Journal of Statistical Physics, Vol. 122, No. 3, February 2006 (© 2006)
DOI: 10.1007/s10955-005-8016-8

A Theory on Flat Histogram Monte Carlo Algorithms

Faming Liang[1]

Received November 12, 2004; accepted August 10, 2005; Published Online: January 20, 2006

The flat histogram Monte Carlo algorithms have been successfully used in many problems in scientific computing.However, there is no a rigorous theory for the convergence of the algorithms. In this paper, a modified flat histogram algorithm is presented and its convergence is studied. The convergence of the multicanonical algorithm and the Wang-Landau algorithm is argued based on their relations to the modified algorithm. The numerical results show the superiority of the modified algorithm to the multicanonical and Wang-Landau algorithms.

KEY WORDS: Convergence; Contour Monte Carlo; Multicanonical; Wang-Landau Algorithm.
PACS number: 02.70.Tt, 02.50.Ng

1. INTRODUCTION

The flat histogram Monte Carlo algorithms[1,2] have been successfully used in many problems in scientific computing, e.g., spin glasses simulations[2], protein folding[3,4], Lennard-Jone glass[5], and others. However, there is no a rigorous theory for the convergence of the algorithms. In this paper, we present a modified flat histogram algorithm and study its convergence. We then argue for the convergence of the multicanonical algorithm[1] and the Wang-Landau (WL) algorithm[2] based on their relations to the modified algorithm, although the argument may not be very strict.

The remaining part of this paper is organized as follows. In Section 2, we present the modified algorithm. In Section 3, we discuss the convergence of the flat histogram algorithms. In Section 4, we present the numerical results of the modified

[1] Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA; e-mail: fliang@stat.tamu.edu

algorithm on a simple example with a comparison study with the multicanonical and WL algorithms.

2. CONTOUR MONTE CARLO

Suppose that we are working on the following Boltzmann distribution,

$$f(\boldsymbol{x}) = \frac{1}{Z} \exp\{-H(\boldsymbol{x})/\tau\}, \quad \boldsymbol{x} \in \mathcal{X}, \tag{1}$$

where $Z = \int_{\mathcal{X}} \exp\{-H(\boldsymbol{x})/\tau\} d\boldsymbol{x}$ is the partition function, τ is the temperature, $\mathcal{X}$ is the phase space, and $H(\boldsymbol{x})$ is the energy function. For a complex system, the energy function has often many local energy minima separated by high energy barriers. In the following, we present a modified flat histogram algorithm, the so-called contour Monte Carlo (CMC), for simulation from $f(\boldsymbol{x})$.

Suppose that the phase space has been partitioned according to a chosen parameterization into m disjoint subregions. For example, the partition can be made according to the microcanonical energy, and the m disjoint subregions are as follows: $E_1 = \{\boldsymbol{x} : H(\boldsymbol{x}) \le h_1\}$, $E_2 = \{\boldsymbol{x} : h_1 < H(\boldsymbol{x}) \le h_2\}$, ..., $E_{m-1} = \{\boldsymbol{x} : h_{m-2} < H(\boldsymbol{x}) \le h_{m-1}\}$, and $E_m = \{\boldsymbol{x} : H(\boldsymbol{x}) > h_{m-1}\}$, where $h_1, \ldots, h_{m-1}$ are $m-1$ specified real numbers. Let $\psi(\boldsymbol{x})$ be a non-negative function defined on the phase space with $0 < \int_{\mathcal{X}} \psi(\boldsymbol{x}) d\boldsymbol{x} < \infty$, and $g_i = \int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x}$. If $\mathcal{X}$ is finite and $\psi(\boldsymbol{x}) \equiv 1$, then g_i is the number of configurations contained in the subregion E_i. If $\psi(\boldsymbol{x}) = \exp\{-H(\boldsymbol{x})/\tau\}$, then g_i is the partition function of the truncated distribution of $f(\boldsymbol{x})$ on the subregion E_i. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$, where π_i denotes the desired (relative) sampling frequency of the subregion E_i, $0 < \pi_i < 1$, and $\sum_{i=1}^m \pi_i = 1$. The CMC simulation can be described as follows. Let $\widehat{g}_i^{(t)}$ denote the working estimate of g_i at iteration t, $\widehat{\boldsymbol{g}}^{(t)} = (\widehat{g}_1^{(t)}, \ldots, \widehat{g}_m^{(t)})$, and

$$\widehat{f}^{(t)}(\boldsymbol{x}) = \frac{1}{Z_t} \sum_{i=1}^{m} \frac{\psi(\boldsymbol{x})}{\widehat{g}_i^{(t)}} I(\boldsymbol{x} \in E_i), \tag{2}$$

denote the working density at iteration t, where Z_t is the partition function of $\widehat{f}^{(t)}(\boldsymbol{x})$, and $I(\cdot)$ is the indicator function. In the working density (2), if the phase space is partitioned according to the energy function, each subregion E_i associates with a different weight $\widehat{g}_i$. In this sense, the algorithm is called contour Monte Carlo. Let $\boldsymbol{x}_k^{(t)}$, $k = 1, \ldots, L$, denote samples drawn from $\widehat{f}^{(t)}(\boldsymbol{x})$, and $\boldsymbol{y}^{(t)} = (y_1^{(t)}, \ldots, y_m^{(t)})$ denote the realized sampling frequency of the m subregions by the L samples, where $y_i^{(t)} = \frac{1}{L} \sum_{k=1}^{L} I(\boldsymbol{x}_k^{(t)} \in E_i)$. One iteration of CMC consists of the following two steps. Note that in the initial iteration, we have $t = 0$ and $\log \widehat{g}_1^{(0)} = \cdots = \log \widehat{g}_m^{(0)} = 0$.

(a) **(Sampling)** Draw sample $\boldsymbol{x}_k^{(t)}$, $k = 1, \ldots, L$, from the working density $\widehat{f}^{(t)}(\boldsymbol{x})$ as defined in (2).

(b) **(Weight updating)** Update the working estimates of g_i's in the following manner,

$$\log \widehat{g}_i^{(t+1)} = \log \widehat{g}_i^{(t)} + \delta_t (y_i^{(t)} - \pi_i), \quad i = 1, \ldots, m \tag{3}$$

where δ_t is called the weight modification factor.

In CMC, $\{\delta_t : t = 0, 1, 2, \ldots\}$, which will be simplified to $\{\delta_t\}$ later, is a sequence of positive, non-increasing numbers satisfying the condition

$$\sum_{t=0}^{\infty} \delta_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \delta_t^2 < \infty. \tag{4}$$

The algorithm iterates till δ_t has been a very small number, say, a number less than 10^{-8}.

A brief explanation of condition (4) is as follows. The first condition is a necessary condition for the convergence of $\log \widehat{g}_i^{(t)}$ to $\log(g_i)$. If $\sum_{t=0}^{\infty} \delta_t = C < \infty$, then, as follows from (3),

$$\sum_{t=0}^{\infty} |\log \widehat{g}_i^{(t+1)} - \log \widehat{g}_i^{(t)}| \le \sum_{t=0}^{\infty} \delta_t |y_i^{(t)} - \pi_i| \le \sum_{t=0}^{\infty} \delta_t = C < \infty,$$

where the second inequality holds because both $y_i^{(t)}$ and π_i lie in the interval $[0, 1]$. Thus, the value of $\log \widehat{g}_i^{(t)}$ does not reach $\log(g_i)$ as $t \to \infty$ if, for example, the initial point $\log \widehat{g}_i^{(0)}$ is sufficiently far away from $\log(g_i)$. On the other hand, neither should the number δ_t be too large. Otherwise, the random errors will prevent convergence. It turns out that the condition $\sum_{t=0}^{\infty} \delta_t^2 < \infty$ asymptotically damps the effect of the random errors introduced by $y_i^{(t)}$'s. When it holds, we have $\delta_t |y_i^{(t)} - \pi_i| \le \delta_t \to 0$, as $t \to \infty$.

There are many ways to choose the sequence $\{\delta_t\}$ to satisfy the condition (4). For example, in this paper we set

$$\delta_t = \frac{\kappa}{\max(\kappa, t)}, \quad t = 0, 1, 2, \ldots, \tag{5}$$

for some specified value of κ. A large value of κ will allow the sampler to reach all subregions very quickly even for a large system. The choice of κ should not affect the convergence of the algorithm. However, a good choice of κ may make the algorithm perform more stably and save computation time. This is illustrated in Section IV by a numerical example.

At each iteration, the samples $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_L^{(t)}$ can be drawn using a conventional MCMC algorithm, say, the Metropolis-Hastings (MH)[6] algorithm, starting with

the last sample obtained in the preceding iteration. The sample size L should be chosen such that the equalities

$$E\left(y_i^{(t)}\right) = S_i^{(t)} \Big/ S^{(t)}, \quad i = 1, \ldots, m, \tag{6}$$

hold for large t, where $S_i^{(t)} = \int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x} / \widehat{g}_i^{(t)}$ and $S^{(t)} = \sum_{j=1}^{m} S_j^{(t)}$. According to the standard MCMC theory, we know that (6) holds when L is large, even if the samples $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_L^{(t)}$ are correlated. In practice, L should be chosen according to the autocorrelation time of the sample sequence. The longer autocorrelation time the sequence has, the large value of L we should choose. For example, we can set L to a multiple of the integrated autocorrelation time of the sample sequence. Fortunately, when t is large, CMC is reduced to an approximately "random walk" in the space of subregions (if each subregion is regarded as a "point") with the visiting frequency to each of the subregions being proportional to the corresponding desired sampling frequency. Hence, L is not required to be very large at this stage. When t is small, we do not care too much whether the equalities in (6) hold or not, as the work at this stage is just to prepare "initial" values for the latter stages. The self-adjusting mechanism of CMC, i.e., adjusting the value of $1/\widehat{g}_i^{(t)}$ and thus, the sampling probability of the subregion E_i in the adverse direction of the realized sampling frequency of the subregion E_i, warrants the success of the "initialization" process. Note that the multicanonical algorithm and the Wang-Landau algorithm have the same self-adjusting mechanism. In our experience, a value of L between 10 and 100 is appropriate for most problems. Our numerical results reported in Section 4 indicate that the choice of L does not affect the accuracy of the CMC estimates significantly.

If the desired sampling distribution is chosen to be a uniform distribution, i.e., $\pi_1 = \cdots = \pi_m = 1/m$, the weight updating step can be simplified to

$$\log \widehat{g}_i^{(t+1)} = \log \widehat{g}_i^{(t)} + \delta_t y_i^{(t)}, \quad i = 1, \ldots, m, \tag{7}$$

as adding to or subtracting from $\log \widehat{g}_i^{(t)}$'s a constant will not change $\widehat{f}^{(t)}(\boldsymbol{x})$ and the simulation process.

The desired sampling distribution $\boldsymbol{\pi}$ can be chosen to bias sampling to the low energy region, although in this paper the algorithm is demonstrated by setting it to be the uniform distribution. As shown in Ref.[7,8], biasing sampling to the low energy region often results in a run with improved ergodicity. This makes CMC attractive for use in hard optimization problems. This also makes CMC more flexible than the multicanonical and WL algorithms. In the latter two algorithms, each subregion has to be sampled equally.

At last, we would like to mention that the generalization of $\psi(\boldsymbol{x})$ from a constant to a general non-negative function is useful. First, it extends the application of the flat histogram algorithms to continuum systems. Second, it leads to a great

deal of applications of the flat histogram algorithms in model selection[9], and it is of interest to statisticians.

3. CONVERGENCE PROPERTY OF FLAT HISTOGRAM ALGORITHMS

CMC falls into the category of stochastic approximation algorithms[10−−12]. Based on the theory of stochastic approximation, we provide a proof for its convergence. The proof is simple, which is just to verify the conditions given in Blum[11]. (See Appendix for the details.) In CMC, the phase space partition can be made blindly. This may lead to that some of the subregions are empty. In this paper, a subregion E_i is called empty if $\int_{E_i} \psi(\boldsymbol{x})d\boldsymbol{x} = 0$. As $t \to \infty$, we have

$$\log \widehat{g}_i^{(t)} \to \begin{cases} c + \log\left(\int_{E_i} \psi(\boldsymbol{x})d\boldsymbol{x}\right) - \log(\pi_i + \nu), & E_i \neq \emptyset, \\ -\infty, & E_i = \emptyset, \end{cases} \tag{8}$$

where $\nu = \sum_{j\in\{i:E_i=\emptyset\}} \pi_j/(m - m_0)$ and m_0 is the number of empty subregions, and c is a constant which can be determined with some extra information on the system. For example, in simulation from an Ising model, if $\psi(\boldsymbol{x}) = 1$ then c can be determined with the information that $\sum_{i=1}^m g_i$ is known. See Appendix for more discussions on the issue.

After convergence, CMC will be reduced to a random walk in the space of subregions with the visiting frequency to each of the subregions being proportional to the corresponding desired sampling frequency. Based on this observation, the convergence of CMC can be diagnosed as follows. Let $\widehat{\pi}_i$ denote the sampling frequency of the subregion E_i in the simulation. We define

$$\epsilon_f(E_i) = \begin{cases} \dfrac{\widehat{\pi}_i - (\pi_i + \nu)}{\pi_i + \nu} \times 100\%, & E_i \neq \emptyset, \\ 0, & E_i = \emptyset, \end{cases}$$

for $i = 1, \ldots, m$, for assessing the difference of the realized sampling distribution from the desired one. If $\max_{i=1}^m \epsilon_f(E_i)$ is large, say, greater than 10 percents, the convergence of the simulation should be questioned. In this case, CMC should be re-run with more iterations or a large value of κ. The number of iterations and κ should be tuned such that $\max_{i=1}^m \epsilon_f(E_i)$ falls into a satisfactory interval, say, $\pm 5\%$.

CMC is closely related to the multicanonical algorithm[1,13]. If $\mathcal{X}$ is finite, $\psi(\boldsymbol{x}) \equiv 1$, $\pi_1 = \cdots = \pi_m = \frac{1}{m}$, and the weight updating scheme is modified appropriately, CMC can be reduced to the multicanonical algorithm. In the multicanonical algorithm, the initial estimate of g_i can be obtained via a short simulation from the distribution $f(\boldsymbol{x})$, and then the estimate $\log(\widehat{g}_i^{(t)})$ evolves with iterations

as follows:

$$\log\left(\widehat{g}_i^{(t+1)}\right) = c + \log\left(\widehat{g}_i^{(t)}\right) + \log\left(\widehat{\pi}^{(t)}(E_i) + \alpha_i\right), \quad i = 1, \ldots, m \tag{9}$$

where the constant c is introduced to ensure that $\log(\widehat{g}_i^{(t+1)})$ is an estimate of $\log(g_i)$, $\widehat{\pi}^{(t)}(E_i)$ is the (unnormalized) sampling frequency of the subregion E_i at iteration t, and $\alpha_1, \ldots, \alpha_m$ are small positive numbers which serve as "prior counts" to smooth out the estimates $\widehat{g}_i$'s. Since the multicanonical algorithm assumes that the simulation from the working density $\widehat{f}^{(t)}(\boldsymbol{x})$ has reached equilibrium before proceeding to the weight updating step, the relation

$$\widehat{\pi}^{(t)}(E_i) \propto \frac{\int_{E_i} \psi(\boldsymbol{x})dx}{\widehat{g}_i^{(t)}} = \frac{g_i}{\widehat{g}_i^{(t)}}, \quad i = 1, \ldots, m, \tag{10}$$

should approximately hold. Substituting (10) into (9), we have

$$\log\left(\widehat{g}_i^{(t+1)}\right) = c + \log(g_i), \quad i = 1, \ldots, m, \tag{11}$$

which implies the validity of the algorithm for estimating g_i's (up to a multiplicative constant).

CMC is also closely related to the WL algorithm[2]. If $\mathcal{X}$ is finite, $\psi(\boldsymbol{x}) \equiv 1$, $\pi_1 = \cdots = \pi_m = 1/m$, each $\boldsymbol{x}_1^{(t)}$ is drawn with only one MH step (i.e., $L = 1$), and the sequence $\{\delta_t\}$ is specified appropriately, then CMC can be reduced to the WL algorithm. The WL simulation consists of a number of stages. Each stage associates with a different value of δ_t. Let s denotes the total number of stages, $t_{(i)}$'s denote the change points of stages, and $\delta_{(i)}$ denote the common value of δ_t at stage i. Then δ_t can be expressed as a piecewise constant function of t,

$$\delta_t = \begin{cases} \delta_{(1)}, \ t_{(0)} \le t < t_{(1)}, \\ \delta_{(2)}, \ t_{(1)} \le t < t_{(2)}, \\ \cdots \\ \delta_{(s)}, \ t_{(s-1)} \le t < t_{(s)}, \end{cases} \tag{12}$$

where $t_{(0)} = 1$, and $t_{(s)}$ is the total number of iterations of the run. In WL, $\{\delta_{(i)}, i = 1, \ldots, s\}$ is usually set to a geometrically decreasing sequence, say,

$$\delta_{(i+1)} = \frac{1}{2}\delta_{(i)}, \quad i = 1, 2, \ldots, \tag{13}$$

as suggested in Wang and Landau[2]. The $t_{(i)}$'s are chosen such that the samples generated between the iterations $t_{(i-1)}$ and $t_{(i)}$ form a flat histogram in the space of subregions. Clearly, the sequence $\{\delta_t\}$ defined in (12) and (13) violates the condition (4). Due to the violation, the above theory established for the CMC algorithm is not directly applicable to the WL algorithm. However, we note that the CMC theory is still relevant to the WL algorithm in some sense if the sequence

Table I. The unnormalized mass function of the 10-state distribution

x	1	2	3	4	5	6	7	8	9	10
$P(x)$	1	100	2	1	3	3	1	200	2	1

$\{\delta_t\}$ is modified as follows. Let $L'_{(i)} = t_{(i)} - t_{(i-1)}$ denote the number of iterations performed at stage i. If the equality $L'_{(i+1)}/L'_{(i)} = \delta_{(i)}/\delta_{(i+1)}$ holds for all stages $i = 1, 2, \ldots$; that is, $L'_{(i)}$ increases geometrically with the rate $\delta_{(i)}/\delta_{(i+1)}$, then the condition (4) is satisfied. If we further assume that $\boldsymbol{x}_1^{(t)}$ is an exact sample drawn from $\widehat{f}^{(t)}(\boldsymbol{x})$, then the CMC theory is completely applicable to the WL algorithm. However, the assumption that $\boldsymbol{x}_1^{(t)}$ is an exact sample of $\widehat{\pi}^{(t)}(\boldsymbol{x})$ is questionable. We note that under this assumption, a theoretical study has been done by Zhou and Bhatt[14], where an analytic proof is established for the the convergence of the WL algorithm. They show the convergence of the density-of-state estimates by showing the convergence of the histograms, while our argument is made for the density-of-state estimates directly.

4. NUMERICAL RESULTS

In the following example we compare the efficiency of the multicanonical, WL and CMC algorithms. The distribution of the example consists of 10 states with the unnormalized mass function $P(x)$ as specified in Table 1. It has two modes which are well separated by low mass states.

The state space was partitioned according to the mass function into the following five subregions: $E_1 = \{8\}$, $E_2 = \{2\}$, $E_3 = \{5, 6\}$, $E_4 = \{3, 9\}$, and $E_5 = \{1, 4, 7, 10\}$. The three algorithms, multicanonical, Wang-Landau and CMC, are all applied to estimate the density of states for this example. The true value of $\boldsymbol{g}$ is $\boldsymbol{g} = (1, 1, 2, 2, 4)$, which is equal to the number of states in the respective subregions. The following statistic is defined to assess the accuracy of the estimate,

$$\epsilon_g = \sqrt{\sum_{i=1}^{m} (\widehat{g}_i - g_i)^2 / g_i},$$

where $\widehat{\boldsymbol{g}} = (\widehat{g}_1, \ldots, \widehat{g}_m)$ denotes an estimate of $\boldsymbol{g}$ obtained using above algorithms. Since the energy function is evaluated once in each MH step, we measure the CPU cost of each run by the total number of MH steps or, equivalently, the total number of energy evaluations performed in the run. In the following, we denote by N the total number of energy evaluations of a run, and denote by n the total number of iterations.

To evaluate the performance of the multicanonical algorithm, the algorithm was run with different values of $N = 50000$, 100000, 150000, 200000, and

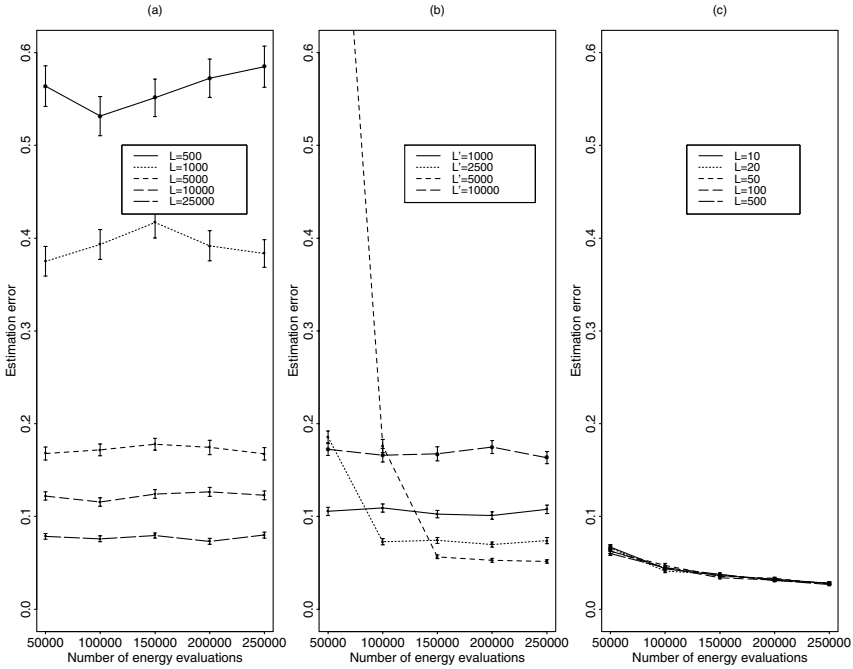

Fig. 1. Comparison of (a) the multicanonical algorithm, (b) the WL algorithm, and (c) the CMC algorithm. The vertical segments show the ±one-standard deviation of $\bar{\epsilon}_g$'s.

250000. For each value of N, different choices of L and n are tried. The choices of L include $L = 500$, 1000, 5000, 10000, and 25000. The n is set accordingly such that the equality $N = nL$ holds for all runs. For each choice of (N, L), the algorithm was run for 100 times independently. In each run, g_i's are estimated in Eq. (9) with the prior counts $\alpha_1 = \cdots = \alpha_5 = 0.1$, and ϵ_g is calculated for assessing the accuracy of the estimates. Let $\bar{\epsilon}_g$ denote the average of ϵ_g's over the 100 runs. Figure 1(a) shows the $\bar{\epsilon}_g$'s resulted from the 25 choices of (N, L). From the plot, it is easy to see that for the multicanonical algorithm, the accuracy of the estimates is mainly determined by L. The larger value of L, the higher accuracy of the estimates. For a given value of L, the accuracy of the estimates can not be significantly improved by increasing the value of N. This observation may seem a little bit surprise to us. But it can be understood from Eqs. (10) and (11) as follows. These two equations imply the Markovian property of the estimates $\widehat{\boldsymbol{g}}^{(t)}$'s. Conditional on $\widehat{\boldsymbol{g}}^{(t)}$, the next iteration estimate $\widehat{\boldsymbol{g}}^{(t+1)}$ is independent of the previous estimates $\widehat{\boldsymbol{g}}^{(t-1)}, \ldots, \widehat{\boldsymbol{g}}^{(1)}$. Suppose that the estimate $\widehat{\boldsymbol{g}}^{(t)}$ has reached a certain accuracy. In order to improve further its accuracy, the only way is to increase the value of L in the later iterations of the run. Otherwise, the accuracy of the follow-

on estimates, $\widehat{\boldsymbol{g}}^{(t+1)}$, $\widehat{\boldsymbol{g}}^{(t+2)}$, ..., will be limited by the accuracy of $\widehat{\pi}(E_i)$'s, which is determined by L, the number of MH steps performed in each iteration. Hence, in the multicanonical algorithm, the estimates can only reach a certain accuracy limited by L, the estimates will not be improved with further more iterations if L does not increase with iterations.

To evaluate the performance of the WL algorithm, the algorithm was run with the same choices of N as for the multicanonical algorithm. For simplicity, in each run we set $\delta_{(1)} = 1$, $\delta_{(i+1)} = 0.5\delta_{(i)}$, and $L'_{(i)}$'s to a constant which has been large enough such that a flat histogram can be formed during each stage of the simulation. Let L' denote the constant. The choices of L' we tried include $L' = 1000$, 2500, 5000, and 10000. Given N and L', n is set to the quotient N/L'. For each choice of (N, L'), the WL algorithm was run for 100 times independently. Figure 1(b) shows $\bar{\epsilon}_g$'s resulted from the 20 choices of (N, L'). This plot shows that the accuracy of the WL estimates are mainly determined by L'. For a given value of L', once the estimate has reached a certain accuracy, it can not be improved significantly with further more iterations. This is consistent with the finding of Yan and Pablo[15]. This phenomenon can be understood from the design of the algorithm. If $L'_{(1)} = \cdots = L'_{(s)} = \cdots = L'$ is finite, and $\{\delta_{(i)}\}$ is a sequence decreasing geometrically, then the tail sum $\sum_{t=T+1}^{\infty} \delta_t < \infty$ for any value of T. Hence, The large number of configurations generated towards the end of the simulation make only a small contribution to the estimates. For the WL algorithm, we have also tried the choice of $\{\delta_t\}$ that $L'_{(i)}$ increases geometrically with the rate $\delta_{(i)}/\delta_{(i+1)}$. The tendency that the estimates can be improved continuously has been observed. However, this leads to an explosion of the total number of iterations required by the simulation.

To evaluate the performance of the CMC algorithm, the algorithm was also run with the same choices of N as for the above two algorithms. For each value N, the following choices of L are considered, $L = 10$, 20, 50, 100, and 500. For each choice of (N, L), n is set to the quotient $n = N/L$, κ is set to 10, and CMC was run for 100 times independently. Figure 1(c) shows the $\bar{\epsilon}_g$'s resulted from the 25 choices of (N, L). This plot indicates that CMC produces more accurate estimates than the multicanonical and WL algorithms for all choices of (N, L) given above. More importantly, these estimates can be improved continuously by increasing the value of N. The plot also indicates that for a given value of N, the choice of L has not much effect on the estimates.

As mentioned before, the convergence of CMC can be diagnosed by examining the values of $\epsilon_f(E_i)$'s at the end of the simulation. Figure 2 summarizes the values of $\epsilon_f(E_i)$'s calculated for the 100 runs with the setting $N = 100000$, $L = 10$, and $\kappa = 10$. The five box-plots correspond to the five subregions respectively. All $\epsilon_f(E_i)$'s lie in the interval of $\pm 3\%$. This indicates that all of the 100 runs have converged. The plots for the other choices of (N, L) are similar.

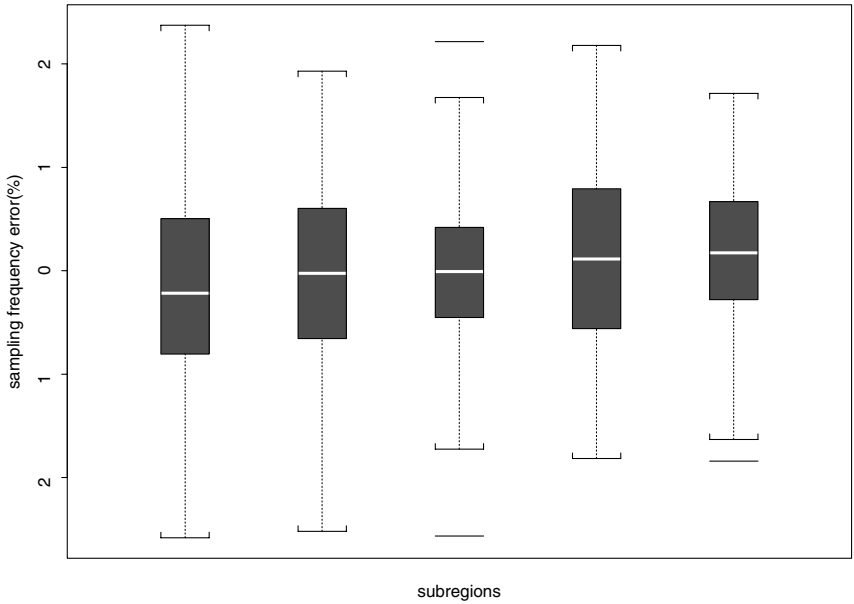

Fig. 2. Box-plots of $\epsilon_f(E_i)$'s obtained in the 100 CMC runs with the setting $N = 100000$, $L = 10$, and $\kappa = 10$.

Later, CMC was re-run with different values of $\kappa = 5$, 10 and 20. The choices of (N, L) are the same as described above. For each choice of (N, L, κ), the algorithm was also run for 100 times independently. The numerical results of these runs are summarized in figure 3. The plots show that κ affects the accuracy of the estimates significantly. For this example, the choice $\kappa = 5$ outperforms the other two choices $\kappa = 10$ and $\kappa = 20$. The choice $\kappa = 2$ has also been tried, but the results are inferior to that produced with $\kappa = 5$. On the choice of κ, our suggestion is as follows. Try a number of choices for κ and choose the minimum one which produces satisfactory $\epsilon_f(E_i)$'s.

APPENDIX A

The appendix is organized as follows. In Section A, we describe a theorem for the convergence of the CMC algorithm. In Section B, we present a generalized version of Blum's theorem[11], which shows the convergence for a multidimensional stochastic approximation algorithm. In Section C, we prove the theorem given in Section A based on the generalized Blum's theorem.

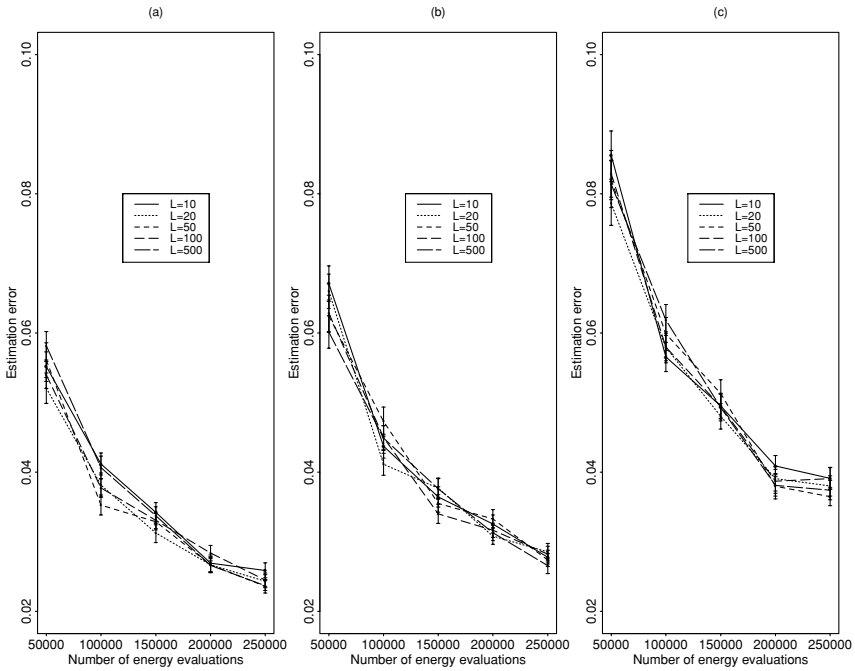

Fig. 3. Computational results of CMC with different choices of κ. (a) $\kappa = 5$; (b) $\kappa = 10$; (c) $\kappa = 20$.

A.1. A Convergence Theorem for the CMC Algorithm

Without loss of generality, we only show the convergence presented in Eq. (8) for the case that all subregions are non-empty or, equivalently, $\nu = 0$. Extension to the case $\nu \neq 0$ is trivial, because changing step (b) of the CMC algorithm to (b') (given below) will not change the simulation process.

(b') set $\log \widehat{g}_i^{(t+1)} = \log \widehat{g}_i^{(t)} + \delta_t \big[y_i^{(t)} - (\pi_i + \nu) \big]$ for all non-empty subregions.

Note that the empty subregions will never be visited during the simulation. To simplify notations, in the following we denote $\log \widehat{g}_i^{(t)}$ by $\theta_i^{(t)}$, and denote $(\log \widehat{g}_1^{(t)}, \ldots, \log \widehat{g}_m^{(t)})$ by $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_m^{(t)})$.

Theorem 1. *Let $E_1, \ldots, E_m$ be a partition of the phase space $\mathcal{X}$, $\psi(\boldsymbol{x})$ be a non-negative function defined on $\mathcal{X}$ with $0 < \int_{\mathcal{X}} \psi(\boldsymbol{x}) d\boldsymbol{x} < \infty$, and $\theta^{(0)}$ be an arbitrary m-vector. Let $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_L^{(t)}$ be samples drawn from the distribution*

$$\widehat{f}^{(t)}(\boldsymbol{x}) \propto \sum_{i=1}^{m} \frac{\psi(\boldsymbol{x})}{e^{\theta_i^{(t)}}} I(\boldsymbol{x} \in E_i), \quad t = 0, 1, 2, \ldots. \tag{14}$$

Define $\boldsymbol{y}^{(t)} = (y_1^{(t)}, \ldots, y_m^{(t)})$, *where*

$$y_i^{(t)} = \frac{1}{L} \sum_{k=1}^{L} I\left(\boldsymbol{x}_k^{(t)} \in E_i\right)$$

Let $\theta^{(t)}$ *be iterated in the following manner,*

$$\theta^{(t+1)} = \theta^{(t)} + \delta_t(\boldsymbol{y}^{(t)} - \boldsymbol{\pi}), \tag{15}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ *be a m-vector with* $0 < \pi_i < 1$ *and* $\sum_{i=1}^{m} \pi_i = 1$. *If* $\{\delta_t\}$ *satisfies the following condition*

$$\sum_{t=0}^{\infty} \delta_t = \infty, \quad \sum_{t=0}^{\infty} \delta_t^2 < \infty.$$

then

$$P\left\{\lim_{t\to\infty} \theta_i^{(t)} = c + \log\left(\int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x}\right) - \log(\pi_i)\right\} = 1, \quad i = 1, \ldots, m, \tag{16}$$

where c is an arbitrary constant.

Since the distribution $\widehat{f^{(t)}}(\boldsymbol{x})$ defined in (14) is invariant with respect to a shift transformation of $\theta^{(t)}$; that is, $(\theta_1^{(t)}, \ldots, \theta_m^{(t)})$ and $(\theta_1^{(t)} + a, \ldots, \theta_m^{(t)} + a)$ (a is an arbitrary constant) result in the same distribution $\widehat{f^{(t)}}(\boldsymbol{x})$, the constant c in (16) can not be determined with the samples drawn from $\widehat{f^{(t)}}(\boldsymbol{x})$ only. To determine the value of c, we need extra information on the system. For example, in simulation from an Ising model of size $l \times l$, if $\psi(\boldsymbol{x}) = 1$, then c can be determined as

$$c = \log\left(\sum_i \pi_i e^{\theta_i^{(t)}}\right) - l^2 \log(2),$$

with the information that $\sum_i g_i = \sum_i \int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x} = 2^{l^2}$. We note that the multicanonical and WL algorithms both suffer from the same problem, i.e., g_i's can only be determined up to a multiplicative constant.

A.2. BLUM'S Theorem on the Convergence of a Multivariate Stochastic Approximation Algorithm

Lemma 2. *Let* z_t *be a sequence of integrable random variables which satisfy the condition*

$$\sum_{t=1}^{\infty} E\{E(z_{t+1} - z_t | z_1, \ldots, z_t)^+\} < \infty,$$

and are bounded below uniformly in t. Then z_t converges almost surely to a random variable.

In the lemma, $(z)^+$ is defined by $(z)^+ = \frac{1}{2}(z + |z|)$. This lemma corresponds to the corollary stated in Section 3 of Blum's paper [11].

Theorem 3 is a slight generation of Theorem 1 of Blum's paper[11]. Blum shows the convergence of a stochastic approximation sequence to a single point of the parameter space, while Theorem 3 shows the convergence of a stochastic approximation sequence to a subset of the parameter space by adding one more condition (A.3) as stated below. However, the proofs are almost the same. The condition (A.3) is a necessary condition for (20) and (23). It is obvious that a single-point solution set satisfies the condition (A.3) automatically.

Let $\theta = (\theta_1, \ldots, \theta_m)$ be a point in Θ, where Θ is a real m-dimensional vector space spanned by m orthogonal unit vectors. Let $\boldsymbol{y}_\theta = (y_{\theta,1}, \ldots, y_{\theta,m})$ be a vector of m random variables with corresponding families of distributions $\{F_\theta^{(1)}\}$, ..., $\{F_\theta^{(m)}\}$, each depending on m real variables $\theta = (\theta_1, \ldots, \theta_m)$. Let $\mu_i(\theta) = \int_{-\infty}^{\infty} y dF_\theta^{(i)}$, $i = 1, \ldots, m$ be the corresponding expectation functions. Here it is assumed that the distributions $\{F_\theta^{(i)}\}$ and the expectation functions $\mu_i(\theta)$ are unknown; however, it is possible to make an observation on the random vector $\boldsymbol{y}_\theta$ for any choice of $\theta \in \Theta$. Let $\boldsymbol{\mu}(\theta) = (\mu_1(\theta), \ldots, \mu_m(\theta))$. Let $\omega(\theta)$ be a real-valued function defined on Θ and possessing continuous partial derivatives of the first and second order, the vector of first partial derivatives will be denoted by $\boldsymbol{d}(\theta)$ and the matrix of second partial derivatives by $\boldsymbol{A}(\theta)$; that is,

$$\boldsymbol{d}(\theta) = \left(\frac{\partial \omega}{\partial \theta_i}\right)\Bigg|_\theta, \quad \boldsymbol{A}(\theta) = \left(\frac{\partial^2 \omega}{\partial \theta_i \partial \theta_j}\right)\Bigg|_\theta.$$

Then, for any real number δ, we have by Taylor's theorem

$$\omega(\theta + \delta(\boldsymbol{y}_\theta - \boldsymbol{\pi})) = \omega(\theta) + \delta \boldsymbol{d}^T(\theta)(\boldsymbol{y}_\theta - \boldsymbol{\pi}) + \frac{1}{2}\delta^2[(\boldsymbol{y}_\theta - \boldsymbol{\pi})^T + \boldsymbol{A}(\theta\xi\delta(\boldsymbol{y}_\theta - \boldsymbol{\pi}))(\boldsymbol{y}_\theta - \boldsymbol{\pi})],$$

where ξ is a real number with $0 \le \xi \le 1$, $\boldsymbol{\pi}$ is a known m-vector, and $\boldsymbol{b}^T$ denotes the transpose of the vector $\boldsymbol{b}$. Consequently we may take expectations on both sides to obtain

$$E\omega(\theta + \delta(\boldsymbol{y}_\theta - \boldsymbol{\pi})) = \omega(\theta) + \delta \boldsymbol{d}^T(\theta)(\boldsymbol{\mu}(\theta) - \boldsymbol{\pi}) + \frac{1}{2}\delta^2 E[(\boldsymbol{y}_\theta - \boldsymbol{\pi})^T \boldsymbol{A}(\theta + \xi\delta(\boldsymbol{y}_\theta - \boldsymbol{\pi}))(\boldsymbol{y}_\theta - \boldsymbol{\pi})]. \quad (17)$$

To simplify writing we employ the following notations:

$$u(\theta) = \boldsymbol{d}^T(\theta)(\boldsymbol{\mu}(\theta) - \boldsymbol{\pi}), \quad v_\delta(\theta) = E[(\boldsymbol{y}_\theta - \boldsymbol{\pi})^T \boldsymbol{A}(\theta + \xi\delta(\boldsymbol{y}_\theta - \boldsymbol{\pi}))(\boldsymbol{y}_\theta - \boldsymbol{\pi})].$$

Let $\|\boldsymbol{b}\|$ denotes the norm of the vector $\boldsymbol{b}$. Consider now the following conditions:

$A.1$: $\sum_{t=1}^{\infty} \delta_t = \infty, \quad \sum_{t=1}^{\infty} \delta_t^2 < \infty$.
$A.2$: $\omega(\theta) \geq 0$.
$A.3$: There exists a nonempty set $\Theta_0 = \{\theta : \mu_i(\theta) = \pi_i, i = 1, \ldots, m; \theta \in \Theta\}$, $\omega(\theta)$ is a constant over the set Θ_0, and the interior set of Θ_0 is empty.
$A.4$: $\sup_{\{\theta \notin \Theta_0, \inf_{\widetilde{\theta} \in \Theta_0} \|\theta - \widetilde{\theta}\| > \epsilon\}} u(\theta) < 0$ for every $\epsilon > 0$.
$A.5$: $\inf_{\{\theta \notin \Theta_0, \inf_{\widetilde{\theta} \in \Theta_0} \|\theta - \widetilde{\theta}\| > \epsilon\}} |\omega(\theta) - \omega(\widetilde{\theta})| > 0$ for every $\epsilon > 0$.
$A.6$: $v_\delta(\theta) \leq V < \infty$ for every number δ.

Theorem 3. *Assume the following conditions are satisfied: (i) the sequence $\{\delta_t\}$ satisfies $A.1$; (ii) there exists a real-valued function $\omega(\theta)$ with continuous first and second partial derivatives satisfying $A.2, \ldots, A.6$; (iii) the sequence $\theta^{(t)}$ iterates in the following manner,*

$$\theta^{(t+1)} = \theta^{(t)} + \delta_t(\boldsymbol{y}_{\theta^{(t)}} - \boldsymbol{\pi}). \tag{18}$$

Then $\theta^{(t)}$ converges almost surely to the set Θ_0; that is,

$$P\Big\{\lim_{t \to \infty} \theta^{(t)} \in \Theta_0\Big\} = 1.$$

Proof: To simplify notations, we let $z_t = \omega(\theta^{(t)})$ and $z_0 = \omega(\widetilde{\theta})$ for every $\widetilde{\theta} \in \Theta_0$, $u_t = u(\theta^{(t)})$, and $v_t = v_{\delta_t}(\theta^{(t)})$. From equation (17) one obtains

$$E(z_{t+1}|z_1, \ldots, z_t) = z_t + \delta_t E(u_t|z_1, \ldots, z_t) + \frac{\delta_t^2}{2} E(v_t|z_1, \ldots, z_t) \quad \text{a.s.} \tag{19}$$

Since $\boldsymbol{\mu}(\widetilde{\theta}) - \boldsymbol{\pi} = \mathbf{0}$ for every $\widetilde{\theta} \in \Theta_0$, we have, by virtue of conditions $A.2$–$A.6$,

$$E(u_t|z_1, \ldots, z_t) \leq 0 \quad \text{a.s.}, \qquad E(v_t|z_1, \ldots, z_t) \leq V \quad \text{a.s.}, \tag{20}$$

both for all t. Hence,

$$E(z_{t+1} - z_t|z_1, \ldots, z_t) \leq \frac{1}{2}\delta_t^2 V \quad \text{a.s.} \tag{21}$$

By conditions $A.1$ and $A.2$ and Lemma 2 one obtains

$$P(z_t \text{ converges}) = 1. \tag{22}$$

Taking expectations on both sides of (19) and iterating, we have

$$E(z_{t+1}) = z_1 + \sum_{j=1}^{t} \delta_j E(u_j) + \frac{1}{2}\sum_{j=1}^{t} \delta_j^2 E(v_j).$$

From the condition $A.2$ and from the property of expectation of conditional expectation it follows that

$$E(z_t) \geq 0, \quad E(u_t) \leq 0, \quad E(v_t) \leq V \quad (n = 1, 2, \ldots).$$

Since V is non-negative and the series $\sum_{t=1}^{\infty} \delta_t^2$ converges, the nonpositive term series $\sum_{t=1}^{\infty} \delta_t E(u_t)$ also converges. By virtue of the fact that $\sum_{t=1}^{\infty} \delta_t = \infty$ we have

$$\lim_{t\to\infty} \sup E(u_t) = 0, \quad \lim_{t\to\infty} \inf E(|u_t|) = 0.$$

Let $\{t_r\}$ be an infinite sequence of integers such that

$$\lim_{r\to\infty} E(|u_{t_r}|) = 0.$$

Then u_{t_r} converges to zero in probability due to Markov's inequality and there exists a further subsequence, say $\{u_{m_r}\}$ such that

$$P\Big\{\lim_{r\to\infty} u_{m_r} = 0\Big\} = 1.$$

From condition $A.4$ it follows that $P\{\lim_{r\to\infty} \theta^{(m_r)} \in \Theta_0\} = 1$. Since z_t is a continuous function of $\theta^{(t)}$ it follows from (22) and condition (A.3) that

$$P\Big\{\lim_{t\to\infty} z_t = z_0\Big\} = 1. \tag{23}$$

Now consider a sample sequence $\{\theta^{(t)}\}$ such that for the corresponding sequence $\{z_t\}$ one has $\lim_{t\to\infty} z_t = z_0$. From condition $A.5$ it is obvious that for such a sequence one must have

$$P\Big\{\lim_{t\to\infty} \theta^{(t)} \in \Theta_0\Big\} = 1;$$

otherwise it will lead to a contradiction of $A.5$. The proof is completed. □

A.3. Proof of Theorem 1

Proof: To prove this theorem, we only need to verify the conditions of Theorem 3.

Let $F_{\theta^{(t)}}^{(i)}$ be a Bernoulli distribution with the success probability

$$p_{\theta^{(t)}}^{(i)} = \frac{\int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x} / e^{\theta_i^{(t)}}}{\sum_{k=1}^{m} \Big[\int_{E_k} \psi(\boldsymbol{x}) d\boldsymbol{x} / e^{\theta_k^{(t)}}\Big]},$$

for $i = 1, \ldots, m$. For simplicity, we define $S_i = \int_{E_i} \psi(\boldsymbol{x}) d\boldsymbol{x} / e^{\theta_i^{(t)}}$ and $S = \sum_{k=1}^{m} S_k$. Assuming that $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_L^{(t)}$ are drawn from $\widehat{f}^{(t)}(\boldsymbol{x})$ using a MCMC

algorithm, we have

$$\mu_i(\theta) = Ey_i^{(t)} = p_{\theta^{(t)}}^{(i)} = \frac{S_i}{S}, \quad i = 1, \ldots, m. \tag{24}$$

Here L can be a reasonably large number.

To simplify notations, in the following we will denote $\theta^{(t)}$ by θ and $\theta_i^{(t)}$ by θ_i, by dropping the superscript t.

(i) By the assumption for $\{\delta_t\}$, condition $A.1$ is satisfied.

(ii) Let $\omega(\theta) = \frac{1}{2}\sum_{k=1}^{m}(\frac{S_k}{S} - \pi_k)^2$. Therefore, the condition $A.2$ is satisfied. As shown below, $\omega(\theta)$ has continuous partial derivatives of the first and second order.

(iii) Solving the system of equations formed by (24), we have

$$\Theta_0 = \left\{(\theta_1, \ldots, \theta_m) : \theta_i = c + \log\left(\int_{E_i} \psi(\boldsymbol{x})d\boldsymbol{x}\right) - \log(\pi_i), i = 1, \ldots, m; \theta \in \Theta\right\},$$

where $c = -\log(S)$. It is obvious that Θ_0 is nonempty and $\omega(\widetilde{\theta}) = 0$ for every $\widetilde{\theta} \in \Theta_0$. The one-to-one correspondence of c and S implies the completeness of Θ_0; that is, Θ_0 has included all θ's which solve the equation $\boldsymbol{\mu}(\theta) = \boldsymbol{\pi}$.

The set Θ_0 forms a line in the space Θ, as it contains only one free parameter c. Therefore, the interior set of Θ_0 is empty. Condition $A.3$ is satisfied.

(iv) To show condition $A.4$ is satisfied, we have the following calculations.

$$\begin{aligned}
\frac{\partial S}{\partial \theta_i} &= \frac{\partial S_i}{\partial \theta_i} = -S_i, \\
\frac{\partial S_i}{\partial \theta_j} &= \frac{\partial S_j}{\partial \theta_i} = 0, \\
\frac{\partial\left(\frac{S_i}{S}\right)}{\partial \theta_i} &= -\frac{S_i}{S}(1 - \frac{S_i}{S}), \\
\frac{\partial\left(\frac{S_i}{S}\right)}{\partial \theta_j} &= \frac{\partial\left(\frac{S_j}{S}\right)}{\partial \theta_i} = \frac{S_i S_j}{S^2},
\end{aligned} \tag{25}$$

for $i, j = 1, \ldots, m$ and $i \neq j$.

$$\begin{aligned}\frac{\partial \omega(\theta)}{\partial \theta_i} &= \frac{1}{2}\sum_{k=1}^{m}\frac{\partial\left(\frac{S_k}{S}-\pi_k\right)^2}{\partial\theta_i}\\ &= \sum_{j\neq i}\left(\frac{S_j}{S}-\pi_j\right)\frac{S_iS_j}{S^2}-\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\left(1-\frac{S_i}{S}\right)\\ &= \sum_{j=1}^{m}\left(\frac{S_j}{S}-\pi_j\right)\frac{S_iS_j}{S^2}-\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\\ &= M\frac{S_i}{S}-\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S},\end{aligned} \tag{26}$$

for $i = 1, \ldots, m$, where one defines $M = \sum_{j=1}^{m}(\frac{S_j}{S}-\pi_j)\frac{S_j}{S}$. Thus,

$$\begin{aligned}u(\theta) = \boldsymbol{d}^T(\theta)(\boldsymbol{\mu}(\theta)-\boldsymbol{\pi}) &= \sum_{i=1}^{m}\left[M\frac{S_i}{S}-\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\right]\left(\frac{S_i}{S}-\pi_i\right)\\ &= -\left\{\sum_{i=1}^{m}\left(\frac{S_i}{S}-\pi_i\right)^2\frac{S_i}{S}-\left[\sum_{i=1}^{m}\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\right]^2\right\}\\ &= -\sigma_\eta^2 \leq 0,\end{aligned}$$

where σ_η^2 denotes the variance of the discrete distribution defined in the following table,

State (η)	$\frac{S_1}{S}-\pi_1$	$\cdots$	$\frac{S_m}{S}-\pi_m$
Prob.	$\frac{S_1}{S}$	$\cdots$	$\frac{S_m}{S}$

If $\theta \in \Theta_0$, $u(\theta) = 0$; otherwise $u(\theta) < 0$. Therefore, condition $A.4$ is satisfied.

(v) By the construction of $\omega(\theta)$ and the completeness of Θ_0, it is obvious that condition $A.5$ is satisfied.

(vi) Based on (26) and (25), we have the following calculations:

$$\begin{aligned}\frac{\partial M}{\partial\theta_i} &= \sum_{k\neq i}\left[2\frac{S_iS_k^2}{S^3}-\pi_k\frac{S_iS_k}{S^2}\right]-2\frac{S_i^2}{S^2}\left(1-\frac{S_i}{S}\right)+\pi_i\frac{S_i}{S}\left(1-\frac{S_i}{S}\right)\\ &= \frac{S_i}{S}\left[\sum_{k=1}^{m}\frac{S_k^2}{S^2}+M-2\frac{S_i}{S}+\pi_i\right],\end{aligned} \tag{27}$$

$$\frac{\partial^2\omega(\theta)}{\partial\theta_i^2} = \frac{\partial\left[M\frac{S_i}{S} - \left(\frac{S_i}{S} - \pi_i\right)\frac{S_i}{S}\right]}{\partial\theta_i}$$

$$= -M\frac{S_i}{S}\left(1 - \frac{S_i}{S}\right) + \frac{S_i}{S}\frac{\partial M}{\partial\theta_i} + 2\frac{S_i^2}{S^2}\left(1 - \frac{S_i}{S}\right) - \pi_i\frac{S_i}{S}\left(1 - \frac{S_i}{S}\right)$$

$$= \frac{S_i^2}{S^2}\left[\sum_{k=1}^{m}\frac{S_k^2}{S^2} + 2M - 4\frac{S_i}{S} + 2\pi_i + 2\right] - \frac{S_i}{S}(M + \pi_i), \tag{28}$$

and

$$\frac{\partial^2\omega(\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial\left[M\frac{S_i}{S} - \left(\frac{S_i}{S} - \pi_i\right)\frac{S_i}{S}\right]}{\partial\theta_j}$$

$$= \frac{S_iS_j}{S^2}M + \frac{S_i}{S}\frac{\partial M}{\partial\theta_j} - \left[2\frac{S_i^2S_j}{S^3} - \pi_i\frac{S_iS_j}{S^2}\right] \tag{29}$$

$$= \frac{S_iS_j}{S^2}\left[\sum_{k=1}^{m}\frac{S_k^2}{S^2} + 2M - 2\frac{S_i}{S} - 2\frac{S_j}{S} + \pi_i + \pi_j\right].$$

Since $0 \le S_i/S \le 1$, $0 \le \pi_i \le 1$ and $|M| \le 1$, both $|\frac{\partial^2\omega(\theta)}{\partial\theta_i^2}|$ and $|\frac{\partial^2\omega(\theta)}{\partial\theta_i\partial\theta_j}|$ are bounded above by a constant. Therefore, $v_\delta(\theta)$ is bounded above by a constant for every number δ. Hence, the condition $A.6$ is satisfied.

The proof is completed. □

ACKNOWLEDGMENTS

The author thank the referees for their critial comments which have led to great improvements of this paper. The author's research was partially supported by grants from the National Science Foundation (DMS-0405748) and the National Cancer Institute (CA104620).

REFERENCES

1. B.A. Berg and T. Neuhaus, *Phys. Lett. B.* **267**:291 (1991); *Phys. Rev. Lett.* **68:**9 (1992); B.A. Berg and T. Celik, *Phys. Rev. Lett.* **69:**2292 (1992).
2. F. Wang and D.P. Landau, *Phys. Rev. Lett.* **86:**2050 (2001); *Phys. Rev. E.* **64:**56101 (2001).
3. U.H.E. Hansmann and Y. Okamoto, *Journal of Computational Chemistry* **18:**920 (1997).
4. N. Rathore and J.J. de Pablo, *Journal of Chemical Physics* **116:**7225 (2002).
5. R. Faller and J.J. de Pablo, *Journal of Chemical Physics* **119:**4405 (2003).

6. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *J. Chem. Phys.* **21:**1087 (1953); W.K. Hastings, *Biometrika* **57:**97 (1970).
7. B. Hesselbo and R. B. Stinchcombe, *Phys. Rev. Lett.* **74**:2151 (1995).
8. F. Liang, Phys. Rev. E **69**:66701 (2004).
9. F. Liang, unpublished.
10. H. Robbins and S. Monro, Ann. Math. Stat. **22**:400 (1951).
11. J. R. Blum, Ann. Math. Stat. **25**:737 (1954).
12. A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York: Springer-Verlag (1990).
13. B. A. Berg, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, eds. K. T. Fang, F. J. Hickernell, and H. Niederreiter, New York: Springer, pp. 168 (2000).
14. C. Zhou and R. N. Bhatt, cond-mat/0306711 (2003).
15. Q. Yan and J. J. de Pablo, *Phys. Rev. Lett.* **90**:035701 (2003).