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

More Aggressive Updating than Gibbs Sampler

Kazuya Takabatake[†] and Shotaro Akaho[†]

†Neuroscience Research Institute
AIST Central 2, Umezono 1-1-1, Tsukuba 305-8568 Japan
Email: k.takabatake@aist.go.jp

Abstract—A Markov chain Monte-Carlo sampler in which components are more frequently updated than in Gibbs sampler is proposed. Proposed sampler and Gibbs sampler are compared by using a probabilistic reasoning problem. The sample mean drawn from the proposed sampler shows faster convergence to the true value than it of Gibbs sampler.

1. Introduction

The most straightforward way to draw samples from a given target distribution is using the joint probability table of the target distribution. However the size of this table is proportional to the exponential of the dimension of the random variable, therefore this straightforward way is impractical for large-dimensional random variables.

Markov chain Monte Carlo(MCMC)[5] is a class of random number generators which uses Markov chains whose distribution converges to the target distribution. Some MCMC effectively work in cases that components of multidimensional variable X have sparse Markov network[4].

Let X be a target variable [1], V be the range of value X takes, D_V be the set of distributions on V, π be X's distribution, $\{X^{(t)}\}(t = 0, 1, ...)$ be a Markov chain, and $p^{(t)}$ be the distribution of $X^{(t)}$.

Any distribution $p \in D_V$ can be represented as a vector whose x-th element is $p(x)$. Let $W^{(t)}$ be the transition [2] from $X^{(t-1)}$ to $X^{(t)}(p^{(t-1)}$ to $p^{(t)})$. $W^{(t)}$ can be represented as a matrix whose (x, y) element is $\Pr(X^{(t)} = y|X^{(t-1)} = x)$. Then

$$p^{(t)} = p^{(t-1)} W^{(t)} \tag{1}$$

therefore

$$p^{(t)} = p^{(0)} W^{(1)} ... W^{(t)} \tag{2}$$

holds.

The mission of MCMC is generating $X^{(t)}$ whose distribution $p^{(t)}$ converges to the target distribution π. In designing MCMC, the sequence $\{W^{(t)}\}(t = 1, 2, ...)$ is designed. Under some conditions, we can design such $\{W^{(t)}\}$ without knowing the complete table of the target distribution π.

One important point is that some limited finite length of samples which are drawn by MCMC sampler should have diversification. If the MCMC sampler draws samples from some limited area in V, the samples actually does not reflect the target variable X, in other words such samples are not the samples drawn from the target distribution π. One simple policy to keep diversification of samples is "Do not successively draw the same sample".

In this paper, we firstly review the Metropolis-Hastings algorithm[5], single-component-update MCMC and their special case of the Gibbs sampler[2, 5]. Then we propose new MCMC sampler which update components more aggressively than Gibbs sampler and show the comparison between the proposed sampler and Gibbs sampler using a probabilistic reasoning problem.

2. Review of MCMC around Gibbs Sampler

2.1. Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm[5] is the following algorithm.

step0 Prepare an arbitrary sequence of conditional distribution $\{q^{(t)}(X'|X)\}(t = 1, 2, ...)$, which is called proposal distribution, where X' is a candidate variable for next time. Set an arbitrary value to x. Set $t = 0$.

step1 Generate a random number x' according to $q^{(t)}(x'|x)$.

step2 Set $x = x'$ with probability

$$\alpha^{(t)}(x, x') = \min\left(1, \frac{\pi(x')q^{(t)}(x|x')}{\pi(x)q^{(t)}(x'|x)}\right), \tag{3}$$

which is called acceptance probability, otherwise keep x as it is.

step3 Set $t = t + 1$ and go to step1.

This algorithm simulates the Markov chain whose transition matrix is

$$W^{(t)}_{xy} = \begin{cases} q^{(t)}(y|x)\alpha^{(t)}(x, y) & x \neq y \\ 1 - \sum_{x \neq y} q^{(t)}(y|x)\alpha^{(t)}(x, y) & x = y \end{cases}. \tag{4}$$

This transition matrix suffices the following so-called detailed balance equation[5] for all x, y, t.

$$\pi_x W^{(t)}_{xy} = \pi_y W^{(t)}_{yx} \tag{5}$$

[1] For the sake of simplicity, all random variables in this paper take finite discrete values.

[2] Any transition is a linear operator $D_V \to D_V$.

Summing up eq.(5) about x, we get

$$\pi W^{(t)} = \pi \tag{6}$$

i.e. the transition by $W^{(t)}$ does not move π. It is known that if the Markov chain is weakly ergodic[1, 3] then the distribution $p^{(t)}$ converges to π when $t \to \infty$.

In designing a Metropolis-Hastings algorithm, it is important that x' is accepted frequently. If x' is rarely accepted, the Markov chain stays at the same x for long time therefore $X^{(t)}$ travels only limited area in V in finite time.

2.2. Single-component-update MCMC

Single-component-update MCMC is used in cases that X is multi-dimensional i.e. $X = (X_0, ..., X_{N-1})$. It is a MCMC in each transition only one component is updated. Assume that i-th component X_i is updated and $p^{(t-1)}$ change to $p^{(t)}$. X_{-i} denotes the joint variable of all components of X except for i-th component X_i [3].

Single-component-update MCMC's transition matrix has the following form

$$W^{(t)}_{(x_i,x_{-i})^{(t)}(y_i,y_{-i})} = \delta(x_{-i}, y_{-i}) d^{(t)}(y_i|x) \tag{7}$$

where d is some conditional distribution, δ is Kronecker delta:

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}. \tag{8}$$

X_{-i} does not change therefore the marginal distribution of X_{-i} does not change:

$$p^{(t)}(x_{-i}) = p^{(t-1)}(x_{-i}). \tag{9}$$

By the transition, only the distribution of X_i conditioned by X_{-i} changes:

$$p^{(t)}(y_i|y_{-i}) = \sum_{x_i} p^{(t-1)}(x_i|y_{-i}) d^{(t)}(y_i|x_i, y_{-i}). \tag{10}$$

Let $p^{(t)}_{|x_{-i}}$ be the vector whose x_i-th component is $p^{(t)}(x_i|x_{-i})$ and $W^{(t)}_{|x_{-i}}$ be the matrix, which we name *local transition matrix*, whose (x_i, y_i) component is $d^{(t)}(y_i|x_i, x_{-i})$ then we can rewrite eq.(10) in a vector form.

$$p^{(t)}_{|x_{-i}} = p^{(t-1)}_{|x_{-i}} W^{(t)}_{|x_{-i}} \tag{11}$$

To $p^{(t)}$ converges to π, $W^{(t)}$ should not move π i.e. $\pi W^{(t)} = \pi$. In single-component-update MCMC, it is equivalent to

$$\pi_{|x_{-i}} W^{(t)}_{|x_{-i}} = \pi_{|x_{-i}}. \tag{12}$$

One way to make $W^{(t)}$ which suffice the constraint above is adopting Metropolis-Hastings algorithm. It is called single-component Metropolis-Hastings[5]. In the case that $X = (X_0, ...X_{N-1})$ is a loosely coupled Markov network[4], X_i depends on only a few components among X_{-i} therefore we can easily get $\pi(x'_i|x_{-i})$ or $\pi(x_i|x_{-i})$ from local information. This is the major merit of the single-component Metropolis-Hastings.

There are several ways to select the component updated at time t[5]. Let $i(t)$ be the suffix of the component updated at time t. In this paper, we adopt sequential-update:

$$i(t) = t \mod N. \tag{13}$$

2.3. Gibbs Sampler

Gibbs sampler[2, 5] is a special case of Single-component Metropolis-Hastings. Its proposal distribution is

$$q^{(t)}(x'_i, x'_{-i}|x_i, x_{-i}) = \delta(x'_{-i}, x_{-i})\pi(x'_i|x_{-i}) \tag{14}$$

therefore Gibbs sampler's local transition matrix is

$$G_{|x_{-i}} = \begin{pmatrix} \pi_{|x_{-i}} \\ \vdots \\ \pi_{|x_{-i}} \end{pmatrix} \tag{15}$$

whose rows are the same. It is clear that this local transition suffices eq.(12) because it moves any distribution to $\pi_{|x_{-i}}$.

The acceptance probability is

$$\begin{aligned} \alpha^{(t)}((x_i, x_{-i}), (x'_i, x_{-i})) &= \min\left(1, \frac{\pi(x'_i, x_{-i})\pi(x_i|x_{-i})}{\pi(x_i, x_{-i})\pi(x'_i|x_{-i})}\right) \\ &= 1. \end{aligned} \tag{16}$$

The information about the target distribution π required to perform the Gibbs sampler is the full conditional distribution[5] $\pi(x_i|x_{-i})$ in eq.(14).

One interesting property of Gibbs sampler is that $X^{(t)}_i$ does not refer $X^{(t-1)}_i$. Actually this property specifies what Gibbs sampler is. Given π, the transition has the following properties is unique and it is Gibbs sampler.

- The transition updates only one component.
- $X^{(t)}_i$ does not refer $X^{(t-1)}_i$.
- The transition does not move π, i.e. $\pi W = \pi$

Because the second constraint means that the local transition matrix has the same row vectors v. From the third constraint it has to suffice eq.(12).

$$\pi_{|x_{-i}} \begin{pmatrix} v \\ \vdots \\ v \end{pmatrix} = v = \pi_{|x_{-i}}. \tag{17}$$

This local transition matrix is identical to Gibbs sampler's.

Besides mentioned above, Gibbs sampler is also specified as a greedy algorithm which minimize the Kullback-Leibler divergence $KL(p^{(t)}||\pi)$ in each transition[7].

[3] We use the same notation about the values X take. Readers are expected to interpret symbols for distributions with x_i or x_{-i} as appropriate conditional or marginal distributions. For example

$$\pi(x_i|x_{-i}) = \Pr(X_i = x_i|X_{-i} = x_{-i}).$$

3. More Aggressive Updating than Gibbs ampler

Eq.(16) shows that the candidate x' is always accepted in step2 in section 2.1 therefore the Gibbs sampler most aggressively updates the components among single-component-update Metropolis-Hastings. However if we do not limit ourselves in the frame of Metropolis-Hastings algorithm, we can design a MCMC which more aggressively updates the components than Gibbs sampler.

In Metropolis-Hastings algorithm, transition does not change the value of any component in the following two cases: "x' is rejected" or "$x' = x$".

The first case does not happen in Gibbs sampler however the second case might happen. In cases that the range of values each component takes is small, especially in the binary cases, the second case frequently happens.

In eq.(12), the diagonal elements in the local transition matrix $W^{(t)}_{|x_{-i}} = \pi_{|x_{-i}}$ are the probability i-th component does not change therefore if we reduce diagonal elements with keeping eq.(12) we get the local transition matrix with less staying at x_i.

One way to reduce diagonal elements of the local transition matrix W, which does not move p, is

$$W' = (1+\lambda)W - \lambda I \quad (\lambda \geq 0) \tag{18}$$

It is easy to show $pW' = p$ because both W and identical matrix I does not move p. The diagonal elements of W' are $(1+\lambda)W_{ii} - \lambda$. If we take λ larger we can reduce the diagonal elements of W' however any element of W has to be non-negative therefore

$$\lambda \leq \frac{w}{1-w} \tag{19}$$

where w is the minimum diagonal element of W.

Substituting Gibbs sampler's transition matrix to W in eq.(18), we get For each x_{-i}, we get

$$G'_{|x_{-i}} = (1+\lambda_{x_{-i}})G_{|x_{-i}} - \lambda_{x_{-i}} I \quad \lambda_{x_{-i}} = \frac{w_{x_{-i}}}{1-w_{x_{-i}}} \tag{20}$$

for each x_{-i}, where $l_{x_{-i}}$ is the smallest element of $\pi_{|x_{-i}}$. This is the local transition matrix we propose here.

4. Experiment

We show some comparison between Gibbs sampler and the proposed algorithm here. We applied those two algorithms to a probabilistic reasoning problem[4].

Figure 1 shows the Markov network[4] used in the probabilistic reasoning, where each X_i is binary. Each clique's local potential[2] was determined by assigning random numbers.

The reasoning task here is getting some X_i's distribution under the condition $X_{10} = 0$. This task is performed as follows.

- Fix the value of X_{10} to 0.

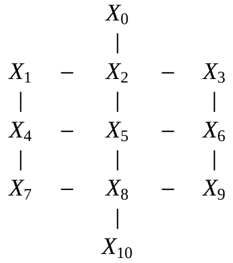

Figure 1: Markov network used in the experiment

- Sequentially and periodically make the local transition with $X_0, ..., X_9$ and get the sequence of samples $x^{(0)}, x^{(1)}, ...$

- Assume the random numbers above are drawn from the posterior distribution $P(X_i|X_{10} = 0)$. Then any expectation of $E_{P(X_i|X_{10}=0)}(f(X_i))$ is estimated by the sample mean

$$E_{P(X_i|X_{10}=0)}[f(X_i)] \sim \frac{1}{n}\sum_{t=1}^{n} f(x_i^{(t)}) \tag{21}$$

In the case to get $P(X_i = 1|X_{10} = 0) = E_{P(X_i|X_{10}=0)}[f(X_i)]$, f is just the identical function.

We took 10 traces of sample mean with 10 different random number seeds. All components were initialized to 0 at the beginning of each trace. Figure 2 shows the these traces. The bold lines shows the true value of $P(X_0 = 1|X_{10} = 0)$. As is shown, the proposed sampler shows faster convergence to the true value than Gibbs sampler. And we see the proposed sampler flips x_0 more frequently than Gibbs sampler. Gibbs sampler tends to keep x_0 unchanged.

Figure 3 shows the case that true value of $P(X_i = 1|X_{10} = 0)$ is close to 0. No improvement is seen in this case. This phenomenon comes from that the Gibbs sampler's local transition has a small diagonal element. We discuss this degradation in the next section.

5. Discussion

In Monte-Carlo integration, successive samples generated by the sampler are used for taking sample mean therefor diversification of samples seen in the limited length of successive samples is important. In some cases, diversification of samples is more important than independency among them[8]. In this paper, eq.(18) is used to reduce the diagonal elements of the local transition matrix. However this diagonal element reduction does not work in cases there is just one small diagonal element. Akaho[8] shows another diagonal element reduction method to improve this defect.

In the cases that Gibbs sampler's local transition matrix in eq.(15) has a small diagonal element $w_{x_{-i}}$, λ_{-i} is small

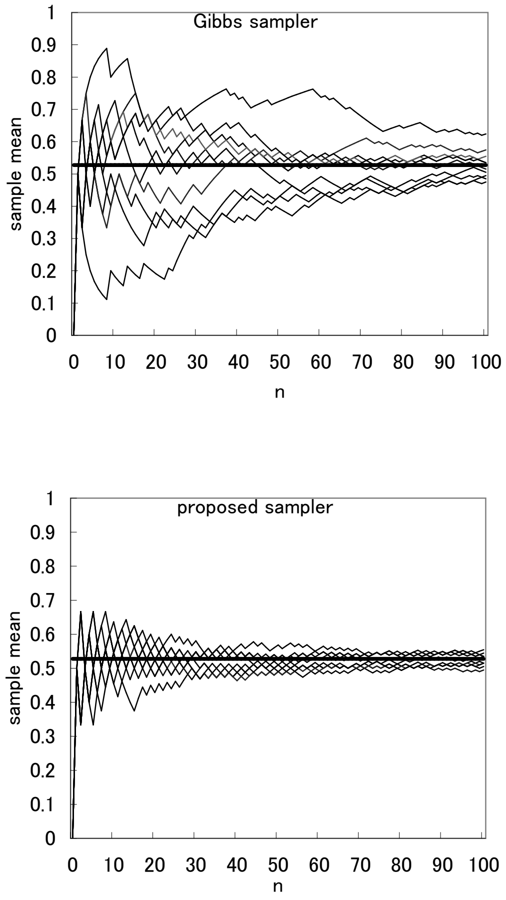

Figure 2: sample mean of x_0

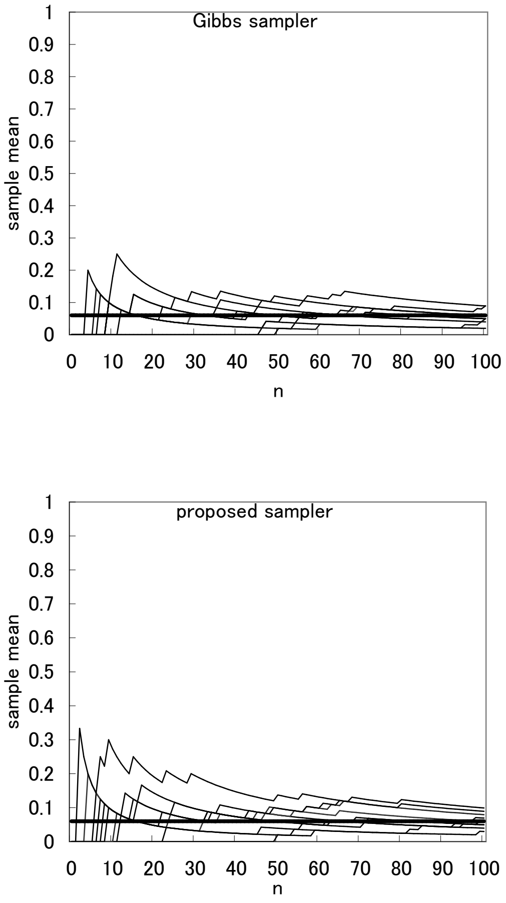

Figure 3: sample mean of x_7

and the proposed sampler's local transition matrix $G'_{|x_{-i}}$ in eq.(18) is almost same as Gibbs sampler's local transition matrix $G_{|x_{-i}}$. In such cases we can not expect the improvement. In the case X_i is binary, this degradation happens X_i has small entropy, in other words X_i takes almost always the same value. Therefore in probabilistic reasoning applications, if you have less interest in such less informative components X_i, this degradation does not become problematic.

Acknowledgement

This work is supported by JSPS.KAKENHI14658106.

References

[1] E. Seneta, "Non-negative Matrices and Markov Chains, Second Edition", Springer-Verlag, 1981.

[2] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 6, pp.721-741, 1984.

[3] P. J. M. van Laarhoven, E. H. L. Aarts, "Simulated Annealing: Theory and Applications", Kluwer Academic Publishers, 1987.

[4] J. Pearl, "Probabilistic Reasoning in Intelligent Systems", Morgan Kaufmann Publishers, 1988.

[5] W. R. Gilks, S. Richardson, D. J. Spiegelhalter, "Introducing Markov chain Monte Carlo", In: W. R. Gilks, S. Richardson, D. J. Spiegelhalter(ed.), "Markov Chain Monte Carlo in Practice", Chapman & Hall, pp.1-19, 1996.

[6] W. R. Gilks, "Full conditional distributions", In: W. R. Gilks, S. Richardson, D. J. Spiegelhalter(ed.), "Markov Chain Monte Carlo in Practice", Chapman & Hall, pp.1-19, 1996.

[7] K. Takabatake, "Information Geometry of Gibbs Sampler", *WSEAS Transactions on Systems*, Issue 2, Vol.3, pp.449-455, 2004.

[8] S. Akaho, K. Takabatake, "Active Update in single Component MCMC for Discrete Variables", to appear in NOLTA2004