Asymptotic Behavior of Stochastic Complexity of Complete Bipartite Graph-type Boltzmann Machines

Yu Nishiyama[1] and Sumio Watanabe[2]

[1] Department of Computational Intelligence and Systems Science,
Tokyo Institute of Technology,
4259, Nagatuta, Midori-ku, Yokohama, 226-8503 Japan
nishiyudesu@cs.pi.titech.ac.jp
[2] Precision and Intelligence Laboratory, Tokyo Institute of Technology,
4259, Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan
swatanab@pi.titech.ac.jp

Abstract. In singular statistical models, it was shown that Bayes learning is effective. However, on Bayes learning, calculation containing the Bayes posterior distribution requires huge computational costs. To overcome the problem, mean field approximation (or equally variational Bayes method) was proposed. Recently, the generalization error and stochastic complexity in mean field approximation have been theoretically studied. In this paper, we treat the complete bipartite graph-type Boltzmann machines and derive the upper bound of the asymptotic stochastic complexity in mean field approximation.

1 Introduction

Boltzmann machines are the learning models which were proposed by Hinton in 1984 and one of the machines that can learn the stochastic relationship between the input and output. From the mathematical viewpoint, Boltzmann machines belong to the family of Bayesian networks or Graphical models, which are well known and used in information science. In addition, Boltzmann machines can be regarded as the spin system that has the random interactions studied in statistical physics. Therefore, we can say that Boltzmann machines are the important models both from the aspect of the application to information science and the relationship between information science and statistical physics. However, as for the Boltzmann machines, the mapping from the parameters to the models is not one-to-one. Then, it follows that there are a lot of parameters on which determinants of Fisher information matrices become 0 and that neither the maximum likelihood estimator nor posterior converge to normal distribution. In consequence, we can not apply the conventional asymptotic theory of regular statistical models to Boltzmann machines. It means that the basic and important methods such as the model selection or the statistical test have not been established yet.

In general, when a learning machine has the degenerate Fisher information matrix, the model is called a singular statistical model. It is known that if a learning machine has hidden variables or hierarchical structures, the model tends to be singular statistical model. In some singular statistical models, the asymptotic behavior of Bayes generalization error was mathematically derived by use of algebraic geometrical methods[1][2][3]. The results indicate that singular statistical models have better generalization performance than those of regular and show the effectiveness of Bayes learning. However, on the Bayes learning, the calculation which contains Bayes posterior distribution requires huge computational costs since Bayes posterior distribution contains the high-dimensional integral. To avoid the huge computational costs, mean field approximation (or variational Bayes method), which is originally known in statistical physics, was proposed. In the mean field approximation, we approximate Bayes posterior distribution by the trial distribution whose parameters are all independent of each other. The trial distribution is optimized to be the nearest from the Bayes posterior distribution in the meaning of Kullback distance. The algorithms based on the mean field approximation have shown the effectiveness to practical information systems by virtue of the efficiency of the computational costs.

Recently, the asymptotic behavior of stochastic complexity and generalization error in the mean field approximation have been theoretically studied. The theoretical results can teach us the accuracy of the mean field approximation of Bayes learning and the difference from the regular statistical models. Furthermore, the results lead to a criterion to judge whether iteration values in the algorithms converge to local minima or global minimum and also lead to a foundation of the model selection proposed in singular statistical models. The asymptotic behavior of stochastic complexity has been studied and clarified in the learning models of reduced rank regressions[4], normal mixture models[5], hidden Markov models[6], stochastic context-free grammer[7] and neural networks[8].

In this paper, we treat complete bipartite graph-type Boltzmann machines and derive mathematically the upper bound of the asymptotic stochastic complexity. The upper bound can be obtained by restricting the trial distribution especially to normal distribution family[8]. Now, the result in this paper is extended to general Boltzmann machines and the upper bound is also obtained[9].

2 Learning Theory

2.1 Bayes Learning

We denote the framework of Bayes learning here. Generally we express datum of something as an M dimensional vector $x \in R^M$. We assume that n data $X^n = (X_1, X_2, ..., X_n)$ are independently and identically taken from the true distribution $q(x)dx$. Then, we prepare a conditional probability density function $p(x|\theta)$ named learning model and an a priori distribution $\varphi(\theta)$, where $\theta \in R^d$ is

a model parameter. We compose the Bayes posterior distribution as

$$p(\theta|X^n) = \frac{1}{Z(X^n)}\varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta), \tag{1}$$

where $Z(X^n)$ is the normalization constant. The Bayes predictive distribution of x is defined as

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)d\theta,$$

where this equation means the average of the learning model over the Bayes posterior distribution. We expect that the Bayes predictive distribution approaches the true distribution $q(x)$ after learning n data. To know the difference between the predictive distribution and the true distribution, the Bayes generalization error is defined by

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

where $E_{X^n}\{\}$ means the average over all sets of n training sample.

2.2 Learning Theory of Singular Model

Bayes posterior distribution defined by eq.(1) is rewritten as

$$p(\theta|X^n) = \frac{e^{-n\tilde{H}_n(\theta)}}{\bar{Z}(X^n)},$$

where $\bar{Z}(X^n)$ is the different normalization constant from $Z(X^n)$ and $\tilde{H}_n(\theta)$ is

$$\tilde{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(X_i)}{p(X_i|\theta)} - \frac{1}{n}\log\varphi(\theta) = H_n(\theta) - \frac{1}{n}\log\varphi(\theta). \tag{2}$$

Here $H_n(\theta)$ is the empirical Kullback information. Then, we define $F(n)$ as

$$F(n) = E_{X^n}\{-\log\bar{Z}(X^n)\} = E_{X^n}\{-\log\int e^{-n\tilde{H}_n(\theta)}d\theta\}.$$

It is known that $F(n+1) - F(n) = G(n)$ holds [**?**]. This equation tells us that it is equivalent to calculate $G(n)$ and $F(n)$. This important value $F(n)$ is called stochastic complexity. It is known that $F(n)$ can be asymptotically expanded as

$$F(n) \sim \lambda\log n - (m-1)\log\log n + O(1) \quad (n\to\infty), \tag{3}$$

where $-\lambda$ and m are respectively the nearest pole to origin and the order of $-\lambda$ of zeta function $J(z)$[1]. $J(z)$ is

$$J(z) = \int H(\theta)^z\varphi(\theta)d\theta, \tag{4}$$

where z is a complex number and $H(\theta)$ is the Kullback information as follows.

$$H(\theta) = \int q(x)\log\frac{q(x)}{p(x|\theta)}dx. \tag{5}$$

2.3 Mean Field Approximation

Let us introduce the framework of mean field approximation. For arbitrary probability density function $f(\theta)$, by using Jensen's inequality, the stochastic complexity $F(n)$ has the following upper bound.

$$\begin{aligned} F(n) &= E_{X^n}\{-\log \int f(\theta)\frac{e^{-nH_n(\theta)}\varphi(\theta)}{f(\theta)}d\theta\} \\ &\leq E_{X^n}\{-\int f(\theta)\log\frac{e^{-nH_n(\theta)}\varphi(\theta)}{f(\theta)}d\theta\} \\ &= E_{X^n}\{\int f(\theta)\log f(\theta)d\theta + n\int f(\theta)\tilde{H}_n(\theta)d\theta\}. \end{aligned} \tag{6}$$

If we assume the parameters in $f(\theta)$ are independent of each other, that is, $\bar{f}(\theta) = \prod_{j=1}^{d} f_j(\theta_j)$, then $\bar{f}(\theta)$ which minimizes the last equation in eq.(6) is called the mean field approximation on Bayes posterior distribution. The minimum value of the last equation in eq.(6)

$$\bar{F}(n) = E_{X^n}[\min_{\bar{f}(\theta)}\{\int \bar{f}(\theta)\log\bar{f}(\theta)d\theta + n\int \bar{f}(\theta)\tilde{H}_n(\theta)d\theta\}] \tag{7}$$

is called the stochastic complexity in mean field approximation.

Moreover, the stochastic complexity in mean field approximation $\bar{F}(n)$ has the following upper bound.

$$\begin{aligned} \bar{F}(n) &\leq \min_{\tilde{f}(\theta)}\{\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta + n\int \tilde{f}(\theta)E_{X^n}[\tilde{H}_n(\theta)]d\theta\} \\ &= \min_{\tilde{f}(\theta)}\{\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta + n\int \tilde{f}(\theta)\tilde{H}(\theta)d\theta\}, \end{aligned} \tag{8}$$

where $\tilde{H}(\theta)$ is given by

$$\tilde{H}(\theta) = H(\theta) - \frac{1}{n}\log\varphi(\theta). \tag{9}$$

3 Learning Model

3.1 Learning Model

The learning model which we consider in this paper is the complete bipartite graph-type Boltzmann machines that are shown in Fig.1. $\{x_i\}_{i=1}^{M}$ represent both the input and output units. $\{y_i\}_{i=1}^{K}$ represent the hidden units. There are the connections only between $\{x_i\}_{i=1}^{M}$ and $\{y_i\}_{i=1}^{K}$. Each value of $\{x_i\}_{i=1}^{M}$, $\{y_i\}_{i=1}^{K}$ takes 1 or -1. Then, the learning model of the complete bipartite graph-type

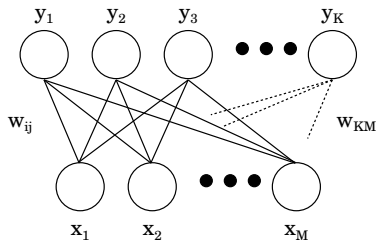

Fig. 1. Complete Bipartite Graph-type Boltzmann Machines.

Boltzmann machines is mathematically shown as follows under the model parameter θ in **2** is w.

$$p(x|w) = \frac{\sum_y \exp(\sum_{i=1}^K \sum_{j=1}^M w_{ij} x_j y_i)}{\sum_x \sum_y \exp(\sum_{i=1}^K \sum_{j=1}^M w_{ij} x_j y_i)}$$
$$= \frac{\prod_{i=1}^K \sum_{y_i} \exp(\sum_{j=1}^M w_{ij} x_j y_i)}{\sum_x \prod_{i=1}^K \sum_{y_i} \exp(\sum_{j=1}^M w_{ij} x_j y_i)} = \frac{\prod_{i=1}^K \cosh(\sum_{j=1}^M w_{ij} x_j)}{Z(w)}. \quad (10)$$

Here $\sum_y$ means $\sum_{y_1} \sum_{y_2} \sum_{y_3} \cdots \sum_{y_K}$ and each sum takes $\{+1, -1\}$. It is also applied to $\sum_x$. The distribution is marginalized by hidden units $\{y_i\}_{i=1}^K$ since the hidden units are unobservable. $Z(w)$ is the normalization constant.

It is noted that the learning model in eq.(10) is a discrete distribution because each $\{x_i\}_{i=1}^M$ takes $\{-1, +1\}$. Therefore, the number of combinations of input and output values is 2^M in total. It implies that the number of model parameters $2^M - 1$ is enough to describe the learning model $p(x|w)$. Hence, we consider the case when the number of hidden units K satisfies $KM \leq 2^M - 1$. In addition, when $M = 1$, the learning model $p(x|w)$ becomes

$$q(x) = p(x|w) = \frac{\prod_{i=1}^H \cosh(w_{i1})}{Z(w)} = \frac{1}{2^M}, \quad (11)$$

where we used $\cosh x$ is even function generally. From eq.(11), $p(x|w)$ doesn't depend on the value of model parameters. Hence, we assume $M \geq 2$.

Hereafter, we consider the problem under these two conditions.

3.2 True Distribution

We assume that the true distribution $q(x)$ in **2.1** is included in the learning model of eq.(10) and the true number of hidden units is expressed as K^* ($\leq K$). Then, there is at least a true model parameter w^* such that w^* satisfies

$$w_{ij}^* = 0 \ \text{ for } \ i \in \{K^* + 1, \cdots, K\}.$$

Under the true parameter w^*, the true distribution $q(x)$ becomes

$$p(x|w^*) = \frac{\prod_{i=1}^{K^*} \cosh(\sum_{j=1}^{M} w_{ij}^* x_j)}{Z(w^*)}. \tag{12}$$

4 Main Theorem

4.1 Methods

We assume that the a priori distribution $\varphi(w)$ in eq.(9) is the normal distribution defined by

$$\varphi(w) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^{KM} \exp\{\frac{-\sum_{i=1}^{K}\sum_{j=1}^{M}(w_{ij} - \hat{w}_{ij})^2}{2\sigma^2}\}. \tag{13}$$

We also assume that the trial distribution $\tilde{f}(w)$ in eq.(8) is restricted to the normal distribution family as follows.

$$\tilde{f}(w) = \frac{\exp\{-\sum_{i=1}^{K}\sum_{j=1}^{M} N_{ij}(w_{ij} - \hat{w}_{ij})^2\}}{Z(N)}. \tag{14}$$

We can obtain the upper bound of eq.(8) by restricting $\tilde{f}(w)$ especially to the normal distribution family and optimizing $\{N_{ij}\}$ and $\{\hat{w}_{ij}\}$. It is noted that the normal distribution is one of the simplest distributions in terms of the application to practical systems.

4.2 Main Theorem

[Theorem] The stochastic complexity of complete bipartite graph-type Boltzmann machines in mean field approximation has the following upper bound.

$$\bar{F}(n) \leq \frac{KM + K^*M}{4} \log n + C. \tag{15}$$

Here C is the constant. M, K and K^* are respectively the number of input and output units, hidden units of the learning model and hidden units of true distribution.

4.3 Proof of Main Theorem

We use the following lemma. According to the methods given in **4.1**, we do not have to limit the learning model to the Boltzmann machines and can evaluate generally the upper bound of $\bar{F}(n)$. The following lemma is about this. We represent parameter θ instead of w.

[Lemma] For Kullback information $H(\theta)$ and $\theta \in R^d$, if there exists a value of parameter $\hat{\theta}$ such that the number of elements of the set

$$\{\ i\ \ ;\ \ H(\hat{\theta}) = 0\ \text{ and }\ \frac{\partial^2 H(\theta)}{\partial \theta_i^2}\Big|_{\hat{\theta}} \neq 0\}$$

is less than or equal to r, the general stochastic complexity in mean field approximation has the following upper bound.

$$\bar{F}(n) \leq \frac{d+r}{4}\log n + O(1).$$

[Proof of Lemma]
We restrict $\tilde{f}(\theta)$ in eq.(8) to the normal distribuiton

$$\tilde{f}(\theta) = \frac{\exp\{-\sum_{i=1}^d N_i(\theta_i - \hat{\theta}_i)^2\}}{Z(N)}. \tag{16}$$

Then, the first term of the right-hand side in eq.(8) becomes, from the calculation of gaussian entropy,

$$\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta = \frac{1}{2}\sum_{i=1}^d \log N_i - \frac{d}{2}\log\pi - \frac{d}{2}. \tag{17}$$

By substituting eq.(17) to the right-hand side in eq.(8), we have

$$\bar{F}(n) \leq \frac{1}{2}\sum_{i=1}^d \log N_i + n\int \tilde{f}(\theta)\tilde{H}(\theta)d\theta + C_1. \tag{18}$$

Besides, by substituting the a priori distribution

$$\varphi(\theta) = \frac{1}{Z(\sigma)}\exp\{-\frac{\sum_{i=1}^d(\theta_i - \hat{\theta}_i)^2}{2\sigma^2}\} \tag{19}$$

to eq.(18), we have

$$\bar{F}(n) \leq \frac{1}{2}\sum_{i=1}^d \log N_i + \frac{1}{2\sigma^2}\sum_{i=1}^d \frac{1}{2N_i} + n\int \tilde{f}(\theta)H(\theta)d\theta + C_2. \tag{20}$$

Hereafter, we optimize parameters $\{N_i\}$ to minimize the right-hand side in eq.(20) in the meaning of the coefficient of the asymptotic form($n \to \infty$). If we expand $H(\theta)$ in eq.(20) around $\hat{\theta}$, we have

$$n\int f_0(\theta)\{\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k H(\theta)}{\partial\theta^k}\Big|_{\hat{\theta}}(\theta-\hat{\theta})^k\}d\theta$$

$$= nH(\hat{\theta}) + \frac{n}{2!}\sum_{i=1}^{d}\frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}}\frac{1}{2N_i} + \frac{n}{4!}\sum_{i=1}^{d}\frac{\partial^4 H(\theta)}{\partial\theta_i^4}\Big|_{\hat{\theta}}\frac{1}{2N_i}\frac{3}{2N_i}$$
$$+\frac{n}{4!}\underbrace{\sum_{j=1}^{d}\sum_{i=1}^{d}}_{j\neq i}\frac{\partial^4 H(\theta)}{\partial\theta_i^2\partial\theta_j^2}\Big|_{\hat{\theta}}(\frac{1}{2N_i})(\frac{1}{2N_j}) + \text{higher order}. \quad (21)$$

We optimize $\{N_i\}$ as

$$N_i = n \ \text{ for } \{\ i\ ;\ \frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}} \neq 0\}, \qquad N_i = n^{\frac{1}{2}} \ \text{ for } \{\ i\ ;\ \frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}} = 0\}. \quad (22)$$

Then, from $nH(\hat{\theta}) = 0$, eq.(21) becomes constant order. Hence, eq.(20) becomes

$$\bar{F}(n) \leq \frac{1}{2}\sum_{j=1}^{r}\log n + \frac{1}{4}\sum_{j=r+1}^{d}\log n + O(1) = \frac{d+r}{4}\log n + O(1).$$

The lemma's equation was derived.

This lemma implies that the upper bound of stochastic complexity is given only if we calculate the second order differential or equally Fisher information of learning models. we obtain the result in **4.2** by applying this lemma to complete bipartite graph-type Boltzmann machines.

[Proof of Theorem]
The Kullback information of complete bipartite graph-type Boltzmann machines is given by substituting eqs.(10) and (12) to eq.(5) under changing $\int dx$ to Σ_x. The first and second order differentials are calculated as

$$\frac{\partial H(w)}{\partial w_{\alpha\beta}}\Big|_{\hat{w}} = \left\langle\ t_{\alpha\beta}\ \right\rangle_{\hat{w}} - \left\langle\ t_{\alpha\beta}\ \right\rangle_{w^*}, \qquad \frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{\hat{w}} = \left\langle\ t_{\alpha\beta}^2\ \right\rangle_{w^*} - \left\langle\ t_{\alpha\beta}\ \right\rangle_{\hat{w}}^2, \quad (23)$$

where $t_{\alpha\beta}$ and the average $\left\langle\ f(x,w_1)\ \right\rangle_{w_2}$ are respectively

$$t_{\alpha\beta} = \tanh(\sum_{j=1}^{M} w_{\alpha j}x_j)x_\beta, \qquad \left\langle\ f(x,w_1)\ \right\rangle_{w_2} = \sum_x f(x,w_1)\ p(x|w_2).$$

If we choose the parameter $\hat{w}$ which satisfies $H(\hat{w}) = 0$, from $p(x|w^*) = p(x|\hat{w})$ and $\left\langle\ \right\rangle_{w^*} = \left\langle\ \right\rangle_{\hat{w}}$, the second order differential becomes the variance as

$$\frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{\hat{w}} = \left\langle\ (t_{\alpha\beta} - \left\langle\ t_{\alpha\beta}\ \right\rangle_{\hat{w}})^2\ \right\rangle_{\hat{w}}. \quad (24)$$

Hereafter, we choose the parameter $\hat{w}$ as w^* especially. When $\alpha \in \{K^* + 1, \cdots, K\}$, since $w^*_{\alpha\beta} = 0$ and $t_{\alpha\beta} = 0$, eq.(24) becomes

$$\frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{w^*} = 0 \quad \text{ for } \alpha \in \{K^* + 1, \cdots, K\}. \quad (25)$$

Therefore, we can set $r = K^*M$ and $d = KM$ in the lemma and the theorem was proved.

5 Discussion

5.1 About Lower Bound

In section 4, we gave the upper bound of stochastic complexity of the Boltzmann machines in mean field approximation. Although we do not obtain the rigorous lower bound yet, we evaluate the lower bound especially when we restrict the trial distribution $\tilde{f}(w)$ to the normal distribution whose mean value is w^*. In contrast with eq.(25), if $w^*_{\alpha\beta} \neq 0$ $\alpha \in \{1, 2, \cdots, K^*\}$, the second order differential becomes

$$\frac{\partial^2 H(w)}{\partial w^2_{\alpha\beta}}\Big|_{w^*} \neq 0 \quad \text{for } \alpha \in \{1, 2, \cdots, K^*\}. \tag{26}$$

Therefore, the bound in section 4.2 is the minimum in the meaning of optimizing the second order differential.

Subsequently, in section 4, we optimized the parameters $\{N_i\}$ as eq.(22). However, if there exist the parameters $\{N_i\}$ which satisfy

$$\{\, i\,; \;\; \frac{\partial^2 H(\theta)}{\partial \theta_i^2}\Big|_{\hat{\theta}} = \frac{\partial^4 H(\theta)}{\partial \theta_i^2 \partial \theta_j^2}\Big|_{\hat{\theta}} = 0\} \quad \text{for } {}^{\forall} j \in \{1, \cdots, d\}, \tag{27}$$

we can optimize parameters as e.g. $N_i = n^{\frac{1}{3}}$ instead of $N_i = n^{\frac{1}{2}}$ and obtain smaller bound. Therefore, we consider whether there exist the parameters $N_{\alpha\beta}$ which satisfy eq.(27) or not in the case of the Boltzmann machines. To satisfy eq.(27), we can limit to $\alpha \in \{K^*+1, \cdots, K\}$ because of eqs.(25) and (26). Then, the fourth order differential of the Boltzmann machines is calculated as

$$\frac{\partial^4 H(w)}{\partial w^2_{\alpha\delta} \partial w^2_{\alpha\beta}}\Big|_{w^*} = 2(1 - \left\langle\, x_\beta x_\delta \,\right\rangle^2_{w^*}) \quad \text{for} \quad \delta \neq \beta. \tag{28}$$

When $\delta \neq \beta$, $\left\langle\, x_\beta x_\delta \,\right\rangle_{w^*}$ satisfies

$$-\left\langle\, x_\beta x_\beta \,\right\rangle_{w^*} < \left\langle\, x_\beta x_\delta \,\right\rangle_{w^*} < \left\langle\, x_\beta x_\beta \,\right\rangle_{w^*} \Longleftrightarrow -1 < \left\langle\, x_\beta x_\delta \,\right\rangle_{w^*} < 1. \tag{29}$$

Therefore, eq.(28) does not become 0 and there exists no parameter which satisfies eq.(27).

5.2 Comparison with Regular Statistical Model

If we compare the upper bound in **4.2** with the regular statistical model expressed as

$$F(n) = \frac{KM}{2} \log n + O(1),$$

where KM is the number of parameters which describe the Boltzmann machines, we know that the coefficient of $\log n$ term in **4.2** is smaller than that of the regular statistical model because of $K^* \leq K$. Furthermore, in other study

of the Boltzmann machines, the upper bound of stochastic complexity on Bayes learning instead of the mean field approximation was derived by using algebraic geometrical method[10]. The coefficients of both upper bounds accord with each other. It indicates that the upper bound of Bayes stochastic complexity is attainable by the mean field approximation even when the trial distribution is restricted to the simple normal distribution.

6 Conclusion

We derived the upper bound of stochastic complexity of complete bipartite graph-type Boltzmann machines in mean field approximation.

7 Acknowledgement

This research was supported by the Ministry of Education, Science, Sports, and Culture in Japan, Grant-in-aid for scientific research 15500130.

References

1. S. Watanabe, "Algebraic analysis for nonidentifiable learning machines", Neural Computation, vol.13, no.4, pp.899-933, 2001.
2. Miki Aoyagi and Sumio Watanabe, "The generalization error of reduced rank regression in bayesian estimation", Proc. of ISITA2004, pp.1068-1073, Italy, 2004.
3. K. Yamazaki and S. Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity", International Journal of Neural Networks, 16(7), pp.1029-1038, 2003.
4. S. Nakajima and S. Watanabe, "Generalization Error and Free Energy of Linear Neural Networks in Variational Bayes Approach", Proc. of ICONIP2005, pp.55-60, Taiwan, 2005.
5. K. Watanabe and S. Watanabe, "Lower bounds of stochastic complexities in variational Bayes learning of gaussian mixture models", Proc. IEEE conference on Cybernetics and Intelligent Systems, pp.99-104, 2004.
6. Tikara Hosino, Kazuho Watanabe and Sumio Watanabe, "Stochastic Complexity of Variational Bayesian Hidden Markov Models", Proc. of IJCNN2005, Canada, 2005.
7. Tikara Hosino, Kazuho Watanabe and Sumio Watanabe, "Stochastic Complexity of Stochastic Context Free Grammer on Variational Bayesian method", IEICE Technicalreport, NC2005-49, October, 2005.
8. Nobuhiro Nakano and Sumio Watanabe, "Stochastic Complexity of Layered Neural Networks in Mean Field Approximation", Proc. of ICONIP2005, pp.332-337, Taiwan, 2005.
9. Yu Nishiyama and Sumio Watanabe, "Asymptotic Behavior of Free Energy of General Boltzmann Machines in Mean Field Approximation", IEICE Technicalreport, to appear, July, 2006.
10. K. Yamazaki and S. Watanabe, "Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities", IEEE Trans. Neural Networks, vol.16(2), pp.312-324, 2005.