Locally Optimal Sampler

Ting-Li Chen [*] Shang-Ying Shiu [†]

Abstract

Let $\mathcal{X}$ be a finite space and π be an underlying probability on $\mathcal{X}$. For any real-valued function f defined on $\mathcal{X}$, we are interested in calculating the expectation of f under π. Let $X_0, X_1, \ldots, X_n, \ldots$ be a Markov chain generated by some transition matrix P with invariant distribution π. The time average, $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$, is a reasonable approximation to the expectation, $E_\pi[f(X)]$. In this paper, we propose an MCMC algorithm using a locally optimal transition matrix. From our simulation studies, our proposed method outperformed two famous MCMC algorithms, the Gibbs Sampling and the Metropolis-Hastings algorithm.

Key Words: Markov Chain Monte Carlo, Metropolis-Hastings algorithm, Gibbs Sampling

1. Introduction

Markov chain Monte Carlo methods (MCMC) are algorithms for sampling from probability distributions. In many applications, we only know about a function proportional to the density, while the normalizing constant is usually difficult to compute. The idea of MCMC is to generate a Markov chain with the transition probability between two states that only depends on the ratio of their densities. If the targeted distribution is invariant to the transition matrix, and if the Markov chain is irreducible and aperiodic, the equilibrium distribution will equal to the targeted one. Therefore if we run the Markov chain for a sufficiently long time, the states visited by the chain can be viewed as samples from the target distribution.

The most popular MCMC algorithms are the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs Sampling (Geman and Geman, 1984). Are they the most efficient MCMC algorithms? The evaluation of the performance of MCMC algorithms depends on the comparison criterion. Asymptotic variance, spectral gap, geometric ergodicity, and second largest eigenvalue in absolute value are used as criteria for various approximations. In many applications, the quantity of interest is the expected value. In that case, asymptotic variance is a natural criterion. Based on average case analysis, Chen et al. (2011) derived the optimal transition matrices with respect to the asymptotic variance. However, the theoretical result can not be applied in practice directly due to the size of state space.

In this paper, we proposed an MCMC algorithm based on using the optimal transition matrix locally. We first introduce some preliminary results in Section 2. The main algorithm is stated in Section 3, and our simulation studies are presented in Section 4. We have our conclusion remark in Section 5.

2. Preliminary results

Let $\mathcal{X}$ be a finite state space and π be an underlying probability defined on $\mathcal{X}$ with $0 < \pi_1 \leq \pi_2 \leq \cdots \leq \pi_N$. For any real-valued function f defined on $\mathcal{X}$, the quantity of interest is the expectation of f under π, denoted by $\pi(f) = \sum_{x\in\mathcal{X}} f(x)\pi(x)$. Suppose that

[*] Institute of Statistical Science, Academia Sinica, 128 Academia Road Sec.2, Nankang District, Taipei, 11529 Taiwan

[†] Department of Statistics, National Taipei University, 151, University Rd., San Shia District, New Taipei City, 23741 Taiwan

$X_0, X_1, \ldots, X_n, \ldots$ is a Markov chain generated by some transition matrix P with invariant distribution π, then the time average $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$ is a reasonable approximation to $\pi(f)$. To measure how good P is, let $\nu(f, P)$ be the asymptotic variance:

$$\nu(f,P) \doteq \lim_{n\to\infty} n\mathrm{E}_\mu \left[\frac{\sum_{k=0}^{n-1} f(X_k)}{n} - \pi(f)\right]^2,$$

where μ is any initial distribution. On the average case analysis, the performance of P is measured by averaging $\nu(f, P)$ over standardized f:

$$\int_{f:\pi(f)=0,\pi(f^2)=1} \nu(f,P)dS(f), \tag{1}$$

where $dS(f)$ stands for the uniform measure on the normalized surface area. Hwang (2005) proved that

$$\int_{f\in\mathcal{N},\pi(f^2)=1} \nu(f,P)dS(f) = \frac{2}{N-1}\text{Trace}(I-P)^{-1} - 1. \tag{2}$$

Therefore, the main goal becomes to find p minimizing the trace of $(I-P)^{-1}$. Chen et al. (2011) proved that

$$\min_{P_{N\times N}} (\text{Trace}(I - P_{N\times N})^{-1}) = \sum_{i=1}^{N}(i-1)\pi_i. \tag{3}$$

They further proved that there are 2^{N-1} transition matrices can reach the above lower bound. They also provided a method to construct all the optimal transition matrices.

One of the optimal transition matrix is

$$P = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \frac{\pi_1}{\pi_N} & \ldots & \frac{\pi_2-\pi_1}{\pi_N} & \ldots & \frac{\pi_N-\pi_{N-1}}{\pi_N} \end{pmatrix}. \tag{4}$$

When $\pi_1 = \pi_2 = \ldots = \pi_N$, the corresponding Markov process is deterministic. Intuitively, this is indeed the optimal MCMC to approximate any expectation, since it will not visit the same state again until all other states being visited.

3. Locally Optimal Sampling

In practice, MCMC algorithms are applied to sample distributions when the state space is huge. Otherwise, it is much more efficient to sample distributions directly. When the state space is huge, it is almost impossible to order the states with respect to their probabilities. The optimal transition matrices derived by Chen et al. (2011) are constructed with the assumption that the order of the probabilities of all states is known. Therefore, this theoretic result can not be applied directly. However, this method can be adopted as a local updating rule. In Gibbs Sampling (Geman and Geman, 1984), the local update is sampled from the conditional probabilities. To apply the result by Chen et al. (2011), we can construct the optimal transition matrix based on the conditional probabilities and update each site with this locally optimal transition.

Suppose that the distribution to be sampled is $\pi(x_1, x_2, \ldots, x_d)$, where $x_i \in \Lambda_i = \{1, 2, \ldots, s_i\}$. Take one of the optimal transition matrices, (4), as an example, the locally optimal sampling is proposed as follows:

1. Select a random initial $\mathbf{x}^{(\mathbf{0})} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})$, where each $x_i^{(0)}$ is sampled uniformly from Λ_i.
2. At time t, select a random index i_t uniformly from $\{1, 2, \ldots, d\}$.
3. Compute $f(y) = \pi(x_1^{(t-1)}, x_2^{(t-1)}, \ldots, x_{i_t-1}^{(t-1)}, y, x_{i_t+1}^{(t-1)}, \ldots, x_d^{(t-1)})$ for all $y \in \Lambda_{i_t}$.
4. Order $f(y)$. Let $g(y)$ be the order of y over f, and denote $y_{(j)}$ for the y with $g(y) = j$.
5. (a) If $g(x_{i_t}^{(t-1)}) \neq s_{i_t}$, assign $x_{i_t}^{(t)} = y_{(g(x_{t_i}^{(t-1)})+1)}$.

   (b) If $g(x_{i_t}^{(t-1)}) = s_{i_t}$, sample z from the multinomial distribution with probability

   $$\{\frac{f(y_{(1)})}{f(y_{(s_{i_t})})}, \frac{f(y_{(2)}) - f(y_{(1)})}{f(y_{(s_{i_t})})}, \ldots, \frac{f(y_{(s_{i_t})}) - f(y_{(s_{i_t}-1)})}{f(y_{(s_{i_t})})}\},$$

   and assign $x_{i_t}^{(t)} = y_{(z)}$.

   (c) $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i_t$.
6. Go back to step 2 until t is large enough.

Note that we select a random site at step 2. One can modify step 2 to update all sites in a preselected order.

(4) is just one of the 2^{N-2} optimal transition matrices. In theory, we can use any optimal transition matrix for local updates. However, it might not be easy to write a code with a particular optimal transition matrix. (4) is an ideal transition matrix, as every state, except the one with the largest probability, goes to the next one deterministically. The following one is another good choice for local updates.

$$P = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ \frac{\pi_1}{\pi_2} & 0 & \cdots & 0 & \frac{\pi_2 - \pi_1}{\pi_2} \\ 0 & \frac{\pi_2}{\pi_3} & \cdots & 0 & \frac{\pi_3 - \pi_2}{\pi_3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\pi_{N-1}}{\pi_N} & \frac{\pi_N - \pi_{N-1}}{\pi_N} \end{pmatrix}. \tag{5}$$

In fact, this matrix is adjoint to (4). Each state is either going to the one right behind itself or going to the largest one.

4. Simulation

Example 1: First, we experimented with the Potts model, a generalization of the Ising model:

$$\Pr(\mathbf{x}) = \frac{1}{Z} \exp\{\frac{1}{T} \sum_{i=1}^{d} 1_{\{x_i = x_{i+1}\}}\}, \tag{6}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, $x_i \in \{1, 2, \ldots, s\}$, x_{d+1} is defined to be equal to x_1, and Z is the normalizing constant. We choose $d = 6$ and $s = 2, 3, \ldots, 6$, so that the size of state space is not too huge, and we can theoretically compute the averaged asymptotic variance. The result with $T =!$ is presented in Table 1.

When s is smaller, the Metropolis-Hastings algorithm outperformed the Gibbs sampler, and vice versa when s is larger. When $s = 2$, Our proposed Locally Optimal Sampler

T=1,d=6	s=2	s=3	s=4	s=5	s=6
Optimal	0.7748	0.8717	0.9264	0.9575	0.9751
Gibbs	2.9116	1.7747	1.4695	1.3358	1.2614
M-H	1.8744	1.5946	1.4507	1.3634	1.3042
LOS	1.8744	1.2837	1.1473	1.0964	1.0709

Table 1: The asymptotic variances of Gibbs Sampler, Metropolis-Hastings algorithm and Locally Optimal Sampler, respectively

(LOS) is identically the same as the Metropolis-Hastings algorithm. For other s, LOS outperforms the other two methods. We also experimented with various T, and we found that LOS always had the smallest averaged asymptotic variance in our experiments.

Example 2: Next, we experimented with another generalization of the Ising model:

$$\Pr(\mathbf{x}) = \frac{1}{Z}\exp\{-\frac{1}{T}\sum_{i=1}^{d}|x_i - x_{i+1}|\}, \tag{7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, $x_i \in \{1, 2, \ldots, s\}$, x_{d+1} is defined to be equal to x_1, and Z is the normalizing constant. Again, We choose $d = 6$ and $s = 2, 3, \ldots, 6$. The result with $T = 1$ is presented in Table 2.

T=1,d=6	s=2	s=3	s=4	s=5	s=6
Optimal	0.5496	0.7434	0.8528	0.9150	0.9501
Gibbs	4.8232	2.6515	2.0010	1.7105	1.5492
M-H	2.7488	2.4146	2.3111	2.2941	2.3049
LOS	2.7488	1.6053	1.2970	1.1756	1.1160

Table 2: The asymptotic variances of Gibbs Sampler, Metropolis-Hastings and Locally Optimal Sampler, respectively

From Table 2, LOS again has the smallest averaged asymptotic variances among three, and we observe similar results in different T's.

Example 3: Now we experimented with the Potts model (6) again, but with a much larger state space. We choose $T = 1$, $s = 5$ and $d = 100$. With this setup, the size of the state space is roughly 7.9×10^{69}, and it is impossible to theoretically compute the averaged asymptotic variance, which requires the eigenvalues of the transition matrix of size $5^{100} \times 5^{100}$. Instead, we selected four statistics and computed their sample variances based on 1000 trials of simulated MCMC. The four statistics are the longest run length, the number of runs, the number of sites being 1, and the longest run length of state 5. For each MCMC algorithm, we performed 1000 markov chains, and we recorded the sample means and the sample variances in each iteration. The results are shown in Figure 1 and Figure 2. The red line is the result by the Gibbs Sampler, the blue one is by the Metropolis-Hastings algorithm, and the black one is by the Locally Optimal Sampler. The dash-line is the result when the site visited is random, and the solid-line is the one when the site visited is by a pre-selected order.

From the simulation results, the Metropolis-Hastings algorithm always produced much larger sample variances. Our proposed LOS almost always produced smaller variances. For “the number of sites being 1” and “the longest run length of state 5” , LOS drastically reduced the sample variances to smaller than half of those from the other two samplers.

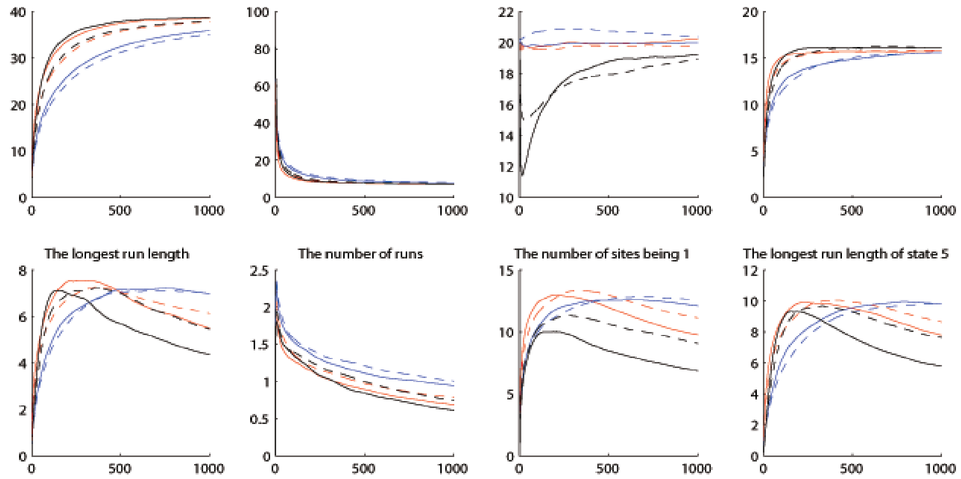

Figure 1: Sample means and variances by the Gibbs Sampler (red), the Metropolis-Hastings algorithm (blue) and Locally Optimal Sampler (black) for the first 1000 iterations.

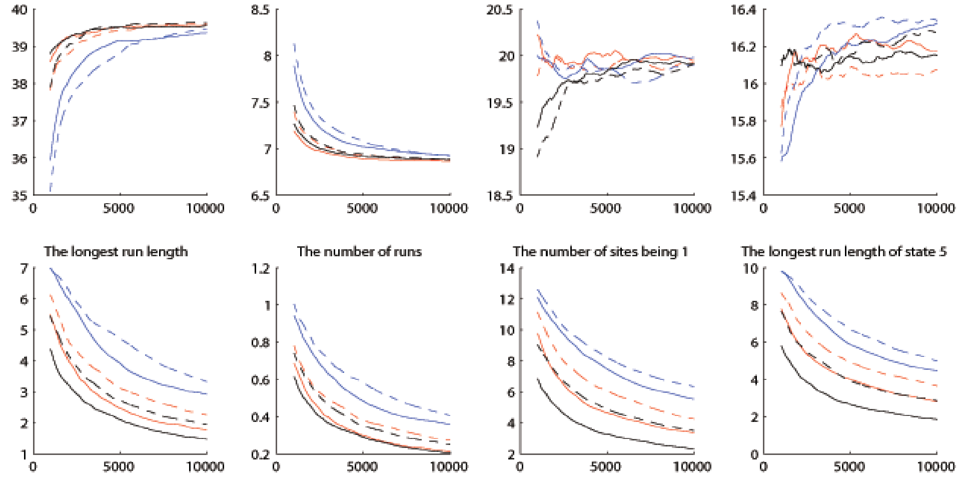

Figure 2: Sample means and variances by the Gibbs Sampler (red), the Metropolis-Hastings algorithm (blue) and Locally Optimal Sampler (black) for iterations 1001 to 10000.

In these two statistics, LOS needed only about 1/3 of iterations to produce similar sample variance by the Gibbs sampler. Note that most of the LOS updates are deterministic moves (except when the state is a local maximum), which are computationally more efficient compared to the probabilistic moves adopted in Metropolis-Hastings and Gibbs Sampler.

5. Conclusion

In this paper, we proposed a locally optimal sampler using the optimal transition matrix by Chen et al. (2011). LOS is very efficient. From our simulation studies, it outperformed the two most popular MCMC algorithms, the Gibbs Sampling and the Metropolis-Hastings algorithm.

References

Chen, T.-L., Chen, W.-K., Hwang, C.-R., and Pai, H.-M. (2011). On the optimal transition matrix for mcmc sampling. manuscript.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Hastings, W. K. (1970). Monte-carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–106.

Hwang, C.-R. (2005). Accelerating monte carlo markov processes. *COSMOS*, 1(1):87–94.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.