Management Science and Financial Engineering
Vol 19, No 2, November 2013, pp.49-53 http://dx.doi.org/10.7737/MSFE.2013.19.2.049
ISSN 2287-2043 | EISSN 2287-2361 |

Improved MCMC Simulation for Low-Dimensional Multi-Modal Distributions

Hyunwoong Ji
Department of Industrial Engineering, Seoul National University
Jaewook Lee
Department of Industrial Engineering, Seoul National University
Namhyoung Kim*
Department of Industrial Engineering, Seoul National University

(Received: October 30, 2013 / Revised: November 10, 2013 / Accepted: November 11, 2013)

ABSTRACT

A Markov-chain Monte Carlo sampling algorithm samples a new point around the latest sample due to the Markov property, which prevents it from sampling from multi-modal distributions since the corresponding chain often fails to search entire support of the target distribution. In this paper, to overcome this problem, mode switching scheme is applied to the conventional MCMC algorithms. The algorithm separates the reducible Markov chain into several mutually exclusive classes and use mode switching scheme to increase mixing rate. Simulation results are given to illustrate the algorithm with promising results.

Keywords: Markov-Chain Monte Carlo, Support Partitioning, Random Walk Metropolis-Hastings, Markov Chain Mixing Time

* Corresponding Author, E-mail: namhyoung@snu.ac.kr

1. INTRODUCTION

Markov chains which are constructed using Markov chain Monte Carlo algorithm have the desired distribution as its target distribution. Constructing a Markov chain with the desired properties is usually possible. However, difficulties of using MCMC algorithms are related to determining when to stop sampling. In the case of sampling from a multi-modal distribution, the obtained Markov chain is usually reducible that the distribution of the Markov chain even does not converge to the target distribution, which is more difficult problem.

Metropolis Adjusted Langevin algorithm or Gibbs sampling can be a solution for the problem (Roberts and Stramer, 2002). However the both two algorithms have a defect. The former has two unspecified parameters to improve mixing rate, but it is difficult to find the optimal two parameters. Also the method is applicable to a special type of distributions. The latter uses coordinate-wise sampling from its conditional densities to avoid multi-dimensional problem, but getting closed form of conditional densities is not always possible.

Therefore, a novel MCMC algorithm which is applicable to any probability distribution is proposed to solve the multi-modal problem. The algorithm separates the reducible Markov chain into communication classes which are mutually exclusive, and set the classes as modes. To increase mixing rate, mode switching with the classes is used and points are sampled from each mode.

This study is organized in the following ways. Section 2 explains the main idea of the proposed method. Next section addresses the step by step algorithm of it. Experimental result appears in Section 4. Section 5 discusses major results and implications of our study.

2. MAIN IDEA

The conventional MCMC algorithms are applicable practically only when the constructed Markov chain is

irreducible (Andrieu *et al*., 2003). However, if a target distribution has multi-mode and, especially, when the distances among modes are long, constructing the corresponding irreducible Markov chain can be a hard problem (Kalogeropoulos *et al*., 2006). Therefore, if general algorithms are applied to multi-modal distribution, the Markov chains obtained by the algorithms cover only one communication class (i.e. all samples are nearby a certain mode), which is an undesirable consequence.

Our purpose is finding an algorithm which make use of any of the existing MCMC algorithms and also applicable to sampling from multi-mode distribution. To use the existing MCMC algorithm, the resulting Markov chain of it should be irreducible, so we add transition procedure artificially from a communication class to any another communication class. By doing so, the constructed Markov chain has only one communication class which means the chain is irreducible.

To realize the idea, transition probabilities from a communication class to another class should be set. To make the distributions of the chain converge to the target distribution, the transition probability should be the ratio between probabilities of two communication classes. On the first try, probabilities of each communication classes were estimated by using non-parametric method including Kernel smoothing and naïve histogram method. When it comes to one dimensional case, the probabilities estimated appropriately in terms of accuracy, but it suffers from hard computation. Also the computational cost will increase exponentially as dimension of support of a target distribution increase (Epanechnikov, 1969), so non parametric method is neglected to estimate the transition probability.

The reason of the hard computation is that the non-parametric approach estimates the exact value of the probabilities of communication classes. However, we do not need all the information and only information that we need is merely the ratio between probabilities of two communication classes.

By using a summation of f(x) at each point, which is also computed in most MCMC algorithms, we can estimate the ratio between probabilities of two communication classes. As a consequence, computational time has decreased impressively, with similar or increased performance relative to applying non-parametric method on it. More details are in Section 3.1 and Section 3.2.

3. ALGORITHM

Step 1: Support Partitioning

Partition a support of a target distribution into proper subset of a communication class. For further convenience, rename the proper subset of a communication class as sub-communication class. Let the number of sub-communication classes be **N** and the i-th sub-communication class be A_j. Therefore, the support can be expressed as $S = A_1 \cup A_2 \cup \cdots \cup A_N$.

Step 2: Initialization

For each sub-communication class, initialize $\hat{W}_l$ which approximate

$$Wi = \int_{x \in A_i} f(x)\,dx$$

where f(x) is a target distribution. For each i = 1, ⋯, N, set p_i which is a point on A_i. Let sample set Σ be an empty Markov chain.

Step 3: Do Step 3.1 to Step 3.3 L times

Step 3.1: Mode selection

Sample a number from the integers 1 to N with probability $\hat{W}_l's$. Let the sample be j. Add p_j, the point on $\hat{W}_l$ to Σ, the Markov chain.

Step 3.2: With starting point p_j of Step 3.1, generate M-1 samples by using a MCMC-algorithm from $f_{A_j}(x)$ where

$$f_{A_j}(x) = \begin{cases} f(x)/k_{A_j}, & x \in A_j \\ 0, & x \notin A_j \end{cases}$$

$$k_{A_j} = \int_{x \in A_i} f(x)dx$$

Step 3.3 Add the M-1 samples to Σ. Update $\widehat{W}_l$ by using the latest M samples of Σ.

There are a lot of ways to updating the approximated weight $\hat{W}_l$ of Step 3.3. In the following two sub sections, two methods are introduced. The first method is simple, but computationally expensive. The second method is more complicated but computationally efficient.

3.1 Weight Updating Method 1

For each A_i, $f_{A_i}(x)$ can be estimated through density estimation using M samples of Step 3.3. Let the estimated density be $\hat{f}_{A_i}$. $f_{A_i}(x)$ can be rewritten as $f_{A_i}(x) = f(x)/\int_{x \in A_i} f(x)dx$. Therefore, $\sum_{j=1}^{M} f(x_j)/\hat{f}_{A_i}(x_j)$ can be a proper estimator of $\int_{x \in A_i} f(x)$. However, this method is extravagant in terms of computation, because density function is estimated to estimate $\int_{x \in A_i} f(x)$ which is just a number.

3.2 Weight Updating Method 2

As once mentioned in Section 2, we only need the ratio of values of probability among the sub-communication classes, not the exact values of them. The second method is based on the point. To explain the second method, some equations and definitions are needed.

$$wi = \int_{\{x \in A_i\}} f(x)dx$$

The conditional distribution is given as follows.

$$f_{A_i}(x) = \frac{f(x)}{w_i}$$

$$\frac{1}{N}\sum_{x_{i\in A_j}} \frac{1}{f(x)} \approx \int_{\{x\in A_j\}} \frac{1}{f(x)} f_{A_j}(x)dx$$

$$= \int_{\{x\in A_j\}} \frac{1}{f(x)} \frac{f(x)}{w_j} dx = \int_{\{x\in A_j\}} \frac{1}{f(x)} \frac{f(x)}{w_j} dx = \frac{1}{w_j} V_{A_j}$$

$$\therefore \frac{w_j}{V_{A_j}} : \frac{w_k}{V_{A_k}} = \left\{\sum_{x_{i\in A_j}} \frac{1}{f(x)}\right\}^{-1} : \left\{\sum_{x_{i\in A_k}} \frac{1}{f(x)}\right\}^{-1}$$

where $V_{A_j} = \int_{\{x\in A_j\}} 1\, dx$.

Because of the term $\frac{1}{f(x)}$, for all $x \in A_j$ $f(x)$ should be a positive number. In other words, the sub-communication class should be a subset of support of the corresponding target distribution, which means this algorithm pre-assumes that the support of the target distribution is known. Besides, what we need is $w_j : w_k$, not $\frac{w_j}{V_{A_j}} : \frac{w_k}{V_{A_k}}$, so this algorithm also pre-assumes that the volume of each sub-communication class is known.

The interesting point is that for all sample points, the value of target density is computed in most of MCMC algorithms. Therefore, if such algorithms are applied to our algorithms, only the additional L×M additions are needed compared to just using the algorithm.

4. EXPERIMENTAL RESULTS

Two toy problems are addressed in this section. The first problem is sampling from an 1-dimensional mixture distribution consist of uniform, Gaussian and exponential distribution. The second problem is sampling from 2-dimensional Gaussian mixture distribution.

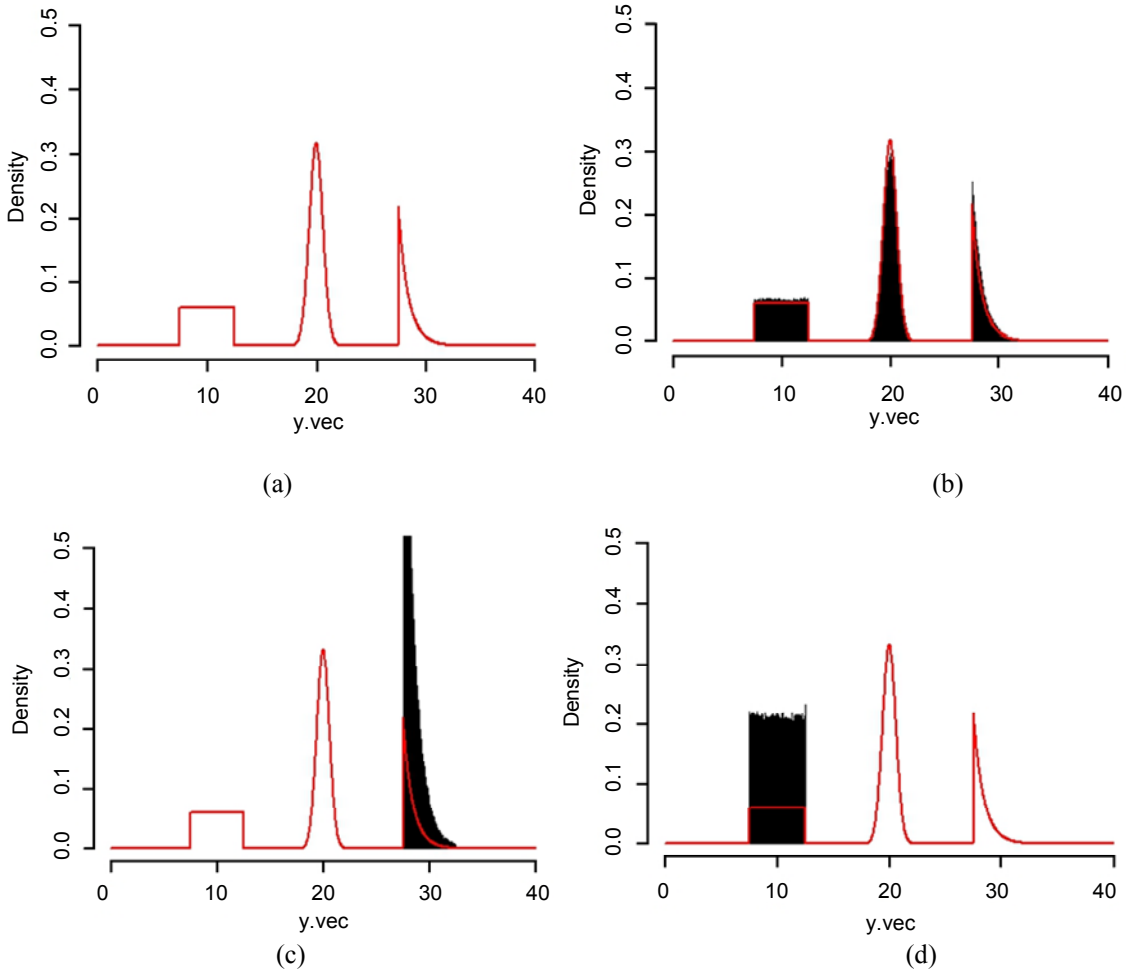

Figure 1. (a) Target Distribution of the First Problem (b) Histogram of 1,000,000 Samples Generated by the Proposed algorithm (c) Histogram of 1,000,000 Samples Generated by the Metropolis-Hastings Algorithm (d) Histogram of 1,000,000 Samples Generated by the Metropolis-Hastings Algorithm

In the first problem, the target distribution, f(x), has density 0.3×U(x: 7.5, 12.5)+0.5×N(x: 20, 0.6)+0.2×E(x-30: 1.1), where U(x: 7.5, 12.5) is a uniform distribution with two boundaries, 7.5 and 12.5, N(x: 20, 0.6) is a normal distribution with mean, 20, and variance, 0.6 and E(x: 1.1) is an exponential distribution with rate parameter, 1.1. Figure 1(a) describes the probability density of the first toy problem.

We have compared the proposed algorithm with random walk Metropolis-Hastings algorithm. Figure 1(b) ~Figure 1(d) describes the histogram of 1,000,000 samples generated by the proposed algorithm and 2 Markov chains of length 1,000,000 acquired by the random walk Metropolis-Hastings algorithm, respectively. Throughout the figures, the exact target distribution is depicted as the red curve to determine whether each chain is converges to the chain or not.

There are some diagnostics that determine when it is safe to stop sampling. The diagnostics include methods of Gelman and Rubin (1992), Geweke (1992), and Heidelberger and Welch (1983) Cowles *et al*. (1996). However, such methods are applicable only when the acquired Markov chain is irreducible (Cowles *et al*., 1996), which is not the case of our interests. Therefore, the histograms of Figure 1(b) to Figure 1(d) are suggested to compare two algorithms. It is clear that the proposed algorithm outperforms the random walk Metropolis-Hastings algorithm by comparing Figure 1(b) with Figure 1(c) and Figure 1(d).

In the second problem, the target distribution, f(x) has density

$$0.3 \times N(\mu_1, \Sigma) + 0.7 \times N(\mu_2, \Sigma)$$

where $\mu_1 = (-3, -2)$, $\mu_2 = (2, 2)$ and $\Sigma = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$.

Figure 2 describes the probability density of the second toy problem.

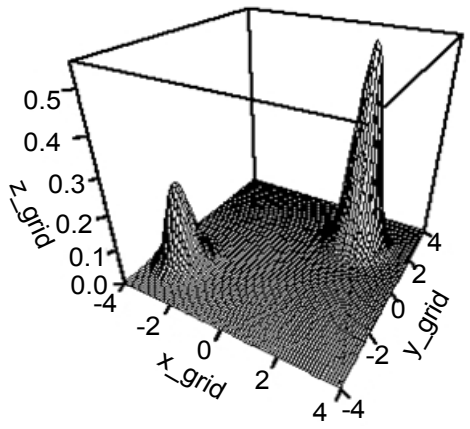

Figure 2. Target Distribution of the Second Problem

In this problem, it is not appropriate to determine convergence of the Markov chain graphically, because the distribution is defined on 2-dimensional support. Therefore, distance between empirical mean of the Markov chain and the expectation of the target distribution is used as an alternative convergence measure. Table 1 shows the performance of the proposed and RMH (random walk Metropolis-Hastings algorithm).

Table 1. Performance Comparisons by Estimator of First Moment

	$\hat{E}(x_1)$	$\hat{E}(x_2)$	$E(x_1)$	$E(x_2)$	Distance
Proposed	0.5349	0.8285	0.5	0.8	0.04505
	0.4820	0.7842			0.02395
	0.5167	0.8072			0.01818
RHM	-3.1634	-1.9145			4.55949
	2.0579	1.9686			1.94748
	2.0960	2.21472			2.13275

We have generated 3 Markov chain for each algorithm. According to Euclidean distance between empirical mean, $\hat{\mu}$, and true mean, μ of Table 1, it is concluded that proposed algorithm outperforms RHM. An interesting thing is empirical mean of RHM tend to be nearby $\mu_1 = (-3, -2)$ or $\mu_2 = (2, 2)$. This is because Markov chains of RHM are reducible and the corresponding communication classes are comprised of two disjoint areas; one of which contains $\mu_1 = (-3, 2)$ and the other of which contains $\mu_2 = (2, 2)$.

5. CONCLUSIONS

By monitoring the 1-dimensional case in Section 4, the suggested method works well relative to random walk Metro-Hastings algorithm. We obtained consistent result when treating the 2-dimensional case. We expect that the proposed method can be used in various applications like Bayesian statistics, physics and computational finance (Rannala, 2002; Binder and Heermann, 2010; Kim and Lee, 2013). While the proposed method improved sampling performance, something should be considered as future works.

In further researches, one can seek for a better criterion for measuring the result. Throughout this paper, we judge the performance of two algorithms by monitoring and estimation error. However, it is impossible to use the former measure on multi-dimensional distribution and the latter measure even does not exist when dealing with practical distribution. Therefore, finding a criterion which measures a similarity between a distribution of Markov chain and the corresponding target distribution, when the target distribution is multi-modal could be another research subject. Also, there is a need for finding communication classes with respect to the target density and the Markov chain of it, which needs further research.

ACKNOWLEDGEMENT

This work was supported by Research Settlement Fund for the new faculty of SNU.

REFERENCES

Andrieu, C. and N. de Freitas, A. Doucet, and M. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning* 50, 1/2 (2003), 5-43.

Binder, K. and D. W. Heermann, "Monte Carlo simulation in statistical physics: an introduction," Springer, 2010.

Cowles, M. K. and B. P. Carlin, "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association* 91, 434 (1996), 883-904.

Epanechnikov, V., "Non-Parametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications* 14, 1 (1969), 153-158.

Kalogeropoulos, K., G. Roberts, and P. Dellaportas, "Irreducible Markov Chain Monte Carlo Schemes for Partially Observed Diffusions," *Nonlinear Statistical Signal Processing Workshop*, IEEE, (2006) 216-219.

Kim, N. and J. Lee, "No-arbitrage implied volatility functions: Empirical evidence from KOSPI 200 index options," *Journal of Empirical Finance* 21 (2013), 36-53.

Rannala, B., "Identifiability of Parameters in MCMC Bayesian Inference of Phylogeny," *Systematic Biology* 51, 5 (2002), 754-760.

Roberts, G. O. and O. Stramer, "Langevin Diffusions and Metropolis-Hastings Algorithms," *Methodology And Computing In Applied Probability* 4, 4 (2002), 337-357.