社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

特異モデルのベイズ学習における交換モンテカルロ法について

永田　賢二†　　渡辺　澄夫††

† 東京工業大学 総合理工学研究科 知能システム科学専攻 〒 226-8503 横浜市緑区長津田 4259
†† 東京工業大学 精密工学研究所 〒 226-8503 横浜市緑区長津田 4259
E-mail: †kenji.nagata@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

あらまし　神経回路網、混合正規分布、ベイズネットワーク、隠れマルコフモデル、ボルツマンマシンなどの特異モデルと呼ばれる学習モデルが情報科学において広く用いられている。これらの学習モデルでは、ベイズ学習が汎化能力に優れていることが解明されているが、実際にベイズ事後分布を従来のモンテカルロ法で実現すると計算量が爆発するという問題を有している。本論文では、特異モデルのベイズ学習において交換モンテカルロ法が有効であることを提案し、従来のモンテカルロ法よりも精度良くベイズ事後分布を実現することを実験的に明らかにする。
キーワード　交換モンテカルロ法、特異モデル、ベイズ事後分布

Exchange Monte Carlo Method for Bayesian Lerning of Singular Learning Machines

Kenji NAGATA† and Sumio WATANABE††

† Department of Computer Science Tokyo Institute of Technology, 4259
Nagatsuda,Midori-ku,Yokohama,226-8503 Japan
†† PI Lab., Tokyo Institute of Technology.
E-mail: †kenji.nagata@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

Abstract　A lot of singular learning machines such as neural networks, normal mixtures, Bayesian networks and hidden Markov models are widely used in practical information systems. In these learning machines, it was clarified that the Bayesian learning provides the better generalization performance than the maximum likelihood method. However, it needs huge computational costs to realize the Bayesian posterior distribution by the conventional Monte Carlo method. In this paper, we propose that the exchange Monte Carlo method is appropriate for the Bayesian learning of singular learning machines, and experimentally show that it attains the better posterior distribution than the conventional Monte Carlo method.
Key words　Exchange Monte Carlo Method, Singular Learning Machines , Bayesian Posterior Distribution

1. Introduction

A lot of learning machines with hierarchical structures such as neural networks, normal mixtures, Bayesian networks, hidden Markov models and Boltzmann machines are called singular learning machines because they have singularities in their parameter spaces. They are widely used for pattern recognition, data mining and gene analysis. Recently, an algebraic geometrical method for singular learning machines has recently been established [1] [2]. By this method, in the Bayesian learning, the generalization errors of singular learning machines were proved to be less than those of regular statistical models [3] [4].

In the Bayesian learning, it is necessary to realize the Bayesian posterior distribution accurately near the singularities. A Markov Chain Monte Carlo (MCMC) method is often used to generate a sequence of Markov chain that converges to the target distribution. Recently, it has been shown that the Metropolis algorithm, one of the MCMC methods, needs huge computational costs to approximate the Bayesian posterior distributions of singular learning machines [5]. This is because the Bayesian posteriors of the singular learning machines are widely and complexly distributed in the parameter space.

On the other hand, an improved MCMC method has recently been developed, which is surveyed in [6]. This method gives us a general strategy to overcome the problem of huge computational costs. An exchange Monte Carlo (MC) method is well known as one of the extended ensemble method [7], where effectiveness has been shown in many fields.

In this paper, we propose that the exchange MC method is appropriate to approximate the Bayesian posterior distribution of singular learning machines, and experimentally show that the numerical stochastic complexity is better approximated by the exchange MC method than by the Metropolis method.

This paper consists of six sections. In Section II and III, we introduce the frameworks of the Bayesian learning and the MCMC method respectively. In Section IV, the simulation results of the exchange MC method are described. Finally, discussion and conclusion are followed in Section V and Section VI.

2. Frameworks of Bayesian learning

Let $X^n = (X_1, X_2, ..., X_n)$ be n training samples independently and identically taken from the true distribution $q(x)$. In the Bayesian learning of a learning machine $p(x|w)$ whose parameter is w, the prior distribution $\varphi(w)$ of the parameter w needs to be set. Then the posterior distribution $p(w|X^n)$ is defined by the given dataset X^n and the prior distribution $\varphi(w)$ as follows,

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

where $Z(X^n)$ is the normalization constant. In the Bayesian learning, the predictive distribution $p(x|X^n)$ is given by averaging the learning machine over the posterior distribution,

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

which estimates the true density function of X given dataset X^n.

The Bayesian posterior distribution can be rewritten as

$$p(w|X^n) = \frac{1}{Z_0(X^n)}\exp(-nH_n(w))\varphi(w), \qquad (1)$$

where $H_n(w)$ is the empirical Kullback-Leibler divergence,

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

and $Z_0(X^n)$ is another normalization constant. Hereafter, the posterior distribution is represented by Eq.(1).

The stochastic complexity $F(X^n)$ is defined by

$$F(X^n) = -\log Z_0(X^n),$$

which depends on the learning machine $p(x|w)$ and the prior distribution $\varphi(w)$. The stochastic complexity is used as a criterion for the model selection and the hyperparameter optimization [8] [9].

The average stochastic complexity $F(n)$ is defined by

$$F(n) = E_{X^n}\left[F(X^n)\right],$$

where $E_{X^n}[\cdot]$ denotes the expectation value over all sets of training samples. It was proved that $F(n)$ has the following asymptotic form [1] [2],

$$F(n) = \lambda\log n - (m-1)\log\log n + O(1), \qquad (2)$$

where λ and m are the rational number and the natural number respectively. In regular statistical models, λ is equal to the half of the number of parameters and $m = 1$. On the contrary, in the singular learning machines, λ is not larger than the half of the number of parameters and $m \geqq 1$. Since the increase of the average stochastic complexity is equal to the generalization error, the Bayesian learning provides the better generalization performance than the maximum likelihood method.

However, in the Bayesian learning, we need to compute the expectation over the posterior distribution, which usually cannot be carried out analytically. Hence, the MCMC method is applied to the Bayesian learning in singular learning machines.

3. Markov Chain Monte Carlo method

By the MCMC method, we obtain the sample sequence which converges in law to the random variable subject to a target probability distribution $P(w)$. The Metropolis algorithm is well known as one of MCMC methods.

3.1 Metropolis algorithm

The function $\hat{H}(w)$ is defined as

$$P(w) \propto \exp(-\hat{H}(w)).$$

One starts at an arbitrary point in the parameter space to be sampled, w_0. The general iteration at any point in the sequence w_k is to repeat the following cycle K times:

1. Select a new trial position $w' = w_k + \Delta w$, where Δw is randomly chosen from a symmetric step distribution.
2. Calculate the ratio $r = \exp(-(\hat{H}(w') - \hat{H}(w_k)))$.
3. Accept the trial position, that is, set $w_{k+1} = w'$ if $r \geqq 1$, or with probability r, if $r < 1$, otherwise, stay put, $w_{k+1} = w_k$.

When the Metropolis algorithm is employed in the computation of expectation over the Bayesian posterior distribution of a singular learning machine, it requires vast computational resources [5]. The characteristic time to generate a sample sequence which converges to the Bayesian posterior distribution increases rapidly as the number n of the training samples

increases. This is caused by the fact that the target distribution is complexly distributed in the parameter space.

3.2 Exchange Monte Carlo method

Recently, various improvements have been made on MC algorithms to overcome the problem of complex local minima which are separated from each other by the high energy barriers. The exchange MC method is proposed as one of the improved algorithms [7].

The exchange MC method treats a compound system which consists of non-interacting L sample sequences of the system concerned. The elements of the l-th sample sequence $\{w_l\}$ converge in law to the random variable which is subject to the following probability distribution

$$P_l(w) \propto \exp(-t_l \hat{H}(w)) \quad (1 \leqq l \leqq L),$$

where $t_1 < t_2 < \cdots < t_L$. Given a set of the temperatures $\{t_l\}$, the simultaneous distribution for finding $\{w_l\} = \{w_1, w_2, \cdots, w_L\}$ is expressed as a sample product formula

$$P(\{w_l\}; \{t_l\}) = \prod_{l=1}^{L} P_l(w_l). \tag{3}$$

The exchange MC method is based on two types of updating in constructing a Markov chain. One is conventional update based on the Metropolis algorithm for each target distribution $P_l(w_l)$. In addition to the Metropolis algorithm, we carry out the position exchange between two sequences, that is, $\{w_l, w_{l+1}\} \to \{w_{l+1}, w_l\}$. The transition probability $P(w_l, w_{l+1}; t_l, t_{l+1})$ is defined by

$$P(w_l, w_{l+1}; t_l, t_{l+1}) = \min(1, \exp(-\Delta)),$$

where

$$\Delta(w_l, w_{l+1}; t_l, t_{l+1}) = (t_{l+1} - t_l)(\hat{H}(w_l) - \hat{H}(w_{l+1})).$$

Under these updates, the simultaneous distribution of Eq.(3) is invariant because these updates satisfy the detailed balance condition for the distribution of Eq.(3). Consequently, the following two steps are carried out in alternate shifts:

(1) Each sequence is simulated simultaneously and independently for a few iterations by Metropolis algorithm.

(2) Two positions are exchanged with the probability $P(w_l, w_{l+1}; t_l, t_{l+1})$.

The exchange MC method is widely used for many applications and its effectiveness has been clarified. In this paper, we propose that the exchange MC method is appropriate for computing the Bayesian learning in the singular learning machines, and show its effectiveness by experimental results.

4. Experiment

In this section, we clarify the effectiveness of the exchange MC method by comparing the theoretical value of the stochastic complexities with the experimental ones.

In these experiments, we consider the case of averaging over all sets of training samples, that is to say, the Bayesian posterior distribution and the stochastic complexity are respectively rewritten as

$$p(w|X^n) \propto \exp(-nH(w))\varphi(w)$$
$$F(n) = -\log \int \exp(-nH(w))\varphi(w)dw,$$

where H(w) is the Kullback-Leibler divergence from the true distribution $q(x)$ to the learning machine $p(x|w)$ as follows,

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

It is well known that $H(w) \geqq 0$ and that $H(w) = 0$ is equivalent to $q(x) = p(x|w)$. The dimension of the parameter space is denoted by d, that is, $w = \{w_1, \cdots, w^d\}$, and $H(w)$ and $\varphi(w)$ are respectively defined by

$$H(w) = \prod_{j=1}^{d} (w^j)^2$$
$$\varphi(w) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{d}(w^j)^2\right).$$

Then, the asymptotic stochastic complexity is theoretically given by

$$F(n) = \frac{1}{2}\log n - (d-1)\log\log n + \frac{d-1}{2}\log\pi + \frac{d}{2}\log 2 - \log(1 + O(\frac{1}{\log n})). \tag{4}$$

Note that the last term, $O(\frac{1}{\log n})$, is in proportion to $1/\log n$. For example, if d is 2, the last term can be calculated as follows,

$$\frac{4\log 2 - \gamma}{\log n},$$

where γ is Eular constant. By comparing the numerical $F(n)$ obtained by the MCMC method with the theoreical one, we can measure the accuracy of the MCMC method.

4.1 Experimental calculation for stochastic complexity

One of the advantages of the MCMC method is to be able to compute the expectation over a density function not normalized. However, it is generally difficult to compute the normalization constant by using the MCMC method in spite of its importance. Therefore, a lot of calculation methods of normalization constant (free energy, marginal likelihood or stochastic complexity) have been proposed [10]. One of the methods is described as follows.

Let us introduce the function $f(n, t)$,

$$f(n, t) = -\log \int \exp(-ntH(w))\varphi(w)dw.$$

Then it immediately follows that $f(n,0) = 0$. Our purpose is to calculate $F(n) = f(n,1)$. It can be expressed by

$$\begin{aligned} f(n,1) &= \left\{ \sum_{l=1}^{L-1} f(n,t_{l+1}) - f(n,t_l) \right\} \\ &= \sum_{l=1}^{L-1} \log \frac{\int e^{-nt_{l+1}H(w)}\varphi(w)dw}{\int e^{-nt_lH(w)}\varphi(w)dw} \\ &= \sum_{l=1}^{L-1} \log \frac{\int e^{-n(t_{l+1}-t_l)H(w)}e^{-nt_lH(w)}\varphi(w)dw}{\int e^{-nt_lH(w)}\varphi(w)dw}, \end{aligned}$$

where the set of temperatures $\{t_l\}$ is determined by $t_1 = 0$ and $t_L = 1$. By using the probability distribution $Q_l(w)$,

$$Q_l(w) \propto \exp(-nt_lH(w))\varphi(w),$$

$f(n,1)$ is rewritten as

$$f(n,1) = \sum_{l=1}^{L-1} \log \int \exp(-n(t_{l+1} - t_l)H(w))Q_l(w)dw.$$

This integration can be regarded as the expectation of $\exp(-n(t_{l+1} - t_l)H(w))$ over the distribution Q_l. Hence, the value of the stochastic complexity can be calculated by generating the L sample sequences from the probability distribution $Q_l : l = 1, 2, \cdots, L$.

4. 2 Experimental results

In the experiments, the number L of the set of the temperatures $\{t_1, \cdots, t_L\}$ is configured as 32, and the temperature t_l is defined as

$$t_l = \begin{cases} 0 & (\text{if } l = 1) \\ 2^{-L+l} & (\text{otherwise}). \end{cases}$$

Note that $t_L = 1$. The initial value of the parameter w is randomly selected from the prior distribution $\varphi(w)$. For calculating the expectation, we use the last fifty percents of the sample sequence in order to reduce the influence of the initial value. Δw of each Metropolis algorithm is randomly chosen from the uniform distribution with the range $[-D(n,t), D(n,t)]$, and the value of $D(n,t)$ is optimized so that the acceptance ratio of the Metropolis algorithm is within the range of sixty percents to eighty percents. An iteration for Step 1 of the exchange MC method is set as one. In the exchange MC method, the rule for selecting the exchange pairs is $\{(w_1, w_2), (w_3, w_4), \cdots, (w_{31}, w_{32})\}$ if the number k of MC iteration is odd, and $\{(w_2, w_3), (w_4, w_5), \cdots, (w_{30}, w_{31})\}$ otherwise. For the evaluation of the algorithm, we define the error rate G as $(\hat{f} - \hat{f}_0)/\hat{f}_0$, where $\hat{f}_0$ and $\hat{f}$ are the theoretical value of the stochastic complexity and the experimental one respectively.

Firstly, we compare the behavior of the sample sequence w in the parameter space between using the exchange MC method and using the Metropolis algorithm. we set the number n of the training samples 10000, the number K of the MC iterations 10000 and the number d of the dimensions 2. Figure 1 and Figure 2 respectively show the result of the exchange MC method and that of the Metropolis algorithm. Comparing two results, it is clear that the sample sequence generated by the Metropolis method is localized in the parameter space, whereas that by the exchange MC method can accurately approximate the target distribution in the parameter space.

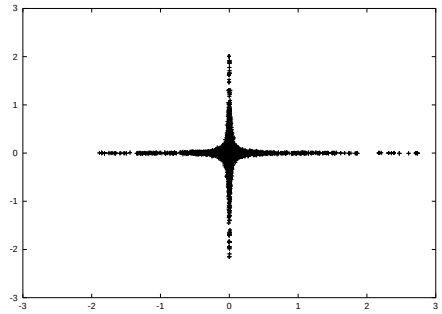

図 1 Behavior of the sample sequence in the parameter space using the exchange MC method in the case that $n = 10000$, $K = 10000$ and $d = 2$.

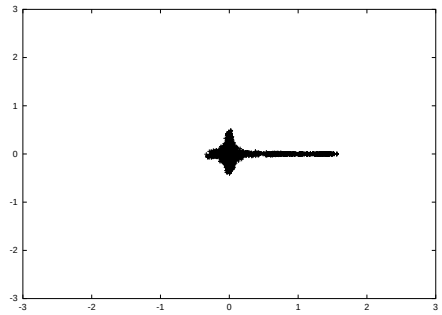

図 2 Behavior of the sample sequence in the parameter space using the Metropolis algorithm in the case that $n = 10000$, $K = 10000$ and $d = 2$.

Secondly, in order to evaluate the accuracy of the algorithm quantitatively, we calculate the error rates of the stochastic complexity. Figure 3 and Figure 4 respectively show the error rates of the stochastic complexity using the exchange MC method and using the Metropolis algorithm. The horizontal axis shows the common logarithm of the MC iterations and the vertical one the error rate of the stochastic complexity. In these figures, we set the number n of the training samples 100000 and the number d of the dimensions 2. Since the MCMC method is the probabilistic algorithm, we average the error rates over 20 independent trials. The results of the Metropolis algorithm are biased in the small MC iterations because of the influence of the initial value. On the contrary, the exchange MC method provides the small bias and variance.

In the same way, we study the dependence on the number of the training samples. The experimental results are shown

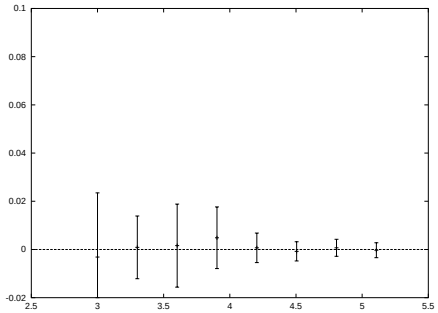

図 3 The error rate of the stochastic complexity against the MC iterations obtained by the exchange MC method in the case that $n = 100000$ and $d = 2$.

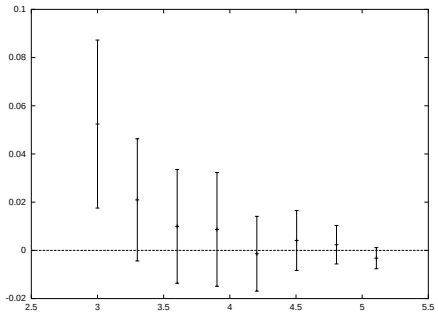

図 4 The error rate of the stochastic complexity against the MC iterations obtained by the Metropolis algorithm in the case that $n = 100000$ and $d = 2$.

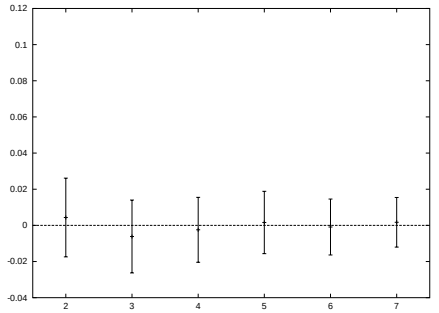

図 5 The error rate of the stochastic complexity against the number of training samples obtained by the exchange MC method in the case that $K = 8000$ and $d = 2$.

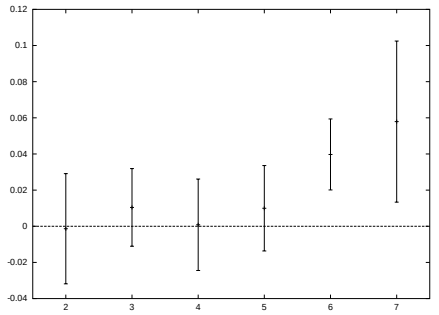

図 6 The error rate of the stochastic complexity against the number of training samples obtained by the Metropolis algorithm in the case that $K = 8000$ and $d = 2$.

in Figure 5 and Figure 6. In these experiments, we set the number K of the MC iterations 8000 and the number d of dimensions 2. The horizontal axis shows the common logarithm of n and the vertical one the error rate of the stochastic complexity. When the number n of the training samples increases, the target distribution is more localized in the parameter space except the analytic set $H(w) = 0$. Therefore, in order to carry out the Metropolis algorithm effectively, the value of $D(n,t)$ has to be small against n in spite of the fact that the extension of the target distribution in the analytic set $H(w) = 0$ is constant against n. By the reason, in the Metropolis algorithm, the biased value of the stochastic complexity is obtained for the large n. On the contrary, the exchange MC method gives us the correct value of the stochastic complexity compared to the theoretical value.

Finally, we study the dependence on the number of dimensions. We set the number n of the training samples 100000 and the number K of the MC iterations 8000. Figure 7 and Figure 8 show the results of the experiments. The horizontal axis shows the number d of dimensions and the vertical one the error rate of the stochastic complexity. In the case of Eq.(4), the dimensionality of the analytic set of $H(w) = 0$ is $d - 1$. Therefore, it is more difficult to generate samples from the target distribution as the number d of dimensions increases. The influence is found by the result of the Metropolis algorithm, which show the biased value of the stochastic complexity in the large d. However, we can see that the exchange MC method overcomes the influence by the experimental results.

5. Discussion

In this paper, we proposed that the exchange MC method is appropriate for the Bayesian learning of the singular learning machines and clarified its effectiveness experimentally by comparing the theoretical value of the stochastic complexity with the experimental one. As a result, we found that the experimental value of the stochastic complexity using the exchange MC method converges to the theoretical value in the smaller number of MC iterations than using the Metropolis algorithm. This is caused by the fact that the exchange between the sample sequence in the small t and in the large t overcomes the difficulty of sampling from the complex Bayesian posterior distribution. As mentioned in Section III.B, the Bayesian posterior distribution is widely and complexly distributed in the parameter space. However, in small t, the target distribution is not so complex. Therefore, it is easy to generate the sample sequence from the target distribution in small t comparing to that in large t.

Let us discuss two points in association with this paper.

First, we discuss the prediction accuracy of the Bayesian

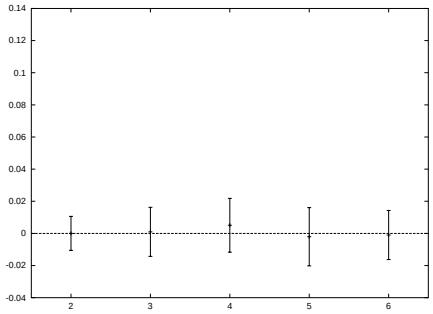

図 7 The error rate of the stochastic complexity against the number of dimensions obtained by the exchange MC method in the case that $n = 100000$ and $K = 8000$.

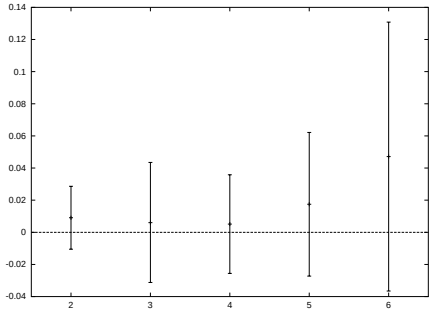

図 8 The error rate of the stochastic complexity against the number of dimensions obtained by the Metropolis algorithm in the case that $n = 100000$ and $K = 8000$.

learning using the exchange MC method. In our experiments, we calculate the stochastic complexity in order to evaluate the effectiveness of the exchange MC method. However, the calculation of the stochastic complexity needs a lot of sample sequences from the target distribution of the different temperature t, while the calculation of the predictive distribution needs only one sample sequence from the Bayesian posterior distribution, that is, the target distribution of $t = 1$. Therefore, in order to evaluate the prediction accuracy, we have to calculate the expectation value over the Bayesian posterior distribution, which is one of important future studies.

Second, we discuss the model selection problems. As we mentioned in Section II, the stochastic complexity is used as a criterion for the model selection. In particular, for the singular learning machines, since the coefficient λ of the stochastic complexity include the information of the true model, the information criterion for singular learning machines has been proposed recently [11]. The results in this paper show that the exchange MC method is efficient for calculating the stochastic complexity. Consequently, by these result, we can see that the exchange MC method is efficiently used for the model selection problem.

6. Conclusion

In this paper, we proposed that the exchange MC method was appropriate for the Bayesian learning of singular learning machines and clarified the effectiveness of the exchange MC method by calculating the stochastic complexity. As a result, we found the following properties.

- The exchange MC method provides the sample sequence which converges to the Bayesian posterior distribution in the smaller number of MC iterations than the Metropolis algorithm.
- The exchange MC method also achieves the above effect in the large number of training samples and in the higher-dimensional space of the parameters.

In the future works, the following assignments should be addressed,

- Applying the exchange MC method to the practical Bayesian learning of singular learning machines.
- Clarifying the effectiveness of the exchange MC method for the model selection problem.
- Comparing the property of the exchange MC method to those of other MCMC methods.

This work was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research 15500130.

文　献

[1] S.Watanabe, “Algebraic analysis for nonidentifiable learning machines,” Neural Computation, Vol.13, No.4, pp.899-933, 2001.

[2] S.Watanabe, “Algebraic geometrical methods for hierarchical learning machines,” Neural Networks, Vol.14, No.8, pp.1049-1060, 2001.

[3] M.Aoyagi, S.Watanabe, “The generalization error of reduced rank regression in Bayesian estimation.” In Proc. of ISITA, pages 1068-1073, Parma, Italy, 2004.

[4] K.Yamazaki, S,Watanabe, “Singularities in mixture models and upper bounds of stochastic complexity,” Neural Networks, Vol.16, No.7, pp.1029-1038, 2003.

[5] N.Nakano, K.Takahashi, S.Watanabe, “On the Evaluation Criterion of the MCMC Method in Singular Learning Machines” , Trans. of IEICE, Vol.J88-D-2, No.10, pp.2011-2020, 2005.

[6] Y.Iba, “Extended Ensemble Monte Carlo”, International Journal of Modern Physics, C12, pp.623-656, 2001.

[7] K.Hukushima, K.Nemoto, “Exchange Monte Carlo Method and Application to Spin Glass Simulation”, Journal of the Physical Society of Japan, Vol.65, No.6, pp.1604-1608, 1996.

[8] H.Akaike, “Likelihood and Bayes procedure”, Bayesian Statistics, University Press, pp.143-166, 1980.

[9] G.Schwarz, “Estimating the dimension of a model”, Annals of Statistics, Vol.6, No.2, pp.461-464, 1978.

[10] Y.Ogata, “A Monte Carlo Method for an Objective Bayesian Procedure.” , Ann. Inst. Statis. Math. , 42(3), pp.403-433, 1990.

[11] K.Yamazaki, K.Nagata, S.watanabe, “A New Method of Model Selection Based on Learning Coefficient”, in Proc. of NOLTA, pp.389-392, 2005.