

Reversible Markov Chain in the Multimodal Problem of the Gibbs Sampler

Guang lu Gong

Tsinghua University

Department of Applied Mathematics

Min ping Qian and Jun Xie

Peking University

Department of Probability and Statistics

Abstract

Markov chain Monte Carlo methods such as the Gibbs sampler enable the fitting of models of virtually unlimited complexity, and as such have revolutionized the practice of Bayesian data analysis. However those schedules fail in some problems suffering from convergence difficulties of the Markov chain. In this paper, for a distribution $\pi(X)$ which has several rather deep local probability traps, we construct a reversible Markov chain and obtain the distribution ratios among the different traps. This method overcomes the convergence difficulties of being trapped in one of the local minimum and generates more accurate samples of $\pi(X)$. We illustrate our approach with some examples.¹

1 Introduction

The Gibbs sampler is a Markov chain Monte Carlo (MCMC) method to treat complicated models. Through this approach we can form sample-based estimations, predictions, diagnosis *etc*, so as to resolve intractable models of non-linear and high dimensional. In this article, we introduce a reversible Markov

¹AMS 1991 subject classifications:

Key words and phrases: Gibbs sampler; Metastability; Multimodality; Reversible Markov chain; Stochastic Ising Model.

chain algorithm to deal with the convergence difficulty of the multimodal problems in the Gibbs sampler, for which the convergence to the invariant measure, the distribution of samples required, needs an unreasonably long time, since there is very little chance for the Markov chain to go from one deep trap to the others.

We introduce the Gibbs sampler algorithm and discuss the convergence problem first. Assume that we focus on a rather complex distribution such as the posterior of Bayes. Let $\pi(X) = \pi(x_1, \dots, x_m)$, $X \in \mathcal{X} \subseteq R^m$ be a joint density. To induce the features of this distribution, we can construct irreducible Markov chains, which have $\pi(X)$ as their equilibrium distribution. Suppose $\pi(x_i|x_j, j \neq i)$ denotes the induced full conditional density for each of the components x_i , given values of the other components. Then a Markov chain is defined by the transition probability

$$P(X, Y) = \prod_{l=1}^m \pi(y_l|x_j, j > l, y_j, j < l), \quad X, Y \in \mathcal{X}.$$

Such a transition probability allows one generating an 1-dimension random variable at a time, successively for m times, to transit from one m -dimensional random vector to another at next time. The iteration of such random vectors produces a sequence of m -dimensional random vectors $X^{(0)}, X^{(1)}, \dots, X^{(t)}, \dots$ which is then a realization path of the Markov chain X with transition probability $P(X, Y)$.

Under suitable regular conditions, we have the asymptotic results for the Markov chain

$$\begin{aligned} X^{(t)} &\xrightarrow[t \rightarrow \infty]{d} X \sim \pi(X); \\ \frac{1}{t} \sum_{i=1}^t f(X^{(i)}) &\xrightarrow[t \rightarrow \infty]{} E_{\pi}\{f(X)\}, \text{ almost surely,} \end{aligned} \tag{1}$$

where f is a π -integrable function. For a large time t , $X^{(t)}$ can be considered as a random sample from $\pi(X)$ and the average of $f(X^{(i)})$ over a Markov chain path approximates the expectation of the function in the state space. However, in practice, how can we decide whether t has been already sufficiently long? The convergence of the Gibbs sample has been discussed considerably, see for example Gelman and Rubin (1992), Geweke (1992), Schervish and Carlin (1992), Roberts and Polson (1994), Tierney (1994), Diaconis (1995), Liu *et al.* (1995) and Rosenthal (1995). But useful bounds on rates of convergence are still difficult to be found, especially in the multimodal problems.

For a simple one dimensional discrete probability distribution, with two humps, Figure 1 shows its shape. Figure 2 is obtained from the Markov chain of the stochastic Ising model. It demonstrates two chains' transition behavior starting from different values. In a single run, the chain would be trapped at one of the modes for a very long time. It is shown that much higher order of time is needed for going ergodicity between two modes than going "local ergodicity" in one mode (eg. see Chen *et al.* (1995)). Besag and Green (1993) use an auxiliary variate to treat the difficulties of multimodality. Here we present another way by using reversible Markov chain method. The idea is, first, to symmetrize the transition of the Gibbs sampling such that the Markov chain becomes reversible; second, to stop when a run achieves local ergodicity in each trap; third, to evaluate the ratios of the distribution among different modes by simple simulations; finally, to reject the samples by the proportions above to get the samples in the target distribution.

In section 2, we introduce our symmetrizing method of Markov chains with the equilibrium distribution $\pi(X)$, then estimate the ratios between modes. In section 3, we give two numerical simulation examples to demonstrate the effectiveness of this scheme. In section 4, for the two dimensional Ising model, which is a typical example of the multimodality, a numerical example is shown that our algorithm can achieve the pretty ratios between variant modes. In the final section, we present the theoretical background of the algorithm.

2 Reversible Markov Chain

In the practice of the MCMC algorithms, running several parallel chains, if we find their outputs are not consistent, then we have to be careful with the inferences of the Gibbs sampling. It usually means that the equilibrium distribution $\pi(X)$ has humps. Suppose that the state space can be divided into two subsets B_1 and B_2 , with "centers" widely separated and hard to be transferd each other by the Markov chain (the multi-humps can be done similarly). We propose our scheme below.

2.1 \diamond Restriction in One of the Modes For Reversible Chains

Let us show the conclusion explicitly in the case of discrete state space for simplicity. A Markov chain with equilibrium distribution $\pi(X)$ is called reversible

if its transition probability ($P(X, Y)$) satisfies the following equation

$$\pi(X) P(X, Y) = \pi(Y) P(Y, X) \quad X, Y \in \mathcal{X}. \quad (2)$$

For a reversible Markov chain, we can restrict it in a subset B_i of the state space, with $\pi(X)$ still being an invariant measure on B_i . In fact, let

$$P_i(X, Y) = \begin{cases} P(X, Y) , & Y \neq X , X, Y \in B_i, \\ P(X, X) + \sum_{Y \notin B_i} P(X, Y) , & Y = X , X \in B_i, \end{cases} \quad (3)$$

we still have

$$\pi(X) P_i(X, Y) = \pi(Y) P_i(Y, X), \quad X, Y \in B_i.$$

Define

$$\pi_i(X) = \frac{\pi(X)}{\pi(B_i)}, \quad X \in B_i,$$

where

$$\pi(B_i) = \sum_{X \in B_i} \pi(X).$$

Then $\pi_i(X)$ is a probability invariant distribution of $(P_i(X, Y))$, $X, Y \in B_i$.

In the 2-mode problems we discussed, the transition probabilities from B_1 to B_2 and B_2 to B_1 are very low and the chain is often trapped at B_1 or B_2 . So starting at a point in B_i before the chain leaving B_i , it is the same as we run the chain by $(P_i(X, Y))$, $i = 1, 2$. When the Markov chain looks like being stationary in B_1 or B_2 (in fact, we do not know B_i 's in advance, and practically "being stationary" means that the chain stably oscillating at some area) then the samples are just considered approximately drawing from $\pi_i(X)$ ($i = 1, 2$) respectively. Thus we obtain the samples of $\pi(X)$ except a constant ratio. A further discussion shows that Markov chains of $(P_i(X, Y))$ ($i = 1, 2, \dots$) go to stationary much faster than that of $(P(X, Y))$.

2.2 Construct a Reversible Markov Chain

For an arbitrary transition probability ($P(X, Y)$), define

$$P^*(X, Y) = \frac{1}{2} (P(X, Y) + P^-(X, Y)),$$

where

$$P^-(X, Y) = \frac{\pi(Y) P(Y, X)}{\pi(X)},$$

then we obtain the reversibility

$$\pi(X) P^*(X, Y) = \pi(Y) P^*(Y, X).$$

However, we expect that $(P^*(X, Y))$ keeps the advantage of the Gibbs sampler i.e. generating only one entry at a time. Fortunately $P^-(X, Y)$ is also a transition of the type of Gibbs sampling, which generates a 1-dimension random variable at a time and is only different from $(P(X, Y))$ by the order of changing variables. This can be seen by Lemma 1.

Lemma 1. $P^-(X, Y) = \prod_{l=1}^m \pi(y_l | y_j, j > l; x_j, j < l).$

Proof.

$$\begin{aligned} & \frac{\pi(Y) p(Y, X)}{\pi(X)} \\ &= \frac{\pi(y_1, \dots, y_m) \pi(x_1 | y_2, \dots, y_m) \pi(x_2 | x_1, y_3, \dots, y_m) \cdots \pi(x_m | x_1, \dots, x_{m-1})}{\pi(x_1, \dots, x_m)} \\ &= \frac{\pi(x_m | x_1, \dots, x_{m-1}) \pi(x_{m-1} | x_1, \dots, x_{m-2}, y_m) \cdots \pi(x_1 | y_2, \dots, y_m) \pi(y_1, \dots, y_m)}{\pi(x_1, \dots, x_m)} \\ &= \frac{\pi(x_1, \dots, x_m) \pi(x_1, \dots, x_{m-1}, y_m) \cdots \pi(x_1, y_2, \dots, y_m) \pi(y_1, \dots, y_m)}{\pi(x_1, \dots, x_m) \pi(x_1, \dots, x_{m-1}) \pi(x_1, \dots, x_{m-2}, y_m) \cdots \pi(y_2, \dots, y_m)} \\ &= \pi(y_m | x_1, \dots, x_{m-1}) \pi(y_{m-1} | x_1, \dots, x_{m-2}, y_m) \cdots \pi(y_1 | y_2, \dots, y_m). \end{aligned}$$

Lemma 1 tells that

$$P^*(X, Y) = \frac{1}{2} \prod_{l=1}^m \pi(y_l | x_j, j > l, y_j, j < l) + \frac{1}{2} \prod_{l=1}^m \pi(y_l | y_j, j > l, x_j, j < l). \quad (4)$$

For simulating, this is almost the same as original Gibbs sampling, and it just need to toss a coin to determine the order of generating 1-dimension random variables. If the coin has its face up, then from $X = (x_1, \dots, x_m)$, we generate the next vector $Y = (y_1, \dots, y_m)$ by the order y_1, \dots, y_m . Otherwise we generate Y by the reverse order y_m, y_{m-1}, \dots, y_1 .

2.3 Obtain the Ratios of Two Modes

Let

$$\frac{\pi_i(X)}{\pi(X)} = c_i, \quad \forall X \in B_i, i = 1, 2,$$

then we have

$$\pi(B_i) = \frac{1}{c_i} \text{ and}$$

$$\sum_i \pi(B_i) = \sum_i \frac{1}{c_i} = 1.$$

By the reversibility, using formula (2), in the case of $P(X, \tilde{X}) > 0$, we have

$$\frac{\pi(X)}{\pi(\tilde{X})} = \frac{P(\tilde{X}, X)}{P(X, \tilde{X})}, \text{ for } X \in B_1, \tilde{X} \in B_2,$$

which implies

$$\frac{c_1}{c_2} = \frac{\pi_1(X)}{\pi(X)} \bigg/ \frac{\pi_2(\tilde{X})}{\pi(\tilde{X})} = \frac{\pi_1(X) P(X, \tilde{X})}{\pi_2(\tilde{X}) P(\tilde{X}, X)}. \quad (5)$$

While in the case of $P(X, \tilde{X}) = 0$, letting $X = Y_1, Y_2, \dots, Y_k = \tilde{X}$ be a path from X to \tilde{X} , we also have

$$\pi(X) \prod_{l=1}^{k-1} P(Y_l, Y_{l+1}) = \pi(\tilde{X}) \prod_{l=1}^{k-1} P(Y_{l+1}, Y_l),$$

which leads to

$$\frac{c_1}{c_2} = \frac{\pi_1(X) \prod_{l=1}^{k-1} P(Y_l, Y_{l+1})}{\pi_2(\tilde{X}) \prod_{l=1}^{k-1} P(Y_{l+1}, Y_l)}. \quad (6)$$

We have known the transition probability $P(X, \tilde{X})$, so the ratio c_1/c_2 can be determined by working out the values of $\pi_i(\cdot)$.

In practical calculation, the estimation $\hat{\pi}_i(X)$ of $\pi_i(X)$ can be obtained by a simple simulation of random variable with the distribution density $P_i(X, \cdot)$ for fixed X . The following Lemma 2 explains the reason for this.

Lemma 2. *In the continuous case, suppose that the restricted transition density in a basin B_i satisfies $P_i(X, Y) > 0, \forall X, Y \in B_i$, then*

$$\pi_i(X) = \left(E^{P_i} \left(\frac{1}{P_i(\xi, X)} \right) \right)^{-1},$$

where E^{P_i} means the expectation about the distribution $P_i(X, \cdot)$. Furthermore suppose ξ^1, \dots, ξ^n is an i.i.d. sequence with this distribution then $\pi_i(X)$ has the estimation

$$\hat{\pi}_i(X) = \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{P_i(\xi^j, X)} \right]^{-1}. \quad (7)$$

Proof. Recalling the formula (2)

$$\pi_i(X) P_i(X, Y) = \pi_i(Y) P_i(Y, X), \quad X, Y \in B_i,$$

we have

$$\pi_i(Y) = \frac{\pi_i(X)}{P_i(Y, X)} P_i(X, Y),$$

since $P_i(X, Y) > 0$. Then

$$1 = \int_{B_i} \pi_i(Y) dY = \int_{B_i} \frac{\pi_i(X)}{P_i(Y, X)} P_i(X, Y) dY = \pi_i(X) \int_{B_i} \frac{1}{P_i(Y, X)} P_i(X, Y) dY,$$

and

$$\begin{aligned} \pi_i(X) &= \left[\int_{B_i} \frac{1}{P_i(Y, X)} P_i(X, Y) dY \right]^{-1} \\ &= \left(E^{P_i} \left(\frac{1}{P_i(\xi, X)} \right) \right)^{-1}. \end{aligned}$$

By large number law, $\pi_i(X)$ can be estimated by averaging $1/P_i(Y, X)$ over the sample values of Y , i.e. we have the estimation (7).

Picking “good” points $X \in B_1$ and $\tilde{X} \in B_2$, which means *that...* we use (7) to calculate the value $\hat{\pi}_i(X)$, then the ratio is approximately given by

$$\left(\frac{c_1}{c_2} \right)^\wedge = \frac{\hat{\pi}_1(X) P(X, \tilde{X})}{\hat{\pi}_2(\tilde{X}) P(\tilde{X}, X)}. \quad (8)$$

2.4 Basic Algorithm

Now we list the algorithm of the reversible Markov chain Gibbs sampler:

Step 1. Choose a number of initials from the state space. The initials may be uniformly distributed or with distribution according to some known conditions.

Step 2. Run several Markov chains beginning with these initials and make decision by exploratory data analysis (EDA) of a point cloud of multivariate observations. We can use the exploratory, particularly graphical, toolkit(((?))) developed by the multivariate EDA and visual EDA communities. If the results are consistent among different chains, then the ordinary Gibbs sampling method works. Otherwise we find some humps in the state space and continue the following steps.

Step 3. Once there exist humps, partition the state space into several local basins B_1, B_2, \dots, B_k . In fact the different areas of sample paths that occur in Step 2 have constructed the basin $B_i, i = 1, 2, \dots, k$.

Step 4. In each basin B_i , following (4) we construct reversible Markov chains. Then judge stability in the local basin and obtain the samples in it.

Step 5. Pick one point X_i which has the relatively large frequency in each basin B_i and calculate the value $\hat{\pi}_i(X_i)$ by (7). Then use (8) to estimate the ratio $\left(\frac{c_1}{c_i}\right)^\wedge$ of the two basins B_1 and $B_i, i = 2, \dots, k$. Furthermore by

$$\sum_{i=1}^k \frac{1}{c_i} = 1,$$

we have the asymptotic value $\hat{c}_1, \dots, \hat{c}_k$.

Step 6. Accept the samples from different local basins according to the proportions of

$$\left(\frac{1}{\hat{c}_1}, \frac{1}{\hat{c}_2}, \dots, \frac{1}{\hat{c}_k} \right)$$

to obtain the samples of the whole distribution $\pi(X)$. An estimation of the value $E_\pi\{f(X)\}$ can also be given by the mixture average of the samples.

Remark: In practice, for the Bayes posterior distribution $\pi(X)$, the main modes will often be induced separately by the prior and by the likelihood and it will then be possible to locate them by deterministic hill climbing from the knowledge of the posterior distribution. Once we have located the modes our algorithm works better.

3 Examples of mixtured normal distributions

In this section, we provide two numerical examples to demonstrate the reversible Markov chains in the Gibbs sampler in more detail.

3.1 Two modes problem in the state space \mathbf{R}^2

Suppose a surface formed by a mixture of two bivariate normals with locations, widely separated relative to the spreads. Denote the global distribution by $\pi(x, y)$. We do not need to know its explicit form, but assume that the conditional distributions (or the generation scheme of the Markov chain) are available. Picking starting value (x^0, y^0) , we draw the successive random values as below:

$$\begin{aligned}x^1 &\sim \pi_1(x|y^0), \\y^1 &\sim \pi_2(y|x^1).\end{aligned}$$

Here $\pi_1(x|y)$ and $\pi_2(y|x)$ are both 1-dimension mixtured normal distributions, of which random numbers can be easily generated. In fact, the conditional distributions are given as follows

$$\begin{aligned}\pi_1(x|y) &= \frac{e^{-\frac{y^2}{4}}}{e^{-\frac{y^2}{4}} + 3e^{-\frac{(y-20)^2}{4}}} N(20, 2) + \frac{3e^{-\frac{(y-20)^2}{4}}}{e^{-\frac{y^2}{4}} + 3e^{-\frac{(y-20)^2}{4}}} N(0.5 \times (y - 20), 1.5), \\ \pi_2(y|x) &= \frac{e^{-\frac{(x-20)^2}{4}}}{e^{-\frac{(x-20)^2}{4}} + 3e^{-\frac{x^2}{4}}} N(0, 2) + \frac{3e^{-\frac{x^2}{4}}}{e^{-\frac{(x-20)^2}{4}} + 3e^{-\frac{x^2}{4}}} N(20 + 0.5 \times x, 1.5),\end{aligned}\tag{9}$$

where $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2 .

Choosing initial points from uniform distribution in \mathbf{R}^2 , we obtain two basins by the Markov chains. Figure 3 and 4 show the sample points of two Markov chain paths, which form the two basins respectively. In Figure 5, from $(x^0, y^0) = (20, 5)$ and $(x^0, y^0) = (25, 2)$ the changing of the samples deviation has the same trend. We judge that after a few thousands steps the chains get stationary in a local area. The methods of Geweke or those of Gelman and Rubin can also be used here to judge stationary in the local basins. However, the serious problem is that the chain can not move from one basin to the other in the finite time of our simulations.

From samples of our case, the distribution in one local basin is easy to estimate. Using the histogram to overview the distribution in one basin, we obtain *normal* asymptotically. Then the means and covariance matrices are estimated through the samples by the usual way. The results are listed below.

	Running steps	Initials	$\hat{\mu}$	$\hat{\Sigma}$
1	2107	x=20 y=5	20.00 -0.05	2.02, 0.09 0.09, 1.97
2	2025	x=5 y=15	-0.06 19.90	2.01, 1.05 1.05, 2.08

where $\hat{\mu}$ is the mean and $\hat{\Sigma}$ is the covariance matrix based on samples.

Let

$$P^* ((x^0, y^0), (x^1, y^1)) = \frac{1}{2} \pi_1 (x^1 | y^0) \pi_2 (y^1 | x^1) + \frac{1}{2} \pi_2 (y^1 | x^0) \pi_1 (x^1 | y^1),$$

and let $\hat{\pi}_1 (\xi)$, $\hat{\pi}_2 (\eta)$ denote the sample-based local distribution starting at points $\xi = (\xi_x, \xi_y) = (20.00, -0.05)$ and $\eta = (\eta_x, \eta_y) = (-0.06, 19.90)$, then

$$\frac{P^* (\xi, \eta)}{P^* (\eta, \xi)} = \frac{\pi_1 (\eta_x | \xi_y) \pi_2 (\eta_y | \eta_x) + \pi_2 (\eta_y | \xi_x) \pi_1 (\eta_x | \eta_y)}{\pi_1 (\xi_x | \eta_y) \pi_2 (\xi_y | \xi_x) + \pi_2 (\xi_y | \eta_x) \pi_1 (\xi_x | \xi_y)} \approx 3.4585.$$

Furthermore

$$\frac{\hat{\pi}_1 (\xi)}{\hat{\pi}_2 (\eta)} = \frac{\frac{1}{2\pi|\hat{\Sigma}_1|^{\frac{1}{2}}}}{\frac{1}{2\pi|\hat{\Sigma}_2|^{\frac{1}{2}}}} = \frac{|\hat{\Sigma}_2|^{\frac{1}{2}}}{|\hat{\Sigma}_1|^{\frac{1}{2}}} \approx 0.8804,$$

hence by (8)

$$\left(\frac{c_1}{c_2} \right)^\wedge = \frac{\hat{\pi}_1 (\xi) P^* (\xi, \eta)}{\hat{\pi}_2 (\eta) P^* (\eta, \xi)} = 3.04$$

Figure 6 shows the mixture samples from the two chains with the ratio $c_1 / c_2 = 3.00$ (????)

In fact the real invariant distribution of the Markov chain is

$$\pi(x, y) \propto N\left(\begin{pmatrix} 20 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right) + 3N\left(\begin{pmatrix} 0 \\ 20 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right).$$

We can see the accepted samples are consistent with the equilibrium distribution well. (((better to give a table to show how good this samples fit the distribution))))))

3.2 High-dimensional Surface of Three Modes

Assuming $X = (x_1, \dots, x_m)$, $m = 50$, is a multivariate and $\pi(X)$ is a 50-dimensional density. A Markov chain with the equilibrium distribution $\pi(X)$ is constructed by the transition probability:

$$P(X, Y) = \pi(y_1|x_2, \dots, x_m) \pi(y_2|y_1, x_3, \dots, x_m) \dots \pi(y_m|y_1, \dots, y_{m-1}).$$

In the special case, $\pi(X)$ has humps and the Markov chain ($P(X, Y)$)'s 1-dimension full conditional distribution $\pi(x_i|x_j, j \neq i)$ is a mixture of three normal distributions. Similar with formula (9)((*where???*))), we have known

$$\pi(x_i|x_j, j \neq i) = \alpha_1 N(\mu_1, \Sigma_1) + \alpha_2 N(\mu_2, \Sigma_2) + \alpha_3 N(\mu_3, \Sigma_3).((\text{???}))$$

For each $l = 1, 2, 3$, μ_l and Σ_l are respectively mean vector and covariance matrix related with $x_j, j \neq i$. At the same time, $\alpha_1, \alpha_2, \alpha_3$ are the coefficients determined both by x_j and the distribution ratios of the state space.

Analyzing the transition function, we assume that there exist three basins and we distribute many initials in the state space for the Markov chains. Using the first two entries to show the sample points, Figure 7, 8 and 9 demonstrate three chains from different starting points. We can see that there are three local areas B_1, B_2 and B_3 and in each basin the Markov chain converges fast (seen from a point cloud) but it is also trapped in the basin. Partial estimated values from the Markov chains are listed below.

	1	2	3
Running steps	30000	30000	30000
Initials	x_1, x_2 20.32, 2.17	x_1, x_2 19.00, 19.00	x_1, x_2 0.00, -19.00
Estimated means	0.002, 0.003	20.58, 19.58	-0.008, -20.009
True means	0.00, 0.00	20.50, 19.50	0.00, -20.00
Estimated covariances	2.00, 0.01 0.01, 1.98	1.56, 0.83 0.83, 1.55	0.98, 0.00 0.00, 1.99
True covariances	2.00, 0.00 0.00, 2.00	2.00, 1.00 1.00, 2.00	1.00, 0.00 0.00, 2.00

According to Step 5 of the algorithm, we choose the estimated mean $\hat{\mu}_i$ in each basin B_i to calculate the value

$$\hat{\pi}_i(\hat{\mu}_i) = \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{P^*(\xi^j, \hat{\mu}_i)} \right]^{-1}, \quad i = 1, 2, 3,$$

where $n = 32767$ in the simulation and

$$P^*(X, Y) = \frac{1}{2}\pi(y_1|x_2, \dots, x_m)\pi(y_2|y_1, x_3, \dots, x_m)\dots\pi(y_m|y_1, \dots, y_{m-1}) + \frac{1}{2}\pi(y_m|x_1, \dots, x_{m-1})\pi(y_{m-1}|x_1, \dots, x_{m-2}, y_m)\dots\pi(y_1|y_2, \dots, y_{m-1}).$$

Furthermore, we calculate $P^*(\hat{\mu}_1, \hat{\mu}_2)$ through its logarithmic value, since it is too small. By (8) we obtain that

$$\left(\frac{c_1}{c_2}\right)^\wedge = 2.7 \text{ and } \left(\frac{c_1}{c_3}\right)^\wedge = 6.(((\text{how could be exactly 6 ???})))$$

In fact the joint distribution is

$$\pi(X) \propto N_{50}(\mu^1, \Sigma^1) + 3N_{50}(\mu^2, \Sigma^2) + 6N_{50}(\mu^3, \Sigma^3),$$

where $N_{50}(\mu, \Sigma)$ is the 50-dimension normal distribution. The estimated ratios are the asymptotic of the true value $c_1/c_2 = 3$ and $c_1/c_3 = 6$.

4 The Two Dimensional Ising Model

Through reversible Markov chain, we get a scheme to treat the multimodal problems. Another example appears in some dynamical systems where the metastable behavior is observed and the system will stay in one of the metastable sets for a long time, just like what discussed above, and the Markov chain is then restricted in a local trap, which prevents us from getting global informations. A clear picture is provided by studying the metastable behavior of the two dimensional stochastic Ising model. In this section, we will use the scheme introduced in Section 2 to simulate the two dimensional Ising model and analyze the output.

The two dimensional Ising model can be set up as follows. Let Λ be the two dimensional lattice torus, namely $\{1, \dots, N\} \times \{1, \dots, N\}$ with periodic boundary condition. Associate to each site $x \in \Lambda$ either a +1 spin or a -1 spin, denoted by $\eta(x)$. The collection $\{\eta(x)\}_{x \in \Lambda}$ is called a configuration and denoted by η . Define

$$\begin{aligned} +\underline{1} &= \{\eta(x) = +1, \text{ for all } x \in \Lambda\}, \\ -\underline{1} &= \{\eta(x) = -1, \text{ for all } x \in \Lambda\}. \end{aligned}$$

Let S be the set of all configuration η 's, i.e. $S = \{-1, +1\}^\Lambda$. For each configuration $\eta \in S$, assign the Hamiltonian

$$H(\eta) = -\frac{1}{2} \sum_{\langle x, y \rangle} \eta(x) \eta(y) - \frac{h}{2} \sum_{x \in \Lambda} \eta(x) - \frac{h}{2} N^2 + N^2,$$

where the first sum runs over the pairs of neighboring sets of Λ by counting $\langle x, y \rangle$ and $\langle y, x \rangle$ as the same. We can see that $H(-\underline{1}) = 0$.

The discrete time stochastic Ising model is defined as the Markov chain on $S = \{-1, +1\}^\Lambda$ with the transition probability

$$P^\beta(\xi, \eta) = \begin{cases} N^{-2} [1 + \exp(-\beta(H(\xi) - H(\eta)))]^{-1}, & \text{if } \eta = \xi^x, x \in \Lambda, \\ 0, & \text{if } \eta \neq \xi^x, \forall x \in \Lambda, \eta \neq \xi, \\ N^{-2} \sum_{x \in \Lambda} [1 + \exp(\beta(H(\xi) - H(\xi^x)))]^{-1}, & \text{if } \eta = \xi, \end{cases}$$

where ξ^x is defined by ξ by changing the spin $\xi(x)$ at x only, namely

$$\xi^x(y) = \begin{cases} \xi(y), & \text{if } y \neq x, \\ -\xi(y), & \text{if } y = x. \end{cases}$$

Here β is a parameter which is interpreted as the inverse temperature in physics and h is the strength of the external field. We denote the stochastic Ising model exclusively by $\{X_n, n \geq 0\}$.

This Markov chain $\{X_n, n \geq 0\}$ is of high dimensional, in fact every state η is determined by the spins in $N \times N$ sites. But the transition probability $(P^\beta(\xi, \eta))_{(\xi, \eta)}$ implies that only one site transfers each time and the Markov chain is reversible. This Markov chain is also positive recurrent since the state space is finite and all states are connected. But there exist basins in the state space which trap the Markov chain for a long time. The Ising model is a typical multimodal problem, which is analyzed in Chen *et al.* (1995). The state space S has a hierarchic structure and for a fixed level we have a clear picture about the *attractors* and their *basins*. That is we can decide how many modes there exist and where they are in the state space. Hence we restrict a Markov chain in each basin to generate the samples then estimate the distribution ratio between two basins.

In a simple case we assume $\beta = 2.0$, $N = 5$ and $h = 0$. Giving the initial states from the uniform distribution in the state space, we construct many parallel chains. It shows that starting from an arbitrary configuration the chain falls into $+\underline{1}$ or $-\underline{1}$ quickly and then transfers around these two states. We can see the transition behavior of the Markov chains through the shape of

the configurations.

Running steps: 30000

initial state

+ + - - +
+ - - + -
- - - - +
+ + + + +
- - + + -

Harmiltonian: 28.00(((???)))

↓

+ + + + -
+ + + + +
+ + + + +
+ + + + +
+ + + + +

Harmiltonian: 4.00

staying steps: 81

↓

+ + + + +
+ + + + +
+ + + + +
+ + + + +
+ + + + +

Harmiltonian: 0

staying steps: 29644

Running steps: 30000
initial state

+	+	-	-	-	
-	+	-	-	-	Hamiltonian: 22.00
-	-	+	-	+	
-	+	+	+	+	
+	+	+	-	+	
		↓			
-	-	-	-	-	
-	-	-	-	-	
-	-	-	-	-	Hamiltonian: 4.00 staying steps: 12
-	-	-	-	+	
-	-	-	-	-	
		↓			
-	-	-	-	-	
-	-	-	-	-	
-	-	-	-	-	Hamiltonian: 0 staying steps: 28602
-	-	-	-	-	
-	-	-	-	-	

Figure 10 and 11 show the changing of the Hamiltonian of two Markov chains which fall into $+1$ or -1 's basin. They appear stationary.

We obtain two basins "centered" on $+1$ or -1 from the simulation. Within $n = 30000$ steps the chain doesn't exit from its basin. Hence we can only generate the samples in those local areas. If we run the chain for more times, for one real path, after $n = 1759052$ steps the states $+1$ and -1 are connected. But the staying time around the configuration $+1$ is 282381 and the time around -1 is 1153683, which shows the wrong ratio of the two states. The results are shown in the table below.

Steps	30000	30000	1759052
Staying at $+1$	29644	0	282381
Staying at -1	0	28602	1153683
Ratio	—	—	0.245

Then we follow the algorithm in Section 2. Assume B_1 and B_2 are the basins around the configurations $+1$ and -1 respectively, which are constructed from

the simulation. By formula (6)

$$\frac{c_1}{c_2} = \frac{\pi_1(+\underline{1}) \prod_{l=0}^{k-1} P^\beta(\xi_l, \xi_{l+1})}{\pi_2(-\underline{1}) \prod_{l=0}^{k-1} P^\beta(\xi_{l+1}, \xi_l)},$$

where $+\underline{1} = \xi_0, \xi_1, \dots, \xi_k = -\underline{1}$ is a path from $+\underline{1}$ to $-\underline{1}$. The form of $(P^\beta(\xi, \eta))_{(\xi, \eta)}$ implies

$$\frac{c_1}{c_2} = \frac{\pi_1(+\underline{1})}{\pi_2(-\underline{1})} e^{\beta(H(+\underline{1}) - H(-\underline{1}))}.$$

The estimation $\hat{\pi}_1(+\underline{1})$ and $\hat{\pi}_2(-\underline{1})$ are calculated by the frequency, which leads to

$$\frac{\hat{\pi}_1(+\underline{1})}{\hat{\pi}_2(-\underline{1})} = 1.03.$$

Furthermore we know $H(-\underline{1}) = H(+\underline{1}) = 0$, hence the estimated value

$$\left(\frac{c_1}{c_2}\right)^\wedge = \frac{\hat{\pi}_1(+\underline{1})}{\hat{\pi}_2(-\underline{1})} = 1.03.$$

It means that the ratio of accepting the samples from the two basins B_1 and B_2 is almost 1. The conclusion is in consist with the theoretic results. In fact, the invariant distribution of the Ising model is the Gibbs distribution

$$\pi(\xi) = \frac{e^{-\beta H(\xi)}}{Z_\beta},$$

where $Z_\beta = \sum_{\eta \in S} e^{-\beta H(\eta)}$ is the partition function. It is easy to learn that $+\underline{1}$ and $-\underline{1}$ are the points with the lowest Hamiltonian, that is they have the highest probability. Since we choose $h = 0$, $+\underline{1}$ and $-\underline{1}$ and their respective basins B_1 and B_2 are symmetric. The ratio of B_1 and B_2 in the state space is just 1.

5 Theoretical Background

(((((maybe cancel this part!!!!))))))

In this section, we add some facts related to the algorithm motivated and outlined in Section 2. They are, in multimodal problems, (a) the Markov chain restricted in a good local basin has geometric convergence rate and (b) the estimated ratio between every two basins converges to its real value.

The following two lemmas are referred to Roberts and Polson (1994).

Lemma 3. *Suppose that all states are communicated each other. Then the finite state space Gibbs sampler algorithm converges geometrically to π in L^1 .*

In the continuous case, we need some notes. Assume $\mathcal{X} \subseteq \mathbf{R}^m$ is the state space,

$$p(X, Y) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}$$

is the Markov chain's transition kernel.

Lemma 4. *Suppose that there is a non-negative function $q : \mathcal{X} \rightarrow \mathbf{R}$ such that $q(X) > 0$ on a set of positive Lebesgue measure, and for some $N \in \mathbf{N}$*

$$p^{(N)}(X, Y) \geq q(Y)$$

for all X in the domain of the Markov chain, $\forall Y \in \mathcal{X}$. Then the Markov chain is aperiodic and positive recurrent and $\exists M < \infty, 0 < \rho < 1$, such that

$$\|p^{(t)}(X_0, \cdot) - \pi(\cdot)\| = \int_{\mathcal{X}} |p^{(t)}(X_0, Y) - \pi(Y)| dY < M\rho^t,$$

where the ratio

$$\rho \leq (1 - \int_{\mathcal{X}} q(Y) dY)^{1/N}.$$

In our algorithm, the basins are obtained by the sample paths of Markov chains. Hence in the finite states case, the states in one of the basins must communicate. While in the continuous case, we can denote a basin as

$$B_i = \{X \in \mathcal{X} \mid p(\xi, X) > \delta, \text{ or } \exists n > 0, p^{(n)}(\xi, X) > \delta^n, \delta > 0\},$$

where ξ is the "center" of a local area in the simulation. The realization of a Markov chain must satisfy that for $\forall X, Y \in B_i$, the transition probability has a lower bound, *i.e.* there exist $q(Y)$ and N in B_i such that

$$p^{(N)}(X, Y) > q(Y), \text{ for } \forall X, Y \in B_i \text{ and} \\ 0 < \int_{B_i} q(Y) dY < \int_{B_i} p^{(N)}(X, Y) dY < 1.$$

Therefore we have the following theorem.

Theorem 5. *The algorithm stated in Section 2 has the geometric convergence rate in each local basin.*

For the ratio of the two basins B_1 and B_i , we calculate the value c_1/c_i according to (7) and (8). In the finite case, by the ergodicity

$$\hat{\pi}_i(X) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{X\}}(X^k) \xrightarrow[n \rightarrow \infty]{} \pi_i(X) \text{ a.s.} \quad (10)$$

and in the continuous case the formula

$$\hat{\pi}_i(X) = \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{P_i(\xi^j, X)} \right]^{-1} \xrightarrow[n \rightarrow \infty]{} \pi_i(X) \text{ a.s.} \quad (11)$$

is also true by the large number law. So we give the fact below.

Theorem 6. *The estimated ratio*

$$\left(\frac{c_1}{c_i} \right)^\wedge = \frac{\hat{\pi}_1(X) P(X, \tilde{X})}{\hat{\pi}_i(\tilde{X}) P(\tilde{X}, X)} \xrightarrow[n \rightarrow \infty]{} \frac{c_1}{c_i} \text{ a.s.}$$

where $P(\cdot, \cdot)$ is the reversible transition function and $X \in B_1, \tilde{X} \in B_i$. c_1/c_i is the same with (5).

Note. The convergence of $\hat{\pi}_i(X)$ to $\pi_i(X)$ is easier to get. In a local basin the Markov chains converge geometrically so the formula (10) is obtained. On the other hand, the convergence of (11) is the rule of the i.i.d. sequence we can also get its rate by the central limit theorem and the Berry-Esseen inequality. The details are not discussed here.

References

- [1] Baldi, P., Frigessi, A. and Piccioni, M. (1993) Importance sampling for Gibbs random fields. *Ann. Appl. Prob.*, 3, 914-933.
- [2] Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B*, 55, 25-37.
- [3] Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of change point problems. *Appl. Statist.*, 41, 389-405.
- [4] Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 57, No. 3, 473-484.
- [5] Chen, D., Feng, J., and Qian, M. P. (1995) The metastable behavior of the two dimensional Ising model. In *Dirichlet Forms and Stochastic Processes* (edited by Z. M. Ma, M. Röckner, J. A. Yan, W. de Gruyer, Berlin) pp. 73-86.
- [6] Diaconis, P. and Stroock D. (1991) Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.* Vol 1, No.1, 36-61.
- [7] Ferrari, P. A., Frigessi, A. and Schonmann, R. H. (1993) Convergence of some partially parallel Gibbs samplers with annealing. *Ann. Appl. Prob.*, 3, 137-153.
- [8] Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Soc.* Vol. 7, No. 4, 457-511.
- [9] Gelman, A. and Rubin, D. B. (1992) A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 627-633. Oxford: Oxford University Press.
- [10] Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp 169-193. Oxford: Oxford University Press.
- [11] Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: Applications of Gibbs sampling in medicine. *J. R. Statist. Soc. B*, 55, 39-52.

- [12] Hwang, C. R., Hwang, S. Y. and Sheu, S. J. (1993) Accelerating Gaussian diffusions. *Ann. Appl. Prob.*, 3, 897-913.
- [13] Ingrassia, S. (1994) On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds. *Ann. Appl. Prob.*, 4, 347-389.
- [14] Liu, J., Wong, W. H. and Kong, A. (1995) Covariance structure and convergence rate of the Gibbs sampler with various scan. *J. R. Statist. Soc. B*, 57, No. 1.
- [15] Nummelin, E. (1984) *General Irreducible Markov Chains and Non-negative Operators*. Cambridge: Cambridge University Press.
- [16] Roberts, G. O. and Smith, A. F. M. (1992) Some convergence theory for Markov chain Monte Carlo. Preprint.
- [17] Roberts, G. O. and Polson, N. G. (1994) On the geometric convergence of the Gibbs sampler. *J. R. Statist. Soc. B*, 56, No. 2, 377-384.
- [18] Roberts, G. O. and Smith, A. F. M. (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Processes Appl.*, 49,207-216.
- [19] Schervish, M. J. and Carlin, B. P. (1992) On the convergence of successive substitution sampling. *J. Comp. Graph. Statist.*, Vol 1, No.2, 111-127.
- [20] Rosenthal, J. S. (1995) Rates of convergence for Gibbs sampling for variance component models. *Ann. of Statist.* Vol. 23, No. 3, 740-761.
- [21] Smith, A. F. M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 55,3-23.
- [22] Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Stat.*
- [23] Yu, Bin (1992) Density estimation in the L^∞ norm for dependent data with applications to the Gibbs sampler. *Ann. Statist.*

² AMS 1991 subject classifications:

Key words and phrases: Gibbs sampler; Metastability; Multimodality; Reversible Markov chain; Stochastic Ising Model.