論文 / 著書情報
Article / Book Information

題目(和文)	特異モデルにおける確率的複雑さの漸近展開
Title(English)	Asymptotic expansion of stochastic complexities in singular learning machines
著者(和文)	山崎啓介
Author(English)	Keisuke Yamazaki
出典(和文)	学位:工学博士, 学位授与機関:東京工業大学, 報告番号:甲第5823号, 授与年月日:2004年3月26日, 学位の種別:課程博士, 審査員:渡辺澄夫
Citation(English)	Degree:Doctor of Engineering, Conferring organization: Tokyo Institute of Technology, Report number:甲第5823号, Conferred date:2004/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Asymptotic Expansion of Stochastic Complexities in Singular Learning Machines

Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology

Keisuke Yamazaki

2004

Contents

Chapter 1

Introduction

In the information engineering field, complicated hierarchical models such as multi-layered perceptrons, gaussian mixtures and graphical models are mainly used. These models have the wide range of application. For example, many kinds of gaussian mixtures are employed in pattern recognition, automatic data clustering, discovery of knowledge from data samples. In spite of these applications, the theoretical properties of these models have not yet been clarified.

All statistical models fall into two typical categories, *identifiable* and *non-identifiable*. In general, if the parameter of the learning model is uniquely determined by its behavior, the model is identifiable. If otherwise, non-identifiable. The learning model is represented by the probability distribution $p(x|w)$, which has the parameter w. A non-identifiable model has many true parameters, because its mapping is not one-to-one. When we assume that the class of functions $p(x|\cdot)$ includes the true distribution $q(x)$, the true parameters $\{w; p(x|w) = q(x)\}$ are not one point in its parameter space. In general, the set has many singular points. At these points, the Fisher information matrices are not positive definite. Hence, the log likelihood cannot be approximated by any quadratic form of the parameter in the neighborhood of these singularities. The method of regular statistical models, which are identifiable, cannot be applied to these models. That is why there is no mathematical foundation for non-identifiable models.

The importance of analysis of non-identifiable models has recently been pointed out (Hartigan, 1985; Amari & Ozeki, 2001). In some models such as mixture models, the maximum likelihood estimator often diverges. It has been proposed that, by choosing a locally conic parameterization, the asymptotic behavior of the log likelihood ratio of the maximum likelihood method can be analyzed based on the theory of empirical processes (Dacunha-Castelle & Gassiart, 1997). It was proven that the maximum likelihood method produces very small training errors and very large generalization errors (Hagiwara, 2002). It has been made well known by numerous experiments that the Bayesian estimation is more useful than the maximum likelihood method (Akaike, 1980; Mackay, 1992).

In order to clarify the behavior of the non-identifiable model, we have proven the relation between the Bayesian generalization error and the singularities in the parameter space, by using algebraic geometry (Watanabe, 1999). In the algebraic geometrical method, we assume that the Kullback information of the non-identifiable model is an analytic function. Then, we refer to the model as the singular model. We found that the asymptotic expansion of the stochastic complexity (Rissanen, 1986) depends on the largest pole of the zeta function of the Kullback information and the a priori distribution. It is well known that the stochastic complexity determines the generalization error (Levin et. al., 1990). This method provides a mathematical foundation of Bayesian estimation when the number of training samples is sufficiently large. Based on the algebraic geometrical method, we clarified the properties of singular models such as multi-layer perceptrons. In general, hierarchical learning models with Bayesian estimation achieve the more precise inference than regular statistical models, even if the true distribution is not contained in the finite parametric models (Watanabe, 2001b).

In this thesis, we discuss the relationship between some singular models and the stochastic complexities, which is equal to the minus type II likelihood or the free energy, is the most important observable in Bayesian statistics. For example, the increase of the stochastic complexity is equal to the generalization error. In addition, we can carry out the model selection and the hyperparameter optimization using the stochastic complexity. This thesis establishes the following stuffs.

(1) An algorithm to clarify the stochastic complexities in singular learning machines.

(2) A unified perspective of some singular learning machines in terms of Bayesian networks.

(3) A mathematical foundation for analysis of singular learning machines.

This thesis consists of nine chapters. In the second chapter, we introduce the standard framework for Bayesian estimation and summarize the algebraic geometrical method and the properties of the stochastic complexity. In the third chapter, the relationship between the stochastic complexity and the volume-dimension is stated. Using this relationship, we are able to construct a probabilistic algorithm for calculation of the stochastic complexity. According to the algorithm, we can obtain the coefficient of the stochastic complexity in general singular models. However, in terms of the model selection, the relationship between the increase of the model size and that of its stochastic complexity is needed. Thus, in the fourth to seventh chapters, we analyze some concrete singular models such as mixture models, Boltzmann machines, Bayesian networks and hidden Markov models, and clarify their stochastic complexities in the mathematical rigorous way. In the eighth and ninth chapters, we discuss and conclude our results.

Chapter 2

Bayes Estimation

In this chapter, we introduce the standard framework for Bayesian estimation. This is well known in statistical learning theory. Then, we summarize the algebraic geometrical method and some mathematical properties of the stochastic complexity.

2.1 Bayesian Learning and Stochastic Complexity

Let $X^n = (X_1, X_2, \cdots, X_n)$ be a set of training samples. The number of training samples is n. These and the testing samples are independently and identically taken from the true probability distribution $q(x)$. Let the learning machine be $p(x|w)$, which has a parameter w. For example, the parameter corresponds to the mean vector and the variance when $p(x|w)$ is a gaussian distribution. In general, we determine the optimal w from a loss function. This process is referred to as 'learning'. In Bayesian learning, however, we construct the predictive distribution $p(x|X^n)$ from the training samples, the learning machine and an a priori distribution. The a priori probability distribution $\varphi(w)$ is given on the set of parameters W. Then, the a posteriori probability distribution is defined by

$$p(w|X^n) = \frac{1}{Z_0(X^n)} \varphi(w) \prod_{i=1}^{n} p(X_i|w),$$

where Z_0 is a normalizing constant. The empirical Kullback information is given by

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

Then, $p(w|X^n)$ is rewritten as

$$p(w|X^n) = \frac{1}{Z(X^n)} \exp(-nH_n(w)) \varphi(w),$$

where the normalizing constant $Z(X^n)$ is given by

$$Z(X^n) = \int \exp(-nH_n(w))\varphi(w)dw.$$

The stochastic complexity is defined by

$$F(X^n) = -\log Z(X^n).$$

In Bayesian learning, we must prepare the learning model and the a priori distribution. Then, it is a problem which sized model or which prior makes the prediction precise. This is referred to as the model selection problem. We can select the optimal model and the hyperparameter that determine the a priori distribution by minimizing $-\log Z_0(X^n)$ because it is a minus log marginal likelihood of the learning model and the prior. This is equivalent to minimizing the stochastic complexity, since

$$\begin{aligned} -\log Z_0(X^n) &= -\log Z(X^n) + S(X^n), \\ S(X^n) &= -\sum_{i=1}^{n} \log q(X_i), \end{aligned}$$

where the empirical entropy $S(X^n)$ is independent of the learners. The average stochastic complexity $F(n)$ is defined by

$$F(n) = -E_{X^n}\Big[\log Z(X^n)\Big], \qquad (2.1)$$

where E_{X^n} stands for the expectation value over all sets of training samples.

The Bayesian predictive distribution $p(x|X^n)$ is given by

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

In Bayesian learning, we use all parameters w according to $p(w|X^n)$. It is different from other estimations which determine the optimal parameter such as the maximum likelihood estimation. The generalization error $G(n)$ is the average Kullback information from the true distribution to the Bayesian predictive distribution,

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

It is very important to clarify the behavior of $G(n)$, when the number of training samples is sufficiently large. The relation between $G(n)$ and $F(n)$ is,

$$G(n) = F(n+1) - F(n). \tag{2.2}$$

This relation is well known (Levin et.al., 1990; Yamanishi, 1998; Watanabe, 1999) and allows that the generalization error to be calculated from the average stochastic complexity. When $F(n)$ is obtained as

$$F(n) = \lambda\log n + o(\log n),$$

the model's generalization error is given by

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

This λ is referred to as the learning coefficient of the machine.

If a learning machine is an identifiable and regular statistical model, it is proven (Schwarz, 1978) that asymptotically

$$F(n) = \frac{d}{2}\log n + const,$$

holds, where d is the dimension of the parameter space W. However, for models that are non-identifiable and non-regular such as artificial neural networks, the different results are derived. The asymptotic expansion of $F(n)$ is

$$F(n) = \lambda \log n - (m-1) \log \log n + const,$$

where the rational number $(-\lambda)$ and the natural number m are respectively the maximum pole and its order of the zeta function of the Kullback information and the a priori distribution (Watanabe, 2001a). It is hard to find the largest pole since the resolution of singularities is needed. According to this formula, however, the upper bound of the constant λ can be derived in some models such as multi-layer neural networks (Watanabe, 2001b). In this thesis, first, we construct the probabilistic algorithm for calculation of λ. Next, we evaluate the upper bounds of λ in mixture models, Boltzmann machines, Bayesian networks and hidden Markov models by finding a pole of the zeta function.

2.2 Algebraic Geometry of the Stochastic Complexity

We define the Kullback information from the true distribution $q(x)$ to the learner $p(x|w)$ by

$$H(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx. \tag{2.3}$$

This function is equal to zero iff $q(x) = p(x|w)$ and not less than zero for all w. Thus, it indicates the distance from $q(x)$ to $p(x|w)$. However, it does not satisfy the symmetric law. The asymptotic form of the stochastic complexity greatly relates to the singularities of the parameter set $\{w; H(w) = 0\}$. Note the important and nontrivial relation that was clarified by the algebraic geometrical method (Watanabe, 2001a; Watanabe,2001b).

First, assume that the Kullback information $H(w)$ is an analytic function of w in the support of the a priori distribution. If the learner is in a redundant state in comparison with the true distribution, the set $\{w \in W; H(w) = 0\}$ includes quite complicated singularities. The algebraic geometry is the only means by which we can analyze the effect of singularities. We need the function $J(z)$ of complex variable z, which is defined by

$$J(z) = \int H(w)^z \varphi(w) dw. \tag{2.4}$$

This function is called the zeta function of $H(w)$ and the a priori distribution $\varphi(w)$. It is a holomorphic function in the region $Re(z) > 0$, and can be analytically continued to the meromorphic function on the entire complex plane, whose poles are all real, negative and rational numbers. This continuation is ensured by the existence of the b-function (Watanabe, 2001a).

Let $0 > -\lambda_1 > -\lambda_2 > \cdots$ be the sequence of poles of the zeta function ordered from the origin to minus infinity, and $m_1, m_2, \cdots$ be the respective orders of the poles. The inverse Mellin transform of $J(z)$ gives the state density function $(t > 0)$,

$$v(t) = \int \delta(t - H(w)) \varphi(w) dw.$$

Since $J(z)$ is a meromorphic function, the asymptotic expansion of $v(t)$ is given by

$$v(t) = \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} c_{km} t^{\lambda_k - 1} (-\log t)^{m-1},$$

for $t \longrightarrow 0$. It has been proven that $\mathcal{F}(n)$ defined by

$$\mathcal{F}(n) = -\log \int \exp(-nH(w)) \varphi(w) dw$$

is the upper bound of $F(n)$ (Watanabe, 2001a). We can apply the asymptotic expansion of $v(t)$ to $\mathcal{F}(n)$; then, this is rewritten by

$$\begin{aligned} \mathcal{F}(n) &= -\log \int_0^{\infty} \exp(-t) v(\frac{t}{n}) \frac{dt}{n} \\ &= \lambda_1 \log n - (m_1 - 1) \log \log n + \text{const.}, \end{aligned} \tag{2.5}$$

for $n \longrightarrow \infty$. The coefficient of leading term in $\mathcal{F}(n)$ is λ_1, the absolute value of the largest pole. In fact, we can calculate λ_1 and m_1 by using the resolution of singularities in algebraic geometry (Atiyah, 1970; Hironaka, 1964; Watanabe, 2001a). However, it is generally difficult to find the complete resolution map. We can alternatively find a partial resolution of singularities. This gives us a pole $-\mu$ of zeta function $J(z)$. Then we obtain the upper bounds of the stochastic complexity, since μ is the upper bound of λ_1. In this thesis, we present the nontrivial upper bound of stochastic complexity of some singular models derived by the algebraic geometrical method.

2.3 Basic Properties of Stochastic Complexity

Let us summarize some basic properties of stochastic complexity. These properties are trivial but very useful in the proofs.

First, define a function $\mathcal{F}(S, \psi)$ by

$$\mathcal{F}(S, \psi) = -\log \int \exp(-nS(w))\psi(w)dw,$$

where S is a function of w and ψ is a positive function over all w. This is well defined even if $\psi(w)$ is not a probability density function.

(Proposition. 1) Using Jensen's inequality, we can show easily that the following inequality holds (Opper & Haussler,1995; Watanabe, 1999),

$$F(n) \leq \mathcal{F}(H, \varphi), \tag{2.6}$$

where $H(w)$ is the Kullback information defined by the equation (2.3).

(Proposition. 2) If the functions H_1, H_2 and the positive function φ_1, φ_2 satisfy

$$H_1(w) \leq H_2(w) \quad (\forall w \in W),$$
$$\varphi_1(w) \geq \varphi_2(w) \quad (\forall w \in W),$$

then the following inequality immediately holds,

$$\mathcal{F}(H_1, \varphi_1) \leq \mathcal{F}(H_2, \varphi_2).$$

This inequality also claims that, if the integrated region in the parameter set is $U \supset V$,

$$-\log \int_U \exp(-nK(w))\psi(w)dw \leq -\log \int_V \exp(-nK(w))\psi(w)dw,$$

holds. From this inequality, we obtain the upper bound of the stochastic complexity. Based on this property, it is sufficient to consider only the restricted parameter set.

(Proposition. 3) If $w = (w_1, w_2)$, assume that H and φ are separated into two functions of each other,

$$\begin{aligned} H(w_1, w_2) &= H_1(w_1) + H_2(w_2), \\ \varphi(w_1, w_2) &= \varphi_1(w_1)\ \varphi_2(w_2). \end{aligned}$$

The following equality holds,

$$\mathcal{F}(H, \varphi) = \mathcal{F}(H_1, \varphi_1) + \mathcal{F}(H_2, \varphi_2).$$

Define the zeta functions by

$$\begin{aligned} J(z) &= \int H(w)\varphi(w)dw, \\ J_i(z) &= \int H_i(w_i)\varphi_i(w_i)dw_i \quad (i = 1, 2). \end{aligned}$$

Let $-\mu$, $-\mu_1$, $-\mu_2$ be the largest poles of J, J_1 and J_2 respectively. The property claims that

$$\mu = \mu_1 + \mu_2. \tag{2.7}$$

Chapter 3

Learning Coefficient and Volume-Dimension

In this chapter, we introduce a relationship between a learning coefficient and a volume-dimension. Then, using the relationship, we construct a new probabilistic algorithm to calculate the learning coefficient.

3.1 Volume-dimension

Let the volume of subset of the parameter space $V(t)$ $(0 \leq t < \infty)$ be

$$V(t) = \int_{H(w)<t} \varphi(w)dw. \tag{3.1}$$

This function means the volume of the parameter set according to the probability measure $\varphi(w)dw$, where the Kullback information $H(w)$ represented by (2.3) is not larger than t. We assume that $\varphi(w)$ is not equal to zero on $\{w^*; p(x|w^*) = q(x)\}$. In this section, we prove the following theorem.

Theorem 1 *Assume that* $0 < \alpha$ $(\alpha \neq 1)$ *is an arbitrary constant, then*

$$\lambda = \lim_{t \to +0} \frac{\log\{V(\alpha t)/V(t)\}}{\log \alpha},$$

where λ *is the coefficient of the learning curve.*

(Remark) This λ is also the coefficient of the leading term in the stochastic complexity.

This theorem claims that the learning coefficient is equal to the volume-dimension in the parameter space. In the field of fractal geometry, this dimension is referred to as the box counting dimension.

(Proof of Theorem 1) We define the function $\Theta(\cdot)$ as

$$\Theta(y) = \begin{cases} 1 & (y \geq 0) \\ 0 & (y < 0) \end{cases} .$$

Then, $V(t)$ is rewritten as

$$V(t) = \int \Theta(t - H(w))\varphi(w)dw.$$

Let $\delta(\cdot)$ be Dirac's delta function. According to the relationship of distribution $\Theta(t)' = \delta(t)$,

$$V'(t) = \frac{dV}{dt} = \int \delta(t - H(w))\varphi(w)dw.$$

Since the number of t where $\{w; H(w) = t\}$ is an analytic set and has singularities is finite, $V'(t)$ is well-defined. Thus, it follows that

$$\begin{aligned} J(z) &= \int dt \int dw\ \delta(t - H(w))\ t^z\ \varphi(w) \\ &= \int t^z V'(t)dt. \end{aligned}$$

$J(z)$ is the Mellin transform of $V'(t)$. Using a property of Mellin transform and the fact that the largest pole of $J(z)$ is rational $-\lambda$, and that its multiplicity m is counting number, we can derive that $V'(t)$ has the following asymptotic expansion at $t \to +0$ (Watanabe, 2001b),

$$V'(t) = c_1 t^{\lambda-1}(-\log t)^{m-1} + r_1(t), \tag{3.2}$$

where $c_1 > 0$ is a constant and $r_1(t)$ satisfies

$$\lim_{t \to +0} \frac{r_1(t)}{t^{\lambda-1}(-\log t)^{m-1}} = 0.$$

Since $V(0) = 0$,

$$V(t) = \int_0^t V'(s)ds.$$

In general, when a function has the asymptotic expansion, the primitive function can be asymptotically expanded and calculated by integrating each term. By using the function

$$f(\lambda, m, t) = \int_0^t s^{\lambda-1}(-\log s)^{m-1}ds,$$

and the equation(3.2), it follows that

$$V(t) = c_1 f(\lambda, m, t) + r_2(t),$$

where $r_2(t)$ satisfies

$$\lim_{t\to+0} \frac{r_2(t)}{f(\lambda, m, t)} = 0.$$

Using integration by part, we obtain the recurrence formula of m in $f(\lambda, m, t)$,

$$\begin{aligned} f(\lambda, m, t) &= \frac{1}{\lambda}t^{\lambda}(-\log t)^{m-1} \\ &\quad + \frac{m-1}{\lambda} f(\lambda, m-1, t). \end{aligned}$$

Since m is a counting number and

$$f(\lambda, 1, t) = \frac{t^{\lambda}}{\lambda},$$

$f(\lambda, m, t)$ is finite summation,

$$f(\lambda, m, t) = \sum_{k=1}^{m} \frac{(m-1)!}{\lambda^k (m-k)!} t^{\lambda}(-\log t)^{m-k}.$$

Therefore,

$$V(t) = c_2 t^{\lambda}(-\log t)^{m-1} + r_3(t),$$

where $c_2 > 0$ is a constant and $r_3(t)$ satisfies

$$\lim_{t\to+0} \frac{r_3(t)}{t^{\lambda}(-\log t)^{m-1}} = 0.$$

Define

$$A = \lim_{t \to +0} V(\alpha t)/V(t).$$

Then,

$$\begin{aligned} A &= \lim_{t \to +0} \frac{c_2(\alpha t)^{\lambda}(-\log(\alpha t))^{m-1} + r_3(t)}{c_2 t^{\lambda}(-\log t)^{m-1} + r_3(t)} \\ &= \alpha^{\lambda}, \end{aligned}$$

which completes the proof of Theorem 1. (End of Proof)

(Remark) From this result, it is clear that this method calculates λ independent of m. If $m \geq 2$, however, the convergence at $t \to 0$ is weaker.

3.2 Proposed Algorithm

In this section, we present the new algorithm to calculate λ based on Theorem 1. The main goal of this algorithm is to approximate the volume $V(t)$. However, it is not easy to build the uniform distribution on the analytic set $\{w; H(w) < t\}$ since the set has complex singularities at $t \to +0$. The algorithm gets over this problem.

Assume the set of the parameter space is $W = [-1, 1]^d$ and it includes the subset of $\{w^*; p(x|w^*) = q(x)\}$. We define an a priori distribution $\varphi(w)$ is the uniform distribution on W. It is known that λ does not depend on $\varphi(w)$ when the prior satisfies $\varphi(w) > 0$ on $\{w; H(w) = 0\}$. Let the Kullback information $H(w)$ be given. Let $0 < \alpha < 1$, $0 < T < 1$, $N >> 1$ be constants. We denote the volume (the number of elements) of the set S as $|S|$. Let us obtain the sequence $\{\lambda_n; n = 1, 2, 3, ...\}$ whose mean value converges to λ from the following algorithm.

Algorithm to calculate the learning coefficient

Step 1.
Let $t = T$, $n = 1$.

Obtain parameters $w_1, w_2, ...,$ from the uniform random distribution on $W = [-1,1]^d$ till the volume of

$$P_1 = \{i; H(w_i) \leq t\}$$

is equal to N ($|P_1|$=N). Calculate $|Q_1|$ such that

$$Q_1 = \{i \in P_1; H(w_i) \leq \alpha t\}.$$

From

$$\frac{V(\alpha t)}{V(t)} \cong \frac{|Q_1|}{N}$$

and Theorem 1,

$$\lambda_1 = \frac{\log\{|Q_1|/N\}}{\log \alpha}.$$

Step 2.

Let $n := n + 1$ and $t := \alpha t$.

Select the elements from P_{n-1} in a random order, and let them be u_i. Obtain $w_1, w_2, \cdots$ such that

$$w_i = u_i + v_i \quad (i = 1, 2, 3, \cdots),$$

where v_i are random variables on $[-\frac{t}{2}(1-\alpha), \frac{t}{2}(1-\alpha)]^d$. Continue it till $|P_n| = N$, where

$$P_n = \{i; H(w_i) \leq t\}.$$

The elements in P_n are samples from the uniform distribution on $\{w; H(w) < t\}$. Calculate $|Q_n|$ such that

$$Q_n = \{i \in P_n; H(w_i) \leq \alpha t\}.$$

From

$$\frac{V(\alpha t)}{V(t)} \cong \frac{|Q_n|}{N}$$

and Theorem 1,

$$\lambda_n = \frac{\log\{|Q_n|/N\}}{\log \alpha}.$$

Step 3.
Repeat Step 2.

According to Theorem 1, the mean value of the sequence λ_1, λ_2, λ_3, $\cdots$ converges to λ.

3.3 Verification of the Algorithm

In this section, we apply the algorithm to tree-layered perceptrons and verify the effectiveness.

3.3.1 Evaluation: Three-Layered Perceptron

We apply the presented algorithm to tree-layered perceptrons which have one input unit, one output unit and K hidden units. We can depict the model as

$$\begin{aligned} p(y|x,w) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(y-f(x,w))^2), \\ f(x,w) &= \sum_{k=1}^{K} a_k \tanh(b_k x), \end{aligned}$$

where $x, y \in R$ are input and output respectively. The number of the parameter is $2K$ and

$$w = \{a_k, b_k; k = 1, 2, ..., K\}.$$

In our experiments, we assume that the input x is taken from the uniform distribution on $[-1, 1]$ and the true distribution is

$$q(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2}).$$

This means $f(x, w^*) = 0$. In other words, $\forall a_k = 0$ or $\forall b_k = 0$. Then,

$$\begin{aligned} H(w) &= \int_{-1}^{1} \frac{dx}{2} \int dy\, p(y|x, w^*) \log \frac{p(y|x, w^*)}{p(y|x, w)} \\ &= \frac{1}{4} \int_{-1}^{1} f(x, w)^2 dx. \end{aligned} \tag{3.3}$$

When we calculate the learning coefficient, we can rewrite the Kullback information as the algebraic function at the neighborhood of the parameter set that determines the smallest λ by dividing the parameter space (Watanabe, 1998). Define

$$H_0(w) = \sum_{k=1}^{K} \{ \sum_{j=1}^{K} a_j \, (b_j)^{2k-1} \}^2.$$

Then, there are constants $c_1, c_2 > 0$ in origin-neighborhood such that

$$c_1 H_0(w) \leq H(w) \leq c_2 H_0(w). \tag{3.4}$$

In this case, λ, m from the zeta function of $H(w)$ is equal to that of $H_0(w)$. The paper (Watanabe, 1998) shows that

$$\sum_{k=1}^{K} \frac{1}{4k-2} \leq \lambda \leq \frac{\sqrt{K}}{2}. \tag{3.5}$$

Moreover, the paper (Watanabe, 2001b) shows that $\lambda = 1/2$, $m = 2$ when $K = 1$ and $\lambda = \frac{2}{3}$, $m = 1$ when $K = 2$. Our algorithm is able to calculate general models. However, we apply it to these models in order to verify its effectiveness.

3.3.2 Experimental Results: Algebraic Kullback Information

We show the results when $K = 1, 2$ (Figure 3.1, 3.2).

The horizontal axis shows the number of iterations (n in the algorithm), whereas the vertical axis the value of λ_n. We apply our algorithm in 100 times changing the seed of random number. We show the mean values and

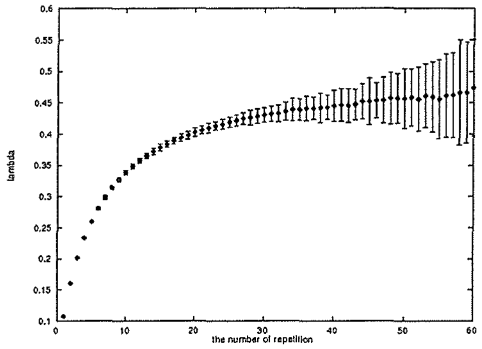

Figure 3.1: Convergence of λ_n(algebraic Kullback information,$K = 1$)

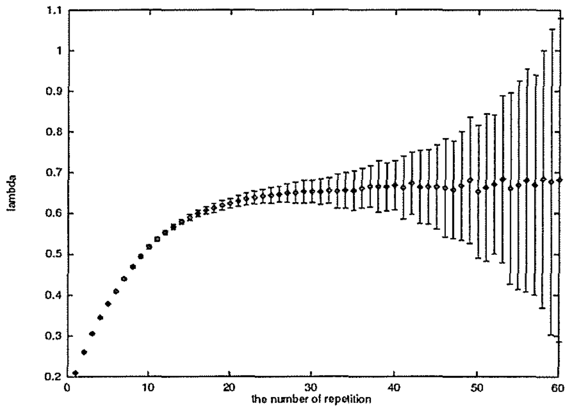

Figure 3.2: Convergence of λ_n(algebraic Kullback information,$K = 2$)

standard deviations as "Mean Value ± 2×Standard Deviation" in the figures. The conditions of the experiments are that $\alpha = 2/3$, $T = 0.5$ and $N = 1000000$. According to the equation (3.4), we use $H_0(w)$ instead of $H(w)$. The runtimes of Figure 3.1 and 3.2 are about 300 minutes by an computer on the market (450MHz). The mean value attains the true value though the standard deviation grows. It is not easy to determine whether the approximate value converges or not. Then, we show the figures when $N = 4000000$ (Figure 3.3, 3.4).

It is clear that the values of standard deviation are a half of the previous cases. The runtime is quadruple. When we assess the convergence and need the precise value, it is sufficient to try a large number of samples N. From our result, the error times $1/\sqrt{a}$ when the number of samples times a.

From the Figure 3.3 and 3.4, λ is estimated in $0.46 \pm 0.05, 0.66 \pm 0.06$, respectively. These results almost consist with the theoretical value 1/2, 2/3. Though the case when $K \geq 3$ is still unknown, our results (Figure 3.5 and 3.6) satisfy the bounds (3.5).

3.3.3 Experimental Results: Analytic Kullback Information

In general, we can find the algebraic expression such as the equation (3.4) when $H(w)$ is an analytic function. However, it is not easy to do it since the expression depends on the learning machine. Then, we introduce the method that does not need $H_0(w)$. It is generally hard to integrate the Kullback information (3.3). By using samples $\{x_a\}$ from the uniform distribution on the integral region, we construct the function,

$$H_r(w) = \frac{1}{2A} \sum_{a=1}^{A} f(x_a, w)^2. \tag{3.6}$$

We use this function instead of $H(w)$. Although the result is more precise when the number of $\{x_a\}$ is larger, the runtime is also larger. Theoretically,

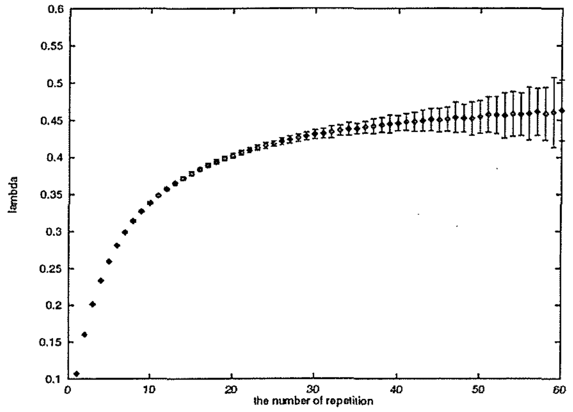

Figure 3.3: Convergence of λ_n(algebraic Kullback information, $K = 1$, $4N$)

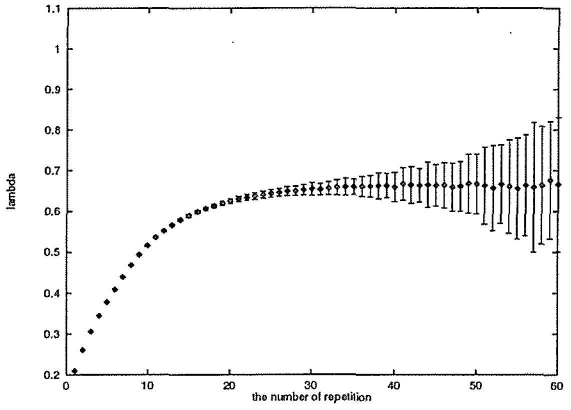

Figure 3.4: Convergence of λ_n(algebraic Kullback information,$K = 2$, $4N$)

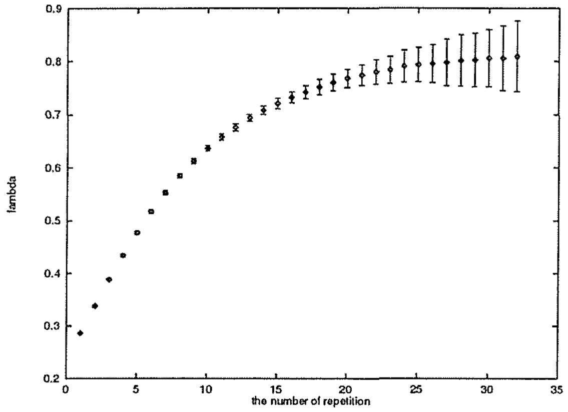

Figure 3.5: Convergence of λ_n(algebraic Kullback information,$K = 3$)

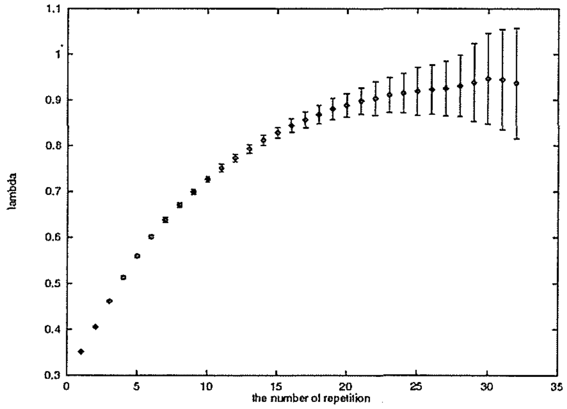

Figure 3.6: Convergence of λ_n(algebraic Kullback information,$K = 4$)

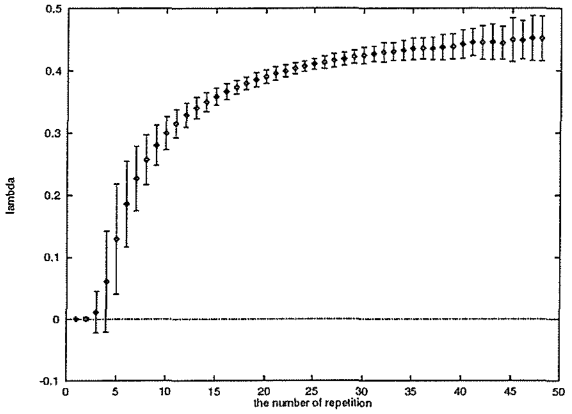

Figure 3.7: Convergence of λ_n(analytic Kullback information,$K = 1$)

it is sufficient to exist constants $c_1, c_2 > 0$ such that

$$c_1 H_r(w) \leq H(w) \leq c_2 H_r(w). \tag{3.7}$$

In order to satisfy the condition, $\{x_a\}$ is not in specific region and it seems sufficient that the number of $\{x_a\}$ is larger than the number of parameters. In our experiments, $A = 100$. We show the results (Figure 3.7, 3.8 and 3.9).

The values of λ is the same as the previous experiments. This means the method without $H_0(w)$ is also effective.

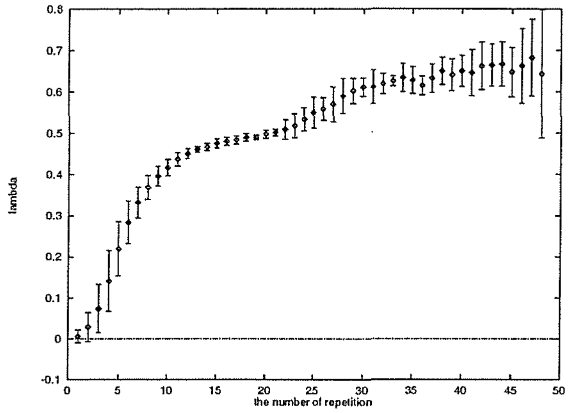

Figure 3.8: Convergence of λ_n(analytic Kullback information,$K = 2$)

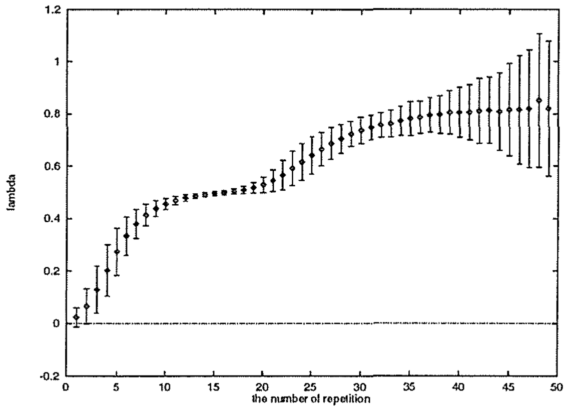

Figure 3.9: Convergence of λ_n(analytic Kullback information,$K = 3$)

Chapter 4

Mixture Models

In this chapter, we introduce the mixture model and derive a theorem of the stochastic complexity. A learning machine which is a mixture of several distributions is referred to as a mixture model. When these distributions are normal distributions, the learning machine is a gaussian mixture. It is employed not only to estimate the true distribution but also to discriminate the category to which the datum belongs. Thus, the model is used in a lot of information processing fields, for example, pattern recognition, automatic data clustering, data mining, etc. When components of distributions are neural networks, the machine is called a mixture of experts (Jacobs & Jordan, 1991). The learning algorithms using EM method are developed. In spite of the wide range of application and learning algorithms, the model selection and the properties of generalization are not yet clarified. The results of this chapter construct the mathematical foundation to analyze them.

4.1 Mixtures of Several Learning Machines

Let $f(x|b)$ be a conditional probability density function of $x \in R^N$ with a given parameter $b \in R^M$. The learning machine $p(x|w)$ made of their mixture is defined by

$$p(x|w) = \sum_{k=1}^{K} a_k f(x|b_k), \tag{4.1}$$

where $\{a_k \in R\}$ is the set of coefficients which satisfy $a_k \geq 0$ and

$$\sum_{k=1}^{K} a_k = 1.$$

The parameter of the machine $p(x|w)$ is $w = \{a_k, b_k\}$. The set of all parameters is denoted by $W = \{w\}$. The probability distribution $f(x|b_k)$ and the integer K are respectively called the component of the mixture and the number of components.

If each component $f(x|b_k)$ is equal to the normal distribution given by

$$f(x|b_k) = \frac{1}{(2\pi\sigma_k)^{1/2}} \exp(-\frac{\|x - m_k\|^2}{2\sigma_k^2}), \tag{4.2}$$

then the learning machine $p(x|w)$ is called a gaussian mixture or a normal mixture. It should be emphasized that, when we study some clustering methods or a competitive neural network, we implicitly consider the gaussian mixture. Hence the gaussian mixtures have a lot of applications to information processing systems. In some applications, the parameter is restricted to the averages of each components,

$$\begin{aligned} b_k &= m_k \in R^N, & (4.3)\\ \sigma_k &= \text{const.} \quad (k = 1, 2, \cdots, K), & (4.4) \end{aligned}$$

whereas in the other applications the parameter consists of both averages and standard deviations,

$$\begin{aligned} b_k &= (m_k, \sigma_k) \quad (k = 1, 2, \cdots, K), & (4.5)\\ m_k &\in R^N, \sigma_k > 0 \quad (k = 1, 2, \cdots, K). & (4.6) \end{aligned}$$

In this chapter, we consider a general mixture model which contains both cases as special cases. The dimension of the parameter b_k is denoted by M.

In Theorem 1, we prove the upper bound of the stochastic complexity for the general mixture represented by the equation (4.1). The upper bounds of the stochastic complexities of the both gaussian mixtures are obtained as corollaries of the Theorem 1.

4.2 Stochastic Complexity of Mixture Models

We assume the three general conditions, (A4.1),(A4.2), and (A4.3).

(A4.1) Firstly, we assume that the learning machine can attain the true model. The learner is given by

$$p(x|w) = \sum_{k=1}^{K} a_k f(x|b_k). \tag{4.7}$$

The set of parameters in the learning model $p(x|w)$ is defined by

$$W = \{a, \{b_k\}; a \in A, b_k \in B\},$$

where the set A is defined by

$$A = \{a = (a_1, a_2, .., a_{K-1}); a_k \geq 0, \sum_{k=1}^{K-1} a_k \leq 1\},$$

and B is a fixed open subset of R^M. Note that a_K is not a parameter but a function of $a = (a_1, a_2, ..., a_{K-1})$ determined by

$$a_K = 1 - \sum_{k=1}^{K-1} a_k.$$

We assume that the true distribution $q(x)$ is given by

$$q(x) = \sum_{k=1}^{H} a_k^* f(x|b_k^*), \tag{4.8}$$

where $H < K$, $a_1^*, a_2^*, \cdots, a_H^* > 0$, $\sum_{k=1}^{H} a_k^* = 1$, and $b_1^*, b_2^*, \cdots, b_H^* \in B$.

(A4.2) Secondly, we assume that the a priori probability distribution is positive on a true parameter. For a constant $\epsilon > 0$, we define the subset of parameter $A(\epsilon) \subset A$ by

$$\begin{aligned} A(\epsilon) &= \{(a_1, a_2, ..., a_{K-1}) \in A; \\ &\quad |a_k - a_k^*| \leq \epsilon \ \ (1 \leq k \leq H), \\ &\quad 0 \leq a_k \leq \epsilon \ \ (H+1 \leq k \leq K-1)\} \end{aligned}$$

and $B_k(\epsilon) \subset B$ $(1 \leq k \leq H)$ by

$$B_k(\epsilon) = \{b \in B; \|b_k - b_k^*\| \leq \epsilon\}.$$

We assume that there exists a constant $\epsilon > 0$ such that

$$\inf_{A(\epsilon), B_k(\epsilon)} \varphi(a, b_1, b_2, \cdots, b_K) > 0,$$

where '$\inf\limits_{A(\epsilon), B_k(\epsilon)}$' denotes the infimum value of $\varphi(w)$ in the region,

$$\begin{aligned} a &\in A(\epsilon), \\ b_k &\in B_k(\epsilon) \quad (1 \leq k \leq H-1), \\ b_k &\in B_H(\epsilon) \quad (H \leq k \leq K). \end{aligned}$$

(A4.3) Thirdly, we assume that the distribution of the single component is analytic function of its parameter. Assume that $f(x|b) > 0$ for arbitrary $x \in R^N$ and $b \in B$, and that the Kullback information

$$D(b_k^*||b) = \int f(x|b_k^*) \log \frac{f(x|b_k^*)}{f(x|b)} dx$$

is a twice continuously differentiable function of b in a neighborhood of b_k^* for all $1 \leq k \leq H$.

Theorem 2 *Assume the three conditions (A4.1), (A4.2), and (A4.3). Then, for an arbitrary natural number n, the stochastic complexity satisfies the inequality*

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= \begin{cases} (K+H-1)/2 & (\text{ } \mathit{if} \, M=1) \\ (M(H+1)+2K-H-3)/2 & (M \geq 2) \end{cases} \end{aligned}$$

where C is a constant independent of n.

In particular, the upper bounds of stochastic complexities of gaussian mixtures are immediately obtained as corollaries of Theorem 2.

Corollary 1 *If the learning machine is a gaussian mixture given by equations (4.2), (4.3), and (4.4), and if the true distribution is given by the equation (4.8), then*

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= \begin{cases} (K+H-1)/2 & (\text{ } \mathit{if} \, N=1) \\ (N(H+1)+2K-H-3)/2 & (N \geq 2) \end{cases} \end{aligned}$$

where N is the dimension of the input space.

Corollary 2 *If the learning machine is a normal mixture with standard deviations given by equations (4.2), (4.5),and (4.6), and if the true distribution is given by the equation (4.8), then*

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= ((N+1)(H+1)+2K-H-3)/2. \end{aligned}$$

where N is the dimension of inputs.

Let d be the dimension of the parameter w in the learning machine $p(x|w)$. Then Theorem 2 claims that the coefficient of $\log n$ is smaller than

$$\frac{d}{2} = \frac{K-1+MK}{2}.$$

From the equation (2.2) and Theorem 2, if the generalization error $G(n)$ has an asymptotic expansion, then it should satisfy

$$G(n) \leq \frac{\mu}{n}.$$

Hence the generalization error of the mixture model is smaller than that of regular statistical model.

Theorem 1 holds under the condition that the true distribution is completely represented by the mixture of H components ($H \leq K$). In practical applications, since the true distribution can not be represented by any learning machine with finite components in general, it might seem that the conditions (A4.1), (A4.2), and (A4.3) correspond to a special case. However, even if the true distribution is not strictly contained in the learning machine with finite components, Theorem 2 shows the advantage of hierarchical learning machines than regular statistical models (Watanabe, 2001b), because singularities make the sum of the function approximation errors and statistical estimation errors.

4.3 Singularities in Mixture Models

Before the proof of the theorem, let us confirm that mixture models are singular. We illustrate the shape of the true parameters in the parameter space. According to the equation (4.1), the simplest mixture model is written as

$$p(x|w) = a_1 f(x|b_1) + a_2 f(x|b_2),$$

where $a_2 = 1 - a_1$. This learning machine has two components ($K = 2$). For simplicity, the dimension of b_i is one (Figure 4.1 (a)). Assume that $q(x) = f(x|b^*)$, where b^* is a constant. This true distribution has one component (Figure 4.1 (b)). Then, the set of true parameters is

$$\{a_1 = 1, b_1 = b^*\} \cup \{a_1 = 0, b_2 = b^*\} \cup \{b_1 = b_2 = b^*\}.$$

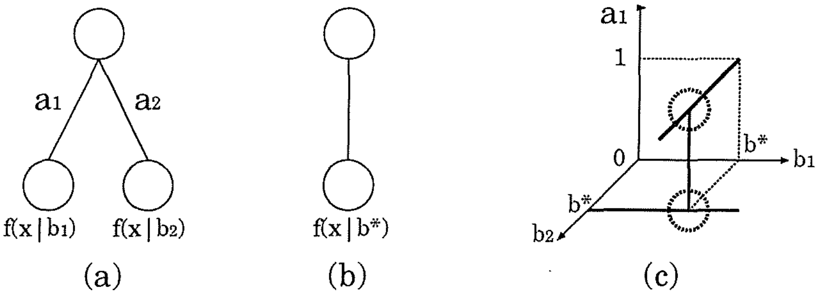

Figure 4.1: (a) The learning machine, (b) The true distribution, (c) The parameter space

This set has singularities $(a_1, b_1, b_2) = (1, b^*, b^*), (0, b^*, b^*)$ (Figure 4.1 (c)). Therefore mixture models are singular even in this simple example.

4.4 Proof of Theorem 2

In this section, we prove Theorem 2.

By the equation (2.6), we have the inequality,

$$F(n) \leq -\log \int \exp(-nH(w))\varphi(w)dw.$$

The Kullback information $H(w)$ from the true distribution $q(x)$ to the learner $p(x|w)$ is rewritten as

$$H(w) = \int \Big\{ \sum_{k=1}^{H} a_k^* f(x|b_k^*) \Big\} \log \frac{\sum_{k=1}^{H} a_k^* f(x|b_k^*)}{\sum_{k=1}^{K} a_k f(x|b_k)} dx.$$

Let us divide the parameter w into $w = (w_1, w_2, w_3)$, where

$$\begin{aligned} w_1 &= (a_1, a_2, \cdots, a_{H-1}), \\ w_2 &= (b_1, b_2, \cdots, b_H, b_K), \\ w_3 &= (a_H, a_{H+1}, \cdots, a_{K-1}, b_{H+1}, \cdots, b_{K-1}). \end{aligned}$$

We introduce three functions.

$$\begin{aligned} H_1(w_1) &= \sum_{k=1}^{H-1} a_k^* \log \frac{a_k^*}{a_k} + a_H^* \log \frac{a_H^*}{1-\sum_{k=1}^{H-1} a_k}, \\ H_2(w_2) &= \sum_{k=1}^{H} a_k^* D(b_k^*||b_k) + a_H^* D(b_H^*||b_K), \\ H_3(w_3) &= \sum_{k=H+1}^{K-1} a_H^* \frac{a_k}{a_H} D(b_H^*||b_k). \end{aligned}$$

If $K = H+1$, then we define $H_3(w_3) = 0$. At first, we show the following lemma.

Lemma 1 *For arbitrary* $w \in W$,

$$H(w) \le H_1(w_1) + H_2(w_2) + H_3(w_3).$$

(Proof of Lemma 1) In general, the following log-sum inequality holds: For arbitrary sequences of positive numbers $\{c_k, k = 1, 2, ..., H\}$ and $\{d_k, k = 1, 2, ..., H\}$,

$$\{\sum_{k=1}^{H} c_k\} \log \frac{\sum_{k=1}^{H} c_k}{\sum_{k=1}^{H} d_k} \le \sum_{k=1}^{H} c_k \log \frac{c_k}{d_k}.$$

By substituting c_k, d_k by

$$\begin{aligned} c_k &= a_k^* f(x|b_k^*) \quad (1 \le k \le H), \\ d_k &= a_k f(x|b_k) \quad (1 \le k \le H-1), \\ d_H &= \sum_{k=H}^{K} a_k f(x|b_k), \end{aligned}$$

it follows that

$$\begin{aligned} H(w) \le & \sum_{k=1}^{H-1} \int a_k^* f(x|b_k^*) \log \frac{a_k^* f(x|b_k^*)}{a_k f(x|b_k)} dx \\ & + \int a_H^* f(x|b_H^*) \log \frac{a_H^* f(x|b_H^*)}{\sum_{k=H}^{K} a_k f(x|b_k)} dx. \end{aligned}$$

Hence, we have an inequality,

$$H(w) \leq R(w) + S(w) + T(w), \tag{4.9}$$

where

$$\begin{aligned} R(w) &= \sum_{k=1}^{H-1} a_k^* \log \frac{a_k^*}{a_k} + a_H^* \log \frac{a_H^*}{\sum\limits_{k=H}^{K} a_k}, \\ S(w) &= \sum_{k=1}^{H-1} a_k^* \int f(x|b_k^*) \log \frac{f(x|b_k^*)}{f(x|b_k)} dx, \\ T(w) &= a_H^* \int f(x|b_H^*) \log \frac{f(x|b_H^*)}{\sum\limits_{k=H}^{K} \alpha_k f(x|b_k)} dx. \end{aligned}$$

Here we used a notation,

$$\alpha_k = \frac{a_k}{\sum\limits_{j=H}^{K} a_j} \quad (k = H, H+1, \cdots, K).$$

First, by using $\sum_{k=1}^{K} a_k = 1$,

$$R(w) = H_1(w_1).$$

Second, by the definition of the Kullback information,

$$S(w) = \sum_{k=1}^{H-1} a_k^* D(b^*||b_k).$$

Third, since $\sum_{k=H}^{K} \alpha_k = 1$, we can apply Jensen's inequality to

$$T(w) = -a_H^* \int f(x|b_H^*) \log\{\sum_{k=H}^{K} \alpha_k \frac{f(x|b_k)}{f(x|b_H^*)}\} dx.$$

It follows that

$$\begin{aligned} T(w) &\leq -\sum_{k=H}^{K} a_H^* \alpha_k \int f(x|b_H^*) \log \frac{f(x|b_k)}{f(x|b_H^*)} dx, \\ &= \sum_{k=H}^{K} a_H^* \alpha_k \int f(x|b_H^*) \log \frac{f(x|b_H^*)}{f(x|b_k)} dx. \end{aligned}$$

Then, by using

$$\begin{aligned} \alpha_H &\leq 1, \quad \alpha_K \leq 1, \\ \alpha_k &\leq \frac{a_k}{a_H} \quad (k = H+1, H+2, \cdots, K-1), \end{aligned}$$

it follows that

$$T(w) \leq a_H^* D(b_H^* || b_H) + a_H^* D(b_H^* || b_K) + H_3(w_3).$$

Hence, by summing up the above results, it follows that

$$\begin{aligned} H(w) &\leq R(w) + S(w) + T(w) \\ &\leq H_1(w_1) + H_2(w_2) + H_3(w_3), \end{aligned}$$

which completes the Lemma 1. (End of Proof)

By using Lemma 1, we can divide the stochastic complexity into three parts. Let $\epsilon > 0$ be a constant used in the assumption (A4.2). We define three sets of parameters

$$\begin{aligned} W_1 &= \{w_1; |a_k - a_k| \leq \epsilon, (k = 1, 2, \cdots, H-1)\} \\ W_2 &= \{w_2; \|b_k - b_k^*\| \leq \epsilon, (k = 1, 2, \cdots, H, K)\} \\ W_3 &= \{w_3; |a_H - a_H^*| \leq \epsilon, \\ &\quad 0 \leq a_k \leq \epsilon, (k = H, H+1, \cdots, K-1), \\ &\quad \|b_k - B_H^*\| \leq \epsilon(k = H+1, H+2, \cdots, K-1)\} \end{aligned}$$

Then, since $\epsilon > 0$ is a sufficiently small constant, w_1, w_2, and w_3 become free variables from each others. In other words, they can be determined without restriction from other variables. According to the division of the parameter $w = (w_1, w_2, w_3)$, we define three partial stochastic complexities $(j = 1, 2, 3)$ by

$$F_j(n) = -\log \int_{W_j'} \exp(-nH_j(w_j)) dw_j,$$

where the integrated regions W_1', W_2', W_3' are taken such that $W_1' \subset W_1$, $W_2' \subset W_2$, $W_3' \subset W_3$, and that

$$W_1' \times W_2' \times W_3' \subset \text{supp } \varphi,$$

where supp φ is the support of the a priori probability distribution. From the assumption of (A4.2),

$$\eta \equiv \inf_{w \in W_1 \times W_2 \times W_3} \varphi(w) > 0.$$

Then the stochastic complexity is bounded by

$$F(n) \le -\log \eta - \sum_{j=1}^{3} \log \int_{W_j} \exp(-nH_j(w_j)) dw_j$$

Therefore,

$$F(n) \le F_1(n) + F_2(n) + F_3(n) + const.,$$

In order to prove Theorem 2, it is sufficient to bound each $F_j(n)$ $(j = 1, 2, 3)$.

It is easy to bound $F_1(n)$ and $F_2(n)$, because they can be bounded by the stochastic complexities of identifiable learning machines.

Lemma 2 *Two partial stochastic complexities satisfy the inequalities,*

$$\begin{aligned} F_1(n) &\le \frac{H-1}{2} \log n + C_1, \\ F_2(n) &\le \frac{(H+1)M}{2} \log n + C_2, \end{aligned}$$

where C_1 and C_2 are constants independent of n.

(Proof of Lemma 2) Let $f(s)$ be a real function of $s \in R^L$ which satisfies

$$f(s) \le c\|s - s_0\|^2$$

in some open set U which contains $s_0 \in R^L$, where $c > 0$ is a constant. $H_1(w)$ and $H_2(w)$ respectively satisfies this condition by putting '$f(s) = H_1(w_1)$, $L = H - 1$' and '$f(s) = H_2(w_2)$, $L = M(H+1)$'. The function

$$S(n) = -\log \int_U \exp(-nf(s)) ds$$

satisfies

$$\begin{aligned} S(n) &\leq -\log \int_U \exp(-cn\|s-s_0\|^2)ds \\ &= \frac{L}{2}\log n - \log \int_{U_n} \exp(-c\|y\|^2)dy \qquad (4.10) \end{aligned}$$

where $U_n = \{y; y/\sqrt{n} + s_0 \in U\}$ converges to R^L as n tends to infinity. By using Lebesgue's convergence theorem, the second term of the right side of the equation (4.10) converges to the constant. (End of Proof)

On the other hand, the set $\{w_3; H_3(w_3) = 0\}$ contains singularities, we need the algebraic geometrical method.

Lemma 3 *The third partial stochastic complexity satisfies the inequality,*

$$F_3(n) \leq (K-H-1)\min\{1, \frac{M}{2}\}\log n + C_3,$$

where C_3 is a constant independent of n.

(Proof of Lemma 3) When $K = H+1$, then $H_3(w_3) = 0$. Hence Lemma 3 holds. Thus, we can assume $K \geq H+2$. By the assumption (A4.3), $D(b_H^*||b_k)$ is a twice continuously differentiable function of b_k. Therefore, in a sufficiently small open set $U \subset R^M$ which contains b_H^*, there exists $c_0 > 0$ such that for any $b_k \in U$,

$$D(b^H||b_k) \leq c_0\|b_H^* - b_k\|^2.$$

Hence, if $b_k \in U$ $(k = H+1, H+2, \cdots, K-1)$, then, by using $c_1 = c_0 a_H^*/(a_H^* - \epsilon)$, if $w_3 \in W_3$,

$$H_3(w_3) \leq H_4(w_3),$$

where

$$H_4(w_3) = \sum_{k=H+1}^{K-1} c_1 a_k \|b_H^* - b_k\|^2.$$

Based on the algebraic geometrical method given in the equations (2.4), (2.5), it is sufficient to show that, with a small constant $\epsilon > 0$,

$$\begin{aligned} J(z) &= \int_{a_H^*-\epsilon}^{a_H^*+\epsilon} da_H \Big[\prod_{k=H+1}^{K-1} \int_0^{\epsilon} da_k \Big] \\ &\quad \times \Big[\prod_{k=H+1}^{K-1} \int_U db_k \Big] \; H_4(w_3)^z \end{aligned}$$

has a pole at $z = -\mu$ which satisfies

$$\mu \leq (K - H - 1)\min\{1, \frac{M}{2}\}.$$

First, a set of parameter S is defined by,

$$S = \{w_3; 0 \leq a_{H+1} \leq \epsilon, a_k \leq a_{H+1}, b_k \in U\}.$$

Then, S is contained in the integrated region of $J(z)$ and the largest pole of

$$J_1(z) = \int_S H_4(w_3)^z dw_3$$

is smaller than the largest pole of $J(z)$. Let us define a new variable w_4 and a mapping

$$g : w_4 = (\alpha, \{\alpha_k\}_{k=H+2}^{K-1}, a_H, \{b_k\}_{k=H+1}^{K-1}) \mapsto w_3$$

by

$$\begin{aligned} a_{H+1} &= \alpha, \\ a_k &= \alpha_k \alpha \quad (k = H+2, H+3, \cdots, K-1). \end{aligned}$$

This mapping is a blowing-up in algebraic geometry. Then the function $H_4(g(w_4))$ divided by α is a constant function of α,

$$H_5(\{\alpha_k\}, a_H, \{b_k\}) \equiv H_4(g(w_4))/\alpha.$$

The Jacobian $|g'(w_4)|$ of the mapping g is

$$|g'(w_4)| = \alpha^{K-H-2}.$$

Thus we can integrate the variable α,

$$\begin{aligned} J_1(z) &= \int_0^\epsilon \alpha^{z+K-H-2} \hat{J}_1(z) d\alpha \\ &= \frac{\epsilon^{K-H-1}}{z+K-H-1} \hat{J}_1(z) \\ \hat{J}_1(z) &= \int H_5(\{\alpha_k\}, a_H, \{b_k\})^z da_H \prod_{k=H+2}^{K-1} d\alpha_k \prod_{k=H+1}^{K-1} db_k. \end{aligned}$$

If z is real and larger than the largest pole of $\hat{J}_1(z)$, the function $\hat{J}_1(z)$ is not equal to zero. Thus the largest pole of $J_1(z)$ is not smaller than $z = -(K-H-1)$. Hence

$$\mu \leq K - H - 1. \tag{4.11}$$

Second, using a small constant $\delta > 0$, the set of parameters T is defined by

$$T = \{w_3; 0 \leq b_{k,j} - b^*_{k,j} \leq b_{H+1,1} - b^*_{H+1,1} \leq \delta, \quad (k,j) \neq (H+1,1)\}$$

Then, the largest pole of

$$J_2(z) = \int_T H_4(w_3)^z dw_3$$

is not larger than the largest pole of $J(z)$. A new variable $(\alpha, \{\beta_{kj}\})$ and the mapping g are defined by

$$g : w_5 = \{\{a_i\}, \alpha, \{\beta_{k,j}\}\} \mapsto w_3$$

where $H \leq i \leq K-1$, $H+1 \leq k \leq K-1$, $1 \leq j \leq M$ and

$$\begin{aligned} b^*_{H,1} - b_{H+1,1} &= \alpha \\ b^*_{H,j} - b_{k,j} &= \alpha\beta_{k,j} \quad ((k,j) \neq (H+1,1)) \end{aligned}$$

Then, the function $H_3(g(w_5))$ divided by α^2 is a constant function of α,

$$H_6(\{a_i\}, \{\beta_{k,j}\}) \equiv H_3(g(w_5))/\alpha^2.$$

The Jacobian is

$$|g'| = \alpha^{M(K-H-1)-1}.$$

Hence,

$$\begin{aligned} J_2(z) &= \int_{a_H^*-\epsilon}^{a_H^*+\epsilon} da_H \Big[\prod_{k=H+1}^{K-1} \int_0^{\epsilon} da_k \Big] \\ &\quad \times \int_0^{\delta} d\alpha \; \alpha^{2z+M(K-H-1)} \\ &\quad \times \Big[\prod_{(k,j)\neq(H+1,1)} \int_0^1 d\beta_{k,j} \Big] \; H_6(\{a_i\}, \{\beta_{k,j}\})^z \end{aligned}$$

has a pole at $z = -M(K-H-1)/2$, resulting that

$$\mu \leq \frac{M(K-H-1)}{2}. \tag{4.12}$$

By combining the inequalities (4.11) and (4.12), we obtain Lemma 3. (End of Proof)

Now, Theorem 2 can be proven by Lemma 1, 2, and 3.

(Proof of Theorem 2) By combining the above lemmas,

$$\begin{aligned} F(n) &\leq \{(H-1)/2 + (H+1)M/2 + \min\{1, M/2\}(K-H-1)\} \log n \\ &= \log n \begin{cases} \frac{1}{2}(H+K-1) & (\text{If } M=1) \\ \frac{1}{2}(M(H+1)+2K-H-3) & (\text{Otherwise}) \end{cases} \end{aligned}$$

which completes Theorem 2. (End of Proof).

(Remark) In the proof, we have used the property that the stochastic complexity becomes larger when we restrict the region of the parameter. This seems to contradict the fact that the more complex learning machine has the larger stochastic complexity. However, this is not a contradiction. In the proof, we used the property that, if $U \supset$, then

$$-\log \int_U \exp(-nK(w))\psi(w)dw \leq -\log \int_V \exp(-nK(w))\psi(w)dw.$$

Remark that, in this inequality, even if $\psi(w)$ is a probability density function on U, it is not a probability density function on V in general. Therefore, this inequality does not compare the complexity of the learning machine using the parameter set U with that of the same learning machine using the parameter set V.

Chapter 5

Boltzmann Machines

In this chapter, we introduce the Boltzmann machines (Rumellhart & McClelland, 1986) and derive a theorem of the stochastic complexity. Boltzmann machines were devised as spin systems in statistical physics for a long time. There are sophisticated learning algorithms using simulated annealing (Ackley et. al., 1985). The model consists of several hidden and observable units, which take the values $\{\pm 1\}$, and the connections between them. In general, it has connections between any two units. However, there are a lot of difficulties to analyze the models because of the connections. In this chapter, let us consider the model which has no connections between hidden units. We refer to this structure as complete bipartite graph-type.

5.1 Complete Bipartite Graph-type Boltzmann Machines

Let observable units and hidden units be $x = \{x_j\}_{j=1}^{M} \in \{-1,1\}^M$ and $h = \{h_i\}_{i=1}^{K} \in \{-1,1\}^K$ respectively. The learning machine $p(x|w)$ is defined by

$$\begin{aligned} p(x|w) &= \frac{\rho_K(x,w)}{Z_K(w)}, & (5.1) \\ \rho_K(x,w) &= \sum_h^K \left\{ \exp\left(\sum_{i=1}^{K} \sum_{j=1}^{M} w_{ij} h_i x_j \right) \right\}, & (5.2) \end{aligned}$$

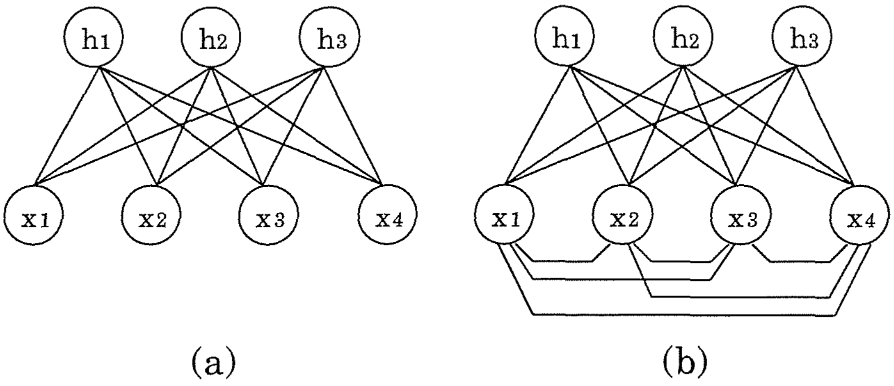

Figure 5.1: Example (K,M)=(3,4). (a) Type I, (b) Type II

$$Z_K(w) = \sum_x \rho_K(x,w), \tag{5.3}$$

where $\sum_h^K$ stands for $\sum_{h_1=\pm1}\cdots\sum_{h_K=\pm1}$ and $\sum_x$ stands for $\sum_{x_1=\pm1}\cdots\sum_{x_M=\pm1}$ respectively, and the parameter w is given by

$$w = \{w_{ij}\} \in R^{K\times M} \qquad (1 \le i \le K, 1 \le j \le M). \tag{5.4}$$

Boltzmann machines generally have all connections between units. However we consider models that have all connections between observable units and hidden units, and no connections between observable units or hidden units. We call this model Type I (Figure 5.1, a). The graph is called the complete bipartite one. We also consider a model defined by

$$p(x|w) = \frac{\gamma_K(x,w)}{Y_K(w)}, \tag{5.5}$$

$$\gamma_K(x,w) = \sum_h^K \left\{ \exp\left(\sum_{i=1}^K \sum_{j=1}^M w_{ij}h_i x_j + \sum_{k=1}^M \sum_{k<l}^M v_{kl}x_k x_l \right) \right\}, \tag{5.6}$$

$$Y_K(w) = \sum_x \gamma_K(x,w), \tag{5.7}$$

where the parameter w is given by

$$w = \{w_{ij}, v_{kl}\} \in R^{K\times M + M(M-1)/2} \quad (1 \le i \le K, 1 \le j \le M, 1 \le k < l \le M). \tag{5.8}$$

This model has all connections not only between observable units and hidden units but among observable units. We call this model Type II (Figure 5.1, b).

We prove the upper bounds of the stochastic complexities for Boltzmann machines Type I in Theorem 3, and those Type II in Corollary 3.

5.2 Stochastic Complexity of Boltzmann Machines

Let us assume two general conditions, (A5.1) and (A5.2).

(A5.1) First, assume that the learning model includes the true distribution. The learner is given by

$$p(x|w) = \frac{\rho_K(x,w)}{Z_K(w)}, \tag{5.9}$$

where $\rho_K(x,w)$ and $Z_K(w)$ is defined by

$$\begin{aligned} \rho_K(x,w) &= \prod_{i=1}^{K} \left\{ e^{-\sum_{j=1}^{M} w_{ij}x_j} + e^{\sum_{j=1}^{M} w_{ij}x_j} \right\}, & (5.10) \\ Z_K(w) &= \sum_{x} \rho_K(x,w). & (5.11) \end{aligned}$$

The set of parameters in $p(x|w)$ is respectively defined by

$$W = \{w_{ij}; 1 \leq i \leq K, 1 \leq j \leq M\}, \tag{5.12}$$

which is a fixed open set of $R^{M\times K}$. Also assume that the true distribution $q(x)$ is given by

$$q(x) = \frac{\rho_H(x,w^*)}{Z_H(w^*)}, \tag{5.13}$$

where $H < K$ and $w^* = \{w^*_{ij}\} \in R^{M\times H}$.

(A5.2) Second, assume that the a priori probability distribution is positive on a true parameter. For a constant $\epsilon > 0$, we define the subset of parameter

$W(\epsilon) \subset W$ by

$$W(\epsilon) = \{\{w_{ij}\} \in W \quad ; \quad |w_{ij} - w^*_{ij}| \leq \epsilon \ (1 \leq i \leq H), \\ |w_{ij}| \leq \epsilon \ (H+1 \leq i \leq K)\}.$$

Assume that there is a constant $\epsilon > 0$ such that

$$\inf_{W(\epsilon)} \varphi(w) > 0,$$

where '$\inf_{W(\epsilon)}$' denotes the infimum value of $\varphi(w)$ in $w \in W(\epsilon)$.

Theorem 3 *Assume the two conditions (A5.1) and (A5.2). Then, for arbitrary natural number n, the stochastic complexity satisfies the inequality*

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= \begin{cases} (K+3H+1)/4 & (\textit{ if } M=2) \\ (K+H)M/4 & (M \geq 3), \end{cases} \end{aligned}$$

where C is a constant independent of n.

Corollary 3 *If the learning machine is Type II, and the true distribution is given by the equations (5.6), (5.7) and*

$$q(x) = \frac{\gamma_H(x,w^*)}{Y_H(w^*)},$$

where $H < K$ and $w^* = \{w^*_{ij}, v^*_{kl}\} \in R^{H \times M + M(M-1)/2}$, then

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= \begin{cases} (K+3H+3)/4 & (\textit{ if } M=2) \\ (K+H+M-1)M/4 & (M \geq 3). \end{cases} \end{aligned}$$

(Remark 1) When the number of observable units is one ($M = 1$),

$$q(x) = p(x|w) = \frac{1}{2^M},$$

for all w. Then, $F(n) = 0$. We will consider $M \geq 2$.

(Remark 2) The case $M \geq 3$ is actually applicable to $M = 2$. However, the upper bound is looser than the case of $M = 2$ in Theorem 3 and Corollary 3.

Let d be the dimension of the parameter w in the learner $p(x|w)$. It is well known that the coefficient of $\log n$ equals $d/2$ for regular statistical models. In these models, coefficients are $KM/2$ and $KM/2 + M(M-1)/4$, respectively. Theorem 3 and Corollary 3 claim that the coefficient μ is smaller than that of the regular model.

5.3 Singularities in Boltzmann Machines

Before the proof of the theorem, let us confirm that Boltzmann machines are singular models. We will illustrate the shape of the true parameters of Boltzmann machines in the parameter space. According to the equations (5.1)-(5.3), the simplest Boltzmann machine is written as

$$p(x|w) = \frac{e^{-wx} + e^{wx}}{\sum_{x'=\pm 1} (e^{-wx'} + e^{wx'})}.$$

This learning machine has one observable unit x, one hidden unit and one connection w between them ($M = 1, K = 1$). Assume that $q(x) = 1/2$. In this learner, all parameters are true since $x \in \{\pm 1\}$. Based on this example, we consider a learning model that has one hidden unit, two observable units x_1 and x_2 and the connections w_1 and w_2 between these hidden and observable units ($M = 2, K = 1$) (Figure 5.2 (a)). The learning model is

$$p(x_1, x_2|w_{11}, w_{12}) = \frac{e^{-w_{11}x_1 - w_{12}x_2} + e^{w_{11}x_1 + w_{12}x_1}}{\sum_{x'_1=\pm 1} \sum_{x'_2=\pm 1} \left(e^{-w_{11}x'_1 - w_{12}x'_2} + e^{w_{11}x'_1 + w_{12}x'_2}\right)}.$$

Assume that $q(x) = 1/4$ (Figure 5.2 (b)). Then, the set of true parameters is

$$\{w_{11} = 0\} \cup \{w_{12} = 0\}.$$

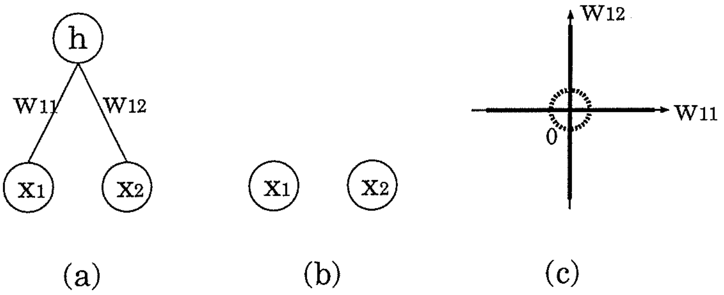

Figure 5.2: (a) The learning machine, (b) The true distribution, (c) The parameter space

This set has a singularity $(w_{11}, w_{12}) = (0, 0)$ (Figure 5.2 (c)). Therefore, Boltzmann machines are singular even in this simple case.

5.4 Proof of Theorem 3

Let the true distribution $q(x)$ and the learner $p(x|w)$ be written as

$$\begin{aligned} q(x) &= \frac{\rho(x, w^*)}{Z(w^*)}, \\ p(x|w) &= \frac{\rho(x, w)}{Z(w)}, \end{aligned}$$

where

$$\begin{aligned} w, w^* &\in R^d, \\ \rho(x, w) &> 0, \\ Z(w) &= \sum_x \rho(x, w). \end{aligned}$$

Let us divide the parameter w and the true parameter w^* into two parts,

$$\begin{aligned} w &= (w_\alpha, w_\beta), \\ w^* &= (w^*_\alpha, w^*_\beta), \end{aligned}$$

such that dimensions of w_α and w_α^* are the same. Assume the following two conditions (a5.1) and (a5.2).

(a5.1) First, assume that the function $\rho(x, w)$ is the product such that

$$\begin{aligned}\rho(x,w) &= \sigma_\alpha(x,w_\alpha)\sigma_\beta(x,w_\beta),\\ \rho(x,w^*) &= \sigma_\alpha(x,w_\alpha^*)\sigma_\beta(x,w_\beta^*),\end{aligned}$$

where $\sigma_\alpha(x, w_\alpha) > 0$ and $\sigma_\beta(x, w_\beta) > 0$.

(a5.2) Second, assume that probability distribution functions defined by

$$\begin{aligned}p_\alpha(x|w_\alpha) &= \frac{\sigma_\alpha(x,w_\alpha)}{\sum_{x'}\sigma_\alpha(x',w_\alpha)},\\ p_\beta(x|w_\beta) &= \frac{\sigma_\beta(x,w_\beta)}{\sum_{x'}\sigma_\beta(x',w_\beta)}\end{aligned}$$

are finite.

Let us define two Kullback informations,

$$\begin{aligned}H_\alpha(w_\alpha) &= \sum_x p_\alpha(x|w_\alpha^*)\log\frac{p_\alpha(x|w_\alpha^*)}{p_\alpha(x|w_\alpha)},\\ H_\beta(w_\beta) &= \sum_x p_\beta(x|w_\beta^*)\log\frac{p_\beta(x|w_\beta^*)}{p_\beta(x|w_\beta)}.\end{aligned}$$

We can derive the following lemma.

Lemma 4 *If $\|w - w^*\| \leq \delta$, where δ is a sufficiently small constant, there exists a sufficiently large constant $N > 0$, such that*

$$H(w) \leq N\{H_\alpha(w_\alpha) + H_\beta(w_\beta)\}. \tag{5.14}$$

In order to prove Lemma 4, we need to look at Lemmas 5, and 6.

Lemma 5 *Let $r(x)$ be a probability distribution function of $x \in \{\pm 1\}^M$ which satisfies*

$$r(x) > 0.$$

Let $F(x)$ be a real function of x. Then, there exists a constant C independent of x, such that

$$\log \sum_x r(x) e^{F(x)} \le \sum_x r(x) F(x) + C \sum_x r(x) F(x)^2.$$

(Proof of Lemma 5)

Let a function $A(y)$ $(y \ge 0)$ be

$$\begin{aligned} A(y) &= \log B(y), \\ B(y) &= \sum_x r(x) e^{yF(x)}. \end{aligned}$$

Using the mean-value theorem, we obtain

$$A(y) = A(0) + yA'(0) + \frac{1}{2} y^2 A''(y^*),$$

where $0 \le y^* \le y$. By substituting y for $y = 1$, it follows that

$$\begin{aligned} A(1) &= 0 + \frac{B'(0)}{B(0)} + \frac{1}{2}\left[\frac{B''(y^*)}{B(y^*)} - \left(\frac{B'(y^*)}{B(y^*)} \right)^2 \right] \\ &\le \sum_x r(x) F(x) + \frac{1}{2} \frac{B''(y^*)}{B(y^*)}. \end{aligned}$$

However,

$$\begin{aligned} \frac{B''(y^*)}{B(y^*)} &= \frac{\sum_x r(x) F(x)^2 e^{y^* F(x)}}{\sum_x r(x) e^{y^* F(x)}} \\ &\le \frac{\sum_x r(x) e^{y^* F(x)} \sum_x F(x)^2}{\sum_x r(x) e^{y^* F(x)}} \\ &\le C_1 \sum_x r(x) F(x)^2, \end{aligned}$$

where

$$C_1 = \frac{1}{\min_x r(x)}.$$

Therefore,

$$\log \sum_x r(x) e^{F(x)} = A(1) \le \sum_x r(x) F(x) + \frac{1}{2C_1} \sum_x r(x) F(x)^2,$$

which completes Lemma 5. (End of Proof)

Lemma 6 *If $\|w - w^*\| \leq \delta$, where δ is a sufficiently small constant and $w, w^* \in R^d$, there exists a sufficiently large constant L_1 independent of w, such that*

$$\sum_x p(x|w^*) \left(\log \frac{p(x|w^*)}{p(x|w)} \right)^2 \leq L_1 \sum_x p(x|w^*) \left(\log \frac{p(x|w^*)}{p(x|w)} \right).$$

(Proof of Lemma 6)

Let $S(t)$ be the function defined by

$$S(t) = t + e^{-t} - 1.$$

Using the mean-value theorem, we can easily show that, when $|t| \leq \epsilon$, where ϵ is a sufficiently small constant,

$$S(t) \geq Ct^2,$$

where C is a sufficiently small constant independent of t. Here, let us define the function,

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

By using the mean-value theorem, it follows that

$$f(x, w) = f(x, w^*) + \sum_{j=1}^{d} (w_j - w_j^*) \frac{\partial f}{\partial w_j}(x, \hat{w}_j),$$

where $0 \leq \hat{w}_j \leq w_j$. From $f(x, w^*) = 0$ and $\|w - w^*\| \leq \delta$,

$$\begin{aligned} |f(x, w)| &\leq \sum_{j=1}^{d} |w_j - w_j^*| |\frac{\partial f}{\partial w_j}(x, \hat{w}_j)| \\ &\leq \sum_{j=1}^{d} |w_j - w_j^*| \sup_{0 \leq \hat{w} \leq w} |\frac{\partial f}{\partial w_j}(x, \hat{w}_j)| \\ &\leq \epsilon, \end{aligned}$$

where ϵ is independent of x and w. Therefore, using $S(t)$, we obtain

$$\begin{aligned} L_1 \sum_x p(x|w^*) \left(\log \frac{p(x|w^*)}{p(x|w)} \right) &= L_1 \sum_x p(x|w^*) S \left(\log \frac{p(x|w^*)}{p(x|w)} \right) \\ &\geq \sum_x p(x|w^*) \left(\log \frac{p(x|w^*)}{p(x|w)} \right)^2, \end{aligned}$$

where L_1 is C^{-1}. (End of Proof)

Second, using Lemmas 5 and 6, we can derive Lemma 4.

(Proof of Lemma 4)

Define some functions as

$$\begin{aligned} f_\alpha(x, w_\alpha) &= \log \frac{p_\alpha(x|w_\alpha)}{p_\alpha(x|w_\alpha^*)}, \\ f_\beta(x, w_\beta) &= \log \frac{p_\beta(x|w_\beta)}{p_\beta(x|w_\beta^*)}, \\ Y(w_\alpha, w_\beta) &= \sum_x p_\alpha(x|w_\alpha) p_\beta(x|w_\beta). \end{aligned}$$

We can easily obtain the following equation,

$$p_\alpha(x|w_\alpha) p_\beta(x|w_\beta) = p_\alpha(x|w_\alpha^*) p_\beta(x|w_\beta^*) e^{f_\alpha(x,w_\alpha)+f_\beta(x,w_\beta)}.$$

The Kullback information given by the equation (2.3) is rewritten as

$$\begin{aligned} H(w) &= \sum_x q(x) \log \frac{Y(w_\alpha, w_\beta) p_\alpha(x|w_\alpha^*) p_\beta(x|w_\beta^*)}{Y(w_\alpha^*, w_\beta^*) p_\alpha(x|w_\alpha) p_\beta(x|w_\beta)} \\ &= \sum_x q(x) \log \frac{p_\alpha(x|w_\alpha^*) p_\beta(x|w_\beta^*)}{p_\alpha(x|w_\alpha^*) p_\beta(x|w_\beta^*) e^{f_\alpha(x,w_\alpha)+f_\beta(x,w_\beta)}} \\ &\quad + \log \frac{Y(w_\alpha, w_\beta)}{Y(w_\alpha^*, w_\beta^*)} \\ &= -\sum_x q(x) \left\{ f_\alpha(x, w_\alpha) + f_\beta(x, w_\beta) \right\} \\ &\quad + \log \sum_x q(x) e^{f_\alpha(x,w_\alpha)+f_\beta(x,w_\beta)}. \end{aligned} \tag{5.15}$$

By applying Lemma 5 to the second term of the equation (5.15), we get

$$\begin{aligned} H(w) &\leq C\sum_x q(x)\left\{f_\alpha(x,w_\alpha)+f_\beta(x,w_\beta)\right\}^2 \\ &\leq 2C\left\{\sum_x q(x)\left\{-f_\alpha(x,w_\alpha)\right\}^2+\sum_x q(x)\left\{-f_\beta(x,w_\beta)\right\}^2\right\} \quad (5.16) \end{aligned}$$

where C is a constant independent of x. Since we assumed that $p_\alpha(x|w_\alpha)$ and $p_\beta(x|w_\beta)$ are finite (a5.2),

$$\begin{aligned} p(x|w^*) &\leq L_2 p_\alpha(x|w^*_\alpha), \\ p(x|w^*) &\leq L_2 p_\beta(x|w^*_\beta), \end{aligned}$$

where L_2 is a sufficiently large constant. Thus, we apply these inequalities and $q(x)=p(x|w^*)$ to the equation (5.16).

$$H(w)\leq 2CL_2\left\{\sum_x p_\alpha(x|w^*_\alpha)\left\{-f_\alpha(x,w_\alpha)\right\}^2+\sum_x p_\beta(x|w^*_\beta)\left\{-f_\beta(x,w_\beta)\right\}^2\right\}.$$

Moreover, by applying Lemma 6 to each term, we can show

$$H(w)\leq -2CL_2\left\{L_1\sum_x p_\alpha(x|w^*_\alpha)f_\alpha(x,w_\alpha)+L_1\sum_x p_\beta(x|w^*_\beta)f_\beta(x,w_\beta)\right\}.$$

This means

$$H(w)\leq N[H_\alpha(w_\alpha)+H_\beta(w_\beta)].$$

(End of Proof)

- Using Lemma 4, we prove Theorem 3.

(Proof of Theorem 3) By the equation (2.6), we have the inequality,

$$F(n)\leq -\log\int \exp(-nH(w))\varphi(w)dw.$$

The Kullback information $H(w)$ from the true distribution $q(x)$ to the learner $p(x|w)$ is rewritten as

$$H(w)=\sum_x \frac{\rho_H(x,w^*)}{Z_H(w^*)}\log\frac{Z_K(w)\rho_H(x,w^*)}{Z_H(w^*)\rho_K(x,w)}.$$

Let us divide the parameter w into $w = (w_1, w_2)$, where

$$\begin{aligned} w_1 &= \{w_{ij}\} \quad (1 \leq i \leq H, \ \ 1 \leq j \leq M), \\ w_2 &= \{w_{ij}\} \quad (H+1 \leq i \leq K, \ \ 1 \leq j \leq M). \end{aligned}$$

Let us introduce two functions.

$$\begin{aligned} H_1(w_1) &= \sum_x \frac{\rho_H(x,w^*)}{Z_H(w^*)} \log \frac{Z_\alpha(w_1)\rho_H(x,w^*)}{Z_H(w^*)\rho_\alpha(x,w_1)}, \\ H_2(w_2) &= \sum_x \frac{1}{2^M} \log \frac{Z_\beta(w_2)}{2^M \rho_\beta(x,w_2)}, \end{aligned}$$

where

$$\begin{aligned} \rho_\alpha(x,w_1) &= \prod_{i=1}^{H} \left\{ e^{-\sum_{j=1}^m w_{ij}x_j} + e^{\sum_{j=1}^m w_{ij}x_j} \right\}, \\ \rho_\beta(x,w_2) &= \prod_{i=H+1}^{K} \left\{ e^{-\sum_{j=1}^m w_{ij}x_j} + e^{\sum_{j=1}^m w_{ij}x_j} \right\}, \\ Z_\alpha(w_1) &= \sum_x \rho_\alpha(x,w_1), \\ Z_\beta(w_2) &= \sum_x \rho_\beta(x,w_2). \end{aligned}$$

By substituting w_α, w_β, w^*_α and w^*_β in Lemma 4 by

$$\begin{aligned} w_\alpha &= w_1, \\ w_\beta &= w_2, \\ w^*_\alpha &= w^*, \\ w^*_\beta &= 0, \end{aligned}$$

it follows that

$$H(w) \leq N\left[H_1(w_1) + H_2(w_2)\right],$$

where N is a sufficiently large constant. We define two sets of parameters

$$\begin{aligned} W_1 &= \{w_1; |w_{ij} - w^*_{ij}| \leq \epsilon, (1 \leq i \leq H, 1 \leq j \leq M)\}, \\ W_2 &= \{w_2; |w_{ij}| \leq \epsilon, (H+1 \leq i \leq K, 1 \leq j \leq M)\}, \end{aligned}$$

where $\epsilon > 0$ is a sufficiently small constant. Then, w_1 and w_2 are free variables from each other. We define partial stochastic complexities by

$$F_k(n) = -\log \int_{W'_k} \exp(-nH_k(w_k))dw_k \quad (k = 1, 2),$$

where the integrated regions W'_1 and W'_2 are taken such that $W'_1 \subset W_1$ and $W'_2 \subset W_2$, and that

$$W'_1 \times W'_2 \subset \mathrm{supp}\varphi,$$

where suppφ is the support of the a priori probability distribution. From the assumption (A5.2),

$$\eta \equiv \inf_{w \in W_1 \times W_2} \varphi(w) > 0.$$

The stochastic complexity is bounded by

$$F(n) \leq -\log \eta - \sum_{k=1}^{2} \log \int_{W_k} \exp(-nH_k(w_k))dw_k.$$

Thus,

$$F(n) \leq F_1(n) + F_2(n) + const..$$

In order to prove Theorem 3, it is sufficient to bind each $F_k(n) \ \ (k = 1, 2)$. It is easy to bind $F_1(n)$, because it can be bound by the stochastic complexity of identifiable models.

Lemma 7 *A partial stochastic complexity satisfies the inequality,*

$$F_1(n) \leq \frac{HM}{2} \log n + C_1,$$

where C_1 is a constant independent of n.

(Proof of Lemma 7)

In an open set W_1, which contains w^*, it follows that

$$H_1(w_1) \leq c\|w_1 - w^*\|^2,$$

where $c > 0$ is a constant. Thus, $F_1(n)$ satisfies

$$\begin{aligned} F_1(n) &\leq -\log \int_{W_1} \exp(-cn\|w_1 - w^*\|^2) dw_1 \\ &= \frac{HM}{2} \log n - \log \int_{W_{1n}} \exp(-c\|y\|^2) dy, \qquad (5.17) \end{aligned}$$

where $W_{1n} = \{y; y/\sqrt{n} + w^* \in W_1\}$ converges to R^{HM} as n tends to infinity. By using Lebesgue's convergence theorem, the second term of the right side of the equation (5.17) converges to the constant. (End of Proof)

However, as the set $\{w_2; H_2(w_2) = 0\}$ includes singularities, we need the algebraic geometrical method.

First, let us look at the case of $M = 2$ and then, general M.

Lemma 8 *When the number of observable units is $M = 2$, a partial stochastic complexity satisfies the inequality,*

$$F_2(n) \leq \frac{K - H + 1}{4} \log n + C_{21},$$

where C_{21} is a constant independent of n.

(Proof of Lemma 8)

We can describe $H_2(w_2)$ as

$$\begin{aligned} H_2(w_2) &= \frac{1}{4} \sum_x \log \frac{\sum_{x'} \prod_{i=H+1}^{K} (e^{-w_{i1}x'_1 - w_{i2}x'_2} + e^{w_{i1}x'_1 + w_{i2}x'_2})}{4 \prod_{i=H+1}^{K} (e^{-w_{i1}x_1 - w_{i2}x_2} + e^{w_{i1}x_1 + w_{i2}x_2})}, \\ &= -\frac{1}{4} \sum_x \log \frac{4 \prod_{i=H+1}^{K} \cosh(w_{i1}x_1 + w_{i2}x_2)}{\sum_{x'} \prod_{i=H+1}^{K} \cosh(w_{i1}x'_1 + w_{i2}x'_2)}. \end{aligned}$$

By using

$$\cosh(\alpha \pm \beta) = \cosh \alpha \cosh \beta \pm \sinh \alpha \sinh \beta,$$

$$\begin{aligned} \prod_{i=H+1}^{K} \cosh(w_{i1}x_1 + w_{i2}x_2) = \prod_{i=H+1}^{K} &\{\cosh w_{i1} \cosh w_{i2} \\ &+ \mathrm{sgn}(x_1 x_2) \sinh w_{i1} \sinh w_{i2}\}. \end{aligned}$$

Let us define some functions,

$$\begin{aligned} \alpha_i &= \cosh w_{i1} \cosh w_{i2}, \\ \beta_i &= \sinh w_{i1} \sinh w_{i2}, \\ \prod_{i=H+1}^{K} (\alpha_i \pm \beta_i) &= \pm g_{odd}(w_2) + g_{even}(w_2) + \prod_{i=H+1}^{K} \alpha_i, \end{aligned}$$

where $g_{odd}(w_2)$ is the sum of terms that have odd β_i as the factor, and $g_{even}(w_2)$ is the sum of terms that have non-zero even β_i. For example,

If $K = H+2$ then

$$g_{odd}(w_2) = \alpha_{H+1}\beta_{H+2} + \beta_{H+1}\alpha_{H+2}.$$

If $K = H+3$ then

$$g_{odd}(w_2) = \alpha_{H+1}\alpha_{H+2}\beta_{H+3} + \alpha_{H+1}\beta_{H+2}\alpha_{H+3} + \beta_{H+1}\alpha_{H+2}\alpha_{H+3} + \beta_{H+1}\beta_{H+2}\beta_{H+3}.$$

Substituting ±1 for each x, we can rewrite $H_2(w_2)$ as

$$\begin{aligned} H_2(w_2) &= -\frac{1}{4}\log \frac{g_{num}(w_2)}{\left[\prod_{i=H+1}^{K} \alpha_i + g_{even}(w_2)\right]^4} \\ &= -\frac{1}{2}\log\left[1 - g(w_2)^2\right], \\ g_{num}(w_2) &= \left[\prod_{i=H+1}^{K} \alpha_i + g_{even}(w_2) + g_{odd}(w_2)\right]^2 \\ &\quad \times \left[\prod_{i=H+1}^{K} \alpha_i + g_{even}(w_2) - g_{odd}(w_2)\right]^2, \\ g(w_2) &= \frac{g_{odd}(w_2)}{\prod_{i=H+1}^{k} \alpha_i + g_{even}(w_2)}. \end{aligned}$$

Based on the property of the stochastic complexity (Proposition. 2), it is sufficient to consider the neighborhood W_2. In this neighborhood, the following equations hold,

$$\lim_{w_2 \to 0} g_{odd}(w_2) = 0,$$
$$\lim_{w_2 \to 0} g_{even}(w_2) = 0.$$

We can derive

$$g(w_2) = \frac{g_{odd}(w_2)}{\prod_{i=H+1}^{K} \alpha_i} + r_1(w_2),$$
$$\lim_{w_2 \to 0} r_1(w_2) \frac{\prod_{i=H+1}^{K} \alpha_i}{g_{odd}(w_2)} = 0,$$

and

$$\begin{aligned} \log(1+x) &= x + r_2(x), \\ \lim_{x \to 0} \frac{r_2(x)}{x} &= 0. \end{aligned} \tag{5.18}$$

Thus,

$$\begin{aligned} H_2(w_2) &\leq \frac{1}{2} \left(\frac{g_{odd}(w_2)}{\prod_{i=H+1}^{K} \alpha_i} \right)^2 + r_2(w_2) \\ &= \frac{1}{2} \left(h_{odd}(w_2) \right)^2 + r_2(w_2), \end{aligned}$$
$$\lim_{w_2 \to 0} r_2(w_2) \left(\frac{\prod_{i=H+1}^{K} \alpha_i}{g_{odd}(w_2)} \right)^2 = 0, \tag{5.19}$$

where $h_{odd}(w_2)$ is the sum of terms that have odd $\tanh w_{i1} \tanh w_{i2}$ as the factor. Because $y = \tanh x$ is one-to-one mapping, its Jacobian is positive definite. Therefore, we regard $\tanh x$ as x,

$$\tanh x = x \tag{5.20}$$

Let us define

$$\begin{aligned} \gamma_i &= w_{i1} w_{i2}, \\ \prod_{i=H+1}^{K} (\gamma_i + 1) &= f_{odd}(w_2) + f_{even}(w_2) + 1, \end{aligned}$$

where $f_{odd}(w_2)$ is the sum of terms that have odd γ_i as the factor, and $f_{even}(w_2)$ is the sum of terms that have non-zero even γ_i as the factor.

For example,

If $K = H + 2$, then

$$f_{odd}(w_2) = \gamma_{H+1} + \gamma_{H+2}.$$

If $K = H + 3$, then

$$f_{odd}(w_2) = \gamma_{H+1} + \gamma_{H+2} + \gamma_{H+3} + \gamma_{H+1}\gamma_{H+2}\gamma_{H+3}.$$

We can rewrite the inequality (5.19) as

$$H_2(w_2) \le \frac{1}{2} f_{odd}(w_2)^2 + r_2(w_2),$$

In the neighborhood W_2, there are two constants, c_1, c_2 and

$$\begin{aligned} \frac{1}{2}\left(f_{odd}(w_2)\right)^2 + r_2(w_2) &\le H_{21}(w_2), \\ H_{21}(w_2) &= c_1 \left(\sum_{i=H+1}^{K} w_{i1}w_{i2} \right)^2 + c_2 \sum_{i=H+2}^{K} \sum_{i<j} \left(w_{i1}w_{j1}\right)^2. \end{aligned}$$

If $K \le H + 2$, then, the second term can be defined to be equal to zero in $H_{21}(w_2)$. We obtain

$$H_2(w_2) \le H_{21}(w_2).$$

According to the property of the stochastic complexity (Proposition. 2),

$$\begin{aligned} F_2(n) &\le F_{21}(n), \\ F_{21}(n) &= -\log \int_{W_2} \exp(-nH_{21}(w_2))dw_2. \end{aligned}$$

In order to clarify the asymptotic expansion of $F_{21}(n)$, consider the zeta function,

$$J_{11}(z) = \int_{W_2} H_{21}(w_2)^z dw_2.$$

Based on the algebraic geometrical method, it is sufficient to prove that $J_{11}(z)$ has the pole $-(K - H + 1)/4$.

Let us define a new variable w_3 and a mapping

$$g : w_3 = (\omega, \{\omega_{i1}\}_{i=H+2}^{K}, \{\omega_{i2}\}_{i=H+1}^{K}) \mapsto w_2$$

by

$$\begin{aligned} w_{H+1,1} &= \omega^2 - \omega \sum_{H+2}^{K} \omega_{i1}\omega_{i2}, \\ w_{i1} &= \omega\omega_{i1} \quad (i = H+2, H+3, \cdots, K), \\ w_{H+1,2} &= \omega_{H+1,2}, \\ w_{i2} &= \omega_{H+1,2}\omega_{i2} \quad (i = H+2, H+3, \cdots, K). \end{aligned}$$

This mapping is a blowing-up in algebraic geometry. This is a partial resolution of a singularity. Then, the function $H_{21}(g(w_3))$ has the factor ω^4. We define the function

$$H_{22}(\{\omega_{i1}\}, \{\omega_{i2}\}) \equiv H_{21}(g(w_3))/\omega^4.$$

This is a constant function of ω. The Jacobian $|g'(w_3)|$ of the mapping g is

$$|g'(w_3)| = \omega^{K-H}.$$

Thus, we can integrate the variable ω,

$$\begin{aligned} J_{12}(z) &= \int_0^{\epsilon} \omega^{4z+K-H} \hat{J}_{12}(z) d\omega \\ &= \frac{\epsilon^{4z+K-H+1}}{4z+K-H+1} \hat{J}_{12}(z), \\ \hat{J}_{12}(z) &= \int H_{22}(\{\omega_{i1}\}, \{\omega_{i2}\})^z d\omega_{H+1,2} \prod_{i=H+2}^{K} d\omega_{i1}\omega_{i2}. \end{aligned}$$

If z is real and larger than the largest pole of $\hat{J}_{12}(z)$, the function $\hat{J}_{12}(z)$ is not equal to zero (Watanabe, 2001b). Thus, the largest pole of $J_{11}(z)$ is not smaller than $z = -(K-H+1)/4$, which completes Lemma 8. (End of Proof)

Second, let us look at the case of $M \geq 3$ and divide the parameter w_2 into $w_2 = (u_{H+1}, u_{H+2}, \cdots, u_K)$, where

$$u_i = \{w_{ij}\} \quad (1 \leq j \leq M).$$

Define the sets of parameters, the Kullback informations and the stochastic complexities.

$$\begin{aligned}
W_{2i} &= \{u_i; |w_{ij}| \leq \epsilon, (1 \leq j \leq M)\}, \\
H_{2i}(u_i) &= \sum_x \frac{1}{2^M} \log \frac{\sum_{x'} \left(e^{-\sum_{j=1}^M w_{ij}x'_j} + e^{\sum_{j=1}^M w_{ij}x'_j}\right)}{2^M \left(e^{-\sum_{j=1}^M w_{ij}x_j} + e^{\sum_{j=1}^M w_{ij}x_j}\right)}, \\
F_{2i}(n) &= -\log \int_{W_{2i}} \exp\left(-nH_{2i}(u_i)\right) du_i \quad (H+1 \leq i \leq K).
\end{aligned}$$

By applying Lemma 4 to $H_2(w_2)$ recursively, we obtain

$$H_2(w_2) \leq N \left[\sum_{i=H+1}^{K} H_{2i}(u_i) \right].$$

Because of the property of the stochastic complexity (Proposition. 2),

$$F_2(n) \leq \sum_{i=H+1}^{K} F_{2i}(n) + const.. \tag{5.21}$$

Thus, it is sufficient to derive the asymptotic expansion of $F_{2i}(n)$.

Lemma 9 *When the number of observable units is $M \geq 3$, partial stochastic complexities satisfy the inequalities,*

$$F_{2i}(n) \leq \frac{M}{4} \log n + C_{2i} \quad (H+1 \leq i \leq K),$$

where C_{2i} are constants independent of n.

(Proof of Lemma 9)

Because $F_{2k}(n)$ corresponds to the stochastic complexity of the Boltzmann machine with one hidden unit, we can rewrite the parameters $\{w_{ij}\}$ as $\{w_j\}$ and $H_{2i}(u_i)$ as

$$H_{2i}(u_i) = -\frac{1}{2^M} \sum_x \log \frac{2^M \cosh\left(\sum_{j=1}^M w_j x_j\right)}{\sum_{x'} \cosh\left(\sum_{j=1}^M w_j x'_j\right)}.$$

By substituting ±1 for each x', we obtain

$$\begin{aligned} H_{2i}(u_i) &= -\frac{1}{2^M}\log f(u_i) + \log \prod_{j=1}^{M} \cosh w_j, \\ f(u_i) &= \prod_{x} \cosh\left(\sum_{j=1}^{M} w_j x_j\right). \end{aligned}$$

By applying the addition theorems of cosh and sinh to $f(u_i)$, we can describe it as

$$f(u_i) = \left[\prod_{j=1}^{M} \cosh w_j\right]^{2^M} + f_1(u_i),$$

where $f_1(u_i)$ is a quantic of $\cosh w_j$ and $\sinh w_j$ with degree $2^M M$. The function $f_1(u_i)$ is the sum of terms that have the even multiplied $\sinh w_j$ as the factor. Let us define the function of u_i,

$$\tau(u_i) = \prod_{j=1}^{M} \cosh w_j.$$

By using this $\tau(u_i)$, we can rewrite $f_1(u_i)$ as

$$f_1(u_i) = \sum_{k=1}^{2^M-1} h_k(u_i)\left[\tau(u_i)\right]^k,$$

where $h_k(u_i)$ is a quantic of $\cosh w_j$ and $\sinh w_j$ with degree $(2^M - k)M$. The function $h_k(u_i)$ is also the sum of terms that have even multiplied $\sinh w_j$ as the factor. In the expansion of $f(u_i)$, the term that has single $\sinh w_j x_j$ vanishes because $x_j = \pm 1$. Thus,

$$h_{2^M-1}(u_i) = 0.$$

Then, $k = 2^M - 2$ is the highest degree of $h_k(u_i)$.

$$h_{2^M-2}(u_i) = c_{2^M-2} \sum_{k=H+1}^{M} \sum_{k<l} \left(\sinh w_k \sinh w_l \prod_{j\neq k,l}^{M} \cosh w_j\right)^2,$$

where c_{2M-2} is a constant independent of u_i. Because

$$\left(\frac{\partial}{\partial w_k}\right)^2 \left(\frac{\partial}{\partial w_l}\right)^2 f(u_i)\Big|_{u_i=0} < 0,$$

we obtain

$$c_{2M-2} < 0.$$

Thus, there is a constant $c_1 > 0$ such that

$$f_1(u_i) \geq -c_1 \sum_{k=H+1}^{M} \sum_{k<l} \left(\sinh w_k \sinh w_l \prod_{j\neq k,l}^{M} \cosh w_j\right)^2 \tau^{2^M-2}(u_i)$$

in the neighborhood W_{2i}. Then, we can rewrite $H_{i2}(u_i)$ as

$$\begin{aligned} H_{2i}(u_i) &= -\frac{1}{2^M}\log f(u_i) + \log\tau(u_i) \\ &\leq -\frac{1}{2^M}\log\left(\tau^{2^M}(u_i) - c_1 \sum_{k=H+1}^{M}\sum_{k<l} g_{kl}^2(u_i)\tau^{2^M-2}(u_i)\right) \\ &\quad +\frac{1}{2^M}\log\tau^{2^M}(u_i) \\ &= -\frac{1}{2^M}\log\left(1 - c_1\sum_{k=H+1}^{M}\sum_{k<l}(\tanh w_k \tanh w_l)^2\right), \\ g_{kl}(u_i) &= \sinh w_k \sinh w_l \prod_{j\neq k,l}^{M}\cosh w_j. \end{aligned}$$

By using the same procedures as in (5.18) and (5.20), there is a constant $c_2 > 0$ such that

$$\begin{aligned} H_{2i}(u_i) &\leq H_{3i}(u_i), \\ H_{3i}(u_i) &= c_2 \sum_{k=H+1}^{M}\sum_{k<l}(w_k w_l)^2. \end{aligned}$$

According to the property of the stochastic complexity (Proposition. 2),

$$\begin{aligned} F_{2i}(n) &\leq F_{3i}(n), \\ F_{3i}(n) &= -\log\int_{W_{2i}}\exp(-nH_{3i}(u_i))du_i. \end{aligned}$$

In order to obtain the asymptotic expansion of $F_{2i}(n)$,let us consider the zeta function,

$$J_{2i}(z) = \int_{W_{2i}} H_{3i}(u_i)^z du_i.$$

It is sufficient to prove that $J_{2i}(z)$ has the pole $\mu = M/4$.

Let us define a new variable v_i and a mapping

$$g : v_i = (\nu, \{\nu_j\}_{j=1}^{M}) \mapsto u_i$$

by

$$\begin{aligned} w_1 &= \nu, \\ w_j &= \nu\nu_j \quad (j = 2, 3, \cdots, M). \end{aligned}$$

Then, the function $H_{3i}(u_i)$ divided by ν^4 is a constant function of ν,

$$H_{4i}(\{\nu_j\}) \equiv H_{3i}(g(u_i))/\nu^4.$$

The Jacobian $|g'(u_i)|$ of the mapping g is

$$|g'(u_i)| = \nu^{M-1}.$$

Thus, we can integrate the variable ν,

$$\begin{aligned} J_{2i}(z) &= \int_0^{\epsilon} \nu^{4z+M-1} \hat{J}_{2i}(z) d\nu \\ &= \frac{\epsilon^{4z+M-1}}{4z + M} \hat{J}_{2i}(z), \\ \hat{J}_{2i} &= \int H_{4i}(\{\nu_j\})^z \prod_{j=2}^{M} d\nu_j. \end{aligned}$$

If z is real and larger than the the largest pole of $\hat{J}_{2i}(z)$, the function $\hat{J}_{2i}(z)$ is not equal to zero. Thus, the largest pole of J_{2i} is not smaller than $z = -M/4$. Therefore, we can obtain Lemma 9. (End of Proof)

Using the equation (5.21) and Lemma 9, we obtain the following lemma,

Lemma 10 *When the number of observable units is $M \geq 3$, a partial stochastic complexity satisfies the inequality,*

$$F_2(n) \leq \frac{(K-H)M}{4} \log n + C_{22},$$

where C_{22} is a constant independent of n.

By combining Lemmas 7, 8 and 10 and the proposition of the stochastic complexity (Proposition. 2), we can consequently show

$$\begin{aligned} F(n) &\leq C + \mu \log n \\ \mu &= \begin{cases} (K+3H+1)/4 & (\text{ if } M=2) \\ (K+H)M/4 & (M \geq 3), \end{cases} \end{aligned}$$

which completes Theorem 3. (End of Proof)

Now let us prove Corollary 3.

(Proof of Corollary 3)

Divide the parameter w into

$$\begin{aligned} w &= (w_1, w_2, v), \\ w_1 &= \{w_{ij}\} \quad (1 \leq i \leq H, 1 \leq j \leq M), \\ w_2 &= \{w_{ij}\} \quad (H+1 \leq i \leq K, 1 \leq j \leq M), \\ v &= \{v_{kl}\} \quad (1 \leq k < l \leq M), \end{aligned}$$

and the true parameter w^* into

$$\begin{aligned} w^* &= (u^*, v^*), \\ u^* &= \{w^*_{ij}\} \quad (1 \leq i \leq H, 1 \leq j \leq M), \\ v^* &= \{v^*_{kl}\} \quad (1 \leq k < l \leq M). \end{aligned}$$

Define the subset of parameter $W(\epsilon) \subset W$ by

$$\begin{aligned} W(\epsilon) = \{\{w_{ij}, v_{kl}\} \;\; ; \;\; & |w_{ij} - w^*_{ij}| \leq \epsilon \;\; (1 \leq i \leq H), \\ & |w_{ij}| \leq \epsilon \;\; (H+1 \leq i \leq M), \\ & |v_{kl} - v^*_{kl}| \leq \epsilon \;\; (1 \leq k < l \leq M)\}. \end{aligned}$$

We can rewrite the equation (5.6) as

$$\begin{aligned}
\gamma_K(x,w) &= \rho_\alpha(x,w_1)\rho_\beta(x,w_2)\varrho_M(x,\{v_{kl}\}), \\
\varrho_M(x,v) &= \exp\left(\sum_{k=1}^{M}\sum_{k<l} v_{kl}x_k x_l\right), \\
Z_M(v) &= \sum_x \varrho_M(x,v).
\end{aligned}$$

Because the assumptions (a1) and (a2) satisfied, it follows that

$$\begin{aligned}
H(w) &= N[H_1(w_1)+H_2(w_2)+H_3(v)], \\
H_3(v) &= \sum_x \frac{\varrho_M(x,v^*)}{Z_M(v^*)}\log\frac{Z_M(v)\varrho_M(x,v^*)}{Z_M(v^*)\varrho_M(x,v)}.
\end{aligned}$$

We need to clarify a partial stochastic complexity defined by

$$\begin{aligned}
F_3(n) &= -\log\int_{W_3}\exp(-nH_3(v))dv, \\
W_3 &= \{v_{kl}; |v_{kl}-v^*_{kl}|\leq \epsilon \quad (1\leq k<l\leq M)\}.
\end{aligned}$$

Using the same procedures as the proof for Lemma 7, we can immediately derive

$$F_3(n) = \frac{M(M-1)}{4}\log n + C_3, \tag{5.22}$$

where C_3 is a constant independent of n, since the number of parameters of v is equal to $M(M-1)/2$. By combining the equation (5.22) and Theorem 3, we obtain

$$\begin{aligned}
F(n) &\leq C+\mu\log n \\
\mu &= \begin{cases} (K+3H+3)/4 & (\text{ if } M=2) \\ (K+H+M-1)M/4 & (M\geq 3). \end{cases}
\end{aligned}$$

(End of Proof)

Chapter 6

Bayesian Networks

In this chapter, we introduce Bayesian networks and derive a theorem of the stochastic complexity. The graphical model consists of several observable and hidden nodes and the connections between them. The nodes mean the random variables and the connections do the correlations or causal relationships. Especially, the models are referred to as Bayesian networks when the graphs are directed acyclic (Pearl, 1988). They are widely used in the data mining, the fault diagnosis of a system and software of accessibility options. There is an algorithm so called belief propagation to calculate the conditional probability among several nodes. However, the computation of the algorithm is NP hard (Cooper, 1990) except for particular structure models such as singly connected graphs (Lauritzen & Spiegelhalter, 1988). Thus, some approximations using the junction tree, the variational method or the mean field approximation are developed. Despite these learning algorithms, the properties of the generalization are still unknown. The results in this chapter provide a theoretical base for it.

6.1 General Naive Bayesian Networks

Let x be observable nodes, and $h = \{h_k\}_{k=1}^{K}$ be hidden nodes. Let us assume each hidden node h_k has T_k states, and describe that $h_k \in \{1, 2, \cdots, T_k\}$.

Then, the learning model is defined by

$$p(x|w) = \sum_{i_1=1}^{T_1} \cdots \sum_{i_K=1}^{T_K} a_{1i_1} \cdots a_{Ki_K} F(x|b_{i_1 i_2 \cdots i_K}), \tag{6.1}$$

where $F(x|b_{i_1 \cdots i_K})$ is some conditional probability. The parameter w is given by

$$\begin{aligned} w &= \{a, b\}, \\ a &= \{a_{ki_k}\} \quad (1 \leq k \leq K, 2 \leq i_k \leq T_k), \\ b &= \{b_{i_1 i_2 \cdots i_K, j}\} \quad (1 \leq j \leq M). \end{aligned}$$

Then,

$$a_{k1} = 1 - \sum_{i=2}^{T_k} a_{ki} \quad (1 \leq k \leq K). \tag{6.2}$$

The dimension of w is

$$\sum_{k=1}^{K} (T_k - 1) + M \prod_{k=1}^{K} T_k. \tag{6.3}$$

We show the upper bounds of the stochastic complexities of the model represented by equations (6.1) and (6.2).

(Remark 1) If $F(x|b_{i_1 \cdots i_K})$ is given by

$$\begin{aligned} F(x|b_{i_1 i_2 \cdots i_K}) &= \prod_{j=1}^{N} \prod_{l=1}^{Y_j} (b_{i_1 i_2 \cdots i_K, jl})^{\delta(x_j - l)}, \qquad (6.4) \\ \delta(n) &= \begin{cases} 1 & (\text{ if } n = 0) \\ 0 & (\text{otherwise}), \end{cases} \end{aligned}$$

the model is the Bayesian network, which has observable nodes $x = \{x_j\}_{j=1}^{N}$. We assume that each observable node x_j has Y_j states and describe $x_j \in \{1, 2, \cdots, Y_j\}$ (Figure 6.1). Then,

$$b_{i_1 i_2 \cdots i_K, j1} = 1 - \sum_{l=2}^{Y_j} b_{i_1 i_2 \cdots i_K, jl} \quad (1 \leq j \leq N). \tag{6.5}$$

The dimension of the parameter in $F(x|b_{i_1 \cdots i_K})$ is

$$M = \sum_{j=1}^{N} (Y_j - 1). \tag{6.6}$$

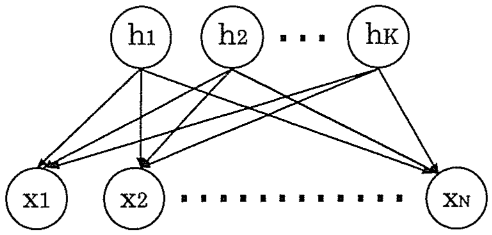

Figure 6.1: The Bayesian network.

6.2 Stochastic Complexity of Bayesian Networks

We assume the following three conditions, (A6.1), (A6.2) and (A6.3).
(A6.1) First, assume that the learning model attains the true distribution. The true distribution has H hidden nodes ($H \leq K$), and each hidden node h_k has S_k states, where $S_k \leq T_k$. In other words, there exists the true parameter w^*, such that

$$\begin{aligned} q(x) &= p(x|w^*), \\ p(x|w^*) &= \sum_{i_1=1}^{S_1} \cdots \sum_{i_H=1}^{S_H} a^*_{1i_1} \cdots a^*_{Hi_H} F(x|b^*_{i_1 \cdots i_H}). \end{aligned} \qquad (6.7)$$

Thus

$$\begin{aligned} w^* &= \{a^*, b^*\}, \\ a^* &= \{a^*_{ki_k}\} \quad (1 \leq k \leq H, 2 \leq i_k \leq S_k), \\ b^* &= \{b^*_{i_1 i_2 \cdots i_H, j}\} \quad (1 \leq j \leq M), \end{aligned}$$

and

$$a^*_{k1} = 1 - \sum_{i=2}^{S_k} a^*_{ki} \quad (1 \leq k \leq H). \qquad (6.8)$$

(Remark 2) In the case of Bayesian networks,

$$F(x|b^*_{i_1 i_2 \cdots i_H}) = \prod_{j=1}^{N} \prod_{l=1}^{Y_j} (b^*_{i_1 i_2 \cdots i_H, jl})^{\delta(x_j - l)}, \qquad (6.9)$$

$$b^*_{i_1 i_2 \cdots i_H, j1} = 1 - \sum_{l=2}^{Y_j} b^*_{i_1 \cdots i_H, jl} \quad (1 \leq j \leq N). \tag{6.10}$$

(A6.2) Second, assume that the a priori probability distribution is positive on the true parameter. For a constant $\epsilon > 0$, let us define the subset of parameter $W(\epsilon) \subset W$ by

$$\begin{aligned} W(\epsilon) = \{\{a, b\} \in W; \\ & |a_{ki_k} - a^*_{ki_k}| \leq \epsilon \ (1 \leq k \leq H, 2 \leq i_k \leq S_k), \\ & |a_{ki_k}| \leq \epsilon \ (\text{otherwise}), \\ & |b_{i_1 i_2 \cdots i_H 11 \cdots 1, j} - b^*_{i_1 i_2 \cdots i_H, j}| \leq \epsilon \\ & \quad (1 \leq i_m \leq S_m, 1 \leq m \leq H, 1 \leq j \leq M), \\ & |b_{i_1 i_2 \cdots i_H \cdots i_K, j} - b^*_{11 \cdots 1, j}| \leq \epsilon \ (\text{otherwise})\}. \end{aligned}$$

Suppose that there is a constant $\epsilon > 0$ such that

$$\inf_{W(\epsilon)} \varphi(w) > 0,$$

where 'inf$_{W(\epsilon)}$' denotes the infimum value of $\varphi(w)$ in $w \in W(\epsilon)$.

(A6.3) Third, let us define the Kullback informations

$$\begin{aligned} & D(i_1 i_2 \cdots i_H || i_1 i_2 \cdots i_H, i_{H+1} \cdots i_K) \\ & = \int dx F(x | b^*_{i_1 \cdots i_H}) \log \frac{F(x | b^*_{i_1 \cdots i_H})}{F(x | b_{i_1 \cdots i_H i_{H+1} \cdots i_K})}. \end{aligned}$$

We assume that they are analytic and finite on the support of $\varphi(w)$.

Theorem 4 *Assume the conditions, (A6.1), (A6.2) and (A6.3). If the learning machine is given by equations (6.1) and (6.2), and the true distribution is given by equations (6.7) and (6.8), then for arbitrary natural number n, the stochastic complexity satisfies the inequality*

$$\begin{aligned} & F(n) \leq C + \mu \log n, \\ & \mu = \frac{1}{2} M \prod_{k=1}^{H} S_k - \frac{1}{2} \sum_{k=1}^{H} S_k + \frac{1}{2} H + \sum_{k=1}^{K} T_k - K, \end{aligned}$$

where C is a constant independent of n.

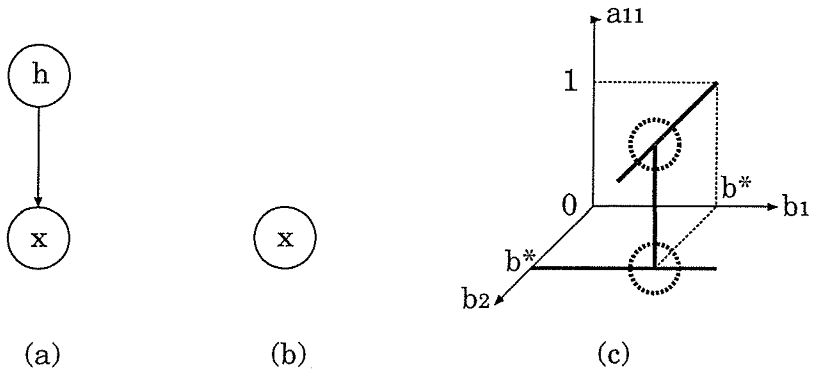

Figure 6.2: (a) The learning machine, (b) The true distribution, (c) The parameter space

6.3 Singularities in Bayesian Networks

Before the proof of the theorem, let us confirm that Bayesian networks are singular. We illustrate the shape of the true parameters in the parameter space. According to the equations (6.1), (6.2), (6.4) and (6.5), the simplest model is written as

$$p(x|w) = a_{11}(b_1^x(1-b_1)^{1-x}) + a_{12}(b_2^x(1-b_2)^{1-x}),$$

where $a_{12} = 1 - a_{11}$ and $x \in \{0, 1\}$. This model has one hidden node and one observable node (Figure 4.1 (a)). Assume that

$$q(x) = b^{*x}(1-b^*)^{1-x}.$$

This true distribution has only one observable node (Figure 6.2 (b)). Then, the set of the true parameters is

$$\{a_{11} = 1, b_1 = b^*\} \cup \{a_{11} = 0, b_2 = b^*\} \cup \{b_1 = b_2 = b^*\}.$$

This set has singularities $(a_{11}, b_1, b_2) = (1, b^*, b^*), (0, b^*, b^*)$. Therefore Bayesian networks are singular even in this simple example.

6.4 Proof of Theorem 4

(Remark 3) If the model is the Bayesian network, we replace $\int dx$ with $\sum_{x_1=1}^{Y_1}\sum_{x_2=1}^{Y_2}\cdots\sum_{x_N=1}^{Y_N}$. The following proof is correctly derived independently of this replacing.

The Kullback information (2.3) is rewritten as

$$H(w)=\int dx\left[\left\{\prod_{k=1}^{H}\sum_{i_k=1}^{S_k}\right\}a^*_{1i_1}\cdots a^*_{Hi_H}F(x|b^*_{i_1i_2\cdots i_H})\right]$$
$$\times\log\frac{\left\{\prod_{k=1}^{H}\sum_{i_k=1}^{S_k}\right\}a^*_{1i_1}\cdots a^*_{Hi_H}F(x|b^*_{i_1\cdots i_H})}{\left\{\prod_{k=1}^{K}\sum_{i_k=1}^{T_k}\right\}a_{1i_1}\cdots a_{Ki_K}F(x|b_{i_1\cdots i_K})},$$

where

$$\left\{\prod_{k=1}^{K}\sum_{i_k=1}^{T_k}\right\}\equiv\sum_{i_1=1}^{T_1}\sum_{i_2=1}^{T_2}\cdots\sum_{i_K=1}^{T_K}.$$

Let us divide the parameter w into $w=\{w_1,w_2\}$, where

$$\begin{aligned}w_1 &= \{a_{ki_k};1\le k\le H, 2\le i_k\le S_k,\\ &\quad b_{i_1i_2\cdots i_K,j};1\le i_k\le S_k, 1\le k\le H,\\ &\quad i_{H+1}=i_{H+2}=\cdots=i_K=1,\\ &\quad 1\le j\le M\},\\ w_2 &= \{a_{ki},b_{i_1i_2\cdots i_K,j};\text{otherwise}\},\end{aligned}$$

and define two functions,

$$H_1(w_1)=\left\{\prod_{k=1}^{H}\sum_{i_k=1}^{S_k}\right\}\left[a^*_{1i_1}\cdots a^*_{Hi_H}\log\frac{a^*_{1i_1}\cdots a^*_{Hi_H}}{\gamma_{1i_1}\cdots\gamma_{Hi_H}}\right]$$
$$+\left\{\prod_{k=1}^{H}\sum_{i_k=1}^{S_k}\right\}a^*_{1i_1}\cdots a^*_{Hi_H}$$
$$\times D(i_1\cdots i_H||i_1\cdots i_H,11\cdots 1),$$
$$H_2(w_2)=\sum_{k=1}^{H}\sum_{i_k=S_k+1}^{T_k}c_{ki_k}a_{ki_k}+\sum_{k=H+1}^{K}\sum_{i_k=1}^{T_k}c_{ki_k}a_{ki_k}$$

$$
\begin{aligned}
&+c_0 \prod_{k=1}^{H} a^*_{k1} \left\{ \prod_{k=1}^{K} \sum_{i_k=1}^{T_k} \right\} \chi(i_1, \cdots, i_K) \\
&\quad \times D(1 \cdots 1 || i_1 \cdots i_H, i_{H+1} \cdots i_K),
\end{aligned}
$$

where $\{c_{ki_k}\}$ and c_0 are positive constants and

$$
\gamma_{ki_k} = \begin{cases} a_{ki_k} & (i_k \neq 1) \\ 1 - \sum_{i'_k=2}^{S_k} a_{ki'_k} & (i_k = 1) \end{cases},
$$

$$
\chi(i_1, \cdots, i_K) = \begin{cases} 0 & (a_{ki_k} \in W_1; 1 \leq k \leq H, \\ & \quad i_{H+1} = \cdots = i_K = 1) \\ a_{\bar{k}i_{\bar{k}}} & \left(\bar{k} = \min_k \{k; a_{ki_k} \in W_2\}\right) \end{cases}.
$$

Let us prove the following lemma,

Lemma 11 *For arbitrary $w \in W(\epsilon)$,*

$$
H(w) \leq H_1(w_1) + H_2(w_2).
$$

(Proof of Lemma 11)

In this proof, we use the notations for summations,

$$
\begin{aligned}
\sum_1 &\equiv \left\{ \prod_{k=1}^{H} \sum_{i_k=1}^{S_k} \right\} \left\{ \prod_{k=H+1}^{K} \sum_{i_k=1}^{1} \right\} - \left\{ \prod_{k=1}^{K} \sum_{i_k=1}^{1} \right\}, \\
\sum_2 &\equiv \left\{ \prod_{k=1}^{K} \sum_{i_k=1}^{T_k} \right\} - \left\{ \prod_{k=1}^{H} \sum_{i_k=1}^{S_k} \right\} \left\{ \prod_{k=H+1}^{K} \sum_{i_k=1}^{1} \right\} \\
&\quad + \left\{ \prod_{k=1}^{K} \sum_{i_k=1}^{1} \right\}.
\end{aligned}
$$

In general, the following log-sum inequality holds: For arbitrary sequences of positive numbers $\{d_k, k = 1, 2, \cdots, I\}$ and $\{e_k, k = 1, 2, \cdots, I\}$,

$$
\left\{ \sum_{k=1}^{I} d_k \right\} \log \frac{\left\{ \sum_{k=1}^{I} d_k \right\}}{\left\{ \sum_{k=1}^{I} e_k \right\}} \leq \sum_{k=1}^{I} \left\{ d_k \log \frac{d_k}{e_k} \right\}.
$$

Thus, for arbitrary sequences of positive numbers $\{d_k, k = 1, 2, \cdots, I\}$ and $\{e_k, k = 1, 2, \cdots, I'\}$, where $I < I'$, it follows that

$$\begin{aligned}
&\left\{\sum_{k=1}^{I} d_k\right\} \log \frac{\left\{\sum_{k=1}^{I} d_k\right\}}{\left\{\sum_{k=1}^{I'} e_k\right\}} \\
&\leq \left\{\sum_{k=1}^{I} - \sum_{k=1}^{1}\right\} \left\{d_k \log \frac{d_k}{e_k}\right\} \\
&\quad +d_1 \log \frac{d_1}{\{\sum_{k=1}^{I'} - \sum_{k=1}^{I} + \sum_{k=1}^{1}\} e_k}.
\end{aligned}$$

Using this inequality, we obtain

$$\begin{aligned}
H(w) \leq \int dx \Bigg[&\sum_{1} a^*_{1i_1} \cdots a^*_{Hi_H} F(x|b^*_{i_1 \cdots i_H}) \\
&\times \log \frac{a_{1i_1} \cdots a^*_{Hi_H} F(x|b^*_{i_1 \cdots i_H})}{a_{1i_1} \cdots a_{K1} F(x|b_{i_1 \cdots i_K})} \\
&+a^*_{11} a^*_{21} \cdots a^*_{H1} F(x|b^*_{11 \cdots 1}) \\
&\times \log \frac{a^*_{11} a^*_{21} \cdots a^*_{H1} F(x|b^*_{11 \cdots 1})}{Z_1(x)} \Bigg],
\end{aligned}$$

where

$$Z_1(x) = \sum_{2} a_{1i_1} a_{2i_2} \cdots a_{Ki_K} F(x|b_{i_1 i_2 \cdots i_K}).$$

Let us define two functions,

$$\begin{aligned}
R_1(w) = \int dx \Bigg[&\sum_{1} a^*_{1i_1} \cdots a^*_{Hi_H} F(x|b^*_{i_1 i_2 \cdots i_H}) \\
&\times \log \frac{a^*_{1i_1} \cdots a^*_{Hi_H} F(x|b^*_{i_1 \cdots i_H})}{a_{1i_1} \cdots a_{K1} F(x|b_{i_1 \cdots i_K})} \Bigg], \\
R_2(w) = \int dx \; &a^*_{11} a^*_{21} \cdots a^*_{H1} F(x|b^*_{11 \cdots 1}) \\
&\times \log \frac{a^*_{11} a^*_{21} \cdots a^*_{H1} F(x|b^*_{11 \cdots 1})}{Z_1(x)}.
\end{aligned}$$

Then,

$$H(w) \leq R_1(w) + R_2(w). \tag{6.11}$$

Let us use the following notations,

$$\begin{aligned} \rho_{i_1 i_2 \cdots i_K} &= \frac{a_{i_1} a_{i_2} \cdots a_{i_K}}{\sigma} \\ \sigma &= \sum_2 a_{1i_1} a_{2i_2} \cdots a_{Ki_K}. \end{aligned}$$

Then $\rho_{i_1 i_2 \cdots i_K}$ is a probability distribution on the set of suffixes of $\sum_2$. We can rewrite $R_2(w)$ as

$$\begin{aligned} R_2(w) &= \int dx a^*_{11} a^*_{21} \cdots a^*_{H1} F(x|b^*_{11\cdots 1}) \\ &\quad \times \left\{ \log \frac{a^*_{11} a^*_{21} \cdots a^*_{H1}}{\sigma} + \log \frac{F(x|b^*_{11\cdots 1})}{Z_2(x)} \right\}, \end{aligned}$$

where

$$Z_2(x) = \sum_2 \rho_{i_1 i_2 \cdots i_K} F(x|b_{i_1 i_2 \cdots i_K}).$$

Applying Jensen's inequality to $R_2(w)$, we obtain

$$\begin{aligned} R_2(w) \leq & \sum_2 \rho_{i_1 \cdots i_K} D(11 \cdots 1 || i_1 \cdots i_H, i_{H+1} \cdots i_K) \\ & + a^*_{11} a^*_{21} \cdots a^*_{H1} \log \frac{a^*_{11} a^*_{21} \cdots a^*_{H1}}{\sigma}. \end{aligned} \quad (6.12)$$

In the region $W(\epsilon)$, there is a constant c_1 such that

$$\begin{aligned} \rho_{i_1 i_2 \cdots i_K} &\leq \frac{a_{ki_k}}{\sigma} \qquad (1 \leq \forall k \leq K) \\ &\leq \frac{a_{ki_k}}{c_1}. \end{aligned} \quad (6.13)$$

We can easily obtain

$$\begin{aligned} a^*_{k1} \log \frac{a^*_{k1}}{a_{k1}} &\leq \left(1 - \sum_{i_k=2}^{S_k} a^*_{ki_k}\right) \log \frac{1 - \sum_{i_k=2}^{S_k} a^*_{ki_k}}{1 - \sum_{i_k=2}^{S_k} a_{ki_k}} \\ &\quad + c_{2k} \sum_{i_k=S_k+1}^{T_k} a_{ki_k}, \end{aligned} \quad (6.14)$$

for $w \in W(\epsilon)$ and $k \leq H$, where c_{2k} is a positive constant, and

$$\log \frac{1}{a_{k1}} \leq c_{3k} \sum_{i_k=S_k+1}^{T_k} a_{ki_k}, \quad (6.15)$$

for $w \in W(\epsilon)$ and $k \geq H+1$, where c_{3k} is a positive constant. In the inequality (6.12), each $\rho_{i_1 \cdots i_K}$ have $a_{ki_k} \in W_2$ as the factor. Using $0 \leq \rho_{i_1 i_2 \cdots i_K} \leq 1$, (6.13),(6.14) and (6.15), we obtain

$$\begin{aligned}
R_1(w) \leq & \sum_{k=1}^{H} \sum_{i_k=S_k+1}^{T_k} c''_{ki_k} a_{ki_k} + \sum_{k=H+1}^{K} \sum_{i_k=2}^{T_k} c''_{ki_k} a_{ki_k} \\
& + \sum_{1} a^*_{1i_1} a^*_{2i_2} \cdots a^*_{Hi_H} \log \frac{a^*_{1i_1} a^*_{2i_2} \cdots a^*_{Hi_H}}{\gamma_{1i_1} \gamma_{2i_2} \cdots \gamma_{Hi_H}} \\
& + \sum_{1} a^*_{1i_1} \cdot \cdot a^*_{Hi_H} D(i_1 \cdots i_H || i_1 \cdots i_H, 1 \cdots 1), \\
R_2(w) \leq & a^*_{11} a^*_{21} \cdots a^*_{H1} \log \frac{a^*_{11} a^*_{21} \cdots a^*_{H1}}{\gamma_{11} \gamma_{22} \cdots \gamma_{H1}} \\
& + \sum_{k=1}^{H} \sum_{i_k=S_k+1}^{T_k} c'_{ki_k} a_{ki_k} + \sum_{k=H+1}^{K} \sum_{i_k=2}^{T_k} c'_{ki_k} a_{ki_k} \\
& + \frac{1}{c_1} \left\{ \prod_{k=1}^{K} \sum_{i_k=1}^{T_k} \right\} \chi(i_1, i_2, \cdots, i_K) \\
& \times D(11 \cdots 1 || i_1 i_2 \cdots i_H, i_{H+1} \cdots i_K) \\
& + D(11 \cdots 1 || 11 \cdots 1, 11 \cdots 1),
\end{aligned}$$

where $\{c'_{ki_k}\}$ and $\{c''_{ki_k}\}$ are positive constants. By combining the above inequalities, (6.11) and $R_1(w)+R_2(w) \leq H_1(w_1)+H_2(w_2)$, we obtain Lemma 11. (End of Proof)

Let us define two sets of the parameters

$$\begin{aligned}
W_1 = & \{w_1; |a_{ki_k} - a^*_{ki_k}| \leq \epsilon \\
& \qquad (1 \leq k \leq H, 2 \leq i_k \leq S_k), \\
& |b_{i_1 i_2 \cdots i_H 11 \cdots 1, j} - b^*_{i_1 i_2 \cdots i_H, j}| \leq \epsilon \\
& \qquad (1 \leq k \leq H, 1 \leq i_k \leq S_k, 1 \leq j \leq M)\}, \\
W_2 \ = & \ \{w_2; |a_{ki_k}| \leq \epsilon, |b_{i_1 i_2 \cdots i_K, j} - b^*_{11 \cdots 1, j}| \leq \epsilon \\
& \qquad \text{(otherwise)}\}.
\end{aligned}$$

$w_1 \in W_1$ and $w_2 \in W_2$ are free variables from each other. Also define the partial stochastic complexities,

$$F_i(n) = -\log \int_{W_i'} \exp(-nH_i(w_i))dw_i \quad (i = 1, 2),$$

where the integrated region W_1 and W_2 are taken such that $W_1' \subset W_1$ and $W_2' \subset W_2$, respectively, and that

$$W_1' \times W_2' \subset \text{supp}\varphi(w),$$

where $\varphi(w)$ is the support of the a priori distribution. From the assumption (A6.2),

$$\eta \equiv \inf_{w \in W_1 \times W_2} \varphi(w) > 0.$$

The stochastic complexity is bounded by

$$F(n) \leq -\log \eta - \sum_{i=1}^{2} \log \int_{W_i} \exp(-nH_i(w_i))dw_i.$$

Thus,

$$F(n) \leq F_1(n) + F_2(n) + const.$$

In order to prove Theorem 4, it is sufficient to bound each $F_i(n) \ \ (i = 1, 2)$. It is easy to bound $F_1(n)$, because it is not larger than the stochastic complexity of identifiable models.

Lemma 12 *A partial stochastic complexity satisfies the inequality,*

$$F_1(n) \leq \frac{1}{2} \left\{ M \prod_{k=1}^{H} S_k + \sum_{k=1}^{H} (S_k - 1) \right\} \log n + C_1,$$

where C_1 is a constant independent of n.

(Proof of Lemma 12)

In an open set W_1, which contains w^*, it follows that

$$H_1(w_1) \leq c\|w_1 - w^*\|^2,$$

where $c > 0$ is a constant. Thus, $F_1(n)$ satisfies

$$\begin{aligned} F_1(n) &\leq -\log \int_{W_1} \exp(-cn\|w_1 - w^*\|^2) dw_1 \\ &= \frac{\bar{d}_1}{2} \log n - \log \int_{W_{1n}} \exp(-c\|y\|^2) dy, \qquad (6.16) \\ \bar{d}_1 &= \left\{ M \prod_{k=1}^{H} S_k + \sum_{k=1}^{H} (S_k - 1) \right\}, \end{aligned}$$

where $\bar{d}_1$ is the dimension of W_1 and $W_{1n} = \{y; y/\sqrt{n} + w^* \in W_1\}$ converges to $R^{\bar{d}_1}$ as n tends to infinity. By using Lebesgue's convergence theorem, the second term of the right side of the equation (6.16) converges to the constant. (End of Proof)

However, the set $\{w_2; H_2(w_2) = 0\}$ includes singularities, we apply the algebraic geometrical method to $F_2(n)$.

Lemma 13 *The second partial stochastic complexity satisfies the inequality,*

$$F_2(n) \leq \left\{ \sum_{k=1}^{H} (T_k - S_k) + \sum_{k=H+1}^{K} (T_k - 1) \right\} \log n + C_2,$$

where C_2 is a constant independent of n.

(Proof of Lemma 13)

In order to clarify the asymptotic expansion of $F_2(n)$, we consider the zeta function,

$$J(z) = \int_{W_2} H_2(w_2)^z dw_2.$$

Based on the algebraic geometrical method, we need to show that this zeta function has a pole,

$$z = -\left\{ \sum_{k=1}^{H} (T_k - S_k) + \sum_{k=H+1}^{K} (T_k - 1) \right\}.$$

According to the definition of $\chi(i_1, i_2, \cdots, i_K)$, all the terms of $H_2(w_2)$ have a_{ki_k} as the factor. Now we define a variable w_3 and a mapping

$$g : w_3 = (\omega, \{\omega_{ki_k}\}, \{b_{i_1 i_2 \cdots i_K, j}\}) \mapsto w_2$$

by

$$\begin{aligned}
\omega &= a_{KT_K}, \\
\omega\omega_{ki_k} &= a_{ki_k} \ (1 \le k \le H, S_k + 1 \le i_k \le T_k), \\
\omega\omega_{ki_k} &= a_{ki_k} \ (H+1 \le k \le K-1, 2 \le i_k \le T_k), \\
\omega\omega_{Ki_K} &= a_{Ki_K} \ (2 \le i_K \le T_K - 1).
\end{aligned}$$

This mapping is called a blow-up in algebraic geometry. The function $H(g(w_3))$ divided ω is a constant function of ω,

$$H_3(\{\omega_{ki_k}\}, \{b_{i_1\cdots i_K,j}\}) \equiv H_2(w_2)/\omega.$$

The Jacobian $|g'(w_3)|$ of the mapping g is

$$\begin{aligned}
|g'(w_3)| &= \omega^{\bar{d}_2}, \\
\bar{d}_2 &\equiv \sum_{k=1}^{H}(T_k - S_k) + \sum_{k=H+1}^{K}(T_k - 1) - 1.
\end{aligned}$$

Thus we can integrate the variable ω,

$$\begin{aligned}
J(z) &= \int_0^{\epsilon} \omega^{z+\bar{d}_2} \hat{J}(z) d\omega \\
&= \frac{\epsilon^{z+\bar{d}_2}}{z + \bar{d}_2 + 1} \hat{J}(z), \\
\hat{J}(z) &= \int H_3(\{\omega_{ki_k}\}, \{b_{i_1\cdots i_K,j}\})^z \prod d\omega_{ki_k} \prod db_{i_1\cdots i_K,j}.
\end{aligned}$$

If z is real and larger than the largest pole of $\hat{J}(z)$, the function $\hat{J}(z)$ is not equal to zero. Thus the largest pole of $J(z)$ is not smaller than $z = -(\bar{d}_2 + 1)$, which completes the proof of Lemma 13. (End of Proof)

(Proof of Theorem 4)

Combining Lemma 11-13, and the properties of the stochastic complexity (Proposition. 2, 3), we obtain Theorem 4. (End of Proof)

Chapter 7

Hidden Markov Models

In this chapter, we introduce hidden Markov models (Rabiner & Juang, 1986) and derive theorems of the stochastic complexity. Since a sequence of observations are generated by hidden states of the model, the model is robust to nonlinear time scaling, and used for learning time series. Thus, it is used in many areas, such as speech recognition, natural language processing, bioinformatics, system identification, etc. There are a lot of algorithms to calculate the probability of the observation given the model, to find the most likely state trajectory given the model and observations and to adjust the parameters of the model to maximize the probability of the observations, for example, the forward-backward algorithm, the Viterbi algorithm and the Baum-Welch algorithm, respectively. However, the properties of prediction are not still clarified. The results of this chapter provide a foundation to analyze them.

7.1 General Hidden Markov Models

Let $x = \{y_1, y_2, \cdots, y_T\} \in R^T$ be an observed time sequence. Suppose that the model has K hidden states. Let us define the transition probabilities by

$$A_K = \{a_{ij}\} \quad (1 \leq i, j \leq K),$$

where a_{ij} is the transition probability from the ith state to the jth state. Assume that

$$\begin{aligned}
&0 \leq a_{ij} \leq 1 \quad (i \neq j), \\
&\sum_{j \neq i}^{K} a_{ij} \leq 1, \\
&a_{ii} = 1 - \sum_{j \neq i}^{K} a_{ij}.
\end{aligned}$$

In an HMM, each state has an observation probability. We use the probability density function $f(y_t|b_i)$, where $b_i \in R^M$, for the observation probability. This function means the probability of y_t in the ith state. b_i is its parameter. Let $B_K(t)$ $(t \geq 2)$ the matrices such that

$$B_K(t) = \begin{pmatrix} a_{11}f(y_t|b_1) & a_{12}f(y_t|b_2) & \cdots & a_{1K}f(y_t|b_K) \\ a_{21}f(y_t|b_1) & a_{22}f(y_t|b_2) & \cdots & a_{2K}f(y_t|b_K) \\ \vdots & \vdots & \ddots & \vdots \\ a_{K1}f(y_t|b_1) & a_{K2}f(y_t|b_2) & \cdots & a_{KK}f(y_t|b_K) \end{pmatrix},$$

where $a_{ij}f(y_t|b_j)$ is the probability of y_t in the jth state after transiting from the ith state. Especially, we assume the initial state ($t = 1$) is the first one. From the initial state probability, let us define

$$\pi_K = (f(y_1|b_1), 0, \cdots, 0).$$

We use the notation,

$$v_K = (1, 1, \cdots, 1)^T,$$

where $(\cdot)^T$ means the transposed matrix. The hidden Markov model is defined by

$$p(x|w) = \pi_K \prod_{t=2}^{T} B_K(t) v_K, \qquad (7.1)$$

$$\prod_{t=2}^{T} B_K(t) \equiv B_K(2) B_K(3) \cdots B_K(T), \qquad (7.2)$$

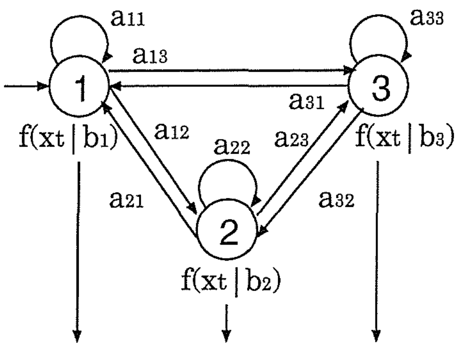

Figure 7.1: The hidden Markov model ($K = 3$).

where the parameter w is given by

$$\begin{aligned} w &= \{a, b\}, \\ a &= \{a_{ij}\} \quad (1 \le i, j \le K, i \neq j), \\ b &= \{b_k\} \quad (1 \le k \le K). \end{aligned}$$

The dimension of w is

$$K(K + M - 1).$$

We show the upper bounds of the stochastic complexities of hidden Markov models represented by (7.1) (Figure 7.1).

7.2 Stochastic Complexity of HMMs

We assume the following four conditions, (A7.1), (A7.2),(A7.3) and (A7.4).

(A7.1) First, assume that the length of observations is fixed and each sample observation x is independent and identical. Then, T is a constant.

(A7.2) Second, assume that the learning machine attains the true distribution. The true distribution has H hidden states, where $H \le K$. By using the true parameter w^*, it is written as

$$\begin{aligned} q(x) &= p(x|w^*), \\ p(x|w^*) &= \pi_H^* \prod_{t=2}^{T} B_H^*(t) v_H, \\ \pi_H^* &= \left(f(y_1|b_1^*), 0, \cdots, 0\right), \\ B_H^*(t) &= \begin{pmatrix} a_{11}^* f(y_t|b_1^*) & \cdots & a_{1H}^* f(y_t|b_H^*) \\ \vdots & \ddots & \vdots \\ a_{H1}^* f(y_t|b_1^*) & \cdots & a_{HH}^* f(y_t|b_H^*) \end{pmatrix}. \end{aligned} \tag{7.3}$$

The true parameter is defined by

$$\begin{aligned} w^* &= \{a^*, b^*\}, \\ a^* &= \{a_{ij}^*\} \quad (1 \le i, j \le H, i \ne j), \\ b^* &= \{b_k^*\} \quad (1 \le k \le H). \end{aligned}$$

Then,

$$a_{ii}^* = 1 - \sum_{j \ne i}^{H} a_{ij}^*.$$

(A7.3) Third, assume that the a priori probability distribution is positive on the true parameter. For a constant $\epsilon > 0$, let us define the subset of parameter $W(\epsilon) \subset W$ by

$$\begin{aligned} W(\epsilon) = \{\{a, b\} \in W \quad ; \quad & |a_{ij} - a_{ij}^*| \le \epsilon \ \ (1 \le i, j \le H, i \ne j), \\ & |a_{ij}| \le \epsilon \ \ (\text{otherwise}), \\ & |b_k - b_k^*| \le \epsilon \ \ (1 \le k \le H), \\ & |b_k - b_1^*| \le \epsilon \ \ (H + 1 \le k \le K)\}. \end{aligned}$$

Then, there is a constant $\epsilon > 0$ such that

$$\inf_{W(\epsilon)} \varphi(w) > 0,$$

where 'inf$_{W(\epsilon)}$' denotes the infimum value of $\varphi(w)$ in $w \in W(\epsilon)$.

(A7.4) At last, let us define the Kullback informations,

$$D_t(b_i^*||b_j) = \int f(y_t|b_i^*) \log \frac{f(y_t|b_i^*)}{f(y_t|b_j)} dy_t.$$

We assume that they are analytic and finite on the support of $\varphi(w)$.

Theorem 5 *Assume the conditions from (A7.1) to (A7.4). If the hidden Markov model (7.1) learns the true distribution (7.3), for arbitrary natural number n, the stochastic complexity satisfies the inequality*

$$\begin{aligned} F(n) &\leq C + \mu \log n, \\ \mu &= \frac{1}{2} H \{2K - H + M - 1\}, \end{aligned}$$

where C is a constant independent of n.

Let d be the dimension of the parameter w. It is well known that regular model's μ is equal to $d/2$. Therefore, the corresponding regular model has the coefficient of its stochastic complexity,

$$\bar{\mu} = \frac{1}{2} K \{K + M - 1\}$$

which is far larger than that of Theorem 5.

If the connections defined by the transition probability (7.1) are sparse, such as

$$\exists (i,j) \, ; a_{ij} = 0,$$

we can obtain the more tight bound. Let L_q be the number of non-zero a_{ij}, where $1 \leq i,j \leq H$ and L_r be the number of non-zero a_{ij}, where $1 \leq i \leq H$ and $H+1 \leq j \leq K$.

Theorem 6 *Assume the conditions from (A7.1) to (A7.4). If the sparse hidden Markov model (7.1) learn the sparse true distribution (7.3), for arbitrary natural number n, the stochastic complexity satisfies the inequality*

$$\begin{aligned} F(n) &\leq C + \mu' \log n, \\ \mu' &= \frac{HM + L_q}{2} + L_r, \end{aligned}$$

where C is a constant independent of n.

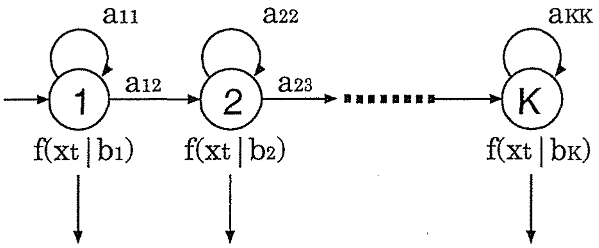

Figure 7.2: The left-to-right model.

(Remark) From the condition (A7.2), if $a_{ij} = 0$, where $1 \leq i, j \leq H$, it follows that $a^*_{ij} = 0$.

Let us apply Theorem 6 to a simple left-to-right model, which is used for speech recognition, bioinformatics and so on (Figure 7.2). This model has the parameters only $a_{i,i+1}$. Then, $L_q = H - 1$ and $L_r = 1$. We can easily derive the following corollary.

Corollary 4 *Assume that the learner is a K-state left-to-right model, and that the true distribution is H-state one. The upper bound of the stochastic complexity has the coefficient,*

$$\mu'' = \frac{HM + H + 1}{2}.$$

This result clarifies that the upper bound of the stochastic complexity does not increase when the learner includes the true. Note that the result is independent of the number of learner's states K.

7.3 Singularities in HMMs

Before the proof of the theorem, let us confirm that hidden Markov models are singular. We illustrate the shape of the true parameters in the parameter space. According to the equation (7.1), The simple hidden Markov model is written as

$$p(x|w) = f(y_1|b_1)\{a_{11}f(y_2|b_1) + a_{12}f(y_2|b_2)\},$$

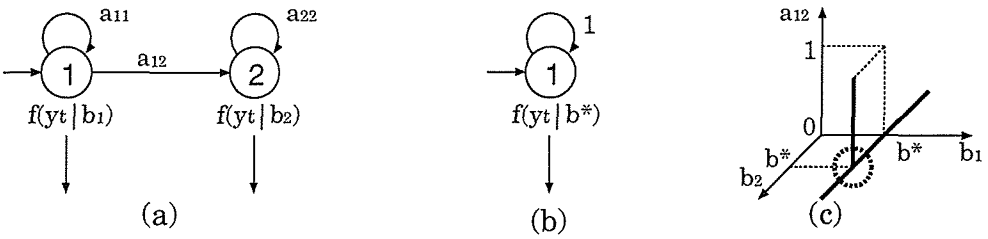

Figure 7.3: (a) The learning machine, (b) The true distribution, (c) The parameter space

where $a_{11} = 1 - a_{12}$, the dimension of b_i is one and the length of observations is two $T = 2$ for simplicity. This model has two hidden states (Figure 7.3 (a)). Assume that

$$q(x) = f(y_1|b^*)f(y_2|b^*).$$

This true distribution has one hidden state (Figure 7.3 (b)). Then, the set of the true parameters is

$$\{a_{12} = 0, b_1 = b^*\} \cup \{b_1 = b_2 = b^*\}.$$

This set has a singularity $(a_{12}, b_1, b_2) = (0, b^*, b^*)$. Therefore, hidden Markov models are singular even in this simple example.

7.4 Proof of Theorem 5, 6

We use the following notation,

$$\gamma_{ij} = \begin{cases} a_{ij} & (i \neq j) \\ 1 - \sum_{j' \neq i}^{H} a_{ij'} & (i = j) \end{cases}$$

γ_{ii} has the term $-a_{ij'}$, whose suffix $j'(\neq i)$ is from one to H, not to K. Let us introduce the following lemma,

Lemma 14 *In the region $W(\epsilon)$, there exists constants $c_{0i} > 0$ such that*

$$a_{ii}^* \log \frac{a_{ii}^*}{a_{ii}} \leq a_{ii}^* \log \frac{a_{ii}^*}{\gamma_{ii}} + c_{0i} \sum_{j=H+1}^{K} a_{ij}.$$

(Proof of Lemma 5)

Suppose that

$$\begin{aligned} y_1 &= \sum_{j=1,j\neq i}^{H} a_{ij}, \\ y_2 &= \sum_{j=H+1,j\neq i}^{K} a_{ij}, \\ y^* &= \sum_{j=1,j\neq i}^{H} a^*_{ij}. \end{aligned}$$

It follows that

$$\begin{aligned} (1-y^*)\log\frac{1-y^*}{1-y_1-y_2} &= (1-y^*)\log\frac{1-y^*}{1-y_1} + (1-y^*)\log\frac{1}{1-y_2/(1-y_1)} \\ &\leq (1-y^*)\log\frac{1-y^*}{1-y_1} + \frac{1-y^*}{1-\max y_1}y_2. \end{aligned}$$

In the region $W(\epsilon)$, there are constants $c_{0i} > 0$ such that

$$\frac{1-y^*}{1-\max y_1} \leq c_{0i},$$

which completes the proof of Lemma 14. (End of Proof)

The Kullback information (2.3) is rewritten as

$$H(w) = \int dx \pi^*_H \prod_{t=2}^{T} B^*_H(t) v_H \log\frac{\pi^*_H \prod_{t=2}^{T} B^*_H(t) v_H}{\pi_K \prod_{t=2}^{T} B_K(t) v_K},$$

where

$$\int dx \equiv \prod_{t=1}^{T} \int dy_t.$$

Let us divide the parameter w into $w = \{w_1, w_2\}$, where

$$\begin{aligned} w_1 &= \{a_{ij}; 1 \leq i,j \leq H, i \neq j, \\ &\quad\ b_k; 1 \leq k \leq H\}, \\ w_2 &= \{a_{ij}; \text{otherwise}, \\ &\quad\ b_k; H+1 \leq k \leq K\}, \end{aligned}$$

and define two functions,

$$\begin{aligned} H_1(w_1) &= \int dx \pi_H^* \prod_{t=2}^{T} B_H^*(t) v_H \log \frac{\pi_H^* \prod_{t=2}^T B_H^*(t) v_H}{\pi_K \prod_{t=2}^T B_H'(t) v_K}, \\ H_2(w_2) &= \sum_{i=1}^{H} \sum_{j=H+1}^{K} c_{ij} a_{ij} \left\{ C_{ij} + \sum_{l=1, l \neq j}^{K} a_{jl} + \sum_{k=H+1}^{K} \sum_{t=1}^{T} D_t(b_1^* || b_k) \right\}, \end{aligned}$$

where $\{c_{ij}\}$ and $\{C_{ij}\}$ are constants $(\forall c_{ij}, C_{ij} > 0)$ and

$$B_H'(t) = \begin{pmatrix} \gamma_{11} f(y_t|b_1) & \gamma_{12} f(y_t|b_2) & \cdots & \gamma_{1H} f(y_t|b_H) \\ \gamma_{21} f(y_t|b_1) & \gamma_{22} f(y_t|b_2) & \cdots & \gamma_{2H} f(y_t|b_H) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{H1} f(y_t|b_1) & \gamma_{H2} f(y_t|b_2) & \cdots & \gamma_{HH} f(y_t|b_H) \end{pmatrix}.$$

Let us prove the following lemma,

Lemma 15 *For arbitrary $w \in W(\epsilon)$,*

$$H(w) \leq H_1(w_1) + H_2(w_2).$$

(Proof of Lemma 15)

Let $B_K(t)$ divide into two matrices,

$$\begin{aligned} B_K(t) &= B_q(t) + B_r(t) \quad (t \geq 2), \\ B_q(t) &= \begin{pmatrix} \begin{matrix} a_{11} f(y_t|b_1) & \cdots & a_{1H} f(y_t|b_H) \\ \vdots & \ddots & \vdots \\ a_{H1} f(y_t|b_1) & \cdots & a_{HH} f(y_t|b_H) \end{matrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \\ B_r(t) &= \begin{pmatrix} & \mathbf{0} & & a_{1,H+1} f(y_t|b_{H+1}) & \cdots & a_{1K} f(y_t|b_K) \\ & & & \vdots & & \vdots \\ a_{H+1,1} f(y_t|b_1) & \cdots & & a_{H+1,H+1} f(y_t|b_{H+1}) & \cdots & a_{H+1,K} f(y_t|b_K) \\ \vdots & & & \vdots & \ddots & \vdots \\ a_{K1} f(y_t|b_1) & \cdots & & a_{K,H+1} f(y_t|b_{H+1}) & \cdots & a_{KK} f(y_t|b_K) \end{pmatrix}. \end{aligned}$$

By using these definitions, the learning model (7.1) is written as

$$\begin{aligned} p(x|w) &= \pi_K \prod_{t=2}^{T} (B_q(t) + B_r(t))\, v_K \\ &= \pi_K \prod_{t=2}^{T} B_q(t) v_K + p_r(x|w), \end{aligned}$$

where $p_r(x|w)$ is the sum of the rest terms.

In general, the following log-sum inequality holds: For arbitrary sequences of positive numbers $\{d_k, k = 1, 2, \cdots, I\}$ and $\{e_k, k = 1, 2, \cdots, I\}$,

$$\left\{\sum_{k=1}^{I} d_k\right\} \log \frac{\left\{\sum_{k=1}^{I} d_k\right\}}{\left\{\sum_{k=1}^{I} e_k\right\}} \leq \sum_{k=1}^{I} \left\{ d_k \log \frac{d_k}{e_k} \right\}.$$

Hence, for arbitrary sequence of positive numbers $\{h_k; k = 1, 2, \cdots, I'\}$, where $I < I'$,

$$\begin{aligned} \left\{\sum_{k=1}^{I} d_k\right\} \log \frac{\left\{\sum_{k=1}^{I} d_k\right\}}{\left\{\sum_{k=1}^{I'} h_k\right\}} &\leq \sum_{k=1}^{I} \left\{ d_k \log \frac{d_k}{e_k} \right\} - d_1 \log \frac{d_1}{h_1} \\ &\quad + d_1 \log \frac{d_1}{\sum_{k=1}^{I'} h_k - \sum_{k=1}^{I} h_k + h_1}. \end{aligned}$$

Applying this inequality to $H(w)$, we obtain

$$\begin{aligned} H(w) &\leq \sum_{i_1=1}^{H} \cdots \sum_{i_T=1}^{H} \int dx a^*_{1i_1} \cdots a^*_{i_{T-1}i_T} f(y_1|b^*_1) \prod_{t=2}^{T} f(y_t|b^*_{i_t}) \\ &\qquad \times \log \frac{a^*_{1i_1} \cdots a^*_{i_{T-1}i_T} f(y_1|b^*_1) \prod_{t=2}^{T} f(y_t|b^*_{i_t})}{a_{1i_1} \cdots a_{i_{T-1}i_T} f(y_1|b_1) \prod_{t=2}^{T} f(y_t|b_{i_t})} \\ &\quad - \int dx a_{11}^{*T-1} \prod_{t=1}^{T} f(y_t|b^*_1) \log \frac{a_{11}^{*T-1} \prod_{t=1}^{T} f(y_t|b^*_1)}{a_{11}^{T-1} \prod_{t=1}^{T} f(y_t|b_1)} \\ &\quad + \int dx a_{11}^{*T-1} \prod_{t=1}^{T} f(y_t|b^*_1) \log \frac{a_{11}^{*T-1} \prod_{t=1}^{T} f(y_t|b^*_1)}{Z_1(x)}, \end{aligned} \tag{7.4}$$

where

$$Z_1(x) = p_r(x|w) + a_{11}^{*T-1} \prod_{t=1}^{T} f(y_t|b_1).$$

We use the following notations of summations,

$$\begin{aligned}\sum_1 &\equiv \left\{\sum_{i_2=1}^{H}\cdots\sum_{i_T=1}^{H}-\sum_{i_2=1}^{1}\cdots\sum_{i_T=1}^{1}\right\},\\ \sum_2 &\equiv \left\{\sum_{i_2=1}^{K}\cdots\sum_{i_T=1}^{K}\right\}-\sum_1.\end{aligned}$$

We rewrite $Z_1(x)$ as

$$Z_1(x)=\sum_2 a_{1i_2}\cdots a_{i_{T-1}i_T}\prod_{t=1}^{T} f(y_t|b_{i_t}),$$

where

$$b_{i_1}\equiv b_1.$$

Let divide the right side of (7.4) into two functions,

$$\begin{aligned}R_1(w) &= \sum_1\int dx a^*_{1i_1}\cdots a^*_{i_{T-1}i_T}\prod_{t=1}^{T} f(y_t|b^*_{i_t})\log\frac{\prod_{t=1}^{T} f(y_t|b^*_{i_t})}{\prod_{t=1}^{T} f(y_t|b_{i_t})}\\ &\quad+\sum_1 a^*_{1i_1}\cdots a^*_{i_{T-1}i_T}\log\frac{a^*_{1i_1}\cdots a^*_{i_{T-1}i_T}}{a_{1i_1}\cdots a_{i_{T-1}i_T}},\\ R_2(w) &= \int dx a_{11}^{*T-1}\left\{\prod_{t=1}^{T} f(y_t|b^*_1)\right\}\log\frac{a_{11}^{*T-1}\prod_{t=1}^{T} f(y_t|b^*_1)}{Z_1(x)},\end{aligned}$$

where

$$b^*_{i_1}\equiv b^*_1.$$

Then, we can rewrite (7.4) as

$$H(w)\leq R_1(w)+R_2(w).$$

We use the following notations,

$$\begin{aligned}\rho_{i_2i_3\cdots i_T} &= \frac{a_{1i_2}a_{i_2i_3}\cdots a_{i_{T-1}i_T}}{\sigma},\\ \sigma &= \sum_2 a_{1i_2}\cdots a_{i_{T-1}i_T}.\end{aligned}$$

We can rewrite $R_2(w)$ as

$$R_2(w) = \int dx a_{11}^{*T-1} \left\{ \prod_{t=1}^{T} f(y_t|b_1^*) \right\} \left\{ \log \frac{a_{11}^{*T-1}}{\sigma} + \log \frac{\prod_{t=1}^{T} f(y_t|b_1^*)}{Z_2(x)} \right\},$$

where

$$Z_2(x) = \sum_{1} \rho_{i_2 i_3 \cdots i_T} \prod_{t=1}^{T} f(y_t|b_{i_t}).$$

Applying Jensen's inequality to $R_2(w)$, we obtain

$$\begin{aligned} R_2(w) &\leq a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{\sigma} \\ &\quad + \sum_{2} \rho_{i_2 \cdots i_T} \int dx \left\{ \prod_{t=1}^{T} f(y_t|b_1^*) \right\} \log \frac{\prod_{t=1}^{T} f(y_t|b_1^*)}{\prod_{t=1}^{T} f(y_t|b_{i_t})}. \end{aligned} \tag{7.5}$$

Because σ includes the term a_{11}^{T-1},

$$\sigma \geq a_{11}^{T-1} = \left\{ 1 - \sum_{j=2}^{K} a_{1j} \right\}^{T-1},$$

it follows that

$$a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{\sigma} \leq a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{a_{11}^{T-1}}.$$

Then, using Lemma 14, we obtain

$$a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{\sigma} \leq a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{\gamma_{11}^{T-1}} + c_{01} \sum_{j=H+1}^{K} a_{1j}, \tag{7.6}$$

where c_{01} is a constant. In the region $W(\epsilon)$, there is a constant ϵ_1 such that

$$\sigma \geq \{1 - \epsilon_1\}^{T-1}.$$

Thus, there is a constant c_1 such that

$$\begin{aligned} \rho_{i_2 i_3 \cdots i_T} &\leq \frac{a_{ij}}{\sigma} \quad (1 \leq \forall i, j \leq K) \\ &\leq \frac{a_{ij}}{c_1}. \end{aligned} \tag{7.7}$$

Divide the Kullback information into ones of each y_t,

$$\int dx \left\{ \prod_{t=1}^{T} f(y_t|b_1^*) \right\} \log \frac{\prod_{t=1}^{T} f(y_t|b_1^*)}{\prod_{t=1}^{T} f(y_t|b_{i_t})} = \sum_{t=1}^{T} D_t(b_1^*||b_{i_t}). \tag{7.8}$$

As we consider the region $W(\epsilon)$, there also exist constants $C_i > 0$ such that

$$D_t(b_1^*||b_i) \leq C_i \quad (1 \leq i \leq H). \tag{7.9}$$

In the inequality (7.5), each $\rho_{i_2 \cdots i_T}$ have

$$a_{ij} \quad (1 \leq i \leq H, H+1 \leq j \leq K)$$

as the factor, because the term is produced by $B_r(t)$. Hence, by using from (7.6) to (7.9), it follows that

$$\begin{aligned} R_2(w) \;\leq\;& a_{11}^{*T-1} \log \frac{a_{11}^{*T-1}}{\gamma_{11}^{T-1}} \\ & + \sum_{i=1}^{H} \sum_{j=H+1}^{K} c_{ij} a_{ij} \left\{ C'_{ij} + \sum_{l=1, l \neq j}^{K} a_{jl} + \sum_{k=H+1}^{K} \sum_{t=1}^{T} D_t(b_1^*||b_k) \right\} \\ & + \int dx \left\{ \prod_{t=1}^{T} f(y_t|b_1^*) \right\} \log \frac{\prod_{t=1}^{T} f(y_t|b_1^*)}{\prod_{t=1}^{T} f(y_t|b_1)}, \end{aligned}$$

where $\{C'_{ij}\}$ are constants.

Using Lemma 14, we can rewrite $R_1(w)$ as

$$\begin{aligned} R_1(w) \;\leq\;& \sum_{i=1}^{H} \sum_{j=H+1}^{K} c'_{0i} a_{ij} \\ & + \sum_{1} \int dx a_{1i_1}^* \cdots a_{i_{T-1}i_T}^* \prod_{t=1}^{T} f(y_t|b_{i_t}^*) \\ & \times \log \frac{\prod_{t=1}^{T} f(y_t|b_{i_t}^*)}{\prod_{t=1}^{T} f(y_t|b_{i_t})} \\ & + \sum_{1} a_{1i_1}^* \cdots a_{i_{T-1}i_T}^* \log \frac{a_{1i_1}^* \cdots a_{i_{T-1}i_T}^*}{\gamma_{1i_1} \cdots \gamma_{i_{T-1}i_T}}, \end{aligned}$$

where $\{c'_{0i}\}$ are constants.

Combining the inequalities of $R_1(w)$ and $R_2(w)$, we can obtain Lemma 15. (End of Proof)

Let us define two sets of the parameters

$$\begin{aligned} W_1 &= \{w_1; |a_{ij} - a^*_{ij}| \leq \epsilon \quad (1 \leq i,j \leq H, i \neq j), \\ & \qquad |b_k - b^*_k| \leq \epsilon \quad (1 \leq k \leq H)\}, \\ W_2 &= \{w_2; |a_{ij}| \leq \epsilon \quad (\text{otherwise}), \\ & \qquad |b_k - b^*_1| \leq \epsilon \quad (H+1 \leq k \leq K)\}. \end{aligned}$$

$w_1 \in W_1$ and $w_2 \in W_2$ are free variables from each other. Define the partial stochastic complexities,

$$F_i(n) = -\log \int_{W'_i} \exp(-nH_i(w_i))dw_i \quad (i = 1, 2),$$

where the integrated region W'_1 and W'_2 are taken such that $W'_1 \subset W_1$ and $W'_2 \subset W_2$, respectively, and that

$$W'_1 \times W'_2 \subset supp\varphi(w),$$

where $\varphi(w)$ is the support of the a priori distribution. From the assumption (A7.2),

$$\eta \equiv \inf_{w \in W_1 \times W_2} \varphi(w) > 0.$$

The stochastic complexity is bound by

$$F(n) \leq -\log \eta - \sum_{i=1}^{2} \log \int_{W_i} \exp(-nH_i(w_i))dw_i.$$

Thus,

$$F(n) \leq F_1(n) + F_2(n) + const.$$

In order to prove Theorem 5, it is sufficient to bind each $F_i(n)$ $(i = 1, 2)$. It is easy to bind $F_1(n)$, because it can be bound by the stochastic complexity of identifiable models.

Lemma 16 *A partial stochastic complexity satisfies the inequality,*

$$F_1(n) \leq \frac{H(H+M-1)}{2} \log n + C_1,$$

where C_1 is a constant independent of n.

(Proof of Lemma 16)

Because $H_1(w_1)$ corresponds to the Kullback information between the regular models, it follows that

$$H_1(w_1) \leq c\|w_1 - w^*\|^2$$

in an open set W_1, which contains w^*, where c is a constant. From (Proposition 2),

$$\begin{aligned} F_1(n) &\leq -\log \int_{W_1} \exp(-cn\|w_1 - w^*\|^2) dw_1 \\ &= \frac{H(H+M-1)}{2} \log n - \log \int_{W_{1n}} \exp(-c\|y\|^2) dy, \end{aligned}$$

where $W_{1n} = \{y; y/\sqrt{n} + w^* \in W_1\}$ converges to $R^{H(H+M-1)}$ as n tends to infinity. By using Lebesgue's convergence theorem, the second term of the right side converges to the constant. (End of Proof)

However, the set $\{w_2; H_2(w_2) = 0\}$ includes singularities, we apply the algebraic geometrical method to $F_2(n)$.

Lemma 17 *The second partial stochastic complexity satisfies the inequality,*

$$F_2(n) \leq H(K-H) \log n + C_2,$$

where C_2 is a constant independent of n.

(Proof of Lemma 17)

In order to clarify the asymptotic expansion of $F_2(n)$, we consider the zeta function,

$$J(z) = \int_{W_2'} H_2(w_2)^z dw_2.$$

Based on the algebraic geometrical method, we need to show that this zeta function has a pole,

$$z = -H(K-H)$$

Now, we define a variable w_3 and a mapping

$$g : w_3 = (\omega, \{\omega_{ij}\}, \{a_{ij}\}, \{b_k\}) \mapsto w_2 \tag{7.10}$$

by

$$\begin{aligned} \omega &= a_{1K}, \\ \omega\omega_{ij} &= a_{ij} \quad (2 \le i \le H, H+1 \le j \le K), \\ \omega\omega_{1j} &= a_{1j} \quad (H+1 \le j \le K-1). \end{aligned}$$

This mapping is called a blow-up in algebraic geometry. The function $H(g(w_3))$ divided ω is a constant function of ω,

$$H_3(\{\omega_{ij}\}, \{a_{ij}\}, \{b_k\}) \equiv H_2(w_2)/\omega.$$

The Jacobian $|g'(w_3)|$ of the mapping g is

$$|g'(w_3)| = \omega^{H(K-H)-1}.$$

Thus, we can integrate the variable ω,

$$\begin{aligned} J(z) &= \int_0^\epsilon \omega^{z+H(K-H)-1} \hat{J}(z) d\omega \\ &= \frac{\epsilon^{z+H(K-H)-1}}{z+H(K-H)} \hat{J}(z), \\ \hat{J}(z) &= \int H_3(\{\omega_{ij}\}, \{a_{ij}\}, \{b_k\})^z d\omega_{ij} da_{ij} db_k. \end{aligned}$$

If z is real and larger than the largest pole of $\hat{J}(z)$, the function $\hat{J}(z)$ is not equal to zero. Thus the largest pole of $J(z)$ is not smaller than $z = -H(K-H)$, which completes the proof of Lemma 17. (End of Proof)

(Proof of Theorem 5)

By combining Lemma 15-17, and the properties of the stochastic complexity (Proposition. 2, 3), we can obtain

$$\frac{H(H+M-1)}{2} + H(K-H) = \frac{H(2K-H+M-1)}{2},$$

which completes the proof of Theorem 5. (End of Proof)

(Proof of Theorem 6)

From the proof of Theorem 5, we can immediately obtain Theorem 6. According to the proof of Lemma 16, when the model is sparse, we can derive

$$F_1(n) \leq \frac{HM+L_q}{2} \log n + const.$$

Since the blow-up (7.10) depends on the number of non-zero a_{ij}, where $1 \leq i \leq H$ and $H+1 \leq j \leq K$, we directly obtain

$$F_2(n) \leq L_r \log n + const.$$

Combining these upper bounds, we accomplish the proof. (End of Proof)

Chapter 8

Discussion

8.1 Model Selection Problem

In this thesis, we assume several conditions in each model such as (A4.1), (A4.2), ..., (A7.4). They fall into three categories.

(C1) The learning machine $p(x|w)$ can attain the true distribution q(x).

(C2) An a priori distribution $\varphi(w)$ is positive on the true parameters.

(C3) Otherwise.

We can rewrite (C1) as that there exist the true parameters w^* such that $q(x) = p(x|w^*)$. It might seem particular from the practical point of view, since there can be some cases when the true distribution is not contained in any learning models. However, we are able to use a machine which almost attains the true distribution. Therefore, the case under the condition (C1) is important. Then, the second one (C2) is necessary to assure the existence of the true parameters.

Let us discuss the model selection. In the Bayesian estimation, the stochastic complexity is a criterion to select the optimal model for effective prediction. In spite of practical application, the mathematical analysis of the stochastic complexity in singular models were not constructed. Therefore, there exist some approximate methods such as the Laplace approximation, variational Bayes and Markov chain Monte Carlo (MCMC) to calculate it. Especially, the result of Laplace approximation is same as the BIC (Bayesian Information Criterion) formula. Our results clarify the theoretical values in singular models. All theorems claim that the coefficient of the stochastic complexity, λ is far smaller than $d/2$ (that of BIC), where d is the dimension of the parameter. This means that BIC does not reflect the effect of the stochastic complexity in singular models. Our results are mathematical foundations for the new criterion in the model selection. In addition, λ is also the coefficient of the Bayes generalization error. The smaller coefficient makes prediction more precise. Thus, our results certify the effectiveness of singular models compared with the regular statistical model which has the same dimension of the parameter. The condition (C2) is needed to attain this λ, since it is proven that λ is equal to $d/2$ when we use the Jeffreys' prior (Watanabe, 2001c). The Jeffreys' prior is equal to zero on the true parameters, $\varphi(w^*) = 0$. It does not satisfy (C2).

8.2 A Unified View of Singular Learning Machines

We obtained a unified view of some singular models. We are able to refer to mixture models and hidden Markov models as Bayesian networks representing by (6.1). It is easily to show that the equation of mixture models (4.1) and that of hidden Markov model (7.1) are the particular cases of (6.1). Figure 8.1 depicts them as Bayesian networks. We refer to the class of these models as *Bayesian network class*. Suppose that the hidden node h has K states. This means that the mixture model has K components and that the hidden Markov model has K hidden states, respectively. The observable node x_i or y_i is continuous. In our analysis of these three models, we used the Jensen's inequality, the log-sum inequality and almost the same blow-ups. Therefore,

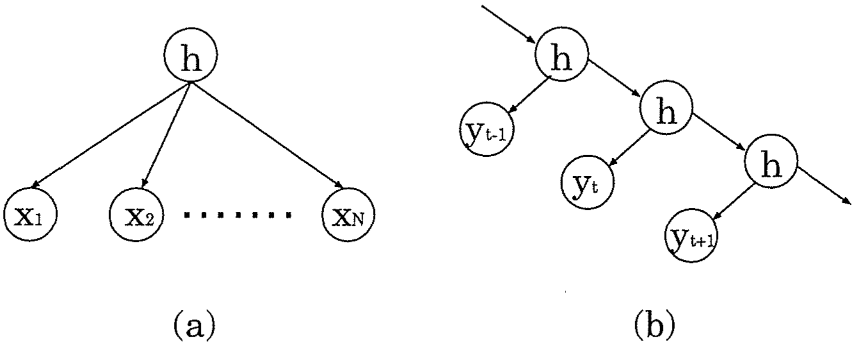

Figure 8.1: (a) The mixture model, (b) The hidden Markov model

we obtained the algorithm for analysis of the *Bayesian network class* (Figure 8.2).

Algorithm to calculate the stochastic complexity in Bayesian network class

Step 1. By using the Jensen's and log-sum inequalities, the Kullback information is divided into two parts, $H_1(w_1)$ and $H_2(w_2)$.

Then, $H_1(w_1)$ means the Kullback information from the true distribution to the learning machine which has the same sized parameter space as that of the true distribution. The original Kullback information $H(w)$ has the relationship, $H(w) \leq H_1(w_1) + H_2(w_2)$, where $w = \{w_1, w_2\}$. Let $J_i(z)$ be the zeta function of $H_i(w_i)$, and λ_i is its pole. From **Proposition.2, 3**, the summation of poles μ_1, μ_2 is an upper bound of the largest pole in the zeta function of $H(w)$.

Step 2. Calculate the pole μ_1.

It is easy to find it, since $H_1(w_1)$ is equivalent to that of statistical regular models. Let d_1 is the dimension of w_1. Then, $\mu_1 = d_1/2$.

Step 3. Calculate a pole μ_2.

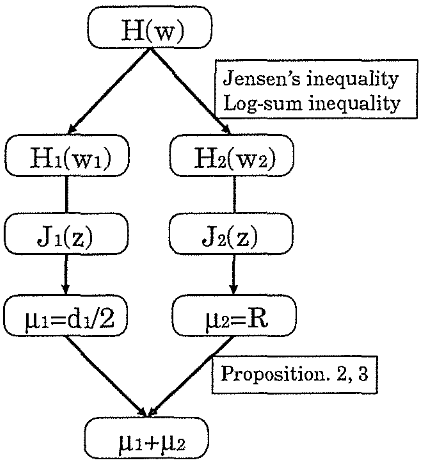

Figure 8.2: The algorithm for the Bayesian network class

The algebraic geometrical method is needed to find it, since $H_2(w_2)$ has singularities in the parameter space. In the *Bayesian network class*, $H_2(w_2)$ has the following upper bound.

$$H_2(w_2) \le \sum_{i=1}^{R} a_i \Psi(w'),$$

where $w' = w_2 \backslash \{a_i\}$ and $\Psi(w')$ is a positive analytic function. Thus, it is easy to find a blow-up.

$$\begin{aligned} \alpha &= a_1, \\ \alpha\alpha_i &= a_i \quad (i \neq 1). \end{aligned}$$

Therefore, μ_2 is equal to the number of the terms, R. Actually, the learning machine $p(x|w)$ is equal to the true distribution $q(x)$ in terms of functions when $\{a_i\}$ are all zeros. In other words, R is the minimum size of the parameters such that $p(x|w)$ attains $q(x)$ when the parameters are equal to zeros. We can describe R as the following.

$$\mu_2 = R = \min \dim \{\{a_i\} \subset w \;\; ; \;\; \{a_i\} = 0 \;\text{ s.t. }\; p(x|w) = q(x)\}.$$

Step 4. The summation $\mu_1 + \mu_2$ is the upper bound of λ.

This algorithm provides the relationship between the size of the learning machine and the stochastic complexity. If the other model is represented by Bayesian network (6.1), its stochastic complexity must be clarified by using this algorithm. Moreover, remark that we use the similar algorithm for Boltzmann machines which are not in the *Bayesian network class.* However, the inequalities in (**Step 1**) and the value of μ_2 are different. They depend on the models.

8.3 Future Works

At last, let us discuss the evaluation of the conventional methods. As we mentioned, we are able to evaluate the Laplace approximation and BIC, since we obtained the theoretical upper bounds of the stochastic complexity. As a result, the approximation is not appropriate in singular models. This means our results provide the mathematical foundation to evaluate the conventional methods for calculation of the stochastic complexity. It is our future study to compare our results with the MCMC and the variational Bayes. Then, we will be able to construct an algorithm to optimize them. Our ultimate goal is to propose the new algorithm for calculation of the stochastic complexity whose effectiveness is guaranteed mathematically. This thesis will give the first step of it.

Chapter 9

Conclusion

This thesis has established the following stuffs.

(1) A method to clarify the stochastic complexity in some concrete singular learning machines.

(2) A unified perspective to analyze the singularities in learning machines such as mixture models and hidden Markov models, in terms of Bayesian networks.

(3) A mathematical foundation to evaluate a criterion for the model selection, a method to design the optimal model and a learning algorithm.

Acknowledgment

Without patient encouragement and support of many people, this thesis could not have been realized.

This thesis is a collection of work conducted at the Watanabe laboratory, Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology from 2001 to 2004.

First and foremost, the author would like to express his sincere gratitude to his esteemed supervisor, Dr. Sumio Watanabe, Professor of Precision and Intelligence Laboratory, Tokyo Institute of Technology, for his constant supervision and fruitful discussions.

The author would also like to thank the members of my PhD committee who monitored my work and took effort in reading and providing me with valuable comments on earlier version of this thesis: Professor Katsumi Nitta, Professor Osamu Hasegawa, Professor Yasuharu Koike, and Professor Yoshiyuki Kabashima.

The author wishes to express his thanks to Dr. Nobuhiko Ogura for his valuable suggestion and technical advice. He would also like to express his gratitude to all the staffs of Watanabe group.

Finally, the author would like to express his deepest appreciation to his parents for their heartful encouragement and support, and would like to dedicate this thesis to them.

References

Ackley, D.H, Hinton, G.E, and Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines, *Cognitive Science*, 9, 147-169.

Akaike,H. (1980) Likelihood and Bayes procedure. *Bayesian Statistics*, (Bernald J.M. eds.) University Press, Valencia, Spain, 143-166, 1980.

Albizuri, F.X, d'Anjou, A., Grana, M. and Larranaga, P. (1997) Structure of the high-order Boltzmann machine from independence maps. *IEEE Trans. on Neural Networks*, 8 (6), 1351-1358.

Amari, S. and Ozeki, T. (2001) Differential and algebraic geometry of multilayer perceptrons. *IEICE Trans*, E84-A (1), 31-38.

Atiyah, M. F. (1970) Resolution of singularities and division of distributions. *Communications of Pure and Applied Mathematics*, 13, 145-150.

Cooper, G.F. (1990) The computational complexity of probabilistic inference using Bayesian networks. *Artificial Intelligence*, 42, 393-405.

Dacunha-Castelle,D. and Gassiat,E. (1997) Testing in locally conic models, and application to mixture models. *Probability and Statistics*, 1, 285-317, .

Hagiwara, K. (2002) On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, 14, 1979-2002.

Hartigan, J.A. (1985) A Failure of likelihood asymptotics for normal mixtures. *Proceedings of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, 2, 807-810.

Hironaka, H. (1964) Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, 79, 109-326.

Jacobs, R. and Jordan, M. (1991) Adaptive mixtures of local experts. *Neural Computation*, 3 (1), 79-87.

Lauritzen, S. and Spiegelhalter, D. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50, 157-224.

Levin, E., Tishby, N., and Solla,S.A. (1990) A statistical approaches to learning and generalization in layered neural networks. *Proc. of IEEE*, 78 (10) 1568-1674.

Mackay, D.J. (1992) Bayesian interpolation. *Neural Computation*, 4 (2), pp.415-447.

Opper, M., & Haussler, D. (1995) Bounds for predictive errors in the statistical mechanics of supervised learning. *Physical Review Letters*, 75 (20), 3772-3775.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, CA, 1988.

Rabiner, L. R. & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*,3 (1), 4-16.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statis-*

tics, 14, 1080-1100.

Rumellhart, D.E., and McClelland, J.L. (1986) Parallel distributed processing. The MIT Press, Cambridge, Massachusetts.

Rusakov, D, and Geiger, D. (2002) Asymptotic model selection for naive Bayesian networks, Proc. of Conference on Uncertainty in Artificial Intelligence.

Schwarz,G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461-464.

Watanabe,S. (1998) On the generalization error by a layered statistical model with Bayesian estimation. *IEICE Trans.*, Vol. J81-A, No. 10, 1442-1452, 1998.

Watanabe,S. (1999) Algebraic analysis for singular statistical estimation. *Lecture Notes on Computer Science*, 1720, 39-50, Springer.

Watanabe,S. (2001a) Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13 (4), 899-933.

Watanabe,S. (2001b) Algebraic geometrical methods for hierarchical learning machines. Neural Networks, 14 (8), 1049-1060.

Watanabe,S.(2002) Resolution of singularities and weak convergence of Bayesian stochastic complexity. Proc. of International Workshop on Singular Models and Geometric Methods in Statistics. Institute of Statistical Mathematics, pp.156-165.

Yamanishi, K. (1998) A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. on Information Theory*, 44

(4), 1424-1439.

Yamazaki,K, and Watanabe,S. (2002) A probabilistic algorithm to calculate the learning curves of hierarchical learning machines with singularities, *Trans. on IEICE*, J85-D-2(3), 363-372. The English version will appear in *Electronics and Communications in Japan.*

Yamazaki,K, and Watanabe,S. (2003a) Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks*, 16, 1029-1038.

Yamazaki,K, and Watanabe,S. (2003b) Singularities in complete bipartite graph-type Boltzmann machines and upper bounds of stochastic complexities, submitted.

Yamazaki,K, and Watanabe,S. (2003c) Stochastic complexity of Bayesian networks, in Proceedings of 19th Conference on Uncertainty in Artificial Intelligence, 59-599.

Yamazaki,K, and Watanabe,S. (2003d) Stochastic complexities of hidden Markov models, in Proceedings of 13th Conference on Nueral Networks for Signal Processing, 179-188.