Lecture Notes for STAT546: Computational Statistics

—Lecture 10: Monte Carlo

Faming Liang

Purdue University

September 11, 2024

Multicanonical Monte Carlo

Multicanonical Monte Carlo (Berg and Neuhaus, 1991, 1992) seeks to draw samples in an ensemble where each configuration with energy $u = H(x)$ is assigned a weight

$$w_m(u) \propto \frac{1}{g(u)} = e^{-S(u)},$$

where $S(u) = \log(g(u))$ is called the microcanonical entropy. A simulation with this weight function will yield a uniform distribution of energy:

$$P_m(u) \propto g(u) w_m(u) = \text{constant},$$

and lead to a free random walk in the space of energy. This allows the sampler to escape any energy barriers, and to explore any regions of the sample space even for those with small $g(u)$'s. The samples generated in the simulation will form a flat histogram in the space of energy, hence, the multicanonical algorithm is also called a flat histogram Monte Carlo algorithm.

Suppose that the energy function U only takes values on a finite set $\{u_1, \ldots, u_m\}$. Let $x_1, \ldots, x_N$ denote the MCMC samples drawn from $f_T(x)$, and let $N_T(i) = \#\{x_j : H(x_j) = u_i\}$ denote the number of samples with energy u_i. As $N \to \infty$,

$$N_T(i)/N \approx \frac{1}{Z(T)} g(u_i) e^{-u_i/T}, \quad i = 1, \ldots, m,$$

then the spectral density can be estimated by

$$\widehat{g}(u_i) = \frac{N_T(i) e^{u_i/T}}{\sum_{j=1}^{m} N_T(j) e^{u_j/T}}, \quad i = 1, \ldots, m.$$

In practice, the temperature T should be sufficiently high such that each value of the energy can be visited with reasonably large frequency.

Given the initial spectral density estimate, the multicanonical algorithm iterates between the following two steps:

1. Run a MCMC sampler, say, the MH algorithm, sufficiently long according to the current weighting function

$$w_m^{(t)}(x) \propto \frac{1}{\widehat{g}_t(H(x))}, \tag{1}$$

   where t indexes the stages of the simulation.

2. Update the spectral density estimate by

$$\log\left(\widehat{g}_{t+1}(u_i)\right) = c + \log\left(\widehat{g}_t(u_i)\right) + \log\left(\widehat{\pi}_t(i) + \alpha_i\right), \quad i = 1, \ldots, m, \tag{2}$$

   where the constant c is introduced to ensure that $\log(\widehat{g}_{t+1})$ is an estimate of $\log(g)$, and $\widehat{\pi}_t(i)$ is the relative sampling frequency of the energy u_i at stage t, and $\alpha_1, \ldots, \alpha_m$ are small positive constants which serve as "prior values" to smooth out the estimate $\widehat{g}$.

Since the number of iterations performed in step 1 is sufficiently large, it is reasonable to assume that the simulation has reached equilibrium, and thus

$$\widehat{\pi}_t(i) \propto \frac{g(u_i)}{\widehat{g}_t(u_i)}, \quad i = 1, \ldots, m. \tag{3}$$

Substituting (3) into (2), then

$$\log\left(\widehat{g}_{t+1}(u_i)\right) = c + \log(g(u_i)), \quad i = 1, \ldots, m, \tag{4}$$

which implies the validity of the algorithm for estimating $g(u)$ (up to a multicanonical constant). On the other hand, in equation (4), the independence of $\widehat{g}_{t+1}(u)$ on the previous estimate $\widehat{g}_t(u)$ implies that the spectral density estimate can only reach limited accuracy, which is determined by the length of the simulation performed in step 1. After certain stage, increasing the number of stages will not improve accuracy of the spectral density estimate.

$1/k$-ensemble sampling

Similar to multicanonical Monte Carlo, $1/k$-ensemble sampling (Hesselbo and Stinchcombe, 1995) seeks to draw samples in an ensemble where each configuration x with energy $u = H(x)$ is assigned a weight

$$w_{1/k}(u) \propto \frac{1}{k(u)},$$

where $k(u) = \sum_{u' \leq u} g(u')$, i.e., the cumulative spectral density function of the distribution. Hence, $1/k$-ensemble sampling will produce the following distribution of energy:

$$P_{1/k}(u) \propto \frac{g(u)}{k(u)} = \frac{d \log k(u)}{du}.$$

Since, in many physical systems, $k(u)$ is a rapidly increasing function of u, $\log k(u) \approx \log g(u)$ for a wide range of u, and the simulation will lead to an approximately random walk in the space of entropy. Recall that $S(u) = \log g(u)$ is called the microcanonical entropy of the system. Comparing to multicanonical Monte Carlo, $1/k$-ensemble sampling is designed to spend more time in exploring low energy regions, hence, it is potentially more suitable for optimization problems. Improvement over multicanonical Monte Carlo has been observed in ergodicity of the simulations for the Ising model and travel salesman problems (Hesselbo and Stinchcombe, 1995).

Wang-Landau algorithm

Like multicanonical Monte Carlo, the Wang-Landau algorithm (Wang and Landau, 2001) seeks to draw samples in an ensemble where each configuration with energy u is assigned a weight

$$w_m(u) \propto \frac{1}{g(u)},$$

where $g(u)$ is the spectral density. The difference between the two algorithms is on their learning procedures for the spectral density. Suppose that the sample space $\mathcal{X}$ is finite and the energy function $H(x)$ takes values on a finite set $\{u_1, \ldots, u_m\}$. The simulation of the Wang-Landau algorithm consists of several stages. In the first stage, it starts with an initial setting of $\widehat{g}(u_1), \ldots, \widehat{g}(u_m)$, say $\widehat{g}(u_1) = \ldots = \widehat{g}(u_m) = 1$, and a random sample x_0 drawn from $\mathcal{X}$, and then iterates between the following steps:

The Wang-Landau algorithm

- Simulate a sample x by a single Metropolis update which admits the invariant distribution $\widehat{f}(x) \propto 1/\widehat{g}(H(x))$.
- Set $\widehat{g}(u_i) \leftarrow \widehat{g}(u_i)\delta^{I(H(x)=u_i)}$ for $i = 1, \ldots, m$, where δ is a modification factor greater than 1 and $I(\cdot)$ is the indicator function.

The algorithm iterates till a flat histogram has been produced in the space of energy. A histogram is usually considered to be flat if the sampling frequency of each u_i is not less than 80% of the average sampling frequency. Once this condition is satisfied, the estimates $\widehat{g}(u_i)$'s and the current sample x are passed on to the next stage as initial values, the modification factor is reduced to a smaller value according to a specified scheme, say, $\delta \leftarrow \sqrt{\delta}$, and the sampler collector is resumed. The next stage simulation is then started, continuing until the new histogram is flat again. The process is repeated until δ is very close to 1, say, $\log(\delta) < 10^{-8}$.

Stochastic approximation Monte Carlo

Consider the problem of sampling from the distribution (**??**). Suppose the sample space $\mathcal{X}$ has been partitioned according to a function $\lambda(x)$ into m subregions, $E_1 = \{x : \lambda(x) \leq u_1\}$, $E_2 = \{x : u_1 < \lambda(x) \leq u_2\}, \ldots$, $E_{m-1} = \{x : u_{m-2} < \lambda(x) \leq u_{m-1}\}$, $E_m = \{x : \lambda(x) \geq u_{m-1}\}$, where $-\infty < u_1 < \cdots < u_{m-1} < \infty$. Here $\lambda(\cdot)$ can be any function of x, such as a component of x, the energy function $H(x)$, etc. Let $\psi(x)$ be a non-negative function with $0 < \int_{\mathcal{X}} \psi(x) dx < \infty$, which is called the working function of SAMC. In practice, one often sets $\psi(x) = \exp(-H(x)/\tau)$. Let $g_i = \int_{E_i} \psi(x) dx$ for $i = 1, \ldots, m$, and $\boldsymbol{g} = (g_1, \ldots, g_m)$. The subregion E_i is called an empty subregion if $g_i = 0$.

An inappropriate specification of the cutoff points u_i's may result in some empty subregions. Technically, SAMC allows for the existence of empty subregions in simulations. To present the idea clearly, we temporarily assume that all subregions are nonempty; that is, assuming $g_i > 0$ for all $i = 1, \ldots, m$. SAMC seeks to sample from the distribution

$$f_g(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x)}{g_i} I(x \in E_i), \tag{5}$$

where π_i's are pre-specified frequency values such that $\pi_i > 0$ for all i and $\sum_{i=1}^{m} \pi_i = 1$. It is easy to see, if $g_1, \ldots, g_m$ are known, sampling from $f_g(x)$ will result in a "random walk" in the space of subregions with each subregion being visited with a frequency proportional to π_i. The distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ is called the desired sampling distribution of the subregions.

Let $\theta_t^{(i)}$ denote the estimate of $\log(g_i/\pi_i)$ obtained at iteration t, and let $\theta_t = (\theta_t^{(1)}, \ldots, \theta_t^{(m)})$. Since m is finite, the partition $E_1, \ldots, E_m$ is fixed, and $0 < \int_{\mathcal{X}} \psi(x) < \infty$, there must exist a number $\epsilon > 0$ such that

$$\epsilon < \min_i \log\left(\frac{g_i}{\pi_i}\right) < \max_i \log\left(\frac{g_i}{\pi_i}\right) < \frac{1}{\epsilon},$$

which implies that θ_t can be restricted to taking values on a compact set. Henceforth, this compact set will be denoted by Θ. In practice, Θ can be set to a huge set, say, $\Theta = [-10^{100}, 10^{100}]^m$. As a practical matter, this is equivalent to set $\Theta = \mathbb{R}^m$. Otherwise, if one assumes $\Theta = \mathbb{R}^m$, a varying truncation version of this algorithm can be considered as in Liang (2009e). For both mathematical and practical simplicity, Θ is restricted to be compact in this chapter.

Let $\{\gamma_t\}$ denote the gain factor sequence, which satisfies the condition (A_1):

(A_1) The sequence $\{\gamma_t\}$ is positive and non-increasing, and satisfies the conditions:

$$(i) \quad \overline{\lim_{t\to\infty}} |\gamma_t^{-1} - \gamma_{t+1}^{-1}| < \infty, \; (ii) \quad \sum_{t=1}^{\infty} \gamma_t = \infty, \; (iii) \quad \sum_{t=1}^{\infty} \gamma_t^{\eta} < \infty, \tag{6}$$

for some $\eta \in (1, 2]$.

In practice, one often sets

$$\gamma_t = \frac{t_0}{\max\{t_0, t^{\xi}\}}, \quad t = 0, 1, 2, \ldots, \tag{7}$$

for some pre-specified values $t_0 > 1$ and $\frac{1}{2} < \xi \leq 1$.

SAMC starts with a random sample x_0 generated in the space $\mathcal{X}$ and an initial estimate $\theta_0 = (\theta_0^{(1)}, \ldots, \theta_0^{(m)}) = (0, \ldots, 0)$, then iterates between the following steps:

The SAMC algorithm

(a) *Sampling*: Simulate a sample x_{t+1} by a single MH update which admits the following distribution as its invariant distribution:

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{\exp(\theta_t^{(i)})} I(x \in E_i). \tag{8}$$

(a.1) Generate y in the sample space $\mathcal{X}$ according to a proposal distribution $q(x_t, y)$.

(a.2) Calculate the ratio

$$r = e^{\theta_t^{(J(x_t))} - \theta_t^{(J(y))}} \frac{\psi(y) q(y, x_t)}{\psi(x_t) q(x_t, y)}.$$

(a.3) Accept the proposal with probability $\min(1, r)$. If it is accepted, set $x_{t+1} = y$; otherwise, set $x_{t+1} = x_t$.

(b) *Weight updating*: For $i = 1, \ldots, m$, set

$$\theta^{(i)}_{t+\frac{1}{2}} = \theta^{(i)}_t + \gamma_{t+1}\left(I_{\{x_{t+1}\in E_i\}} - \pi_i\right). \tag{9}$$

If $\theta_{t+\frac{1}{2}} \in \Theta$, set $\theta_{t+1} = \theta_{t+\frac{1}{2}}$; otherwise, set $\theta_{t+1} = \theta_{t+\frac{1}{2}} + \boldsymbol{c}^*$, where $\boldsymbol{c}^* = (c^*, \ldots, c^*)$ can be any constant vector satisfying the condition $\theta_{t+\frac{1}{2}} + \boldsymbol{c}^* \in \Theta$.

Remark 1:

In the weight updating step, $\theta_{t+\frac{1}{2}}$ is adjusted by adding a constant vector $\boldsymbol{c}^*$ when $\theta_{t+\frac{1}{2}} \notin \Theta$. The validity of this adjustment is simply due to the fact that $f_{\theta_t}(x)$ is invariant with respect to a location shift of θ_t.

The compactness constraint on θ_t should only apply to the components of θ for which the corresponding subregions are unempty. In practice, one can place an indicator on each subregion, indicating whether or not the subregion has been visited or is known to be unempty. The compactness check for $\theta_{t+\frac{1}{2}}$ should be done only for the components for which the corresponding subregions have been visited or are known to be unempty.

Remark 2:

The explanation for the condition (A_1) can be found in advanced books on stochastic approximation, see, e.g., Nevel′son and Has′minskiĭ (1973). The condition $\sum_{t=1}^{\infty} \gamma_t = \infty$ is necessary for the convergence of θ_t. Otherwise, it follows from step (b) that, assuming the adjustment of $\theta_{t+\frac{1}{2}}$ did not occur,

$$\sum_{t=0}^{\infty} |\theta_{t+1}^{(i)} - \theta_t^{(i)}| \leq \sum_{t=0}^{\infty} \gamma_{t+1} |I_{\{x^{(t+1)} \in E_i\}} - \pi_i| \leq \sum_{t=0}^{\infty} \gamma_{t+1} < \infty.$$

Thus, θ_t cannot reach $\log(\boldsymbol{g}/\boldsymbol{\pi})$ if, for example, the initial point θ_0 is sufficiently far away from $\log(\boldsymbol{g}/\boldsymbol{\pi})$. On the other hand, γ_t can not be too large. An overly large γ_t will prevent convergence. It turns out that the third condition in (6) asymptotically damps the effect of random errors introduced by new samples. When it holds, we have $\gamma_{t+1} |I_{\{x_{t+1} \in E_i\}} - \pi_i| \leq \gamma_{t+1} \to 0$ as $t \to \infty$.

Remark 3:

A remarkable feature of the SAMC algorithm is that it possesses the self-adjusting mechanism: If a proposed move is rejected at an iteration, then the weight of the subregion that the current sample belongs to will be adjusted to a larger value, and the total rejection probability of the next iteration will be reduced. This mechanism enables the algorithm to escape from local energy minima very quickly. The SAMC algorithm represents a significant advance for simulations of complex systems for which the energy landscape is rugged.

Convergence

Theorem 1

Assume (A_1) and the drift condition (B_2) hold. Then, as $t \to \infty$,

$$\theta_t^{(i)} \to \theta_*^{(i)} = \begin{cases} C + \log(\int_{E_i} \psi(x)dx) - \log(\pi_i + \nu), & \text{if } E_i \neq \emptyset, \\ -\infty. & \text{if } E_i = \emptyset, \end{cases} \tag{10}$$

where C is an arbitrary constant, $\nu = \sum_{j \in \{i : E_i = \emptyset\}} \pi_j/(m - m_0)$, and m_0 is the number of empty subregions.

To ease verification of the drift condition, one may assume further that the proposal distribution $q(x, y)$ satisfies the following local positive condition:

(A_2) For every $x \in \mathcal{X}$, there exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that

$$\|x - y\| \leq \epsilon_1 \Longrightarrow q(x, y) \geq \epsilon_2, \tag{11}$$

where $\|x - y\|$ denotes a certain distance measure between x and y.

Convergence Rate

The following theorem concerns the convergence rate of θ_t, which gives a L^2 upper bound for the mean squared error of θ_t.

Theorem 2

Assume the gain factor sequence is chosen in (7) and the drift condition (B_2) *holds. Then there exists a constant* λ *such that*

$$E\|\theta_t - \theta_*\|^2 \leq \lambda\gamma_t,$$

where $\theta_* = (\theta_*^{(1)}, \ldots, \theta_*^{(m)})$ *is as specified in (10).*

Monte Carlo Integration

In addition to estimating the normalizing constants g_i's, SAMC can be conveniently used for Monte Carlo integration, estimating the expectation $E_f \rho(x) = \int_{\mathcal{X}} \rho(x) f(x)$ for an integrable function $\rho(x)$. Let (x_1, θ_1), ..., (x_n, θ_n) denote the samples generated by SAMC during the first n iterations. Let $y_1, \ldots, y_{n'}$ denote the distinct samples among $x_1, \ldots, x_n$. Generate a random variable/vector Y such that

$$P(Y = y_i) = \frac{\sum_{t=1}^{n} \exp\{\theta_t^{(J(x_t))}\} I(x_t = y_i)}{\sum_{t=1}^{n} \exp\{\theta_t^{(J(x_t))}\}}, \quad i = 1, \ldots, n', \quad (12)$$

where $I(\cdot)$ is the indicator function. Under the assumptions (A_1), (A_2) and the compactness of $\mathcal{X}$, Liang (2009b) showed that Y is asymptotically distributed as $f(\cdot)$.

Theorem 3

Assume (A_1) and the drift condition (B_2) hold. For a set of samples generated by SAMC, the random variable/vector Y generated in (12) is asymptotically distributed as $f(\cdot)$.

This theorem implies that for an integrable function $\rho(x)$, $E_f\rho(x)$ can be estimated by

$$\widehat{E_f\rho(x)} = \frac{\sum_{t=1}^{n} \exp\{\theta_t^{(J(x_t))}\} h(x_t)}{\sum_{t=1}^{n} \exp\{\theta_t^{(J(x_t))}\}}. \tag{13}$$

As $n \to \infty$, $\widehat{E_f\rho(x)} \to E_f\rho(x)$ for the same reason that the usual importance sampling estimate converges (Geweke, 1989).

Some Implementation Issues

For an effective implementation of SAMC, several issues need to be considered.

- *Sample space partition.* This can be done according to our goal and the complexity of the given problem. For example, if we aim to minimize the energy function, the sample space can be partitioned according to the energy function. The maximum energy difference in each subregion should be bounded by a reasonable number, say, 2, which ensures that the local MH moves within the same subregion have a reasonable acceptance rate. Note that within the same subregion, sampling from the working density (8) is reduced to sampling from $\psi(x)$. If our goal is model selection, then the sample space can be partitioned according to the index of models, as illustrated in §**??**.

- *The desired sampling distribution.* If our goal is to estimate $\boldsymbol{g}$, then we may set the desired distribution to be uniform. However, if our goal is optimization, then we may set the desired sampling distribution biased to low energy regions. As illustrated by Hesselbo and Stinchcombe (1995) and Liang (2005b), biasing sampling to low energy regions often improves the ergodicity of the simulation.
- *The choice of the gain factor sequence and the number of iterations.* To estimate $\boldsymbol{g}$, γ_t should be very close to 0 at the end of simulations. Otherwise, the resulting estimates will have a large variation. Under the setting of (7), the speed of γ_t going to zero is controlled by ξ and t_0. In practice, one often fixes ξ to 1 and choose t_0 according to the complexity of the problem. The more complex the problem, the larger the value of t_0 one should choose. A large t_0 will force the sampler to reach all subregions quickly, even in the presence of multiple local energy minima.

Example I

The distribution consists of 10 states with the unnormalized mass function $P(x)$ being given below. It has two modes which are well separated by low mass states.

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

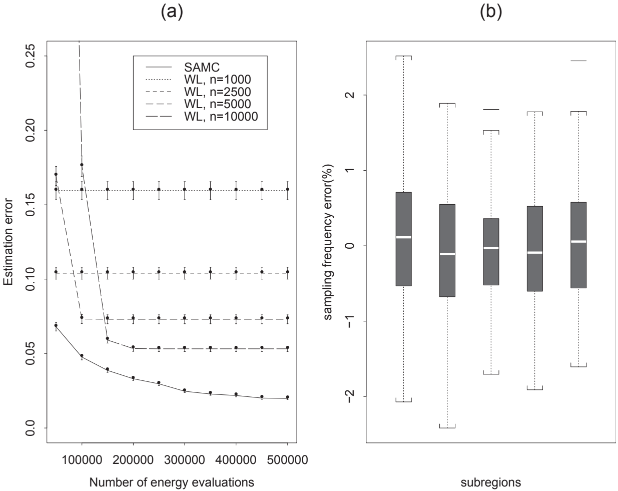

Figure 1: Comparison of the WL and SAMC algorithms. (a) Average $\epsilon_e(t)$ curves obtained by SAMC and WL. The vertical bars show the ±one-standard-deviation of the average of the estimates. (b) Box-plots of $\{\epsilon_f(E_i)\}$ obtained in 100 runs of SAMC. (Liang, Liu and Carroll, 2007)

Table 1: Comparison of SAMC and MH for the 10-state example. (Liang, 2009b)

Algorithm	Bias ($\times 10^{-3}$)	Standard Error ($\times 10^{-3}$)	CPU time (seconds)
SAMC	-0.528	1.513	0.38
MH	-3.685	4.634	0.20

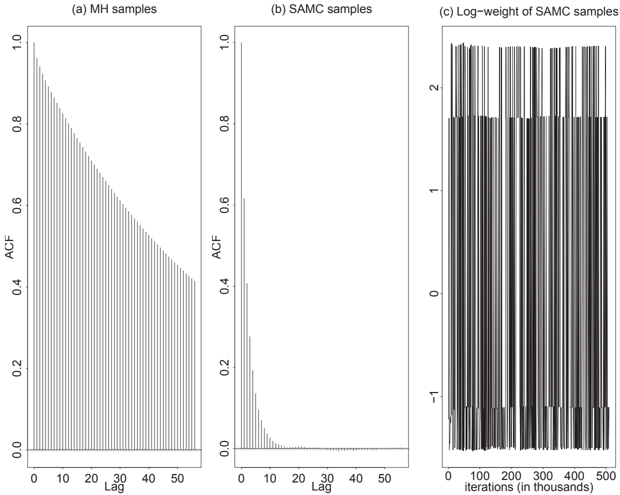

Figure 2: Computational results for the 10-state example. (a) Autocorrelation plot of the MH samples. (b) Autocorrelation plot of the SAMC samples. (c) Log-weights of the SAMC samples. (Liang, 2009b)

Example II

This problem is to sample from a multimodal distribution defined by $f(\boldsymbol{x}) \propto \exp\{-H(\boldsymbol{x})\}$, where $\boldsymbol{x} = (x_1, x_2) \in [1.1, 1.1]^2$ and

$$\begin{aligned} H(\boldsymbol{x}) = & - \left\{x_1 \sin(20x_2) + x_2 \sin(20x_1)\right\}^2 \cosh\left\{\sin(10x_1)x_1\right\} \\ & - \left\{x_1 \cos(10x_2) - x_2 \sin(10x_1)\right\}^2 \cosh\left\{\cos(20x_2)x_2\right\}. \end{aligned}$$

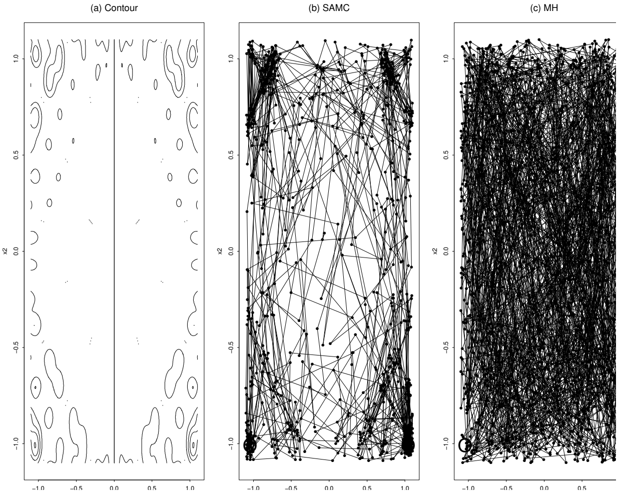

Figure 3: (a) Contour of $H(\boldsymbol{x})$. (b) Sample path of SAMC. (c) Sample path of MH at the temperature $T = 5$. (Liang, Liu and Carroll, 2007)