Open Access

# Multidimensional Integration of a Positive Function Using Markov Chain Monte Carlo

Guthrie Miller*

*Los Alamos National Laboratory, Group RP-2, USA*

**Abstract:** A new Markov Chain Monte Carlo algorithm that allows parallel processing has been described in a previous paper ("Markov Chain Monte Carlo Calculations Allowing Parallel Processing Using a Variant of the Metropolis Algorithm") that appeared in this journal in 2010. In this second follow-on paper, the problem of calculating the normalization integral of the distribution function is considered. In the usual Markov Chain Monte Carlo calculations this normalization integral is not necessary; however, this integral is needed for Bayesian hypothesis testing and is a key quantity (the partition function) in statistical physics. Three different methods of calculating this integral are considered: an importance-sampling method, a reference-hypothesis method, and a direct method of integration over the random-walk region. This latter method is shown to provide the normalization integral in situations where the other methods fail.

**Keywords:** Metropolis algorithm, parallel processing, Markov Chain Monte Carlo (MCMC), Bayesian hypothesis testing, statistical mechanics, partition function.

## 1. INTRODUCTION

This is the third paper is a series that considers Markov Chain Monte Carlo (MCMC) algorithms suitable for parallel processing. The reader should refer to the preceding papers [1, 2] for general background and literature references.

Markov Chain Monte Carlo is a numerical technique that allows the evaluation of expectation value integrals of the form

$$E[g(\theta)] = \frac{\int d\theta\, f(\theta) g(\theta)}{\int d\theta\, f(\theta)}. \tag{1}$$

where $f$ is a positive function, $\theta$ is a multidimensional parameter, and $g(\theta)$ is an arbitrary function of $\theta$. Markov Chain Monte Carlo does not naturally provide the value of the "normalization integral" denominator in Eq. (1); however, this denominator is very important for Bayesian hypothesis testing and as the partition function in Statistical Physics.

A review (as of 1996) of Bayesian hypothesis testing is given in Chapter 10 of ref. [3]. Importance sampling is the basic approach considered, where

$$\int d\theta\, f(\theta) = \int d\theta\, g(\theta) \frac{f(\theta)}{g(\theta)} \cong \frac{1}{T} \sum_t \frac{f(\theta_t)}{g(\theta_t)}$$

with $g$ some probability function (positive and normalized) from which samples $\theta_t$ can be drawn. The method with $g$ the function $f$ itself is the first existing method considered in this paper. In the conclusion of Chapter 10 the author (Adrian Raftery) states "Research on this topic is at an early stage and much remains to be done." A more recent textbook [6] does not add anything new on this subject.

In the present paper, three methods for determining the normalization integral are considered. The first is the importance sampling method mentioned above. The second is a reference-hypothesis method used in ref. [7]. The third is a direct method of integration over the random-walk region.

These methods make use of the MCMC algorithms discussed in the previous papers [1, 2]: the MRT algorithm originating in the 1953 paper [4] by Nicolas Metropolis, Adriana Rosenbluth, Marshall Rosenbluth, Agusta Teller, and Edward Teller, the B algorithm proposed by Barker [5] in 1965, and the new algorithm introduced in Refs. 1. and 2. Instead of a single candidate for the next position of the chain, multiple candidates are being considered. The MRT or B algorithms do not improve by having multiple candidates. The new algorithm is noteworthy because it benefits greatly from having multiple candidates and is therefore suitable for parallel processing, taking good advantage of the multiprocessor computing environments that will be more and more prevalent in the future.

## 2. STATEMENT OF THE PROBLEM AND SOLUTIONS CONSIDERED

We are considering the multidimensional integral of a positive function $f(\theta)$. All dimensions of $\theta$ have domain 0 to 1. The integral $I$

$$I = \int_0^1 d\theta\, f(\theta)$$

can be thought of as the normalization integral that converts $f$ into a probability distribution.

The first solution considered is to write

*Address correspondence to this author at the Los Alamos National Laboratory, Group RP-2, USA; Tel: 505 667 5547; Fax: 505 665 6071; E-mails: guthriemiller@earthlink.net; guthrie@lanl.gov

$$I = \frac{\int d\theta\, f(\theta)}{\int d\theta\, f(\theta)/f(\theta)} = \frac{1}{E(1/f)}, \tag{2}$$

where the expectation value denoted by $E$ is calculated using MCMC.

The second solution considered is to add another variable that determines the hypothesis: the normal one or a reference hypothesis. With probability $P_{H1}$ (the probability of being in the normal hypothesis in the assumed steady state distribution, taken to have the value 1/2) the hypothesis is assumed to be the normal hypothesis, and with probability $P_{H0}$ it is the reference hypothesis, where $f(\theta)$ has a constant reference value $f_0$. Then, using normal MCMC, one calculates the fraction of the time the chain is in the reference hypothesis rather than the normal hypothesis, and

$$I = f_0 \frac{P_{H0}}{P_{H1}} \frac{E(I_{hyp})}{1 - E(I_{hyp})}, \tag{3}$$

where the function $I_{hyp}(\theta)$ takes the value 1 when the chain is in the normal hypothesis and 0 when the chain is in the reference hypothesis.

The third solution considered is to continue the chain some number of iterations after equilibration with all parameters simultaneously varied. There are $l$ candidates for the next position of the chain $\theta_i$ for $i$ = 1… $l$ that are generated from a multidimensional random walk distribution with interval size $\Delta_j$, where $j$ runs over all the $n$ dimensions of $\theta$. Because the point $i$ = 0 (the current chain position) is always at or near a point of high $f$ values and not randomly distributed in the entire random walk interval like the other points, it is left out, and the solution for the normalization integral is given by

$$I \cong \prod_{j=1}^{n} \Delta_j \frac{1}{l} \sum_{i=1}^{l} f(\theta_i), \tag{4}$$

averaged over all chain iterations in the continuation run. This method can be applied with any of the algorithms, even for $l$ = 1 (a single candidate) because of the averaging over a large number of chain iterations, but it requires that the random-walk step size be large compared with the size of the support region of the distribution function $f$.

## 3. ANALYSIS OF THE REFERENCE HYPOTHESIS SOLUTION FOR FINITE, INTEGER-VALUED MARKOV CHAINS

To understand more clearly the use of a reference hypothesis, let us first consider an integer-valued Markov Chain, as was done in ref. [2]. The basic ideas and formalism are given in ref. [2]. The eigenvalues and eigenvectors of the transition matrix are calculated. The time to reach steady state is obtained from the magnitude of the second largest eigenvalue. This is because the initial state can be represented as a linear combination of eigenvectors, and each eigenvector has time dependence $\lambda^t$, where $\lambda$ is the eigenvalue and the $t$ is the chain iteration number. The largest eigenvalue is always 1, and this eigenvalue-1 eigenvector gives the steady state. The second largest eigenvalue is the longest persisting of all the others. It has time dependence exp(-$t/\tau$), where $\tau$ = - 1/ln($\lambda$).

The first approach to the task of including a reference hypothesis is to expand the parameter space from $i$ = 1,…$n$ to $i$ = 0,…$n$, where $f_0$ is the probability of being in the reference hypothesis. It is convenient to order the indices $i$ = 1,…$n$, 0 with 0 last. We define a new transition matrix given in block form by

$$A = \begin{bmatrix} A_{i,j} & A_{i,0} \\ A_{0,j} & * \end{bmatrix}, \tag{5}$$

where $A_{i,j}$ is the transition matrix for the normal hypothesis as discussed in ref. [1], however with the constant $a$ = 1/2. which is the proposal probability of remaining in the same hypothesis, rather than switching. For the B algorithm, the $n$-dimensional column matrix $A_{i,0}$ is given by

$$A_{i,0} = \left[ \frac{1}{2n} \frac{\overline{f}_i}{f_0 + \overline{f}_i} \right],$$

the $n$-dimensional row matrix $A_{0,j}$ is given by

$$A_{0,j} = \left[ \frac{1}{2n} \frac{\overline{f}_0}{\overline{f}_0 + \overline{f}_j} \right],$$

and $*$ denotes the scalar needed for normalization of the last column, given by 1 minus the sum of the off-diagonal terms. Similar formulas apply for the MRT algorithm. The diagonal elements of the old transition matrix $A_{i,j}$ are replaced by the new values taking into account the changed multiplying factor $a$ = 1/2.

For the new algorithm, this time including $i$ = 0,

$$A_{i,j} \propto \overline{f}_i$$

with column by column normalization,

$$A_{i,j} = \frac{\overline{f}_i}{\sum_{i'=0}^{n} \overline{f}_{i'}},$$

for all the nonzero elements of the transition array.

In Fig. (**1**) below is shown the eigenvalue-1 eigenvector with $\overline{f}_0 = 0.1$, using the new algorithm. The full width of the random walk for the normal hypothesis ($l$ = 20) is 5 times the standard deviation of assumed steady state distribution ($\sigma$ = 4).

The other two algorithms gave identical agreement between the eigenvalue-1 eigenvector and the assumed steady state as they must.

In Table **1** are shown the times to reach a steady state, calculated from the second largest eigenvalue (the largest is always 1) for various widths of the assumed steady state distribution. The ratio of $l$ to $\sigma$ is kept constant at 5.

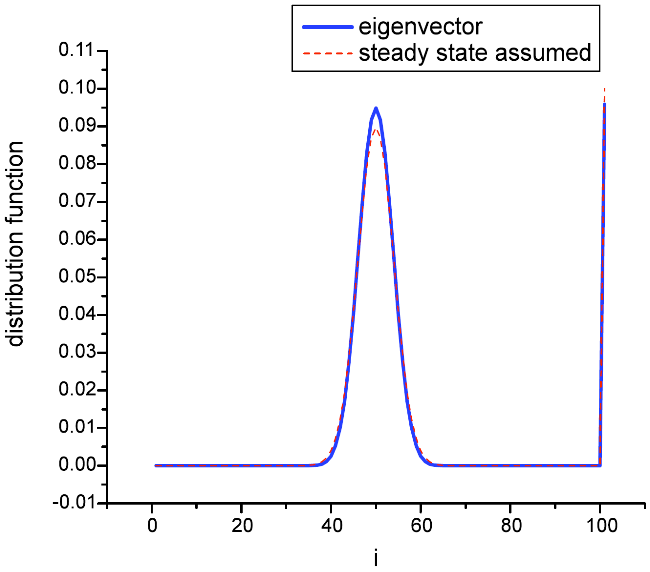

**Fig. (1).** Eigenvalue-1 eigenvector, using the new algorithm.

**Table 1. Time to Reach Steady State (3τ) for the Three MCMC Algorithms for Different Width Steady State Distribution Functions, with (nhyp = 2) and without (nhyp = 1) a Reference Hypothesis**

| Algorithm | $\sigma$ | $l$ | nhyp | $3\tau$ |
|---|---|---|---|---|
| MRT | 2 | 10 | 1 | 6.4 |
| | | | 2 | 11.1 |
| | 4 | 20 | 1 | 7.1 |
| | | | 2 | 13.6 |
| | 8 | 40 | 1 | 7.4 |
| | | | 2 | 22.5 |
| B | 2 | 10 | 1 | 8.4 |
| | | | 2 | 16 |
| | 4 | 20 | 1 | 9.2 |
| | | | 2 | 17.8 |
| | 8 | 40 | 1 | 9.7 |
| | | | 2 | 29.8 |
| new | 2 | 10 | 1 | 1.6 |
| | | | 2 | 1.5 |
| | 4 | 20 | 1 | 1.8 |
| | | | 2 | 1.6 |
| | 8 | 40 | 1 | 1.8 |
| | | | 2 | 1.7 |

In this first approach to the construction of a transition matrix with a reference hypothesis, from the reference hypothesis, transitions occur to the entire space of the normal hypothesis. In the continuous multidimensional parameter case, this needs to happen even with grouping of parameters. An alternative is to have transitions from the reference hypothesis occur to some small random walk interval around a current point. This means that even in the reference hypothesis we need to remember a current point.

To do this we need to expand the parameter space from $i = 1,\ldots n$ to $i = 1,\ldots 2n$, where the points $i = \mathrm{n} + 1,\ldots 2n$ are in the reference hypothesis but record the position of the normal hypothesis. The transition matrix is given in block form by

$$A = \begin{bmatrix} A_{i,j} & A_{i,0} \\ A_{0,j} & A_{0,0} \end{bmatrix}. \tag{6}$$

All the blocks are now $n \times n$ square matrices with the same template of nonzero elements, namely that of the original transition matrix $A_{i,j}$ for the normal hypothesis, although other, simpler templates are possible (for example a diagonal for the off-diagonal blocks, where the parameters do not change when switching hypotheses). For the B algorithm, the $n \times n$ square matrix $A_{i,0}$ is given by

$$A_{i,0} = \left[ \frac{1}{2l} \frac{\bar{f}_i}{\bar{f}_0 + \bar{f}_i} \right],$$

where the nonzero elements exist in a diagonal band of size $l$ just like in the original transition matrix $A_{i,j}$. The $n \times n$ square matrix matrix $A_{0,j}$ is given by

$$A_{0,j} = \left[ \frac{1}{2l} \frac{\bar{f}_0}{\bar{f}_0 + \bar{f}_j} \right],$$

where the nonzero elements are in a transposed position relative to $A_{i,0}$.

The $n \times n$ square matrix matrix $A_{0,0}$ is given by

$$A_{0,0} = \left[ \frac{1}{2l} \frac{\bar{f}_0}{\bar{f}_0 + \bar{f}_0} \right],$$

where, again, the nonzero elements exist in a diagonal band of size $l$ just like in the original transition matrix $A_{i,j}$.

Another way of thinking of this structure is that we have increased the parameter space to be the product of the original space times the space of another integer variable, taking only two values, which determine the hypothesis. The expanded parameter space is ordered so that the parameters of the normal hypothesis come first followed by the parameters of the reference hypothesis. The distribution function in the reference hypothesis is a constant $\bar{f}_0$.

Fig. (**2**) shows the eigenvector with eigenvalue 1, with the same conditions as for Fig. (**1**) except that the probability of the reference hypothesis $\bar{f}_0$ is 0.5 rather than 0.1 (to improve the appearance of the plots).

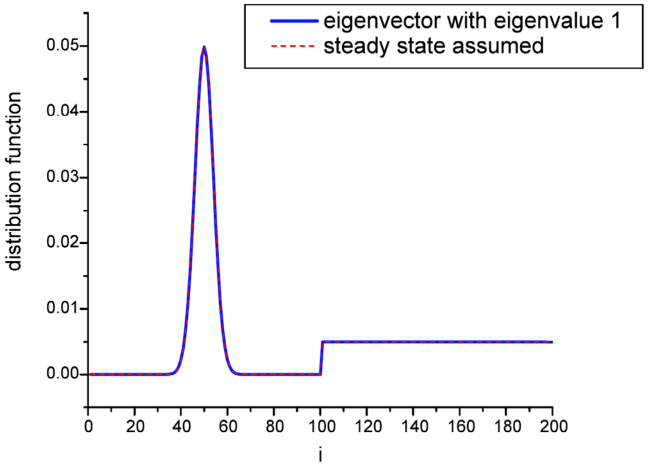

**Fig. (2).** Eigenvector with eigenvalue 1, using the B and MRT algorithms (results indistinguishable).

The new algorithm does not work well in this situation because the desired steady state distribution function does not have a limited size support region. For the reference hypothesis it occupies the entire space. Fig. (**3**) shows the eigenvector with eigenvalue 1 using the new algorithm.

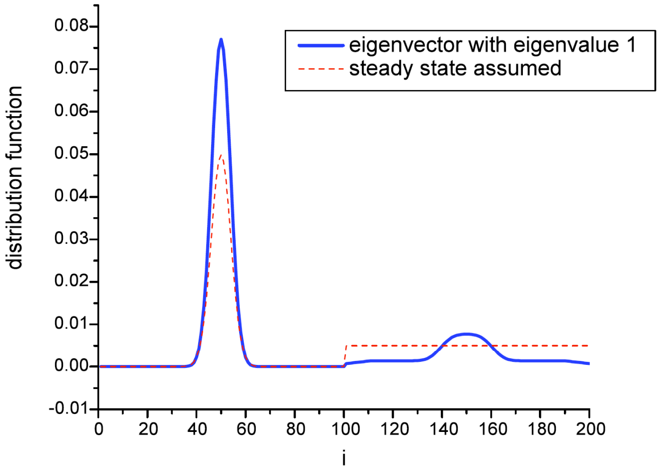

**Fig. (3).** Eigenvector with eigenvalue 1, using the new algorithm.

For narrow steady state distribution functions, the MRT and B algorithms require many more iterations to reach a steady state with a reference hypothesis than without one. This is illustrated in Table **2**.

To understand the long times to reach steady state, for example 933 in the second line of Table **1**, we look at the distribution function after 100 iterations when the distribution is initially concentrated at $i$ = 1 and when the distribution is initially concentrated at $i$ = 50. These are shown in Fig. (**4**).

Even though after 100 iterations the normal hypothesis has reached the desired steady state distribution in both the upper and lower plots, the fraction of the time the chain is in the normal hypothesis is low relative to the desired steady state distribution in the upper plot and high in the lower plot. It takes a long time for the chain to diffusively explore the entire space while in the reference hypothesis. This long time is necessary for the distribution in the reference hypothesis to flatten out.

**Table 2. Time to Reach Steady State (3τ) for the MRT and B MCMC Algorithms for Different Width Steady State Distribution Functions, with (nhyp = 2) and without (nhyp = 1) a Reference Hypothesis. Narrow Distribution Functions Require a Long Time to Reach Steady State**

| Algorithm | *s* | *l* | nhyp | 3*t* |
|---|---|---|---|---|
| MRT | 2 | 10 | 1 | 6.9 |
| | | | 2 | 933 |
| | 4 | 20 | 1 | 7.5 |
| | | | 2 | 221 |
| | 8 | 40 | 1 | 7.9 |
| | | | 2 | 44 |
| | | | | |
| B | 2 | 10 | 1 | 8.9 |
| | | | 2 | 1806 |
| | 4 | 20 | 1 | 9.8 |
| | | | 2 | 413 |
| | 8 | 40 | 1 | 10.3 |
| | | | 2 | 78 |

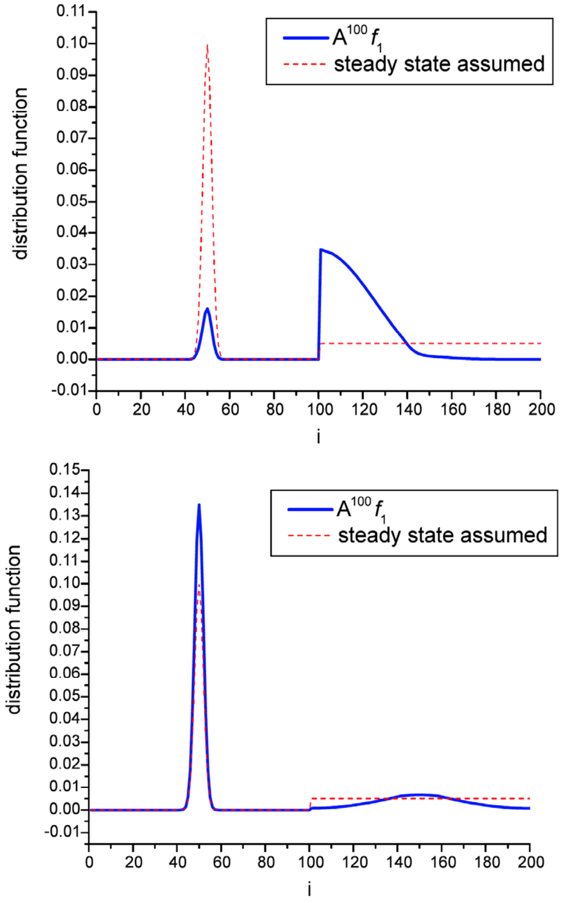

**Fig. (4).** Distribution function after 100 iterations with the chain initially at i = 1 (above) and i = 50 (below), calculated using the MRT algorithm.

## 4. TESTS OF THE METHODS FOR CONTINUOUS VARIABLES

For the test problem $f$ is given by a Gaussian in $n$ dimensions with the same central point $\theta = 0.33$ and the same standard deviation $\sigma$ in all dimensions. There are two versions of the reference hypothesis method, corresponding to the first and second examples in Section 3, which will be denoted as method 2-A and 2-B. We will use both versions. Somewhat surprisingly, given the results of Section 3, the two versions behave similarly. Some improvement is obtained using the new algorithm and method 2-A. The new algorithm is not applicable for method 2-B as has been discussed. The distribution function in the reference hypothesis is chosen so that the chain would be expected to spend equal time in the two hypotheses. If grouping of parameters is used (only 1 group moved at each iteration), a group consists of a single parameter; however, the hypothesis is probabilistically switched for every chain iteration. The first 10% (burnfrac) of the run is discarded.

To demonstrate the first and second methods of calculating the normalization integral given by Eq. (2), the dimensionality is chosen to be 1 and $\sigma$ is chosen to be 0.2. The MRT algorithm is used with one candidate. The random walk step size is 0.5. This is an unchallenging MCMC problem, and for a normal run equilibration occurs by about 10 iterations of the chain. However to calculate the normalization integral, 50000 iterations are used and the results shown in Table **3** are obtained. To demonstrate convergence, two MCMC runs, denoted by MCMC$_1$ and MCMC$_2$ are done. The first starts with all $\theta$ variables equal to 0 and the second with all $\theta$ variables initially equal to 1 and with a different random number seed. The correct answer is given by

$$I = \sqrt{2\pi}\sigma = 0.501 .$$

**Table 3. Calculation of the Normalization Integral for f Given by a One-Dimensional Gaussian with Standard Deviation 0.2, an Unchallenging MCMC Problem. The Correct Answer is I = 0.501. The MRT Algorithm is Used with 50000 Iterations**

| Method | *Normalization Integral I* | |
|---|---|---|
| | MCMC$_1$ | MCMC$_2$ |
| 1 | 0.524 | 0.426 |
| 2-A | 0.435 | 0.486 |
| 2-B | 0.427 | 0.529 |

To demonstrate the second method, a six-dimensional Gaussian is used with $\sigma = 0.01$. This would be a challenging integration problem using other methods of numerical integration. The MRT algorithm is used with a single candidate. The random walk step size Δ is 0.10. For a normal run, equilibration occurs by about 400 iterations of the chain. However to calculate the normalization integral, many more iterations are used and the results shown in Table **4** are obtained. The correct answer is given by

$$I = \left(\sqrt{2\pi}\sigma\right)^6 = 2.48 \times 10^{-10} .$$

**Table 4. Calculation of the Normalization Integral for f Given by a Six-Dimensional Gaussian with Standard Deviation 0.01. The Correct Answer is I = 2.48 × 10$^{-10}$**

| Method | Algorithm | Grouping | #Iterations | Normalization Integral *I* | |
|---|---|---|---|---|---|
| | | | | MCMC$_1$ | MCMC$_2$ |
| 1 | MRT | yes | $5 \times 10^8$ | $3.26 \times 10^{-3}$ | $1.16 \times 10^{-4}$ |
| 2-A | MRT | no | $5 \times 10^8$ | $3.41 \times 10^{-10}$ | $2.71 \times 10^{-10}$ |
| 2-A | new ($l = 8$) | no | $1 \times 10^8$ | $3.30 \times 10^{-10}$ | $2.69 \times 10^{-10}$ |
| 2-B | MRT | yes | $5 \times 10^8$ | $2.32 \times 10^{-10}$ | $1.76 \times 10^{-10}$ |

As can be seen, the first method fails badly. Upon reflection, it is obvious that the expectation of $1/f$ with respect to $f$, cannot converge well unless $f$ is significantly greater than 0 at all points, which is not usually the case. In fact, we are most interested in distributions whose region of support occupies only a minute fraction of the entire space.

The second, reference-hypothesis, method is marginally acceptable in this case. The chain initially remains in one hypothesis or the other for a long time, seeming to be stuck there forever. However, the chain finally begins moving slowly moves back and forth between the two hypotheses, while the distribution in the normal hypothesis remains the same.

To demonstrate the third method, a six-dimensional Gaussian is used with very small $\sigma$ equal to 1 × 10$^{-6}$. This would be a very challenging integration problem using other methods of numerical integration. The new algorithm is used with 8 candidates per iteration and the MRT algorithm with 1 candidate per iteration. In the initial run, with probability 1/2 the random walk step size Δ is 1 × 10$^{-5}$, and with probability 1/2 candidates are chosen from the entire space (Δ = 1). To calculate the normalization integral, after the initial run achieves equilibration the chain is continued for an equal number of iterations with only a small-step-size random walk and without coordinate grouping (all coordinates varied at each iteration). Only the final continuation-run iterations are used to calculate the normalization integral using Eq. (4). The correct answer is given by

$$I = \left(\sqrt{2\pi}\sigma\right)^6 = 2.48 \times 10^{-34} .$$

Fig. (**5**) shows the value of the normalization integral obtained from Eq. (4) as the random walk interval Δ is varied from 5 to 10 times the standard deviation of the Gaussian.

As we would expect, the direct integration method gives a good result as long as the random walk step size is many times larger then the support region of $f$. As the random walk step size is increased, the acceptance fraction during the continuation run becomes smaller, increasing the autocorrelation of the chain, and causing the observed erratic variation of the result.

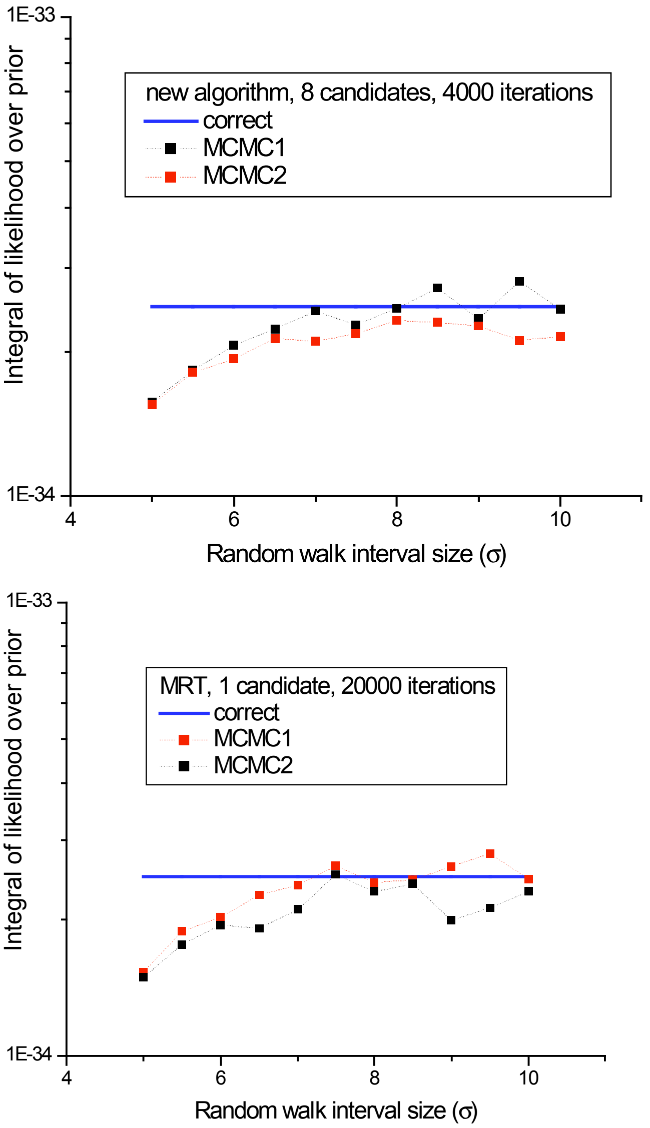

**Fig. (5).** Normalization integral calculated using direct integration over the random-walk region with the new algorithm and the MRT algorithm, as a function of the random walk step size.

## 5. DISCUSSION AND CONCLUSIONS

The normalization integral of the distribution function is not required in normal MCMC and its value is not naturally calculated. However, this quantity is sometimes important in practice, for example in Bayesian hypothesis testing. We have discussed three different methods of calculating the normalization integral. However, for the situation we are most interested in, where the support region of the distribution function occupies only a minute fraction of the entire space, only one of these methods works. With this method, the value of the normalization integral can be obtained by doubling the length of the normal MCMC run. Additional runs varying the size of the random walk step would be necessary to guarantee that the step sizes are sufficiently large. The new method can take advantage of parallel processing, and a very significant reduction of computer run time is possible.

## REFERENCES

[1] Miller G. Markov chain monte carlo calculations allowing parallel processing using a variant of the metropolis algorithm. Open Numer Methods J 2010; 2: 12-7.

[2] Miller G. Markov chain monte carlo calculations allowing parallel processing—II. Open Numer Methods J 2011; x: xx-xx.

[3] Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain monte carlo in practice. USA: Chapman and Hall 1996.

[4] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculation by fast computing machines. J Chem Phys 1953; 21(6): 1087-92.

[5] Barker AA. Monte carlo calculations of the radial distribution functions for a proton-electron plasma. Aust J Phys 1965; 18: 119-33.

[6] Marin J-M, Robert CP. Bayesian core--a practical approach to computational bayesian statistics. USA: Springer 2007.

[7] Miller G, Martz H, Little T, Bertelli L. Bayesian hypothesis testing--use in interpretation of measurements. Health Phys 2008; 94(3): 248-54.

Received: April 14, 2011 Revised: June 27, 2011 Accepted: August 10, 2011