Generalization Performance of Exchange Monte Carlo Method for Normal Mixture Models

Kenji Nagata[1] and Sumio Watanabe[2]

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, MailBox R2-5,4259, Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan
kenji.nagata@cs.pi.titech.ac.jp
[2] P&I Lab.,Tokyo Institute of Technology
swatanab@pi.titech.ac.jp

Abstract. A normal mixture model, which belongs to singular learning machines, is widely used in statistical pattern recognition. In singular learning machines, the Bayesian learning provides the better generalization performance than the maximum likelihood estimation. However, it needs huge computational cost 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 in singular learning machines, and experimentally show that it provides better generalization performance in the Bayesian learning of a normal mixture model than the conventional Monte Carlo method.

1 Introduction

A normal mixture model is a learning machine which estimates the target probability density by sum of normal distributions. This learning machine is widely used in statistical pattern recognition and data clustering. Normal mixture models belong to singular learning machines because they have singular points where the Fisher information matrices are degenerate. In singular learning machines, it is well known that the Bayesian learning provides better generalization performance than the maximum likelihood estimation that tends to produce a learning machine overfitting the data[1].

In the Bayesian learning, it is necessary to realize the Bayesian posterior distribution accurately around the singular points. 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 cost to approximate the Bayesian posterior distributions of singular learning machines [3]. 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 based on the idea of an extended ensemble method, which is surveyed in [4]. This method gives us a general strategy for overcoming the problem of huge computational cost. An exchange Monte Carlo (MC) method is well known as

one of the extended ensemble method [5], and its effectiveness has been shown in a spin glass [5], a strongly correlated system, an optimization problem and many other applications.

In this paper, we propose that the exchange MC method is appropriate to compute the Bayesian learning in singular learning machines, and experimentally show that the exchange MC method provides better generalization performance in the Bayesian learning of a normal mixture model than the Metropolis algorithm.

This paper consists of five sections. In Section 2, we introduce the normal mixture models and the frameworks of the Bayesian learning and the MCMC method respectively. In Section 3, the exchange MC method and its application to the Bayesian learning are described. In Section 4, we state the experimental result. Finally, discussion and conclusion are followed in Section 5.

2 Background

2.1 Normal Mixture Models

Suppose that $g(x|b, \Sigma)$ is a density function of an M-dimensional normal distribution whose mean is $b \in R^M$ and variance-covariance matrix is $\Sigma \in R^{M \times M}$. A normal mixture model $p(x|w)$ of an M-dimensional input $x \in R^M$ with a parameter vector w is defined by $p(x|w) = \sum_{k=1}^{K} a_k g(x|b_k, \Sigma_k)$, where the integer K is the number of components and $\{a_k | a_k \geq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of coefficients. The parameter w of this learning machine is $w = \{a_k, b_k, \Sigma_k\}_{k=1}^{K}$.

In some applications, the parameter is confined to the mean of each component and it is supposed that there are no correlation between each input dimension. In this case, the learning machine is rewritten by

$$p(x|w) = \sum_{k=1}^{K} \frac{a_k}{\sqrt{2\pi\sigma_k^2}^M} \exp\left(-\frac{||x - b_k||^2}{2\sigma_k^2}\right),$$

where $\sigma_k > 0$ is a constant. This means $w = \{a_k, b_k\}_{k=1}^{K}$. Hereafter, we consider learning in this type of normal mixture models.

Normal mixture models belong to singular learning machines because they have singular points in their parameter space. Let us illustrate the singularities by the simplest example. Assume that the true distribution $q(x)$ of a one-dimensional input x is defined by $q(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-b^*)^2}{2}\right)$. This distribution has one component. Also assume that a learning machine is defined by $p(x|w) = \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{(x-b_1)^2}{2}\right) + \frac{1-a}{\sqrt{2\pi}} \exp\left(-\frac{(x-b_2)^2}{2}\right)$, which has two components. The set of true parameter is $\{a = 1, b_1 = b^*\} \cup \{a = 0, b_2 = b^*\} \cup \{b_1 = b_2 = b^*\}$. This set has singular points where two sets of true parameters are crossing (Figure 1). At a singular point, Fisher information matrix is degenerate. Therefore, it is generally difficult to clarify the property of learning for singular learning machines theoretically.

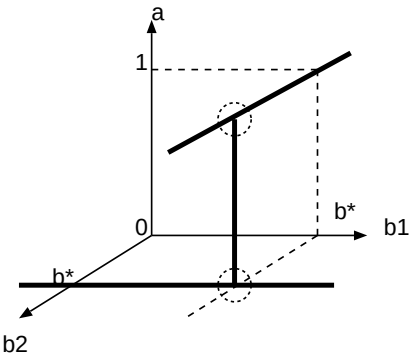

Fig. 1. Singularity of normal mixture models. Thick lines are set of true parameters. Circle center indicates a singular point.

2.2 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, which is also known as the marginal likelihood or as the evidence. 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 generalization error is defined by

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

which indicates the Kullback-Leibler divergence from the true distribution to the predictive distribution. The averaged generalization error has the following asymptotic form,

$$E_{X^n}[G(X^n)] = \frac{\lambda}{n} - \frac{m-1}{n\log n} + O\left(\frac{1}{n\log n}\right),$$

where the notation $E_{X^n}[]$ shows the value of expectation over all sets of training samples. The values of rational number λ and natural number m depend on the learning machine and the prior distribution. Recently, an algebraic geometrical

method for singular learning machines has been established, and their learning coefficients have been clarified [1]. According to the results, in the redundant case, the upper bound of λ for the normal mixture models is $\lambda \leq \frac{MK_0+K_0-1}{2} + \frac{K-K_0}{2}$[2], where K_0 and K are respectively the number of components for the true distribution and for a learning machine, and M is the dimension for data.

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.

2.3 Markov Chain Monte Carlo method for Bayesian Learning

The MCMC method is the algorithm to obtain the sample sequence which converges in law to the random variable subject to a target probability distribution. The Metropolis algorithm is well known as one of MCMC methods[9]. Given the target density function, the Metropolis algorithm can be applied even if the normalization constant is not clarified. Therefore, the sample sequence from the Bayesian posterior distribution can be obtained by the Metropolis algorithm.

However, 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 [3]. 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.

The Bayesian posterior distribution has most of its density around the true parameters. The variance of this distribution becomes small as the number n of training samples increases. As we mentioned in Section 2.1, in singular learning machines, the set of true parameter(s) is not a point but an analytic set like Figure 1. Therefore, the Bayesian posterior distribution for a singular learning machine is complexly distributed in the parameter space. On the contrary, Metropolis algorithm is based on local updating. Moreover, the smaller the variance of target distribution is, the more local the updating of Metropolis algorithm has to become. Consequently, it requires huge cost to generate a sample sequences to converge to the Bayesian posterior distribution for a singular learning machine by the Metropolis algorithm.

3 Proposal

In this paper, we propose that the exchange MC method is appropriate for Bayesian learning in singular learning machines.

3.1 Exchange Monte Carlo method

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

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

where $t_1 < t_2 < \cdots < t_L$. Given a set of the temperatures $\{t\}$, the simultaneous distribution for finding $\{w\} = \{w_1, w_2, \cdots, w_L\}$ is expressed as a simple product formula by $P(\{w\}; \{t\}) = \prod_{l=1}^{L} P_l(w)$. The exchange MC method is based on two types of updating in constructing a Markov chain. One is conventional updates based on the Metropolis algorithm for each target distribution $P_l(w)$. 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

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

Under these updates, the simultaneous distribution is invariant because these updates satisfy the detailed balance condition for the simultaneous distribution.

Consequently, the following two steps are carried out in alternate shifts:

1. Each sequence is simulated simultaneously and independently for a few iteration by Metropolis algorithm.
2. Two positions are exchanged with the probability $P(w_l, w_{l+1}; t_l, t_{l+1})$.

3.2 Application to the Bayesian Learning

The exchange MC method can be applied to Bayesian learning by defining the probability distribution $p_l(w|X^n)$ as

$$p_l(w|X^n) = \frac{1}{Z_l(X^n)} \varphi(w) \left(\prod_{i=1}^{n} p(X_i|w) \right)^{t_l}.$$

As mentioned in Section 2.3, the Bayesian posterior distribution, which is equal to $p_l(w|X^n)$ for $t_l = 1$, is complexly distributed. In the case that $0 < t_l < 1$, the distribution $p_l(w|X^n)$ is distributed less complexly than the Bayesian posterior distribution. Moreover, the distribution $p_l(w|X^n)$ for $t_l = 0$ has no complexity because it is equal to the prior distribution. Therefore, by using the distribution $p_l(w|X^n)$ for $0 \leq t \leq 1$ as the target distribution for the exchange MC method, we expect to obtain a sample sequence to converge to the posterior distribution in less samples than the Metropolis algorithm. 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 present the experimental results where the Bayesian learning is simulated for the mixture model with the 3-dimensional gaussian component

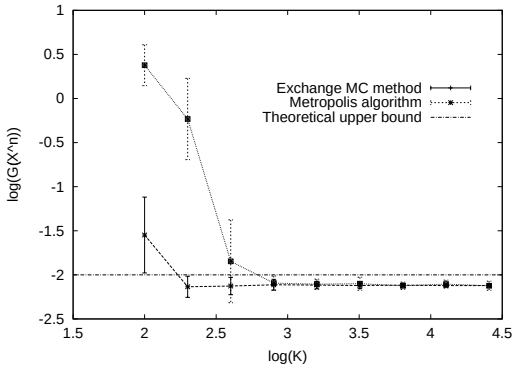

Fig. 2. The results of generalization error using the exchange MC method and the Metropolis algorithm.

$c(x|b) = \frac{1}{(2\pi)^{3/2}} \exp(-\frac{||x-b||^2}{2})$. This means $M = 3$. We compare the exchange MC method with the Metropolis algorithm by applying each method to sample producing from the Bayesian posterior distribution.

In these experiments, the number K_0 of components in the true distribution is set as 2, and the number K in the learning machine 5. The true distribution is set to $q(x) = 0.52 * c(x|(-1.19, 1.43, 3.50)^T) + 0.48 * c(x|(3.54, 2.01, 2.35)^T)$. We prepare a sample set with the sample size $n = 500$ from this true distribution. The prior distributions for the parameter a_k and b_k are respectively defined as a uniform distribution with the range $[0, 1]$ and a 3-dimensional standard gaussian distribution.

The number L of the set of the temperatures $\{t_1, \cdots, t_L\}$ is configured as 42, and the temperature t_l is defined as

$$t_l = \begin{cases} 0 & (\text{if } l = 1) \\ (1.25)^{-L+l} & (\text{otherwise}). \end{cases}$$

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. An iteration for Step 1 of the exchange MC method is set as 1. In the exchange MC method, the rule for selecting the exchange pairs is $\{(w_1, w_2), (w_3, w_4), \cdots, (w_{41}, w_{42})\}$ if the number k of MC iteration is odd, and $\{(w_2, w_3), (w_4, w_5), \cdots, (w_{40}, w_{41})\}$ otherwise.

Firstly, for the evaluation of the algorithm, we calculate the generalization error, which is approximated by $\frac{1}{n'} \sum_{i=1}^{n'} \log \frac{q(x_i')}{p(x_i'|X^n)}$ with test data $\{x_i'\}_{i=1}^{n'=2500}$ generated from the true distribution. Figure 2 shows the average of the generalization errors. The horizontal axis shows the base-10 logarithm of MC iteration, and the vertical one the base-10 logarithm of the generalization error. The value of MC iteration is changed from 100 to 25600. The horizontal line shows the theoretical upper bound of the generalization error. Comparing two

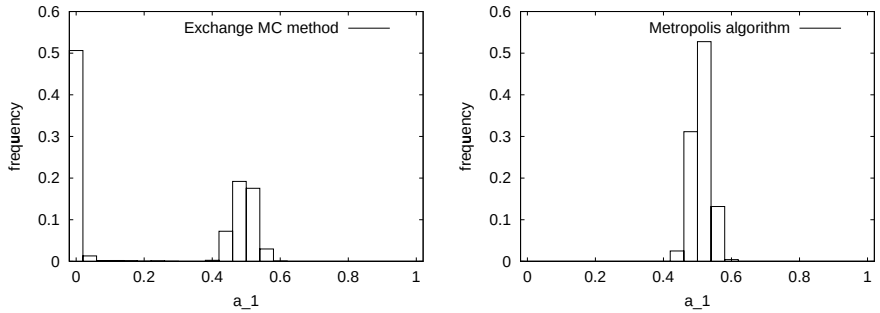

Fig. 3. The histograms of the parameter a_1. The left is obtained by the exchange MC method and the right by the Metropolis algorithm.

results of each algorithm, convergence of generalization error for the exchange MC method is faster than that for Metropolis algorithm. However, after each algorithm is converged, there is little difference between two algorithms. Note that the computational cost of the exchange MC method is higher than that of the Metropolis algorithm. However, by the parallel processing, we can make the computational time of the exchange MC method equal to that of the Metropolis algorithm.

Secondly, we compare the distribution of the sample sequences obtained by the exchange MC method with that by the Metropolis algorithm. Figure 3 shows the histograms of the parameter a_1, the coefficient of one of 5 components. The left part of Figure 3 is obtained by the exchange MC method and the right part the Metropolis algorithm. Considering the true parameters, the marginal distribution for the parameter a_1 has peaks near $a_1 = 0$ and near $a_1 = 0.5$. Consequently, the exchange MC method generates the correct histogram while the Metropolis algorithm generates the localized histogram.

5 Discussion and Conclusion

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 simulating learning of the normal mixture model. As a result, we found that the experimental value of the generalization error using the exchange MC method converges in the smaller number of MC iterations than using the Metropolis algorithm. Moreover, after converging, the exchange MC method can approximate the Bayesian posterior distribution more accurately than the Metropolis algorithm.

In this section, we discuss experimental results. In the setting of Section 3.2, the transition probability for the exchange MC method depends on minus of the logarithm likelihood, that is,

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

$$L(w) = -\sum_{i=1}^{n} \log p(X_i|w)$$

We assume $t_{l+1} > t_l$ in Section 3.1. Hence, if $L(w_l) < L(w_{l+1})$, two positions,w_l and w_{l+1}, are exchanged with probability 1. Therefore, the exchange MC method works to make the likelihood of the parameter w_L for $t = 1$ become large preferentially. This is why the value of the generalization error using the exchange MC method converges fast.

After converging, all the samples $\{w_l\}$ tend to be near the true parameters. On the true parameters, the value of likelihood for any parameters are equal. Therefore, all combination of two samples w_l and w_{l+1} are exchanged frequently. This is why the exchange MC method can realize the Bayesian posterior distribution more accurately than the Metropolis algorithm.

However, in spite of the fact that the Metropolis algorithm produce the localized sample sequence, the generalization error has less difference between the exchange MC method and the Metropolis algorithm. One of our future works is to clarify the relationship between the convergence accuracy of a sample sequence and the generalization error or other expectation values.

Acknowledgement:This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 18-5809 and for Scientific Research 15500130, 2006.

References

1. S.Watanabe, "Algebraic analysis for nonidentifiable learning machines," Neural Computation, Vol.13, No.4, pp.899-933, 2001.
2. K.Yamazaki, S,Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," Neural Networks, Vol.16, No.7, pp.1029-1038, 2003.
3. 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.
4. Y.Iba, "Extended Ensemble Monte Carlo", International Journal of Modern Physics, C12, pp.623-656, 2001.
5. 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.
6. P.Sengupta, A.W.Sandvik, D.K.Campbell, "Bond-order-wave phase and quantum phase transitions in the one dimensional extended Hubbard model", Physical Review B, vol.65, 155113, 2002.
7. K.Pinn, C.Wieczerkowski, "Number of magic squares from parallel tempering Monte Carlo", Int. J. Mod. Phys. C9, 541, 1998.
8. K.Hukushima, "Extended ensemble Monte Carlo approach to hardly relaxing problems", Computer Physics Communications, 147, pp.77-82, 2002.
9. K.Nagata, S.Watanabe, "Exchange Monte Carlo Method for Bayesian Learning in Singular Learning Machines", Proc of International Joint Conference on Neural Networks 2006 (IJCNN2006), to appear.