The Annals of Statistics
2010, Vol. 38, No. 5, 2823–2856
DOI: 10.1214/10-AOS807

TRAJECTORY AVERAGING FOR STOCHASTIC APPROXIMATION MCMC ALGORITHMS[1]

By Faming Liang

Texas A&M University

The subject of stochastic approximation was founded by Robbins and Monro [*Ann. Math. Statist.* **22** (1951) 400–407]. After five decades of continual development, it has developed into an important area in systems control and optimization, and it has also served as a prototype for the development of adaptive algorithms for on-line estimation and control of stochastic systems. Recently, it has been used in statistics with Markov chain Monte Carlo for solving maximum likelihood estimation problems and for general simulation and optimizations. In this paper, we first show that the trajectory averaging estimator is asymptotically efficient for the stochastic approximation MCMC (SAMCMC) algorithm under mild conditions, and then apply this result to the stochastic approximation Monte Carlo algorithm [Liang, Liu and Carroll *J. Amer. Statist. Assoc.* **102** (2007) 305–320]. The application of the trajectory averaging estimator to other stochastic approximation MCMC algorithms, for example, a stochastic approximation MLE algorithm for missing data problems, is also considered in the paper.

1. Introduction. Robbins and Monro (1951) introduced the stochastic approximation algorithm to solve the integration equation

$$h(\theta) = \int_{\mathcal{X}} H(\theta, x) f_\theta(x)\, dx = 0, \tag{1}$$

where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ is a parameter vector and $f_\theta(x)$, $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$, is a density function depending on θ. The d_θ and d_x denote the dimensions of θ and x, respectively. The stochastic approximation algorithm is an iterative recursive algorithm, whose each iteration consists of two steps:

Received May 2009; revised November 2009.
[1]Supported in part by NSF Grants DMS-06-07755, CMMI-0926803 and the Award KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).
AMS 2000 subject classifications. 60J22, 65C05.
Key words and phrases. Asymptotic efficiency, convergence, Markov chain Monte Carlo, stochastic approximation Monte Carlo, trajectory averaging.

Stochastic approximation algorithm.

- Generate $X_{k+1} \sim f_{\theta_k}(x)$, where k indexes the iteration.
- Set $\theta_{k+1} = \theta_k + a_k H(\theta_k, X_{k+1})$, where $a_k > 0$ is called the gain factor.

The stochastic approximation algorithm is often studied by rewriting it as follows:

$$\theta_{k+1} = \theta_k + a_k[h(\theta_k) + \varepsilon_{k+1}], \tag{2}$$

where $h(\theta_k) = \int_{\mathcal{X}} H(\theta_k, x) f_{\theta_k}(x)\, dx$ corresponds to the mean effect of $H(\theta_k, X_{k+1})$, and $\varepsilon_{k+1} = H(\theta_k, X_{k+1}) - h(\theta_k)$ is called the observation noise. In the literature of stochastic approximation, $h(\theta)$ is also called the mean field function. It is well known that the optimal convergence rate of (2) can be achieved with $a_k = -F^{-1}/k$, where $F = \partial h(\theta^*)/\partial\theta$, and θ^* denotes the zero point of $h(\theta)$. In this case, (2) is reduced to Newton's algorithm. Unfortunately, it is often impossible to use this algorithm, as the matrix F is generally unknown.

Although an optimal convergence rate of θ_k cannot be obtained in general, in a sequence of fundamental papers Ruppert (1988), Polyak (1990) and Polyak and Juditsky (1992) showed that the trajectory averaging estimator is asymptotically efficient; that is, $\bar{\theta}_n = \sum_{k=1}^{n} \theta_k/n$ can converge in distribution to a normal random variable with mean θ^* and covariance matrix Σ, where Σ is the smallest possible covariance matrix in an appropriate sense. The trajectory averaging estimator requires $\{a_k\}$ to be relatively large, decreasing slower than $O(1/k)$. As discussed by Polyak and Juditsky (1992), trajectory averaging is based on a paradoxical principle: a slow algorithm having less than optimal convergence rate must be averaged.

Recently, the trajectory averaging technique has been further explored in a variety of papers [see, e.g., Chen (1993), Kushner and Yang (1993, 1995), Dippon and Renz (1997), Wang, Chong and Kulkarni (1997), Tang, L'Ecuyer and Chen (1999), Pelletier (2000) and Kushner and Yin (2003)] with different assumptions for the observation noise. However, up to our knowledge, it has not yet been explored for stochastic approximation MCMC (SAMCMC) algorithms [Benveniste, Métivier and Priouret (1990), Chen (2002), Kushner and Yin (2003), Andrieu, Moulines and Priouret (2005), Andrieu and Moulines (2006)]. The stochastic approximation MCMC algorithms refer to a class of stochastic approximation algorithms for which the sample is generated at each iteration via a Markov transition kernel; that is, $\{x_{k+1}\}$ is generated via a family of Markov transition kernel $\{P_{\theta_k}(x_k, \cdot)\}$ controlled by $\{\theta_k\}$. Recently, the stochastic approximation MCMC algorithms have been used in statistics for solving maximum likelihood estimation problems [Younes (1989, 1999), Moyeed and Baddeley (1991), Gu and Kong (1998), Gu and Zhu (2001)], and for general simulation and optimizations [Liang,

Liu and Carroll (2007), Atchadé and Liu (2010)]. It is worth to point out that in comparison with conventional MCMC algorithms, for example, the Metropolis–Hastings algorithm [Metropolis et al. (1953), Hastings (1970)], parallel tempering [Geyer (1991)], and simulated tempering [Marinari and Parisi (1992), Geyer and Thompson (1995)], the stochastic approximation Monte Carlo (SAMC) algorithm [Liang, Liu and Carroll (2007)] has significant advantages in simulations of complex systems for which the energy landscape is rugged. As explained later (in Section 3), SAMC is essentially immune to the local trap problem due to its self-adaptive nature inherited from the stochastic approximation algorithm. SAMC has been successfully applied to many statistical problems, such as p-value evaluation for resampling-based tests [Yu and Liang (2009)], Bayesian model selection [Liang (2009), Atchadé and Liu (2010)] and spatial model estimation [Liang (2007a)], among others.

In this paper, we explore the theory of trajectory averaging for stochastic approximation MCMC algorithms, motivated by their wide applications. Although Chen (1993, 2002) considered the case where the observation noise can be state dependent, that is, the observation noise ε_{k+1} depends on $\theta_0, \ldots, \theta_k$, their results are not directly applicable to the stochastic approximation MCMC algorithms due to some reasons as explained in Section 5. The theory established by Kushner and Yin (2003) can potentially be extended to the stochastic approximation MCMC algorithm, but, as mentioned in Kushner and Yin [(2003), page 375] the extension is not straightforward and more work needs to be done to deal with the complicated structure of the Markov transition kernel. In this paper, we propose a novel decomposition of the observation noise for the stochastic approximation MCMC algorithms. Based on the proposed decomposition, we show the trajectory averaging estimator is asymptotically efficient for the stochastic approximation MCMC algorithms, and then apply this result to the SAMC algorithm. These results are presented in Lemma A.5, Theorems 2.3 and 3.2, respectively. The application of the trajectory averaging technique to other stochastic approximation MCMC algorithms, for example, a stochastic approximation MLE algorithm for missing data problems, is also considered in the paper.

The remainder of this paper is organized as follows. In Section 2, we present our main theoretical result that the trajectory averaging estimator is asymptotically efficient for the stochastic approximation MCMC algorithms. In Section 3, we apply the trajectory averaging technique to the SAMC algorithm. In Section 4, we apply the trajectory averaging technique to a stochastic approximation MLE algorithm for missing data problems. In Section 5, we conclude the paper with a brief discussion.

2. Trajectory averaging for a general stochastic approximation MCMC algorithm.

2.1. *A varying truncation stochastic approximation MCMC algorithm.* To show the convergence of the stochastic approximation algorithm, restrictive conditions on the observation noise and mean field function are required. For example, one often assumes the noise to be mutually independent or to be a martingale difference sequence, and imposes a sever restriction on the growth rate of the mean field function. These conditions are usually not satisfied in practice. See Chen [(2002), Chapter 1] for more discussions on this issue. To remove the growth rate restriction on the mean field function and to weaken the conditions imposed on noise, Chen and Zhu (1986) proposed a varying truncation version for the stochastic approximation algorithm. The convergence of the modified algorithm can be shown for a wide class of the mean filed function under a truly weak condition on noise; see, for example, Chen, Guo and Gao (1988) and Andrieu, Moulines and Priouret (2005). The latter gives a proof for the convergence of the modified algorithm with Markov state-dependent noise under some conditions that are easy to verify.

Following Andrieu, Moulines and Priouret (2005), we consider the following varying truncation stochastic approximation MCMC algorithm. Let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of Θ such that

$$\bigcup_{s\geq 0} \mathcal{K}_s = \Theta \quad \text{and} \quad \mathcal{K}_s \subset \operatorname{int}(\mathcal{K}_{s+1}), \qquad s \geq 0, \tag{3}$$

where $\operatorname{int}(A)$ denotes the interior of set A. Let $\{a_k\}$ and $\{b_k\}$ be two monotone, nonincreasing, positive sequences. Let $\mathcal{X}_0$ be a subset of $\mathcal{X}$, and let $\mathcal{T}: \mathcal{X} \times \Theta \to \mathcal{X}_0 \times \mathcal{K}_0$ be a measurable function which maps a point (x, θ) in $\mathcal{X} \times \Theta$ to a random point in $\mathcal{X}_0 \times \mathcal{K}_0$; that is, both x and θ will be reinitialized in $\mathcal{X}_0 \times \mathcal{K}_0$. As shown in Lemma A.5, for the stochastic approximation MCMC algorithm, when the number of iterations becomes large, the observation noise ε_k can be decomposed as

$$\varepsilon_k = e_k + \nu_k + \varsigma_k, \tag{4}$$

where $\{e_k\}$ forms a martingale difference sequence, and the expectation of the other two terms will go to zero in certain forms. In Theorems 2.2 and 2.3, we show that $\{e_k\}$ leads to the asymptotic normality of the trajectory averaging estimator $\bar{\theta}_k$, and $\{\nu_k\}$ and $\{\varsigma_k\}$ can vanish or be ignored when the asymptotic distribution of $\bar{\theta}_k$ is considered.

Let σ_k denote the number of truncations performed until iteration k and $\sigma_0 = 0$. The varying truncation stochastic approximation MCMC algorithm starts with a random choice of (θ_0, x_0) in the space $\mathcal{K}_0 \times \mathcal{X}_0$, and then iterates between the following steps:

Varying truncation stochastic approximation MCMC algorithm.

- Draw sample x_{k+1} with a Markov transition kernel, P_{θ_k}, which admits $f_{\theta_k}(x)$ as the invariant distribution.
- Set $\theta_{k+1/2} = \theta_k + a_k H(\theta_k, x_{k+1})$.
- If $\|\theta_{k+1/2} - \theta_k\| \le b_k$ and $\theta_{k+1/2} \in \mathcal{K}_{\sigma_k}$, where $\|z\|$ denote the Euclidean norm of the vector z, then set $(\theta_{k+1}, x_{k+1}) = (\theta_{k+1/2}, x_{k+1})$ and $\sigma_{k+1} = \sigma_k$; otherwise, set $(\theta_{k+1}, x_{k+1}) = \mathcal{T}(\theta_k, x_k)$ and $\sigma_{k+1} = \sigma_k + 1$.

As depicted by the algorithm, the varying truncation mechanism works in an adaptive manner as follows: when the current estimate of the parameter wanders outside the active truncation set or when the difference between two successive estimates is greater than a time-dependent threshold, then the algorithm is reinitialized with a smaller initial value of the gain factor and a larger truncation set. This mechanism enables the algorithm to select an appropriate gain factor sequence and an appropriate starting point, and thus to confine the recursion to a compact set; that is, the number of reinitializations is almost surely finite for every $(\theta_0, x_0) \in \mathcal{K}_0 \times \mathcal{X}_0$. This result is formally stated in Theorem 2.1, which plays a crucial role for establishing asymptotic efficiency of the trajectory averaging estimator.

Regarding the varying truncation scheme, one can naturally propose many variations. For example, one may not change the truncation set when only the condition $\|\theta_{k+1/2} - \theta_k\| \le b_k$ is violated, and, instead of jumping forward in a unique gain factor sequence, one may start with a different gain factor sequence (smaller than the previous one) when the reinitialization occurs. In either case, the proof for the theorems presented in Section 2.2 follows similarly.

2.2. *Theoretical results on the trajectory averaging estimator.* The asymptotic efficiency of $\bar{\theta}_k$ can be analyzed under the following conditions.

Lyapunov condition on $h(\theta)$. Let $\langle x, y\rangle$ denote the Euclidean inner product.

(A_1) Θ is an open set, the function $h : \Theta \to \mathbb{R}^d$ is continuous, and there exists a continuously differentiable function $v : \Theta \to [0, \infty)$ such that:

(i) There exists $M_0 > 0$ such that

$$\mathcal{L} = \{\theta \in \Theta, \langle \nabla v(\theta), h(\theta)\rangle = 0\} \subset \{\theta \in \Theta, v(\theta) < M_0\}. \tag{5}$$

(ii) There exists $M_1 \in (M_0, \infty)$ such that $\mathcal{V}_{M_1}$ is a compact set, where $\mathcal{V}_M = \{\theta \in \Theta, v(\theta) \le M\}$.

(iii) For any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla v(\theta), h(\theta)\rangle < 0$.

(iv) The closure of $v(\mathcal{L})$ has an empty interior.

This condition assumes the existence of a global Lyapunov function v for the mean field h. If h is a gradient field, that is, $h = -\nabla J$ for some lower bounded real-valued and differentiable function $J(\theta)$, then v can be set to J, provided that J is continuously differentiable. This is typical for stochastic optimization problems, for example, machine learning [Tadić (1997)], where a continuously differentiable objective function $J(\theta)$ is minimized.

Stability condition on $h(\theta)$.

(A_2) The mean field function $h(\theta)$ is measurable and locally bounded. There exist a stable matrix F (i.e., all eigenvalues of F are with negative real parts), $\gamma > 0$, $\rho \in (0, 1]$, and a constant c such that, for any $\theta^* \in \mathcal{L}$,

$$\|h(\theta) - F(\theta - \theta^*)\| \le c\|\theta - \theta^*\|^{1+\rho} \qquad \forall \theta \in \{\theta : \|\theta - \theta^*\| \le \gamma\},$$

where $\mathcal{L}$ is defined in (5).

This condition constrains the behavior of the mean field function around the solution points. It makes the trajectory averaging estimator sensible both theoretically and practically. If $h(\theta)$ is differentiable, the matrix F can be chosen to be the partial derivative of $h(\theta)$, that is, $\partial h(\theta)/\partial\theta$. Otherwise, certain approximation may be needed.

Drift condition on the transition kernel P_θ. Before giving details of this condition, we first define some terms and notation. Assume that a transition kernel P_θ is irreducible, aperiodic, and has a stationary distribution on a sample space denoted by $\mathcal{X}$. A set $\mathbf{C} \subset \mathcal{X}$ is said to be small if there exist a probability measure ν on $\mathcal{X}$, a positive integer l and $\delta > 0$ such that

$$P_\theta^l(x, A) \ge \delta\nu(A) \qquad \forall x \in \mathbf{C}, \forall A \in \mathcal{B}_\mathcal{X},$$

where $\mathcal{B}_\mathcal{X}$ is the Borel set defined on $\mathcal{X}$. A function $V : \mathcal{X} \to [1, \infty)$ is said to be a drift function outside $\mathbf{C}$ if there exist positive constants $\lambda < 1$ and b such that

$$P_\theta V(x) \le \lambda V(x) + bI(x \in \mathbf{C}) \qquad \forall x \in \mathcal{X},$$

where $P_\theta V(x) = \int_\mathcal{X} P_\theta(x, y)V(y)\,dy$. For a function $g : \mathcal{X} \to \mathbb{R}^d$, define the norm

$$\|g\|_V = \sup_{x\in\mathcal{X}} \frac{\|g(x)\|}{V(x)}$$

and define the set $\mathcal{L}_V = \{g : \mathcal{X} \to \mathbb{R}^d, \sup_{x\in\mathcal{X}}\|g\|_V < \infty\}$. Given the terms and notation introduced above, the drift condition can be specified as follows.

(A_3) For any given $\theta \in \Theta$, the transition kernel P_θ is irreducible and aperiodic. In addition, there exists 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 $\mathbf{C} \subset \mathcal{X}$, an integer l, constants $0 < \lambda < 1$, b, ς, $\delta > 0$ and a probability measure ν such that

$$\sup_{\theta \in \mathcal{K}} P_\theta^l V^\alpha(x) \leq \lambda V^\alpha(x) + bI(x \in \mathbf{C}) \qquad \forall x \in \mathcal{X}, \tag{6}$$

$$\sup_{\theta \in \mathcal{K}} P_\theta V^\alpha(x) \leq \varsigma V^\alpha(x) \qquad \forall x \in \mathcal{X}, \tag{7}$$

$$\inf_{\theta \in \mathcal{K}} P_\theta^l(x, A) \geq \delta \nu(A) \qquad \forall x \in \mathbf{C}, \forall A \in \mathcal{B}_\mathcal{X}. \tag{8}$$

(ii) There exists a constant $c > 0$ such that, for all $x \in \mathcal{X}$,

$$\sup_{\theta \in \mathcal{K}} \|H(\theta, x)\|_V \leq c, \tag{9}$$

$$\sup_{(\theta, \theta') \in \mathcal{K}} \|H(\theta, x) - H(\theta', x)\|_V \leq c\|\theta - \theta'\|. \tag{10}$$

(iii) There exists a constant $c > 0$ such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$\|P_\theta g - P_{\theta'} g\|_V \leq c\|g\|_V \|\theta - \theta'\| \qquad \forall g \in \mathcal{L}_V, \tag{11}$$

$$\|P_\theta g - P_{\theta'} g\|_{V^\alpha} \leq c\|g\|_{V^\alpha} \|\theta - \theta'\| \qquad \forall g \in \mathcal{L}_{V^\alpha}. \tag{12}$$

Assumption (A_3)(i) is classical in the literature of Markov chain. It implies the existence of a stationary distribution $f_\theta(x)$ for all $\theta \in \Theta$ and V^α-uniform ergodicity [Andrieu, Moulines and Priouret (2005)]. Assumption (A_3)(ii) gives conditions on the bound of $H(\theta, x)$. This is a critical condition for the observation noise. As seen later in Lemmas A.1 and A.5, it directly leads to the boundedness of some terms decomposed from the observation noise. For some algorithms, for example, SAMC, for which $H(\theta, x)$ is a bounded function, the drift function can be simply set as $V(x) = 1$.

Conditions on the step-sizes.

(A_4) The sequences $\{a_k\}$ and $\{b_k\}$ are nonincreasing, positive and satisfy the conditions:

$$\begin{aligned} &\sum_{k=1}^{\infty} a_k = \infty, \qquad \lim_{k \to \infty} (k a_k) = \infty, \\ &\frac{a_{k+1} - a_k}{a_k} = o(a_{k+1}), \qquad b_k = O(a_k^{(1+\tau)/2}), \end{aligned} \tag{13}$$

for some $\tau \in (0, 1]$,

$$\sum_{k=1}^{\infty} \frac{a_k^{(1+\tau)/2}}{\sqrt{k}} < \infty, \tag{14}$$

and for some constants $\alpha \geq 2$ as defined in condition (A_3),

$$\sum_{i=1}^{\infty} \{a_i b_i + (b_i^{-1} a_i)^{\alpha}\} < \infty. \tag{15}$$

It follows from (14) that

$$\sum_{i=[k/2]}^{k} \frac{a_i^{(1+\tau)/2}}{\sqrt{i}} = o(1),$$

where $[z]$ denotes the integer part of z. Since a_k is nonincreasing, we have

$$a_k^{(1+\tau)/2} \sum_{i=[k/2]}^{k} \frac{1}{\sqrt{i}} = o(1),$$

and thus $a_k^{(1+\tau)/2}\sqrt{k} = o(1)$, or $a_k = O(k^{-\eta})$ for $\eta \in (\frac{1}{2}, 1)$. For instance, $a_k = C_1/k^{\eta}$ for some constants $C_1 > 0$ and $\eta \in (\frac{1}{2}, 1)$, then we can set $b_k = C_2/k^{\xi}$ for some constants $C_2 > 0$ and $\xi \in (\frac{1}{2}, \eta - \frac{1}{\alpha})$, which satisfies (13) and (15). Under this setting, the existence of τ is obvious.

Theorem 2.1 concerns the convergence of the general stochastic approximation MCMC algorithm. The proof follows directly from Theorems 5.4, 5.5 and Proposition 6.1 of Andrieu, Moulines and Priouret (2005).

THEOREM 2.1. *Assume conditions* (A_1), (A_3) *and* (A_4) *hold. Let* k_σ *denote the iteration number at which the* σ*th truncation occurs in the stochastic approximation MCMC simulation. Let* $\mathcal{X}_0 \subset \mathcal{X}$ *be such that* $\sup_{x \in \mathcal{X}_0} V(x) < \infty$ *and that* $\mathcal{K}_0 \subset \mathcal{V}_{M_0}$*, where* $\mathcal{V}_{M_0}$ *is defined in* (A_1)*. Then there exists almost surely a number, denoted by* σ_s*, such that* $k_{\sigma_s} < \infty$ *and* $k_{\sigma_s+1} = \infty$*; that is,* $\{\theta_k\}$ *has no truncation for* $k \geq k_{\sigma_s}$*, or mathematically,*

$$\theta_{k+1} = \theta_k + a_k H(\theta_k, x_{k+1}) \qquad \forall k \geq k_{\sigma_s}.$$

In addition, we have

$$\theta_k \to \theta^* \qquad a.s.$$

for some point $\theta^* \in \mathcal{L}$.

Theorem 2.2 concerns the asymptotic normality of $\bar{\theta}_k$.

THEOREM 2.2. *Assume conditions* (A_1), (A_2), (A_3) *and* (A_4) *hold. Let* $\mathcal{X}_0 \subset \mathcal{X}$ *be such that* $\sup_{x \in \mathcal{X}_0} V(x) < \infty$ *and that* $\mathcal{K}_0 \subset \mathcal{V}_{M_0}$*, where* $\mathcal{V}_{M_0}$ *is defined in* (A_1)*. Then*

$$\sqrt{k}(\bar{\theta}_k - \theta^*) \longrightarrow N(\mathbf{0}, \Gamma)$$

for some point $\theta^* \in \Theta$*, where* $\Gamma = F^{-1}Q(F^{-1})^T$*,* $F = \partial h(\theta^*)/\partial\theta$ *is negative definite,* $Q = \lim_{k\to\infty} E(e_k e_k^T)$*, and* e_k *is as defined in (4).*

Below we consider the asymptotic efficiency of $\bar{\theta}_k$. As already mentioned, the asymptotic efficiency of the trajectory averaging estimator has been studied by quite a few authors. Tang, L'Ecuyer and Chen (1999) gives the following definition for the asymptotic efficient estimator that can be resulted from a stochastic approximation algorithm.

DEFINITION 2.1. Consider the stochastic approximation algorithm (2). Let $\{Z_n\}_{n\geq 0}$, given as a function of $\{\theta_n\}_{n\geq 0}$, be a sequence of estimators of θ^*. The algorithm $\{Z_n\}_{n\geq 0}$ is said to be asymptotically efficient if

$$\sqrt{n}(Z_n - \theta^*) \longrightarrow N(\mathbf{0}, F^{-1}\tilde{Q}(F^{-1})^T), \tag{16}$$

where $F = \partial h(y^*)/\partial y$, and $\tilde{Q}$ is the asymptotic covariance matrix of $(1/\sqrt{n}) \times \sum_{k=1}^{n} \varepsilon_k$.

As mentioned in Tang, L'Ecuyer and Chen (1999), $\tilde{Q}$ is the smallest possible limit covariance matrix that an estimator based on the stochastic approximation algorithm (2) can achieve. If $\theta_k \to \theta^*$ and $\{\varepsilon_k\}$ forms or asymptotically forms a martingale difference sequence, then we have $\tilde{Q} = \lim_{k\to\infty} E(\varepsilon_k \varepsilon_k^T)$. In the next theorem, we show that the asymptotic covariance matrix Q established in Theorem 2.2 is the same as $\tilde{Q}$, and thus the trajectory averaging estimator $\bar{\theta}_k$ is asymptotically efficient.

THEOREM 2.3. *Assume conditions* (A_1), (A_2), (A_3) *and* (A_4) *hold. Let* $\mathcal{X}_0 \subset \mathcal{X}$ *be such that* $\sup_{x\in\mathcal{X}_0} V(x) < \infty$ *and that* $\mathcal{K}_0 \subset \mathcal{V}_{M_0}$, *where* $\mathcal{V}_{M_0}$ *is defined in* (A_1). *Then* $\bar{\theta}_k$ *is asymptotically efficient.*

As implied by Theorem 2.3, the convergence rate of $\bar{\theta}_k$, which is measured by the asymptotic covariance matrix Γ, is independent of the choice of the gain factor sequence as long as the condition (A_4) is satisfied. The asymptotic efficiency of $\bar{\theta}_k$ can also be interpreted in terms of Fisher information theory. Refer to Pelletier [(2000), Section 3] and the references therein for more discussions on this issue.

Trajectory averaging enables smoothing of the behavior of the algorithm but at the same time, it slows down the numerical convergence because it takes longer for the algorithm to forget the first iterates. An alternative idea would be to consider moving window averaging algorithms, see, for example, Kushner and Yang (1993) and Kushner and Yin (2003), Chapter 11. Extension of their results to the general stochastic approximation MCMC algorithm will be of great interest.

3. Trajectory averaging for the stochastic approximation Monte Carlo algorithm.

3.1. *The SAMC algorithm.* Suppose that we are interested in sampling from the following distribution

$$f(x) = c\psi(x), \qquad x \in \mathcal{X}, \tag{17}$$

where c is an unknown constant, $\mathcal{X} \subset \mathbb{R}^{d_x}$ is the sample space. The basic idea of SAMC stems from the Wang–Landau algorithm [Wang and Landau (2001), Liang (2005)] and can be briefly explained as follows. 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 sample from the trial distribution

$$f_\omega(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x)}{\omega_i} I_{\{x \in E_i\}}, \tag{18}$$

where π_i's are prespecified constants such that $\pi_i > 0$ for all i and $\sum_{i=1}^m \pi_i = 1$, and $I_{\{x \in E_i\}} = 1$ if $x \in E_i$ and 0 otherwise. For example, if the sample space is partitioned according to the energy function into the following subregions: $E_1 = \{x : -\log(\psi(x)) < u_1\}$, $E_2 = \{x : u_1 \le -\log(\psi(x)) < u_2\}, \ldots, E_m = \{x : -\log(\psi(x)) > u_{m-1}\}$, where $-\infty < u_1 < \cdots < u_{m-1} < \infty$ are the user-specified numbers, then sampling from $f_\omega(x)$ would result in a random walk (by viewing each subregion as a "point") in the space of energy with each subregion being sampled with probability π_i. Here, without loss of generality, we assume that each subregion is unempty; that is, assuming $\int_{E_i} \psi(x)\,dx > 0$ for all $i = 1, \ldots, m$. Therefore, sampling from (18) essentially avoids the local-trap problem suffered by the conventional MCMC algorithms. This is attractive, but ω_i's are unknown. SAMC provides a dynamic way to estimate ω_i's under the framework of the stochastic approximation MCMC algorithm.

In what follows we describe how ω can be estimated by SAMC. Since $f_\omega(x)$ is invariant with respect to a scale change of ω, it suffices to estimate $\omega_1, \ldots, \omega_{m-1}$ by fixing ω_m to a known constant provided $\omega_m > 0$. Let $\theta_k^{(i)}$ denote the working estimate of $\log(\omega_i/\pi_i)$ obtained at iteration k, and let $\theta_k = (\theta_k^{(1)}, \ldots, \theta_k^{(m-1)})$. Why this reparameterization is used will be explained at the end of this subsection. Let $\{a_k\}$ denote the gain factor sequence, and let $\{\mathcal{K}_s, s \ge 0\}$ denote a sequence of compact subsets of Θ as defined in (3). For this algorithm, $\{\mathcal{K}_s, s \ge 0\}$ can be chosen as follows. Define

$$v(\theta) = -\log\left(1 - \frac{1}{2}\sum_{j=1}^{m-1}\left(\frac{S_j}{S} - \pi_j\right)^2\right), \tag{19}$$

where $S_i = \int_{E_i} \psi(x)\,dx / \exp(\theta^{(i)})$ for $i = 1, \ldots, m-1$, and $S = \sum_{i=1}^{m-1} S_i + \int_{E_i} \psi(x)\,dx$. Clearly, $v(\theta)$ is continuous in θ, and $\mathcal{V}_M = \{\theta : v(\theta) \le M\}$ for any $M \in (0, \infty)$ forms a compact subset of Θ. Therefore, $\{\mathcal{V}_{M_s}, s \ge 0\}$, $0 < M_0 < M_1 < \cdots$, is an appropriate choice of $\{\mathcal{K}_s, s \ge 0\}$. For the SAMC algorithm, as seen below, $\|H(\theta_k, X_{k+1})\| = \|(I_{\{x_{k+1} \in E_1\}} - \pi_1, \ldots, I_{\{x_{k+1} \in E_{m-1}\}} - \pi_{m-1})^T\|$ is bounded by the constant $\sqrt{2}$, so we can set the drift function $V(x) = 1$. Hence, the initial sample x_0 can be drawn arbitrarily from $\mathcal{X}_0 = \mathcal{X}$, while leaving the condition $\sup_{x \in \mathcal{X}_0} V(x) < \infty$ holds. In summary, SAMC starts with an initial estimate of $\theta_0 \in \mathcal{K}_0$, and a random sample drawn arbitrarily from the space $\mathcal{X}$, and then iterates between the following steps.

SAMC algorithm.

(a) (Sampling.) Simulate a sample x_{k+1} by a single MH update with the target distribution

$$f_{\theta_k}(x) \propto \sum_{i=1}^{m-1} \frac{\psi(x)}{e^{\theta_k^{(i)}}} I_{\{x \in E_i\}} + \psi(x) I_{\{x \in E_m\}}, \tag{20}$$

provided that E_m is nonempty. In practice, E_m can be replaced by any other unempty subregion.

(a.1) Generate y according to a proposal distribution $q(x_k, y)$.

(a.2) Calculate the ratio

$$r = e^{\theta_k^{(J(x_k))} - \theta_k^{(J(y))}} \frac{\psi(y) q(y, x_k)}{\psi(x_k) q(x_k, y)},$$

where $J(z)$ denotes the index of the subregion that the sample z belongs to.

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

(b) (Weight updating.) Set

$$\theta_{k+1/2}^{(i)} = \theta_k^{(i)} + a_{k+1}(I_{\{x_{k+1} \in E_i\}} - \pi_i), \qquad i = 1, \ldots, m-1. \tag{21}$$

(c) (Varying truncation.) If $\theta_{k+1/2} \in \mathcal{K}_{\sigma_k}$, then set $(\theta_{k+1}, x_{k+1}) = (\theta_{k+1/2}, x_{k+1})$ and $\sigma_{k+1} = \sigma_k$; otherwise, set $(\theta_{k+1}, x_{k+1}) = \mathcal{T}(\theta_k, x_k)$ and $\sigma_{k+1} = \sigma_k + 1$, where σ_k and $\mathcal{T}(\cdot, \cdot)$ are as defined in Section 2.

SAMC sampling is driven by its self-adjusting mechanism, which, consequently, implies the superiority of SAMC in sample space exploration. The self-adjusting mechanism can be explained as follows: if a subregion is visited at iteration k, θ_k will be updated accordingly such that the probability that this subregion (other subregions) will be revisited at the next iterations will decrease (increase). Mathematically, if $x_{k+1} \in E_i$, then $\theta_{k+1/2}^{(i)} \leftarrow \theta_k^{(i)} + a_{k+1}(1 - \pi_i)$ and $\theta_{k+1/2}^{(j)} \leftarrow \theta_k^{(j)} - a_{k+1}\pi_j$ for $j \ne i$. Note that the linear

adjustment on θ transforms to a multiplying adjustment on ω. This also explains why SAMC works on the logarithm of ω. Working on the logarithm enables ω to be adjusted quickly according to the distribution of the samples. Otherwise, learning of ω would be very slow due to the linear nature of stochastic approximation. Including π_i in the transformation $\log(\omega_i/\pi_i)$ facilitates our computation, for example, the ratio r in step (a.2).

The self-adjusting mechanism has led to successful applications of SAMC for many hard computational problems, including phylogenetic tree reconstruction [Cheon and Liang (2007, 2009)], neural network training [Liang (2007b)], Bayesian network learning [Liang and Zhang (2009)], among others.

3.2. *Trajectory averaging for SAMC.* To show that the trajectory averaging estimator is asymptotically efficient for SAMC, we assume the following conditions.

(C_1) The MH transition kernel used in the sampling step satisfies the drift condition (A_3).

To ensure the drift condition to be satisfied, Liang, Liu and Carroll (2007) restrict the sample space $\mathcal{X}$ to be a compact set, assume $f(x)$ to be bounded away from 0 and ∞, and choose the proposal distribution $q(x, y)$ to satisfy the local positive condition: for every $x \in \mathcal{X}$, there exist positive ε_1 and ε_2 such that

$$\|x - y\| \le \varepsilon_1 \quad \Longrightarrow \quad q(x, y) \ge \varepsilon_2. \tag{22}$$

If the compactness condition on $\mathcal{X}$ is removed, we may need to impose some constraints on the tails of the target distribution $f(x)$ and the proposal distribution $q(x, y)$ as done by Andrieu, Moulines and Priouret (2005).

(C_2) The sequence $\{a_k\}$ satisfies the following conditions:

$$\sum_{k=1}^{\infty} a_k = \infty, \qquad \lim_{k\to\infty} (k a_k) = \infty,$$

$$\frac{a_{k+1} - a_k}{a_k} = o(a_{k+1}), \qquad \sum_{k=1}^{\infty} \frac{a_k^{(1+\tau)/2}}{\sqrt{k}} < \infty$$

for some $\tau \in (0, 1]$.

For the SAMC algorithm, as previously discussed, $\|H(\theta_k, X_{k+1})\|$ is bounded by the constant $\sqrt{2}$, so we can set $V(x) = 1$ and set α to any a large number in condition (A_3). Furthermore, given a choice of $a_k = O(k^{-\eta})$ for some $\eta \in (1/2, 1)$, there always exists a sequence $\{b_k\}$, for example, $b_k = 2a_k^{(1+\tau)/2}$ for some $\tau \in (0, 1]$, such that the inequality $\|\theta_{k+1/2} - \theta_k\| = \|a_k H(\theta_k, X_{k+1})\| \le$

b_k holds for all iterations. Hence, a specification of the sequence $\{b_k\}$ can be omitted for the SAMC algorithm.

Theorem 3.1 concerns the convergence of SAMC. In the first part, it states that k_{σ_s} is almost surely finite; that is, $\{\theta_k\}$ can be included in a compact set almost surely. In the second part, it states the convergence of θ_k to a solution θ^*. We note that for SAMC, the same convergence result has been established by Liang, Liu and Carroll (2007) under (C_1) and a relaxed condition of (C_2), where $\{a_k\}$ is allowed to decrease at a rate of $O(1/k)$. Since the focus of this paper is on the asymptotic efficiency of $\bar{\theta}_k$, the convergence of $\{\theta_k\}$ is only stated under a slower decreasing rate of $\{a_k\}$. We also note that for SAMC, we have assumed, without loss of generality, that all subregions are unempty. For the empty subregions, no adaptation of $\{\theta_k\}$ occurs for the corresponding components in the run. Therefore, the convergence of $\{\theta_k\}$ should only be measured for the components corresponding to the nonempty subregions.

THEOREM 3.1. *Assume conditions* (C_1) *and* (C_2) *hold. Then there exists (a.s.) a number, denoted by* σ_s*, such that* $k_{\sigma_s} < \infty$*,* $k_{\sigma_s+1} = \infty$*, and* $\{\theta_k\}$ *given by the SAMC algorithm has no truncation for* $k \geq k_{\sigma_s}$*, that is,*

$$\theta_{k+1} = \theta_k + a_k H(\theta_k, x_{k+1}) \qquad \forall k \geq k_{\sigma_s} \tag{23}$$

and

$$\theta_k \to \theta^* \qquad a.s., \tag{24}$$

where $H(\theta_k, x_{k+1}) = (I_{\{x_{k+1}\in E_1\}} - \pi_1, \ldots, I_{\{x_{k+1}\in E_{m-1}\}} - \pi_{m-1})^T$*, and* $\theta^* = (\log(\omega_1/\pi_1) - \log(\omega_m/\pi_m), \ldots, \log(\omega_{m-1}/\pi_{m-1}) - \log(\omega_m/\pi_m))^T$.

Theorem 3.2 concerns the asymptotic normality and efficiency of $\bar{\theta}_k$.

THEOREM 3.2. *Assume conditions* (C_1) *and* (C_2)*. Then* $\bar{\theta}_k$ *is asymptotically efficient; that is,*

$$\sqrt{k}(\bar{\theta}_k - \theta^*) \longrightarrow N(\mathbf{0}, \Gamma) \qquad \text{as } k \to \infty,$$

where $\Gamma = F^{-1}Q(F^{-1})^T$*,* $F = \partial h(\theta^*)/\partial\theta$ *is negative definite and* $Q = \lim_{k\to\infty} E(e_k e_k^T)$.

The above theorems address some theoretical issues of SAMC. For practical issues, please refer to Liang, Liu and Carroll (2007), where issues, such as how to partition the sample space, how to choose the desired sampling distribution, and how to diagnose the convergence, have been discussed at length. An issue particularly related to the trajectory averaging estimator is

the length of the burn-in period. To remove the effect of the early iterates, the following estimator:

$$\bar{\theta}_k^{(b)} = \frac{1}{k-k_0} \sum_{i=k_0+1}^{k} \theta_i,$$

instead of $\bar{\theta}_k$, is often used in practice, where k_0 is the so-called length of the burn-in period. It is obvious that the choice of k_0 should be based on the diagnosis for the convergence of the simulation. Just like monitoring convergence of MCMC simulations, monitoring convergence of SAMC simulations should be based on multiple runs [Liang, Liu and Carroll (2007)]. In practice, if only a single run was made, we suggest to look at the plot of $\widehat{\boldsymbol{\pi}}$ to choose k_0 from where $\widehat{\boldsymbol{\pi}}_k$ has been approximately stable. Here, we denote by $\widehat{\boldsymbol{\pi}}_k$ the sampling frequencies of the respective subregions realized by iteration k. It follows from Theorem 3.1 that $\widehat{\boldsymbol{\pi}}_k \to \boldsymbol{\pi}$ when the number of iterations, k, becomes large.

Trajectory averaging can directly benefit one's inference in many applications of SAMC. A typical example is Bayesian model selection, where the ratio ω_i/ω_j just corresponds to the Bayesian factor of two models if one partitions the sample space according to the model index and imposes an uniform prior on the model space as done in Liang (2009). Another example is inference for the spatial models with intractable normalizing constants, for which Liang, Liu and Carroll (2007) has demonstrated how SAMC can be used to estimate the normalizing constants for these models and how the estimate can then be used for inference of the model parameters. An improved estimate of the normalizing constant function would definitely benefit one's inference for the model.

4. Trajectory averaging for a stochastic approximation MLE algorithm. Consider the standard missing data problem:

- y is the observed incomplete data.
- $f(x,\theta)$ is the complete data likelihood, that is, the likelihood of the complete data (x,y) obtained by augmenting the observed data y with the missing data x. The dependence of $f(x,\theta)$ on y is here implicit.
- $p(x,\theta)$ is the predictive distribution of the missing data x given the observed data y, that is, the predictive likelihood.

Our goal is to find the maximum likelihood estimator of θ. This problem has been considered by a few authors under the framework of stochastic approximation; see, for example, Younes (1989), Gu and Kong (1998) and Delyon, Lavielle and Moulines (1999). A basic algorithm proposed by Younes (1989) for the problem can be written as

$$\theta_{k+1} = \theta_k + a_k \partial_\theta \log f(X_{k+1}, \theta_k), \tag{25}$$

where the missing data X_{k+1} can be imputed using a MCMC algorithm, such as the Metropolis–Hastings algorithm. Under standard regularity conditions, we have

$$h(\theta) = E_\theta[\partial_\theta \log f(X, \theta)] = \partial_\theta l(\theta),$$

where $l(\theta)$ is the log-likelihood function of the incomplete data.

To show that the trajectory averaging estimator is asymptotically efficient for a varying truncation version of the algorithm (25), we assume (A_3), (A_4) and some regularity conditions for the distribution $f(x, \theta)$. The conditions (A_1) and (A_2) can be easily verified with the following settings:

- The Lyapunov function $v(\theta)$ can be chosen as $v(\theta) = -l(\theta) + C$, where C is chosen such that $v(\theta) > 0$. Thus,

  $$\langle \nabla v(\theta), h(\theta) \rangle = -\|\partial_\theta l(\theta)\|^2.$$

  The set of stationary points of (25), $\{\theta : \langle \nabla v(\theta), h(\theta) \rangle = 0\}$, coincides with the set of the solutions $\{\theta : \partial_\theta l(\theta) = 0\}$. Then the condition (A_1) can be verified by verifying that $l(\theta)$ is continuously differentiable (this is problem dependent).
- The matrix F trivially is the Hessian matrix of $l(\theta)$. Then (A_2) can be verified using the Taylor expansion.

In summary, we have the following theorem.

THEOREM 4.1. *Assume conditions* (A_3) *and* (A_4) *hold. Then the estimator* $\bar{\theta}_k$ *generated by a varying truncation version of algorithm (25) is asymptotically efficient.*

In practice, to ensure the drift condition to be satisfied, we may follow Andrieu, Moulines and Priouret (2005) to impose some constraints on the tails of the distribution $f(x, \theta)$ and the proposal distribution $q(x, y)$. Alternatively, we can follow Liang, Liu and Carroll (2007) to choose a proposal satisfying the local positive condition (22) and to restrict the sample space $\mathcal{X}$ to be compact. For example, we may set $\mathcal{X}$ to a huge space, say, $\mathcal{X} = [-10^{100}, 10^{100}]^{d_x}$. As a practical matter, this is equivalent to setting $\mathcal{X} = \mathbb{R}^{d_x}$.

5. Conclusion. In this paper, we have shown that the trajectory averaging estimator is asymptotically efficient for a general stochastic approximation MCMC algorithm under mild conditions, and then applied this result to the stochastic approximation Monte Carlo algorithm and a stochastic approximation MLE algorithm.

The main difference between this work and the work published in the literature, for example, Polyak and Juditsky (1992) and Chen (1993), are

at the conditions on the observation noise. In the literature, it is usually assumed directly that the observation noise has the decomposition $\varepsilon_k = e_k + \nu_k$, where $\{e_k\}$ forms a martingale difference sequence and ν_k is a higher order term of $o(a_k^{1/2})$. As shown in Lemma A.5, the stochastic approximation MCMC algorithm does not satisfy this decomposition.

APPENDIX A: PROOFS OF THEOREMS 2.2 AND 2.3

Lemma A.1 is a partial restatement of Proposition 6.1 of Andrieu, Moulines and Priouret (2005).

LEMMA A.1. *Assume condition* (A_3) *holds. Then the following results hold:*

(B_1) *For any* $\theta \in \Theta$, *the Markov kernel* P_θ *has a single stationary distribution* f_θ. *In addition,* $H : \Theta \times \mathcal{X} \to \Theta$ *is measurable for all* $\theta \in \Theta$, $\int_{\mathcal{X}} \|H(\theta, x)\| f_\theta(x)\, dx < \infty$.

(B_2) *For any* $\theta \in \Theta$, *the Poisson equation* $u(\theta, x) - P_\theta u(\theta, x) = H(\theta, x) - h(\theta)$ *has a solution* $u(\theta, x)$, *where* $P_\theta u(\theta, x) = \int_{\mathcal{X}} u(\theta, x') P_\theta(x, x')\, dx'$. *There exist a function* $V : \mathcal{X} \to [1, \infty)$ *such that* $\{x \in \mathcal{X}, V(x) < \infty\} \neq \varnothing$, *and a constant* $\beta \in (0, 1]$ *such that for any compact subset* $\mathcal{K} \subset \Theta$, *the following holds:*

$$
\begin{aligned}
&\text{(i)} \quad \sup_{\theta \in \mathcal{K}} \|H(\theta, x)\|_V < \infty, \\
&\text{(ii)} \quad \sup_{\theta \in \mathcal{K}} (\|u(\theta, x)\|_V + \|P_\theta u(\theta, x)\|_V) < \infty, \\
&\text{(iii)} \quad \sup_{(\theta, \theta') \in \mathcal{K} \times \mathcal{K}} \|\theta - \theta'\|^{-\beta} (\|u(\theta, x) - u(\theta', x)\|_V \\
&\qquad\qquad + \|P_\theta u(\theta, x) - P_{\theta'} u(\theta', x)\|_V) < \infty.
\end{aligned} \tag{26}
$$

Lemma A.2 is a restatement of Proposition 5.1 of Andrieu, Moulines and Priouret (2005).

LEMMA A.2. *Assume conditions* (A_1), (A_3) *and* (A_4) *hold. Let* $\mathcal{X}_0 \subset \mathcal{X}$ *be such that* $\sup_{x \in \mathcal{X}_0} V(x) < \infty$ *and that* $\mathcal{K}_0 \subset \mathcal{V}_{M_0}$, *where* $\mathcal{V}_0$ *is defined in* (A_1). *Then* $\sup_k E[V^\alpha(X_k) I(k \geq k_{\sigma_s})] < \infty$, *where* $\alpha \geq 2$ *is defined in condition* (A_3) *and* k_{σ_s} *is defined in Theorem 2.1.*

Lemma A.3 is a restatement of Corollary 2.1.10 of Duflo (1997), pages 46 and 47.

LEMMA A.3. *Let* $\{S_{ni}, \mathcal{G}_{ni}, 1 \leq i \leq k_n, n \geq 1\}$ *be a zero-mean, square-integrable martingale array with differences* υ_{ni}, *where* $\mathcal{G}_{ni}$ *denotes the* σ*-field. Suppose that the following assumptions apply:*

(i) *The σ-fields are nested: $\mathcal{G}_{ni} \subseteq \mathcal{G}_{n+1,i}$ for $1 \le i \le k_n$, $n \ge 1$.*

(ii) $\sum_{i=1}^{k_n} E(\upsilon_{ni}\upsilon_{ni}^T|\mathcal{G}_{n,i-1}) \to \Lambda$ *in probability, where Λ is a positive definite matrix.*

(iii) *For any $\varepsilon > 0$, $\sum_{i=1}^{k_n} E[\|\upsilon_{ni}\|^2 I_{(\|\upsilon_{ni}\|\ge\varepsilon)}|\mathcal{G}_{n,i-1}] \to 0$ in probability.*

Then $S_{nk_n} = \sum_{i=1}^{k_n} \upsilon_{ni} \to N(0,\Lambda)$ in distribution.

DEFINITION A.1. For $\varrho \in (0,\infty)$, a sequence $\{X_n, n \ge 1\}$ of random variables is said to be residually Cesàro ϱ-integrable [RCI(ϱ), in short] if

$$\sup_{n\ge1} \frac{1}{n} \sum_{i=1}^{n} E|X_i| < \infty$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E(|X_i| - i^{\varrho}) I(|X_i| > i^{\varrho}) = 0.$$

Lemma A.4 is a restatement of Theorem 2.1 of Chandra and Goswami (2006).

LEMMA A.4. *Let $\{X_n, n \ge 1\}$ be a sequence of nonnegative random variables satisfying $E(X_iX_j) \le E(X_i)E(X_j)$ for all $i \ne j$ and let $S_n = \sum_{i=1}^n X_i$. If $\{X_n, n \ge 1\}$ is* RCI(ϱ) *for some $\varrho \in (0,1)$, then*

$$\frac{1}{n}[S_n - E(S_n)] \to 0 \qquad \text{in probability.}$$

LEMMA A.5. *Assume conditions* (A_1), (A_3) *and* (A_4) *hold. Let $\mathcal{X}_0 \subset \mathcal{X}$ be such that $\sup_{x\in\mathcal{X}_0} V(x) < \infty$ and that $\mathcal{K}_0 \subset \mathcal{V}_{M_0}$, where $\mathcal{V}_0$ is defined in* (A_1). *If $k_{\sigma_s} < \infty$, which is defined in Theorem 2.1, then there exist $\mathbb{R}^d$-valued random processes $\{e_k\}_{k\ge k_{\sigma_s}}$, $\{\nu_k\}_{k\ge k_{\sigma_s}}$ and $\{\varsigma_k\}_{k\ge k_{\sigma_s}}$ defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that:*

(i) $\varepsilon_k = e_k + \nu_k + \varsigma_k$ *for $k \ge k_{\sigma_s}$.*

(ii) $\{e_k\}_{k\ge k_{\sigma_s}}$ *is a martingale difference sequence, and $\frac{1}{\sqrt{n}} \sum_{k=k_{\sigma_s}}^n e_k \longrightarrow N(0,Q)$ in distribution, where $Q = \lim_{k\to\infty} E(e_ke_k^T)$.*

(iii) $\frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^k E\|\nu_i\| \to 0$, *as $k \to \infty$.*

(iv) $E\|\sum_{i=k_{\sigma_s}}^k a_i\varsigma_i\| \to 0$, *as $k \to \infty$.*

PROOF. (i) Let $\varepsilon_{k_{\sigma_s}} = \nu_{k_{\sigma_s}} = \varsigma_{k_{\sigma_s}} = 0$, and

$$e_{k+1} = u(\theta_k, x_{k+1}) - P_{\theta_k}u(\theta_k, x_k),$$
$$\nu_{k+1} = [P_{\theta_{k+1}}u(\theta_{k+1}, x_{k+1}) - P_{\theta_k}u(\theta_k, x_{k+1})]$$

$$+\frac{a_{k+2}-a_{k+1}}{a_{k+1}}P_{\theta_{k+1}}u(\theta_{k+1},x_{k+1}), \tag{27}$$

$$\tilde{\varsigma}_{k+1}=a_{k+1}P_{\theta_k}u(\theta_k,x_k),$$

$$\varsigma_{k+1}=\frac{1}{a_{k+1}}(\tilde{\varsigma}_{k+1}-\tilde{\varsigma}_{k+2}).$$

It is easy to verify that (i) holds by noticing the Poisson equation given in (B_2).

(ii) By (27), we have

$$E(e_{k+1}|\mathcal{F}_k)=E(u(\theta_k,x_{k+1})|\mathcal{F}_k)-P_{\theta_k}u(\theta_k,x_k)=0,$$

where $\{\mathcal{F}_k\}_{k\geq k_{\sigma_s}}$ is a family of σ-algebras satisfying $\sigma\{\theta_{k_{\sigma_s}},x_{k_{\sigma_s}}\}\subseteq\mathcal{F}_0$ and $\sigma\{\theta_{k_{\sigma_s}},\theta_{k_{\sigma_s}+1},\ldots,\theta_k;x_{k_{\sigma_s}},x_{k_{\sigma_s+1}},\ldots,x_k\}\subseteq\mathcal{F}_k\subseteq\mathcal{F}_{k+1}$ for all $k\geq k_{\sigma_s}$. Hence, $\{e_k\}_{k\geq k_{\sigma_s}}$ forms a martingale difference sequence.

When $k_{\sigma_s}<\infty$, there exists a compact set $\mathcal{K}$ such that $\theta_k\in\mathcal{K}$ for all $k\geq 0$. Following from Lemmas A.1 and A.2, $\{e_k\}_{k\geq k_{\sigma_s}}$ is e_k is uniformly square integrable with respect to k, and the martingale $s_n=\sum_{k=1}^n e_k$ is square integrable for all n.

By (27), we have

$$\begin{aligned}E(e_{k+1}e_{k+1}^T|\mathcal{F}_k)&=E[u(\theta_k,x_{k+1})u(\theta_k,x_{k+1})^T|\mathcal{F}_k]\\&\quad-P_{\theta_k}u(\theta_k,x_k)P_{\theta_k}u(\theta_k,x_k)^T\\&\triangleq l(\theta_k,x_k).\end{aligned} \tag{28}$$

Following from Lemmas A.1 and A.2, $\|l(\theta_k,x_k)\|$ is uniformly integrable with respect to k. Hence, $\{l(\theta_k,x_k),k\geq k_{\sigma_s}\}$ is RCI(ϱ) for any $\varrho>0$ (Definition A.1). Since $\{E(e_{k+1}e_{k+1}^T|\mathcal{F}_k)-E(e_{k+1}e_{k+1}^T)\}$ forms a martingale difference sequence, the correlation coefficient $\mathrm{Corr}(l(\theta_i,x_i),\ l(\theta_j,x_j))=0$ for all $i\neq j$. By Lemma A.4, we have, as $n\to\infty$,

$$\frac{1}{n}\sum_{k=k_{\sigma_s}}^{n}l(\theta_k,x_k)\to\frac{1}{n}\sum_{k=k_{\sigma_s}}^{n}El(\theta_k,x_k)\qquad\text{in probability.} \tag{29}$$

Now we show that $El(\theta_k,x_k)$ also converges. It follows from (A_1) and (B_2) that $l(\theta,x)$ is continuous in θ. By the convergence of θ_k, we can conclude that $l(\theta_k,x)$ converges to $l(\theta^*,x)$ for any $x\in\mathcal{X}$. Following from Lemmas A.1, A.2 and Lebesgue's dominated convergence theorem, $El(\theta_k,x_k)$ converges to $El(\theta^*,x)$. Combining with (29), we obtain

$$\frac{1}{n}\sum_{k=k_{\sigma_s}}^{n}l(\theta_k,x_k)\to El(\theta^*,x)=\lim_{k\to\infty}E(e_ke_k^T)\qquad\text{in probability.} \tag{30}$$

Since $\|e_k\|$ can be uniformly bounded by an integrable function $cV(x)$, the Lindeberg condition is satisfied, that is,

$$\sum_{i=k_{\sigma_s}}^{n} E\left[\frac{\|e_i\|^2}{n} I_{(\|e_i\|/\sqrt{n}\geq\varepsilon)}\Big|\mathcal{F}_{i-1}\right] \to 0 \qquad \text{as } n\to\infty.$$

Following from Lemma A.3, we have $\sum_{i=k_{\sigma_s}}^{n} e_i/\sqrt{n} \to N(0,Q)$ by identifying $e_i/\sqrt{n}$ to υ_{ni}, n to k_n, and $\mathcal{F}_i$ to $\mathcal{G}_{ni}$.

(iii) By condition (A_4), we have

$$\frac{a_{k+2}-a_{k+1}}{a_{k+1}} = o(a_{k+2}).$$

By (27) and (26), there exists a constant c_1 such that the following inequality holds:

$$\|\nu_{k+1}\|_V \leq c_1\|\theta_{k+1}-\theta_k\| + o(a_{k+2}) = c_1\|a_k H(\theta_k, x_{k+1})\| + o(a_{k+2}),$$

which implies, by (26), that there exists a constant c_2 such that

$$\|\nu_{k+1}\|_{V^2} \leq c_2 a_k. \tag{31}$$

Since $V(x)$ is square integrable, ν_k is uniformly integrable with respect to k and there exists a constant c_3 such that

$$\sum_{k=k_{\sigma_s}}^{\infty} \frac{E\|\nu_k\|}{\sqrt{k}} \leq c_3 \sum_{k=k_{\sigma_s}}^{\infty} \frac{a_k}{\sqrt{k}} < \infty,$$

where the last inequality follows from condition (A_4). Therefore, (iii) holds by Kronecker's lemma.

(iv) A straightforward calculation shows that

$$\sum_{i=k_{\sigma_s}}^{k} a_i\varsigma_i = -\tilde{\varsigma}_{k+1} = -a_{k+1}P_{\theta_k}u(\theta_k, x_k).$$

By Lemmas A.1 and A.2, $E\|P_{\theta_k}u(\theta_k,x_k)\|$ is uniformly bounded with respect to k. Therefore, (iv) holds. □

By Theorem 2.1, we have

$$\theta_{k+1}-\theta^* = (\theta_k-\theta^*) + a_k h(\theta_k) + a_k\varepsilon_{k+1} \qquad \forall k \geq k_{\sigma_s}. \tag{32}$$

To facilitate the theoretical analysis for the random process $\{\theta_k\}$, we define a reduced random process: $\{\tilde{\theta}_k\}_{k\geq 0}$, where

$$\tilde{\theta}_k = \begin{cases} \theta_k + \tilde{\varsigma}_k, & k > k_{\sigma_s}, \\ \theta_k, & 0 \leq k \leq k_{\sigma_s}, \end{cases} \tag{33}$$

which is equivalent to set $\tilde{\varsigma}_k = 0$ for all $k = 0, \ldots, k_{\sigma_s}$. For convenience, we also define

$$\tilde{\varepsilon}_k = e_k + \nu_k, \qquad k > k_{\sigma_s}. \tag{34}$$

It is easy to verify that

$$\begin{aligned}\tilde{\theta}_{k+1} - \theta^* &= (I + a_k F)(\tilde{\theta}_k - \theta^*) \\ &\quad + a_k(h(\theta_k) - F(\tilde{\theta}_k - \theta^*)) + a_k\tilde{\varepsilon}_{k+1} \qquad \forall k \geq k_{\sigma_s},\end{aligned} \tag{35}$$

which implies

$$\begin{aligned}\tilde{\theta}_{k+1} - \theta^* &= \Phi_{k,k_{\sigma_s}}(\tilde{\theta}_{k_{\sigma_s}} - \theta^*) + \sum_{j=k_{\sigma_s}}^{k} \Phi_{k,j+1} a_j \tilde{\varepsilon}_{j+1} \\ &\quad + \sum_{j=k_{\sigma_s}}^{k} \Phi_{k,j+1} a_j (h(\theta_j) - F(\tilde{\theta}_j - \theta^*)) \qquad \forall k \geq k_{\sigma_s},\end{aligned} \tag{36}$$

where $\Phi_{k,j} = \prod_{i=j}^{k}(I + a_i F)$ if $k \geq j$, and $\Phi_{j,j+1} = I$, and I denotes the identity matrix.

For γ specified in (A_2) and a deterministic integer k_0, define the stopping time $\mu = \min\{j : j \geq k_0, \|\theta_j - \theta^*\| \geq \gamma\}$ if $\|\theta_{k_0} - \theta^*\| < \gamma$ and 0 if $\|\theta_{k_0} - \theta^*\| \geq \gamma$. Define

$$A = \{i : k_{\sigma_s} < k_0 \leq i < \mu\}, \tag{37}$$

and let $I_A(k)$ denote the indicator function; $I_A(k) = 1$ if $k \in A$ and 0 otherwise. Therefore, for all $k \geq k_0$,

$$\begin{aligned}&(\tilde{\theta}_{k+1} - \theta^*) I_A(k+1) \\ &\quad = \Phi_{k,k_0}(\tilde{\theta}_{k_0} - \theta^*) I_A(k+1) + \left[\sum_{j=k_0}^{k} \Phi_{k,j+1} a_j \tilde{\varepsilon}_{j+1} I_A(j)\right] I_A(k+1) \\ &\qquad + \left[\sum_{j=k_0}^{k} \Phi_{k,j+1} a_j (h(\theta_j) - F(\tilde{\theta}_j - \theta^*)) I_A(j)\right] I_A(k+1).\end{aligned} \tag{38}$$

Including the terms $I_A(j)$ in (38) facilitates our use of some results published in Chen (2002) in the later proofs, but it does not change equality of (38). Note that if $I_A(k+1) = 1$, then $I_A(j) = 1$ for all $j = k_0, \ldots, k$.

LEMMA A.6. (i) *The following estimate takes place:*

$$\frac{a_j}{a_k} \leq \exp\left(o(1)\sum_{i=j}^{k} a_i\right) \qquad \forall k \geq j, \forall j \geq 1, \tag{39}$$

where $o(1)$ *denotes a magnitude that tends to zero as* $j \to \infty$.

(ii) *Let c be a positive constant, then there exists another constant c_1 such that*

$$\sum_{i=1}^{k} a_i^r \exp\left(-c\sum_{j=i+1}^{k} a_j\right) \le c_1 \qquad \forall k \ge 1, \forall r \ge 1. \tag{40}$$

(iii) *There exist constants $c_0 > 0$ and $c > 0$ such that*

$$\|\Phi_{k,j}\| \le c_0 \exp\left\{-c\sum_{i=j}^{k} a_i\right\} \qquad \forall k \ge j, \forall j \ge 0. \tag{41}$$

(iv) *Let $G_{k,j} = \sum_{i=j}^{k}(a_{j-1} - a_i)\Phi_{i-1,j} + F^{-1}\Phi_{k,j}$. Then $G_{k,j}$ is uniformly bounded with respect to both k and j for $1 \le j \le k$, and*

$$\frac{1}{k}\sum_{j=1}^{k} \|G_{k,j}\| \longrightarrow 0 \qquad \text{as } k \to \infty. \tag{42}$$

PROOF. Parts (i) and (iv) are a restatement of Lemma 3.4.1 of Chen (2002). The proof of part (ii) can be found in the proof of Lemma 3.3.2 of Chen (2002). The proof of part (iii) can be found in the proof of Lemma 3.1.1 of Chen (2002). □

LEMMA A.7. *If conditions* (A_1)–(A_4) *hold, then*

$$\frac{1}{a_{k+1}} E\|(\theta_{k+1} - \theta^*) I_A(k+1)\|^2$$

is uniformly bounded with respect to k, where the set A is as defined in (37).

PROOF. By (33) and (27), we have

$$\begin{aligned}\frac{1}{a_{k+1}}\|\theta_{k+1} - \theta^*\|^2 &= \frac{1}{a_{k+1}}\|\tilde{\theta}_{k+1} - \theta^* - \tilde{\varsigma}_{k+1}\|^2 \\ &\le \frac{2}{a_{k+1}}\|\tilde{\theta}_{k+1} - \theta^*\|^2 + 2a_{k+1}\|P_{\theta_k} u(\theta_k, x_k)\|^2.\end{aligned}$$

Following from (B_2) and Lemma A.2, it is easy to see that $E\|P_{\theta_k} u(\theta_k, x_k)\|^2$ is uniformly bounded with respect to k. Hence, to prove the lemma, it suffices to prove that $\frac{1}{a_{k+1}} E\|(\tilde{\theta}_{k+1} - \theta^*) I_A(k+1)\|^2$ is uniformly bounded with respect to k.

By (33), (A_2) and (B_2), there exist constants c_1 and c_2 such that

$$\begin{aligned}&\|h(\theta_j) - F(\tilde{\theta}_j - \theta^*)\| I_A(j) \\ &\quad = \|h(\theta_j) - F(\theta_j - \theta^*) - F\tilde{\varsigma}_j\| I_A(j) \\ &\quad \le \|h(\theta_j) - F(\theta_j - \theta^*)\| I_A(j) + c_2 a_j \|P_{\theta_{j-1}} u(\theta_{j-1}, x_{j-1})\| \\ &\quad \le c_1 \|\theta_j - \theta^*\|^{1+\rho} + c_2 a_j \|P_{\theta_{j-1}} u(\theta_{j-1}, x_{j-1})\|.\end{aligned} \tag{43}$$

In addition, we have

$$E\|\tilde{\theta}_{k_0}-\theta^*\|^2 I_A(k_0) = E\|\theta_{k_0}-\theta^*+\tilde{\varsigma}_{k_0}\|^2 I_A(k_0) \le 2\|\theta_{k_0}-\theta^*\|^2 I_A(k_0) + 2E\|\tilde{\varsigma}_{k_0}\|^2. \tag{44}$$

It is easy to see from (26) and (27) that $\tilde{\varsigma}_{k_0}$ is square integrable. Hence, following from (37), there exists a constant $\tilde{\gamma}$ such that

$$E\|\tilde{\theta}_{k_0}-\theta^*\|^2 I_A(k_0) \le \tilde{\gamma}. \tag{45}$$

By (38), (41), (43) and (45), and following Chen [(2002), page 141] we have

$$\begin{aligned}
&\frac{1}{a_{k+1}} E\|(\tilde{\theta}_{k+1}-\theta^*) I_A(k+1)\|^2 \\
&\quad \le \frac{5c_0\tilde{\gamma}}{a_{k+1}} \exp\left(-2c\sum_{i=k_0}^{k} a_i\right) \\
&\qquad + \frac{5c_0^2}{a_{k+1}} \sum_{i=k_0}^{k}\sum_{j=k_0}^{k} \left[\exp\left(-c\sum_{s=j+1}^{k} a_s\right) a_j \exp\left(-c\sum_{s=i+1}^{k} a_s\right) a_i \|E e_{i+1} e_{j+1}^T\|\right] \\
&\qquad + \frac{5c_0^2}{a_{k+1}} \sum_{i=k_0}^{k}\sum_{j=k_0}^{k} \left[\exp\left(-c\sum_{s=j+1}^{k} a_s\right) a_j \exp\left(-c\sum_{s=i+1}^{k} a_s\right) a_i E\|\nu_{i+1}\nu_{j+1}^T\|\right] \\
&\qquad + \frac{5c_0^2c_2^2}{a_{k+1}} \sum_{i=k_0}^{k}\sum_{j=k_0}^{k} \left[\exp\left(-c\sum_{s=j+1}^{k} a_s\right) a_j^2 \exp\left(-c\sum_{s=i+1}^{k} a_s\right)\right. \\
&\qquad\qquad \left. \times a_i^2 E\|P_{\theta_{i-1}} u(\theta_{i-1}, x_{i-1})(P_{\theta_{j-1}} u(\theta_{j-1}, x_{j-1}))^T\|\right] \\
&\qquad + \frac{5c_0^2c_1^2}{a_{k+1}} E\left[\sum_{j=k_0}^{k} \exp\left(-c\sum_{s=j+1}^{k} a_s\right) a_j \|\theta_j-\theta^*\|^{1+\rho} I_A(j)\right]^2 \\
&\quad \stackrel{\triangle}{=} I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned}$$

By (39), there exists a constant c_3 such that such that

$$\|I_1\| \le \frac{5c_0c_3\tilde{\gamma}}{a_{k_0}} \exp\left(o(1)\sum_{i=k_0}^{k+1} a_i\right) \exp\left(-2c\sum_{i=k_0}^{k} a_i\right),$$

where $o(1)\to 0$ as $k_0\to\infty$. This implies that $o(1)-2c<0$ if k_0 is large enough. Hence, I_1 is bounded if k_0 is large enough.

By (39) and (40), for large enough k_0, there exists a constant c_4 such that

$$\sum_{j=k_0}^{k} \frac{a_j^2}{a_{k+1}} \exp\left(-c \sum_{s=j+1}^{k} a_s\right) \le \sum_{j=k_0}^{k} a_j \exp\left(-\frac{c}{2} \sum_{s=j+1}^{k} a_s\right) \le c_4. \tag{46}$$

Since $\{e_i\}$ forms a martingale difference sequence (Lemma A.5),

$$Ee_i e_j^T = E(E(e_i|\mathcal{F}_{i-1})e_j^T) = 0 \qquad \forall i > j,$$

which implies that

$$\begin{aligned} I_2 &= \frac{5c_0^2}{a_{k+1}} \sum_{i=k_0}^{k} \left[a_i^2 \exp\left(-2c \sum_{s=j+1}^{k} a_s\right) E\|e_i\|^2 \right] \\ &\le 5c_0^2 \sup_i E\|e_i\|^2 \sum_{i=k_0}^{k} \left[a_i^2 \exp\left(-2c \sum_{s=j+1}^{k} a_s\right) \right]. \end{aligned}$$

Since $\{\|e_i\|, i \ge 1\}$ is uniformly bounded by a function $cV(x)$ which is square integrable, $\sup_i E\|e_i\|^2$ is bounded by a constant. Furthermore, by (40), I_2 is uniformly bounded with respect to k.

By (27), (26) and condition (A_4), there exist a constant c_0 and a constant $\tau \in (0,1)$ such that the following inequality holds:

$$\|\nu_{k+1}\|_V \le c_0\|\theta_{k+1} - \theta_k\| + o(a_{k+2}) \le c_0 b_k + o(a_{k+2}) = O(a_k^{(1+\tau)/2}). \tag{47}$$

This, by (B_1) and the Cauchy–Schwarz inequality, further implies that there exists a constant c_0' such that

$$E\|\nu_{i+1}\nu_{j+1}^T\| \le c_0' a_i^{(1+\tau)/2} a_j^{(1+\tau)/2}. \tag{48}$$

Therefore, there exists a constant c_5 such that

$$\begin{aligned} I_3 &= 5c_0^2 \sum_{i=k_0}^{k} \sum_{j=k_0}^{k} \left[\exp\left(-c \sum_{s=j+1}^{k} a_s\right) \frac{a_j}{\sqrt{a_{k+1}}} \right. \\ &\qquad \left. \times \exp\left(-c \sum_{s=i+1}^{k} a_s\right) \frac{a_i}{\sqrt{a_{k+1}}} O(a_i^{(1+\tau)/2}) O(a_j^{(1+\tau)/2}) \right] \\ &\le 5c_0^2 c_5 \sum_{i=k_0}^{k} \sum_{j=k_0}^{k} \left[\exp\left(-\frac{c}{2} \sum_{s=j+1}^{k} a_s\right) {a_j}^{1/2} \right. \\ &\qquad \left. \times \exp\left(-\frac{c}{2} \sum_{s=i+1}^{k} a_s\right) {a_i}^{1/2} a_i^{(1+\tau)/2} a_j^{(1+\tau)/2} \right] \\ &= 5c_0^2 c_5 \left\{ \sum_{j=k_0}^{k} \left[{a_j}^{1+\tau/2} \exp\left(-\frac{c}{2} \sum_{s=j+1}^{k} a_s\right) \right] \right\}^2. \end{aligned}$$

By (40), I_3 is uniformly bounded with respect to k.

Following from Lemmas A.1 and A.2, $E\|P_{\theta_{i-1}}u(\theta_{i-1},x_{i-1})(P_{\theta_{j-1}}u(\theta_{j-1},x_{j-1}))^T\|$ is uniformly bounded with respect to k. Therefore, there exists a constant c_6 such that

$$I_4 = 5c_0^2c_2^2c_6\sum_{i=k_0}^{k}\sum_{j=k_0}^{k}\left[\exp\left(-c\sum_{s=j+1}^{k}a_s\right)\frac{a_j^2}{\sqrt{a_{k+1}}}\exp\left(-c\sum_{s=i+1}^{k}a_s\right)\frac{a_i^2}{\sqrt{a_{k+1}}}\right]$$
$$\le 5c_0^2c_2^2c_6\left\{\sum_{j=k_0}^{k}\left[{a_j}^{3/2}\exp\left(-\frac{c}{2}\sum_{s=j+1}^{k}a_s\right)\right]\right\}^2.$$

By (40), I_4 is uniformly bounded with respect to k.

The proof for the uniform boundedness of I_5 can be found in the proof of Lemma 3.4.3 of Chen (2002), pages 143 and 144. □

LEMMA A.8. *If conditions* (A_1)–(A_4) *hold, then as* $k\to\infty$,

$$\frac{1}{\sqrt{k}}\sum_{i=k_{\sigma_s}}^{k}\|h(\theta_i)-F(\tilde{\theta}_i-\theta^*)\|\longrightarrow 0 \qquad \textit{in probability.}$$

PROOF. By (33) and (27), there exists a constant c such that

$$\frac{1}{\sqrt{k}}\sum_{i=k_{\sigma_s}}^{k}\|h(\theta_i)-F(\tilde{\theta}_i-\theta^*)\|$$
$$\le \frac{1}{\sqrt{k}}\sum_{i=k_{\sigma_s}}^{k}\|h(\theta_i)-F(\theta_i-\theta^*)\|+\frac{c}{\sqrt{k}}\sum_{i=k_{\sigma_s}}^{k}a_i\|P_{\theta_{i-1}}u(\theta_{i-1},x_{i-1})\|$$
$$\triangleq I_1+I_2.$$

To prove the lemma, it suffices to prove that I_1 and I_2 both converge to zero in probability as $k\to\infty$.

Following from Lemmas A.1 and A.2, $E\|P_{\theta_k}u(\theta_k,x)\|$ is uniformly bounded for all $k\ge k_{\sigma_s}$. This implies, by condition (A_4), there exists a constant c such that

$$\sum_{i=1}^{\infty}\frac{a_iE\|P_{\theta_{i-1}}u(\theta_{i-1},x_{i-1})\|}{\sqrt{i}}<c\sum_{i=1}^{\infty}\frac{a_i}{\sqrt{i}}<\infty.$$

By Kronecker's lemma, $E(I_2)\to 0$, and thus $I_2\to 0$ in probability.

The convergence $I_1\to 0$ can be established as in Chen [(2002), Lemma 3.4.4] using the condition (A_2) and Lemma A.7. □

Proof of Theorem 2.2. By Theorem 2.1, θ_k converges to the zero point θ^* almost surely and

$$\theta_{k+1} = \theta_k + a_k H(\theta_k, x_{k+1}) \qquad \forall k \geq k_{\sigma_s}.$$

Consequently, we have, by (33),

$$\begin{aligned} \sqrt{k}(\bar{\theta}_k - \theta^*) &= o(1) + \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} (\theta_i - \theta^*) \\ &= o(1) + \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} (\tilde{\theta}_i - \theta^*) - \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} \tilde{\varsigma}_i, \end{aligned} \tag{49}$$

where $o(1) \to 0$ as $k \to \infty$.

Condition (A_4) implies $\frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} a_i \to 0$ by Kronecker's lemma. Following Lemmas A.1 and A.2, there exists a constant c such that

$$\frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} E\|\tilde{\varsigma}_i\| \leq \frac{c}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} a_{i+1} \to 0. \tag{50}$$

Therefore, $\frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} \tilde{\varsigma}_i \to 0$ in probability as $k \to \infty$.

By (36), (49) and (50), we have

$$\begin{aligned} \sqrt{k}(\bar{\theta}_k - \theta^*) &= o_p(1) + \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} \Phi_{i-1,k_{\sigma_s}} (\tilde{\theta}_{k_{\sigma_s}} - \theta^*) \\ &\quad + \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} \sum_{j=k_{\sigma_s}}^{i-1} \Phi_{i-1,j+1} a_j \tilde{\varepsilon}_{j+1} \\ &\quad + \frac{1}{\sqrt{k}} \sum_{i=k_{\sigma_s}}^{k} \sum_{j=k_{\sigma_s}}^{i-1} \Phi_{i-1,j+1} a_j (h(\theta_j) - F(\tilde{\theta}_j - \theta^*)) \\ &\triangleq o_p(1) + I_1 + I_2 + I_3, \end{aligned} \tag{51}$$

where $o_p(\cdot)$ means

$$Y_k = o_p(Z_k) \quad \text{if and only if} \quad Y_k / Z_k \to 0 \qquad \text{in probability, as } k \to \infty.$$

By noticing that $\Phi_{k,j} = \Phi_{k-1,j} + a_k F \Phi_{k-1,j}$, we have

$$\Phi_{k,j} = I + \sum_{i=j}^{k} a_i F \Phi_{i-1,j} \quad \text{and} \quad F^{-1} \Phi_{k,j} = F^{-1} + \sum_{i=j}^{k} a_i \Phi_{i-1,j},$$

and thus

$$a_{j-1}\sum_{i=j}^{k}\Phi_{i-1,j}=\sum_{i=j}^{k}(a_{j-1}-a_i)\Phi_{i-1,j}+\sum_{i=j}^{k}a_i\Phi_{i-1,j}.$$

By the definition of $G_{k,j}$ given in Lemma A.6(iv), we have

$$a_{j-1}\sum_{i=j}^{k}\Phi_{i-1,j}=-F^{-1}+G_{k,j}, \tag{52}$$

which implies

$$I_1=\frac{1}{\sqrt{k}a_{k_{\sigma_s}-1}}(-F^{-1}+G_{k,k_{\sigma_s}})(\tilde{\theta}_{k_{\sigma_s}}-\theta^*).$$

By Lemma A.6, $G_{k,j}$ is bounded. Therefore, $I_1\to 0$ as $k\to\infty$. The above arguments also imply that there exists a constant $c_0>0$ such that

$$\left\|a_j\sum_{i=j+1}^{k}\Phi_{i-1,j+1}\right\|<c_0 \qquad \forall k,\forall j<k. \tag{53}$$

By (53), we have

$$\begin{aligned}\|I_3\|&=\frac{1}{\sqrt{k}}\left\|\sum_{j=k_{\sigma_s}}^{k}\sum_{i=j+1}^{k}\Phi_{i-1,j+1}a_j(h(\theta_j)-F(\tilde{\theta}_j-\theta^*))\right\|\\&\le\frac{c_0}{\sqrt{k}}\sum_{j=k_{\sigma_s}}^{k}\|h(\theta_j)-F(\tilde{\theta}_j-\theta^*)\|.\end{aligned}$$

It then follows from Lemma A.8 that I_3 converges to zero in probability as $k\to\infty$.

Now we consider I_2. By (34) and (52),

$$\begin{aligned}I_2&=-\frac{F^{-1}}{\sqrt{k}}\sum_{j=k_{\sigma_s}}^{k}e_{j+1}+\frac{1}{\sqrt{k}}\sum_{j=k_{\sigma_s}}^{k}G_{k,j+1}e_{j+1}\\&\quad+\frac{1}{\sqrt{k}}\sum_{j=k_{\sigma_s}}^{k}(-F^{-1}+G_{k,j+1})\nu_{j+1}\\&\triangleq J_1+J_2+J_3.\end{aligned}$$

Since $\{e_j\}$ is a martingale difference sequence,

$$E(e_i^TG_{k,i}^TG_{k,j}e_j)=E[E(e_i|\mathcal{F}_{i-1})^TG_{k,i}^TG_{k,j}e_j]=0 \qquad \forall i>j,$$

which implies that

$$E\|J_2\|^2 = \frac{1}{k}\sum_{j=k_{\sigma_s}}^{k} E(e_{j+1}^T G_{k,j+1}^T G_{k,j+1} e_{j+1}) \le \frac{1}{k}\sum_{j=k_{\sigma_s}}^{k} \|G_{k,j+1}\|^2 E\|e_{j+1}\|^2.$$

By the uniform boundedness of $\{E\|e_i\|^2, i \ge k_{\sigma_s}\}$, (42) and the uniform boundedness of $G_{k,j}$, there exists a constant c_1 such that

$$E\|J_2\|^2 \le \frac{c_1}{k}\sum_{j=k_{\sigma_s}}^{k} \|G_{k,j+1}\| \to 0 \qquad \text{as } k \to \infty. \tag{54}$$

Therefore, $J_2 \to 0$ in probability as $k \to \infty$.

Since $G_{k,j}$ is uniformly bounded with respect to both k and j, there exists a constant c_2 such that

$$E\|J_3\| \le \frac{c_2}{\sqrt{k}}\sum_{j=k_{\sigma_s}}^{k} E\|\nu_{j+1}\|.$$

Following from Lemma A.5(iii), J_3 converges to zero in probability as $k \to \infty$.

By Lemma A.5, $J_1 \to N(0, S)$ in distribution. Combining with the convergence results of I_1, I_3, J_2 and J_3, we conclude the proof of the theorem.

Proof of Theorem 2.3. Since the order of ς_k is difficult to treat, we consider the following stochastic approximation MCMC algorithm:

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + a_k(h(\theta_k) + \tilde{\varepsilon}_{k+1}), \tag{55}$$

where $\{\tilde{\theta}_k\}$ and $\{\tilde{\varepsilon}_k\}$ are as defined in (33) and (34), respectively. Following from Lemma A.5(ii), $\{\tilde{\varepsilon}_k\}$ forms a sequence of asymptotically unbiased estimator of 0.

Let $\bar{\tilde{\theta}}_n = \sum_{k=1}^n \tilde{\theta}_k/n$. To establish that $\bar{\tilde{\theta}}$ is an asymptotically efficient estimator of θ^*, we will first show (in step 1)

$$\sqrt{n}(\bar{\tilde{\theta}} - \theta^*) \to N(\mathbf{0}, \Gamma), \tag{56}$$

where $\Gamma = F^{-1}Q(F^{-1})^T$, $F = \partial h(\theta^*)/\partial\theta$ and $Q = \lim_{k\to\infty} E(e_k e_k^T)$; and then show (in step 2) that the asymptotic covariance matrix of $\sum_{k=1}^n \tilde{\varepsilon}_k/\sqrt{n}$ is equal to Q.

Step 1. By (34), we have

$$\bar{\tilde{\theta}} = \bar{\theta} + \frac{1}{n}\sum_{k=1}^{n} \tilde{\varsigma}_k. \tag{57}$$

By Lemmas A.1 and A.2, $E\|P_{\theta_{k-1}}u(\theta_{k-1},x_{k-1})\|$ is uniformly bounded for $k\ge k_{\sigma_s}$ and thus there exists a constant c such that

$$E\left\|\frac{1}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}\tilde{\varsigma}_k\right\| = E\left\|\frac{1}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}a_kP_{\theta_{k-1}}u(\theta_{k-1},x_{k-1})\right\| \le \frac{c}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}a_k.$$

By Kronecker's lemma and (A_4), we have $\frac{1}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}a_k\to 0$ in probability. Hence, $\frac{1}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}\tilde{\varsigma}_k = o_p(1)$ and

$$\frac{1}{n}\sum_{k=k_{\sigma_s}}^{n}\tilde{\varsigma}_k = o_p(n^{-1/2}). \tag{58}$$

That is

$$\bar{\bar{\theta}}_n = \bar{\theta}_n + o_p(n^{-1/2}). \tag{59}$$

Following from Theorem 2.2 and Slutsky's theorem, (56) holds.

Step 2. Now we show the asymptotic covariance matrix of $\sum_{k=1}^{n}\tilde{\varepsilon}_k/\sqrt{n}$ is equal to Q. Consider

$$\begin{aligned}
&E\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\tilde{\varepsilon}_k\right)\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\tilde{\varepsilon}_k\right)^T - \frac{1}{n}\left(\sum_{k=1}^{n}E(\tilde{\varepsilon}_k)\right)\left(\sum_{k=1}^{n}E(\tilde{\varepsilon}_k)\right)^T\\
&\quad= \frac{1}{n}\sum_{k=1}^{n}E(\tilde{\varepsilon}_k\tilde{\varepsilon}_k^T) + \frac{1}{n}\sum\sum_{i\ne j}E(\tilde{\varepsilon}_i\tilde{\varepsilon}_j^T) - \frac{1}{n}\left[\sum_{k=1}^{n}E(\tilde{\varepsilon}_k)\right]\left[\sum_{k=1}^{n}E(\tilde{\varepsilon}_k)\right]^T\\
&\quad= (I_1)+(I_2)+(I_3).
\end{aligned}$$

By (34), we have

$$\begin{aligned}
(I_1) &= \frac{1}{n}\sum_{k=1}^{n}E(e_ke_k^T) + \frac{2}{n}\sum_{k=1}^{n}E(e_k\nu_k^T) + \frac{1}{n}\sum_{k=1}^{n}E(\nu_k\nu_k^T)\\
&= (J_1)+(J_2)+(J_3).
\end{aligned}$$

By (47), $\|\nu_k\nu_k^T\|_{V^2} = O(a_k^{1+\tau})$ for $k\ge k_{\sigma_s}$, where $\tau\in(0,1)$ is defined in (A_4). Since $V^2(x)$ is square integrable, there exists a constant c such that

$$\frac{1}{n}\sum_{k=1}^{n}E\|\nu_k\nu_k^T\| \le o(1) + \frac{c}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{k=k_{\sigma_s}}^{n}a_k^{1+\tau},$$

which, by Kronecker's lemma and (A_4), implies $J_3\to 0$ as $n\to\infty$.

Following from Lemmas A.1 and A.2, $\{\|e_k\|\}_{k\ge k_{\sigma_s}}$ is uniformly bounded with respect to k. Therefore, there exists a constant c such that

$$J_2 = \frac{2}{n}\sum_{k=1}^{n}E\|e_k\nu_k^T\| \le o(1) + \frac{c}{n}\sum_{k=k_{\sigma_s}}^{n}E\|\nu_k\|.$$

Following from Lemma A.5(iii), $J_2 \to 0$ as $n \to \infty$.

By (28), $E(e_{k+1}e_{k+1}^T) = El(\theta_k, x_k)$. Since $l(\theta, x)$ is continuous in θ, it follows from Theorem 2.1 that $l(\theta_k, x)$ converges to $l(\theta^*, x)$ for any $x \in \mathcal{X}$. Furthermore, following from Lemma A.2 and Lebesgue's dominated convergence theorem, we conclude that $El(\theta_k, x_k)$ converges to $El(\theta^*, x)$, and thus

$$J_1 \to El(\theta^*, x) = \lim_{k\to\infty} E(e_k e_k^T) = Q.$$

Summarizing the convergence results of J_1, J_2 and J_3, we conclude that $(I_1) \to Q$ as $n \to \infty$.

By (34), for $i \neq j$, $i \geq k_{\sigma_s}$ and $j \geq k_{\sigma_s}$, we have

$$\begin{aligned} E(\tilde{\varepsilon}_i \tilde{\varepsilon}_j^T) &= E\{(e_i + \nu_i)(e_j + \nu_j)^T\} = E(e_i e_j^T + \nu_i \nu_j^T + e_i \nu_j^T + \nu_i e_j^T) \\ &= E(\nu_i \nu_j^T), \end{aligned} \tag{60}$$

where the last equality follows from the result that $\{e_k\}_{k \geq k_{\sigma_s}}$ is a martingale difference sequence [Lemma A.5(ii)]. By (48), there exists a constant c such that

$$E\|\nu_i \nu_j^T\| \leq c a_i^{(1+\tau)/2} a_j^{(1+\tau)/2},$$

which implies that

$$\left\| \frac{1}{n} \sum \sum_{i \neq j} E(\nu_i \nu_j^T) \right\| \leq o(1) + c \left[\frac{1}{\sqrt{n}} \sum_{i=k_{\sigma_s}}^{n} a_i^{(1+\tau)/2} \right] \left[\frac{1}{\sqrt{n}} \sum_{j=k_{\sigma_s}}^{n} a_j^{(1+\tau)/2} \right]. \tag{61}$$

By Kronecker's lemma and (A_4), $\sum_{i=k_{\sigma_s}}^{n} a_i^{(1+\tau)/2} / \sqrt{n} \to 0$ and thus

$$\frac{1}{n} \sum \sum_{i \neq j} E(\nu_i \nu_j^T) \to 0 \qquad \text{as } n \to \infty. \tag{62}$$

In summary of (60) and (62), we have

$$(I_2) = \frac{1}{n} \sum \sum_{i \neq j} E(\tilde{\varepsilon}_i \tilde{\varepsilon}_j^T) \to 0 \qquad \text{as } n \to \infty. \tag{63}$$

By (47), there exists a constant c such that

$$\frac{1}{\sqrt{n}} \left\| \sum_{k=1}^{n} E\nu_k \right\| \leq o(1) + \frac{1}{\sqrt{n}} \sum_{k=k_{\sigma_s}}^{n} E\|\nu_k\| = o(1) + \frac{c}{\sqrt{n}} \sum_{k=k_{\sigma_s}}^{n} a_k^{(1+\tau)/2}.$$

By Kronecker's lemma and (A_4), we have

$$\frac{1}{\sqrt{n}} \left\| \sum_{k=1}^{n} E\nu_k \right\| \to 0 \qquad \text{as } n \to \infty. \tag{64}$$

By Lemma A.1(i) and (ii), where it is shown that $\{e_k\}_{k\geq k_{\sigma_s}}$ is a martingale difference sequence, we have

$$(I_3) = \frac{1}{n}\left[\sum_{k=1}^{n} E(e_k+\nu_k)\right]\left[\sum_{k=1}^{n} E(e_k+\nu_k)\right]^T$$
$$= \left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n} E(\nu_k)\right]\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n} E(\nu_k)\right]^T.$$

Following from (64), we have $(I_3)\to 0$ as $n\to\infty$.

Summarizing the convergence results of (I_1), (I_2) and (I_3), the asymptotic covariance matrix of $\sum_{k=1}^{n}\tilde{\varepsilon}_k/\sqrt{n}$ is equal to Q. Combining with (56), we conclude that $\bar{\tilde{\theta}}_k$ is an asymptotically efficient estimator of θ^*.

Since $\bar{\tilde{\theta}}_k$ and $\bar{\theta}_k$ have the same asymptotic distribution $N(\mathbf{0},\Gamma)$, $\bar{\theta}_k$ is also asymptotically efficient as an estimator of θ^*. This concludes the proof of Theorem 2.3.

APPENDIX B: PROOFS OF THEOREMS 3.1 AND 3.2

The theorems can be proved using Theorems 2.1 and 2.2 by showing that SAMC satisfies the conditions (A_1) and (A_2), as (A_3) is assumed, and (A_4) and and the condition $\sup_{x\in\mathcal{X}_0} V(x)<\infty$ have been verified in the text.

Verification of (A_1). To simplify notation, in the proof we drop the subscript k, denoting x_k by x and denote $\theta_k=(\theta_k^{(1)},\ldots,\theta_k^{(m-1)})$ by $\theta=(\theta^{(1)},\ldots,\theta^{(m-1)})$. Since the invariant distribution of the MH kernel is $f_\theta(x)$, we have for any fixed θ,

$$\begin{aligned} E(I_{\{x\in E_i\}}-\pi_i) &= \int_{\mathcal{X}} (I_{\{x\in E_i\}}-\pi_i) f_\theta(x)\,dx \\ &= \frac{\int_{E_i}\psi(x)\,dx/e^{\theta^{(i)}}}{\sum_{j=1}^{m}[\int_{E_j}\psi(x)\,dx/e^{\theta^{(j)}}]} - \pi_i \\ &= \frac{S_i}{S}-\pi_i \end{aligned} \tag{65}$$

for $i=1,\ldots,m-1$, where $S_i=\int_{E_i}\psi(x)\,dx/e^{\theta^{(i)}}$ and $S=\sum_{i=1}^{m-1}S_i+\int_{E_m}\psi(x)\,dx$. Therefore,

$$h(\theta)=\int_{\mathcal{X}} H(\theta,x)f_\theta(x)\,dx=\left(\frac{S_1}{S}-\pi_1,\ldots,\frac{S_{m-1}}{S}-\pi_{m-1}\right)^T.$$

It follows from (65) that $h(\theta)$ is a continuous function of θ. Let $\Lambda(\theta)=1-\frac{1}{2}\sum_{j=1}^{m-1}(\frac{S_j}{S}-\pi_j)^2$, and define $v(\theta)=-\log(\Lambda(\theta))$ as in (19). As shown

below, $v(\theta)$ is continuously differentiable. Since $0 \le \frac{1}{2}\sum_{j=1}^{m-1}(\frac{S_j}{S}-\pi_j)^2 < \frac{1}{2}[\sum_{j=1}^{m-1}(\frac{S_j}{S})^2+\pi_j^2)] \le 1$ for all $\theta \in \Theta$, $v(\theta)$ takes values in the interval $[0,\infty)$.

Solving the system of equations formed by (65), we have the single solution

$$\theta^{(i)} = c + \log\left(\int_{E_i} \psi(\mathbf{x})\,d\mathbf{x}\right) - \log(\pi_i), \qquad i = 1,\dots,m-1,$$

where $c = -\log(\int_{E_m}\psi(\mathbf{x})\,d\mathbf{x}) + \log(\pi_m)$. It is obvious that $v(\theta^*) = 0$, and $v(\mathcal{L})$ has an empty interior, where θ^* is specified in Theorem 3.1. Therefore, (A_1)(iv) is satisfied.

Given the continuity of $v(\theta)$, for any numbers $M_1 > M_0 > 0$, $\theta^* \in \mathrm{int}(\mathcal{V}_{M_0})$, and $\mathcal{V}_{M_1}$ is a compact set, where $\mathrm{int}(A)$ denotes the interior of the set A. Therefore, (A_1)(i) and (A_1)(ii) are verified.

To verify the condition (A_1)(iii), we have the following calculations:

$$\begin{aligned}
&\frac{\partial S}{\partial\theta^{(i)}} = \frac{\partial S_i}{\partial\theta^{(i)}} = -S_i, \qquad \frac{\partial S_i}{\partial\theta^{(j)}} = \frac{\partial S_j}{\partial\theta^{(i)}} = 0,\\
&\frac{\partial(S_i/S)}{\partial\theta^{(i)}} = -\frac{S_i}{S}\left(1-\frac{S_i}{S}\right), \qquad \frac{\partial(S_i/S)}{\partial\theta^{(j)}} = \frac{\partial(S_j/S)}{\partial\theta^{(j)}} = \frac{S_iS_j}{S^2}
\end{aligned} \tag{66}$$

for $i,j = 1,\dots,m-1$ and $i \ne j$. Let $b = \sum_{j=1}^{m-1} S_j/S$, then we have

$$\begin{aligned}
\frac{\partial v(\theta)}{\partial\theta^{(j)}} &= \frac{1}{2\Lambda(\theta)}\sum_{j=1}^{m-1}\frac{\partial(S_j/S-\pi_j)^2}{\partial\theta^{(j)}}\\
&= \frac{1}{\Lambda(\theta)}\left[\sum_{j\ne i}\left(\frac{S_j}{S}-\pi_j\right)\frac{S_iS_j}{S^2} - \left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\left(1-\frac{S_i}{S}\right)\right]\\
&= \frac{1}{\Lambda(\theta)}\left[\sum_{j=1}^{m-1}\left(\frac{S_j}{S}-\pi_j\right)\frac{S_iS_j}{S^2} - \left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\right]\\
&= \frac{1}{\Lambda(\theta)}\left[b\mu_\xi\frac{S_i}{S} - \left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{S}\right]
\end{aligned}$$

for $i = 1,\dots,m-1$, where it is defined $\mu_\xi = \sum_{j=1}^{m-1}(\frac{S_j}{S}-\pi_j)\frac{S_j}{bS}$. Thus,

$$\begin{aligned}
&\langle\nabla v(\theta), h(\theta)\rangle\\
&\quad= \frac{1}{\Lambda(\theta)}\left[b^2\mu_\xi\sum_{i=1}^{m-1}\left(\frac{S_i}{S}-\pi_i\right)\frac{S_i}{bS} - b\sum_{i=1}^{m-1}\left(\frac{S_i}{S}-\pi_i\right)^2\frac{S_i}{bS}\right]\\
&\quad= -\frac{1}{\Lambda(\theta)}\left[b\sum_{i=1}^{m-1}\left(\frac{S_i}{S}-\pi_i\right)^2\frac{S_i}{bS} - b^2\mu_\xi^2\right]
\end{aligned} \tag{67}$$

$$= -\frac{1}{\Lambda(\theta)}(b\sigma_\xi^2 + b(1-b)\mu_\xi^2) \le 0,$$

where σ_ξ^2 denotes the variance of the discrete distribution defined in the following table:

State (ξ)	$\frac{S_1}{S} - \pi_1$	$\cdots$	$\frac{S_{m-1}}{S} - \pi_{m-1}$
Prob.	$\frac{S_1}{bS}$	$\cdots$	$\frac{S_{m-1}}{bS}$

If $\theta = \theta^*$, $\langle \nabla v(\theta), h(\theta)\rangle = 0$; otherwise, $\langle \nabla v(\theta), h(\theta)\rangle < 0$. Therefore, ($A_1$)(iii) is satisfied.

Verification of (A_2). To verify this condition, we first show that $h(\theta)$ has bounded second derivatives. Continuing the calculation in (66), we have

$$\frac{\partial^2(S_i/S)}{\partial(\theta^{(i)})^2} = \frac{S_i}{S}\left(1-\frac{S_i}{S}\right)\left(1-\frac{2S_i}{S}\right), \qquad \frac{\partial^2(S_i/S)}{\partial\theta^{(j)}\,\partial\theta^{(i)}} = -\frac{S_iS_j}{S^2}\left(1-\frac{2S_i}{S}\right),$$

which implies that the second derivative of $h(\theta)$ is uniformly bounded by noting the inequality $0 < \frac{S_i}{S} < 1$.

Let $F = \partial h(\theta)/\partial\theta$. By (66), we have

$$F = \begin{pmatrix} -\frac{S_1}{S}\left(1-\frac{S_1}{S}\right) & \frac{S_1S_2}{S^2} & \cdots & \frac{S_1S_{m-1}}{S^2} \\ \frac{S_2S_1}{S^2} & -\frac{S_2}{S}\left(1-\frac{S_2}{S}\right) & \cdots & \frac{S_2S_{m-1}}{S^2} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{S_{m-1}S_1}{S^2} & \cdots & \cdots & -\frac{S_{m-1}}{S}\left(1-\frac{S_{m-1}}{S}\right) \end{pmatrix}.$$

Thus, for any nonzero vector $\mathbf{z} = (z_1, \ldots, z_{m-1})^T$,

$$\begin{aligned} \mathbf{z}^T F \mathbf{z} &= -\left[\sum_{i=1}^{m-1} z_i^2 \frac{S_i}{S} - \left(\sum_{i=1}^{m-1} z_i \frac{S_i}{S}\right)^2\right] \\ &= -b\left[\sum_{i=1}^{m-1} z_i^2 \frac{S_i}{bS} - \left(\sum_{i=1}^{m-1} z_i \frac{S_i}{bS}\right)^2\right] - b(1-b)\left(\sum_{i=1}^{m-1} z_i \frac{S_i}{bS}\right)^2 \\ &= -b\,\mathrm{Var}(Z) - b(1-b)(E(Z))^2 < 0, \end{aligned} \tag{68}$$

where $E(Z)$ and $\mathrm{Var}(Z)$ denote, respectively, the mean and variance of the discrete distribution defined by the following table:

State (Z)	z_1	$\cdots$	z_{m-1}
Prob.	$\frac{S_1}{bS}$	$\cdots$	$\frac{S_{m-1}}{bS}$

This implies that the matrix F is negative definite and thus stable. Applying Taylor's expansion to $h(\theta)$ at the point θ^*, we have

$$\|h(\theta) - F(\theta - \theta^*)\| \le c\|\theta - \theta^*\|^{1+\rho},$$

for some constants $\rho \in (0,1]$ and $c > 0$, by noting that $h(\theta^*) = 0$ and that the second derivatives of $h(\theta)$ are uniformly bounded with respect to θ. Therefore, (A_2) is satisfied.

Acknowledgments. The author thanks the Editor, Associate Editor and the Referee for their constructive comments which have led to significant improvement of this paper.

REFERENCES

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070

Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** 283–312. MR2177157

Atchadé, Y. F. and Liu, J. S. (2010). The Wang–Landau algorithm in general state spaces: Applications and convergence analysis. *Statist. Sinica* **20** 209–233.

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer, New York. MR1082341

Chandra, T. K. and Goswami, A. (2006). Cesàro α-integrability and laws of large numbers. II. *J. Theoret. Probab.* **19** 789–816. MR2279604

Chen, H. F. (1993). Asymptotically efficient stochastic approximation. *Stochastics Stochastics Rep.* **45** 1–16. MR1277359

Chen, H. F. (2002). *Stochastic Approximation and Its Applications*. Kluwer Academic, Dordrecht. MR1942427

Chen, H. F., Guo, L. and Gao, A. (1988). Convergence and robustness of the Robbins–Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.* **27** 217–231. MR0931029

Chen, H. F. and Zhu, Y. M. (1986). Stochastic approximation procedures with randomly varying truncations. *Sci. Sinica Ser. A* **29** 914–926. MR0869196

Cheon, S. and Liang, F. (2007). Phylogenetic tree reconstruction using sequential stochastic approximation Monte Carlo. *BioSystems* **91** 94–107.

Cheon, S. and Liang, F. (2009). Bayesian phylogeny analysis via stochastic approximation Monte Carlo. *Mol. Phylog. Evol.* **53** 394–403.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103

Dippon, J. and Renz, J. (1997). Weighted means in stochastic approximation of minima. *SIAM J. Control Optim.* **35** 1811–1827. MR1466929

Duflo, M. (1997). *Random Iterative Models*. Springer, Berlin. MR1485774

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

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte Carlo method for incomplete data estimation problems. *Proc. Natl. Acad. Sci. USA* **95** 7270–7274. MR1630899

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 339–355. MR1841419

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

Kushner, H. J. and Yang, J. (1993). Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM J. Control Optim.* **31** 1045–1062. MR1227546

Kushner, H. J. and Yang, J. (1995). Stochastic approximation with averaging and feedback: Rapidly convergent "on-line" algorithms. *IEEE Trans. Automat. Control* **40** 24–34. MR1344315

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer, New York. MR1993642

Liang, F. (2005). Generalized Wang–Landau algorithm for Monte Carlo computation. *J. Amer. Statist. Assoc.* **100** 1311–1327. MR2236444

Liang, F. (2007a). Continuous contour Monte Carlo for marginal density estimation with an application to a spatial statistical model. *J. Comp. Graph. Statist.* **16** 608–632. MR2351082

Liang, F. (2007b). Annealing stochastic approximation Monte Carlo for neural network training. *Mach. Learn.* **68** 201–233.

Liang, F. (2009). Improving SAMC using smoothing methods: Theory and applications to Bayesian model selection problems. *Ann. Statist.* **37** 2626–2654. MR2541441

Liang, F., Liu, C. and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102** 305–320. MR2345544

Liang, F. and Zhang, J. (2009). Learning Bayesian networks for discrete data. *Comput. Statist. Data Anal.* **53** 865–876.

Marinari, E. and 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. and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1091.

Moyeed, R. A. and Baddeley, A. J. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18** 39–50. MR1115181

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.* **39** 49–72. MR1780908

Polyak, B. T. (1990). New stochastic approximation type procedures. *Avtomat. i Telemekh.* 7 98–107 (in Russian). MR1071220

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407. MR0042668

Ruppert, D. (1988). Efficient estimators from a slowly convergent Robbins–Monro procedure. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell Univ.

Tadić, V. (1997). Convergence of stochastic approximation under general noise and stability conditions. In: *Proceedings of the 36th IEEE Conference on Decision and Control* **3** 2281–2286. IEEE Systems Society, San Diego, CA.

Tang, Q. Y., L'Ecuyer, P. and Chen, H. F. (1999). Asymptotic efficiency of perturbation-analysis-based stochastic approximation with averaging. *SIAM J. Control Optim.* **37** 1822–1847. MR1720140

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

Wang, I.-J., Chong, E. K. P. and Kulkarni, S. R. (1997). Weighted averaging and stochastic approximation. *Math. Control Signals Systems* **10** 41–60. MR1462279

Younes, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Related Fields* **82** 625–645. MR1002904

Younes, L. (1999). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics Stochastics Rep.* **65** 177–228. MR1687636

Yu, K. and Liang, F. (2009). Efficient P-value evaluation for resampling-based tests. Technical report, Dept. Statistics, Texas A&M Univ., College Station, TX.

Department of Statistics
Texas A&M University
College Station, Texas 77843-3143
USA
E-mail: fliang@stat.tamu.edu