6
Singular learning theory

In this chapter we prove the four main formulas in singular learning theory. The formulas which clarify the singular learning process are not only mathematically beautiful but also statistically useful.

Firstly, we introduce the standard form of the log likelihood ratio function. A new foundation is established on which singular learning theory is constructed without the positive definite Fisher information matrix. By using resolution of singularities, there exists a map $w = g(u)$ such that the log likelihood ratio function of any statistical model is represented by

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

where u^k is a normal crossing function and $\xi_n(u)$ is an empirical process which converges to a Gaussian process in law.

Secondly, we prove that, under a natural condition, the stochastic complexity of an arbitrary statistical model can be asymptotically expanded as

$$\begin{aligned} F_n &= -\log \int \prod_{i=1}^{n} p(X_i|w)^{\beta}\ \varphi(w)dw \\ &= n\beta S_n + \lambda \log n - (m-1)\log\log n + F_n^R, \end{aligned}$$

where S_n is the empirical entropy of the true distribution, F_n^R is a random variable which converges to a random variable in law, and $(-\lambda)$ and m are respectively equal to the largest pole and its order of the zeta function of a statistical model,

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

In regular statistical models $\lambda = d/2$ and $m = 1$ where d is the dimension of the parameter space, whereas in singular learning machines $\lambda \neq d/2$ and $m \geq 1$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

in general. The constant λ, the learning coefficient, is equal to the real log canonical threshold of the set of true parameters if $\varphi(w) > 0$ on singularities.

Thirdly, we prove that the means of Bayes generalization error B_{g}, Bayes training error B_{t}, Gibbs generalization error G_{g}, and Gibbs training error G_{t} are respectively given by

$$E[B_{\mathrm{g}}] = \left(\frac{\lambda + \nu\beta - \nu}{\beta}\right)\frac{1}{n} + o\left(\frac{1}{n}\right), \tag{6.1}$$

$$E[B_{\mathrm{t}}] = \left(\frac{\lambda - \nu\beta - \nu}{\beta}\right)\frac{1}{n} + o\left(\frac{1}{n}\right), \tag{6.2}$$

$$E[G_{\mathrm{g}}] = \left(\frac{\lambda + \nu\beta}{\beta}\right)\frac{1}{n} + o\left(\frac{1}{n}\right), \tag{6.3}$$

$$E[G_{\mathrm{t}}] = \left(\frac{\lambda - \nu\beta}{\beta}\right)\frac{1}{n} + o\left(\frac{1}{n}\right), \tag{6.4}$$

where n is the number of random samples, $\beta > 0$ is the inverse temperature of the *a posteriori* distribution, and $\nu = \nu(\beta) > 0$ is the singular fluctuation. From these equations, the equations of states in statistical estimation are derived:

$$E[B_{\mathrm{g}}] - E[B_{\mathrm{t}}] = 2\beta(E[G_{\mathrm{t}}] - E[B_{\mathrm{t}}]) + o\left(\frac{1}{n}\right),$$

$$E[G_{\mathrm{g}}] - E[G_{\mathrm{t}}] = 2\beta(E[G_{\mathrm{t}}] - E[B_{\mathrm{t}}]) + o\left(\frac{1}{n}\right).$$

Although both λ and $\nu(\beta)$ strongly depend on the set of a true distribution, a statistical model, and an *a priori* distribution, the equations of states always hold independent of them. Therefore, based on the equations of states, we can predict Bayes and Gibbs generalization errors from Bayes and Gibbs training errors without any knowledge of the true distribution. Based on random samples, we can evaluate how appropriate a model and an *a priori* distribution are.

And, lastly, we show the symmetry of the generalization and training errors in the maximum likelihood or the maximum *a posteriori* estimation,

$$E[R_{\mathrm{g}}] = \frac{C}{n} + o\left(\frac{1}{n}\right),$$

$$E[R_{\mathrm{t}}] = -\frac{C}{n} + o\left(\frac{1}{n}\right),$$

where R_{g} and R_{t} are the generalization error and the training error of the maximum likelihood or *a posteriori* estimator respectively. In singular statistical models, the constant $C > 0$ is given by the maximum values of a Gaussian

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

process on the set of true parameters, hence C is far larger than $d/2$ in general. In order to make C small, we need a strong improvement by some penalty term, which is not appropriate in singular statistical estimation.

From a historical point of view, the concepts and proofs of this chapter were found by the author of this book.

6.1 Standard form of likelihood ratio function

To establish singular learning theory we need some fundamental conditions.

Definition 6.1 (Fundamental condition (I)) Let $q(x)$ and $p(x|w)$ be probability density functions on $\mathbb{R}^N$ which have the same support. Here $p(x|w)$ is a parametric probability density function for a given parameter $w \in W \subset \mathbb{R}^d$. The set of all parameters W is a compact set in $\mathbb{R}^d$. The Kullback–Leibler distance is defined by

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

We assume

$$W_0 = \{w \in W;\ K(w) = 0\}$$

is not the empty set. It is said that $q(x)$ and $p(x|w)$ satisfy the fundamental condition (I) with index s ($s \geq 2$) if the following conditions are satisfied.
(1) For $f(x, w) = \log(q(x)/p(x|w))$, there exists an open set $W^{(C)} \subset \mathbb{C}^d$ such that:
(1-a) $W \subset W^{(C)}$,
(1-b) $W^{(C)} \ni w \mapsto f(\cdot, w)$ is an $L^s(q)$-valued complex analytic function,
(1-c) $M(x) \equiv \sup_{w \in W^{(C)}} |f(x, w)|$ is contained in $L^s(q)$.
(2) There exists $\epsilon > 0$ such that, for

$$Q(x) \equiv \sup_{K(w) \leq \epsilon} p(x|w),$$

the following integral is finite,

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

Remark 6.1 These are remarks about the fundamental condition (I).
(1) By definition, there exists a real open set $W^{(R)} \subset \mathbb{R}^d$ such that

$$W \subset W^{(R)} \subset W^{(C)}.$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

The log density ratio function $f(x, w)$ is an $L^s(q)$-valued real analytic function on $W^{(R)}$ and an $L^s(q)$-valued complex analytic function on $W^{(C)}$. For a sufficiently small constant $\epsilon > 0$, we define

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

Based on the resolution theorem in Chapters 2 and 3, there exist open sets $W_\epsilon^{(R)} \subset W^{(R)}$, $W_\epsilon^{(C)} \subset W^{(C)}$ and subsets of manifolds $\mathcal{M}$, $\mathcal{M}^{(R)}$, and $\mathcal{M}^{(C)}$ such that

$$W_\epsilon \subset W_\epsilon^{(R)} \subset W_\epsilon^{(C)},$$
$$\mathcal{M} \subset \mathcal{M}^{(R)} \subset \mathcal{M}^{(C)},$$

and that

$$\mathcal{M} \equiv g^{-1}(W_\epsilon),$$
$$\mathcal{M}^{(R)} \equiv g^{-1}\big(W_\epsilon^{(R)}\big),$$
$$\mathcal{M}^{(C)} \equiv g^{-1}\big(W_\epsilon^{(C)}\big),$$

where $w = g(u)$ is the resolution map and its complexification. In the following, we use this notation.

(2) In the fundamental condition (I), we mainly study the case in which

$$W_0 \equiv \{w \in W; K(w) = 0\}$$

is not one point but a set with singularities. In other words, the Fisher information matrix of $p(x|w)$ is degenerate on W_0 in general. However, this condition contains the regular statistical model as a special case.

(3) From conditions (1-b) and (1-c), $f(x, w)$ is represented by the absolutely convergent power series in the neighborhood of arbitrary $w_0 \in W$:

$$f(x, w) = \sum_\alpha a_\alpha(x)(w - w_0)^\alpha.$$

The function $a_\alpha(x) \in L^s(q)$ is bounded by

$$|a_\alpha(x)| \le \frac{M(x)}{R^\alpha},$$

where R is the associated convergence radii.

(4) If $M(x)$ satisfies

$$\int M(x)^2 e^{M(x)} q(x) dx < \infty,$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

then condition (2) is satisfied, because

$$\begin{aligned}\int M(x)^2 Q(x)dx &= \int M(x)^2 \sup_{K(w)\leq\epsilon} p(x|w)dx \\ &= \int M(x)^2 \sup_{K(w)\leq\epsilon} e^{-f(x,w)}q(x)dx \\ &\leq \int M(x)^2 e^{M(x)} q(x)dx.\end{aligned}$$

(5) Let $w = g(u)$ be a real proper analytic function which makes $K(g(u))$ normal crossing. If $q(x)$ and $p(x|w)$ satisfy the fundamental condition (I) with index s, then $q(x)$ and $p(x|g(u))$ also satisfy the same condition, where $\mathbb{R}^d$ and $\mathbb{C}^d$ are replaced by real and complex manifolds respectively.

(6) The condition that W is compact is necessary because, even if the log density ratio function is a real analytic function of the parameter, $|w| = \infty$ is an analytic singularity in general. For this reason, if W is not compact and W_0 contains $|w| = \infty$, the maximum likelihood estimator does not exist in general. For example, if $x = (x_1, x_2)$, $w = (a, b)$, and $p(x_2|x_1, w) \propto \exp(-(x_2 - a\sin(bx_1))^2/2)$, then the maximum likelihood estimator never exists. On the other hand, if $|w| = \infty$ is not a singularity, $\mathbb{R}^d \cup \{|w| = \infty\}$ can be understood as a compact set and the same theorems as this chapter hold. From the mathematical point of view, it is still difficult to construct singular learning theory of a general statistical model in the case when W_0 contains analytic singularities. From the viewpoint of practical applications, we can select W as a sufficiently large compact set.

(7) If a model satisfies the fundamental condition (I) with index $s + t$, $(s \geq 2, t > 0)$, then it automatically satisfies the fundamental condition (I) with index s. The condition with index $s = 6$ is needed to ensure the existence of the asymptotic expansion of the Bayes generalization error in the proof. (See the proof of Theorem 6.8.)

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

(9) The same results as this chapter can be proven based on the weaker conditions. However, to describe clearly the mathematical structure of singular

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

learning theory, we adopt the condition (I). For the case of the weaker condition, see Section 7.8.

Theorem 6.1 *Assume that $q(x)$ and $p(x|w)$ satisfy the fundamental condition (I) with index $s = 2$. There exist a constant $\epsilon > 0$, a real analytic manifold $\mathcal{M}^{(R)}$, and a real analytic function $g : \mathcal{M}^{(R)} \to W_\epsilon^{(R)}$ such that, in every local coordinate U of $\mathcal{M}^{(R)}$,*

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

where $k_1, k_2, \ldots, k_d$ are nonnegative integers. Moreover, there exists an $L^s(q)$-valued real analytic function $a(x, u)$ such that

$$f(x, g(u)) = a(x, u)\, u^k \quad (u \in U), \tag{6.5}$$

and

$$\int a(x, u) q(x) dx = u^k \quad (u \in U). \tag{6.6}$$

Remark 6.2 (1) In this chapter, $\epsilon > 0$ is taken so that Theorem 6.1 holds. The set of parameters W is represented as the union of two subsets. The former is W_ϵ which includes $\{w; K(w) = 0\}$ and the latter is $W \setminus W_\epsilon$ in which $K(w) > \epsilon$. To W_ϵ, we can apply resolution of singularities.
(2) A typical example of $K(w)$, $w = g(u)$, $f(x, g(u))$, and $a(x, u)$ is shown in Example 7.1.

Proof of Theorem 6.1 Existence of $\epsilon > 0$, $\mathcal{M}^{(R)}$, and g is shown by resolution of singularities, Theorem 2.3. Let us prove eq.(6.5). For arbitrary $u \in U$,

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

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

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

Then for any $u \in U'$,

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

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

$$1 \geq e^{-L} \int_{D_L} \frac{f(x, g(u))^2}{2u^{2k}} q(x) dx. \tag{6.7}$$

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

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

where

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

and $\sum_{\alpha \geq k}$ shows the sum of indices that satisfy

$$\alpha_i \geq k_i \quad (i = 1, 2, \ldots, d) \tag{6.8}$$

and $\sum_{\alpha < k}$ shows the sum of indices that do not satisfy at least one of eq.(6.8). Here $a(x, u)$ is an $L^s(q)$-valued real analytic function. From eq.(6.7), for an arbitrary $u^k \neq 0$ $(u \in U')$,

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

Here $|a(x, u)|$ is a bounded function of $u \in U'$. If $b(x, u) \equiv 0$ does not hold, then $|b(x, u)| \to \infty$ $(u \to 0)$, hence we can choose u and D_L so that the above inequality does not hold. Therefore, we have $b(x, u) \equiv 0$, which shows eq.(6.5). From

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

we obtain eq.(6.6). □

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

is defined by

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

The expectation of $K_n(w)$ is equal to the Kullback–Leibler distance, $E[K_n(w)] = K(w)$. For w satisfying $K(w) > 0$, the log likelihood ratio function is given as

$$K_n(w) = K(w) - \sqrt{K(w)/n}\ \psi_n(w),$$

where $\psi_n(w)$ is defined by

$$\psi_n(w) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{K(w) - f(X_i, w)}{\sqrt{K(w)}}. \tag{6.9}$$

Here $\psi_n(w)$ is an empirical process on $\{w \in W; K(w) \geq \epsilon\}$ and converges to a Gaussian process $\psi(w)$ in law. However, if $K(w) = 0$, $\psi_n(w)$ is ill-defined. For the set $W_\epsilon = \{w \in W; K(w) \leq \epsilon\}$, by Theorem 2.3, there exists a manifold such that, in each local coordinate $(u_1, u_2, \ldots, u_d)$, the Kullback–Leibler distance is given by $K(g(u)) = u^{2k}$. Then

$$\sqrt{K(g(u))} = |u^k|,$$

which is not real analytic at $u^k = 0$. Therefore, eq.(6.9) should be replaced by choosing an appropriate branch so that it is a real analytic function at $u^2 = 0$. The following representation is the theoretical foundation on which singular learning theory is constructed.

Main Theorem 6.1 (Standard form of log likelihood ratio function) *Assume that the fundamental condition (I) holds with index $s = 2$. There exist a real analytic manifold $\mathcal{M}^{(R)}$ and a real analytic and proper map $g : \mathcal{M}^{(R)} \to W_\epsilon^{(R)}$ such that*

$$K(g(u)) = u^{2k}$$

on each local coordinate U of $\mathcal{M}^{(R)}$. The log likelihood ratio function is represented in U by

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}}\, u^k\, \xi_n(u), \tag{6.10}$$

where

$$\xi_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{a(X_i, u) - E_X[a(X, u)]\} \tag{6.11}$$

is an empirical process. Equation (6.10) is called the standard form of the log likelihood ratio function.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Proof of Main Theorem 6.1 From Theorems 2.3, 5.9, and 6.1, this Main Theorem is obtained. □

Remark 6.3 (1) From the definition, the empirical process $\xi_n(u)$ satisfies

$$E[\xi_n(u)] = 0$$

and

$$\begin{aligned} E[\xi_n(u)\xi_n(v)] &= E_X[a(X,u)a(X,v)] - E_X[a(X,u)]E_X[a(X,v)] \\ &= E_X[a(X,u)a(X,v)] - u^k v^k. \end{aligned}$$

In particular, if $K(g(u)) = K(g(v)) = 0$, then

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

(2) The empirical process $\xi_n(u)$ is well defined on the manifold even for $K(g(u)) = 0$, whereas the empirical process $\xi_n(g^{-1}(w))$ is ill-defined if $K(w) = 0$ in general, which is one of the reasons why algebraic geometry is necessary in statistical learning theory.

Theorem 6.2 *(1) Assume that the fundamental condition (I) holds with $s = 2$. Then the empirical processes $\psi_n(w)$ on $\{w; K(w) > \epsilon\}$ and $\xi_n(u)$ on $\mathcal{M}$ converge in law to the Gaussian processes $\psi(w)$ and $\xi(u)$, respectively.*
(2) Assume that the fundamental condition (I) holds with $s = 4$. Then the empirical processes satisfy

$$\lim_{n\to\infty} E\Big[\sup_{K(w)>\epsilon} |\psi_n(w)|^{s-2}\Big] = E\Big[\sup_{K(w)>\epsilon} |\psi(w)|^{s-2}\Big] < \infty,$$

$$\lim_{n\to\infty} E\Big[\sup_{u\in\mathcal{M}} |\xi_n(u)|^{s-2}\Big] = E\Big[\sup_{u\in\mathcal{M}} |\xi(u)|^{s-2}\Big] < \infty.$$

Proof of Theorem 6.2 For $\psi_n(w)$, this theorem is immediately derived from Theorems 5.9 and 5.10. Let us prove the theorem for $\xi_n(u)$. The subset $\mathcal{M}$ is compact because W_ϵ is compact and the resolution map $g : \mathcal{M}^{(R)} \to W_\epsilon^{(R)}$ is proper. Therefore $\mathcal{M}$ can be covered by a finite union of local coordinates. From Theorems 5.9 and 5.10, we immediately obtain the theorem. □

Remark 6.4 By the above theorem, the limiting process $\xi(u)$ is a Gaussian process on $\mathcal{M}$ which satisfies

$$E[\xi(u)] = 0$$

and

$$E[\xi(u)\xi(v)] = E_X[a(X,u)a(X,v)] - u^k v^k.$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

In particular, if $K(g(u)) = K(g(v)) = 0$, then

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

In other words, the Gaussian process $\xi(u)$ has the same mean and covariance function as $\xi_n(u)$. Note that the tight Gaussian process is uniquely determined by its mean and covariance.

Theorem 6.3 *Assume that $q(x)$ and $p(x|w)$ satisfy the fundamental condition (I) with index $s = 4$. If $K(g(u)) = 0$, then*

$$E_X[a(x,u)^2] = E[|\xi_n(u)|^2] = E[|\xi(u)|^2] = 2.$$

Proof of Theorem 6.3 It is sufficient to prove $E_X[a(X,u)^2] = 2$ when $K(g(u)) = 0$. Let the Taylor expansion of $f(x, g(u))$ be

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

Then

$$|a_\alpha(x)| \le \frac{M(x)}{R^\alpha},$$

where R are associated convergence radii and

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

Hence

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

where $c_1 > 0$ is a constant. In the same way as in the proof of Theorem 6.1 and with $f(x, g(u)) = a(x,u)u^k$, for arbitrary u ($u^k \ne 0$), we have

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

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

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Then

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

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

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

for any u that satisfies $u^{2k} = 0$. □

6.2 Evidence and stochastic complexity

Definition 6.2 (Evidence) Let $q(x)$ and $p(x|w)$ be probability distributions which satisfy the fundamental condition (I) with $s = 2$. The set $D_n = \{X_1, X_2, \ldots, X_n\}$ consists of random variables which are independently subject to $q(x)dx$. Let $\varphi(w)$ be a probability density function on $\mathbb{R}^d$. The evidence of a pair $p(x|w)$ and $\varphi(w)$ for D_n is defined by

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

Also the stochastic complexity is defined by

$$F_n = -\log Z_n.$$

The normalized evidence and the normalized stochastic complexity are respectively defined by

$$\begin{aligned} Z_n^0 &= \frac{Z_n}{\prod_{i=1}^{n} q(x_i)^{\beta}} \\ &= \int \exp(-n\beta K_n(w))\varphi(w)dw, \end{aligned} \tag{6.12}$$

$$F_n^0 = -\log Z_n^0. \tag{6.13}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Theorem 6.4 *For arbitrary natural number n, the normalized stochastic complexity satisfies*

$$E\big[F_n^0\big] \leq -\log \int \exp(-n\beta K(w))\varphi(w)dw.$$

Proof of Theorem 6.4

$$\begin{aligned} F_n^0 &= -\log \int \exp(-n\beta K_n(w))\varphi(w)dw \\ &= -\log \int \exp(-n\beta(K_n(w) - K(w)) - n\beta K(w))\varphi(w)dw \\ &= -\log \int \exp(-n\beta(K_n(w) - K(w)))\rho(w)dw \\ &\quad -\log \int \exp(-n\beta K(w))\varphi(w)dw, \end{aligned}$$

where

$$\rho(w) = \frac{\exp(-n\beta K(w))\varphi(w)}{\int \exp(-n\beta K(w'))\varphi(w')dw'}.$$

By Jensen's inequality and a definition $K^*(w) \equiv K_n(w) - K(w)$,

$$\int \exp(-n\beta K^*(w))\rho(w)dw \geq \exp\Big(-\int n\beta K^*(w)\rho(w)dw\Big).$$

Using $E[K^*(w)] = 0$, we obtain the theorem. □

Remark 6.5 In the proof of Theorem 6.4, the convergence in law of $\xi_n(u) \to \xi(u)$ is not needed. Therefore, the upper bound of the stochastic complexity can be shown by the weaker condition.

Definition 6.3 (Fundamental condition (II)) Assume that the set of parameters W is a compact set defined by

$$W = \{w \in \mathbb{R}^d; \pi_1(w) \geq 0, \pi_2(w) \geq 0, \ldots, \pi_k(w) \geq 0\},$$

where $\pi_1(w), \pi_2(w), \ldots, \pi_k(w)$ are real analytic functions on some real open set $W^{(R)} \subset \mathbb{R}^d$. The *a priori* probability density function $\varphi(w)$ is given by $\varphi(w) = \varphi_1(w)\varphi_2(w)$ where $\varphi_1(w) > 0$ is a function of class C^∞ and $\varphi_2(w) \geq 0$ is a real analytic function.

Remark 6.6 In singular statistical models, the set of parameters and the *a priori* distribution should be carefully prepared from the theoretical point of view. In particular, their behavior in the neighborhood of $K(w) = 0$ and the boundary of W have to be set naturally. The condition that $\pi_1(w), \pi_2(w), \ldots, \pi_k(w)$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

and $\varphi_1(w)$ are real analytic functions is necessary because, if at least one of them is a function of class C^∞, there is a pathological example. In fact, if $\varphi_1(w) = \exp(-1/w^2)$ $(w \in \mathbb{R}^1)$ and $K(w) = w^2$ in a neighborhood of the origin, then $(d/dw)^k \varphi_1(0) = 0$ for an arbitrary $k \geq 0$, and

$$\int_{-1}^{1} (w^2)^z \exp\left(-\frac{1}{w^2}\right) dw$$

has no pole.

Theorem 6.5 (Partition of parameter space) *Assume the fundamental conditions (I) and (II) with index $s = 2$. Let $\epsilon > 0$ be a constant. By applying Hironaka's resolution theorem (Theorem 2.3) to a real analytic function,*

$$K(w)(\epsilon - K(w))\varphi_2(w) \prod_{j=1}^{k} \pi_j(w),$$

we can find a real analytic manifold $\mathcal{M}^{(R)}$ and a proper and real analytic map $g : \mathcal{M}^{(R)} \to W_\epsilon^{(R)}$ such that all functions

$$K(g(u)), \quad \epsilon - K(g(u)), \quad \varphi_2(g(u)), \quad \pi_1(g(u)), \ldots, \pi_k(g(u))$$

have only normal crossing singularities. By using Remark 2.14 and Theorem 2.11, we can divide the set $W_\epsilon = \{w \in W; K(w) \leq \epsilon\}$ such that the following conditions (1), (2), (3), and (4) are satisfied.

(1) The set of parameters $\mathcal{M} = g^{-1}(W_\epsilon)$ is covered by a finite set

$$\mathcal{M} = \cup_\alpha M_\alpha,$$

where M_α is given by a local coordinate,

$$M_\alpha = [0, b]^d = \{(u_1, u_2, \ldots, u_d) \; ; \; 0 \leq u_1, u_2, \ldots, u_d \leq b\}.$$

(2) In each M_α,

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

where $k_1, k_2, \ldots, k_d$ are nonnegative integers.

(3) There exists a function $\phi(u)$ of class C^∞ such that

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

where $|g'(u)|$ is the absolute value of the Jacobian determinant and

$$\phi(u) > c > 0 \quad (u \in [0, b]^d)$$

is a function of class C^∞, where $c > 0$ is a positive constant.

(4) There exists a set of functions $\{\sigma_\alpha(u)\}$ of class C^∞ which satisfy

$$\begin{aligned} \sigma_\alpha(u) &\geq 0, \\ \sum_\alpha \sigma_\alpha(u) &= 1, \\ \sigma_\alpha(u) &> 0 \quad (u \in [0,b)^d), \\ supp\ \sigma_\alpha(u) &= [0,b]^d, \end{aligned}$$

such that, for an arbitrary integrable function $H(w)$,

$$\begin{aligned} \int_{W_\epsilon} H(w)\varphi(w)dw &= \int_{\mathcal{M}} H(g(u))\varphi(g(u))|g'(u)|du \\ &= \sum_\alpha \int_{M_\alpha} H(g(u))\phi^*(u)u^h du, \end{aligned}$$

where we defined $\phi^*(u)$ by omitting local coordinate α,

$$\phi^*(u) \equiv \sigma_\alpha(u)\phi(u).$$

Moreover there exist constants $C_1 > 0$ such that

$$\begin{aligned} C_1 \sum_\alpha \int_{M_\alpha} H(g(u))\phi(u)u^h du &\leq \int_{W_\epsilon} H(w)\varphi(w)dw \\ &\leq \sum_\alpha \int_{M_\alpha} H(g(u))\phi(u)u^h du. \quad (6.14) \end{aligned}$$

Proof of Theorem 6.5 This theorem is obtained by the resolution theorem (Theorem 2.3), Theorem 2.11, and Remark 2.14. □

Theorem 6.6 *Assume the fundamental conditions (I) and (II) with index $s = 2$. The holomorphic function of $z \in \mathbb{C}$,*

$$\zeta(z) = \int K(w)^z \varphi(w)dw \quad (\mathrm{Re}(z) > 0), \quad (6.15)$$

can be analytically continued to the unique meromorphic function on the entire complex plane whose poles are all real, negative, and rational numbers.

Proof of Theorem 6.6 Let us define

$$\zeta_1(z) = \int_{K(w)<\epsilon} K(w)^z \varphi(w)dw,$$

$$\zeta_2(z) = \int_{K(w)\geq\epsilon} K(w)^z \varphi(w)dw.$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

In an arbitrary neighborhood of $z \in \mathbb{C}$, $|(\partial K(w)^z/\partial z)\varphi(w)|$ is a bounded function on the compact set $\{w \in W; K(w) \geq \epsilon\}$, hence $\zeta_2(z)$ is a holomorphic function on the entire complex plane. Let us study $\zeta_1(z)$. The Kullback–Leibler distance and the *a priori* distribution can be represented as in Theorem 6.5.

$$\zeta_1(z) = \sum_{\alpha} \int_{M_\alpha} u^{2kz}\; u^h\; \phi^*(u) du.$$

Since $\phi^*(u)$ has the finite-order Taylor expansion,

$$\phi^*(u) = \sum_{|j|\leq n} a_j u^j + R_n(u),$$

where n can be taken as large as necessary. In the region $\mathrm{Re}(z) > 0$,

$$\int_{[0,b]^d} u^{2kz+h+j} du = \prod_{p=1}^{d} \frac{b^{2k_p z+h_p+j_p+1}}{(2k_p z + h_p + j_p + 1)}.$$

Hence the function $\zeta_1(z)$ can be analytically continued to the unique meromorphic function and all poles are real, negative, and rational numbers. □

Definition 6.4 (Zeta function and learning coefficient) The meromorphic function $\zeta(z)$ that is analytically continued from eq.(6.15) is called the zeta function of a statistical model. The largest pole and its order are denoted by $(-\lambda)$ and m, respectively, where λ and m are respectively called the learning coefficient and its order. If the Kullback–Leibler distance and the *a priori* distribution are represented as in Theorem 6.5, then the learning coefficient is given by

$$\lambda = \min_{\alpha} \min_{1\leq j\leq d} \left(\frac{h_j + 1}{2k_j} \right), \tag{6.16}$$

and its order m is

$$m = \max_{\alpha}\; \sharp\{j; \lambda = (h_j + 1)/(2k_j)\}, \tag{6.17}$$

where $\sharp$ shows the number of elements of the set S. Let $\{\alpha^*\}$ be a set of all local coordinates in which both the minimization in eq.(6.16) and the maximization in eq.(6.17) are attained. Such a set of local coordinates $\{\alpha^*\}$ is said to be the essential family of local coordinates. For each local coordinate α^* in the essential family of local coordinates, we can assume without loss of generality that u is represented as $u = (x, y)$ such that

$$x = (u_1, u_2, \ldots, u_m),$$
$$y = (u_{m+1}, u_{m+2}, \ldots, u_d),$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

and that

$$\lambda = \frac{h_j+1}{2k_j} \quad (1 \le j \le m),$$
$$\lambda < \frac{h_j+1}{2k_j} \quad (m+1 \le j \le d).$$

For a given function $f(u) = f(x, y)$, we use the notation $f_0(y) \equiv f(0, y)$.

Theorem 6.7 (Convergence in law of evidence) *Assume the fundamental conditions (I) and (II) with index $s = 2$. The constants λ and m are the learning coefficient and its order respectively. Let Z_n^0 be the normalized evidence defined in eq.(6.12). When $n \to \infty$, the following convergence in law holds,*

$$\frac{n^\lambda Z_n^0}{(\log n)^{m-1}} \to \sum_{\alpha^*} \gamma_b \int_0^\infty dt \int_{M_{\alpha^*}} t^{\lambda-1}\, e^{-\beta t+\sqrt{t}\beta\xi_0(y)}\, \phi_0^*(y)dy,$$

where $\gamma_b > 0$ is a constant defined by eq.(4.16).

Proof of Theorem 6.7 The normalized evidence can be divided as

$$Z_n^0 = Z_n^{(1)} + Z_n^{(2)},$$

where

$$Z_n^{(1)} = \int_{K(w)\le\epsilon} e^{-n\beta K_n(w)}\varphi(w)dw,$$
$$Z_n^{(2)} = \int_{K(w)>\epsilon} e^{-n\beta K_n(w)}\varphi(w)dw.$$

Firstly, let us study $Z_n^{(2)}$. If $K(w) > \epsilon$, by using the Cauchy–Schwarz inequality,

$$\begin{aligned} nK_n(w) &= nK(w) - \sqrt{K(w)}\psi_n(w) \\ &\ge \frac{nK(w) - \psi_n(w)^2}{2} \\ &\ge \frac{1}{2}\Big(n\epsilon - \sup_{K(w)>\epsilon} |\psi_n(w)|^2\Big). \end{aligned}$$

Since $\psi_n(w)$ is an empirical process which converges to a Gaussian process with supremum norm in law, $\sup_{K(w)>\epsilon} |\psi_n(w)|^2$ converges in law, and therefore

$$0 \le \frac{n^\lambda}{(\log n)^{m-1}} Z_n^{(2)} \le \frac{n^\lambda e^{-n\beta\epsilon/2}}{(\log n)^{m-1}} \exp\Big(\frac{\beta}{2} \sup_{K(w)>\epsilon} |\psi_n(w)|^2\Big) \tag{6.18}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

converges to zero in probability by Theorem 5.2. Secondly, $Z_n^{(1)}$ is given by

$$Z_n^{(1)} = \sum_{\alpha} \int_0^{\infty} dt \int_{M_\alpha} \exp(-n\beta u^{2k} + \beta\sqrt{n}u^k \xi_n(u))u^h \phi^*(u)du.$$

Let us define

$$Y^{(1)}(\xi_n) \equiv \gamma_b \sum_{\alpha^*} \int_0^{\infty} dt \int dy\ t^{\lambda-1} y^{\mu} e^{-\beta t+\beta\sqrt{t}\xi_{n,0}(y)} \phi_0^*(y).$$

Then $Y^{(1)}(\xi)$ is a continuous function of ξ with respect to the norm $\|\cdot\|$, hence the convergence in law $Y^{(1)}(\xi_n) \to Y^{(1)}(\xi)$ holds. Let us apply Theorem 4.9 to the coordinates α^* with $p = 0,\ r = m,\ f = \xi_n$. Also we apply Theorem 4.8 to the other coordinates with $r = m - 1$ and $f = \xi_n$. Then there exists a constant $C_1 > 0$ such that

$$\left| \frac{n^\lambda Z_n^{(1)}}{(\log n)^{m-1}} - Y^{(1)}(\xi_n) \right| \leq \frac{C_1}{\log n} \sum_{\alpha} e^{\beta\|\xi_n\|^2/2} \{\beta\|\xi_n\|\|\phi^*\| + \|\nabla\phi^*\| + \|\phi^*\|\}.$$

Since $\|\xi_n\|$ converges in law, the right-hand side of this equation converges to zero in probability. Therefore, the convergence in law

$$\frac{n^\lambda}{(\log n)^{m-1}} Z_n^{(1)} \to Y^{(1)}(\xi)$$

holds, which completes the theorem. □

Main Theorem 6.2 (Convergence of stochastic complexity)
(1) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index $s = 2$. Then the following convergence in law holds:

$$\begin{aligned} &F_n^0 - \lambda \log n + (m-1)\log\log n \\ &\quad \to -\log \sum_{\alpha^*} \gamma_b \int_0^{\infty} dt \int t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\xi_0(y)}\ \phi_0^*(y)dy. \end{aligned}$$

(2) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index $s = 4$. Then the following convergence of expectation holds:

$$\begin{aligned} &E\big[F_n^0\big] - \lambda \log n + (m-1)\log\log n \\ &\to -E\Big[\log \sum_{\alpha^*} \gamma_b \int_0^{\infty} dt \int t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\xi_0(y)}\ \phi_0^*(y)dy\Big]. \end{aligned}$$

Proof of Main Theorem 6.2 (1) From Theorem 6.7 and the fact that $-\log(\cdot)$ is a continuous function, the first part is proved by Theorem 5.1.

(2) For the second part, it is sufficient to prove that

$$A_n \equiv -\log \frac{Z_n^0 n^\lambda}{(\log n)^{m-1}}$$

is asymptotically uniformly integrable. By using the same decomposition of Z_n^0 as in the proof of Theorem 6.7,

$$A_n = -\log\Big(\frac{Z_n^{(1)} n^\lambda}{(\log n)^{m-1}} + \frac{Z_n^{(2)} n^\lambda}{(\log n)^{m-1}}\Big).$$

By eq.(6.18) and Theorem 4.8 with $p = 0$ and $r = m$,

$$\begin{aligned} A_n &\geq -\log\Big\{\exp(\beta\|\xi_n\|^2/2)\|\varphi\| + \exp\Big(\frac{\beta}{2}\sup_{K(w)>\epsilon}|\psi_n(w)|^2\Big)\Big\} \\ &\geq -(\beta/2)\max\Big\{\|\xi_n\|^2, \sup_{K(w)>\epsilon}|\psi_n(w)|^2\Big\} + C_2, \end{aligned}$$

where C_2 is a constant and we used the fact that $\log(e^p + e^q) \leq \max\{p, q\} + \log 2$ for arbitrary p, q. On the other hand, by $\phi(u) > 0$ and Theorem 6.5 (4),

$$Z_n^{(2)} \geq C_1 \sum_\alpha \int_0^\infty dt \int du \exp(-n\beta u^{2k} + \sqrt{n}\beta u^k \xi_n) u^h du.$$

Hence, by Theorem 4.8, and $A_n \leq -\log(Z_n^{(2)} n^\lambda/(\log n)^{m-1})$,

$$A_n \leq \beta\|\xi_n\|^2/2 - \log\min|\phi| + C_3,$$

where C_3 is a constant. By Theorem 5.8, $E[\|\xi_n\|^4] < \infty$, $E[\|\psi_n\|^4] < \infty$. Hence A_n is asymptotically uniformly integrable. By Theorem 5.5, we obtain the theorem. □

Corollary 6.1 *(1) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index $s = 2$. Then the following asymptotic expansion holds,*

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

where F_n^R is a random variable which converges to a random variable in law.
(2) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index $s = 4$. Then the following asymptotic expansion of the expectation holds,

$$E[F_n] = n\beta S + \lambda \log n - (m-1)\log\log n + E\big[F_n^R\big],$$

where $E[F_n^R]$ converges to a constant.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

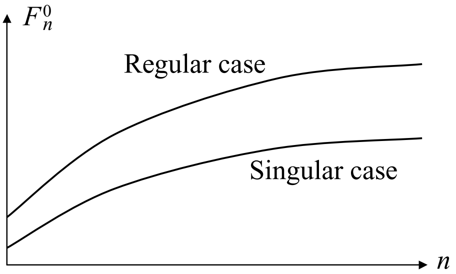

Fig. 6.1. Stochastic complexity

Proof of Corollary 6.1 By the definition

$$F_n^R = -\log \sum_{\alpha^*} \gamma_b \int_0^\infty dt \int du\ t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\xi_{n,0}(y)}\ \phi_0^*(y)dy,$$

this corollary is immediately derived from Main Theorem 6.2. □

Remark 6.7 (1) Figure 6.1 shows the behavior of the normalized stochastic complexity. If the *a priori* distribution is positive on $K(w) = 0$, then the learning coefficient is equal to the real log canonical threshold. Note that the learning coefficient is invariant under a transform

$$\begin{aligned} p(x|w) &\mapsto p(x|g(u)), \\ \varphi(w)dw &\mapsto \varphi(g(u))|g'(u)|du. \end{aligned}$$

(2) If a model is regular, $\lambda = d/2$ and $m = 1$, where d is the dimension of the parameter space. Examples of λ and m in several models are shown in Chapter 7.
(3) Assume that $\beta = 1$. By Main Theorem 6.2 and Theorem 1.2, if the mean of Bayes generalization error $B_{\rm g}$ has an asymptotic expansion, then

$$\begin{aligned} E[B_{\rm g}] &= E\big[F_{n+1}^0\big] - E\big[F_n^0\big], \\ &= \frac{\lambda}{n} + o\Big(\frac{1}{n}\Big). \end{aligned}$$

However, in general, even if a function $f(n)$ has an asymptotic expansion, $f(n+1) - f(n)$ may not have an asymptotic expansion. To prove the Bayes generalization error has an asymptotic expansion, we need to show a more precise result as in the following section.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

6.3 Bayes and Gibbs estimation

In the previous section, we studied the asymptotic expansion of the stochastic complexity. In this section, mathematical relations in the Bayes quartet are proved, which are called equations of states in learning. Throughout this section, we assume the fundamental conditions (I) and (II) with index $s = 6$. Firstly, the main theorems are introduced without proof. Secondly, basic lemmas are prepared. Finally, the main theorems are proved.

6.3.1 Equations of states

Assume that random samples $X_1, X_2, \ldots, X_n$ are independently taken from the probability distribution $q(x)dx$. For a given set of random samples $D_n = \{X_1, X_2, \ldots, X_n\}$, the generalized *a posteriori* distribution is defined by

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

which can be rewritten as

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

where $\beta > 0$ is the inverse temperature.

Definition 6.5 (Bayes quartet) Let $E_w[\cdot]$ be the expectation value using $p(w|D_n)$. Four errors are defined.

(1) Bayes generalization error,

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

(2) Bayes training error,

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

(3) Gibbs generalization error,

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

(4) Gibbs training error,

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

This set of four errors is called the Bayes quartet.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

The most important variable in practical applications among them is the Bayes generalization error because it determines the accuracy of the estimation. However, we prove that there are mathematical relations between them. It is shown in Theorem 1.3 that, by using the log density ratio function

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

and the log likelihood ratio function

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

the Bayes quartet can be rewritten as

$$\begin{aligned} B_{\rm g} &= E_X\Big[-\log E_w[e^{-f(X,w)}]\Big], \\ B_{\rm t} &= \frac{1}{n}\sum_{i=1}^{n} -\log E_w[e^{-f(X_i,w)}], \\ G_{\rm g} &= E_w[K(w)], \\ G_{\rm t} &= E_w[K_n(w)]. \end{aligned}$$

If the true distribution $q(x)$ is contained in the statistical model $p(x|w)$, then the four errors in the Bayes quartet converge to zero in probability when n tends to infinity. In this section, we show how fast random variables in the Bayes quartet tend to zero.

Theorem 6.8 *Assume the fundamental conditions (I) and (II) with $s = 6$.*
(1) There exist random variables $B_{\rm g}^*$, $B_{\rm t}^*$, $G_{\rm g}^*$, and $G_{\rm t}^*$ such that, when $n \to \infty$, the following convergences in law hold:

$$nB_{\rm g} \to B_{\rm g}^*, \quad nB_{\rm t} \to B_{\rm t}^*, \quad nG_{\rm g} \to G_{\rm g}^*, \quad nG_{\rm t} \to G_{\rm t}^*.$$

(2) When $n \to \infty$, the following convergence in probability holds:

$$n(B_{\rm g} - B_{\rm t} - G_{\rm g} + G_{\rm t}) \to 0.$$

(3) Expectation values of the Bayes quartet converge:

$$\begin{aligned} E[nB_{\rm g}] &\to E[B_{\rm g}^*], \\ E[nB_{\rm t}] &\to E[B_{\rm t}^*], \\ E[nG_{\rm g}] &\to E[G_{\rm g}^*], \\ E[nG_{\rm t}] &\to E[G_{\rm t}^*]. \end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Main Theorem 6.3 (Equations of states in statistical estimation) *Assume the fundamental conditions (I) and (II) with $s = 6$. For arbitrary $q(x)$, $p(x|w)$, and $\varphi(w)$, the following equations hold.*

$$E[B_{\mathrm{g}}^*] - E[B_{\mathrm{t}}^*] = 2\beta(E[G_{\mathrm{t}}^*] - E[B_{\mathrm{t}}^*]), \tag{6.19}$$

$$E[G_{\mathrm{g}}^*] - E[G_{\mathrm{t}}^*] = 2\beta(E[G_{\mathrm{t}}^*] - E[B_{\mathrm{t}}^*]). \tag{6.20}$$

Remark 6.8 (1) Main Theorem 6.3 shows that the increases of errors from training to prediction are in proportion to the differences between the Bayes and Gibbs training. We give Main Theorem 6.3 the title **Equations of states in statistical estimation**, because these hold for any true distribution, any statistical model, any *a priori* distribution, and any singularities. If two of these errors are measured by observation, then the other two errors can be estimated without any knowledge of the true distribution.
(2) From the equations of states, widely applicable information criteria (WAIC) are obtained. See Section 8.3.
(3) Although the equations of states hold universally, the four errors themselves strongly depend on a true distribution, a statistical model, an *a priori* distribution, and singularities.

Corollary 6.2 *The two generalization errors can be estimated by the two training errors,*

$$\begin{pmatrix} E[B_{\mathrm{g}}^*] \\ E[G_{\mathrm{g}}^*] \end{pmatrix} = \begin{pmatrix} 1-2\beta & 2\beta \\ -2\beta & 1+2\beta \end{pmatrix} \begin{pmatrix} E[B_{\mathrm{t}}^*] \\ E[G_{\mathrm{t}}^*] \end{pmatrix}. \tag{6.21}$$

Proof of Corollary 6.2 This corollary is directly derived from Main Theorem 6.3. □

Remark 6.9 (1) From eq.(6.21), it follows that

$$\begin{pmatrix} E[G_{\mathrm{t}}^*] \\ E[B_{\mathrm{t}}^*] \end{pmatrix} = \begin{pmatrix} 1-2\beta & 2\beta \\ -2\beta & 1+2\beta \end{pmatrix} \begin{pmatrix} E[G_{\mathrm{g}}^*] \\ E[B_{\mathrm{g}}^*] \end{pmatrix},$$

which shows that there is symmetry in the Bayes quartet.

Theorem 6.9 *Assume the fundamental conditions (I) and (II) with index $s = 6$. When $n \to \infty$, the convergence in probability*

$$nG_{\mathrm{g}} + nG_{\mathrm{t}} - \frac{2\lambda}{\beta} \to 0$$

holds, where λ is the learning coefficient. Moreover,

$$E[G_{\mathrm{g}}^*] + E[G_{\mathrm{t}}^*] = \frac{2\lambda}{\beta}. \tag{6.22}$$

Corollary 6.3 *Assume the fundamental conditions (I) and (II) with* $s = 6$. *The following convergence in probability holds,*

$$nB_{\mathrm{g}} - nB_{\mathrm{t}} + 2nG_{\mathrm{t}} - \frac{2\lambda}{\beta} \to 0,$$

where λ *is the learning coefficient. Moreover,*

$$E[B_{\mathrm{g}}^*] - E[B_{\mathrm{t}}^*] + 2E[G_{\mathrm{t}}^*] = \frac{2\lambda}{\beta}.$$

In particular, if $\beta = 1$, $E[B_{\mathrm{g}}^*] = \lambda$.

Proof of Corollary 6.3 This corollary is derived from Theorem 6.8 (1) and 6.9. □

Definition 6.6 (Empirical variance) The empirical variance V of the log likelihood function is defined by

$$V = \sum_{i=1}^{n} \Big\{ E_w[(\log p(X_i|w))^2\ ] - (\ E_w[\log p(X_i|w)]\ )^2 \Big\}. \tag{6.23}$$

By using the log density ratio function $f(x, w) = \log(q(x)/p(x|w))$, this can be rewritten as

$$V = \sum_{i=1}^{n} \Big\{ E_w[\ f(X_i|w)^2\ ] - E_w[f(X_i|w)]^2 \Big\}. \tag{6.24}$$

Theorem 6.10 *The following convergences in probability hold:*

$$V - 2(nG_{\mathrm{t}} - nB_{\mathrm{t}}) \to 0, \tag{6.25}$$

$$V - 2(nG_{\mathrm{g}} - nB_{\mathrm{g}}) \to 0. \tag{6.26}$$

There exists a constant $\nu = \nu(\beta) > 0$ *such that*

$$\lim_{n\to\infty} E[V] = \frac{2\nu(\beta)}{\beta}$$

and

$$E[B_{\mathrm{g}}^*] = \frac{\lambda}{\beta} + \Big(1 - \frac{1}{\beta}\Big)\nu(\beta), \tag{6.27}$$

$$E[B_{\mathrm{t}}^*] = \frac{\lambda}{\beta} - \Big(1 + \frac{1}{\beta}\Big)\nu(\beta), \tag{6.28}$$

$$E[G_{\mathrm{g}}^*] = \frac{\lambda}{\beta} + \nu(\beta), \tag{6.29}$$

$$E[G_{\mathrm{t}}^*] = \frac{\lambda}{\beta} - \nu(\beta). \tag{6.30}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

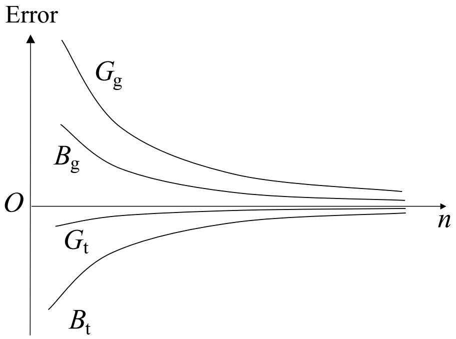

Fig. 6.2. Bayes and Gibbs errors

The constant $\nu(\beta) > 0$ is a singular fluctuation defined in eq.(6.48), which satisfies an inequality,

$$0 \leq \nu(\beta) \leq \frac{E_\xi[\|\xi\|^2]}{8} + \frac{E_\xi[\|\xi\|^2]^{1/2}}{8}\sqrt{E_\xi[\|\xi\|^2] + 16\lambda/\beta},$$

where $\|\xi\|$ is the maximum value of the random process $\xi(u)$.

Remark 6.10 The behavior of the Bayes quartet is shown in Figure 6.2. From Theorem 6.10, the singular fluctuation $\nu(\beta)$ can be represented in several ways:

$$\begin{aligned}\nu(\beta) &= (1/2)(E[B_g^*] - E[B_t^*]) \\ &= (1/2)(E[G_g^*] - E[G_t^*]) \\ &= \beta\,(E[G_g^*] - E[B_g^*]) \\ &= \beta\,(E[G_t^*] - E[B_t^*]) \\ &= \lim_{n\to\infty} \frac{\beta}{2} E[V].\end{aligned}$$

From the equations of states,

$$E[B_g] = E[B_t] + \frac{\beta}{n} E[V] + o(1/n), \tag{6.31}$$

$$E[G_g] = E[G_t] + \frac{\beta}{n} E[V] + o(1/n). \tag{6.32}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Using these relations, $\nu(\beta)$ can be estimated from numerical experiments. By definition, $\nu(\beta)$ is invariant under a birational transform

$$p(x|w) \mapsto p(x|g(u)),$$

$$\varphi(w)dw \mapsto \varphi(g(u))|g'(u)|du.$$

Remark 6.11 (Regular model case) In singular learning machines, λ is not equal to $d/2$ in general, and $\nu(\beta)$ depends on β. In a regular statistical model, $\lambda = \nu(\beta) = d/2$, which means that,

$$E[B_{\rm g}^*] = \frac{d}{2},$$

$$E[B_{\rm t}^*] = -\frac{d}{2},$$

$$E[G_{\rm g}^*] = \Big(1 + \frac{1}{\beta}\Big)\frac{d}{2},$$

$$E[G_{\rm t}^*] = \Big(- 1 + \frac{1}{\beta}\Big)\frac{d}{2},$$

which is a special case of Main Theorem 6.3. This result is obtained by the classical asymptotic theory of the maximum likelihood estimator with positive definite Fisher information matrix. Assume that examples are independently taken from $q(x) = p(x|w_0)$. Let $I_n(w)$ and $I(w)$ be the empirical and mean Fisher information matrices respectively. The expectation of a function $f(w)$ under the *a posteriori* distribution is asymptotically given by

$$E_w[f(w)] \cong \frac{\int f(w)\exp\big(-\frac{n\beta}{2}\big|I_n(\hat{w})^{1/2}(w-\hat{w})\big|^2\big)dw}{\int \exp\big(-\frac{n\beta}{2}\big|I_n(\hat{w})^{1/2}(w-\hat{w})\big|^2\big)dw},$$

where, for a given symmetric matrix I and a vector v, $|I^{1/2}v|^2 = (v, Iv)$. The random variable $\sqrt{n}(\hat{w} - w_0)$ converges in law to the normal distribution with mean zero and covariance matrix $I(w_0)^{-1}$, and the Taylor expansions for w_0 and $\hat{w}$ are given by

$$E[nK(\hat{w})] = E\Big[nK(w_0) + \frac{n}{2}|I(w_0)^{1/2}(\hat{w} - w_0)|^2\Big]$$

$$= \frac{d}{2} + o(1)$$

$$E[nK_n(w_0)] = E\Big[nK_n(\hat{w}) + \frac{n}{2}|I_n(\hat{w})^{1/2}(w_0 - \hat{w})|^2\Big]$$

$$= 0.$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Therefore the Gibbs generalization and training errors are given by

$$\begin{aligned}
E[nG_{\mathrm{g}}] &= E[E_w[nK(w)]] \\
&= E\Big[\frac{n}{2}E_w[|I(\hat{w})^{1/2}(w-\hat{w})|^2]+nK(\hat{w})\Big]+o(1) \\
&= \frac{d}{2\beta}+\frac{d}{2}+o(1), \\
E[nG_{\mathrm{t}}] &= E[E_w[nK_n(w)]] \\
&= E\Big[\frac{n}{2}E_w[|I(\hat{w})^{1/2}(w-\hat{w})|^2]+nK_n(\hat{w})\Big]+o(1) \\
&= \frac{d}{2\beta}-\frac{d}{2}+o(1).
\end{aligned}$$

Consequently, in regular statistical models, $\lambda = \nu(\beta) = d/2$.

6.3.2 Basic lemmas

In this subsection, some basic lemmas are prepared which are used in the proofs of the above theorems.

Note that there are three different expectations. The first is the expectation over the parameter space of the *a posteriori* distribution. The second is the expectation over random samples $D_n = \{X_1, X_2, \ldots, X_n\}$. It is denoted by $E[\ ]$, which is also used for the expectation over the limit process ξ. The last is the expectation over the test sample X. It is denoted by $E_X[\ ]$.

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

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

The four errors of the Bayes quartet in the restricted region are given by

$$\begin{aligned}
B_{\mathrm{g}}(a) &= E_X\Big[-\log E_w[e^{-f(X,w)}|_{K(w)\le a}]\Big], \\
B_{\mathrm{t}}(a) &= \frac{1}{n}\sum_{j=1}^{n} -\log E_w[e^{-f(X_j,w)}|_{K(w)\le a}], \\
G_{\mathrm{g}}(a) &= E_w[K(w)|_{K(w)\le a}], \\
G_{\mathrm{t}}(a) &= E_w[K_n(w)|_{K(w)\le a}].
\end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Since W is compact and $K(w)$ is a real analytic function,

$$\overline{K} = \sup_{w \in W} K(w)$$

is finite, therefore

$$\begin{aligned} B_{\mathrm{g}}(\overline{K}) &= B_{\mathrm{g}}, \\ B_{\mathrm{t}}(\overline{K}) &= B_{\mathrm{t}}, \\ G_{\mathrm{g}}(\overline{K}) &= G_{\mathrm{g}}, \\ G_{\mathrm{t}}(\overline{K}) &= G_{\mathrm{t}}. \end{aligned}$$

Also we define $\eta_n(w)$ for w such that $K(w) > 0$ by

$$\eta_n(w) = \frac{K(w) - K_n(w)}{\sqrt{K(w)}}, \tag{6.33}$$

and

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

Let H_{t} denote $H_{\mathrm{t}}(\overline{K})$.

Lemma 6.1 *(1) For an arbitrary $a > 0$, the following inequalities hold.*

$$\begin{aligned} & B_{\mathrm{t}}(a) \leq G_{\mathrm{t}}(a) \leq \tfrac{3}{2} G_{\mathrm{g}}(a) + \tfrac{1}{2} H_{\mathrm{t}}(a), \\ & 0 \leq B_{\mathrm{g}}(a) \leq G_{\mathrm{g}}(a), \\ & -\tfrac{1}{4} H_{\mathrm{t}}(a) \leq G_{\mathrm{t}}(a). \end{aligned}$$

(2) In particular, by putting $a = \overline{K}$,

$$\begin{aligned} & B_{\mathrm{t}} \leq G_{\mathrm{t}} \leq \tfrac{3}{2} G_{\mathrm{g}} + \tfrac{1}{2} H_{\mathrm{t}}, \\ & 0 \leq B_{\mathrm{g}} \leq G_{\mathrm{g}}, \\ & -\tfrac{1}{4} H_{\mathrm{t}} \leq G_{\mathrm{t}}. \end{aligned}$$

Proof of Lemma 6.1 (1) Because $B_{\mathrm{g}}(a)$ is the Kullback–Leibler distance from $q(x)$ to the Bayes predictive distribution using the restricted *a priori* distribution on $K(w) \leq a$, it follows that $B_{\mathrm{g}}(a) \geq 0$. Using Jensen's inequality,

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

hence $B_{\mathrm{g}}(a) \leq G_{\mathrm{g}}(a)$ and $B_{\mathrm{t}}(a) \leq G_{\mathrm{t}}(a)$. By using the Cauchy–Schwarz inequality,

$$\begin{aligned} K_n(w) &= K(w) - \sqrt{K(w)}\,\eta_n(w) \\ &\leq K(w) + \tfrac{1}{2}\{K(w) + \eta_n(w)^2\}. \end{aligned} \tag{6.34}$$

Therefore $G_{\mathrm{t}}(a) \leq (3G_{\mathrm{g}}(a) + H_{\mathrm{t}}(a))/2$. Also

$$\begin{aligned} K_n(w) &= K(w) - \sqrt{K(w)}\eta_n(w) \\ &\geq \Big(\sqrt{K(w)} - \frac{\eta_n(w)}{2}\Big)^2 - \frac{\eta_n(w)^2}{4} \\ &\geq -\frac{\eta_n(w)^2}{4}. \end{aligned} \tag{6.35}$$

Hence we have $G_{\mathrm{t}}(a) \geq -H_{\mathrm{t}}(a)/4$. (2) is immediately derived from (1). □

Remark 6.12 (1) By Lemma 6.1, if $nH_{\mathrm{t}}(a)$, $nG_{\mathrm{g}}(a)$, and $nB_{\mathrm{t}}(a)$ are asymptotically uniformly integrable (AUI), then $nG_{\mathrm{t}}(a)$ and $nB_{\mathrm{g}}(a)$ are also AUI, for an arbitrary $a > 0$.
(2) In Lemma 6.4, we prove $nH_{\mathrm{t}}(\epsilon)$ is AUI. In Lemma 6.5, we prove $nG_{\mathrm{g}}(\epsilon)$ and $nB_{\mathrm{t}}(\epsilon)$ are AUI. In Lemma 6.2, we show nH_{t} is AUI using Lemma 6.4. In Lemma 6.3, we show nG_{g} and nB_{t} are AUI using Lemma 6.5. Then the four errors in the Bayes quartet are all AUI.
(3) Based on Theorem 5.4 (3), if $E[|X_n|^s] < \infty$, then $X_n^{s-\delta}$ $(\delta > 0)$ is AUI.

Lemma 6.2 *(1) There exists a constant $C_H > 0$ such that*

$$E[(nH_{\mathrm{t}})^3] = C_H < \infty.$$

(2) For an arbitrary $\delta > 0$,

$$P(nH_{\mathrm{t}} > n^{\delta}) \leq \frac{C_H}{n^{3\delta}}. \tag{6.36}$$

(3) nH_{t} is asymptotically uniformly integrable.

Proof of Lemma 6.2 (1) For any $\epsilon > 0$ and $a > 0$,

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

is an empirical process and $f(x,w)$ is a real analytic function of w. Therefore

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

It is proved in Lemma 6.4 that

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

Therefore, by the definition of H_{t}, (1) is obtained.
(2) Let S be a random variable defined by

$$S = \begin{cases} 1 & \text{if } nH_{\mathrm{t}} > n^{\delta} \\ 0 & \text{otherwise.} \end{cases}$$

Then $E[S] = P(nH_{\mathrm{t}} > n^{\delta})$ and

$$C_H = E[(nH_{\mathrm{t}})^3] \geq E[(nH_{\mathrm{t}})^3\ S] \geq E[S]\ n^{3\delta}.$$

(3) is immediately derived from (1). □

Lemma 6.3 *(1) For an arbitrary* $\epsilon > 0$, *the following convergences in probability hold:*

$$n(B_{\mathrm{g}} - B_{\mathrm{g}}(\epsilon)) \to 0,$$
$$n(B_{\mathrm{t}} - B_{\mathrm{t}}(\epsilon)) \to 0,$$
$$n(G_{\mathrm{g}} - G_{\mathrm{g}}(\epsilon)) \to 0,$$
$$n(G_{\mathrm{t}} - G_{\mathrm{t}}(\epsilon)) \to 0.$$

(2) The four errors of the Bayes quartet nB_{g}, nB_{t}, nG_{g}, *and* nG_{t} *are all asymptotically uniformly integrable.*

Proof of Lemma 6.3 We use the notation,

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

By using the Cauchy–Schwarz inequality,

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

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

$$S_1(f(w)) \leq \big(\sup_w f(w)\big)\ e^{-n\beta\epsilon/2} \exp\Big(\frac{\beta}{2}nH_{\mathrm{t}}\Big),$$

$$S_0(g(w)) \geq c_0\ \big(\inf_w g(w)\big)\ n^{-d/2} \exp\Big(-\frac{\beta}{2}nH_{\mathrm{t}}\Big),$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

where $c_0 > 0$ is a constant and d is the dimension of the parameter space. Hence

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

where, by using Theorem 7.2,

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

Then $|\log s(n)| \leq n\beta\epsilon/2 + n\beta H_t + o(n)$. Let $M_n \equiv \sum_{j=1}^{n} M(X_j)/n$. Then $E[M_n^3] \leq E_X[M(X)^3]$, $E_X[M(X)^k]_{\{M(X)>n\}} \leq E_X[M(X)^3]/n^{3-k}$.
(1) Firstly, we study the Bayes generalization error,

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

Thus

$$\begin{aligned} n|B_{\mathrm{g}} - B_{\mathrm{g}}(\epsilon)| &\leq nE_X\Big[\log\Big(1+\frac{S_1(e^{-f(X,w)})}{S_0(e^{-f(X,w)})}\Big) + \log\Big(1+\frac{S_1(1)}{S_0(1)}\Big)\Big] \\ &\leq nE_X[\log(1+s(n)\,e^{2M(X)})] + ns(n) \\ &= ns(n) + nE_X[\log(1+s(n)\,e^{2M(X)})]_{\{2M(X)\leq n\beta\epsilon/4\}} \\ &\quad + nE_X[\log(1+s(n)\,e^{2M(X)})]_{\{2M(X)>n\beta\epsilon/4\}} \\ &\leq ns(n) + ns(n)\exp(n\beta\epsilon/4) \\ &\quad + nE_X[|2M(X)| + |\log s(n)|]_{\{2M(X)>n\beta\epsilon/4\}}. \end{aligned}$$

It follows that $n(B_{\mathrm{g}} - B_{\mathrm{g}}(\epsilon)) \to 0$. Secondly, in the same way, the Bayes training error satisfies

$$n|B_t - B_t(\epsilon)| \leq \sum_{j=1}^{n} \log(1+s(n)\; e^{2M(X_j)}) + n\log(1+s(n)) \equiv L_n. \quad (6.37)$$

We can prove the convergence in mean $E[L_n] \to 0$ because

$$\begin{aligned} E[L_n] &= E[L_n]_{\{H_t\leq\beta\epsilon/4\}} + E[L_n]_{\{H_t>\beta\epsilon/4\}} \\ &\leq nE_X[\log(1+(n^d/c_0)\,e^{2M(X)-n\beta\epsilon/4})] \\ &\quad + \frac{n^{d+1}}{c_0}\exp(-n\beta\epsilon/4) + 2nE[M_n + |\log s(n)|]_{\{H_t>\beta\epsilon/4\}}. \end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Thus we obtain $n(B_{\rm g} - B_{\rm g}(\epsilon)) \to 0$. Thirdly, the Gibbs generalization error can be estimated as

$$\begin{aligned} n|G_{\rm g} - G_{\rm g}(\epsilon)| &\leq \Big| n \frac{S_0(K(w)) + S_1(K(w))}{S_0(1) + S_1(1)} - \frac{nS_0(K(w))}{S_0(1)} \Big| \\ &\leq \frac{nS_1(K(w))}{S_0(1)} + \frac{nS_0(K(w))S_1(1)}{S_0(1)^2} \\ &\leq 2n\,\overline{K}\,s(n), \end{aligned} \tag{6.38}$$

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

$$\begin{aligned} n|G_{\rm t} - G_{\rm t}(\epsilon)| &\leq 2n\;s(n)\;\sup_w |K_n(w)| \\ &\leq 2n\;s(n)\,M_n, \end{aligned}$$

which converges to zero in probability.
(2) Firstly, from Lemma 6.2, $nH_{\rm t}$ is AUI. Secondly, let us prove $nB_{\rm t}$ is AUI. From eq.(6.37),

$$|nB_{\rm t}| \leq |nB_{\rm t}(\epsilon)| + L_n.$$

Moreover, by employing a function

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

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

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

Hence the following always holds:

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

Therefore

$$|nB_{\rm t}| \leq |nB_{\rm t}(\epsilon)| + B^*,$$

where

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

By summing up the above equations,

$$E[|nB_{\rm t}|^{3/2}] \le E[2|nB_{\rm t}(\epsilon)|^{3/2}] + E[2(B^*)^{3/2}].$$

In Lemma 6.5, we prove $E[|nB_{\rm t}(\epsilon)|^{3/2}] < \infty$. By Lemma 6.2 (2) with δ such that $n^\delta = \epsilon\beta n/4$, we have $P(H_{\rm t} > \epsilon\beta/4) \le C'_H/n^3$, hence

$$\begin{aligned} E[(B^*)^{3/2}] &\le E[(B^*)^{3/2}]_{\{H_{\rm t}>\epsilon\beta/4\}} + E[(B^*)^{3/2}]_{\{H_{\rm t}\le\epsilon\beta/4\}} \\ &\le E[(nM_n)^3]^{1/2} E[1]^{1/2}_{\{H_{\rm t}>\epsilon\beta/4\}} \\ &\quad + E[(L_n)^3]^{1/2}_{\{H_{\rm t}\le\epsilon\beta/4\}} < \infty. \end{aligned}$$

Hence $|nB_{\rm t}|$ is AUI. Lastly, we show $nG_{\rm g}$ is AUI. From eq.(6.38),

$$0 \le nG_{\rm g} \le nG_{\rm g}(\epsilon) + 2n\ s(n)\ \overline{K}.$$

Moreover, $nG_{\rm g} \le n\overline{K}$ always, by definition. Therefore

$$nG_{\rm g} \le nG_{\rm g}(\epsilon) + K^*$$

where

$$\begin{aligned} K^* &\equiv \begin{cases} n\overline{K} & (nH_{\rm t} > n^{2/3}) \\ \overline{K}\ n\ s(n) & (nH_{\rm t} \le n^{2/3}) \end{cases} \\ &\le \begin{cases} n\overline{K} & (nH_{\rm t} > n^{2/3}) \\ \overline{K}\ e^{-n\beta\epsilon/3} & (nH_{\rm t} \le n^{2/3}). \end{cases} \end{aligned}$$

Then

$$0 \le E[(nG_{\rm g})^{3/2}] \le E[2(nG_{\rm g}(\epsilon))^{3/2}] + E[2(K^*)^{3/2}].$$

It is proven in Lemma 6.5 that $E[(nG_{\rm g}(\epsilon))^{3/2}] < \infty$. By Lemma 6.2 (2) with $\delta = 2/3$, we have $P(nH_{\rm t} > n^{2/3}) \le C_H/n^2$, hence

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

Hence $nG_{\rm g}$ is AUI. Since $E[(nH_{\rm t})^3] < \infty$, $E[(nB_{\rm t})^{3/2}] < \infty$, and $E[(nG_{\rm g})^{3/2}] < \infty$, all four errors are also AUI by Lemma 6.1. □

Based on Lemma 6.3, $B_{\rm g}(\epsilon)$, $B_{\rm t}(\epsilon)$, $G_{\rm g}(\epsilon)$, and $G_{\rm t}(\epsilon)$ are the major parts of the four errors when $n \to \infty$. The region in the parameter set to be studied is

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

for a sufficiently small $\epsilon > 0$. Since W_ϵ contains singularities of $K(w) = 0$, we need Theorems 6.1 and 6.5. Let us define the supremum norm by

$$\|f\| = \sup_{u \in \mathcal{M}} |f(u)|.$$

There exists an $L^s(q)$-valued analytic function $\mathcal{M} \ni u \mapsto a(x, u) \in L^s(q)$ such that, in each local coordinate,

$$\begin{aligned} f(x, g(u)) &= a(x, u)\, u^k, \\ E_X[a(X, u)] &= u^k, \\ K(g(u)) = 0 \Rightarrow\ & E_X[a(X, u)^2] = 2, \\ E_X[\|a(X)\|^s] &< \infty. \end{aligned}$$

We define $\|a(X)\| = \sup_{u \in \mathcal{M}} |a(X, u)|$. An empirical process $\xi_n(u)$ is defined by eq.(6.11). Then the empirical process satisfies the following lemma.

Lemma 6.4 *(1) Let $s = 6$. The empirical process $\xi_n(u)$ satisfies*

$$E[\|\xi_n\|^s] < \text{const.} < \infty$$
$$E[\|\nabla \xi_n\|^s] < \text{const.} < \infty$$

where the constant does not depend on n, and $\|\nabla \xi_n\| = \sum_{j=1}^{d} \|\partial_j \xi_n\|$.
(2) The random variable $nH_{\mathrm{t}}(\epsilon)$ is asymptotically uniformly integrable.

Proof of Lemma 6.4 (1) This lemma is derived from Theorem 5.8 and fundamental condition (I). (2) is immediately derived from (1). □

Let the Banach space of uniformly bounded and continuous functions on $\mathcal{M}$ be

$$B(\mathcal{M}) = \{f(u)\ ;\ \|f\| < \infty\}.$$

Since $\mathcal{M}$ is compact, $B(\mathcal{M})$ is a separable norm space. The empirical process $\xi_n(u)$ defined on $B(\mathcal{M})$ weakly converges to the tight Gaussian process $\xi(u)$.

Definition 6.7 (Integral over parameters) Let $\xi(u)$ be an arbitrary function on $\mathcal{M}$ of class C^1. We define the mean of $f(u)$ over $\mathcal{M}$ for a given $\xi(u)$ by

$$E_u^\sigma[f(u)|\xi] = \frac{\displaystyle\sum_\alpha \int_{[0,b]^d} f(u)\, Z(u, \xi)\, du}{\displaystyle\sum_\alpha \int_{[0,b]^d} Z(u, \xi)\, du},$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

where $\sum_\alpha$ is the summation over all coordinates of $\mathcal{M}$, $0 \le \sigma \le 1$, and

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

Based on this definition of $E_u^\sigma[\ \ |\xi]$ and the standard form of the log likelihood ratio function, the major parts of the four errors are given by the case $\sigma = 0$,

$$B_{\mathrm{g}}(\epsilon) = E_X\Big[-\log E_u^0[e^{-a(X,u)u^k}|\xi_n]\Big], \tag{6.39}$$

$$B_{\mathrm{t}}(\epsilon) = \frac{1}{n}\sum_{j=1}^{n} - \log E_u^0[e^{-a(X_j,u)u^k}|\xi_n], \tag{6.40}$$

$$G_{\mathrm{g}}(\epsilon) = E_u^0[u^{2k}|\xi_n], \tag{6.41}$$

$$G_{\mathrm{t}}(\epsilon) = E_u^0\Big[u^{2k} - \frac{1}{\sqrt{n}}\ u^k \xi_n(u)\Big|\xi_n\Big]. \tag{6.42}$$

Lemma 6.5 *Assume that $k_1 > 0$, where k_1 is the first coefficient of the multi-index $k = (k_1, k_2, \ldots, k_d)$, and that $0 \le \sigma \le 1$.*
(1) For an arbitrary real analytic function $\xi(u)$ and $a(x, u)$,

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

where $\partial_1 = (\partial/\partial u_1)$, $c_1, c_2 > 0$ are constants.
(2) For the empirical process $\xi_n(u)$,

$$E\big[E_u^\sigma[n\ u^{2k}|\xi_n]\big] < \infty,$$
$$E\big[E_u^\sigma[n^{3/2}\ u^{3k}|\xi_n]\big] < \infty.$$

(3) Random variables $nG_{\mathrm{g}}(\epsilon)$ ard $nB_{\mathrm{t}}(\epsilon)$ are asymptotically uniformly integrable.

Proof of Lemma 6.5 (1) Let $0 \le p \le 3$. We use the notation $g(u) = u_2^{k_2} \cdots u_d^{k_d}$ and $h(u) = u_2^{h_2} \cdots u_d^{h_d}$, which do not depend on u_1. Then

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

$$f(u) = \beta\sqrt{n} u^k \xi(u) - \sigma u^k a(X, u).$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

By eq.(6.14) and given $\phi(u) > 0$, for each $0 \leq p \leq 3$, there exists a constant $c_p > 0$ such that

$$0 \leq E_u^{\sigma}[u^{pk}|\xi] \leq c_p \frac{N_p}{N_0}.$$

By applying partial integration to N_p and using $q = (p-2)k_1 + h_1 + 1$,

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

By the relation

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

and the Cauchy–Schwarz inequality, since $u \in [0, b]^d$, there exists $B > 0$ such that

$$\begin{aligned}
|u_1 \partial_1 f(u)| \leq \frac{\beta n k_1 u^{2k}}{2} + B(\|\xi\|^2 + \|\partial_1 \xi\|^2 \\
+ \|a\| + \|\partial_1 a\|).
\end{aligned}$$

Hence, by $B' = \max\{B, q\}$,

$$\begin{aligned}
\frac{N_p}{N_0} \leq \frac{N_p}{4N_0} + B'(1 + \|\xi\|^2 + \|\partial_1 \xi\|^2 \\
+ \|a\| + \|\partial_1 a\|) \frac{N_{p-2}}{N_0}.
\end{aligned} \tag{6.43}$$

The case $p = 2$ shows the first half of (1). For the latter half, By eq.(6.43) with $p = 3$, using $B'' = 4B'/3$,

$$\frac{N_3}{N_0} \leq B''(1 + \|\xi\|^2 + \|\partial_1 \xi\|^2 + \|a\| + \|\partial_1 a\|) \frac{N_1}{N_0}.$$

Since $N_1/N_0 \leq (N_2/N_0)^{1/2}$ by the Cauchy–Schwarz inequality, there exists $B''' > 0$ such that

$$\frac{N_3}{N_0} \leq B'''(1 + \|\xi\|^2 + \|\partial_1 \xi\|^2 + \|a\| + \|\partial_1 a\|)^{3/2}.$$

In general $(\sum_{i=1}^n |a_i|^2/n)^{1/2} \leq (\sum_{i=1}^n |a_i|^3/n)^{1/3}$, so the latter half of (1) is obtained.

(2) By Lemma 6.4 and the result of (1) of this lemma, part (2) is immediately derived.

(3) By the definition, $nG_{\mathrm{g}}(\epsilon) = E_u^0[n\, u^{2k}|\xi_n]$. Then from (2) of this lemma, $nG_{\mathrm{g}}(\epsilon)$ is asymptotically uniformly integrable (AUI). Let us prove $nB_{\mathrm{t}}(\epsilon)$ is AUI. By using the notation

$$b(s) = -\sum_{j=1}^{n} \log E_u^0[e^{-s\, a(X_j,u)u^k}],$$

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

$$\begin{aligned} nB_{\mathrm{t}}(\epsilon) &= b(1) = b(0) + b'(0) + \tfrac{1}{2} b''(s^*) \\ &= B_1 + B_2, \end{aligned}$$

where

$$B_1 = \sum_{j=1}^{n} E_u^0[a(X_j, u)u^k|\xi_n],$$

$$B_2 = \frac{1}{2n} \sum_{j=1}^{n} E_u^{s^*}[a(X_j, u)^2\; n\; u^{2k}|\xi_n]\Big|_{X=X_j}.$$

The first term $B_1 = nG_{\mathrm{t}}(\epsilon)$. From Lemma 6.1,

$$-\tfrac{1}{4} nH_{\mathrm{t}}(\epsilon) \leq nG_{\mathrm{t}}(\epsilon) \leq \tfrac{1}{2}(3nG_{\mathrm{g}}(\epsilon) + nH_{\mathrm{t}}(\epsilon)).$$

Therefore $E[|B_1|^{3/2}] < \infty$, because $E[(nG_{\mathrm{g}}(\epsilon))^{3/2}] < \infty$ and $E[(nH_{\mathrm{t}}(\epsilon))^3] < \infty$. Moreover,

$$|B_2|^{3/2} \leq \frac{1}{n} \sum_{j=1}^{n} \|a(X_j)^3\| \Big(E_u^{s^*}[n\; u^{2k}|\xi_n]\Big|_{X=X_j} \Big)^{3/2}.$$

By the statements (1) and (2) of this lemma, $E[|B_2|^{3/2}] < \infty$, therefore $nB_{\mathrm{t}}(\epsilon)$ is AUI. □

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Without loss of generality, for each local coordinate, we can assume $u = (x, y)$ $x \in \mathbb{R}^r$, $y \in \mathbb{R}^{r'}$ $(r' = d - r)$, $k = (k, k')$, $h = (h, h')$, and

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

We define $\mu = h' - 2k'\lambda_\alpha \in \mathbb{R}^{r'}$; then

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

hence y^μ is integrable in $[0, b]^{r'}$. Both λ_α and r depend on the local coordinate. Let λ be the smallest λ_α and m be the largest r among the coordinates in which $\lambda = \lambda_\alpha$. Then $(-\lambda)$ and m are respectively equal to the largest pole and its order of the zeta function, as is shown in Definition 6.4. Let α^* be the index of the set of coordinates which satisfy $\lambda_\alpha = \lambda$ and $r = m$. As is shown in Lemma 6.6, only coordinates M_{α^*} affect the four errors. Let $\sum_{\alpha^*}$ be the sum of such coordinates. For a given function $f(u)$, we use the notation $f_0(y) = f(0, y)$. Also $a_0(X, y) = a(X, 0, y)$.

Definition 6.8 The expectation of a function $f(y, t)$ for a given function $\xi(u)$ on the essential family of local coordinates is defined by

$$E_{y,t}[f(y,t)|\xi] = \frac{\displaystyle\sum_{\alpha^*} \int_0^\infty dt \int dy\ f(y,t)\ Z_0(y,t,\xi)}{\displaystyle\sum_{\alpha^*} \int_0^\infty dt \int dy\ Z_0(y,t,\xi)},$$

where $\int dy$ stands for $\int_{[0,b]^{d-m}} dy$ and

$$Z_0(y, t, \xi) = \gamma_b y^\mu\ t^{\lambda-1} e^{-\beta t + \beta\sqrt{t}\ \xi_0(y)} \phi_0^*(y).$$

Here $\gamma_b > 0$ is a constant defined by eq.(4.16).

Lemma 6.6 *Let $p \geq 0$ be a constant. There exists $c_1 > 0$ such that, for an arbitrary C^1-class function $f(u)$ and analytic function $\xi(u)$, the following inequality holds:*

$$\left| E_u^0[(n\ u^{2k})^p\ f(u)|\xi] - \ E_{y,t}[t^p f_0(y)|\xi] \right| \leq \frac{D(\xi, f, \phi^*)}{\log n},$$

where

$$D(\xi, f, \phi^*) \equiv \frac{c_1 e^{2\beta\|\xi\|^2} \|\phi^*\|}{(\min \phi^*)^2} \{\beta \|\nabla\xi\| \|f\phi^*\| + \|\nabla(f\phi^*)\| + \|f\phi^*\|\}$$

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

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Proof of Lemma 6.6 Using Z^p and Y^p in Definition 4.10 and eq.(4.19), we define A, B, and C by

$$A \equiv E_u^0[(n\ u^{2k})^p\ f(u)|\xi] = \frac{\sum_\alpha n^p Z^p(n,\xi,f\phi^*)}{\sum_\alpha Z^0(n,\xi,f\phi^*)},$$

$$B \equiv E_{y,t}[t^p f_0(y)|\xi] = \frac{\sum_{\alpha^*} n^p Y^p(n,\xi,f\phi^*)}{\sum_{\alpha^*} Y^0(n,\xi,f\phi^*)},$$

$$C \equiv \frac{\sum_\alpha n^p Y^p(n,\xi,f\phi^*)}{\sum_\alpha Y^0(n,\xi,f\phi^*)},$$

where $\sum_\alpha$ and $\sum_{\alpha^*}$ denote the sum of all local coordinates and the sum of coordinates in the essential family respectively. To prove the lemma, it is sufficient to show $|A - B| \leq D(\xi, f, \phi^*)/\log n$. Since

$$|A - B| \leq |A - C| + |C - B|,$$

we show the inequalities for $|A - C|$ and $|C - B|$ respectively. The set $(n, \xi, f\phi^*)$ is omitted for simplicity. Firstly, $|A - C|$ is bounded by

$$\begin{aligned} |A - C| &= n^p \left| \frac{\sum_\alpha Z^p}{\sum_\alpha Z^0} - \frac{\sum_\alpha Y^p}{\sum_\alpha Y^0} \right| \\ &= n^p \left| \frac{\sum_\alpha Z^p - \sum_\alpha Y^p}{\sum_\alpha Z^0} + \frac{\sum_\alpha Y^p \{\sum_\alpha Y^0 - \sum_\alpha Z^0\}}{\sum_\alpha Z^0 \sum_\alpha Y^0} \right| \\ &\leq n^p \frac{\sum_\alpha |Z^p - Y^p|}{\sum_\alpha Z^0} + n^p \frac{\sum_\alpha Y^p}{\sum_\alpha Y^0} \times \frac{\sum_\alpha |Y^0 - Z^0|}{\sum_\alpha Z^0}. \end{aligned}$$

For general $a_i, b_i > 0$,

$$\frac{\sum a_i}{\sum b_i} \leq \sum \frac{a_i}{b_i}.$$

Therefore,

$$|A - C| \leq \sum_\alpha \frac{n^p |Z^p - Y^p|}{Z^0} + \sum_\alpha \frac{n^p Y^p}{Y^0} \times \sum_\alpha \frac{|Y^0 - Z^0|}{Z^0}.$$

Then by using Theorems 4.8, 4.9, 4.10, there exist constants $C_1, C_2 > 0$ such that

$$\begin{aligned} |A - C| \leq{}& \frac{C_1 e^{\beta\|\xi\|^2}}{\log n} \frac{\{\beta \|\nabla\xi\| \|f\phi^*\| + \|\nabla(f\phi^*)\| + \|f\phi^*\|\}}{\min \phi^*} \\ &+ \frac{C_2 e^{2\beta\|\xi\|^2}}{\log n} \frac{\|\phi^*\| \{\beta \|\nabla\xi\| \|f\phi^*\| + \|\nabla(f\phi^*)\| + \|f\phi^*\|\}}{(\min \phi^*)^2}. \end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Secondly,

$$|C - B| = n^p \left| \frac{\sum_\alpha Y^p}{\sum_\alpha Y^0} - \frac{\sum_{\alpha^*} Y^p}{\sum_{\alpha^*} Y^0} \right|.$$

Let us use the simplified notation,

$$T^p = \sum_{\alpha^*} Y^p,$$

$$U^p = \sum_{\alpha \backslash \alpha^*} Y^p.$$

Then, by $\sum_\alpha = \sum_{\alpha^*} + \sum_{\alpha\backslash\alpha^*}$,

$$\begin{aligned} |C - B| &= n^p \left| \frac{T^p + U^p}{T^0 + U^0} - \frac{T^p}{T^0} \right| \\ &\leq \frac{n^p U^p}{T^0} + \frac{n^p U^0 T^p}{(T^0)^2}. \end{aligned}$$

By Theorem 4.10, there exists $C_3 > 0$ such that

$$|C - B| \leq \frac{C_3 e^{\beta\|\xi\|^2} \|\phi^*\|}{\min \phi^*} + \frac{C_4 e^{2\beta\|\xi\|^2} \|\phi^*\|^2}{(\min \phi^*)^2}.$$

By combining two results, we obtain the lemma. □

6.3.3 Proof of the theorems

In this subsection, we prove the theorems.

Definition 6.9 (Explicit representation of the Bayes quartet) Four functionals of a given function $\xi(u)$ are defined by

$$B_{\rm g}^*(\xi) \equiv \tfrac{1}{2} E_X[\ E_{y,t}[a_0(X, y) t^{1/2} |\xi]^2\ ], \tag{6.44}$$

$$B_{\rm t}^*(\xi) \equiv G_{\rm t}^*(\xi) - G_{\rm g}^*(\xi) + B_{\rm g}^*(\xi), \tag{6.45}$$

$$G_{\rm g}^*(\xi) \equiv E_{y,t}[t|\xi], \tag{6.46}$$

$$G_{\rm t}^*(\xi) \equiv E_{y,t}[t - t^{1/2}\xi_0(y)|\xi]. \tag{6.47}$$

Note that these four functionals do not depend on n. If $\xi(u)$ is a random process, then the four functionals are random variables. The singular fluctuation is defined by

$$\nu(\beta) = \tfrac{1}{2} E_\xi \Big[E_{y,t}[t^{1/2}\xi_0(y)|\xi] \Big]. \tag{6.48}$$

Proof of Theorems 6.8 In this proof, we use the simplified notation

$$E_u^\sigma[f(u)] = E_u^\sigma[f(u)|\xi_n],$$

$$E_{y,t}[f(y,t)] = E_{y,t}[f(y,t)|\xi_n],$$

in other words, '$|\xi_n$' is omitted. Firstly we prove the following convergences in probability.

$$n B_{\mathrm{g}}(\epsilon) - B_{\mathrm{g}}^*(\xi_n) \to 0, \tag{6.49}$$

$$n B_{\mathrm{t}}(\epsilon) - B_{\mathrm{t}}^*(\xi_n) \to 0, \tag{6.50}$$

$$n G_{\mathrm{g}}(\epsilon) - G_{\mathrm{g}}^*(\xi_n) \to 0, \tag{6.51}$$

$$n G_{\mathrm{t}}(\epsilon) - G_{\mathrm{t}}^*(\xi_n) \to 0. \tag{6.52}$$

Based on eq.(6.41), eq.(6.46), and Lemma 6.6 with $p = 1$,

$$\begin{aligned}|n G_{\mathrm{g}}(\epsilon) - G_{\mathrm{g}}^*(\xi_n)| &= \left|E_u^0[nu^{2k}] - E_{y,t}[t]\right| \\ &\leq \frac{D(\xi_n, 1, \phi^*)}{\log n}.\end{aligned}$$

Because the convergence in law $\xi_n \to \xi$ holds, eq.(6.51) is obtained. Also, based on eq.(6.42), eq.(6.47), and Lemma 6.6 with $p = 1, \frac{1}{2}$,

$$\begin{aligned}|n G_{\mathrm{t}}(\epsilon) - G_{\mathrm{t}}^*(\xi_n)| &= \left|E_u^0[nu^{2k} - \sqrt{n}u^k\xi_n] - E_{y,t}[t - t^{1/2}\xi_0]\right| \\ &\leq \frac{D(\xi_n, 1, \phi^*)}{\log n} + \frac{D(\xi_n, \xi_n, \phi^*)}{\log n}.\end{aligned}$$

Because the convergence in law $\xi_n \to \xi$ holds, eq.(6.52) is obtained. Let us prove eq.(6.49); we define

$$b_{\mathrm{g}}(\sigma) \equiv E_X\Big[-\log E_u^0[e^{-\sigma a(X,u)u^k}]\Big],$$

then it follows that $n B_{\mathrm{g}}(\epsilon) = n b_{\mathrm{g}}(1)$ and there exists $0 < \sigma^* < 1$ such that

$$n B_{\mathrm{g}}(\epsilon) = n b_{\mathrm{g}}(0) + n b_{\mathrm{g}}'(0) + \frac{n}{2} b_{\mathrm{g}}''(0) + \frac{n}{6} b_{\mathrm{g}}^{(3)}(\sigma^*) \tag{6.53}$$

$$\begin{aligned}&= n E_u^0[u^{2k}] - \frac{n}{2} E_X E_u^0[a(X,u)^2 u^{2k}] \\ &\quad + \frac{n}{2} E_X E_u^0[a(X,u)u^k]^2 + \tfrac{1}{6} n b_{\mathrm{g}}^{(3)}(\sigma^*),\end{aligned} \tag{6.54}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

where we used $b_{\mathrm{g}}(0)=0$, and $E_X[a(X,u)]=u^k$ hence $b'_{\mathrm{g}}(0)=E_u^0[u^{2k}]$. The first term on the right-hand side of eq.(6.54) is equal to $nG_{\mathrm{g}}(\epsilon)$. By Lemma 6.6, the following convergence in probability

$$\begin{aligned}&\left|nE_XE_u^0[a(X,u)^2u^{2k}]-E_XE_{y,t}[a_0(X,y)^2t]\right|\\&\leq\frac{E_X[D(\xi,a(X,u)^2,\phi^*)]}{\log n}\to 0\end{aligned}\tag{6.55}$$

holds, where $D(\beta,\xi,a(X,u),\phi^*)$ is defined as in Lemma 6.6. Since $E_X[a_0(X,y)^2]=2$, the sum of the first two terms on the right-hand side of eq.(6.54) converges to zero in probability. For the third term, by using the notation

$$\begin{aligned}\rho(u,v)&=E_X[a(X,u)a(X,v)],\\\rho_0(u,y)&=\rho(u,(0,y)),\\\rho_{00}(y',y)&=\rho((0,y'),(0,y)),\end{aligned}$$

and applying Lemma 6.6,

$$\begin{aligned}&\left|nE_XE_u^0[a(X,u)u^k]^2-E_{y,t}[a_0(X,y)t^{1/2}]^2\right|\\&\leq\left|\sqrt{n}E_u^0\Big[u^k\big(\sqrt{n}E_v^0[\rho(u,v)v^k]-E_{y,t}[\rho_0(u,y)t^{1/2}]\big)\Big]\right|\\&\quad+\left|E_{y,t}\Big[t^{1/2}\big(\sqrt{n}E_u^0[\rho_0(u,y)u^k]-E_{y',t'}[\rho_{00}(y',y)(t't)^{1/2}]\big)\Big]\right|\\&\leq\frac{c_1\sqrt{n}}{\log n}E_u^0[u^k]\;D(\xi_n,\rho(\cdot,\cdot),\phi^*)\\&\quad+\frac{c_1}{\log n}E_{y,t}[t^{1/2}D(\xi_n,\rho(\cdot,y),\phi^*)].\end{aligned}\tag{6.56}$$

Equation (6.56) converges to zero in probability by Lemma 6.5. Therefore the difference between the third term and $B_{\mathrm{g}}^*(\xi_n)$ converges to zero in probability. For the last term, we have

$$\begin{aligned}\left|nb_{\mathrm{g}}^{(3)}(\sigma^*)\right|&=n\Big|E_X\Big\{E_u^{\sigma^*}[a(X,u)^3u^{3k}]+2E_u^{\sigma^*}[a(X,u)u^k]^3\\&\qquad-3E_u^{\sigma^*}[a(X,u)^2u^{2k}]E_u^{\sigma^*}[a(X,u)u]\Big\}\Big|\\&\leq 6nE_X\Big[\|a(X)\|^3\;E_u^{\sigma^*}[u^{3k}]\Big],\end{aligned}$$

where we used Hölder's inequality. By applying Lemma 6.5,

$$\begin{aligned}\left|nb_{\rm g}^{(3)}(\sigma^*)\right| \le \frac{6c_2}{n^{1/2}} E_X\Big[\|a(X)\|^3\,\{1+\|\xi_n\|^3+\|\partial\xi_n\|^3 \\ +\|a(X)\|^{3/2}+\|\partial a(X)\|^{3/2}\}\Big],\end{aligned} \tag{6.57}$$

which shows $nb_{\rm g}^{(3)}(\sigma^*)$ converges to zero in probability, because the fundamental condition (I) with index $s=6$ is assumed. Hence eq.(6.49) is proved. We proceed to the proof of eq.(6.50). An empirical expectation $E_j^*[\ ]$ is simply denoted by

$$E_j^*[f(X_j)] = \frac{1}{n}\sum_{j=1}^{n} f(X_j).$$

By defining

$$b_{\rm t}(\sigma) = E_j^*[-\log E_u^0[e^{-\sigma a(X_j,u)u^k}]],$$

it follows that $nB_{\rm t}(\epsilon)=nb_{\rm t}(1)$ and there exists $0<\sigma^*<1$ such that

$$\begin{aligned}nB_{\rm t}(\epsilon) = nG_{\rm t}(\epsilon) - \frac{n}{2}E_j^*E_u^0[a(X_j,u)^2u^{2k}] \\ +\frac{n}{2}E_j^*E_u^0[a(X_j,u)u^k]^2+\tfrac{1}{6}nb_{\rm t}^{(3)}(\sigma^*).\end{aligned} \tag{6.58}$$

Then, by applying Lemma 6.5, $nb_{\rm t}^{(3)}(\sigma^*)$ converges to zero in probability in the same way as eq.(6.57). In fact,

$$\begin{aligned}\left|nb_{\rm t}^{(3)}(\sigma^*)\right| &= \Big|E_j^*\Big\{E_u^{\sigma^*}[a(X_j,u)^3u^{3k}]+2E_u^{\sigma^*}[a(X_j,u)u^k]^3 \\ &\qquad -3E_u^{\sigma^*}[a(X_j,u)^2u^{2k}]E_u^{\sigma^*}[a(X_j,u)u]\Big\}\Big| \\ &\le 6nE_j^*\Big[\|a(X_j)\|^3\ E_u^{\sigma^*}[u^{3k}]\Big].\end{aligned}$$

By applying Lemma 6.5, using the fundamental condition (7) with $5=6$,

$$\begin{aligned}\left|nb_{\rm t}^{(3)}(\sigma^*)\right| \le \frac{6c_2}{n^{1/2}} E_j^*\Big[\|a(X_j)\|^3\,\{1+\|\xi_n\|^3+\|\partial\xi_n\|^3 \\ +\|a(X_j)\|^{3/2}+\|\partial a(X_j)\|^{3/2}\}\Big],\end{aligned} \tag{6.59}$$

which converges to zero in probability. By the same methods as eq.(6.55) and eq.(6.56), replacing respectively $E_X[\|a(X)^2\|]$ with $E_j^*\|a(X_j)^2\|$ and $\rho(u,v)$

with $\rho_n(u, v) = E_j^* a(X_j, u) a(X_j, v)$,

$$\begin{aligned}
&\Big|\frac{n}{2} E_j^* E_u^0[a(X_j, u)^2 u^{2k}] - G_{\mathrm{g}}^*(\xi_n)\Big| \\
&\le \frac{n}{2}\Big| E_j^* E_u^0[a(X_j, u)^2 u^{2k}] - E_X E_u^0[a(X, u)^2 u^{2k}]\Big| \\
&\quad + \Big|\frac{n}{2} E_X E_u^0[a(X, u)^2 u^{2k}] - G_{\mathrm{g}}^*(\xi_n)\Big| \\
&\le \Big(\sup_u |E_j^* a(X_j, u) - E_X a(X, u)|\Big)\, \frac{n}{2} E_u^0[u^{2k}] \\
&\quad + \Big|\frac{n}{2} E_X E_u^0[a(X, u)^2 u^{2k}] - G_{\mathrm{g}}^*(\xi_n)\Big|,
\end{aligned}$$

which converges to zero by Lemma 6.5 and eq.(6.55). In the same way, the following convergence in probability holds,

$$\tfrac{1}{2} E_j^* E_u^0[a(X_j, u) u^k]^2 - B_{\mathrm{g}}^*(\xi_n) \to 0,$$

and therefore the following convergence in probability also holds:

$$n B_{\mathrm{t}}(\epsilon) - n G_{\mathrm{t}}(\epsilon) + n G_{\mathrm{g}}(\epsilon) - n B_{\mathrm{g}}(\epsilon) \to 0. \tag{6.60}$$

Therefore eq.(6.50) is obtained. By combining eq.(6.49)–eq.(6.52) with Lemma 6.3 (2), we obtain the following convergences in probability:

$$n B_{\mathrm{g}} - B_{\mathrm{g}}^*(\xi_n) \to 0, \tag{6.61}$$
$$n B_{\mathrm{t}} - B_{\mathrm{t}}^*(\xi_n) \to 0, \tag{6.62}$$
$$n G_{\mathrm{g}} - G_{\mathrm{g}}^*(\xi_n) \to 0, \tag{6.63}$$
$$n G_{\mathrm{t}} - G_{\mathrm{t}}^*(\xi_n) \to 0. \tag{6.64}$$

The four functionals $B_{\mathrm{g}}^*(\xi)$, $B_{\mathrm{t}}^*(\xi)$, $G_{\mathrm{g}}^*(\xi)$, and $G_{\mathrm{t}}^*(\xi)$ are continuous functions of $\xi \in B(\mathcal{M})$. From the convergence in law of the empirical process $\xi_n \to \xi$, these convergences in law

$$\begin{aligned}
B_{\mathrm{g}}^*(\xi_n) \to B_{\mathrm{g}}^*(\xi), \quad B_{\mathrm{t}}^*(\xi_n) \to B_{\mathrm{t}}^*(\xi), \\
G_{\mathrm{g}}^*(\xi_n) \to G_{\mathrm{g}}^*(\xi), \quad G_{\mathrm{t}}^*(\xi_n) \to G_{\mathrm{t}}^*(\xi),
\end{aligned}$$

are derived. Therefore Theorem 6.8 (1) and (2) are obtained. Theorem 6.8 (3) is shown because the four errors are asymptotically uniformly integrable by Lemma 6.3. □

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Proof of Main Theorem 6.3 Before proving the theorem, we introduce a property of a Gaussian process. We use the notation

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

$$\int du^* = \sum_{\alpha^*} \gamma_b \int dx\ dy\ \delta(x)\ y^{\mu},$$

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

where $u = (x, y)$. Then, by Definition 6.9,

$$E[B_{\mathrm{g}}^*] = \frac{1}{2\beta^2} E\Big[E_X\Big[\Big(\frac{\int du^* a(X,u) S_\lambda'(\xi(u))}{Z(\xi)}\Big)^2\Big]\Big],$$

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

$$E[G_{\mathrm{g}}^*] = \frac{1}{\beta^2} E\Big[\frac{\int du^* S_\lambda''(\xi(u))}{Z(\xi)}\Big],$$

$$E[G_{\mathrm{t}}^*] = \frac{1}{\beta^2} E\Big[\frac{\int du^* S_\lambda''(\xi(u))}{Z(\xi)}\Big] - \frac{1}{\beta} E\Big[\frac{\int du^*\ \xi(u) S_\lambda'(\xi(u))}{Z(\xi)}\Big].$$

Let $\nu = \nu(\beta)$ be the singular fluctuation in eq.(6.48) and A be a constant,

$$\nu = \frac{1}{2\beta} E\Big[\frac{\int du^*\ \xi(u) S_\lambda'(\xi(u))}{Z(\xi)}\Big], \tag{6.65}$$

$$A = \frac{1}{\beta^2} E\Big[\frac{\int du^* S_\lambda''(\xi(u))}{Z(\xi)}\Big]. \tag{6.66}$$

By eq.(5.24) in Theorem 5.11 and $\rho(u, u) = E_X[a(X, u)^2] = 2$ with $u = (0, y)$, we have

$$E[B_g^*] = A - \frac{1}{\beta}\nu, \tag{6.67}$$

$$E[B_t^*] = A - \Big(2 + \frac{1}{\beta}\Big)\nu, \tag{6.68}$$

$$E[G_g^*] = A, \tag{6.69}$$

$$E[G_t^*] = A - 2\nu. \tag{6.70}$$

By combining these equations to eliminate A and ν, we obtain two equations which do not contain either A or ν, giving Main Theorem 6.3. □

Proof of Theorem 6.9 By eq.(5.22) with $a = \xi_n(u)$

$$\frac{2}{\beta^2}\frac{S_\lambda''(\xi_n(u))}{S_\lambda(\xi_n(u))} - \frac{1}{\beta}\frac{\xi_n(u)S_\lambda'(\xi_n(u))}{S_\lambda(\xi_n(u))} - \frac{2\lambda}{\beta} = 0.$$

By the definitions of $G_{\mathrm{g}}^*(\xi_n)$ and $G_{\mathrm{t}}^*(\xi_n)$,

$$G_{\mathrm{g}}^*(\xi_n) + G_{\mathrm{t}}^*(\xi_n) - \frac{2\lambda}{\beta} = 0.$$

By the convergences in law $nG_{\mathrm{g}} \to G_{\mathrm{g}}^*(\xi)$ and $nG_{\mathrm{t}} \to G_{\mathrm{t}}^*(\xi)$, Theorem 6.9 is obtained. □

Proof of Theorem 6.10 Let $\nu = \nu(\beta)$ be the singular fluctuation in eq.(6.48). By eq.(5.23) in Theorem 5.11, eq.(6.66), and eq.(6.65),

$$A = \frac{\lambda}{\beta} + \nu(\beta).$$

Hence from eqs.(6.67)–(6.70), we obtain eqs.(6.27)–(6.30). From the definition in eq.(6.24),

$$V = nE_j^* E_u^0[a(X_j, u)^2 u^{2k}] - nE_j^* E_u^0[a(X_j, u)u^k]^2 + o_p(1),$$

where $o_p(1)$ is a random variable which converges to zero in probability. Then, based on eq.(6.58) in the proof of Theorem 6.8, the following convergence in probability holds,

$$V - 2(G_{\mathrm{t}}^*(\xi_n) - B_{\mathrm{t}}^*(\xi_n)) \to 0,$$

which gives eqs.(6.25) and (6.26). Therefore V converges in law and is asymptotically uniformly integrable, so, when $n \to \infty$,

$$E[V] \to \frac{2\nu(\beta)}{\beta}.$$

Let us introduce an expectation $\langle\ \rangle$ defined by

$$\langle f(u,t)\rangle = \frac{\int du^*\ f(u,t)\ t^{\lambda-1}\ e^{-\beta t + \beta\sqrt{t}\xi(u)}}{\int du^*\ t^{\lambda-1}\ e^{-\beta t+\beta\sqrt{t}\xi(u)}}. \tag{6.71}$$

Then

$$A = E_\xi[\langle t\rangle],$$
$$\nu(\beta) = \tfrac{1}{2}E_\xi[\langle\sqrt{t}\xi(u)\rangle].$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

By using the Cauchy–Schwarz inequality,

$$\begin{aligned}E_\xi[\langle \sqrt{t}\xi(u)\rangle] &\le E_\xi[\langle t\rangle]^{1/2} E_\xi[\langle \xi(u)^2\rangle]^{1/2}\\ &\le \sqrt{A}\ E_\xi[\|\xi\|^2]^{1/2}.\end{aligned}$$

By combining this inequality with

$$A = \frac{\lambda}{\beta} + \nu(\beta) \le \frac{\lambda}{\beta} + \frac{\sqrt{A}}{2} E_\xi[\|\xi\|^2]^{1/2},$$

we obtain

$$\sqrt{A} \le \frac{E_\xi[\|\xi\|^2]^{1/2}}{4} + \sqrt{E_\xi[\|\xi\|^2]/16 + \lambda/\beta},$$

which completes the proof. □

Remark 6.13 (Singular fluctuation) By using the expectation notation defined in eq.(6.71) we can represent the variance of $a_0(X, y)$,

$$\begin{aligned}\nu(\beta) &= \tfrac{1}{2} E_\xi[\langle \sqrt{t}\xi_0(y)\rangle]\\ &= \frac{\beta}{2} E_\xi E_X\Big[\Big\langle \Big(\sqrt{t}a_0(X, y)\Big)^2 \Big\rangle - \Big\langle \sqrt{t}a_0(X, y)\Big\rangle^2 \Big].\end{aligned}$$

Note that

$$a(x, w) = \frac{\log(q(x)/p(x|w)) - K(w)}{\sqrt{K(w)}}.$$

Although $a(x, w)$ is not well defined at a singularity in the original parameter space, it can be made well-defined on the manifold by resolution of singularities. Both the real log canonical threshold λ and the singular fluctuation $\nu(\beta)$ determine the asymptotic behavior of a statistical model. In regular statistical models, $\lambda = \nu(\beta) = d/2$, where d is the dimension of the parameter space, whereas, in singular statistical models, λ and $\nu(\beta)$ are different from $d/2$ in general.

6.4 Maximum likelihood and *a posteriori*

In this section, we study the estimator $\hat{w}$ which minimizes

$$-\sum_{i}^{n} \log p(X_i|w) + a_n\sigma(w)$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

in a compact set W, which is equal to the parameter that minimizes

$$R_n^0(w) = nK_n(w) + a_n\sigma(w).$$

We assume that $\sigma(w)$ is a C^2-class function of w in an open set which contains W, and that $\{a_n \geq 0\}$ is a nondecreasing sequence. We can assume $\sigma(w) \geq 0$ without loss of generality. If $a_n = 0$ for arbitrary n, then $\hat{w}$ is called the maximum likelihood estimator (MLE) and if $a_n = 1$ for arbitrary n and $\sigma(w) = -\log\varphi(w)$, where $\varphi(w)$ is an *a priori* probability density function, then $\hat{w}$ is called the maximum *a posteriori* estimator (MAP). The generalization and training errors are respectively defined by

$$\begin{aligned} R_{\mathrm{g}} &= K(\hat{w}), \\ R_{\mathrm{t}} &= K_n(\hat{w}). \end{aligned}$$

Although the MAP employs an *a priori* distribution, its generalization error is quite different from that of Bayes estimation.

To study the ML or MAP method, we have to analyze the geometry of the parameter space. Let us assume the fundamental conditions (I) and (II) with index $s = 4$. We prove that, for arbitrary $\epsilon > 0$, $P(K(\hat{w}) > \epsilon)$ is sufficiently small that it does not affect the asymptotic generalization and training errors. To study the event $K(\hat{w}) \leq \epsilon$, we use the resolution of singularities and the standard form of the log likelihood ratio function. Then the Kullback–Leibler distance becomes a normal crossing function defined on a local coordinate $[0, b]^d$ of a manifold

$$u^{2k} = u_1^{2k_1} u_2^{2k_2} \cdots u_r^{2k_r},$$

where r is an integer which satisfies $1 \leq r \leq d$ and $k_1, k_2, \ldots, k_r > 0$ are natural numbers. Without loss of generality, we can assume that $b = 1$. For a given $u \in [0, 1]^d$, the integer a is defined by the number $(1 \leq a \leq r)$ which satisfies

$$\frac{u_a^2}{k_a} \leq \frac{u_i^2}{k_i} \quad (i = 1, 2, \ldots, r). \tag{6.72}$$

Intuitively, a is the number on the axis which is farthest from the given point u. A map $[0, 1]^d \ni u \to (t, v)$, where $t \in \mathbb{R}^1$, $v = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$, is defined by

$$\begin{aligned} t &= u^{2k}, \\ v_i &= \begin{cases} \sqrt{u_i^2 - (k_i/k_a)u_a^2} & (1 \leq i \leq r) \\ u_i & (r < i \leq d). \end{cases} \end{aligned}$$

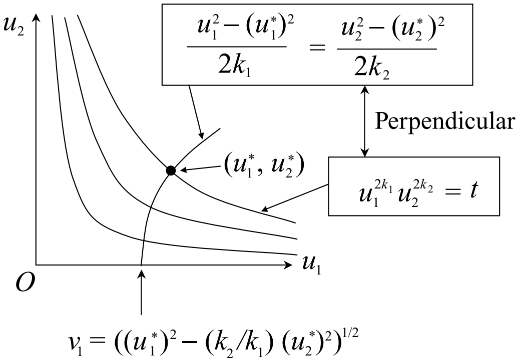
u_2
$\frac{u_1^2-(u_1^*)^2}{2k_1} = \frac{u_2^2-(u_2^*)^2}{2k_2}$
Perpendicular
(u_1^*, u_2^*)
$u_1^{2k_1} u_2^{2k_2} = t$
u_1
O
$v_1 = ((u_1^*)^2 - (k_2/k_1)\,(u_2^*)^2)^{1/2}$

Fig. 6.3. Normal crossing coordinate

Then, by definition, $v_a = 0$. The set V is defined by

$$V \equiv \{v = (v_1, v_2, \ldots, v_d) \in [0, 1]^d;\ v_1 v_2 \cdots v_r = 0\}.$$

Then $[0, 1]^d \ni u \mapsto (t, v) \in T \times V$ is a one-to-one map as in Figure 6.3. Under this correspondence u is identified with (t, v).

Remark 6.14 Note that, if a surface parameterized by t,

$$u_1^{2k_1} u_2^{2k_2} \cdots u_r^{2k_r} = t, \tag{6.73}$$

and the curve parameterized by $(u_1^*, u_2^*, \ldots, u_r^*)$,

$$\frac{u_1^2 - (u_1^*)^2}{2k_1} = \frac{u_2^2 - (u_2^*)^2}{2k_2} = \cdots = \frac{u_r^2 - (u_r^*)^2}{2k_r}, \tag{6.74}$$

have a crossing point, then they are perpendicular to each other in the restricted space $\mathbb{R}^r$ at the crossing point. In fact, the perpendicular vector of eq.(6.73) and the tangent vector at $(u_1, u_2, \ldots, u_r)$ of eq.(6.74) are equal to each other,

$$\left(\frac{k_1}{u_1}, \frac{k_1}{u_2}, \ldots, \frac{k_r}{u_r}\right).$$

The map $u \mapsto (t, v)$ can be understood as a function from a point $(u_1^*, u_2^*, \ldots, u_r^*)$ to the crossing point of eq.(6.74) and $u^{2k} = 0$. It is equal to the limit point of steepest descent dynamics, eq.(1.35).

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Theorem 6.11 *Let $f(u)$ be a function of class C^1 which is defined on an open set that contains $[0,1]^d$. Then as a function of (t, v), $f(t, v)$ satisfies*

$$|f(t,v) - f(0,v)| \leq C\, t^{1/k^*} \|\nabla f\| \quad (0 \leq t < 1),$$

where $k^* = 2(k_1 + \cdots + k_r)$,

$$\|\nabla f\| = \sup_{u \in [0,1]^d} \max_{1 \leq j \leq d} \left| \frac{\partial f}{\partial u_j}(u) \right|,$$

and $C > 0$ is a constant which does not depend on t, s, and f.

Proof of Theorem 6.11 Let $u = (t, v)$ and $u' = (0, v)$. It $t > 0$ then the Jacobian determinant of the map $u \mapsto (t, v)$ is not equal to zero. There exists $u^* \in [0,1]^d$ such that

$$|f(t,v) - f(0,v)| \leq \sum_{j=1}^{r} |u_j - u'_j| \left| \frac{\partial f}{\partial u_j}(u^*) \right|.$$

Hence

$$|f(t,v) - f(0,v)| \leq \|\nabla f\| \sum_{j=1}^{r} |u_j - u'_j|,$$

where

$$\|\nabla f\| = \sum_{j=1}^{d} \max_{u \in [0,1]^d} \left| \frac{\partial f}{\partial u_j} \right|.$$

If $j = a$ then $|u_j - u'_j| = u_a$. If $j \neq a$, then $u_a^2/k_a \leq u_j^2/k_j$,

$$\begin{aligned} |u_j - u'_j| &= \left| u_j - \left(u_j^2 - (k_j/k_a)u_a^2\right)^{1/2} \right| \\ &= \frac{(k_j/k_a)u_a^2}{u_j + \left(u_j^2 - (k_j/k_a)u_a^2\right)^{1/2}} \\ &\leq \sqrt{k_j/k_a}\, u_a. \end{aligned}$$

Hence there exists $C' > 0$ such that

$$|f(t,v) - f(0,v)| \leq C' \|\nabla f\| u_a.$$

On the other hand by eq.(6.72) there exists $C'' > 0$ such that

$$t = u^{2k} \geq C''(u_a)^{2k^*},$$

which completes the theorem. □

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Theorem 6.12 *Assume the fundamental conditions (I) and (II) with index* s ($s \geq 6$) *and that* $a_n/n \to 0$ ($n \to \infty$). *Let* $\psi_n(w)$ *be an empirical process on* $\{w; K(w) > \epsilon\}$,

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

and

$$\|\psi_n\| = \sup_{K(w)>\epsilon} |\psi_n(w)|.$$

Then the following hold.
(1) For a given $\epsilon > 0$, *there exists a constant* $C > 0$, *such that, for arbitrary* $n \geq 1$,

$$P(\|\psi_n\|^2 > n\epsilon) \leq \frac{C}{n^{s/2}},$$
$$P(K(\hat{w}) > \epsilon) \leq \frac{C}{n^{s/2}}.$$

(2) For a given $\epsilon > 0$, *there exists a constant* $C' > 0$, *such that, for arbitrary* $n \geq 1$,

$$E[\|\psi_n\|^2]_{\{\|\psi_n\|^2 > n\epsilon\}} \leq \frac{C'}{n^{s/2-1}},$$
$$E[nK(\hat{w})]_{\{K(\hat{w})>\epsilon\}} \leq \frac{C'}{n^{s/2-1}}.$$

Proof of Theorem 6.12 (1) From Theorem 5.8,

$$E[\|\psi_n\|^s] = C < \infty,$$

and it follows that

$$\begin{aligned} C &\geq E[\|\psi_n\|^s]_{\{\|\psi_n\|^2 > n\epsilon\}} \\ &\geq (n\epsilon)^{s/2} P(\|\psi_n\|^2 > n\epsilon). \end{aligned}$$

Therefore

$$P(\|\psi_n\|^2 > n\epsilon) \leq \frac{C}{(n\epsilon)^{s/2}}. \tag{6.75}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} R_n^0(w) &= nK(w) - \sqrt{nK(w)}\, \psi_n(w) + a_n\sigma(w) \\ &\geq \tfrac{1}{2}(nK(w) - \|\psi_n\|^2) + a_n\sigma(w). \end{aligned}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

A parameter w_0 in the set of true parameters satisfies $K(w_0) = 0$, hence

$$R_n^0(w_0) = a_n \sigma(w_0).$$

Therefore, by the definition of $\hat{w}$, $R_n^0(\hat{w}) \leq R_n^0(w_0)$, and consequently

$$\tfrac{1}{2}(nK(\hat{w}) - \|\psi_n\|^2) + a_n\sigma(\hat{w}) \leq a_n\sigma(w_0).$$

Hence, if $\|\psi_n\|^2 \geq n\epsilon$, there exists a constant $c_1 > 0$ such that

$$nK(\hat{w}) \leq \|\psi_n\|^2 + 4a_n\|\sigma\| \leq c_1\|\psi_n\|^2.$$

Because $a_n/n \to 0$,

$$\begin{aligned} P(K(\hat{w}) > \epsilon) &\leq P(c_1\|\psi_n\|^2 > n\,\epsilon) \\ &\leq \frac{c_2}{(n\epsilon)^{s/2}}. \end{aligned}$$

(2) By using the above results,

$$\begin{aligned} E[\|\psi_n\|^2]_{\{\|\psi_n\|^2 > n\epsilon\}} &\leq \sum_{j=1}^{\infty} E[\|\psi_n\|^2]_{\{jn\epsilon < \|\psi_n\|^2 \leq (j+1)n\epsilon\}} \\ &\leq \sum_{j=1}^{\infty} (j+1)\epsilon n \times \frac{C}{(n\epsilon j)^{s/2}} \\ &\leq \frac{c_3}{(n\epsilon)^{s/2-1}}. \end{aligned}$$

In the same way,

$$\begin{aligned} E[nK(\hat{w})]_{\{K(w) > \epsilon\}} &\leq \sum_{j=1}^{\infty} E[nK(\hat{w})]_{\{jn\epsilon < nK(\hat{w}) \leq (j+1)n\epsilon\}} \\ &\leq \sum_{j=1}^{\infty} (j+1)\epsilon n \times \frac{c_2}{(n\epsilon j)^{s/2}} \\ &\leq \frac{c_4}{(n\epsilon)^{s/2-1}}. \end{aligned}$$

□

Remark 6.15 (Consistency of estimation) By Theorem 6.12, if the fundamental conditions (I) and (II) are satisfied, then both the maximum likelihood estimator and the maximum *a posteriori* estimator converge to the true set of parameters in probability. This property is called consistency of estimation. If the fundamental conditions are not satisfied, then such a model may not have consistency.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Theorem 6.13 *Assume the fundamental conditions (I) and (II) with index s ($s \geq 6$) and that, for an arbitrary $p > 0$, $a_n/n^p \to 0$ ($n \to \infty$). Let $\xi_n(w)$ be an empirical process on $\{w; K(w) < \epsilon\}$,*

$$\xi_n(w) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{a(X_i, u) - E_X[a(X, u)]\},$$

and

$$\|\xi_n\| = \sup_{u \in \mathcal{M}} |\xi_n(u)|.$$

Then the following hold.
(1) For a given $0 < \delta < 1$, there exists a constant $C > 0$, such that, for arbitrary $n \geq 1$,

$$P(\|\xi_n\|^2 > n^\delta) \leq \frac{C}{n^{s\delta/2}},$$

$$P(nK(g(\hat{u})) > n^\delta) \leq \frac{C}{n^{s\delta/2}},$$

where $\hat{u}$ is defined by $\hat{w} = K(g(\hat{u})))$.
(2) For a given $0 < \delta < 1$, there exists a constant $C' > 0$, such that, for arbitrary $n \geq 1$,

$$E[\|\xi_n\|^2]_{\{\|\psi_n\|^2 > n^\delta\}} \leq \frac{C'}{n^{\delta(s/2-1)}},$$

$$E[nK(g(\hat{u}))]_{\{nK(g(\hat{u})) > n^\delta\}} \leq \frac{C'}{n^{\delta(s/2-1)}}.$$

Proof of Theorem 6.13 This theorem is proved in the same way as the previous theorem. Let $\sigma(u) \equiv \sigma(g(u))$.
(1) From Theorem 5.8,

$$E[\|\xi_n\|^s] = C < \infty,$$

and it follows that

$$\begin{aligned} C &\geq E[\|\xi_n\|^s]_{\{\|\xi_n\|^2 > n^\delta\}} \\ &\geq n^{s\delta/2} P(\|\xi_n\|^2 > n^\delta). \end{aligned}$$

Therefore

$$P(\|\xi_n\|^2 > n^\delta) \leq \frac{C}{n^{s\delta/2}}. \tag{6.76}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

By using $K(g(u)) = u^{2k}$ and the Cauchy–Schwarz inequality,

$$\begin{aligned} R_n^0(g(u)) &= nu^{2k} - u^k\ \xi_n(u) + a_n\sigma(u) \\ &\geq \tfrac{1}{2}(nu^{2k} - \|\xi_n\|^2) + a_n\sigma(u). \end{aligned}$$

A parameter u_0 in the set of true parameters satisfies $K(g(u_0)) = u_0^{2k} = 0$, hence

$$R_n^0(g(u_0)) = a_n\sigma(u_0).$$

Therefore, by the definition of $\hat{u}$, $R_n^0(\hat{u}) \leq R_n^0(w_0)$,

$$\tfrac{1}{2}(n\hat{u}^{2k} - \|\xi_n\|^2) + a_n\sigma(\hat{u}) \leq a_n\sigma(u_0).$$

Hence, if $\|\xi_n\|^2 \geq n^\delta$, since a_n is smaller than any power of n^p $(p > 0)$, there exists a constant $c_1 > 0$ such that

$$n\hat{u}^{2k} \leq \|\xi_n\|^2 + 4a_n\|\sigma\| \leq c_1\|\xi_n\|^2.$$

Therefore

$$\begin{aligned} P(nK(g(\hat{u}))) > n^\delta) &\leq P(c_1\|\xi_n\|^2 > n^\delta) \\ &\leq \frac{c_2}{n^{s\delta/2}}. \end{aligned}$$

(2) By using the above results,

$$\begin{aligned} E[\|\xi_n\|^2]_{\{\|\xi_n\|^2 > n^\delta\}} &\leq \sum_{j=1}^{\infty} E[\|\xi_n\|^2]_{\{jn^\delta < \|\xi_n\|^2 \leq (j+1)n^\delta\}} \\ &\leq \sum_{j=1}^{\infty} (j+1)n^\delta \times \frac{C}{(jn^\delta)^{s/2}} \\ &\leq \frac{c_3}{n^{\delta(s/2-1)}} \end{aligned}$$

In the same way,

$$\begin{aligned} E[nK(g(\hat{u}))]_{\{nK(g(\hat{u}))>n^\delta\}} &\leq \sum_{j=1}^{\infty} E[nK(g(\hat{u}))]_{\{jn^\delta < nK(g(\hat{u})) \leq (j+1)n^\delta\}} \\ &\leq \sum_{j=1}^{\infty} (j+1)n \times \frac{c_2}{(jn^\delta)^{s/2}} \\ &\leq \frac{c_4}{n^{\delta(s/2-1)}}. \end{aligned}$$

□

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Main Theorem 6.4 *Assume that $q(x)$ and $p(x|w)$ satisfy the fundamental conditions (I) and (II) with index s $(s \geq 6)$. Let $\{a_n \geq 0\}$ be a nondecreasing sequence that satisfies the condition that, for arbitrary $p > 0$,*

$$\lim_{n\to\infty} \frac{a_n}{n^p} = 0.$$

Let $M = \{M_\alpha\}$ be the manifold found by resolution of singularities and its local coordinate.
(1) If $a_n \equiv 0$, then

$$\lim_{n\to\infty} nE[R_{\rm g}] = \tfrac{1}{4}E\Big[\max_{\alpha} \max_{u\in M_{\alpha 0}} \Big(\max\{0, \xi(u)\}\Big)^2\Big],$$

$$\lim_{n\to\infty} nE[R_{\rm t}] = -\tfrac{1}{4}E\Big[\max_{\alpha} \max_{u\in M_{\alpha 0}} \Big(\max\{0, \xi(u)\}\Big)^2\Big],$$

where $\max_\alpha$ shows the maximization for local coordinates and

$$M_{\alpha 0} = \{u \in M_\alpha; K(g(u)) = 0\}.$$

(2) If $\lim_{n\to\infty} a_n = a^*$, then

$$\lim_{n\to\infty} nE[R_{\rm g}] = \tfrac{1}{4}E\Big[\max_{\alpha}\Big(\max\{0, \xi(u^*)\}\Big)^2\Big],$$

$$\lim_{n\to\infty} nE[R_{\rm t}] = -\tfrac{1}{4}E\Big[\max_{\alpha}\Big(\max\{0, \xi(u^*)\}\Big)^2\Big],$$

where u^* is the parameter in $M_{\alpha 0}$ that maximizes

$$\tfrac{1}{4}\max\{0, \xi(u)\}^2 - a^*\sigma(g(u)).$$

(3) If $\lim_{n\to\infty} a_n = \infty$, then

$$\lim_{n\to\infty} nE[R_{\rm g}] = \tfrac{1}{4}E\Big[\max_{\alpha} \max_{u\in M_{\alpha 00}} \Big(\max\{0, \xi(u)\}\Big)^2\Big],$$

$$\lim_{n\to\infty} nE[R_{\rm t}] = -\tfrac{1}{4}E\Big[\max_{\alpha} \max_{u\in M_{\alpha 00}} \Big(\max\{0, \xi(u)\}\Big)^2\Big],$$

where $M_{\alpha 00}$ is the set of parameters which minimizes $\sigma(g(u))$ in the set $M_{\alpha 0}$.

Proof of Main Theorem 6.4 Let $\epsilon > 0$ be a sufficiently small constant. The proof is divided into several cases.
Case (A), $\|\psi_n\|^2 > n\epsilon$. By the proof of Theorems 6.12, $nK(\hat{w}) \leq c_1\|\psi_n\|^2$. The generalization error of the partial expectation is bounded by Theorem 6.12,

$$E[nK(\hat{w})]_{\{\|\psi_n\|^2 > n\epsilon\}} \leq \frac{C'}{n^2}, \tag{6.77}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Also the training error of the partial expectation is

$$E[nK_n(\hat{w})]_{\{\|\psi_n\|^2>n\epsilon\}} \leq \frac{1}{2}E[3nK(\hat{w}) + \|\psi_n\|^2]_{\{\|\psi_n\|^2>n\epsilon\}} \leq \frac{C''}{n^2}. \tag{6.78}$$

Therefore the event $\|\psi_n\|^2 > n\epsilon$ does not affect the generalization and training errors asymptotically.

Case (B), $\|\psi_n\|^2 \leq n\epsilon$. As is shown by the proofs of Theorems 6.12 and 6.13, if $\|\psi_n\|^2 \leq n\epsilon$, then $K(\hat{w}) \leq c_1 n\epsilon$. Let us use resolution of singularities and Main Theorem 6.1. The function to be minimized, which is called a loss function, is

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

where $u \in [0, 1]^d$. By using parameterization $(t, v) \in T \times S$ of each local coordinate, the loss function is given by

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

where we use the notation

$$\begin{aligned} K(t, v) &= K(g(t, v)) = t^2, \\ K_n(t, v) &= K_n(g(t, v)), \\ R_n^0(t, v) &\equiv R_n^0(g(t, v)), \\ \sigma(t, v) &\equiv \sigma(g(t, v)). \end{aligned}$$

Note that, even if the optimal parameter is on the boundary of $[0, 1]^d$, it asymptotically does not affect the value t because, sufficiently near the point $(0, v)$, the surface $v =$ const. can be taken perpendicular to the boundary. The loss function is rewritten as

$$R_n^0(t, v) = nt^2 - \sqrt{n}\ t\ \xi_n(0, v) + a_n\sigma(0, v) + R_1(t, v),$$

where

$$R_1(t, v) = -\sqrt{n}\ t\ \Big(\xi_n(t, v) - \xi_n(0, v)\Big) + a_n(\sigma(t, v) - \sigma(0, v)).$$

Case (B1), $\|\xi_n\|^2 > n^\delta$ $(0 < \delta \leq 1)$. By Theorem 6.13 and $nK(\hat{u}) < c_1\|\xi_n\|^2$,

$$E[nK(\hat{u})]_{\{\|\xi_n\|^2>n^\delta\}} \leq \frac{c_5}{n^{\delta(s/2-1)}},$$

and

$$\begin{aligned} E[nK_n(\hat{u})]_{\{\|\xi_n\|^2>n^\delta\}} &\leq (1/2)E[3nK(\hat{w}) + \|\xi_n\|^2]_{\{\|\xi_n\|^2>n^\delta\}} && (6.79) \\ &\leq \frac{c_6}{n^{\delta(s/2-1)}}. && (6.80) \end{aligned}$$

Therefore the event $\|\xi_n\|^2 > n^\delta$ $(\delta > 0)$ does not affect the generalization and training errors asymptotically.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Case (B2), $\|\xi_n\|^2 \le n^\delta$ ($\delta > 0$). We know $nK(\hat{u}) = n\hat{t}^2$ is not larger than $c_7 n^\delta$, hence $\hat{t} \le c_8 n^{(\delta-1)/2}$. Thus we can restrict t in the region,

$$T_\delta \equiv \{0 \le t \le c_8 n^{(\delta-1)/2}\}.$$

By using Theorem 6.11,

$$\begin{aligned}\|R_1\| &\equiv \sup_{t \in T_\delta} |R_1(t, v)| \\ &\le \sup_{t \in T_\delta} \{n^{1/2} t^{1+k_0} \|\nabla \xi_n\| + a_n t^{k_0} \|\nabla \sigma\|\} \\ &\le c_9 n^{-\delta/2} \|\nabla \xi_n\| + c_{10} a_n n^{k_0(\delta-1)/2} \|\nabla \sigma\|, \end{aligned} \tag{6.81}$$

where $k_0 = 1/k^*$. Therefore $\|R_1\| \to 0$ in probability. We need to minimize

$$R_n^0(t, v) = nt^2 - \sqrt{n}\ t\ \xi_n(0, v) + a_n \sigma(0, v) + R_1(t, v)$$

in $T_\delta \times V$. For a given v, the parameter t that minimizes $R_n^0(t, v)$ is denoted by $t(v)$.

Case (B2-1), $\xi_n(0, v) \le 0$. If $\xi_n(0, v) \le 0$ then $R_n^0(t(v), v)$ is not larger than the special case $t = 0$,

$$R_n^0(t(v), v) \le a_n \sigma(0, v) + \|R_1\|.$$

On the other hand, by removing the nonnegative term,

$$R_n^0(t(v), v) \ge a_n \sigma(0, v) - \|R_1\|.$$

Therefore

$$|nt(v)^2 - \sqrt{n}\ t(v)\ \xi_n(0, v)| \le 2\|R_1\|.$$

Moreover, since $\xi_n(0, v) \le 0$,

$$|nR_{\mathrm{g}}| \le 2\|R_1\|, \tag{6.82}$$

$$|nR_{\mathrm{t}}| \le 2\|R_1\|. \tag{6.83}$$

Case (B2-2), $\xi(0, v) > 0$. We have

$$R_n^0(t, v) = (\sqrt{n}t - \xi_n(0, v)/2)^2 - \tfrac{1}{4}\xi_n(0, v)^2 + a_n \sigma(0, v) + R_1(t, v).$$

Then $R_n^0(t(v), v)$ is not larger than the special case $t = \xi_n(0, v)/(2\sqrt{n})$,

$$R_n^0(t(v), v) \le -\tfrac{1}{4}\xi_n(0, v)^2 + a_n \sigma(0, v) + \|R_1\|. \tag{6.84}$$

On the other hand, by removing the nonnegative term,

$$R_n^0(t(v), v) \ge -\tfrac{1}{4}\xi_n(0, v)^2 + a_n \sigma(0, v) - \|R_1\|. \tag{6.85}$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

Therefore

$$(\sqrt{n}\, t - \xi_n(0, v)/2)^2 \leq 2\|R_1\|,$$

which means that

$$\xi_n(0, v)^2/4 - 2\|R_1\| \leq nR_{\mathrm{g}} \leq \xi_n(0, v)^2/4 + 2\|R_1\|, \tag{6.86}$$

$$-\xi_n(0, v)^2/4 - 2\|R_1\| \leq nR_{\mathrm{t}} \leq -\xi_n(0, v)^2/4 + 2\|R_1\|. \tag{6.87}$$

Then by using the convergence in law $\xi_n(u) \to \xi_n(u)$, and eq.(6.84) and eq.(6.85), the minimizing procedure for v is divided into three cases. By comparing $-\xi(0, v)^2/4$ with $a_n \sigma(0, v)$, we have following results.

(1) If $a_n \equiv 0$, the following convergences in probability hold:

$$nR_{\mathrm{g}} \to (1/4) \max_{\alpha} \max_{u \in M_{\alpha 0}} \max\{0, \xi(u)^2\},$$

$$nR_{\mathrm{t}} \to -(1/4) \max_{\alpha} \max_{u \in M_{\alpha 0}} \max\{0, \xi(u)^2\}.$$

(3) If $a_n \to \infty$, the following convergences in probability hold:

$$nR_{\mathrm{g}} \to (1/4) \max_{\alpha} \max_{u \in M_{\alpha 00}} \max\{0, \xi(u)^2\},$$

$$nR_{\mathrm{t}} \to -(1/4) \max_{\alpha} \max_{u \in M_{\alpha 00}} \max\{0, \xi(u)^2\}.$$

(2) If $\lim_n a_n = a^*$, the following covergences in probability hold:

$$nR_{\mathrm{g}} \to (1/4) \max_{\alpha} \overset{*}{\max_{u \in M_{\alpha 0}}} \max\{0, \xi(u)^2\},$$

$$nR_{\mathrm{t}} \to -(1/4) \max_{\alpha} \overset{*}{\max_{u \in M_{\alpha 0}}} \max\{0, \xi(u)^2\},$$

where $\overset{*}{\max_{u \in M_{\alpha 0}}}$ shows the maximization of $(1/4)\xi(u)^2 - a^* \sigma(u)$ in the set $M_{\alpha 0}$. Lastly, from eqs.(6.77), (6.78), (6.79), (6.80), (6.82), (6.83), (6.86), (6.87), and $E[\|R_1\|^2] < \infty$, both nR_{g} and nR_{t} are asymptotically uniformly integrable, which completes Main Theorem 6.4. □

Corollary 6.4 *Let $\hat{w}$ be the maximum likelihood or a posteriori estimator.*
(1) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index s $(s \geq 6)$. Then

$$F_n = -\sum_{i=1}^{n} \log p(X_i|\hat{w}) + \lambda \log n - (m-1) \log\log n + F_n^{MR},$$

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

where F_n^{MR} is a random variable which converges to a random variable F^{MR} in law.
(2) Assume that $q(x)$, $p(x|w)$ and $\varphi(w)$ satisfy the fundamental conditions (I) and (II) with index s $(s \geq 6)$. Then the following convergence of expectation holds,

$$E[F_n] = nE_X[E[\log p(X|\hat{w})]] + \lambda \log n - (m-1) \log \log n + E[F_n^{MR}],$$

where $E[R_n^{MR}] \to E[F^{MR}]$.

Proof of Corollary 6.4 From Main Theorem 6.4,

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

converges in law. Its expectation also converges. This corollary is immediately derived from Main Theorem 6.2. □

Remark 6.16 (1) In the equation

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

the generalization error is represented by the training error; however, the left-hand side contains the entropy $+1$, whereas the right-hand side has entropy -1. Therefore an information criterion cannot be directly derived from this equation. To construct an information criterion, we need a constant $C > 0$ such that

$$E[R_\mathrm{g}] = E[R_\mathrm{t}] + \frac{2C}{n} + o\Big(\frac{1}{n}\Big).$$

In a regular statistical model $C = d/2$; however, in a singular model, it depends on the true distribution and a statistical model.
(2) If the set of parameters is not compact, then $|w| = \infty$ may be an analytic singularity. For the behavior of the maximum value of the random process and its application to the maximum likelihood training errors, see, for example, [20, 36, 113]. Although several results were obtained on the asymptotic behavior of the training errors, it is still difficult to know their generalization errors. It is conjectured that the asymptotic generalization error of the maximum likelihood is very large.

Remark 6.17 (Phase transition) It seems that, when $\beta \to \infty$, the Bayes and Gibbs generalization errors converge to that of the maximum likelihood estimation. However, such convergence does not hold in general, even if the set of

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press

parameters is compact. In Gibbs and Bayes estimations, the main part of the *a posteriori* distribution is contained in the essential coordinates which minimize λ and maximize its order. However, in the maximum likelihood estimation, such a restriction is not introduced. Therefore, the asymptotic generalization and training errors may not be continuous at $\beta \to \infty$. Such a phenomenon is called a phase transition in statistical physics. From the viewpoint of statistical physics, a singular learning machine has phase transition at $\beta = \infty$, in general.

https://doi.org/10.1017/CBO9780511800474.007 Published online by Cambridge University Press