2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Active Update in Single Component MCMC for Discrete Variables

Shotaro Akaho[†] and Kazuya Takabatake[†]

[†]Neuroscience Research Institute
AIST (The National Institute of Advanced Industrial Science and Technology)
Tsukuba 305-8568, Japan
Email: s.akaho@aist.go.jp

Abstract—We propose a single component Markov chain Monte Carlo (MCMC) method that converges faster than Gibbs sampler. A basic idea is to change the value of state variables as frequently as possible. However, just increasing the changing probability often leads to high rejection rate. Therefore, we choose a transition probability among a class of distributions that ensure zero rejection rate. We propose a method for calculating a transition matrix with a large changing probability. The method requires just slight computation, and we show that the proposed method converges faster than Gibbs sampler theoretically for simple cases. For larger size problems, we examine the performance by synthetic simulation.

1. Introduction

Graphical models have become increasingly popular for Bayesian inference in various areas of science and engineering. If a graphical model is singly-connected without any loop, a belief propagation method can be applied to obtain exact statistics such as a posterior mean. However, the computation time increases combinatorially for multiply-connected models. Therefore a lot of methods to obtain approximate values have been proposed.

One approach is based on Markov chain Monte Carlo (MCMC)[2, 3]. In MCMC, random samples are generated from a distribution whose limiting distribution is identical to the posterior, and the posterior mean is approximated by their average. MCMC approach is considered to be slower than other methods such as mean field approximation[4]. However, MCMC is still used in various areas of science, because it can approximate the exact value as precisely as possible if computation time permits.

In this paper, we focus on the case of single component MCMC in which the whole random variable is divided into components and the update is performed sequentially for each component. Furthermore, we consider specifically the case of discrete variable, though the basic concept can be applied to continuous variable cases.

Among single component MCMC methods, Gibbs sampler (heat-bath algorithm) is the most commonly used method[1]. The proposal distribution of the Gibbs sampler is identical to the full conditional distribution for the current variable. One of the authors showed that the Gibbs sampler performs best among single component MCMC methods in a greedy sense from an information geometric interpretation[5].

However, fast convergence of distribution is not necessarily equivalent to quick convergence of Monte Carlo integration. Let us consider a simple example. Suppose we have only one binary (0-1) unit and the target distribution is 1/2 for both values. The Gibbs sampling is nothing but a coin flipping in this case, and the mean value is estimated by the frequency of one side of the coin. Although it converges to the true value 1/2, the average of another sequence $01010101\cdots$ converges much faster. We extend this idea to more general cases of single component MCMC for discrete variables.

Intuitively, the basic idea is changing the current state more often. However, if we increase the probability of changing a state, a rejection rate will also increase in general, which results in the same state. In this paper, we consider a method for finding such a distribution with zero rejection rate by just small computation time.

On the acceleration of Gibbs sampler, a lot of contributions have been proposed. However, most of them focus on continuous cases, while our approach is efficient in particular for discrete cases.

2. Single component MCMC

What we want to evaluate is the average of a function $f(X)$ of a random variable X with respect to some distribution $\pi(X)$. In Bayesian inference, X is a set of unobserved variables in graphical models and $\pi(X)$ is a posterior distribution conditioned by observed variables.

MCMC approximates $E_\pi[f(X)]$ by Monte Carlo simulation, $E[f(X)] \simeq (1/T)\sum_{t=0}^{T} f(X(t))$, where the random sequence $X(0), X(1), \ldots, X(T)$ is generated by Markov chain $P(X(t+1) \mid X(t))$, i.e., the current state $X(t+1)$ depends only on the previous state $X(t)$. If the limiting distribution of $P(X(t+1) \mid X(t))$ from any initial solution is equal to $\pi(X)$, the Monte Carlo integration converges to the true value of $E[f(X)]$.

The single component MCMC is a kind of MCMC in which the random variable X is divided into components $\{X_1, \ldots, X_N\}$, and the states of the components are changed one by one as follows: Before the update of the i-th component at time step $t+1$, the current state of X is given by $\{X_i(t), X_{-i}(t)\}$, where $X_{-i}(t)$ is defined by $X_{-i}(t) = \{X_1(t+1), \ldots, X_{i-1}(t+1), X_{i+1}(t), \ldots, X_N(t)\}$, which is all components of the current X except $X_i(t)$. Then $X_i(t+1)$ is generated by an appropriate Markov chain with the transition probability $P(X_i(t+1) \mid X_i(t), X_{-i}(t))$.

Gibbs sampler is a special class of single component MCMC, in which the transition probability is given by $P(X_i(t+1) \mid X_i(t), X_{-i}(t)) = \pi(X_i \mid X_{-i})$, where $\pi(X_i \mid X_{-i}) = \pi(X)/\sum_{X_{-i}} \pi(X)$ is the full conditional probability.

3. Active update scheme

From an information geometrical interpretation, Takabatake showed that the Gibbs sampler moves the current distribution to the closest point of the target distribution $\pi(X)$ among all single component MCMC by which $\pi(X)$ is stationary. Therefore, the distribution converges quickly by the Gibbs sampler in a greedy sense. However, it does not ensure quick convergence of the Monte Carlo integration.

Note that a new value of a state updated by the Gibbs sampler is independent of the old value of the state. Therefore, the state sometimes does not change the value even when the transition probability is small. That causes slow convergence of the Gibbs sampler. One way to improve the performance is increasing the transition probability. However, it will also make the rejection rate large. To avoid the increase of the rejection rate, we find a proposal distribution among a class of distributions which converges to the full conditional distribution used in the Gibbs sampler.

Here, let us introduce some notational conventions. Since we focus on the update of only one component, we omit the subscript for the component X_i unless it is necessary. Let X_i be a k valued discrete variable and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ denote the vector of full conditional probability of the component, $\pi_j = \pi(X_i = j \mid X_{-i} = X_{-i}(t))$, where the values of other components X_{-i} are fixed to $X_{-i}(t)$, and $\sum_{j=1}^{k} \pi_j = 1$. We assume $\pi_j > 0$ for all j.

The transition matrix P is a $k \times k$ matrix, and the (i,j) element represents the probability of transition from the current state i to the next state j. It has to satisfy several conditions as follows:

1. P is a probability matrix in the sense $\sum_{j=1}^{k} P_{ij} = 1$ for all i, and $0 \leq P_{ij} \leq 1$ for all i, j, i.e.,

$$P\mathbf{1}^\top = \mathbf{1}^\top, \qquad O \leq P \leq \mathbf{1}^\top \mathbf{1}, \tag{1}$$

where $\mathbf{1} = (1, \ldots, 1)$ and O is a zero matrix.

2. The target distribution $\boldsymbol{\pi}$ is stationary by P,

$$\boldsymbol{\pi} P = \boldsymbol{\pi}, \tag{2}$$

3. The chain defined by P is irreducible, i.e. for all i, j, there exists $t > 0$ that satisfies $(P^t)_{ij} > 0$.

Let us consider the transition matrix of the Gibbs sampler, $P^{\mathrm{G}} = (\boldsymbol{\pi}^\top, \cdots, \boldsymbol{\pi}^\top)^\top$ where all rows are identical because the next state is independent of the current state. This matrix satisfies all the conditions described above.

3.1. Linear programming

First of all, we explain a probably optimal but computationally intensive approach. The problem can be formulated by the LP problem, $\min_P \sum_{i=1}^{k} P_{ii}$, subject to the constraints (1) and (2). Since the irreducibility condition cannot be written as a linear constraint, we need to check whether the LP solution satisfies the irreducibility.

3.2. Diagonal element reduction

Let us consider a class of transition matrices in the form

$$P^{\mathrm{D}} = (1+\lambda)P^{\mathrm{G}} - \lambda I_k,$$

which reduces the diagonal elements of P^{G} by λ. $\boldsymbol{\pi}$ is stationary by P^{D}, and we would like to take as large λ as possible.

Theorem 1 *P^{D} satisfies all the conditions of transition matrix by which the target distribution $\boldsymbol{\pi}$ is stationary, when λ is taken appropriately. The maximal value of such λ is given by $\lambda = \min_i \pi_i / (1 - \min_i \pi_i)$.*

Although P^{D} can be obtained by very small computation, the effect is slight if there is a small π_i.

3.3. Block matrix partition

In this subsection, we introduce another class of transition matrices based on a matrix partition. Suppose K_1 and K_2 are any partition of the indices $K = \{1, \ldots, k\}$ ($K_1 \cup K_2 = K$, $K_1 \cap K_2 = \emptyset$), first let us consider the matrix in the form of

$$P^{\mathrm{B}}(K) = \begin{pmatrix} a_1 P^{\mathrm{G}}(K_1, K_1) & b_1 P^{\mathrm{G}}(K_1, K_2) \\ b_2 P^{\mathrm{G}}(K_2, K_1) & a_2 P^{\mathrm{G}}(K_2, K_2) \end{pmatrix},$$

where $P^{\mathrm{G}}(K_i, K_j)$ is defined by the submatrix of P^{G} with rows of K_i and columns of K_j.

In particular, if $a_1 = a_2 = b_1 = b_2 = 1$, $P^{\mathrm{B}}(K)$ is identical to P^{G}. Our aim is to find a solution such that a_1 and a_2 are small.

Theorem 2 *$P^{\mathrm{B}}(K)$ satisfies all the conditions of transition matrix by which the target distribution is stationary, when values of a_1, a_2, b_1, b_2 are taken appropriately. a_1 and a_2 are minimized simultaneously, when $b_1 = b_2 = b = 1/\max\{\pi(K_1), \pi(K_2)\}$, where $\pi(K_1) = \sum_{i \in K_1} \pi_i$, $\pi(K_2) = \sum_{i \in K_2} \pi_i$, and they are given by $a_1 = 1 - b\pi(K_1)/\pi(K_2)$, $a_2 = 1 - b\pi(K_2)/\pi(K_1)$. Furthermore, either a_1 or a_2 is equal to zero.*

We apply this partition procedure recursively for a nonzero diagonal block. Here we have two kinds of freedoms: One is the depth of recursion, and the other is the partition of indices at each step. If the transition matrix is calculated off-line, the recursion can be performed up to the deepest level and each partition may also be optimized by some combinatorial method. However, if we have to calculate it on-line, the recursion should be just one or two levels and the partition be performed in a deterministic way. In addition to the computation issue, we need to calculate only the row of P^{B}, not all the elements, particularly in the case of on-line.

3.4. A special case: existence of $\pi_i \geq 1/2$

A lot of numerical experiments of LP suggests that all diagonal elements can be zero except for the case that $\pi_i \geq 1/2$ for some i. In this special case, we can take the following transition matrix P^{A}, $P^{\mathrm{A}}_{ii} = 2 - 1/\pi_i, P^{\mathrm{A}}_{ij} = \pi_j/\pi_i(j \neq i), P^{\mathrm{A}}_{ki} = 1(k \neq i), P^{\mathrm{A}}_{kj} = 0(k, j \neq i)$. All components except the i-th row and the i-th column are zero. The computation cost for P^{A} is very small, and in fact, if the current state is not i, random number generation is not necessary and we just move the state to i.

3.5. Binary case

The case of binary valued variable is special, because all the above transition matrices coincide, $P^{\mathrm{LP}} = P^{\mathrm{D}} = P^{\mathrm{B}} = P^{\mathrm{A}}$.

4. Analysis of convergence

In this section, we analyze some simple cases theoretically, and compare the convergence of Monte Carlo integration of the proposed method with the Gibbs sampler.

4.1. Monte Carlo integration

Here, we evaluate a Monte Carlo integration of the random variable X_i, $r_i(T) = (1/T)\sum_{t=1}^{T} X_i(t)$, where the initial starting point $X_i(0)$ at $t = 0$ is discarded.

The statistics $r_i(T)$ is a estimator of the posterior mean μ_i, and it is the most basic but important statistics in Bayesian inference.

Let us define the estimation bias by $m_i(T) = E[r_i(T) - \mu_i]$, where the expectation is taken with respect to all Monte Carlo samples. Good estimator does not minimize only the amount of the estimation bias, but the squared error defined by $v_i(T) = E[(r_i(T) - \mu_i)^2]$.

Let $m_i^{\mathrm{G}}(T)$ and $v_i^{\mathrm{G}}(T)$ be the values for the Gibbs sampler. In this section, we analyze the case of binary variable in which all the proposed methods coincide as described in sec. 3.5, hence we write the values of the proposed method by $m_i^{\mathrm{A}}(T)$ and $v_i^{\mathrm{A}}(T)$.

4.2. One binary unit

In this subsection, we consider the case that there is only one unobserved unit X_1 that takes binary values 0 and 1. The posterior distribution for X_1 is parametrized just $\pi_1 = P(X_1 = 1 \mid X_{-1})$. Without loss of generality, we assume $\pi_1 \leq 1/2$ and the initial condition is $X_i(0) = 0$. The true value of posterior mean is given by $\mu_1 = \pi_1$.

Theorem 3 *In the case of a single binary unit, the bias of the Monte Carlo integration is asymptotically given by* $m_1^{\mathrm{G}} = 0$, *for the Gibbs sampler, and* $m_1^{\mathrm{A}} \simeq {\pi_1}^2/T$, *for the proposed method,* $where \simeq$ *represents the same* $1/T$ *order.*

The variance values are given by $v_1^{\mathrm{G}} \simeq \pi_1/T$, $v_1^{\mathrm{A}} \simeq \pi_1(1 - 2\pi_1)(1 - \pi_1)/T$ *respectively. When* π_1 *is close to 1/2, the* $1/T$ *order term in* v_1^{A} *vanishes.*

From this theorem, we conclude that the proposed method converges quickly in the sense of squared error particularly when $\pi_1 \simeq 1/2$, though it has a larger bias than the Gibbs sampler.

4.3. Two binary units (XOR type)

In this section, we investigate the case of two binary units. Since there are three free parameters even for the two binary units, it makes much more difficult to analyze in general. Through preliminary numerical simulations, we observed that the difference between the Gibbs sampler and the proposed method is small in the case of the following XOR (exclusive or) type problem. Thus, we analyze this one parameter problem.

p**-XOR problem**: Suppose we have two random variables X_1 and X_2 and let π_{ij} be the joint posterior $\pi_{ij} = P(X_1 = i, X_j = j \mid X_{-i} \cap X_{-j})$. Then the p-XOR problem is defined by $\pi_{00} = \pi_{11} = p/2, \pi_{01} = \pi_{10} = (1 - p)/2$, where $p \leq 1/2$ is a parameter. For the sake of convenience, we assume the initial state is $X_1 = X_2 = 0$ and the single component MCMC is performed in the sequence of X_1, X_2. This problem is difficult for MCMC approach because the chain has to pass a state of a small probability.

Theorem 4 *In the case of* p*-XOR problem, the bias of* X_1 *of the Gibbs and the proposed method are given by* $m_1^{\mathrm{G}} \simeq -(1 - 2p)^2/\{8p(1 - p)\}$, $m_1^{\mathrm{A}} \simeq (1 - 2p + 2p^2)/\{8p(1 - p)\}$, *respectively, and the variance is* $v_1^{\mathrm{G}} \simeq (3 - 8p + 6p^2)/\{8p(1 - p)\}$, $v_1^{\mathrm{A}} \simeq (1 - 2p)(3 - 4p + 2p^2)/\{8p(1 - p)\}$.

This theorem shows qualitatively similar relation to Theorem 3 between the Gibbs sampler and the proposed method. The proposed method has always smaller variance, and the difference is clear as p is close to $1/2$.

Note on MFA Mean field approximation (MFA) is known as a promising approach for Bayesian inference. It converges very quickly and perform effectively in particular when the degree of multiple connection is small such as error correcting codes. However, they do not produce exact solutions in general and perform poorly for strongly multiply-connected models. For the p-XOR problem, if p is less than some critical value p_c, the MFA does not converge to the true value. Therefore, the proposed method improves Gibbs sampler for a large $p(\simeq 1/2)$, and it avoids the phase transition in the MFA for a small $p(\leq p_c)$.

5. Experiments

In this section, we examine the performance of the proposed method for a larger size of problem. We consider the Boltzmann distribution, $P(X_1,\dots,X_N) = (1/Z)\exp(\sum_{i<j} J_{ij}X_iX_j + \sum_i \theta_i X_i)$, where $X_i = \pm 1$.

In the simulation, we prepared $N = 100$ units and generated J_{ij} and θ_i randomly from the Gaussian distribution $\mathcal{N}[0, 1/N]$. In order to obtain the 'true value' of posterior mean, we calculated Monte Carlo integration by 500000 steps of the Gibbs sampler where the first 10000 samples are discarded as burn-in.

5.1. Binary case

Here we compare the Gibbs sampler with the proposed method in the case that each binary variable represents each component. For a fixed configuration of J_{ij} and θ_i, MCMC is performed for 100 different sets of random numbers. The result is shown in Fig. 1. For the MCMC, we discarded the first 100 samples as burn-in, and calculated the average of successive 1000 samples. From the figure, we observe that the error curves are approximately linear and parallel in this log-log plot. Therefore, we introduce the performance index for comparison, $\rho(\mathrm{G},\mathrm{A}) = \exp((\beta^{\mathrm{G}} - \beta^{\mathrm{A}})/\max\{\alpha^{\mathrm{G}}, \alpha^{\mathrm{A}}\})$, where α, β are parameters of linear fitting $\log \mathrm{MSE} = \beta - \alpha \log t$. $\rho(\mathrm{G},\mathrm{A})$ approximately evaluates the ratio of iterations that the Gibbs sampler requires to reach the same accuracy as the proposed method. Thus, for example, if $\rho(\mathrm{G},\mathrm{A}) = 2$, the proposed method converges about two times faster than the Gibbs sampler. We calculated $\rho(\mathrm{G},\mathrm{A})$ for 10 different configurations of J_{ij}, θ_i, and it distributes 2.8 ± 0.8 (mean±std.dev.), which is larger than 1.0 with a significance <1%. The difference of computation time is negligible less than 1% of time except the calculation of the transition probability. As described earlier, the proposed method is sometimes faster because the random number generation is not always necessary.

In the figure, the results of two kinds of the MFA (naive MFA and TAP MFA) are also shown up to 100 iterations. From the simulation, however, they behave in a similar way: first they approach to the true value, then they go apart.

5.2. Grouping

In order to investigate the performance when the variable takes more than two values, we make a group with two variables, hence each group component takes 4 values from $(-1,-1)$ to (1,1).

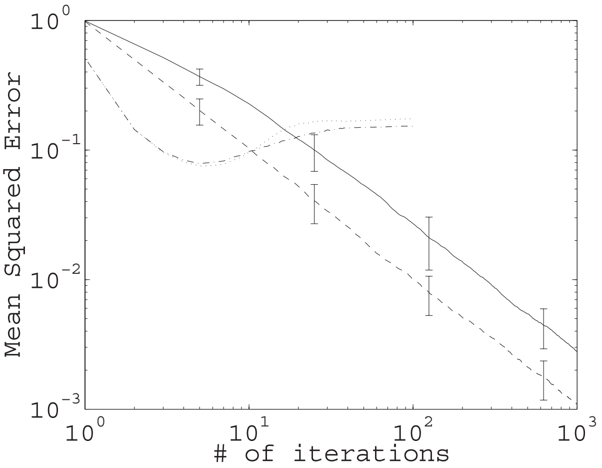

Figure 1: Comparison of the Gibbs sampler, the proposed method, and mean field approximations in binary case. Solid line: Gibbs, Dashed: Proposed, Dot-dashed: Naive MFA, Dotted: TAP MFA. Curves and error bars for MCMC methods represent the mean and standard deviation for 100 runs.

We compared the following three MCMC methods for the group components: the Gibbs sampler P^{G}, the diagonal element reduction P^{D} and the block matrix partition P^{B} of depth two recursion with a fixed partition.

The performance index for 10 different configurations are distributed as $\rho(\mathrm{G},\mathrm{D}) = 1.1 \pm 0.2$ (significant (<5%)), $\rho(\mathrm{G},\mathrm{B}) = 1.8 \pm 0.3$ (significant (<1%)).

The difference of computation time is also small in this case (less than 5% of time), which does not affect the statistical significance.

References

[1] S. Geman, D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. on PAMI*, 6, 721–741.

[2] W.R. Gilks, S. Richardson, D.J. Spiegelhalter (eds) (1996). *Markov Chain Monte Carlo in Pracetice.* Chapman & Hall.

[3] J.S. Liu (2001). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag.

[4] M. Opper, D. Saad (eds) (2001). *Advanced Mean Field Theory*, MIT Press.

[5] K. Takabatake (2004). Information Geometry of Gibbs Sampler. In *Proc. of WSEAS Int. Conference on Neural Networks and Applications*.