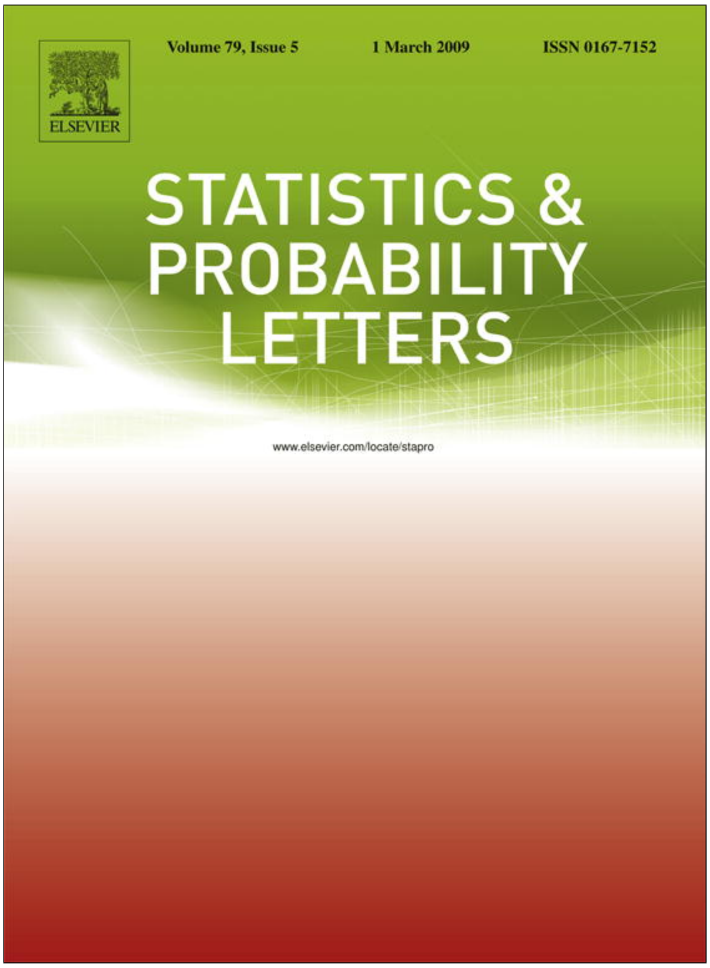

Volume 79, Issue 5
1 March 2009
ISSN 0167-7152
ELSEVIER
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Statistics and Probability Letters 79 (2009) 581–587

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the use of stochastic approximation Monte Carlo for Monte Carlo integration

Faming Liang*

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

ARTICLE INFO

Article history:
Received 18 April 2008
Received in revised form 29 September 2008
Accepted 1 October 2008
Available online 19 October 2008

ABSTRACT

The stochastic approximation Monte Carlo (SAMC) algorithm has recently been proposed as a dynamic optimization algorithm in the literature. In this paper, we show in theory that the samples generated by SAMC can be used for Monte Carlo integration via a dynamically weighted estimator by calling some results from the literature of nonhomogeneous Markov chains. Our numerical results indicate that SAMC can yield significant savings over conventional Monte Carlo algorithms, such as the Metropolis-Hastings algorithm, for the problems for which the energy landscape is rugged.

1. Introduction

As known by many researchers, the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984) are prone to get trapped into local energy minima in simulations from a system for which the energy landscape is rugged. In terms of physics, the negative of the logarithmic density/mass function is called the energy function of the system. To overcome the local trap problem, advanced Monte Carlo algorithms have been proposed, such as parallel tempering (Geyer, 1991), simulated tempering (Marinari and Parisi, 1992), evolutionary Monte Carlo (Liang and Wong, 2001), dynamic weighting (Wong and Liang, 1997), multicanonical sampling (Berg and Neuhaus, 1991), the Wang–Landau algorithm (Wang and Landau, 2001), equi-energy sampler (Kou et al., 2006), SAMC (Liang et al., 2007; Atchadé and Liu, 2007), among others.

Among the aforementioned algorithms, SAMC is a very sophisticated one in both theory and applications. The basic idea of SAMC stems from the Wang–Landau algorithm and can be briefly explained as follows. Let

$$f(x) = c\psi(x), \quad x \in \mathcal{X}, \tag{1}$$

denote the target probability density/mass function, where $\mathcal{X}$ is the sample space and c is an unknown constant. Let $E_1, \ldots, E_m$ denote a partition of $\mathcal{X}$, and let $\omega_i = \int_{E_i} \psi(x)dx$ for $i = 1, \ldots, m$. SAMC seeks to draw samples from the trial distribution

$$f_\omega(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x)}{\omega_i} I(x \in E_i), \tag{2}$$

where $I(\cdot)$ is an indicator function, π_i's are pre-specified constants such that $\pi_i > 0$ for all i and $\sum_{i=1}^m \pi_i = 1$. In Liang et al. (2007), $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ is called the desired sampling distribution of the subregions. If $\omega_1, \ldots, \omega_m$ can be well estimated, sampling from $f_\omega(x)$ will result in a "random walk" in the space of subregions (by regarding each subregion as a "point") with each subregion being sampled with a frequency proportional to π_i. Hence, the local trap problem can be overcome

* Tel.: +1 979 8458885.
E-mail address: fliang@stat.tamu.edu.

0167-7152/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.10.007

essentially, provided that the sample space is partitioned appropriately. SAMC provides a systematic way, as reviewed in Section 2.1, to estimate $\omega_1, \ldots, \omega_m$ under the framework of the stochastic approximation method (Robbins and Monro, 1951).

SAMC has been applied successfully to many hard computational problems, such as phylogenetic tree reconstruction (Cheon and Liang, 2008), neural network training (Liang, 2007), and Bayesian model selection (Liang et al., 2007). However, its use in Monte Carlo integration has not yet been well explored. Suppose that our goal is to estimate the quantity

$$E_f h(x) = \int h(x) f(x) \mathrm{d}x, \tag{3}$$

where $h(x)$ denotes an integrable function with respect to $f(x)$. Liang et al. (2007) proposed a two-stage sampling procedure which can generate (importance) samples used for evaluating the integral $E_f h(x)$. The procedure is as follows:

(i) (*ω-estimation*) Run SAMC to estimate ω_i's.
(ii) (*Resampling*) Simulate samples from $f_{\widehat{\omega}}(x)$ using a MCMC algorithm, and retain each sample with a probability proportional to $\widehat{\omega}_{J(x)}$, where $J(x)$ denotes the index of the subregion the sample x belongs to. Alternatively, we can retain all the samples and the corresponding subregion weights.

Although the procedure is general with respect to the setting of $\boldsymbol{\pi}$, it is not optimal in terms of efficiency, as it discards the samples generated in the ω-estimation stage for use in estimating $E_f h(x)$. In this paper, we show in theory that the samples generated in the ω-estimation stage can be used for estimating $E_f h(x)$. Our numerical results indicate that SAMC can create significant savings over conventional Monte Carlo algorithms for the problems for which the energy landscape is rugged.

2. Use of SAMC for Monte Carlo integration

2.1. A brief review of the SAMC algorithm

Let θ_{ti} denote the working estimate of $\log(\omega_i/\pi_i)$ obtained at iteration t, let $\theta_t = (\theta_{t1}, \ldots, \theta_{tm})$, and let $\{\gamma_t\}$ denote the gain factor sequence which satisfies the condition (A_1) given in Appendix. In this article, we set

$$\gamma_t = \frac{t_0}{\max(t_0, t)}, \quad t = 0, 1, 2, \ldots \tag{4}$$

for some specified values of $t_0 > 1$. A large value of t_0 will allow the sampler to reach all subregions very quickly even for a large system. Let $J(x)$ denote the index of the subregion the sample x belongs to. With the above notations, one iteration of SAMC can be described as follows:

SAMC algorithm:

(i) (MH sampling) Simulate a sample x_t by a single MH update with the target distribution

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{e^{\theta_{ti}}} I(x \in E_i). \tag{5}$$

(i.1) Generate y according to the proposal distribution $q(x_t, y)$.
(i.2) Calculate the ratio

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

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

(ii) (Weight updating) Set

$$\theta^* = \theta_t + \gamma_{t+1}(\mathbf{e}_t - \boldsymbol{\pi}), \tag{6}$$

where $\mathbf{e}_t = (e_{t,1}, \ldots, e_{t,m})$ and $e_{t,i} = 1$ if $x_t \in E_i$ and 0 otherwise. If $\theta^* \in \Theta$, set $\theta_{t+1} = \theta^*$; otherwise, set $\theta_{t+1} = \theta^* + \mathbf{c}^*$, where $\mathbf{c}^* = (c^*, \ldots, c^*)$ can be an arbitrary vector which satisfies the condition $\theta^* + \mathbf{c}^* \in \Theta$.

As implied by Theorem 5.4 of Andrieu et al. (2005), the varying truncation of θ^* can only occur a finite number of times, and thus $\{\theta_t\}$ can be kept in a compact space during simulations. To avoid frequent truncations, Liang et al. (2007) suggested to set Θ to $[-B_\Theta, B_\Theta]^m$ with B_Θ being a huge number, e.g., 10^{100}, which, as a practical matter, is equivalent to setting $\Theta = \mathbb{R}^m$. Note that adding to or subtracting a constant vector from θ_t will not change $f_{\theta_t}(x)$.

To study the convergence of the algorithm, Liang et al. (2007) assumed that the MH updates used in step (i) satisfy the condition (A_2) given in Appendix. In practice, such kinds of updates can be easily designed for both discrete and continuum systems. For example, a sufficient design is to restrict $\mathcal{X}$ to a compact set and to choose a proposal distribution $q(x, y)$ satisfying the condition: For every $x \in \mathcal{X}$, there exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $|x - y| \le \epsilon_1 \Longrightarrow q(x, y) \ge \epsilon_2$. If $\mathcal{X}$ is unbounded, some other constraints can be put on the tails of the target distribution $f(x)$ and the proposal distribution

$q(x, y)$ to ensure the condition (A_2) hold. Refer to Roberts and Tweedie (1996), Rosenthal (1995) and Roberts and Rosenthal (2004) for more discussions on this issue. Under the conditions (A_1) and (A_2), Liang et al. (2007) showed that as $t \to \infty$,

$$\theta_{ti} \to \begin{cases} C + \log(\omega_i) - \log(\pi_i + \nu), & \text{if } E_i \neq \emptyset, \\ -\infty. & \text{if } E_i = \emptyset, \end{cases} \tag{7}$$

holds almost surely, where $\nu = \sum_{j\in\{i:E_i=\emptyset\}} \pi_j/(m-m_0)$, m_0 is the number of empty subregions, and C represents an arbitrary constant. A subregion E_i is called empty if $\int_{E_i} \psi(x)\mathrm{d}x = 0$. In SAMC, the sample space partition can be made blindly, and this may lead to some empty subregions. SAMC allows for the existence of empty subregions. The constant C can be determined by imposing a constraint on θ_t, say, $\sum_{i=1}^m \mathrm{e}^{\theta_{ti}}$ is equal to a known number. Refer to Liang et al. (2007) for practical issues on implementation of SAMC, where the issues, such as how to partition the sample space, how to choose the gain factor sequence, and how to diagnose the convergence of the algorithm, have been discussed at length.

The superiority of SAMC in sample space exploration is due to its self-adjusting mechanism. If a proposal is rejected, the weight of the subregion that the current sample belongs to will be adjusted to a larger value, and thus the proposal of jumping out from the current subregion will be less likely to be rejected in the next iteration. This mechanism enables the system to escape from local traps very quickly. This is very important for the systems with multiple local energy minima.

2.2. Use of SAMC for Monte Carlo integration

In this section, we show that the samples generated by SAMC in the ω-estimation stage can be used directly in evaluation of $E_f h(x)$. To show this property, we first introduce the following lemmas.

Lemma 2.1 (*Billingsley, 1986, P. 218*). *Suppose that $F_t(A) = \int_A f_t(x)\mathrm{d}x$ and $F(A) = \int_A f(x)\mathrm{d}x$ for densities $f_t(x)$ and $f(x)$ defined on $\mathcal{X}$. If $f_t(x)$ converges to $f(x)$ almost surely, then $F_t(A) \longrightarrow F(A)$ as $t \to \infty$ uniformly for any $A \in \mathcal{B}(\mathcal{X})$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel set of the space $\mathcal{X}$.*

Let $\{Z_t, t \geq 0\}$ be a nonhomogeneous Markov chain with the finite state space $S = \{1, 2, \ldots, k\}$, the initial distribution

$$(P(1), P(2), \ldots, P(k)), \quad P(i) > 0,\ i \in S, \tag{8}$$

and the transition matrix

$$P_t = (P_t(j|i)), \quad P_t(j|i) > 0,\ i, j \in S,\ t \geq 1, \tag{9}$$

where $P_t(j|i) = P(Z_t = j|Z_{t-1} = i)$ for $t \geq 1$. Let $P = (P(j|i))$ be an ergodic transition matrix, and let $(p_1, \ldots, p_k)$ be the stationary distribution determined by P.

Lemma 2.2 (*Liu and Yang, 1996; Theorem 7*). *Let $\{Z_t, t \geq 0\}$ be a nonhomogeneous Markov chain with the initial distribution (8) and the transition matrix (9), and let g and g_t, $t \geq 1$, be functions defined on S. If the following conditions hold,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n |P_t(j|i) - P(j|i)| = 0, \quad \text{a.s.}, \forall i, j \in S, \tag{10}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n |g_t(i) - g(i)| = 0, \quad \text{a.s.}, \forall i \in S, \tag{11}$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n g_t(Z_t) = \sum_{i=1}^k p_i g(i), \quad \text{a.s.} \tag{12}$$

Let $(x_1, \theta_1), \ldots, (x_n, \theta_n)$ denote a set of samples generated by SAMC in the ω-estimation stage. Without loss of generality, we assume that in the SAMC simulations, there are no empty subregions and a constraint has been imposed on θ_t such that $C = 0$ holds in (7). Let $J(x_t)$ denote the index of the subregion that the sample x_t belongs to. Then $J(x_1), \ldots, J(x_n)$ forms a sample from a nonhomogeneous Markov chain defined on the finite state space $\{1, \ldots, m\}$, recalling that the sample space has been partitioned into m disjoint subregions $E_1, \ldots, E_m$. The Markov chain is nonhomogeneous as the target distribution $f_{\theta_t}(x)$ changes from iterations to iterations. Let

$$f_\theta(x) = \sum_{i=1}^m \frac{\pi_i \psi(x)}{\omega_i} I(x \in E_i). \tag{13}$$

It follows from (7) that $f_{\theta_t}(x) \to f_\theta(x)$ almost surely. Then, it follows from Lemma 2.1 that

$$\lim_{t\to\infty} P_t(j|i) = P(j|i), \tag{14}$$

where the transition probability $P_t(j|i)$ is defined as

$$P_t(j|i) = \int_{E_i}\int_{E_j}\left[s_t(x, \mathrm{d}y) + I(x \in \mathrm{d}y)\left(1 - \int_{\mathcal{X}} s_t(x, \mathrm{d}z)\right)\right]\mathrm{d}x, \tag{15}$$

$s_t(x, \mathrm{d}y) = q(x, \mathrm{d}y)\min\{1, [f_{\theta_t}(y)q(y, x)]/[f_{\theta_t}(x)q(x, y)]\}$, and $P(j|i)$ can be defined similarly by replacing $f_{\theta_t}(\cdot)$ by $f_\theta(\cdot)$ in (15). Following from (14), (10) holds. It is obvious that $\boldsymbol{\pi}$ forms the stationary distribution of the Markov chain induced by the transition matrix P. Define the function

$$g_t(J(x_t)) = \mathrm{e}^{\theta_{tJ(x_t)}}, \quad t = 0, 1, 2, \ldots. \tag{16}$$

Following from (7), we have

$$\lim_{t\to\infty} g_t(i) = \omega_i/\pi_i, \quad \forall i \in S, \tag{17}$$

which implies (11) holds by defining $g(i) = \omega_i/\pi_i$, $i = 1\ldots, m$. Since both conditions (10) and (11) hold, we have, by Lemma 2.2, the following law of large numbers for the SAMC samples.

Proposition 2.1. *Assume the conditions* (A_1) *and* (A_2). *For a set of samples generated by SAMC, we have*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{e}^{\theta_{tJ(x_t)}} = \sum_{i=1}^{m}\omega_i, \quad a.s. \tag{18}$$

Let $\mathcal{A} \subset \mathcal{X}$ denote an arbitrary Borel set, and let $\mathcal{A}^c$ denote the complementary set of $\mathcal{A}$. Thus, $\tilde{E}_1 = E_1 \cap \mathcal{A}$, $\tilde{E}_2 = E_1 \cap \mathcal{A}^c, \ldots, \tilde{E}_{2m-1} = E_m \cap \mathcal{A}, \tilde{E}_{2m} = E_m \cap \mathcal{A}^c$ form a new partition of the sample space $\mathcal{X}$. In this paper, we call the new partition an induced partition by $\mathcal{A}$, and the subregion $\tilde{E}_i$ an induced subregion by $\mathcal{A}$. Let $(x_1, \theta_1), \ldots, (x_n, \theta_n)$ denote again a set of samples generated by SAMC with the partition $E_1, \ldots, E_m$, and let $\tilde{J}(x_t)$ denote the index of the induced subregion x_t belongs to. Then $\tilde{J}(x_1), \ldots, \tilde{J}(x_n)$ forms a sample from a nonhomogeneous Markov chain defined on the finite state space $\{1, \ldots, 2m\}$. The transition matrices of the Markov chain can be defined similarly to (15). The stationary distribution of the limiting Markov chain is then $(\tilde{p}_1, \ldots, \tilde{p}_{2m})$, where $\tilde{p}_{2i-1} = \pi_i P_f(\mathcal{A}|E_i)$ and $\tilde{p}_{2i} = \pi_i P_f(\mathcal{A}^c|E_i)$ for $i = 1, \ldots, m$, and $P_f(A|B) = \int_{A\cap B} f(x)\mathrm{d}x / \int_B f(x)\mathrm{d}x$. Define

$$\tilde{g}_t(\tilde{J}(x_t)) = \mathrm{e}^{\theta_{tJ(x_t)}} I(x_t \in \mathcal{A}), \quad t = 0, 1, 2, \ldots, \tag{19}$$

where $I(\cdot)$ is the indicator function. Following from (7), we have

$$\lim_{t\to\infty}\tilde{g}_t(2i-1) = \omega_i/\pi_i \quad \text{and} \quad \lim_{t\to\infty}\tilde{g}_t(2i) = 0, \quad i = 1, \ldots, m, \tag{20}$$

which implies (11) holds by defining $\tilde{g}(2i-1) = \omega_i/\pi_i$ and $\tilde{g}(2i) = 0$, $i = 1\ldots, m$. With the similar arguments as for Proposition 2.1, we have Proposition 2.2:

Proposition 2.2. *Assume the conditions* (A_1) *and* (A_2). *For a set of samples generated by SAMC, we have*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{e}^{\theta_{tJ(x_t)}} I(x_t \in \mathcal{A}) = \sum_{i=1}^{m}\omega_i P_f(\mathcal{A}|E_i), \quad a.s. \tag{21}$$

Consider again the samples $(x_1, \theta_1), \ldots, (x_n, \theta_n)$ generated by SAMC with the partition $E_1, \ldots, E_m$. Let $y_1, \ldots, y_{n'}$ denote the distinct samples among $x_1, \ldots, x_n$. Generate a random variable/vector Y such that

$$P(Y = y_i) = \frac{\sum_{t=1}^{n}\mathrm{e}^{\theta_{tJ(x_t)}} I(x_t = y_i)}{\sum_{t=1}^{n}\mathrm{e}^{\theta_{tJ(x_t)}}}, \quad i = 1, \ldots, n', \tag{22}$$

where $I(\cdot)$ is the indicator function. Since Θ has been restricted to a compact set in SAMC, $\theta_{tJ(x_t)}$ is finite. The following theorem shows that Y is asymptotically distributed as $f(\cdot)$.

Theorem 2.1. *Assume the conditions* (A_1) *and* (A_2). *For a set of samples generated by SAMC, the random variable/vector Y generated in* (22) *is asymptotically distributed as* $f(\cdot)$.

Proof. For any Borel set $\mathcal{A} \subset \mathcal{X}$,

$$P\left(Y \in \mathcal{A} | (x_1, \theta_1), \ldots, (x_n, \theta_n)\right) = \frac{\sum_{t=1}^{n} \mathrm{e}^{\theta_{tJ(x_t)}} I(x_t \in \mathcal{A})}{\sum_{t=1}^{n} \mathrm{e}^{\theta_{tJ(x_t)}}}.$$

Following from Propositions 2.1 and 2.2, we have, by noting $P_f(E_i) = \omega_i / \sum_{j=1}^{m} \omega_j$,

$$P\left(Y \in \mathcal{A} | (x_1, \theta_1), \ldots, (x_n, \theta_n)\right) \to \frac{\sum_{i=1}^{m} \omega_i P_f(\mathcal{A}|E_i)}{\sum_{i=1}^{m} \omega_i} = \int_{\mathcal{A}} f(x) \mathrm{d}x$$

which, by Lebesgue's dominated convergence theorem, implies that

$$P(Y \in \mathcal{A}) = E\left[P\left(Y \in \mathcal{A} | (x_1, \theta_1), \ldots, (x_n, \theta_n)\right)\right] \to \int_{\mathcal{A}} f(x) \mathrm{d}x.$$

The proof is completed. □

Theorem 2.1 implies that for an integrable function $h(x)$, the expectation $E_f h(x)$ can be estimated by

$$\widehat{E_f h(x)} = \frac{\sum_{t=1}^{n} \mathrm{e}^{\theta_{tJ(x_t)}} h(x_t)}{\sum_{t=1}^{n} \mathrm{e}^{\theta_{tJ(x_t)}}}. \tag{23}$$

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

We note that SAMC falls into the class of dynamic weighting algorithms, that is, SAMC is (asymptotically) invariant with respect to the importance weights (IWIW) (Wong and Liang, 1997). Let $g_t(x, w)$ be the joint distribution of the sample (x, w) drawn at iteration t, where $w = \exp(\theta_{tJ(x)})$. The concept IWIW can be defined as follows:

The joint distribution $g_t(x, w)$ is said to be correctly weighted with respect to a distribution $f(x)$ if

$$\int w g_t(x, w) \mathrm{d}w \propto f(x). \tag{24}$$

A transition rule is said to satisfy IWIW if it maintains the correctly weighted property for the joint distribution $g_t(x, w)$ whenever an initial joint distribution is correctly weighted.

IWIW is a more general concept than the detailed balance condition satisfied by conventional Monte Carlo algorithms. It is easy to verify that both the MH algorithm and the adaptive Metropolis algorithm (Harrio et al., 2001) satisfy IWIW by assigning each sample an equal weight of 1.

Theorem 2.2. *Assume the conditions* (A_1) *and* (A_2). *Then SAMC asymptotically satisfies IWIW.*

Proof. Let $g_t(x, w_x)$ denote the joint distribution of the sample (x, w_x) generated by SAMC at iteration t, where $w_x = \mathrm{e}^{\theta_{tJ(x)}}$. If x can be regarded as a sample drawn from $f_{\theta_t}(x)$, then (x, w_x) has the joint distribution

$$g_t(x, w_x) = f_{\theta_t}(x) \delta\left(w_x = \mathrm{e}^{\theta_{tJ(x)}}\right), \tag{25}$$

where $\delta(\cdot)$ denotes a Dirac measure and corresponds to the conditional distribution of w_x given the sample x. It is easy to check that (25) is correctly weighted with respect to $f(x)$. The existence of such a sample is obvious, as it can, at least, be obtained after the convergence of θ_t.

Let (y, w_y) denote the sample generated at iteration $t+1$, and let $w'_y = \mathrm{e}^{\theta_{tJ(y)}}$. It follows from (6) that $w_y = \mathrm{e}^{\gamma_{t+1}(1-\pi_{J(y)})} w'_y$. Furthermore, we have

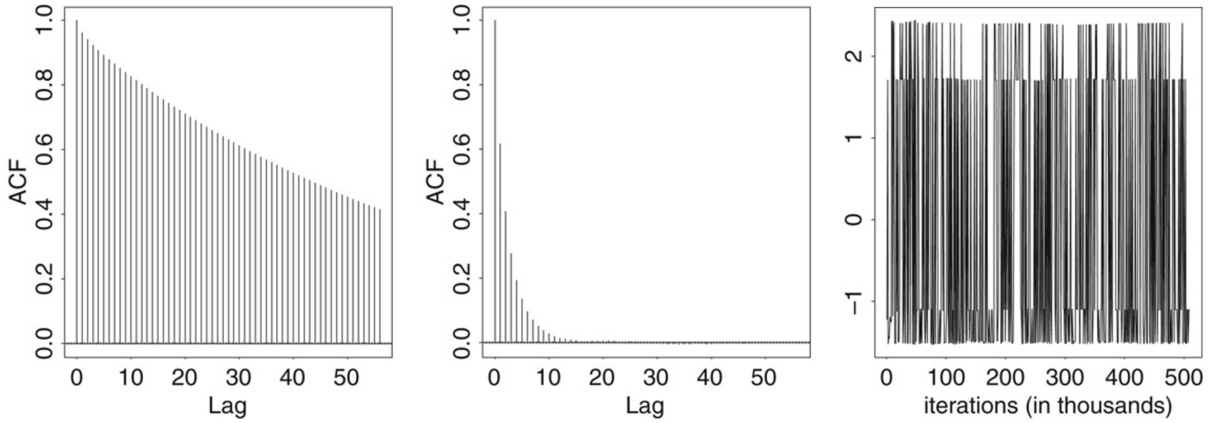
$$\begin{aligned}
\int w_y g_{t+1}\left(y, w_y\right) \mathrm{d}w_y &= \int\int\int \mathrm{e}^{\gamma_{t+1}(1-\pi_{J(y)})} w'_y g_t\left(x, w_x\right) P_{\theta_t}\left((x, w_x) \to \left(y, w'_y\right)\right) \mathrm{d}x \mathrm{d}w_x \mathrm{d}w'_y \\
&= \int\int\int \mathrm{e}^{\gamma_{t+1}(1-\pi_{J(y)})} w'_y \delta\left(w_x = \mathrm{e}^{\theta_{tJ(x)}}\right) f_{\theta_t}(x) P_{\theta_t}\left((x, w_x) \to \left(y, w'_y\right)\right) \mathrm{d}x \mathrm{d}w_x \mathrm{d}w'_y \\
&= \int\int\int \mathrm{e}^{\gamma_{t+1}(1-\pi_{J(y)})} w'_y \delta(w'_y = \mathrm{e}^{\theta_{tJ(y)}}) f_{\theta_t}(y) P_{\theta_t}\left(\left(y, w'_y\right) \to (x, w_x)\right) \mathrm{d}x \mathrm{d}w_x \mathrm{d}w'_y \\
&= \int \mathrm{e}^{\gamma_{t+1}(1-\pi_{J(y)})} w'_y g_t\left(y, w'_y\right) \mathrm{d}w'_y,
\end{aligned} \tag{26}$$

Table 1
The unnormalized mass function of the 10-state distribution.

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

Table 2
Comparison of SAMC and MH for the 10-state example, where the Bias and Standard Error (of the Bias) were calculated based on 100 independent runs, and the CPU times were measured on a 2.8 GHz computer for each run.

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



(a) MH samples.

(b) SAMC samples.

(c) Log-weight of SAMC samples.

Fig. 1. Computational results for the 10-state example. (a) Autocorrelation plot of the MH samples. (b) Autocorrelation plot of the SAMC samples. (c) Log-weight of the SAMC samples.

where $P_{\theta_t}(\cdot \to \cdot)$ denotes the MH transition kernel used in the t-th iteration of SAMC. Since $\gamma_t \to 0$, $\int w_y g_{t+1}(y, w_y)\,\mathrm{d}w_y \propto f(y)$ holds asymptotically by the last line of (26). If $\pi_{J(y)}$ is independent of y, i.e., $\pi_1 = \cdots = \pi_m = 1/m$, then $\int w_y g_{t+1}(y, w_y)\,\mathrm{d}w_y \propto f(y)$ holds exactly. □

Theorem 2.2 suggests that setting $\boldsymbol{\pi}$ to be non-uniform over the subregions may lead to a slightly biased estimate of $E_f h(x)$ for a short run of SAMC for which the gain factor sequence has not yet decreased to be sufficiently small. This is consistent with our numerical results, which are available at the request from the author.

3. An illustrative example

Reconsider Example 3.1 of Liang et al. (2007). The distribution consists of 10 states with the unnormalized mass function $\psi(x)$ as given in Table 1. It has two modes which are well separated by low mass states. Our goal is to estimate $E_f(X)$, the mean of the distribution.

The sample space was partitioned according to the mass function into five subregions: $E_1 = \{8\}$, $E_2 = \{2\}$, $E_3 = \{5, 6\}$, $E_4 = \{3, 9\}$ and $E_5 = \{1, 4, 7, 10\}$. In SAMC simulations, we set $\pi_1 = \cdots = \pi_5 = 1/5$, chose the MH transition proposal matrix as a stochastic matrix with each row being generated independently from the Dirichlet distribution $Dir(1, \ldots, 1)$, and set the gain factor sequence as in (4) with $T_0 = 10$. SAMC was run 100 times independently, and each run consisted of 5.1×10^5 iterations, for which the first 10^4 iterations were discarded for the burn-in process and the samples generated in the remaining iterations were used for estimation.

For comparison, the MH algorithm was also applied to this example with the same transition proposal matrix. The algorithm was run 100 times independently. Each run consisted of 5.1×10^5 iterations, and the samples generated in the last 5×10^5 iterations were used for estimation. The numerical results are summarized in Table 2. They indicate that SAMC is significantly better than MH for this example in terms of standard errors of the estimates. After accounting for the CPU cost, SAMC can still make about 4-fold improvement over MH. Note that for more complex problems, e.g., the phylogeny estimation problem considered in Cheon and Liang (2008), SAMC and MH will cost about the same CPU time for the same number of iterations, because in this case the CPU time used by each algorithm is dominated by the part used for energy evaluation, and the part used for weighting updating in SAMC is ignorable.

The reason why SAMC outperforms MH can be explained by Fig. 1 (a) and (b). For this example, MH mixes very slowly due to the presence of two separated modes, whilst SAMC can still mix very fast due to its self-adjusting mechanism. Fig. 1(c) shows the evolution of the log-weight of SAMC samples. It indicates that the magnitude of the SAMC weights is rather stable.

Note that in the SAMC simulation, the formula (6) was applied to update θ_t, for which the term $-\gamma_{t+1}\boldsymbol{\pi}$ helps to stabilize the magnitude of the importance weights by keeping the sum of θ_{ti}'s unchanged over iterations.

Acknowledgments

The author's research was supported in part by the grant (DMS-0607755) made by the National Science Foundation and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST). The author thanks Professors Chuanhai Liu and Minghui Chen for their early discussions on the topic, and thanks Professor Hira Koul, the associate editor, and the referee for their comments which have led to significant improvement of this paper.

Appendix

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

$$\lim_{t\to\infty} \gamma_t = 0, \qquad \sum_{t=1}^{\infty} \gamma_t = \infty, \qquad \sum_{t=1}^{\infty} \gamma_t^{\eta} < \infty,$$

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

(A_2) Let P_θ denote the MH transition kernel for a given $\theta \in \Theta$. For any $\theta \in \Theta$, P_θ is ψ-irreducible and aperiodic (Meyn and Tweedie, 1995). In addition, there exist a function $V : \mathcal{X} \to [1, \infty)$ and a constant $\alpha \geq 2$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(i) there exist a set $\mathcal{C} \subset \mathcal{X}$, an integer l, constants $0 < \lambda < 1$, b, ζ, $\delta > 0$ and a probability measure ν such that

- $\sup_{\theta\in\mathcal{K}} P_\theta^l V^\alpha(x) \leq \lambda V^\alpha(x) + bI(x \in \mathcal{C}), \quad \forall x \in \mathcal{X},$
- $\sup_{\theta\in\mathcal{K}} P_\theta V^\alpha(x) \leq \zeta V^\alpha(x), \quad \forall x \in \mathcal{X},$
- $\inf_{\theta\in\mathcal{K}} P_\theta^l(x, A) \geq \delta\nu(A), \quad \forall x \in \mathcal{C}, \forall A \in \mathcal{B}_\mathcal{X}$

where $P_\theta V(x) = \int_\mathcal{X} P_\theta(x, y)V(y)\mathrm{d}y$ and $\mathcal{B}_\mathcal{X}$ is the Borel set defined on $\mathcal{X}$.

(ii) there exists a constant c such that for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

- $\|P_\theta g - P_{\theta'} g\|_V \leq c\|g\|_V |\theta - \theta'|, \quad \forall g \in \mathcal{L}_V,$
- $\|P_\theta g - P_{\theta'} g\|_{V^\alpha} \leq c\|g\|_{V^\alpha} |\theta - \theta'|, \quad \forall g \in \mathcal{L}_{V^\alpha},$

where $|z|$ denotes the norm of the vector z, $\|g\|_V = \sup_{x\in\mathcal{X}} |g(x)|/V(x)$, and $\mathcal{L}_V = \{g : \mathcal{X} \to \mathbb{R}^m, \|g\|_V < \infty\}$.

References

Andrieu, C., Moulines, E., Priouret, P., 2005. Stability of stochastic approximation under verifiable conditions. SIAM J. Control Optim. 44, 283–312.

Atchadé, Y.F., Liu, J.S., 2007. The Wang–Landau algorithm in general state spaces: applications and convergence analysis. Technical Report, Department of Statistics, University of Michigan.

Berg, B.A., Neuhaus, T., 1991. Multicanonical algorithms for 1st order phase-transitions. Phys. Lett. B 267, 249–253.

Billingsley, P., 1986. Probability and Measure, 2nd edition. John Wiley & Sons, New York.

Cheon, S., Liang, F., 2008. Phylogenetic tree reconstruction using stochastic approximation Monte Carlo. BioSystems 91, 94–107.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57, 1317–1339.

Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood. In: Keramigas, E.M. (Ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Interface Foundation, Fairfax, pp. 156–163.

Harrio, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli 7, 223–242.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Kou, S.C., Zhou, Q., Wong, W.H., 2006. Equi-energy sampler with applications to statistical inference and statistical mechanics (with discussion). Ann. Statist. 34, 1581–1619.

Liang, F., 2007. Annealing stochastic approximation Monte Carlo for neural network training. Mach. Learning 68, 201–233.

Liang, F., Liu, C., Carroll, R.J., 2007. Stochastic approximation in Monte Carlo computation. J. Amer. Statist. Assoc. 102, 305–320.

Liang, F., Wong, W.H., 2001. Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. J. Amer. Statist. Assoc. 96, 653–666.

Liu, W., Yang, W., 1996. An extension of Shannon–McMillan theorem and some limit properties for nonhomogeneous Markov chains. Stochastic Process. Appl. 61, 129–145.

Marinari, E., Parisi, G., 1992. Simulated tempering: A new Monte Carlo scheme. Europhys. Lett. 19, 451–458.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1091.

Meyn, S., Tweedie, R., 1995. Markov Chains and Stochastic Stability. Springer-Verlag, London.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Statist. 22, 400–407.

Roberts, G.O., Rosenthal, J.S., 2004. General state-space Markov chains and MCMC algorithms. Prob. Surv. 1, 20–71.

Roberts, G.O., Tweedie, R.L., 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis Algorithms. Biometrika 83, 95–110.

Rosenthal, J.S., 1995. Minorization conditions and convergence rate for Markov chain Monte Carlo. J. Amer. Statist. Assoc. 90, 558–566.

Wang, F., Landau, D.P., 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys. Rev. Lett. 86, 2050–2053.

Wong, W.H., Liang, F., 1997. Dynamic weighting in Monte Carlo and optimization. Proc. Natl. Acad. Sci. USA 94, 14220–14224.