is

$$K(a, b, c) = \frac{1}{2}(a^2b^2 + b^2c^2 + c^2a^2) + \cdots .$$

Therefore

$$q(x) = p(x|a, b, c) \Longleftrightarrow a = b = 0, \quad \text{or} \quad b = c = 0, \quad \text{or} \quad c = a = 0.$$

The Fisher information matrix is equal to zero at $(0, 0, 0)$. If we want to judge whether the hidden variable Y is necessary to explain a given set of random samples, we should clarify the effect of the singularity of $K(a, b, c) = 0$.

1.2.3 Conditional probability density

Example 1.4 (Regular model) A probability density function of $(x, y) \in \mathbb{R}^2$,

$$p(x, y|a, b) = q_0(x) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - ax - b)^2), \tag{1.9}$$

is a statistical model, where the set of parameters is $W = \{w = (a, b) \in \mathbb{R}^2\}$ and $q_0(x)$ is a constant probability density function of x. This model is referred to as a line regression model. If the true distribution is $q(x, y)$, the true conditional probability density function

$$q(y|x) = \frac{q(x, y)}{\int q(x, y')\, dy'}$$

is estimated by the conditional probability density function

$$p(x|y, a, b) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - ax - b)^2). \tag{1.10}$$

The two models eq.(1.9) and eq.(1.10) have the same log likelihood ratio function and the same Kullback–Leibler distance, hence the two models are equivalent from a statistical point of view. In other words, estimation of the conditional density function of y for a given x can be understood as the estimation of a joint probability density function of (x, y), if $q(x)$ is not estimated.

If the true distribution is given by $w_0 = (a_0, b_0)$,

$$q(x, y) = p(x, y|a_0, b_0).$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

The log likelihood ratio function is

$$K_n(a,b) = \frac{(a^2-a_0^2)}{2}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) + \frac{(b^2-b_0^2)}{2} + (ab-a_0b_0)\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)$$
$$-(a-a_0)\Big(\frac{1}{n}\sum_{i=1}^n X_iY_i\Big) - (b-b_0)\Big(\frac{1}{n}\sum_{i=1}^n Y_i\Big).$$

The Kullback–Leibler distance is

$$K(a,b) = \frac{1}{2}\int (ax+b-a_0x-b_0)^2\, q_0(x)\, dx$$
$$= \tfrac{1}{2}(w-w_0)\cdot I(w-w_0),$$

where

$$I = \begin{pmatrix} m_2 & m_1 \\ m_1 & 1 \end{pmatrix},$$

and

$$m_i = \int x^i\, q_0(x)\, dx.$$

The Fisher information matrix is always equal to I, which does not depend on the true parameter w_0. It is positive definite if and only if $m_2 \neq m_1^2$. In other words, I is degenerate if and only if the variance of $q_0(x)$ is equal to zero.

Example 1.5 (Singular model) Another example of a statistical model of $y \in \mathbb{R}^1$ for $x \in \mathbb{R}^1$ is

$$p(x,y|a,b) = q_0(x)\,\frac{1}{\sqrt{2\pi}}\,\exp(-\tfrac{1}{2}(y-a\tanh(bx))^2),$$

where the set of parameters is $W = \{(a,b) \in \mathbb{R}^2\}$. This model is the simplest three-layer neural network. If the true distribution is given by $w = (a_0, b_0)$,

$$q(x,y) = p(x,y|a_0,b_0),$$

the log likelihood ratio function is

$$K_n(a,b) = \frac{1}{2n}\sum_{i=1}^n \{(Y_i - a\tanh(bX_i))^2 - (Y_i - a_0\tanh(b_0X_i))^2\},$$

and the Kullback–Leibler distance is

$$K(a,b) = \frac{1}{2}\int (a\tanh(bx) - a_0\tanh(b_0x))^2\, q_0(x)\, dx.$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

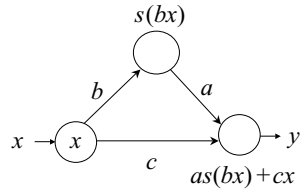


Fig. 1.4. Layered neural network

If $a_0 b_0 \neq 0$, then $K(a, b) = 0$ if and only if $a = a_0$ and $b = b_0$, and the Fisher information matrix $I(a_0, b_0)$ is positive definite. However, if $a_0 b_0 = 0$, then $K(a, b) = 0$ is equivalent to $ab = 0$, and the Fisher information matrix is degenerate. In practical applications of three-layer neural networks, we have to decide whether a three-layer neural network

$$\sum_{h=1}^{H} a_h \tanh(b_h x + c_h)$$

almost approximates the true regression function or not. In such cases, the more precisely the model approximates the true regression function, the more degenerate the Fisher information matrix is. Therefore, we cannot assume that the Fisher information matrix is positive definite in the model evaluation.

Example 1.6 (Layered neural network) Let $x, y \in \mathbb{R}^1$ and $w = (a, b, c) \in \mathbb{R}^3$. The statistical model shown in Figure 1.4,

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}(y - as(bx) - cx)^2),$$

where $s(t) = t + t^2$, is a layered statistical model. If the true distribution is $q(y|x) = p(y|x, 0, 0, 0)$ and if $q(x)$ is the standard normal distribution, then

$$K(a, b, c) = \tfrac{1}{2}(ab + c)^2 + \tfrac{3}{2}a^2 b^2.$$

Hence

$$q(y|x) = p(y|x, w) \Longleftrightarrow ab = c = 0.$$

The Fisher information matrix is equal to zero at $(0, 0, 0)$.

1.3 Statistical estimation methods

In this section, let us introduce some statistical estimation methods, Bayes and Gibbs estimations, the maximum likelihood and *a posteriori* estimations.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

1.3.1 Evidence

Let $D_n = \{X_1, X_2, \ldots, X_n\}$ be a set of random samples. For a given set of a statistical model $p(x|w)$ and an *a priori* probability density function $\varphi(w)$, the *a posteriori* probability density function $p(w|D_n)$ with the inverse temperature $\beta > 0$ is defined by

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

where Z_n is the normalizing constant determined so that $p(w|D_n)$ is a probability density function of w,

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

If $\beta = 1$, $p(w|D_n)$ is called a strict Bayes *a posteriori* density; if $\beta \neq 1$, it is a generalized version. When $\beta \to \infty$, it converges to $\delta(w - \hat{w})$, where $\hat{w}$ is the maximum likelihood estimator. Note that Z_n is a measurable function of D_n, hence it is also a random variable. The random variable Z_n is called the evidence, the marginal likelihood, or the partition function.

Remark 1.10 (Meaning of evidence) If $\beta = 1$, $Z_n = Z_n(X_1, X_2, \ldots, X_n)$ satisfies

$$\int dx_1 dx_2 \cdots dx_n Z_n(x_1, x_2, \ldots, x_n) = 1.$$

Therefore, Z_n with $\beta = 1$ defines a probability density function of D_n for a given pair of $p(x|w)$ and φ. In other words, Z_n can be understood as a likelihood function of the pair $(p(x|w), \varphi(w))$.

The predictive distribution $p(x|D_n)$ is defined by

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

The Bayes estimation is defined by

$$p^*(x) = p(x|D_n),$$

in other words, the Bayes estimation is the map

$$D_n \mapsto p^*(x) = p(x|D_n).$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

The Bayes generalization error $B_{\rm g}$ is the Kullback–Leibler distance from $q(x)$ to $p^*(x)$,

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

Here $B_{\rm g}$ is a measurable function of D_n, hence it is also a random variable. The Bayes training error $B_{\rm t}$ is defined by

$$B_{\rm t} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|D_n)}.$$

In Bayes learning theory, there are several important observables. The stochastic complexity, the minus log marginal likelihood, or the free energy is defined by

$$F_n = -\log Z_n.$$

Since Z_n with $\beta = 1$ can be understood as the likelihood of the pair $p(x|w)$ and $\varphi(w)$, F_n with $\beta = 1$ is the minus log likelihood of them.

In analysis of $B_{\rm g}$, $B_{\rm t}$, and F_n, we have some useful equations. The normalized evidence is defined by

$$Z_n^0 = \frac{Z_n}{\prod_{i=1}^{n} q(X_i)^{\beta}}. \tag{1.11}$$

Then, by using eq.(1.5), the *a posteriori* distribution is rewritten as

$$p(w|D_n) = \frac{1}{Z_n^0} \exp(-n\beta K_n(w))\ \varphi(w),$$

where $K_n(w)$ is the log likelihood ratio function defined in eq.(1.3), and

$$Z_n^0 = \int dw\ \varphi(w)\ \exp(-n\beta K_n(w)).$$

In the same way, the normalized stochastic complexity is defined by

$$F_n^0 = -\log Z_n^0.$$

The empirical entropy is given by

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

Then

$$F_n = F_n^0 - n\beta S_n.$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

By the definition of the predictive distribution, it follows that

$$p(X_{n+1}|D_n) = \frac{\displaystyle\int dw\ \varphi(w)\ p(X_{n+1}|w) \prod_{i=1}^{n} p(X_i|w)^{\beta}}{\displaystyle\int dw\ \varphi(w) \prod_{i=1}^{n} p(X_i|w)^{\beta}}. \tag{1.12}$$

Theorem 1.2 *For an arbitrary natural number n, the Bayes generalization error with* $\beta = 1$ *and its mean satisfy the following equations.*

$$B_{\rm g} = E_{X_{n+1}}\left[F^0_{n+1}\right] - F^0_n,$$
$$E[B_{\rm g}] = E\left[F^0_{n+1}\right] - E[F^0_n].$$

Proof of Theorem 1.2 From eq.(1.12) with $\beta = 1$,

$$p(X_{n+1}|D_n) = \frac{Z_{n+1}}{Z_n}$$

holds for an arbitrary natural number n. The logarithm of this equation results in

$$-\log p(X_{n+1}|D_n) = F_{n+1} - F_n.$$

Also,

$$\frac{p(X_{n+1}|D_n)}{q(X_{n+1})} = \frac{\displaystyle\int dw\ \varphi(w)\ \exp(-(n+1)K_{n+1}(w))}{\displaystyle\int dw\ \varphi(w)\ \exp(-nK_n(w))}$$

shows

$$\frac{p(X_{n+1}|D_n)}{q(X_{n+1})} = \frac{Z^0_{n+1}}{Z^0_n}.$$

Therefore,

$$\log \frac{q(X_{n+1})}{p(X_{n+1}|D_n)} = F^0_{n+1} - F^0_n. \tag{1.13}$$

Based on eq.(1.13), the two equations in the theorem are respectively given by the expectations of X_{n+1} and D_{n+1}. □

This theorem shows that the Bayes generalization error with $\beta = 1$ is equal to the increase of the normalized stochastic complexity.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

1.3.2 Bayes and Gibbs estimations

Let $E_w[\,\cdot\,]$ be the expectation value using the *a posteriori* distribution $p(w|D_n)$. In Bayes estimation, the true distribution is estimated by the predictive distribution $E_w[p(x|w)]$. In the other method of statistical estimation, Gibbs estimation a parameter w is randomly chosen from $p(w|D_n)$, then the true distribution is estimated by $p(x|w)$. Gibbs estimation depends on a random choice of the parameter w. Hence, to study its generalization error, we need the expectation value over random choices of w. Bayes and Gibbs estimations respectively have generalization and training errors. The set of four errors is referred to as the Bayes quartet.

Definition 1.8 (Bayes quartet) For the generalized *a posteriori* distribution $p(w|D_n)$, the four errors are defined as follows.
(1) The Bayes generalization error,

$$B_{\rm g} = E_X\Big[\log \frac{q(X)}{E_w[p(X|w)]}\Big],$$

is the Kullback–Leibler distance from $q(x)$ to the predictive distribution.
(2) The Bayes training error,

$$B_{\rm t} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{E_w[p(X_i|w)]},$$

is the empirical Kullback–Leibler distance from $q(x)$ to the predictive distribution.
(3) The Gibbs generalization error,

$$G_{\rm g} = E_w\Big[E_X\Big[\log \frac{q(X)}{p(X|w)}\Big]\Big],$$

is the mean Kullback–Leibler distance from $q(x)$ to $p(x|w)$.
(4) The Gibbs training error,

$$G_{\rm t} = E_w\Big[\frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)}\Big],$$

is the mean empirical Kullback–Leibler distance from $q(x)$ to $p(x|w)$.

Remark 1.11 The Bayes *a posteriori* distribution $p(w|D_n)$ depends on the set of random samples, D_n. Hence the Bayes quartet is a set of random variables. The most important variable among them is the Bayes generalization error because it is used in practical applications; however, the other variables have important information about statistical estimation. In fact, we prove that there are mathematical relations among them.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Theorem 1.3 (Representation of Bayes quartet) *By using the log density ratio function* $f(x, w) = \log(q(x)/p(x|w))$*, the four errors are rewritten as*

$$B_{\rm g} = E_X\Big[-\log E_w[e^{-f(X,w)}]\Big],$$

$$B_{\rm t} = \frac{1}{n}\sum_{i=1}^{n} -\log E_w[e^{-f(X_i,w)}],$$

$$G_{\rm g} = E_w[K(w)],$$

$$G_{\rm t} = E_w[K_n(w)].$$

Proof of Theorem 1.3 The first and the second equations are derived from

$$\log\frac{q(X)}{E_w[p(X|w)]} = -\log E_w[e^{-f(X,w)}].$$

The third and the fourth equations are derived from the definitions of the Kullback–Leibler distance $K(w)$ and the empirical one $K_n(w)$. □

Remark 1.12 (Generalization errors and square error) Let $p(y|x, w)$ be a conditional probability density of $y \in \mathbb{R}^N$ for a given $x \in \mathbb{R}^M$ defined by

$$p(y|x, w) = \frac{1}{\sqrt{2\pi\sigma^2}^N}\exp\Big(-\frac{1}{2\sigma^2}|y - h(x, w)|^2\Big),$$

where $h(x, w)$ is a function from $\mathbb{R}^M \times \mathbb{R}^d$ to $\mathbb{R}^N$, $|\cdot|$ is the norm of $\mathbb{R}^N$, and $\sigma > 0$ is a constant. Let us compare generalization errors with the square error. If the true conditional distribution is $p(y|x, w_0)$, then the log density ratio function is

$$\begin{aligned} f(x, y, w) &= \frac{1}{2\sigma^2}\{|y - h(x, w)|^2 - |y - h(x, w_0)|^2\} \\ &= \frac{1}{2\sigma^2}\{2(y - h(x, w_0)) \cdot (h(x, w_0) - h(x, w)) \\ &\qquad + |h(x, w_0) - h(x, w)|^2\}. \end{aligned} \tag{1.14}$$

The Kullback–Leibler distance is

$$K(w) = \frac{1}{2\sigma^2}E_X[\,|h(X, w_0) - h(X, w)|^2].$$

The Gibbs generalization error is

$$G_{\rm g} = \frac{1}{2\sigma^2}E_X[E_w[\,|h(X, w_0) - h(X, w)|^2]].$$

The Bayes generalization error is

$$B_{\rm g} = E_X E_Y[-\log E_w[e^{-f}]], \tag{1.15}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $f = f(X, Y, w)$. On the other hand, the regression function of the estimated distribution $E_w[p(y|x, w)]$ is equal to

$$\int y\ E_w[p(y|x, w)]\, dy = E_w\Big[\int yp(y|x, w)dy\Big] = E_w[h(x, w)].$$

Let us define the square error of the estimated and true regression functions by

$$E_{\rm g} = \frac{1}{2\sigma^2} E_X[|E_w[h(x, w)] - h(X, w_0)|^2]. \tag{1.16}$$

In general, $B_{\rm g} \neq E_{\rm g}$. However, asymptotically, $B_{\rm g} \cong E_{\rm g}$. In fact, on a natural assumption, the *a posteriori* distribution $p(w|X_n)$ converges so that $f \to 0$, hence

$$B_{\rm g} = E_X E_Y\Big[-\log E_w\Big[1 - f + \frac{f^2}{2} + o(f^2)\Big]\Big]$$
$$= E_X E_Y\Big(E_w\Big[f - \frac{f^2}{2}\Big] + \tfrac{1}{2} E_w[f]^2 + o(f^2)\Big).$$

By $E_X E_Y E_w(f - f^2/2) = o(f^2)$ using eq.(1.14),

$$B_{\rm g} = \tfrac{1}{2} E_X E_Y[E_w[f(X, Y, w)]^2] + o(f^2)]],$$

hence $B_{\rm g} \cong E_{\rm g}$.

1.3.3 Maximum likelihood and *a posteriori*

Let $q(x)$, $p(x|w)$, and $\varphi(w)$ be the true distribution, a statistical model, and an *a priori* probability density function, respectively. The generalized log likelihood function is given by

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

where $\{a_n\}$ is a sequence of nonnegative real values. By using a log density ratio function

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

the generalized log likelihood function can be rewritten as

$$R_n(w) = R_n^0(w) + nS_n,$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $R_n^0(w)$ is given by

$$R_n^0(w) = \sum_{i=1}^{n} f(X_i, w) - a_n \log \varphi(w). \tag{1.17}$$

Note that, in singular statistical models, sometimes

$$\inf_w R_n(w) = -\infty,$$

which means that there is no parameter that minimizes $R_n(w)$. If a parameter $\hat{w}$ that minimizes $R_n(w)$ exists, then a statistical estimation method

$$D_n \mapsto p(x|\hat{w})$$

is defined. The generalization error $R_{\rm g}$ and the training error $R_{\rm t}$ of this method are respectively defined by

$$R_{\rm g} = \int q(x) \log \frac{q(x)}{p(x|\hat{w})} dx,$$
$$R_{\rm t} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|\hat{w})}.$$

By using $K(w)$ and $K_n(w)$ in equations (1.2) and (1.3) respectively, they can be rewritten as

$$R_{\rm g} = K(\hat{w}),$$
$$R_{\rm t} = K_n(\hat{w}).$$

Definition 1.9 (Maximum likelihood and maximum *a posteriori*)
(1) If $a_n = 0$ for arbitrary n, then $\hat{w}$ is called the maximum likelihood (or ML) estimator and the statistical estimation method is called the maximum likelihood (or ML) method.
(2) If $a_n = 1$ for arbitrary n, then $\hat{w}$ is called the maximum *a posteriori* estimator (or MAP) and the method is called the maximum *a posteriori* (or MAP) method.
(3) If a_n is an increasing function of n, then $\hat{w}$ is the generalized maximum *a posteriori* estimator and the method is called the generalized maximum *a posteriori* method.

Remark 1.13 (Formal relation between Bayes and ML)
(1) If $\beta \to \infty$, both Bayes and Gibbs estimations formally result in the maximum likelihood estimation.
(2) In regular statistical models in which the maximum likelihood estimator has asymptotic normality, the leading terms of the asymptotic generalization

errors of Bayes, ML, and MAP are equal to each other. However, in singular statistical models, they are quite different.

Example 1.7 (Divergence of MLE) Let $g(x|a,\sigma)$ be the normal distribution on $\mathbb{R}^1$,

$$g(x|a,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-a)^2}{2\sigma^2}\Big).$$

Let us study a normal mixture,

$$p(x|a,b,c,\sigma,\rho) = a\ g(x|b,\sigma) + (1-a)\ g(x|c,\rho),$$

where the set of parameters is

$$W = \{(a,b,c,\sigma,\rho)\ ;\ 0 \leq a \leq 1, |b|, |c| < \infty, \sigma, \rho > 0\}.$$

Then the likelihood function for a given D_n

$$L_n(a,b,c,\sigma,\rho) = \prod_{i=1}^{n} p(X_i|a,b,c,\sigma,\rho)$$

is an unbounded function, because

$$\lim_{\rho\to 0} L_n(a,b,X_1,\sigma,\rho) = \infty.$$

Therefore the normal mixture $p(x|a,b,c,\sigma,\rho)$ does not have a maximum likelihood estimator for arbitrary true distribution. To avoid this problem, we should restrict the parameter set or adopt the generalized maximum *a posteriori* method. In singular statistical models, the maximum likelihood estimator often diverges.

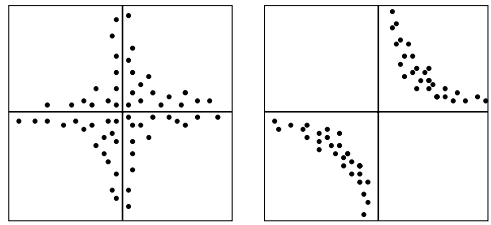
1.4 Four main formulas

In this section, we give an outline of singular learning theory. Because singular learning theory is quite different from regular statistical theory, the reader is advised to read this overview of the results of the book in advance. The equations and explanations in this section are intuitively described, because rigorous definitions and proofs are given in subsequent chapters.

1.4.1 Standard form of log likelihood ratio function

To evaluate how appropriate the statistical models $p(x|w)$ and $\varphi(w)$ are for a given data set $D_n = \{X_1, X_2, \ldots, X_n\}$, we have to study the case when the set

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press



Bayes *a posteriori* Maximum likelihood

Fig. 1.5. Maximum likelihood and Bayes *a posteriori*

of true parameters

$$\begin{aligned} W_0 &= \{w \in W \ ; \ q(x) = p(x|w) \ (\forall x)\} \\ &= \{w \in W; K(w) = 0\} \end{aligned}$$

consists of not one point but a union of several manifolds. If $K(w)$ is a polynomial, then W_0 is called an algebraic set; if $K(w)$ is an analytic function, then W_0 is called an analytic set. If W_0 is not one point, neither the Bayes *a posteriori* distribution nor the distribution of the maximum likelihood estimators converges to the normal distribution. For example, the left-hand side of Figure 1.5 shows a Bayes *a posteriori* distribution when the set of true parameters is $\{(a, b); ab = 0\}$. The right-hand side shows the probability distribution of the maximum likelihood estimator. We need a method to analyze such a singular distribution.

The basic term in statistical learning is the empirical Kullback–Leibler distance,

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

which is a function of $w \in W \subset \mathbb{R}^d$. For $w \in W \setminus W_0$, a random process

$$\psi_n(w) = \sum_{i=1}^{n} \frac{K(w) - f(X_i, w)}{\sqrt{n \ K(w)}}$$

is well-defined. The log likelihood ratio function is rewritten as

$$nK_n(w) = nK(w) - \sqrt{nK(w)} \ \psi_n(w).$$

This representation has two mathematical problems.

(1) (Geometrical problem). In a singular model, W_0 is not one point but a real analytic set hence the log likelihood ratio function cannot be treated locally

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

even if the number of training samples is sufficiently large. Moreover, since the set of true parameters contains complicated singularities, it is difficult to analyze its behavior even in each local neighborhood of W_0.

(2) (Probabilistic problem). When $n \to \infty$, under a natural condition, $\psi_n(w)$ converges in law to a Gaussian process $\psi(w)$ on the set $W \setminus W_0$. However, neither $\psi_n(w)$ nor $\psi(w)$ is well-defined on the set of true parameters W_0. Therefore it is difficult to analyze such a stochastic process near the set of true parameters.

In this book, we propose an algebraic geometrical transform that is powerful enough to overcome these two problems. For a real analytic function $K(w)$, the fundamental theorem in algebraic geometry ensures that there exists a real d-dimensional manifold $\mathcal{M}$ and a real analytic map

$$g : \mathcal{M} \ni u \mapsto w \in W$$

such that, for each coordinate $\mathcal{M}_\alpha$ of $\mathcal{M}$, $K(g(u))$ is a direct product,

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

where $k_1, k_2, \ldots, k_d$ are nonnegative integers. Morevoer, there exists a function $\phi(u) > 0$ and nonnegative integers $h_1, h_2, \ldots, h_d$ such that

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

where $|g'(u)|$ is Jacobian determinant of $w = g(u)$. Note that $k_1, k_2, \ldots, k_d$ and $h_1, h_2, \ldots, h_d$ depend on a local coordinate. By using the notation

$$u = (u_1, u_2, \ldots, u_d),$$
$$k = (k_1, k_2, \ldots, k_d),$$
$$h = (h_1, h_2, \ldots, h_d),$$

the function $K(g(u))$ and the *a priori* distribution $\varphi(g(u))|g'(u)|$ are respectively expressed as

$$K(g(u)) = u^{2k},$$
$$\varphi(g(u))|g'(u)| = \phi(u)|u^h|.$$

The theorem that ensures the existence of such a real analytic manifold $\mathcal{M}$ and a real analytic map $w = g(u)$ is called Hironaka's theorem or resolution of singularities. The function $w = g(u)$ is called a resolution map. In Chapters 2 and 3, we give a rigorous statement of the theorem and a method to find the set $(\mathcal{M}, g)$, respectively. Then by using

$$K(g(u)) = 0 \Longrightarrow f(x, g(u)) = 0,$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

we can prove that there exists a real analytic function $a(x, u)$ such that

$$f(x, g(u)) = a(x, u)\, u^k \quad (\forall x).$$

From the definition of the Kullback–Leibler distance,

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

It follows that

$$\int a(x, u)q(x)dx = u^k.$$

Moreover, by $f(x, g(u)) = \log(q(x)/p(x|g(u)))$,

$$K(g(u)) = \int (f(x, g(u)) + e^{-f(x,g(u))} - 1)q(x)dx.$$

It is easy to show

$$\lim_{t\to 0} \frac{t + e^{-t} - 1}{t^2} \to \frac{1}{2}.$$

Therefore, if $u^{2k} = 0$, then

$$\int a(x, u)^2 q(x)dx = \lim_{u^{2k}\to 0} \frac{2K(g(u))}{u^{2k}} = 2.$$

Here we can introduce a well-defined stochastic process on $\mathcal{M}$,

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

from which we obtain a representation,

$$nK_n(g(u)) = nu^{2k} - \sqrt{n}u^k \xi_n(u). \tag{1.18}$$

By definition, $\xi_n(u)$ satisfies

$$E[\xi_n(u)] = 0 \quad (\forall u \in \mathcal{M}),$$
$$E[\xi_n(u)\xi_n(v)] = E_X[a(X, u)a(X, v)] - u^k v^k \quad (\forall u, v \in \mathcal{M}).$$

If $K(g(u)) = K(g(v)) = 0$, then

$$E[\xi_n(u)\xi_n(v)] = E_X[a(X, u)a(X, v)],$$

and $E[\xi_n(u)^2] = 2$. By the central limit theorem, for each $u \in \mathcal{M}$, $\xi_n(u)$ converges in law to a Gaussian distribution with mean zero and variance 2. In Chapter 5, we prove the convergence in law $\xi_n \to \xi$ as a random variable on the space of bounded and continuous functions on $\mathcal{M}$. Then the Gaussian

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

process $\xi(u)$ is uniquely determined by its mean and covariance. Here we attain the first main formula.

Main Formula I (Standard form of log likelihood ratio function)
Under natural conditions, for an arbitrary singular statistical model, there exist a real analytic manifold $\mathcal{M}$ and a real analytic map $g : \mathcal{M} \to W$ such that the log likelihood ratio function is represented by

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u), \tag{1.19}$$

where $\xi_n(u)$ converges in law to the Gaussian process $\xi(u)$. Also $w = g(u)$ gives the relation

$$\varphi(g(u))|g'(u)| = \phi(u)|u^h|, \tag{1.20}$$

where $\phi(u) > 0$ is a positive real analytic function.

Remark 1.14 (1) Note that the log likelihood ratio function of any singular statistical model can be changed to the standard form by algebraic geometrical transform, which allows $|g'(u)| = 0$.
(2) The integration over the manifold $\mathcal{M}$ can be written as the finite sum of the integrations over local coordinates. There exists a set of functions $\{\sigma_\alpha(u)\}$ such that $\sigma_\alpha(u) \geq 0$, $\sum_\alpha \sigma_\alpha(u) = 1$, and the support of $\sigma_\alpha(u)$ is contained in $\mathcal{M}_\alpha$. By using a function $\phi^*(u) = \phi(u)\sigma_\alpha(u) \geq 0$, where dependence of α in ϕ^* is omitted, for an arbitrary integrable function $F(w)$,

$$\begin{aligned} \int_W F(w)\varphi(w)dw &= \int_{\mathcal{M}} F(g(u))\varphi(g(u))|g'(u)|du \\ &= \sum_\alpha \int F(g(u))\phi^*(u)|u^h|du. \end{aligned} \tag{1.21}$$

(3) In regular statistical models, the set of true parameters consists of one point, $W_0 = \{w_0\}$. By the transform $w = g_0(u) = w_0 + I(w_0)^{1/2}u$

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

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

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

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Here each $\xi_n(k)$ converges in law to the standard normal distribution. This property is called asymptotic normality. If a statistical model has asymptotic normality, Bayes generalization and training errors, MAP, and ML estimations are obtained by using the normal distribution. However, singular statistical models do not have asymptotic normality. The standard form of the log likelihood ratio function, eq.(1.19), is the universal base for singular statistical models.

1.4.2 Evidence of singular model

In singular learning theory, the zeta function of a statistical model plays an important role.

Definition 1.10 (Zeta function of a statistical model) For a given set (p, q, φ), where $p(x|w)$ is a statistical model, $q(x)$ is a true probability distribution, and $\varphi(w)$ is an *a priori* probability density function with compact support, the zeta function $\zeta(z)$ ($z \in \mathbb{C}$) of a statistical model is defined by

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

where $K(w)$ is the Kullback–Leibler distance from $q(x)$ to $p(x|w)$.

By the definition, the zeta function is holomorphic in $Re(z) > 0$. It can be rewritten by using resolution map $w = g(u)$,

$$\zeta(z) = \int_{\mathcal{M}} K(g(u))^z \, \varphi(g(u)) \, |g'(u)| \, du.$$

By using $K(g(u)) = u^{2k}$ and eq.(1.20),(1.21), in each local coordinate,

$$\zeta(z) = \sum_{\alpha} \int_{\mathcal{M}_\alpha} u^{2kz+h} \phi^*(u) du.$$

It is easy to show that

$$\int_0^b u_1^{2k_1 z + h_1} du_1 = \frac{b^{2k_1 z + h_1}}{2k_1 z + h_1 + 1}.$$

Therefore, the Taylor expansion of $\phi^*(u)$ around the origin in arbitrary order shows that $\zeta(z)$ ($\mathrm{Re}(z) > 0$) can be analytically continued to the meromorphic function on the entire complex plane $\mathbb{C}$, whose poles are all real, negative, and rational numbers. They are ordered from the larger to the smaller,

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

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

The largest pole $(-\lambda_1)$ is determined by

$$\lambda_1 = \min_{\alpha} \min_{1 \le j \le d} \Big(\frac{h_j + 1}{2k_j} \Big). \tag{1.22}$$

Let m_k be the order of the pole $(-\lambda_k)$. The order m_1 is the maximum number of the elements of the set $\{j\}$ that attain the minimum of eq.(1.22). Therefore, the zeta function has the Laurent expansion,

$$\zeta(z) = \zeta_0(z) + \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} \frac{c_{km}}{(z + \lambda_k)^{m_k}}, \tag{1.23}$$

where $\zeta_0(z)$ is a holomorphic function and c_{km} is a coefficient. Let the state density function of $t > 0$ be

$$\begin{aligned} v(t) &= \int \delta(t - K(w))\, \varphi(w)\, dw \\ &= \sum_{\alpha} \int \delta(t - u^{2k}) |u^h| \phi^*(u) du. \end{aligned}$$

The zeta function is equal to its Mellin transform,

$$\zeta(z) = \int_0^{\infty} v(t)\, t^z\, dt.$$

Conversely, $v(t)$ is uniquely determined as the inverse Mellin transform of $\zeta(z)$. The inverse Mellin transform of

$$F(z) = \frac{(m-1)!}{(z + \lambda)^m}$$

is equal to

$$f(t) = \begin{cases} t^{\lambda - 1} (-\log t)^{m-1} & (0 < t < 1) \\ 0 & \text{otherwise} \end{cases}.$$

By eq.(1.23), we obtain the asymptotic expansion of $v(t)$ for $t \to 0$,

$$v(t) = \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} c'_{km}\, t^{\lambda_k - 1}\, (-\log t)^{m-1}.$$

This expansion holds for arbitrary $\phi^*(u)$, and c'_{km} is a linear transform of $\phi^*(u)$, therefore there exists a set of Schwartz distributions $\{D_{km}(u)\}$ whose supports are contained in $\mathcal{M}_0 = g^{-1}(W_0)$ such that the asymptotic expansion

$$\delta(t - u^{2k}) u^h \phi^*(u) = \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} D_{km}(u)\, t^{\lambda_k - 1}\, (-\log t)^{m-1}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

holds for $t \to 0$. Let $Y_n(w)dw$ be a measure defined by

$$Y_n(w)dw \equiv \exp(-n\beta K_n(w))\ \varphi(w)\ dw,$$

then we have an asymptotic expansion,

$$\begin{aligned} Y_n(w)dw &= Y_n(g(u))\ |g'(u)|\ du \\ &= \sum_{\alpha} e^{-n\beta u^{2k}+\sqrt{n}\beta u^k \xi_n(u)}\ \phi^*(u)|u^h|du \\ &= \sum_{\alpha} \int_0^{\infty} dt\ \delta(t-u^{2k}) \\ &\quad \times \phi^*(u)|u^h|e^{-n\beta t+\sqrt{nt}\beta\ \xi_n(u)}\ du \\ &= \sum_{\alpha} \sum_{k=1}^{\infty} \sum_{r=0}^{m_k-1} D_{km}(u)du \\ &\quad \times \int_0^{\infty} \frac{dt}{n} \left(\frac{t}{n}\right)^{\lambda_k-1} \left(\log \frac{n}{t}\right)^r\ e^{-\beta t+\sqrt{t}\beta\ \xi_n(u)}. \end{aligned}$$

For simplicity we use the notation $\lambda = \lambda_1, m = m_1$ and

$$du^* = \sum_{\alpha^*} D_{1\,m_1}(u)du, \tag{1.24}$$

where $\sum_{\alpha^*}$ shows the sum of local coordinates that attain the minimum λ and the maximum m in eq.(1.22). Such local coordinates are called essential coordinates in this book. By using the convergence in law $\xi_n(u) \to \xi(u)$, the largest term of the asymptotic expansion of the *a posteriori* distribution is given by

$$Y_n(w)\ dw \cong \frac{(\log n)^{m-1}}{n^{\lambda}}\ du^* \int_0^{\infty} dt\ t^{\lambda-1}\ e^{-\beta t+\sqrt{t}\beta\ \xi(u)}. \tag{1.25}$$

The normalized evidence is

$$Z_n^0 = \int Y_n(w)\ dw.$$

It follows that

$$\begin{aligned} F_n^0 &= -\log Z_n^0 \\ &\cong \lambda\ \log n - (m-1)\log\log n + F^R(\xi), \end{aligned} \tag{1.26}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $F^R(\xi)$ is a random variable

$$F^R(\xi) = -\log\Big(\int du^* \int_0^\infty dt\ t^{\lambda-1}\ e^{-\beta t+\sqrt{t}\beta\ \xi(u)}\Big).$$

We obtain the second main result.

Main Formula II (Convergence of stochastic complexity)
Let $(-\lambda)$ and m be respectively the largest pole and its order of the zeta function

$$\zeta(z) = \int K(w)^z \varphi(w) dw$$

of a statistical model. The normalized stochastic complexity has the following asymptotic expansion,

$$F_n^0 = \lambda \log n - (m-1)\log\log n + F^R(\xi) + o_p(1),$$

where $F^R(\xi)$ is a random variable and $o_p(1)$ is a random variable which satisfies the convergence in probability $o_p(1) \to 0$. Therefore the stochastic complexity F_n has the asymptotic expansion

$$F_n = n\beta S_n + \lambda \log n - (m-1)\log\log n + F^R(\xi) + o_p(1),$$

where S_n is the empirical entropy defined by eq.(1.4).

Remark 1.15 If a model is regular then $K(w)$ is equivalent to $|w|^2$, hence $\lambda = d/2$ and $m = 1$ where d is the dimension of the parameter space. The asymptotic expansion of F_n with $\beta = 1$ in a regular statistical model is well known as the Bayes information criterion (BIC) or the minimum description length (MDL). Hence Main Formula II contains BIC and MDL as a special case. If a model is singular, then $\lambda \neq d/2$ in general. In Chapter 3, we give a method to calculate λ and m, and in Chapter 7, we show examples in several statistical models. The constant λ is an important birational invariant, which is equal to the real log canonical threshold if $\varphi(w) > 0$ at singularities. Therefore Main Formula II claims that the stochastic complexity is asymptotically determined by the algebraic geometrical birational invariant.

1.4.3 Bayes and Gibbs theory

In real-world problems, the true distribution is unknown in general. The third formula is useful because it holds independently of the true distribution $q(x)$.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

The expectation value of an arbitrary function $F(w)$ over the *a posteriori* distribution is defined by

$$E_w[F(w)] = \frac{\int F(w)Y_n(w)dw}{\int Y_n(w)dw},$$

where $Y_n(w) = \exp(-nK_n(w))\varphi(w)$. When the number of training samples goes to infinity, this distribution concentrates on the union of neighborhoods of $K(w) = 0$. In such neighborhoods, the renormalized *a posteriori* distribution $E_{u,t}[\ ]$ is defined for an arbitrary function $A(u, t)$,

$$E_{u,t}[A(u,t)] = \frac{\displaystyle\int du^* \int_0^\infty dt\ A(u,t)\, t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\,\xi(u)}}{\displaystyle\int du^* \int_0^\infty dt\ t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\,\xi(u)}},$$

where du^* is defined in eq.(1.24). Then eq.(1.25) shows the convergence in law

$$E_w[(\sqrt{n}\,f(x,w))^s] \to E_{u,t}[(\sqrt{t}\ a(x,u))^s]$$

for $s > 0$, where the relations of the paramaters are

$$\begin{aligned} w &= g(u), \\ t &= nK(w) = nu^{2k}, \\ f(x,w) &= a(x,u)u^k. \end{aligned}$$

Based on these properties, we can derive the asymptotic behavior of the Bayes quartet from Theorem 1.3. Firstly, Gibbs generalization error is

$$G_{\rm g} = E_w[K(w)] = \frac{1}{n}E_{u,t}[t].$$

Secondly, Gibbs training error is

$$G_{\rm t} = \frac{1}{n}\sum_{i=1}^{n} E_w[f(X_i,w)] = \frac{1}{n}E_{u,t}[\xi(u)t^{1/2}] + o_p\Big(\frac{1}{n}\Big),$$

where $o_p(1/n)$ is a random variable which satisfies the convergence in probability, $n\ o_p(1/n) \to 0$. Thirdly, the Bayes generalization error is

$$\begin{aligned} B_{\rm g} &= E_X[-\log E_w[1 - f(X,w) + \tfrac{1}{2}f(X,w)^2]] + o_p\Big(\frac{1}{n}\Big) \\ &= E_X[-\log(1 - E_w[f(X,w)] + \tfrac{1}{2}E_w[f(X,w)^2])] + o_p(1/n) \end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Then by using $-\log(1-\epsilon) = \epsilon + \epsilon^2/2 + o(\epsilon^2)$ and

$$
\begin{aligned}
E_w[f(X,w)] &= \frac{1}{\sqrt{n}} E_{u,t}[a(X,u)t^{1/2}], \\
E_X[E_w[f(X,w)]] &= E_w[K(w)] = \frac{1}{n} E_{u,t}[t], \\
E_X[E_w[f(X,w)^2]] &= E_X[E_{u,t}[a(X,u)^2 t]] = \frac{2}{n} E_{u,t}[t] + o_p(1/n),
\end{aligned}
$$

where we used $E_X[a(X,u)^2] = 2$, it follows that

$$
B_{\mathrm{g}} = \frac{1}{2n} E_X[E_{u,t}[a(X,u)t^{1/2}]^2] + o_p(1/n). \tag{1.27}
$$

And, lastly, the Bayes training error is

$$
\begin{aligned}
B_{\mathrm{t}} &= \frac{1}{n}\sum_{i=1}^{n}[-\log E_w[1 - f(X_i,w) + \tfrac{1}{2} f(X_i,w)^2]] + o_p\Big(\frac{1}{n}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}[-\log(1 - E_w[f(X_i,w)] + \tfrac{1}{2} E_w[f(X_i,w)^2])] + o_p\Big(\frac{1}{n}\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\big\{E_{u,t}[a(X_i,u)t^{1/2}] - \tfrac{1}{2} E_{u,t}[a(X_i,u)^2 t] \\
&\quad + \tfrac{1}{2} E_{u,t}[a(X_i,u)t^{1/2}]^2\big\} + o_p\Big(\frac{1}{n}\Big) \\
&= G_{\mathrm{t}} - G_{\mathrm{g}} + B_{\mathrm{g}} + o_p(1/n),
\end{aligned}
$$

where we used the law of large numbers

$$
\frac{1}{n}\sum_{i=1}^{n} a(X_i,u)a(X_i,v) = E_X[a(X,u)a(X,v)] + o_p(1)
$$

in the last equation. By using convergence in law $\xi_n(u) \to \xi(u)$, we prove the convergences in law of the Bayes quartet,

$$
\begin{aligned}
nB_{\mathrm{g}} &\to B_{\mathrm{g}}^*, \quad nB_{\mathrm{t}} \to B_{\mathrm{t}}^*, \\
nG_{\mathrm{g}} &\to G_{\mathrm{g}}^*, \quad nG_{\mathrm{t}} \to G_{\mathrm{t}}^*,
\end{aligned}
$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $B_{\rm g}^*$, $B_{\rm t}^*$, $G_{\rm g}^*$, $G_{\rm t}^*$ are random variables represented by the random process $\xi(u)$. Let us introduce the notation ($a \in \mathbb{R}$),

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

$$Z(\xi) = \int du^*\ S_\lambda(\xi(u)).$$

Then

$$S'_\lambda(a) = \beta \int_0^\infty dt\ t^{\lambda-1/2}\ e^{-\beta t + a\beta\sqrt{t}},$$

$$S''_\lambda(a) = \beta^2 \int_0^\infty dt\ t^{\lambda}\ e^{-\beta t + a\beta\sqrt{t}}.$$

Finally we obtain

$$E[B_{\rm g}^*] = \frac{1}{2\beta^2} E\left[E_X\left[\left(\frac{\int du^* a(X,u) S'_\lambda(\xi(u))}{Z(\xi)} \right)^2 \right] \right],$$

$$E[B_{\rm t}^*] = E[B_{\rm g}^*] + E[G_{\rm t}^*] - E[G_{\rm g}^*],$$

$$E[G_{\rm g}^*] = \frac{1}{\beta^2} E\left[\frac{\int du^* S''_\lambda(\xi(u))}{Z(\xi)} \right],$$

$$E[G_{\rm t}^*] = \frac{1}{\beta^2} E\left[\frac{\int du^* S''_\lambda(\xi(u))}{Z(\xi)} \right] - \frac{1}{\beta} E\left[\frac{\int du^*\ \xi(u) S'_\lambda(\xi(u))}{Z(\xi)} \right].$$

These equations show that the expectations of the Bayes quartet are represented by linear sums of three expectation values over the random process $\xi(u)$. On the other hand, $\xi(u)$ is a Gaussian process which is represented by

$$\xi(u) = \sum_{i=1}^\infty b_k(u) g_k,$$

where $\{g_k\}$ is a set of random variables that are independently subject to the standard normal distribution and $b_k(u) = E[\xi(u) g_k]$. By using the partial integration $E[g_k F(g_k)] = E[(\partial/\partial g_k) F(g_k)]$ for an arbitrary integrable function $F(\ )$, we can prove

$$E[B_{\rm g}^*] = \frac{1}{\beta^2} E\Big[\frac{\int du^*\ S''_\lambda(\xi(u))}{Z(\xi)} \Big] - \frac{1}{2\beta^2} E\Big[\frac{\int du^*\ \xi(u)\ S'_\lambda(\xi(u))}{Z(\xi)} \Big].$$

Therefore four errors are given by the linear sums of two expectations of $S'_\lambda(\xi(u))$ and $S''_\lambda(\xi(u))$. By eliminating two expectations from four equations, we obtain two equations which hold for the Bayes quartet.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Main Formula III (Equations of states in statistical estimation) There are two universal relations in Bayes quartet.

$$E[B_{\rm g}^*] - E[B_{\rm t}^*] = 2\beta(E[G_{\rm t}^*] - E[B_{\rm t}^*]), \tag{1.28}$$

$$E[G_{\rm g}^*] - E[G_{\rm t}^*] = 2\beta(E[G_{\rm t}^*] - E[B_{\rm t}^*]). \tag{1.29}$$

These equations hold for an arbitrary true distribution, an arbitrary statistical model, an arbitrary *a priori* distribution, and arbitrary singularities.

Remark 1.16 (1) Main Formula III holds in both regular and singular models. Although the four errors themselves strongly depend on $q(x)$, $p(x|w)$, and $\varphi(w)$, these two equations do not. By this formula, we can estimate the Bayes and Gibbs generalization errors from the Bayes and Gibbs training errors without any knowledge of the true distributions. The constant

$$\nu(\beta) = \beta(E[G_{\rm t}^*] - E[B_{\rm t}^*]) \tag{1.30}$$

is the important birational invariant called a singular fluctuation. Then Main Formula III claims that

$$E[B_{\rm g}^*] = E[B_{\rm t}^*] + 2\nu(\beta), \tag{1.31}$$

$$E[G_{\rm g}^*] = E[G_{\rm t}^*] + 2\nu(\beta). \tag{1.32}$$

We can estimate $\nu(\beta)$ from samples. In fact, by defining two random variables,

$$V_0 = \sum_{i=1}^{n} (\log E_w[p(X_i|w)] - E_w[\log p(X_i|w)]),$$

$$V = \sum_{i=1}^{n} (E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2),$$

we have $V_0 = nG_{\rm t} - nB_{\rm t}$ and $E[V/2]$ is asymptotically equal to $E[nG_{\rm t}] - E[nB_{\rm t}]$,

$$\nu(\beta) = \beta E[V_0] + o(1) = (\beta/2)E[V] + o(1).$$

(2) If a model is regular then, for any $\beta > 0$,

$$\nu(\beta) = d/2, \tag{1.33}$$

where d is the dimension of the parameter space. If a model is regular, both Bayes and Gibbs estimation converge to the maximum likelihood estimation, when $\beta \to \infty$. Then two equations of states result in one equation,

$$E[nR_{\rm g}] = E[nR_{\rm t}] + d, \tag{1.34}$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $R_{\rm g}$ and $R_{\rm t}$ are the generalization and training errors of the maximum likelihood estimator. The equation (1.34) is well known as the Akaike information criterion (AIC) of a regular statistical model, hence Main Formula III contains AIC as a very special case. In singular learning machines, eq.(1.33) does not hold in general, hence AIC cannot be applied. Moreover, Main Formula III holds even if the true distribution is not contained in the model [120].

1.4.4 ML and MAP theory

The last formula concerns the maximum likelihood or *a posteriori* method. Let W be a compact set, and $f(x, w)$ and $\varphi(w)$ be respectively analytic and C^2-class functions of $w \in W$. Then there exists a parameter $\hat{w} \in W$ that minimizes the generalized log likelihood ratio function,

$$R_n^0(w) = \sum_{i=1}^{n} f(X_i, w) - a_n \log \varphi(w),$$

where a_n is a nondecreasing sequence. Note that, if W is not compact, the parameter that minimizes $R_n^0(w)$ does not exist in general. By applying the standard form of the log likelihood ratio function and a simple notation $\sigma(u) = -\log \varphi(g(u))$, in each coordinate, the function $R_n^0(g(u))$ is represented by

$$\frac{1}{n} R_n^0(g(u)) = u^{2k} - \frac{1}{\sqrt{n}}\, u^k\, \xi_n(u) + \frac{a_n}{n}\sigma(u),$$

where $\xi_n(u) \to \xi(u)$ in law. For an arbitrary u, a new parameterization (t, v) is defined by

$$\begin{aligned} t &= u^k, \\ v &= \mathrm{Proj}(u), \end{aligned}$$

where the function Proj( ) maps u to v on the set $\{v; v^{2k} = 0\}$ along the ordinary differential equation $u(T)$ for $T \geq 0$,

$$\frac{d}{dT} u(T) = -\nabla (u(T)^{2k}). \tag{1.35}$$

Here $v = \mathrm{Proj}(u)$ is determined by $v = u(T = \infty)$ for the initial condition $u = u(T = 0)$. More precisely, see Chapter 6 and Figure 6.3. In each local coordinate,

$$\frac{1}{n} R_n^0(g(t, v)) = t^2 - \frac{1}{\sqrt{n}}\, t\, \xi_n(t, v) + \frac{a_n}{n}\sigma(t, v).$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Then we can prove that, for arbitrary C^1-class function $f(u)$ on a compact set, there exist constants $C, \delta > 0$ such that

$$|f(t, v) - f(0, v)| \leq t^{\delta} \|\nabla f\|,$$

where

$$\|\nabla f\| \equiv \sup_j \sup_u \left| \frac{\partial f}{\partial u_j} \right|.$$

Let $\hat{t}$ be the parameter that minimizes $R_n^0(t, v)$, then $\hat{t}$ should be in proportion to $1/\sqrt{n}$, hence

$$\frac{1}{n} R_n^0(g(\hat{t}, v)) = \hat{t}^2 - \frac{1}{\sqrt{n}} \hat{t}\, \xi(0, v) + \frac{a_n}{n} \sigma(0, v) + o_p\left(\frac{1}{n}\right).$$

We can prove that $o_p(1/n)$ does not affect the main terms. Let $\hat{v}$ be the parameter that minimizes $R_n^0(g(\hat{t}, v))$. Then

$$\hat{t} = \frac{1}{2\sqrt{n}} \max_{\alpha} \max\{0, \xi(0, \hat{v})\},$$

where α shows the local coordinate. If $a_n \equiv 0$, then $\hat{v}$ is determined by minimizing

$$-\max_{\alpha} \max\{0, \xi(\hat{v})\}^2.$$

Hence the generalization and training errors are given by

$$R_{\mathrm{g}} = \frac{1}{4n} \left(\max_{u \in \mathcal{M}_0} \{0, \xi(u)\}^2 \right),$$

$$R_{\mathrm{t}} = -\frac{1}{4n} \left(\max_{u \in \mathcal{M}_0} \{0, \xi(u)\}^2 \right),$$

where $\mathcal{M}_0 = g^{-1}(W_0)$ is the set of true parameters. The symmetry of generalization and training errors holds if $a_n / n^p \to \infty$ for arbitrary $p > 0$. Therefore,

$$E[n R_{\mathrm{g}}] = -E[n R_{\mathrm{t}}] + o(1).$$

For the other sequence a_n, the same result is obtained.

Main Formula IV (Symmetry of generalization and training errors)
If the maximum likelihood or generalized maximum *a posteriori* method is applied, the symmetry of generalization and training errors holds,

$$\lim_{n \to \infty} E[n R_{\mathrm{g}}] = -\lim_{n \to \infty} E[n R_{\mathrm{t}}].$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Remark 1.17 (1) In regular statistical models, $E[nR_g] = d/2$ where d is the dimension of the parameter space. In singular statistical models $E[nR_g] >> d/2$ in general, because it is the mean of the maximum value of a Gaussian process. If the parameter space is not compact, then the maximum likelihood estimator sometimes does not exist. Even if it exists, it often diverges for $n \to \infty$, which means that

$$E[nR_t] \to -\infty.$$

In such a case, the behavior of the generalization error is still unknown. It is expected that the symmetry still holds, in which case

$$E[nR_g] \to +\infty.$$

Hence the maximum likelihood method is not appropriate for singular statistical models. Even if the set of parameters is compact, it is still difficult to estimate the generalization error from the training error without knowledge of the true distribution. From a statistical point of view, the maximum likelihood estimator is asymptotically the sufficient statistic in regular models. However, it is not in singular models, because the likelihood function does not converge to the normal distribution.

(2) In singular statistical models, two limiting procedures $n \to \infty$ and $\beta \to \infty$ are not commutative in general. In other words,

$$\lim_{\beta\to\infty}\lim_{n\to\infty} E[nB_g] \neq \lim_{n\to\infty} E[nR_g].$$

In fact, the Bayes generalization error is determined by the sum of the essential local coordinates $\sum_{\alpha^*}$, whereas the maximum likelihood generalization error is determined by the set of all coordinates $\sum_{\alpha}$.

1.5 Overview of this book

The main purpose of this book is to establish the mathematical foundation on which the four main formulas are proved.

In Chapter 2, we introduce singularity theory and explain the resolution theorem which claims that, for an arbitrary analytic function $K(w)$, there exist a manifold and an analytic function $w = g(u)$ such that $K(g(u)) = u^{2k}$.

In Chapter 3, elemental algebraic geometry is explained. The relation between algebra and geometry, Hilbert's basis theorem, projective space, and blow-ups are defined and illustrated. We show how to find the resolution map using recursive blow-ups for a given statistical model.

In Chapter 4, the mathematical relation between the zeta function and the singular integral is clarified. We need Schwartz distribution theory to connect these two concepts. Several inequalities which are used in the following sections are proved.

In Chapter 5, we study the convergence in law of the empirical process to a Gaussian process, $\xi_n(u) \to \xi(u)$. This is the central limit theorem on the functional space. Also we introduce the partial integral on the function space.

Based on mathematical foundations in chapters 2, 3, 4, and 5, the four main formulas are rigorously proved in Chapter 6. These are generalizations of the conventional statistical theory of regular models to singular models. We find two birational invariants, the maximum pole of the zeta function and the singular fluctuation, which determine the statistical learning process.

Chapters 7 and 8 are devoted to applications of this book to statistics and information science.

1.6 Probability theory

In this section, fundamental points of probability theory are summarized. Readers who are familiar with probability theory can skip this section.

Definition 1.11 (Metric space) Let Ω be a set. A function D

$$D : \Omega \times \Omega \ni (x, y) \mapsto D(x, y) \in \mathbb{R}$$

is called a metric if it satisfies the following three conditions.
(1) For arbitrary $x, y \in \Omega$, $D(x, y) = D(y, x) \geq 0$.
(2) $D(x, y) = 0$ if and only if $x = y$.
(3) For arbitrary $x, y, z \in \Omega$, $D(x, y) + D(y, z) \geq D(x, z)$.
A set Ω with a metric is called a metric space. The set of open neighborhoods of a point $x \in \Omega$ is defined by $\{U_\epsilon(x); \epsilon > 0\}$ where

$$U_\epsilon(x) = \{y \in \Omega \ ; \ D(x, y) < \epsilon\}.$$

The topology of the metric space is determined by all open neighborhoods. A metric space Ω is called separable if there exists a countable and dense subset. A set $\{x_n; n = 1, 2, 3, \ldots\}$ is said to be a Cauchy sequence if, for arbitrary $\delta > 0$, there exists M such that

$$m, n > M \Longrightarrow D(x_m, x_n) < \delta.$$

If any Cauchy sequence in a metric space Ω converges in Ω, then Ω is called a complete metric space. A complete and separable metric space is called a Polish space.

Example 1.8 In this book, we need the following metric spaces.
(1) The finite-dimensional real Euclidean space $\mathbb{R}^d$ is a metric space with the metric

$$D(x, y) = |x - y| \equiv \Big(\sum_{i=1}^{d}(x_i - y_i)^2\Big)^{1/2},$$

where $x = (x_i)$, $y = (y_i)$, and $|\cdot|$ is a norm of $\mathbb{R}^d$. The real Euclidean space $\mathbb{R}^d$ is a complete and separable metric space.
(2) A subset of $\mathbb{R}^d$ is a metric space with the same metric. Sometimes a finite or countable subset in $\mathbb{R}^d$ is studied.
(3) Let K be a compact subset in $\mathbb{R}^d$. The set of all continuous function from K to $\mathbb{R}^{d'}$

$$\Omega = \{f\ ;\ f : K \to \mathbb{R}^{d'}\}$$

is a metric space with the metric

$$D(f, g) = \|f - g\| \equiv \max_{x\in K} |f(x) - g(x)|,$$

where $|\cdot|$ is the norm of $\mathbb{R}^{d'}$. By the compactness of K in $\mathbb{R}^d$, it is proved that Ω is a complete and separable metric space.

Definition 1.12 (Probability space) Let Ω be a metric space. A set $\mathcal{B}$ composed of subsets contained in Ω is called a sigma algebra or a completely additive set if it satisfies the following conditions. ($\mathcal{B}$ contains the empty set.)
(1) If $A_1, A_2 \in \mathcal{B}$ then $A_1 \cap A_2 \in \mathcal{B}$.
(2) If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$ (A^c is the complementary set of A).
(3) If $A_1, A_2, A_3 \ldots, \in \mathcal{B}$ then the countable union $\cup_{k=1}^{\infty} A_k \in \mathcal{B}$.
The smallest sigma algebra that contains all open sets of Ω is said to be a Borel field. A pair of a metric space and a sigma algebra $(\Omega, \mathcal{B})$ is called a measurable space. A function P,

$$P : \mathcal{B} \ni A \mapsto 0 \le P(A) \le 1,$$

is called a probability measure if it satisfies
(1) $P(\Omega) = 1$.
(2) For $\{B_k\}$ which satisfies $B_k \cap B_{k'} = \emptyset$ $(k \neq k')$, $P(\cup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} P(B_k)$.

A triple of a metric space, a sigma algebra, and a probability measure $(\Omega, \mathcal{B}, P)$ is called a probability space.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

Remark 1.18 Let $(\mathbb{R}^N, \mathcal{B}, P)$ be a probability space, where $\mathbb{R}^N$ is the N-dimensional real Euclidean space, $\mathcal{B}$ the Borel field, and P a probability distribution. If P is defined by a function $p(x) \geq 0$,

$$P(A) = \int_A p(x)dx \quad (A \in \mathcal{B}),$$

then $p(x)$ is called a probability density function.

Definition 1.13 (Random variable) Let $(\Omega, \mathcal{B}, P)$ be a probability space and $(\Omega_1, \mathcal{B}_1)$ a measurable space. A function

$$X : \Omega \ni \omega \mapsto X(\omega) \in \Omega_1$$

is said to be measurable if $X^{-1}(B_1) \in \mathcal{B}$ for arbitrary $B_1 \in \mathcal{B}_1$. A measurable function X on a probability space is called a random variable. Sometimes X is said to be an Ω_1-valued random variable. By the definition

$$\mu(B_1) = P(X^{-1}(B_1)), \tag{1.36}$$

μ is a probability measure on $(\Omega_1, \mathcal{B}_1)$, hence $(\Omega_1, \mathcal{B}_1, \mu)$ is a probability space. The probability measure μ is called a probability distribution of the random variable X. Then X is said to be subject to μ. Note that μ is the probability distribution on the image space of a function of X. Equation (1.36) can be rewritten as

$$\int_{B_1} \mu(dx) = \int_{X^{-1}(B_1)} P(da).$$

Remark 1.19 (1) In probability theory, the simplified notation

$$P(f(X) > 0) \equiv P(\{\omega \in \Omega; f(X(\omega)) > 0\})$$

is often used. Then by definition, $P(f(X) > 0) = \mu(\{x \in \Omega_1; f(x) > 0\})$.
(2) The probability measure μ to which a random variable X is subject is often denoted by P_X. The map $X \mapsto P_X$ is not one-to-one in general. For example, on a probability space $(\Omega, 2^\Omega, P)$ where $\Omega = \{1, 2, 3, 4\}$ and $P(\{i\}) = 1/4$ $(i = 0, 1, 2, 3)$, two different random variables

$$X(i) = \begin{cases} 0 & (i = 0, 1) \\ 1 & (i = 2, 3) \end{cases}$$

$$Y(i) = \begin{cases} 0 & (i = 0, 2) \\ 1 & (i = 1, 3) \end{cases}$$

are subject to the same probability distribution. Therefore, in general, even if X and Y are subject to the same probability distribution, we cannot predict the realization of Y from a realization of X.

(3) In descriptions of definitions and theorems, sometimes we need only the information of the image space of a random variable X and the probability distribution P_X. In other words, there are some definitions and theorems in which the explicit statement of the probability space $(\Omega, \mathcal{B}, P)$ is not needed. In such cases, the explicit definition of the probability space is omitted, resulting in a statement such as "for Ω_1-valued random variable X which is subject to a probability distribution P_X satisfies the following equality . . ."

Definition 1.14 (Expectation) Let X be a random variable from the probability space $(\Omega, \mathcal{B}, P)$ to $(\Omega_1, \mathcal{B}_1)$ which is subject to the probability distribution P_X. If the integration

$$E[X] = \int X(\omega) P(d\omega) = \int x \; P_X(dx)$$

is well defined and finite in Ω_1, $E[X] \in \Omega_1$ is called the expectation or the mean of X. Let S be a subset of Ω_1. The partial expectation is defined by

$$E[X]_S = \int_{X(\omega) \in S} X(\omega) P(d\omega) = \int_S x \; P_X(dx).$$

Remark 1.20 These are fundamental remarks.
(1) Let $(\Omega_1, \mathcal{B}_1)$ and X be same as Definition 1.14 and $(\Omega_2, \mathcal{B}_2)$ be a measurable space. If $f : \Omega_1 \to \Omega_2$ is a measurable function then $f(X)$ is a random variable on $(\Omega, \mathcal{B}, P)$. The expectation of $f(X)$ is equal to

$$E[f(X)] = \int f(X(\omega)) P(d\omega) = \int f(x) \; P_X(dx).$$

This expectation is often denoted by $E_X[f(X)]$.
(2) Two random variables which have the same probability distribution have the same expectation value. Hence if X and Y have the same probability distribution, we can predict $E[Y]$ based on the information of $E[X]$.
(3) In statistical learning theory, it is important to predict the expectation value of the generalization error from the training error.
(4) If $E[|X|] = C$ then, for arbitrary $M > 0$,

$$\begin{aligned} C &= E[|X|] \geq E[|X|]_{\{|X|>M\}} \\ &\geq M E[1]_{\{|X|>M\}} = M P(|X| > M). \end{aligned}$$

Hence

$$P(|X| > M) \leq \frac{C}{M},$$

which is well known as Chebyshev's inequality. The same derivation is often effective in probability theory.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

(5) The following conditions are equivalent.

$$E[|X|] < \infty \Longleftrightarrow \lim_{M\to\infty} E[|X|]_{\{|X|\geq M\}} = 0.$$

(6) If there exist constants $\delta > 0$ and $M_0 > 0$ such that for an arbitrary $M > M_0$

$$P(|X| \geq M) \leq \frac{1}{M^{1+\delta}},$$

then $E[|X|] < \infty$.

Definition 1.15 (Convergence of random variables) Let $\{X_n\}$ and X be a sequence of random variables and a random variable on a probability space $(\Omega, \mathcal{B}, P)$, respectively.
(1) It is said that X_n converges to X almost surely (almost everywhere), if

$$P\Big(\{\omega \in \Omega\ ;\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\Big) = 1.$$

(2) It is said that X_n converges to X in the mean of order $p > 0$, if

$$\lim_{n\to\infty} E[(X_n - X)^p] = 0.$$

(3) It is said that X_n converges to X in probability, if

$$\lim_{n\to\infty} P(D(X_n, X) > \epsilon) = 0$$

for arbitrary $\epsilon > 0$, where $D(\cdot, \cdot)$ is the metric of the image space of X.

Remark 1.21 There are well-known properties of random variables.
(1) If X_n converges to X almost surely or in the mean of order $p > 0$, then it does in probability.
(2) If X_n converges to X in probability, then it does in law. For the definition of convergence in law, see chapter 5.
(3) In general, "almost surely" is neither sufficient nor necessary condition of "in the means of order $p > 0$."

Remark 1.22 (Limit theorem) From the viewpoint of probability theory, in this book we obtain the limit theorem of the random variables,

$$F_n = -\log \int p(X_1|w)^\beta p(X_2|w)^\beta \cdots p(X_n|w)^\beta \varphi(w) dw,$$

and

$$B_{\mathrm{g}} = E_X \left[-\log \frac{\int p(X|w) p(X_1|w)^\beta \cdots p(X_n|w)^\beta \varphi(w) dw}{\int q(X) p(X_1|w)^\beta \cdots p(X_n|w)^\beta \varphi(w) dw} \right],$$

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press

where $X_1, \ldots, X_n$ are independently subject to the same distribution as X. As the central limit theorem is characterized by the mean and the variance of the random variables, the statistical learning theory is characterized by the largest pole of the zeta function and the singular fluctuation. The large deviation theory indicates that $F_n/n \to pS$, where S is the entropy of X. Main Formulas II and III show more precise results than the large deviation theory.

https://doi.org/10.1017/CBO9780511800474.002 Published online by Cambridge University Press