Adv. Appl. Prob. **46,** 1059–1083 (2014)
Printed in Northern Ireland

WEAK CONVERGENCE RATES OF POPULATION VERSUS SINGLE-CHAIN STOCHASTIC APPROXIMATION MCMC ALGORITHMS

QIFAN SONG,* *Texas A&M University*
MINGQI WU,** *Shell Global Solutions (US) Inc.*
FAMING LIANG,*** *Texas A&M University*

Abstract

In this paper we establish the theory of weak convergence (toward a normal distribution) for both single-chain and population stochastic approximation Markov chain Monte Carlo (MCMC) algorithms (SAMCMC algorithms). Based on the theory, we give an explicit ratio of convergence rates for the population SAMCMC algorithm and the single-chain SAMCMC algorithm. Our results provide a theoretic guarantee that the population SAMCMC algorithms are asymptotically more efficient than the single-chain SAMCMC algorithms when the gain factor sequence decreases slower than $O(1/t)$, where t indexes the number of iterations. This is of interest for practical applications.

Keywords: Asymptotic normality; Markov chain Monte Carlo; stochastic approximation; Metropolis–Hastings algorithm

2010 Mathematics Subject Classification: Primary 60J22
Secondary 65C05

1. Introduction

Robbins and Monro (1951) introduced the stochastic approximation algorithm for solving the integration equation

$$h(\theta) = \int H(\theta, x) f_\theta(x)\,\mathrm{d}x = 0,$$

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 dependent on θ. The stochastic approximation algorithm is a recursive algorithm defined as follows.

(a) Draw sample $x_{t+1} \sim f_{\theta_t}(x)$, where t indexes the iteration.

(b) Set $\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, x_{t+1})$, where γ_{t+1} is called the gain factor.

After six decades of continual development, this algorithm has become an important research area in systems control, and has also served as a prototype for the development of recursive algorithms for online estimation and control of stochastic systems. Recently, the stochastic approximation algorithm has been used with Markov chain Monte Carlo (MCMC), which replaces step (a) by an MCMC sampling step.

Received 19 January 2012; revision received 30 September 2013.

* Current address: Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.
** Postal address: Shell Technology Center Houston, 3333 Highway 6 South, Houston, TX 77082, USA.
*** Current address: Department of Biostatistics, University of Florida, Gainesville, FL 321611, USA. Email address: faliang@ufl.edu

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

(a′) Draw a sample x_{t+1} with a Markov transition kernel $P_{\theta_t}(x_t, \cdot)$, which starts with x_t and admits $f_{\theta_t}(x)$ as the invariant distribution.

In statistics, the stochastic approximation MCMC (SAMCMC) algorithm, which is also known as stochastic approximation with Markov state-dependent noise, has been successfully applied to many problems of general interest, such as maximum likelihood estimation for incomplete data problems (see Younes (1989) and Gu and Kong (1998)), marginal density estimation (see Liang (2007)), and adaptive MCMC (see Haario *et al.* (2001), Andrieu and Moulines (2006), Roberts and Rosenthal (2009), and Atchadé and Fort (2010)).

It is clear that efficiency of the SAMCMC algorithm depends crucially on the mixing rate of the Markov transition kernel P_{θ_t}. Motivated by the success of population MCMC algorithms, see, e.g. Gilks *et al.* (1994), Liu *et al.* (2000), and Liang and Wong (2000), (2001), which can generally converge faster than single-chain MCMC algorithms, we exploit in this paper the performance of a population SAMCMC algorithm, both theoretically and numerically. Our results show that the population SAMCMC algorithm can be asymptotically more efficient than the single-chain SAMCMC algorithm.

Our contribution in this paper is twofold. First, we establish the asymptotic normality for the SAMCMC estimator, which holds for both the population and single-chain SAMCMC algorithms. We note that a similar result has been established in Benveniste *et al.* (1990 Theorem 13, p. 332), but under different conditions for the Markov transition kernel. Our conditions can be easily verified, whereas the conditions given in Benveniste *et al.* (1990) are less verifiable. More importantly, our result is more interpretable than that of Benveniste *et al.* (1990), and this motivates our design of the population SAMCMC algorithm. Second, we propose a general population SAMCMC algorithm, and compare its convergence rate with that of the single-chain SAMCMC algorithm. Our result provides a theoretical guarantee that the population SAMCMC algorithm is asymptotically more efficient than the single-chain SAMCMC algorithm when the gain factor sequence $\{\gamma_t\}$ decreases slower than $O(1/t)$. The theoretical result has been confirmed with a numerical example.

The remainder of this paper is organized as follows. In Section 2 we describe the population SAMCMC algorithm and compare its convergence rate with that of the single-chain SAMCMC algorithm. In Section 3 we study the population stochastic approximation Monte Carlo (Pop-SAMC) algorithm, which is proposed based on the SAMC algorithm by Liang *et al.* (2007) and is a special case of the population SAMCMC algorithm. In Section 4 we present a numerical example in which we compare the performances of SAMC and Pop-SAMC algorithms when sampling from a multimodal distribution. In Section 5 we conclude the paper with a brief discussion.

2. Convergence rates of population versus single-chain SAMCMC algorithms

2.1. Population SAMCMC algorithm

The population SAMCMC algorithm works with a population of samples at each iteration. Let $\boldsymbol{x}_t = (x_t^{(1)}, \ldots, x_t^{(\kappa)})$ denote the population of samples at iteration t, let $\mathcal{X}^\kappa = \mathcal{X} \times \cdots \times \mathcal{X}$ denote the sample space of $\boldsymbol{x}_t$, and let $\mathcal{X}_0^\kappa$ denote a subset of $\mathcal{X}^\kappa$ from which $\boldsymbol{x}_0$ is drawn. The population SAMCMC algorithm starts with a point $(\theta_0, \boldsymbol{x}_0)$ drawn from $\Theta \times \mathcal{X}_0^\kappa$ and then iterates between the following steps.

(a) Draw samples $x_{t+1}^{(1)}, \ldots, x_{t+1}^{(\kappa)}$ with a Markov transition kernel $\boldsymbol{P}_{\theta_t}(\boldsymbol{x}_t, \cdot)$, which starts with $\boldsymbol{x}_t$ and admits $f_{\theta_t}(\boldsymbol{x}) = f_{\theta_t}(x^{(1)}) \cdots f_{\theta_t}(x^{(\kappa)})$ as the invariant distribution.

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

(b) Set $\theta_{t+1} = \theta_t + \gamma_{t+1}\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1})$, where $\boldsymbol{x}_{t+1} = (x_{t+1}^{(1)}, \dots, x_{t+1}^{(\kappa)})$ and

$$\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}) = \frac{1}{\kappa}\sum_{i=1}^{\kappa} H(\theta_t, x_{t+1}^{(i)}).$$

It is easy to see that the population SAMCMC algorithm is actually an SAMCMC algorithm with the mean-field function specified by

$$\begin{aligned} h(\theta) &= \int \boldsymbol{H}(\theta, \boldsymbol{x})\boldsymbol{f}_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int\cdots\int\bigg[\frac{1}{\kappa}\sum_{i=1}^{\kappa} H(\theta, x^{(i)})\bigg] f_\theta(x^{(1)})\cdots f_\theta(x^{(\kappa)})\,\mathrm{d}x^{(1)}\cdots\mathrm{d}x^{(\kappa)} \\ &= 0, \end{aligned}$$

where $\boldsymbol{f}_\theta(\boldsymbol{x}) = f_\theta(x^{(1)})\cdots f_\theta(x^{(\kappa)})$ denotes the joint probability density function of $\boldsymbol{x} = (x^{(1)}, \dots, x^{(\kappa)})$.

If $\kappa = 1$, the algorithm is reduced to the single-chain SAMCMC algorithm. Compared to the single-chain SAMCMC algorithm, the population SAMCMC algorithm has two advantages. First, it provides a more accurate estimate of $h(\theta)$ at each iteration, and this eventually leads to a faster convergence of the algorithm. Note that $H(\theta_t, \boldsymbol{x}_{t+1})$ provides an estimate of $h(\theta_t)$ at iteration t. Second, since a population of Markov chains is run in parallel, the population SAMCMC algorithm is able to incorporate some advanced multiple chain operators, such as the crossover operator (see Liang and Wong (2000), (2001)), the snooker operator (see Gilks *et al.* (1994)), and the gradient operator (see Liu *et al.* (2000)), into simulations. With these operators, the distributed information across the population can then be used in guiding further simulations, and this can accelerate the convergence of the algorithm. However, for illustration purposes, we primarily consider in this paper the single-chain operator, for which we have

$$\boldsymbol{P}_{\theta_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = \prod_{i=1}^{\kappa} P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)}).$$

Extension of our convergence result to the general population SAMCMC algorithm which consists of multiple-chain operators is straightforward and this will be discussed in Section 3.4.

2.2. Main theoretical results

For mathematical simplicity, we assume in this paper that Θ is compact, i.e. the sequence $\{\theta_t\}$ can remain in a compact set. Extension of our results to the case in which $\Theta = \mathbb{R}^{d_\theta}$ trivially follows from the technique of varying truncations studied in Chen (2002) and Andrieu *et al.* (2005), which ensures, almost surely, that the sequence $\{\theta_t\}$ can be included in a compact set. Since Theorems 1 and 2 below are applicable to both the population and single-chain SAMCMC algorithms, we will let X_t denote the sample(s) drawn at iteration t and let $\mathbb{X}$ denote the sample space of X_t. For the population SAMCMC algorithm, we have $\mathbb{X} = \mathcal{X}^\kappa$ and $X_t = \boldsymbol{x}_t$. For the single-chain SAMC algorithm, we have $\mathbb{X} = \mathcal{X}$ and $X_t = x_t$. For any measurable function $f\colon \mathbb{X} \to \mathbb{R}^d$, $\boldsymbol{P}_\theta f(X) = \int_{\mathbb{X}} \boldsymbol{P}_\theta(X, y) f(y)\,\mathrm{d}y$.

Condition 1. (Lyapunov condition on $h(\theta)$.) *Let* $\mathcal{L} = \{\theta \in \Theta\colon h(\theta) = 0\}$.

(A1) *The function* $h\colon \Theta \to \mathbb{R}^d$ *is continuous, and there exists a continuously differentiable function* $v\colon \Theta \to [0, \infty)$ *such that* $v_h(\theta) = \nabla^\top v(\theta)h(\theta) < 0$ *for all* $\theta \in \mathcal{L}^c$, $\sup_{\theta\in\mathcal{K}} v_h(\theta) < 0$ *for any compact set* $\mathcal{K} \subset \mathcal{L}^c$, *and* $\nabla v(\theta)$ *is Lipschitz continuous.*

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

This condition assumes the existence of a global Lyapunov function v for the mean field h. If h is a gradient field, i.e. $h = -\nabla J$ 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.

Condition 2. (Stability condition on $h(\theta)$.)

(A2) *The mean-field function $h(\theta)$ is measurable and locally bounded on Θ. There exists a stable matrix F (i.e. all eigenvalues of F have negative real parts), $\rho > 0$, and a constant c such that, for any $\theta_* \in \mathcal{L}$ (defined in (A1)),*

$$\|h(\theta) - F(\theta - \theta_*)\| \le c\|\theta - \theta_*\|^2 \quad \textit{for all } \theta \in \{\theta \colon \|\theta - \theta_*\| \le \rho\}.$$

This condition constrains the behavior of the mean-field function around the solution points. If $h(\theta)$ is differentiable, the matrix F can be chosen to be the partial derivative of $h(\theta)$, i.e. $\partial h(\theta)/\partial\theta$. Otherwise, a certain approximation may be needed.

Condition 3. (Drift condition on the transition kernel $\boldsymbol{P}_\theta$.) *For a function $g \colon \mathbb{X} \to \mathbb{R}^d$, define the L_∞-norm $\|g\| = \sup_{x\in\mathbb{X}} \|g(x)\|$.*

(A3) *For any given $\theta \in \Theta$, the transition kernel $\boldsymbol{P}_\theta$ is irreducible and aperiodic. In addition, the following statements hold.*

(i) *(Doeblin condition) There exists a constant $\delta > 0$, an integer $l > 0$, and a probability measure ν such that*

$$\inf_{\theta\in\Theta} \boldsymbol{P}_\theta^l(X, A) \ge \delta\nu(A) \quad \textit{for all } X \in \mathbb{X} \textit{ and all } A \in \mathcal{B}_{\mathbb{X}},$$

where $\mathcal{B}_{\mathbb{X}}$ denotes the Borel set of $\mathbb{X}$, i.e. the whole set $\mathbb{X}$ is a small set for each $\boldsymbol{P}_\theta$.

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

$$\sup_{\theta\in\Theta} \|\boldsymbol{H}(\theta, \cdot)\| \le c, \tag{1}$$

$$\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-1} \|\boldsymbol{H}(\theta, \cdot) - \boldsymbol{H}(\theta', \cdot)\| \le c.$$

(iii) *There exists a constant $c > 0$ such that, for all g with $\|g\| < \infty$,*

$$\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-1} \|\boldsymbol{P}_\theta g - \boldsymbol{P}_{\theta'} g\| \le c\|g\|.$$

The Doeblin condition of condition (A3)(i) is equivalent to assuming that the resulting Markov chain has a unique stationary distribution and is uniformly ergodic (see Nummelin (1984, Theorem 6.15)). This condition is slightly stronger than the drift condition assumed in Andrieu *et al.* (2005) and Andrieu and Moulines (2006), which implies the V-uniform ergodicity for $\boldsymbol{P}_\theta$. Condition (A3)(ii) gives conditions on $\boldsymbol{H}(\theta, X)$ which lead directly to the boundedness of the observation noise. It is also worth noting that the property that $\boldsymbol{P}_\theta$ satisfies condition (A3)(i) and (A3)(iii) can be inherited from the corresponding property of the single-chain case. If the conditions hold for the single-chain kernel P_θ then the conditions must hold for $\boldsymbol{P}_\theta$. We refer the reader to the arguments used in the proof of Theorem 4 given in the supplementary paper Song *et al.* (2013).

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

Condition 4. (Conditions on step sizes.)

(A4) (i) *The sequence $\{\gamma_t\}$, which is defined to be $\gamma(t)$ as a function of t and is exchangeable with $\gamma(t)$ in this paper, is positive and nonincreasing, and satisfies the conditions*

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \qquad \frac{\gamma_{t+1} - \gamma_t}{\gamma_t} = O(\gamma_{t+1}^{\tau}), \qquad \sum_{t=1}^{\infty} \frac{\gamma_t^{(1+\tau')/2}}{\sqrt{t}} < \infty, \tag{2}$$

for some $\tau \in [1, 2)$ and $\tau' \in (0, 1)$.

(ii) *The function $\zeta(t) = \gamma(t)^{-1}$ is differentiable such that its derivative varies regularly with exponent $\tilde{\beta} - 1 \geq -1$ (i.e. for any $z > 0$, $\zeta'(zt)/\zeta'(t) \to z^{\tilde{\beta}-1}$ as $t \to \infty$), and either of the following two cases holds:*

(ii.1) *$\gamma(t)$ varies regularly with exponent $(-\beta)$, $\frac{1}{2} < \beta < 1$;*

(ii.2) *for $t \geq 1$, $\gamma(t) = t_0/t$ with $-2\lambda_F t_0 > \max\{1, \tilde{\beta}\}$, where λ_F denotes the largest real part of the eigenvalues of the matrix F (defined in condition (A2)) with $\lambda_F < 0$.*

As shown in Chen (2002, p. 134), the condition $\sum_{t=1}^{\infty} \gamma_t^{(1+\tau')/2}/\sqrt{t} < \infty$, together with the monotonicity of γ_t, implies that $\gamma_t^{(1+\tau')/2} = o(t^{-1/2})$, and, thus,

$$\sum_{t=1}^{\infty} \gamma_t^{1+\tau'} = \sum_t \sqrt{t}\gamma_t^{(1+\tau')/2} \frac{\gamma_t^{(1+\tau')/2}}{\sqrt{t}} < \infty, \tag{3}$$

which is often assumed to hold when studying the convergence of stochastic approximations. While condition (2) is often assumed to hold when studying the weak convergence of the trajectory averaging estimator of θ_t (see, e.g. Chen (2002)). (A4)(ii) can be applied to the usual gains $\gamma_t = t_0/t^{\beta}$, $\frac{1}{2} < \beta \leq 1$. Following Pelletier (1998), we deduce that

$$\left(\frac{\gamma_t}{\gamma_{t+1}}\right)^{1/2} = 1 + \frac{\beta}{2t} + o\left(\frac{1}{t}\right). \tag{4}$$

In terms of γ_t, (4) can be rewritten as

$$\left(\frac{\gamma_t}{\gamma_{t+1}}\right)^{1/2} = 1 + \zeta\gamma_t + o(\gamma_t), \tag{5}$$

where $\zeta = 0$ for the (ii.1) case and $\zeta = 1/2t_0$ for $\beta = 1$ for the (ii.2) case. Clearly, the matrix is $F + \zeta I$, which is still stable.

Theorem 1 below concerns the convergence of the general SAMCMC algorithm.

Theorem 1. *Assume that Θ is compact, and that the conditions (A1), (A3), and (A4)(i) hold. Let the simulation start with a point $(\theta_0, X_0) \in \Theta \times \mathbb{X}_0$, where $\mathbb{X}_0 \subset \mathbb{X}$ such that $\sup_{X \in \mathbb{X}_0} V(X) < \infty$. Then, as $t \to \infty$,*

$$d(\theta_t, \mathcal{L}) \to 0 \quad \textit{almost surely},$$

where $\mathcal{L} = \{\theta \in \Theta \colon h(\theta) = 0\}$ and $d(u, z) = \inf_{z \in \mathbf{z}} \|u - z\|$.

Proof. See Appendix A.

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

To study the convergence rate of θ_t, we rewrite the iterative equation of SAMCMC as

$$\theta_{t+1} = \theta_t + \gamma_t[h(\theta_t) + \xi_{t+1}],$$

where $h(\theta_t) = \int_{\mathbb{X}} \boldsymbol{H}(\theta_t, X) f_{\theta_t}(X)\,\mathrm{d}X$ and $\xi_{t+1} = \boldsymbol{H}(\theta_t, X_{t+1}) - h(\theta_t)$ is called the observation noise. Lemma 1 below concerns the decomposition of the observation noise, whose parts (i) and (iv) are partial restatements of Lemma A.5 of Liang (2010).

Lemma 1. *Assume that the conditions of Theorem 1 hold. Then there exist $\mathbb{R}^{d_\theta}$-valued random processes $\{e_t\}$, $\{\nu_t\}$, and $\{\varsigma_t\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that*

(i) $\xi_t = e_t + \nu_t + \varsigma_t$*;*

(ii) *for any constant $\rho > 0$ (defined in condition (A2)),*

$$\mathbb{E}(e_{t+1} \mid \mathcal{F}_t)\,\mathbf{1}_{\{\|\theta_t - \theta_*\| \le \rho\}} = 0,$$
$$\sup_{t \ge 0} \mathbb{E}(\|e_{t+1}\|^\alpha \mid \mathcal{F}_t)\,\mathbf{1}_{\{\|\theta_t - \theta_*\| \le \rho\}} < \infty,$$

where $\mathcal{F}_t$ is a family of σ-algebras satisfying $\sigma\{\theta_0, X_0; \theta_1, X_1; \ldots; \theta_t, X_t\} = \mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t \ge 0$ and $\alpha \ge 2$ is a constant;

(iii) *almost surely on $\Lambda(\theta_*) = \{\theta_t \to \theta_*\}$, as $n \to \infty$,*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}(e_{t+1}e'_{t+1} \mid \mathcal{F}_t) \to \Gamma \quad \textit{almost surely}, \tag{6}$$

where Γ is a positive definite matrix;

(iv) $\mathbb{E}(\|\nu_t\|^2/\gamma_t)\,\mathbf{1}_{\{\|\theta_t - \theta_*\| \le \rho\}} \to 0$ *as $t \to \infty$;*

(v) $\mathbb{E}\|\gamma_t \varsigma_t\| \to 0$ *as $t \to \infty$.*

Proof. See Appendix B.

Lemma 1 plays a key role in the proof of the following theorem, which concerns the asymptotic normality of θ_t.

Theorem 2. *Assume that Θ is compact and that conditions (A1)–(A4) hold. Conditioned on $\Lambda(\theta_*) = \{\theta_t \to \theta_*\}$,*

$$\frac{\theta_t - \theta_*}{\sqrt{\gamma_t}} \Rightarrow \mathbb{N}(0, \Sigma),$$

where '$\Rightarrow$' denotes the weak convergence, $\mathbb{N}$ is the Gaussian distribution, and

$$\Sigma = \int_0^\infty \mathrm{e}^{(F' + \zeta I)t}\,\Gamma\,\mathrm{e}^{(F + \zeta I)t}\,\mathrm{d}t,$$

with F defined as in (A2), ζ defined as in (5), and Γ defined as in Lemma 1.

Proof. See Appendix B.

Remarks. 1. Theorem 2 has been established in Benveniste *et al.* (1990, Theorem 13, p. 332) but under different assumptions for the Markov transition kernel $\boldsymbol{P}_\theta$. Similar to Andrieu *et al.* (2005), we assume a slightly stronger condition, (A3), that $\boldsymbol{P}_\theta$ satisfies a minorization

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

condition on $\mathbb{X}$. This condition not only ensures the existence of a stationary distribution of $\boldsymbol{P}_\theta$, uniform ergodicity, and the existence and regularity of the solution to the Poisson equation (see, e.g. Meyn and Tweedie (2009)), but also implies boundedness of the moment of the sample X_t. In Benveniste *et al.* (1990), besides some conditions on $\boldsymbol{P}_\theta$, such as the existence and regularity of the solution to the Poisson equation, the authors imposed a moment condition on X_t (see Benveniste *et al.* (1990, Condition A_5, p. 220)). The moment condition is usually very difficult to verify without assumptions on the ergodicity of the Markov chain. Concerning the convergence of the adaptive Markov chain $\{X_t\}$, Andrieu and Moulines (2006) presented a central limit theorem for the average of $\phi(X_t)$, where $\phi(\cdot)$ is a V^r-Lipschitz function for some $r \in [0, \frac{1}{2})$ and $V(\cdot)$ is the drift function. Unlike Andrieu and Moulines (2006), in this paper we present the asymptotic normality for the adaptive stochastic approximation estimator θ_t itself.

2. As shown in Benveniste *et al.* (1990), $(\theta_t - \theta_*)/\sqrt{\gamma_t}$ converges weakly towards the distribution of a stationary Gaussian diffusion with generator

$$\mathrm{d}X_t = (F + \zeta I)X_t + \Gamma^{1/2}\,\mathrm{d}B_t,$$

where B_t is a standard Brownian motion. Therefore, the asymptotic covariance matrix Σ corresponds to the solution of Lyapunov's equation

$$(F + \zeta I)\Sigma + \Sigma(F' + \zeta I) = -\Gamma.$$

An explicit form of the solution can be found in Ziedan (1972), which is omitted here due to its complex formulation.

3. From (20) in the proof of Lemma below, it is not difficult to derive

$$\Gamma = \sum_{k=-\infty}^{\infty} \int H(\theta_*, x)[P_{\theta_*}^k H(\theta_*, x)]^\top \,\mathrm{d}\pi_{\theta_*}(\mathrm{d}x), \tag{7}$$

where π_{θ_*} denotes the invariant distribution of the transition kernel P_{θ_*}. This is the same Γ expression as given in Benveniste *et al.* (1990, Equation (4.4.6), p. 321). Compared to (7), our Γ expression, given in (6), is more interpretable, which corresponds to the asymptotic covariance matrix of e_t. Given the gain factor sequence $\{\gamma_k\}$, the efficiency of an SAMCMC algorithm is determined by Γ. Based on this observation, we show in Theorem 3 below that when $\{\gamma_t\}$ decreases slower than $O(1/t)$, the population SAMCMC algorithm has a smaller asymptotic covariance matrix than the single-chain SAMCMC algorithm and is thus asymptotically more efficient.

4. The condition 'conditioned on $\Lambda(\theta_*)$' accommodates the case in which there exist multiple solutions of the equation $h(\theta) = 0$.

In Theorem 3 below we compare the efficiencies of the population SAMCMC and the single-chain SAMCMC algorithms.

Theorem 3. *Suppose that both the population and single-chain SAMCMC algorithms satisfy the conditions given in Theorem 2. Let θ_t^{p} and θ_t^{s} denote the estimates produced at iteration t by the population and single-chain SAMCMC algorithms, respectively. Given the same gain factor sequence $\{\gamma_t\}$, then $(\theta_t^{\mathrm{p}} - \theta_*)/\sqrt{\gamma_t}$ and $(\theta_{\kappa t}^{\mathrm{s}} - \theta_*)/\sqrt{\kappa\gamma_{\kappa t}}$ have the same asymptotic distribution with convergence rate ratio*

$$\frac{\gamma_t}{\kappa\gamma_{\kappa t}} = \kappa^{\beta-1}, \tag{8}$$

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

where κ denotes the population size, and β is defined in (A4). (Note that $\frac{1}{2} < \beta < 1$ for the (A4)(ii.1) case and $\beta = 1$ for the (A4)(ii.2) case.)

Proof. See Appendix A.

Remarks. 1. When $\beta = 1$ (e.g. $\gamma_t = t_0/t$), the single-chain SAMCMC estimator is as efficient as the population SAMCMC estimator, but this is only true asymptotically. For practical applications (as illustrated in Figure 1(a) and Figure 2 in Section 4), the population SAMCMC estimator can still be more efficient than the single-chain SAMCMC estimator due to the population effect. At each iteration, the population SAMCMC estimator provides a more accurate estimate of $h(\theta_t)$ than the single-chain SAMCMC estimator, and this substantially improves the convergence of the algorithm, especially at the early stage of the simulation.

2. When $\beta < 1$, the population SAMCMC estimator is asymptotically more efficient than the single-chain SAMCMC estimator. See Figure 1(b) in Section 4.

3. The choice of the population size should be balanced with the choice of N, the number of iterations, as the convergence of the algorithm only occurs as $\gamma_t \to 0$. In our experience, $5 \sim 50$ may be a good range for the population size.

3. Population SAMC algorithm

In this section we first give a brief review of the SAMC algorithm, and then describe the population SAMC algorithm and its theoretical properties, including convergence and asymptotic normality.

3.1. The SAMC algorithm

Suppose that we are interested in sampling from a distribution:

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

Here $\mathcal{X}$ is the sample space and c is an unknown constant. Furthermore, we assume that the distribution $f(x)$ is multimodal, which may contain a multitude of modes separated by high energy barriers. It is known that the conventional MCMC algorithms, such as the Metropolis–Hastings algorithm (see Metropolis *et al.* (1953) and Hastings (1970)) and the Gibbs sampler (see Geman and Geman (1984)), are prone to becoming trapped at local modes in simulations from this kind of distribution.

Designing MCMC algorithms that are immune to the local-trap problem has been a long-standing topic in Monte Carlo research. A few significant algorithms have been proposed in this direction, including parallel tempering (see Geyer (1991)), simulated tempering (see Marinari and Parisi (1992)), dynamic weighting (see Wong and Liang (1997)), the Wang–Landau algorithm (see Wang and Landau (2001)), the SAMC algorithm (see Liang *et al.* (2007)), among others. The SAMC algorithm can be described as follows.

Let $E_1, \ldots, E_m$ denote a partition of the sample space $\mathcal{X}$. For example, the sample space can be partitioned according to the energy function of $f(x)$, i.e. $U(x) = -\log\psi(x)$, into the subregions $E_1 = \{x \colon U(x) \le u_1\}$, $E_2 = \{x \colon u_1 < U(x) \le u_2\}, \ldots, E_{m-1} = \{x \colon u_{m-2} < U(x) \le u_{m-1}\}$, and $E_m = \{x \colon U(x) \ge u_m\}$, where $u_1 < u_2 < \cdots < u_{m-1}$ are user-specified numbers. If $\int_{E_i} \psi(x)\,\mathrm{d}x = 0$ then E_i is called an empty subregion. We refer the reader to Liang *et al.* (2007) for more discussions on sample space partitioning. For the time being, we

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

assume that all the subregions are nonempty, that is, $\int_{E_i} \psi(x)\,\mathrm{d}x > 0$ for all $i = 1, \dots, m$. Given the partition, the SAMC algorithms seeks to draw samples from the distribution

$$f_w(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x)}{w_i} \mathbf{1}_{\{x \in E_i\}},$$

where $w_i = \int_{E_i} \psi(x)\,\mathrm{d}x$, and the π_i define the desired sampling frequency for each of the subregions, satisfying the constraints $\pi_i > 0$ for all i and $\sum_{i=1}^{m} \pi_i = 1$. If $w_1, \dots, w_m$ are known, sampling from $f_w(x)$ will lead to 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. Thus, the local-trap problem can be essentially overcome, provided that the sample space is partitioned appropriately.

Since $w_1, \dots, w_m$ are generally unknown, the SAMC algorithm employs the stochastic approximation algorithm to estimate their values. This leads to the following iterative procedure.

(a) (Sampling) Simulate a sample x_{t+1} by running, for one step, the Metropolis–Hastings algorithm which starts with x_t and admits the stationary distribution

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{\mathrm{e}^{\theta_{t,i}}} \mathbf{1}_{\{x \in E_i\}}, \tag{9}$$

where $\theta_t = (\theta_{t,1}, \dots, \theta_{t,m})$ and $\theta_{t,i}$ denotes the working (online) estimator of $\log(w_i/\pi_i)$ at iteration t.

(b) (Weight updating) Set

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, x_{t+1}),$$

where $H(\theta_t, x_{t+1}) = z_{t+1} - \boldsymbol{\pi}$, $z_{t+1} = (\mathbf{1}_{\{x_{t+1} \in E_1\}}, \dots, \mathbf{1}_{\{x_{t+1} \in E_m\}})$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)$, and $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

A remarkable feature of the SAMC algorithm is that it possesses the self-adjusting mechanism, which operates based on past samples. This mechanism penalizes the over-visited subregions and rewards the under-visited subregions, thus enabling the system to escape from local traps very quickly. Mathematically, if a subregion E_i is visited at iteration t, $\theta_{t+1,i}$ will be updated to a larger value, $\theta_{t+1,i} \leftarrow \theta_{t,i} + \gamma_{t+1}(1 - \pi_i)$, such that this subregion has a decreased probability of being visited at the next iteration. On the other hand, for the regions E_j $(j \neq i)$ not visited at iteration t, $\theta_{t+1,j}$ will decrease to a smaller value, $\theta_{t+1,j} \leftarrow \theta_{t,j} - \gamma_{t+1}\pi_j$, such that the chance of visiting these regions will increase at the next iteration. The SAMC algorithm has been successfully applied to many different problems for which the energy landscape is rugged, such as phylogeny inference (see Cheon and Liang (2009)) and Bayesian network learning (see Liang and Zhang (2009)).

3.2. The population SAMC algorithm

The population SAMC (Pop-SAMC) algorithm works as follows. Let $\boldsymbol{x}_t = (x_t^{(1)}, \dots, x_t^{(\kappa)})$ denote the population of samples simulated at iteration t. One iteration of the algorithm consists of the following two steps.

(a) (Population sampling) For $i = 1, \dots, \kappa$, simulate a sample $x_{t+1}^{(i)}$ by running, for one step, the Metropolis–Hasting algorithm which starts with $x_t^{(i)}$ and admits (9) as the invariant distribution. Denote the population of samples by $\boldsymbol{x}_{t+1} = (x_{t+1}^{(1)}, \dots, x_{t+1}^{(\kappa)})$.

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

(b) (Weight updating) Set

$$\theta_{t+1} = \theta_t + \gamma_{t+1}\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}), \tag{10}$$

where $\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}) = \sum_{i=1}^{\kappa} H(\theta_t, x_{t+1}^{(i)})/\kappa$ and $H(\theta_t, x_{t+1}^{(i)})$ is as specified in the SAMC algorithm.

As a special case of the population SAMCMC algorithms, the Pop-SAMC algorithm has a few advantages over the SAMC algorithm. First, since $\boldsymbol{H}(\theta, \boldsymbol{x})$ provides a more accurate estimate of $h(\theta)$ than $H(\theta, x)$ at each iteration, the Pop-SAMC algorithm can converge asymptotically faster than the SAMC algorithm. This is the so-called population effect and will be illustrated in Section 4 through a numerical example. Second, population-based proposals, such as the crossover operator, snooker operator, and gradient operator, can be included in the algorithm to improve efficiency of the sampling step and, thus, the convergence of the algorithm. The only requirement for these operators is that they admit the joint density $f_{\theta_t}(x^{(1)}) \cdots f_{\theta_t}(x^{(\kappa)})$ as the invariant distribution. The weak convergence of the resulting algorithm is discussed at the end of this paper. Third, a smoothing operator can be further introduced to $\boldsymbol{H}(\theta, \boldsymbol{x})$ to improve its accuracy as an estimator of $h(\theta)$. Liang (2009) showed through numerical examples that the smoothing operator can improve the convergence of the SAMC algorithm, if multiple Metropolis–Hastings (MH) updates were allowed at each iteration.

3.3. Theoretical results

Regarding the convergence of θ_t, we note that, for empty subregions, the corresponding components of θ_t will trivially converge to $-\infty$ when the number of iterations goes to ∞. Therefore, without loss of generality, we show in the supplementary paper Song *et al.* (2013) only the convergence of the algorithm for the case in which all subregions are nonempty, that is, $\int_{E_i} \psi(x)\, \mathrm{d}x > 0$ for all $i = 1, \ldots, m$. Extending the proof to the general case is trivial, since replacing (10) by (11) (given below) will not change the process of Pop-SAMC simulation:

$$\theta'_{t+1} = \theta_t + \gamma_{t+1}(\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}) - \boldsymbol{\nu}). \tag{11}$$

Here $\boldsymbol{\nu} = (\nu, \ldots, \nu)$ is an m-vector of ν, where $\nu = \sum_{j \in \{i\colon E_i = \varnothing\}} \pi_j/(m - m_0)$ and m_0 is the number of empty subregions.

In our proof, we assume that Θ is a compact set. As mentioned above for the general SAMCMC algorithm, this assumption is made only for the reason of mathematical simplicity. Extension of our results to the case in which $\Theta = \mathbb{R}^m$ follows trivially from the technique of varying truncations (see Chen (2002), Andrieu *et al.* (2005), and Liang (2010)). Interested readers are referred to Liang (2010) for the details, where the convergence of the SAMC algorithm is studied with $\Theta = \mathbb{R}^m$. In the simulations of this paper, we set $\Theta = [-10^{100}, 10^{100}]^m$, which is practically equivalent to setting $\Theta = \mathbb{R}^m$.

Under the above assumptions, we have the following theorem concerning the convergence of the Pop-SAMC algorithm.

Theorem 4. *Let $P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)})$, $i = 1, \ldots, \kappa$, denote the respective Markov transition kernels used for generating the samples $x_{t+1}^{(1)}, \ldots, x_{t+1}^{(\kappa)}$ at iteration t. Let $\{\gamma_t\}$ be a gain factor sequence satisfying (A4). If Θ is compact, all subregions are nonempty, and each of the transition kernels satisfies (A3)(i), then, as $t \to \infty$,*

$$\theta_t \to \theta_* \quad \textit{almost surely},$$

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

where $\theta_* = (\theta_*^{(1)}, \dots, \theta_*^{(m)})$ *is given by*

$$\theta_*^{(i)} = C + \log\biggl(\int_{E_i} \psi(x)\,\mathrm{d}x\biggr) - \log(\pi_i), \qquad i = 1, \dots, m, \tag{12}$$

with C being a constant.

Proof. See Song *et al.* (2013).

The constant C can be determined by imposing a constraint, e.g. $\sum_{i=1}^{m} \mathrm{e}^{\theta_{ti}}$ is equal to a known number.

Remark. As mentioned above, if some regions are empty, the corresponding components of θ_* will converge to $-\infty$ as $n \to \infty$. In this case, as shown in the supplementary paper Song *et al.* (2013), we have

$$\theta_*^{(i)} = \begin{cases} C + \log\biggl(\int_{E_i} \psi(x)\,\mathrm{d}x\biggr) - \log(\pi_i + \nu) & \text{if } E_i \neq \varnothing, \\ -\infty & \text{if } E_i = \varnothing, \end{cases}$$

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

The Doeblin condition implies the existence of the stationary distribution $f_{\theta_t}(x)$ for each $\theta_t \in \Theta$, and P_θ is uniformly ergodic. For this condition to be satisfied, we assume that $\mathcal{X}$ is compact and $f(x)$ is bounded away from 0 and ∞ on $\mathcal{X}$. This assumption is true for many Bayesian model selection problems, e.g. change-point identification and regression variable selection problems. For these problems, after integrating out model parameters from their posterior, the sample space is reduced to a finite set of models. For continuous systems, one may restrict $\mathcal{X}$ to the region $\{x \colon \psi(x) \geq \psi_{\min}\}$, where $\psi_{\min}$ is sufficiently small such that the region $\{x \colon \psi(x) < \psi_{\min}\}$ is not of interest. For the proposal distribution used in the paper, we assume that it satisfies the local positive condition, that is, there exist two quantities $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $q(x, y) \geq \varepsilon_2$ if $|x - y| \leq \varepsilon_1$, where $q(x, y)$ denotes the proposal mass/density function. In the supplementary paper Song *et al.* (2013), we show that the transition kernel induced by local positive proposal satisfies the Doeblin condition. The local positive condition is quite standard and has been widely used in the study of MCMC convergence; see, e.g. Roberts and Tweedie (1996).

Theorem 5 below concerns the asymptotic normality of θ_t.

Theorem 5. *Assume that the conditions of Theorem 4 hold. Conditioned on* $\Lambda(\theta_*) = \{\theta_t \to \theta_*\}$,

$$\frac{\theta_t - \theta_*}{\sqrt{\gamma_t}} \Rightarrow \mathbb{N}(0, \Sigma),$$

where θ_* *is as defined in (12) and*

$$\Sigma = \int_0^\infty \mathrm{e}^{(F' + \zeta I)t}\,\Gamma\,\mathrm{e}^{(F + \zeta I)t}\,\mathrm{d}t,$$

with F *defined as in (A2),* ζ *defined as in (5), and* Γ *defined as in Lemma 1.*

Proof. See Song *et al.* (2013).

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

Finally, we note that Theorem 3 is also valid for the SAMC and Pop-SAMC algorithms. Here we would like to emphasize that even when the gain factor sequence is chosen as $\gamma_t = O(1/t)$, the Pop-SAMC algorithm still has some numerical advantages in convergence over the SAMC algorithm due to the population effect. See Figure 2 in Section 4.

3.4. Minorization properties of the crossover operator

The Pop-SAMC algorithm works on a population of Markov chains. Its population setting provides a basis for including more global, advanced MCMC operators, such as the crossover operator of the genetic algorithm, in simulations. Without loss of generality, we assume that the crossover operator works only on the first and second chains of the population. The resulting transition kernel can be written as

$$\boldsymbol{P}_{\theta_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = P_{\theta_t \times \theta_t}\{(x_t^{(1)}, x_t^{(2)}), (x_{t+1}^{(1)}, x_{t+1}^{(2)})\} \prod_{i=3}^{\kappa} P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)}), \tag{13}$$

which is a product of $\kappa - 1$ independent transition kernels, where

$$\begin{aligned} &P_{\theta_t \times \theta_t}\{(x_t^{(1)}, x_t^{(2)}), (x_{t+1}^{(1)}, x_{t+1}^{(2)})\} \\ &\quad = (1 - r_{\mathrm{co}}) \prod_{i=1}^{2} P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)}) + r_{\mathrm{co}} P_{\theta_t,\mathrm{co}}\{(x_t^{(1)}, x_t^{(2)}), (x_{t+1}^{(1)}, x_{t+1}^{(2)})\}, \end{aligned}$$

where r_{co} is the probability of applying the crossover kernel $P_{\theta_t,\mathrm{co}}$. Following the proof given in the supplementary paper Song *et al.* (2013), $\prod_{i=1}^{2} P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)})$ is locally positive, which implies that $P_{\theta_t \times \theta_t}$ is locally positive as well, if $r_{\mathrm{co}} < 1$. As long as $\mathcal{X}$ is compact, and $f(x)$ is bounded away from 0 and ∞, (A3)(i) is satisfied by $P_{\theta_t,\mathrm{co}}$. Condition (A3)(ii) is satisfied because it is independent of the kernel used. Condition (A3)(iii) can be verified as follows. Let $s_\theta(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{x}, \boldsymbol{y}) \min\{1, r(\theta, \boldsymbol{x}, \boldsymbol{y})\}$, where $\boldsymbol{x} = (x^{(1)}, x^{(2)})$, $\boldsymbol{y} = (y^{(1)}, y^{(2)})$, and

$$r(\theta, \boldsymbol{x}, \boldsymbol{y}) = \frac{f_\theta(y^{(1)}) f_\theta(y^{(2)})}{f_\theta(x^{(1)}) f_\theta(x^{(2)})} \frac{q(\boldsymbol{y}, \boldsymbol{x})}{q(\boldsymbol{x}, \boldsymbol{y})}$$

is the MH ratio for the crossover operator. It is easy to see that

$$\begin{aligned} \left| \frac{\partial s_\theta(\boldsymbol{x}, \boldsymbol{y})}{\partial \theta_i} \right| &= q(\boldsymbol{x}, \boldsymbol{y}) \mathbf{1}_{\{r(\theta, \boldsymbol{x}, \boldsymbol{y}) < 1\}} r(\theta, \boldsymbol{x}, \boldsymbol{y}) \\ &\quad \times |\mathbf{1}_{\{x^{(1)} \in E_i\}} + \mathbf{1}_{\{x^{(2)} \in E_i\}} - \mathbf{1}_{\{y^{(1)} \in E_i\}} - \mathbf{1}_{\{y^{(2)} \in E_i\}}| \\ &\le 2q(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

The mean value theorem implies that there exists a constant c such that

$$\|s_\theta(\boldsymbol{x}, \boldsymbol{y}) - s_{\theta'}(\boldsymbol{x}, \boldsymbol{y})\| \le c q(\boldsymbol{x}, \boldsymbol{y}) \|\theta - \theta'\|.$$

Following the same argument as in Liang *et al.* (2007), (A3)(iii) is satisfied by $P_{\theta_t,\mathrm{co}}$. This means that each kernel on the right-hand side of (13) satisfies the drift condition (A3). Therefore, the product kernel $\boldsymbol{P}_{\theta_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ satisfies the drift condition. Then the convergence and asymptotic normality of θ_t (Theorem 4 and Theorem 5) still hold for this general Pop-SAMC algorithm with crossover operators. We conjecture that incorporating crossover operators will make the Pop-SAMC algorithm more efficient. How these advanced operators improve the performance of the Pop-SAMC algorithm will be explored elsewhere.

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

4. An illustrative example

To illustrate the performance of the Pop-SAMC algorithm, we study a multimodal example taken from Liang and Wong (2001). The density function over a bivariate $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \sum_{i=1}^{20} \alpha_i \exp\biggl(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top(\boldsymbol{x}-\boldsymbol{\mu}_i)\biggr),$$

where each component has an equal variance $\sigma^2 = 0.01$ and an equal weight $\alpha_1 = \cdots = \alpha_{20} = 0.05$, and the mean vectors $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{20}$ are given in Liang and Wong (2001). Since some components of the mixture distribution are far from others, e.g. the distance between the lower-right component and its nearest neighboring component is 31.4 times the standard deviation, sampling from this distribution is a challenge using existing MCMC algorithms.

We set the sample space $\mathcal{X} = [-10^{100}, 10^{100}]^2$, and then partition it according to the energy function $U(x) = -\log\{p(x)\}$ with an equal energy bandwidth $\Delta u = 0.5$ into the subregions $E_1 = \{x \colon U(x) \le 0\}$, $E_2 = \{x \colon 0 < U(x) \le 0.5\}, \dots, E_{20} = \{x \colon U(x) > 9.0\}$. The Pop-SAMC algorithm was first tested on this example with two gain factor sequences, $\gamma_t = 100/\max(100, t)$ and $\gamma_t = 100/\max(100, t^{0.6})$. In simulations, we set the population size $\kappa = 10$, the number of iterations $N = 10^6$, and the desired sampling distribution to be uniform, i.e. $\pi_1 = \cdots = \pi_{20} = \frac{1}{20}$. The Gaussian random walk proposal distribution was used in the MH sampling step with a covariance matrix of $4I_2$, where I_2 is the 2×2 identity matrix. For a fair comparison with the SAMC algorithm, we initialize the population in a small region $[0, 1] \times [0, 1]$, which is far from the separated components. In Tables 1 and 2 we present the resulting estimates of $P(E_i)$ (i.e. $w_i = \int_{E_i} p(x)\,\mathrm{d}x$) for $i = 2, \dots, 11$, based on 100 independent runs. The computation was done on an Intel® Core™ 2 Duo 3.0GHz computer. As can be seen from the true values of the $P(E_i)$, which were calculated with a total of 2×10^9 samples drawn equally from each of the 20 components of $p(x)$, the subregions

Table 1: A comparison of the efficiencies of the Pop-SAMC and SAMC estimators for the multimodal example with $\gamma_t = t_0/\max\{t_0, t\}$. The numbers in parentheses indicate the standard errors of the $P(E_i)$, $i = 2, \dots, 11$, estimates.

		Pop-SAMC	SAMC	
Setting	True	$(t_0, \tau, N) = (100, 10, 10^6)$	$(t_0, N) = (100, 10^7)$	$(t_0, N) = (1000, 10^7)$
$P(E_2)$	0.2387	0.2383 (0.0003)	0.2390 (0.0003)	0.2382 (0.0008)
$P(E_3)$	0.3027	0.3027 (0.0003)	0.3024 (0.0003)	0.3030 (0.0008)
$P(E_4)$	0.1856	0.1859 (0.0002)	0.1859 (0.0002)	0.1852 (0.0006)
$P(E_5)$	0.1124	0.1124 (0.0001)	0.1121 (0.0001)	0.1126 (0.0004)
$P(E_6)$	0.0663	0.0663 (0.0001)	0.0662 (0.0001)	0.0666 (0.0003)
$P(E_7)$	0.0384	0.0384 (0)	0.0384 (0)	0.0383 (0.0001)
$P(E_8)$	0.0226	0.0226 (0)	0.0225 (0)	0.0227 (0.0001)
$P(E_9)$	0.0134	0.0134 (0)	0.0134 (0)	0.0135 (0.0001)
$P(E_{10})$	0.0080	0.0080 (0)	0.0080 (0)	0.0079 (0)
$P(E_{11})$	0.0048	0.0048 (0)	0.0048 (0)	0.0048 (0)
CPU (s)	—	18	21	21

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

TABLE 2: A comparison of the efficiencies of the Pop-SAMC and SAMC estimators for the multimodal example with $\gamma_t = t_0/\max\{t_0, t^{0.6}\}$. The numbers in parentheses indicate the standard errors of the $P(E_i)$, $i = 2, \dots, 11$, estimates.

		Pop-SAMC	SAMC	
Setting	True	$(t_0, \tau, N) = (100, 10, 10^6)$	$(t_0, N) = (100, 10^7)$	$(t_0, N) = (1000, 10^7)$
$P(E_2)$	0.2387	0.2236 (0.0042)	0.2244 (0.0065)	0.1534 (0.0184)
$P(E_3)$	0.3027	0.3045 (0.0041)	0.3123 (0.0076)	0.3329 (0.0268)
$P(E_4)$	0.1856	0.1909 (0.0035)	0.1850 (0.0054)	0.1815 (0.0207)
$P(E_5)$	0.1124	0.1156 (0.0024)	0.1167 (0.0039)	0.1205 (0.0144)
$P(E_6)$	0.0663	0.0706 (0.0015)	0.0648 (0.0020)	0.0715 (0.0096)
$P(E_7)$	0.0384	0.0387 (0.0008)	0.0390 (0.0013)	0.0542 (0.0071)
$P(E_8)$	0.0226	0.0227 (0.0005)	0.0243 (0.0010)	0.0303 (0.0058)
$P(E_9)$	0.0134	0.0133 (0.0003)	0.0137 (0.0005)	0.0218 (0.0069)
$P(E_{10})$	0.0080	0.0082 (0.0002)	0.0079 (0.0003)	0.0184 (0.0050)
$P(E_{11})$	0.0048	0.0047 (0.0001)	0.0047 (0.0002)	0.0047 (0.0009)
CPU(s)	—	18	23	22

$E_2, \dots, E_{11}$ covered more than 99% of the total mass of the distribution. For comparison, the SAMC algorithm was also applied to this example, but with $N = 10^7$ iterations and four gain factor sequences: $\gamma_t = 100/\max(100, t)$, $\gamma_t = 1000/\max(1000, t)$, $\gamma_t = 100/\max(100, t^{0.6})$, and $\gamma_t = 1000/\max(1000, t^{0.6})$. These settings ensure that each run of the Pop-SAMC and SAMC algorithms consists of the same number of energy evaluations and, thus, costs about the same CPU time. The resulting estimates for the $P(E_i)$ are given in Tables 1 and 2.

Our numerical results agree extremely well with Theorem 3. It follows from the delta method (see, e.g. Casella and Berger (2002)) that the mean-square errors (MSEs) of the $P(E_i)$ estimates should follow the same limiting rule (8) as θ_t. For this example, when the same gain factor sequence $\gamma_t = 100/\max(100, t)$ is used, the SAMC estimator is as efficient as the Pop-SAMC estimator when the number of iterations is large; the two estimators share the same standard errors as reported in Table 1. When the gain factor sequence $\gamma_t = 1000/\max(1000, t)$ is used for SAMC, the runs of the SAMC and Pop-SAMC algorithms end with the same gain factor values. In this case, as expected, the SAMC estimator has larger standard errors than the Pop-SAMC estimator; the relative efficiency of these two estimators is about 3.0^2 $(3 \approx (0.0008 + \cdots + 0.0003)/(0.0003 + \cdots + 0.0001))$, which is close to the theoretical value 10. When the gain factor sequence $\gamma_t = 100/\max(100, t^{0.6})$ is used, the Pop-SAMC estimator is more efficient than the SAMC estimator. The results given in Table 2 show that the relative efficiency of the Pop-SAMC estimator versus the SAMC estimator is about 2.56 $(= 1.6^2$ and $1.6 \approx (0.0065 + \cdots + 0.0002)/(0.0042 + \cdots + 0.0001))$, which agrees well with the theoretical value 2.51 $(= 10^{0.4})$.

We note that the results reported in Tables 1 and 2 are only for the scenario in which the number of iterations is large. For a thorough comparison, we evaluated the MSEs of the Pop-SAMC and SAMC estimators at 100 equally spaced time points, with iterations of $10^4 \sim 10^6$ for the Pop-SAMC algorithm and $10^5 \sim 10^7$ for the SAMC algorithm. The results are shown in Figure 1. The plots indicate that the Pop-SAMC algorithm converges much faster than the SAMC algorithm, even when the gain factor sequence $\gamma_t = t_0/\max(t_0, t)$ is used. As discussed

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

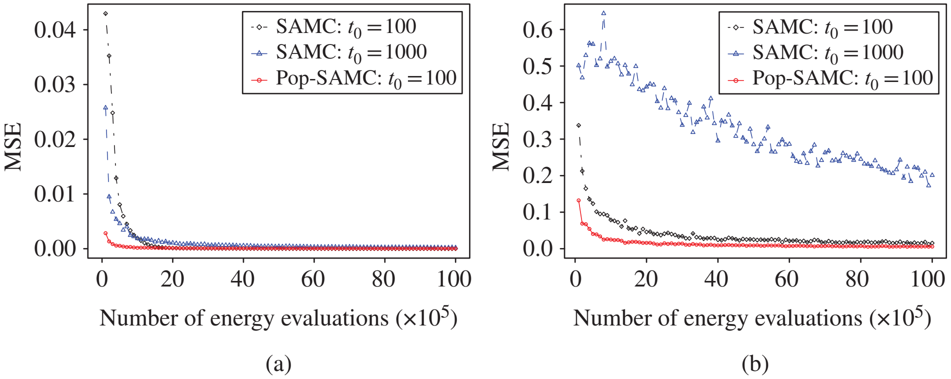

FIGURE 1: Mean-square errors (MSEs) produced by the Pop-SAMC and SAMC algorithms at different iterations. The left plot is produced with $\gamma_t = t_0/\max(t_0, t)$ and the right plot is produced with $\gamma_t = t_0/\max(t_0, t^{0.6})$.

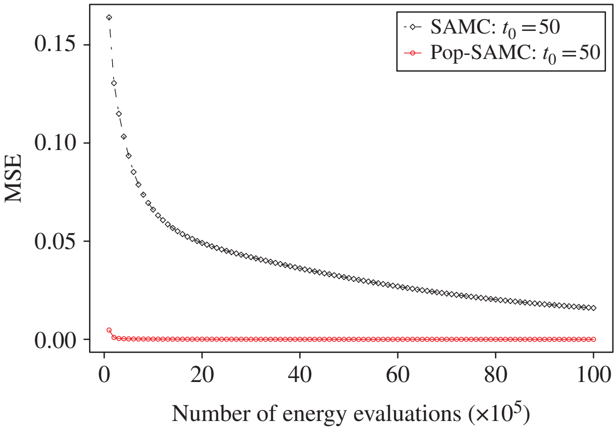

FIGURE 2: Mean-square errors (MSEs) produced by the Pop-SAMC and SAMC algorithms at different iterations with the gain factor sequence $\gamma_t = 50/\max(50, t)$.

previously, this is due to the population effect: the Pop-SAMC algorithm provides a more accurate estimator of $h(\theta_t)$ at each iteration, and this improves its convergence, especially at the early stage of the simulation.

To further explore the population effect of the Pop-SAMC algorithm, both the Pop-SAMC and SAMC algorithms were re-run 100 times with a smaller gain factor sequence $\gamma_t = 50/\max(50, t)$. The results plotted in Figure 2 show that, under this setting, the SAMC algorithm converges very slowly, while the Pop-SAMC algorithm still converges very fast. This experiment shows that the Pop-SAMC algorithm is more robust to the choice of gain factor sequence, and it can work with a smaller gain factor sequence than can the SAMC algorithm.

5. Conclusion

In this paper we proposed a population SAMCMC algorithm and compared its convergence rate with that of the single-chain SAMCMC algorithm. As the main theoretical result, we

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

established the limiting ratio between the L_2 rates of convergence of the two types of SAMCMC algorithm. Our result provided a theoretical guarantee that the population SAMCMC algorithm is asymptotically more efficient than the single-chain SAMC algorithm when the gain factor sequence $\{\gamma_t\}$ decreases slower than $O(1/t)$. This theoretical result was confirmed with a numerical example.

In this paper we also proved the asymptotic normality of SAMCMC estimators under mild conditions. As mentioned previously, the major differences between this work and Benveniste *et al.* (1990) are the assumptions on the Markov transition kernels. Our assumptions are easier to verify than those of Benveniste *et al.* (1990). We note that the work of Chen (2002) and Pelletier (1998) can potentially be extended to SAMCMC algorithms. The major differences between their work and ours are the assumptions on the observation noise. In Chen (2002, Theorem 3.3.2, p. 128) and Pelletier (1998), it is assumed that the observation noise can be decomposed in the form

$$\varepsilon_t = e_t + \nu_t,$$

where $\{e_t\}$ forms a martingale difference sequence and $\{\nu_t\}$ is a higher-order term of $O(\sqrt{\gamma_t})$. However, as shown in Lemma 1, the SAMCMC algorithms do not satisfy this assumption.

Appendix A. Proof of Theorem 1

To prove Theorem 1, we first introduce the following lemmas. Lemma 2 is a combined restatement of Theorem 2 of Andrieu and Moulines (2006), Proposition 6.1 of Andrieu *et al.* (2005), and Lemma 5 of Andrieu and Moulines (2006).

Lemma 2. *Assume that Θ is compact and that condition (A3) holds. Then the following results hold.*

(B1) *For any $\theta \in \Theta$, the Markov kernel P_θ has a single stationary distribution π_θ. In addition, $H\colon \Theta \times \mathbb{X} \to \Theta$ is measurable and, for all $\theta \in \Theta$, $\int_{\mathbb{X}} \|H(\theta, x)\| \pi_\theta(x)\,\mathrm{d}x < \infty$.*

(B2) *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 $P_\theta u_\theta(X) = \int_{\mathbb{X}} u_\theta(y) P_\theta(X, y)\,\mathrm{d}y$. For any $\eta \in (0, 1)$, the following conditions hold:*

(i) $\sup_{\theta\in\Theta}(\|u_\theta(\cdot)\| + \|P_\theta u_\theta(\cdot)\|) < \infty$;

(ii) $\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-\eta}\{\|u_\theta(\cdot) - u_{\theta'}(\cdot)\| + \|P_\theta u_\theta(\cdot) - P_{\theta'} u_{\theta'}(\cdot)\|\} < \infty$.

(B3) *For any $\eta \in (0, 1)$,*

$$\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-\eta} \|h(\theta) - h(\theta')\| < \infty.$$

Tadić (1997) studied the convergence of the SAMCMC algorithm under different conditions from those given in Andrieu *et al.* (2005) and Andrieu and Moulines (2006). We combined some of the results from these three papers to obtain the following lemma, which corresponds to Theorem 4.1 and Lemma 2.2 of Tadić (1997).

Lemma 3. *Assume that the conditions of Theorem 1 hold. Then the following results hold.*

(C1) *There exist $\mathbb{R}^{d_\theta}$-valued random processes $\{\varepsilon_t\}_{t\ge0}$, $\{\varepsilon'_t\}_{t\ge0}$, and $\{\varepsilon''_t\}_{t\ge0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that*

$$\gamma_{t+1}\xi_{t+1} = \varepsilon_{t+1} + \varepsilon'_{t+1} + \varepsilon''_{t+1} - \varepsilon''_t, \qquad t \ge 0, \tag{14}$$

where $\xi_{t+1} = H(\theta_t, X_{t+1}) - h(\theta_t)$.

(C2) *The series* $\sum_{t=0}^{\infty} \|\varepsilon_t'\|$, $\sum_{t=0}^{\infty} \|\varepsilon_t''\|^2$, *and* $\sum_{t=0}^{\infty} \|\varepsilon_{t+1}\|^2$ *all converge almost surely and*

$$\mathbb{E}(\varepsilon_{t+1} \mid \mathcal{F}_t) = 0 \quad \textit{almost surely}, \qquad n \geq 0, \tag{15}$$

where $\{\mathcal{F}_t\}_{t\geq 0}$ *is a family of* σ*-algebras of* $\mathcal{F}$ *which satisfies* $\sigma\{\theta_0\} \subseteq \mathcal{F}_0$ *and* $\sigma\{\varepsilon_t, \varepsilon_t', \varepsilon_t''\} \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, $t \geq 0$.

(C3) *Let* $R_t = R_t' + R_t''$, $t \geq 1$, *where* $R_t' = \gamma_{t+1} \nabla^\top v(\theta_t) \xi_{t+1}$ *and*

$$R_{t+1}'' = \int_0^1 [\nabla v(\theta_t + s(\theta_{t+1} - \theta_t)) - \nabla v(\theta_t)]^\top (\theta_{t+1} - \theta_t)\, \mathrm{d}s.$$

Then $\sum_{t=1}^{\infty} \gamma_t \xi_t$ *and* $\sum_{t=1}^{\infty} R_t$ *converge almost surely.*

Proof. (C1) Let $\varepsilon_0 = \varepsilon_0' = 0$, and let

$$\begin{aligned}
\varepsilon_{t+1} &= \gamma_{t+1}[u_{\theta_t}(x_{t+1}) - P_{\theta_t} u_{\theta_t}(x_t)],\\
\varepsilon_{t+1}' &= \gamma_{t+1}[P_{\theta_{t+1}} u_{\theta_{t+1}}(x_{t+1}) - P_{\theta_t} u_{\theta_t}(x_{t+1})] + (\gamma_{t+2} - \gamma_{t+1}) P_{\theta_{t+1}} u_{\theta_{t+1}}(x_{t+1}),\\
\varepsilon_t'' &= -\gamma_{t+1} P_{\theta_t} u_{\theta_t}(x_t).
\end{aligned}$$

It is easy to verify that (14) is satisfied.

(C2) Since

$$\mathbb{E}(u_{\theta_t}(x_{t+1}) \mid \mathcal{F}_t) = P_{\theta_t} u_{\theta_t}(x_t),$$

we obtain (15). It follows from (B2), (A3), and (A4) that there exist constants $c_3, c_4, c_5, c_6, c_7 \in \mathbb{R}^+$ such that

$$\|\varepsilon_{t+1}\|^2 \leq 2c_3 \gamma_{t+1}^2, \qquad \|\varepsilon_{t+1}''\|^2 \leq c_4 \gamma_{t+1}^2,$$
$$\|\varepsilon_{t+1}'\| \leq c_5 \gamma_{t+1} \|\theta_{t+1} - \theta_t\|^\eta + c_6 \gamma_{t+1}^{1+\tau} \leq c_7 \gamma_{t+1}^{1+\eta},$$

for any $\eta \in (0, 1)$. Following from (3) and setting $\eta \geq \tau'$ (τ' is defined in (A4)), we have

$$\sum_{t=0}^{\infty} \|\varepsilon_{t+1}\|^2 < \infty, \qquad \sum_{t=0}^{\infty} \|\varepsilon_{t+1}'\| < \infty, \qquad \sum_{t=0}^{\infty} \|\varepsilon_{t+1}''\|^2 < \infty,$$

which, by Fubini's theorem, implies that the series $\sum_{t=0}^{\infty} \|\varepsilon_{t+1}\|^2$, $\sum_{t=0}^{\infty} \|\varepsilon_{t+1}'\|$, and $\sum_{t=0}^{\infty} \|\varepsilon_{t+1}''\|^2$ all converge almost surely to some finite-value random variables.

(C3) Let $M = \sup_{\theta\in\Theta} \max\{\|h(\theta)\|, \|\nabla v(\theta)\|\}$, and let L be the Lipschitz constant of $\nabla v(\cdot)$. Since $\sigma\{\theta_t\} \subset \mathcal{F}_t$, it follows from (C2) that $\mathbb{E}(\nabla^\top v(\theta_t)\varepsilon_{t+1} \mid \mathcal{F}_t) = 0$. In addition, we have

$$\sum_{t=0}^{\infty} \mathbb{E}(|\nabla^\top v(\theta_t)\varepsilon_{t+1}|)^2 \leq M^2 \sum_{t=0}^{\infty} \mathbb{E}(\|\varepsilon_{t+1}\|^2) < \infty.$$

It follows from the martingale convergence theorem (see Hall and Heyde (1980, Theorem 2.15)) that both $\sum_{t=0}^{\infty} \varepsilon_{t+1}$ and $\sum_{t=0}^{\infty} \nabla^\top v(\theta_t)\varepsilon_{t+1}$ converge almost surely. By noting C_2 and the following inequalities,

$$\sum_{t=0}^{\infty} |\nabla^\top v(\theta_t)\varepsilon_{t+1}'| \leq M \sum_{t=1}^{\infty} \|\varepsilon_t'\|,$$

$$\sum_{t=1}^{\infty} \gamma_t^2 \|\xi_t\|^2 \leq C\biggl(\sum_{t=1}^{\infty} \|\varepsilon_t\|^2 + \sum_{t=1}^{\infty} \|\varepsilon_t'\|^2 + \sum_{t=0}^{\infty} \|\varepsilon_t''\|^2\biggr),$$

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

for some constant C, both $\sum_{t=0}^{\infty}|\nabla^{\top}v(\theta_t)\varepsilon'_{t+1}|$ and $\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2$ converge. In addition,

$$\begin{aligned}\|R''_{t+1}\| \le L\|\theta_{t+1}-\theta_t\|^2 &= L\|\gamma_{t+1}h(\theta_t)+\gamma_{t+1}\xi_{t+1}\|^2\\ &\le 2L(M^2\gamma_{t+1}^2+\gamma_{t+1}^2\|\xi_{t+1}\|^2),\\ |(\nabla v(\theta_{t+1})-\nabla v(\theta_t))^{\top}\varepsilon''_{t+1}| &\le L\|\theta_{t+1}-\theta_t\|\|\varepsilon''_{t+1}\|,\end{aligned}$$

for all $t \ge 0$. Consequently,

$$\sum_{t=1}^{\infty}|R''_t| \le 2LM^2\sum_{t=1}^{\infty}\gamma_t^2 + 2L\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2 < \infty,$$

$$\begin{aligned}&\sum_{t=0}^{\infty}|(\nabla v(\theta_{t+1})-\nabla v(\theta_t))^{\top}\varepsilon''_{t+1}|\\ &\qquad\le \biggl(2L^2M^2\sum_{t=1}^{\infty}\gamma_t^2+2L^2\sum_{t=1}^{\infty}\gamma_t^2\|\xi_t\|^2\biggr)^{1/2}\biggl(\sum_{t=1}^{\infty}\|\varepsilon''_t\|^2\biggr)^{1/2}\\ &\qquad< \infty.\end{aligned}$$

Since

$$\begin{aligned}\sum_{t=1}^{n}\gamma_t\xi_t &= \sum_{t=1}^{n}\varepsilon_t+\sum_{t=1}^{n}\varepsilon'_t+\varepsilon''_n-\varepsilon''_0,\\ \sum_{t=0}^{n}R'_{t+1} &= \sum_{t=0}^{n}\nabla^{\top}v(\theta_t)\varepsilon_{t+1}+\sum_{t=0}^{n}\nabla^{\top}v(\theta_t)\varepsilon'_{t+1}-\sum_{t=0}^{n}(\nabla v(\theta_{t+1})-\nabla v(\theta_t))^{\top}\varepsilon''_{t+1}\\ &\quad+\nabla^{\top}v(\theta_{n+1})\varepsilon''_{n+1}-\nabla^{\top}v(\theta_0)\varepsilon''_0,\end{aligned}$$

and ε''_n converges to 0 by (C2), it is obvious that $\sum_{t=1}^{\infty}\gamma_t\xi_t$ and $\sum_{t=1}^{\infty}R_t$ converge almost surely. This completes the proof.

Based on Lemma 3, Theorem 1 can be proved in a similar way to Theorem 2.2 of Tadić (1997). Since the paper Tadić (1997) is not openly available, we reproduce the proof of Theorem 1 in the Supplementary paper Song *et al.* (2013).

Appendix B. Proofs of Lemma 1, Theorem 2, and Theorem 3

B.1. Proof of Lemma 1

Lemma 4 below is a restatement of Proposition 6.1 of Andrieu *et al.* (2005). It has a little overlap with (B2).

Lemma 4. *Assume that (A3)(i) and (A3)(iii) hold. Suppose that the family of functions* $\{g_\theta,\ \theta\in\Theta\}$ *satisfies the following condition. For any compact subset* $\mathcal{K}\subset\Theta$,

$$\sup_{\theta\in\mathcal{K}}\|g_\theta(\cdot)\| < \infty, \qquad \sup_{(\theta,\theta')\in\mathcal{K}\times\mathcal{K}}|\theta-\theta'|^{-\iota}\|g_\theta(\cdot)-g_{\theta'}(\cdot)\| < \infty \tag{16}$$

for some $\iota\in(0,1)$. *Let* $u_\theta(x)$ *be the solution to the Poisson equation* $u_\theta(x)-P_\theta u_\theta(x) = g_\theta(x)-\pi_\theta(g_\theta(x))$, *where* $\pi_\theta(g_\theta(x)) = \int_{\mathbb{X}}g_\theta(x)\pi_\theta(x)\,\mathrm{d}x$. *Then, for any compact set* $\mathcal{K}$ *and*

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

any $\iota' \in (0, \iota)$,

$$\sup_{\theta \in \mathcal{K}} (\|u_\theta(\cdot)\| + \|P_\theta u_\theta(\cdot)\|) < \infty,$$

$$\sup_{(\theta,\theta') \in \mathcal{K} \times \mathcal{K}} \|\theta - \theta'\|^{-\iota'} \{\|u_\theta(\cdot) - u_{\theta'}(\cdot)\| + \|P_\theta u_\theta(\cdot) - P_{\theta'} u_{\theta'}(\cdot)\|\} < \infty.$$

Lemma 5 below can be viewed as a partial restatement of Proposition 7 of Andrieu and Moulines (2006), but under different conditions.

Lemma 5. *Assume that* Θ *is compact, and that conditions (A3) and (A4)(i) hold. Let* $\{g_\theta, \theta \in \Theta\}$ *be a family of functions satisfying (16) with* $\iota \in ((1 + \tau')/2, 1)$, *where* τ' *is defined in condition (A4). Then*

$$n^{-1} \sum_{k=1}^{n} \biggl(g_{\theta_k}(X_k) - \int_{\mathbb{X}} g_{\theta_k}(x) \, \mathrm{d}\pi_{\theta_k}(x) \biggr) \to 0 \quad \textit{almost surely}$$

for any starting point (θ_0, X_0).

Proof. Without loss of generality, we assume that g_θ takes values on $\mathbb{R}$. (If g_θ takes values on $\mathbb{R}^d$, the proof proceeds elementwise.) Let $S_n = \sum_{k=1}^{n} [g_{\theta_k}(X_k) - \pi_{\theta_k}(g_{\theta_k}(X_k))]$, where $\pi_{\theta_k}(g_{\theta_k}(X_k)) = \int_{\mathbb{X}} g_{\theta_k}(x) \pi_{\theta_k}(x) \, \mathrm{d}x$. Let $S_n' = \sum_{k=1}^{n} [u_{\theta_k} - P_{\theta_k} u_{\theta_k}]$, where u_{θ_k} is the solution to the Poisson equation

$$u_{\theta_k} - P_{\theta_k} u_{\theta_k} = g_{\theta_k}(X_k) - \pi_{\theta_k}(g_{\theta_k}(X_k)).$$

Furthermore, we decompose S_n into three terms, $S_n = S_n^{(1)} + S_n^{(2)} + S_n^{(3)}$, where

$$\begin{aligned} S_n^{(1)} &= \sum_{k=1}^{n} [u_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1})], \\ S_n^{(2)} &= \sum_{k=1}^{n} [u_{\theta_k}(X_k) - u_{\theta_{k-1}}(X_k)], \\ S_n^{(3)} &= P_{\theta_0} u_{\theta_0}(X_0) - P_{\theta_n} u_{\theta_n}(X_n). \end{aligned}$$

By Lemma 4, for all θ and X, there exists a constant c such that

$$|u_\theta(X)| \le c \quad \text{and} \quad |P_\theta u_\theta(X)| \le c.$$

Let $p > 2$, and let $(1 + \tau')/2 \le \iota' < \iota$ (where τ' is defined in (A4)). Thus, there exists a constant c such that

$$\mathbb{E}(|u_{\theta_{k-1}}(X_k) + P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1})|^p) \le c.$$

Since

$$\begin{aligned} \mathbb{E}(u_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) \mid \mathcal{F}_{k-1}) &= P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) - P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) \\ &= 0, \end{aligned}$$

$\{S_n^{(1)}\}$ is a martingale with the increments upper bounded in L^p. Hence, by Burkholder's inequality (see Hall and Heyde (1980, Theorem 2.10)) and Minkowski's inequality, there exist

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

constants c and c' such that

$$\begin{aligned}\mathbb{E}(|S_n^{(1)}|^p) &\le c\mathbb{E}\biggl(\biggl(\sum_{k=1}^{n}|u_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}u_{\theta_{k-1}}(X_{k-1})|^2\biggr)^{p/2}\biggr)\\ &\le c\biggl\{\sum_{k=1}^{n}[\mathbb{E}(|u_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}}u_{\theta_{k-1}}(X_{k-1})|^p)]^{2/p}\biggr\}^{p/2}\\ &\le c'n^{p/2}.\end{aligned}$$

Now we consider $S_n^{(2)}$. By Lemma 4, the fact that $\|\theta_k - \theta_{k-1}\| = \gamma_k\|H(\theta_{k-1}, X_k)\|$, and (A3)(ii),

$$|S_n^{(2)}| = \biggl|\sum_{k=1}^{n}\{u_{\theta_k}(X_k) - u_{\theta_{k-1}}(X_k)\}\biggr| \le c\sum_{k=1}^{n}\|\theta_k - \theta_{k-1}\|^{\iota'} \le c'\sum_{k=1}^{n}\gamma_k^{\iota'}.$$

Hence, $\mathbb{E}(|S_n^{(2)}|^p) \le c'^p(\sum_{k=1}^{n}\gamma_k^{\iota'})^p$. The third term is also bounded by some constant c: $\mathbb{E}(|S_n^{(3)}|^p) < c$. Therefore, by Minkowski's inequality and Markov's inequality, we can conclude that

$$\mathbb{P}\{n^{-1}|S_n| \ge \delta\} \le C\delta^{-p}\biggl\{n^{-p/2} + \biggl(n^{-1}\sum_{k=1}^{n}\gamma_k^{\iota'}\biggr)^p + n^{-p}\biggr\}, \tag{17}$$

where C denotes a constant. By (17) and the Borel–Cantelli lemma, we have

$$\mathbb{P}\{\sup_{n\ge1} n^{-1}|S_n| \ge \delta\} \le C\delta^{-p}\sum_{n\ge1}\biggl\{n^{-p/2} + n^{-p/2}\biggl(n^{-1/2}\sum_{k=1}^{n}\gamma_k^{\iota'}\biggr)^p + n^{-p}\biggr\}.$$

Then, the SLLN is concluded with Kronecker's lemma, condition (2) and the condition $p > 2$.

Proof of Lemma 1. (i) Define

$$e_{k+1} = u_{\theta_k}(X_{k+1}) - P_{\theta_k}u_{\theta_k}(X_k), \tag{18a}$$

$$\nu_{k+1} = [P_{\theta_{k+1}}u_{\theta_{k+1}}(X_{k+1}) - P_{\theta_k}u_{\theta_k}(X_{k+1})] + \frac{\gamma_{k+2} - \gamma_{k+1}}{\gamma_{k+1}}P_{\theta_{k+1}}u_{\theta_{k+1}}(X_{k+1}), \tag{18b}$$

$$\tilde{\varsigma}_{k+1} = \gamma_{k+1}P_{\theta_k}u_{\theta_k}(X_k), \tag{18c}$$

$$\varsigma_{k+1} = \frac{1}{\gamma_{k+1}}(\tilde{\varsigma}_{k+1} - \tilde{\varsigma}_{k+2}), \tag{18d}$$

where $u_{\cdot}(\cdot)$ is the solution of the Poisson equation (see Lemma 2). It is easy to verify that $H(\theta_k, X_{k+1}) - h(\theta_k) = e_{k+1} + \nu_{k+1} + \varsigma_{k+1}$ holds.

(ii) By (18a), we have

$$\mathbb{E}(e_{k+1} \mid \mathcal{F}_k) = \mathbb{E}(u_{\theta_k}(X_{k+1}) \mid \mathcal{F}_k) - P_{\theta_k}u_{\theta_k}(X_k) = 0.$$

Hence, $\{e_k\}$ forms a martingale difference sequence. It follows from (B2) that

$$\sup_{k\ge0}\mathbb{E}(\|e_{k+1}\|^\alpha \mid \mathcal{F}_k)\,\mathbf{1}_{\{\|\theta_k-\theta_*\|\le\rho\}} < \infty,$$

completing the proof of part (ii).

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

(iii) By (18a), we have

$$\mathbb{E}(e_{k+1}e_{k+1}^\top \mid \mathcal{F}_k) = \mathbb{E}(u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^\top \mid \mathcal{F}_k) - P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^\top$$
$$=: l(\theta_k, X_k).$$

By (B2) and (A3)(i), there exist constants c_1, c_2, c_3, and M such that

$$\|l(\theta_k, X_k)\| \le \mathbb{E}(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^\top\| \mid \mathcal{F}_k) + \|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^\top\| < c_1.$$

For any $\theta_k, \theta_k' \in \Theta$,

$$\begin{aligned}&\|l(\theta_k, X_k) - l(\theta_k', X_k)\|\\ &\quad\le \mathbb{E}(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^\top - u_{\theta_k'}(X_{k+1})u_{\theta_k'}(X_{k+1})^\top\| \mid \mathcal{F}_k)\\ &\qquad + \|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^\top - P_{\theta_k'}u_{\theta_k'}(X_k)P_{\theta_k'}u_{\theta_k'}(X_k)^\top\|.\end{aligned} \tag{19}$$

By (B2)(ii), for any $\eta \in (0, 1)$,

$$\begin{aligned}&\|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^\top - P_{\theta_k'}u_{\theta_k'}(X_k)P_{\theta_k'}u_{\theta_k'}(X_k)^\top\|\\ &\quad\le \|(P_{\theta_k}u_{\theta_k}(X_k) - P_{\theta_k'}u_{\theta_k'}(X_k))P_{\theta_k}u_{\theta_k}(X_k)^\top\|\\ &\qquad + \|P_{\theta_k'}u_{\theta_k'}(X_k)(P_{\theta_k}u_{\theta_k}(X_k)^\top - P_{\theta_k'}u_{\theta_k'}(X_k)^\top)\|\\ &\quad\le c_2\|\theta_k - \theta_k'\|^\eta\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^\top - u_{\theta_k'}(X_{k+1})u_{\theta_k'}(X_{k+1})^\top\| \mid \mathcal{F}_k)\\ &\quad\le \mathbb{E}(\|(u_{\theta_k}(X_{k+1}) - u_{\theta_k'}(X_{k+1}))u_{\theta_k}(X_{k+1})^\top\| \mid \mathcal{F}_k)\\ &\qquad + \mathbb{E}(\|u_{\theta_k'}(X_{k+1})(u_{\theta_k}(X_{k+1})^\top - u_{\theta_k'}(X_{k+1})^\top)\| \mid \mathcal{F}_k)\\ &\quad\le c_3\|\theta_k - \theta_k'\|^\eta.\end{aligned}$$

Substitution into (19) yields $\|l(\theta_k, X_k) - l(\theta_k', X_k)\| \le M\|\theta_k - \theta_k'\|^\eta$ for any $\theta_k, \theta_k' \in \Theta$, where M is a constant.

Let $\iota = \eta \in ((\tau' + 1)/2, 1)$. Then the conditions of Lemma 5 hold and, thus,

$$\frac{1}{n}\sum_{k=1}^{n}[l(\theta_k, X_k) - \pi_{\theta_k}(l(\theta_k, X))] \to 0 \quad \text{almost surely,}$$

where $\pi_{\theta_k}(l(\theta_k, X)) = \int_{\mathbb{X}} l(\theta_k, x)\pi_{\theta_k}(x)\,\mathrm{d}x$. On the other hand, we have

$$\begin{aligned}&\|\pi_{\theta_k}(l(\theta_k, X)) - \pi_{\theta_*}(l(\theta_*, X))\|\\ &\quad\le \|\pi_{\theta_k}(l(\theta_k, X) - l(\theta_*, X))\| + \|\pi_{\theta_k}(l(\theta_*, X)) - \pi_{\theta_*}(l(\theta_*, X))\|\\ &\quad\le M\|\theta_k - \theta_*\|^\eta + \|\pi_{\theta_k}(l(\theta_*, X)) - \pi_{\theta_*}(l(\theta_*, X))\|.\end{aligned}$$

Given $\theta_k \to \theta_*$ almost surely, the first term goes to 0 almost surely as $k \to \infty$. By condition (A3), which implies that the conditions of Proposition 1.3.6 of Atchadé *et al.* (2011) hold, $\pi_{\theta_k}(l(\theta_*, X)) - \pi_{\theta_*}(l(\theta_*, X)) \to 0$ almost surely. Thus, $\| \int_{\mathbb{X}} l(\theta_k, x)\,\mathrm{d}\pi_{\theta_k}(x) - \int_{\mathbb{X}} l(\theta_*, x)\,\mathrm{d}\pi_{\theta_*}(x)\| \to 0$ almost surely and

$$\frac{1}{n}\sum_{k=1}^{n} l(\theta_k, X_k) \to \int_{\mathbb{X}} l(\theta_*, x)\,\mathrm{d}\pi_{\theta_*}(x) = \Gamma \quad \text{almost surely} \tag{20}$$

for some positive definite matrix Γ. This completes the proof of part (iii).

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

(iv) By condition (A4) we have

$$\frac{\gamma_{k+2}-\gamma_{k+1}}{\gamma_{k+1}} = O(\gamma_{k+2}^{\tau})$$

for some value $\tau \in [1, 2)$. By (18), (B2)(i), and (B2)(ii), there exists a constant c_1 such that the inequality

$$\|\nu_{k+1}\| \le c_1\|\theta_{k+1}-\theta_k\| + O(\gamma_{k+2}^{\tau}) = c_1\|\gamma_{k+1}H(\theta_k, X_{k+1})\| + O(\gamma_{k+2}^{\tau})$$

holds, which implies, by (1), that there exists a constant c_2 such that

$$\|\nu_{k+1}\| \le c_2\gamma_{k+1}.$$

Therefore,

$$\mathbb{E}\bigg(\frac{\|\nu_k\|^2}{\gamma_k}\bigg)\mathbf{1}_{\{\|\theta_k-\theta_*\|\le\rho\}} \to 0.$$

This completes the proof of part (iv).

(v) A straightforward calculation shows that

$$\gamma_{k+1}\varsigma_{k+1} = \tilde{\varsigma}_{k+1} - \tilde{\varsigma}_{k+2} = \gamma_{k+1}P_{\theta_k}u_{\theta_k}(X_k) - \gamma_{k+2}P_{\theta_{k+1}}u_{\theta_{k+1}}(X_{k+1}).$$

By (B2), $\mathbb{E}(\|P_{\theta_k}u_{\theta_k}(X_k)\|)$ is uniformly bounded with respect to k. Therefore, (v) holds.

B.2. Proof of Theorem 2

To prove Theorem 2, we introduce Lemma 6, which is a combined restatement of Theorem D.6.4 of Meyn and Tweedie (2009, p. 563) and Theorem 1 of Pelletier (1998).

Lemma 6. *Consider a stochastic approximation algorithm of the form*

$$Z_{k+1} = Z_k + \gamma_{k+1}h(Z_k) + \gamma_{k+1}(\nu_{k+1} + e_{k+1}),$$

where ν_{k+1} and e_{k+1} are noise terms. Assume that $\{\nu_k\}$ and $\{e_k\}$ satisfy (ii)–(iv) given in Lemma 1, and that conditions (A2) and (A4) are satisfied. On the set $\Lambda(z^*) = \{Z_k \to z^*\}$,

$$\frac{Z_k - z^*}{\sqrt{\gamma_k}} \Rightarrow \mathbb{N}(0, \Sigma),$$

where $\mathbb{N}$ is the Gaussian distribution and

$$\Sigma = \int_0^{\infty} \mathrm{e}^{(F'+\zeta I)t}\Gamma\mathrm{e}^{(F+\zeta I)t}\,\mathrm{d}t,$$

with F defined as in (A2), ζ defined as in (5), and Γ defined as in Lemma 1.

Proof of Theorem 2. Rewrite the SAMCMC algorithm in the form

$$\theta_{k+1} - \theta_* = (\theta_k - \theta_*) + \gamma_{k+1}h(\theta_k) + \gamma_{k+1}\xi_{k+1}.$$

To facilitate the theoretical analysis for the random process $\{\theta_k\}$, we define a reduced random process $\{\tilde{\theta}_k\}_{k\ge0}$:

$$\tilde{\theta}_k = \theta_k + \tilde{\varsigma}_{k+1}.$$

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

Here $\tilde{\varsigma}_{k+1}$ is as defined in (18c). Then, for the SAMCMC algorithm, we have

$$\begin{aligned}\tilde{\theta}_{k+1} - \theta_* &= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\theta_k) + \gamma_{k+1}\xi_{k+1} + \tilde{\varsigma}_{k+2} - \tilde{\varsigma}_{k+1} \\ &= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(h(\theta_k) - h(\tilde{\theta}_k) + \xi_{k+1} - \varsigma_{k+1}) \\ &= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(h(\theta_k) - h(\tilde{\theta}_k) + \nu_{k+1} + e_{k+1}) \\ &= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(\tilde{\nu}_{k+1} + e_{k+1}),\end{aligned}$$

where $\tilde{\nu}_{k+1} = \nu_{k+1} + h(\theta_k) - h(\tilde{\theta}_k)$, and ς_{k+1}, ν_{k+1}, and e_{k+1} are defined in (18). Since $h(\cdot)$ is Hölder continuous on Θ (by (B3)) and Θ is compact, there exists a constant M such that $\|h(\theta_k) - h(\tilde{\theta}_k)\| \le M\|\tilde{\theta}_k - \theta_k\|^\eta = M\|\tilde{\varsigma}_{k+1}\|^\eta$ for any $\eta \in (0.5, 1)$. Thus, by (18c), there exists a constant c such that

$$\mathbb{E}\left[\frac{\|h(\theta_k) - h(\tilde{\theta}_k)\|^2}{\gamma_k}\right] \le c\gamma_{k+1}^{2\eta-1}\frac{\gamma_{k+1}}{\gamma_k} \to 0,$$

since $\gamma_{k+1}^{2\eta-1} \to 0$ and $\gamma_{k+1}/\gamma_k \to 1$ as $k \to \infty$. Therefore, $\tilde{\nu}_{k+1} = \nu_{k+1} + h(\theta_k) - h(\tilde{\theta}_k)$ also satisfies property (iv) of Lemma 1.

By Lemma 1 and Lemma 6, we have

$$\frac{\tilde{\theta}_k - \theta_*}{\sqrt{\gamma_k}} \Rightarrow \mathbb{N}(0, \Sigma).$$

By Lemma 2, $\mathbb{E}\|P_{\theta_k} u_{\theta_k}(X_k)\|$ is uniformly bounded with respect to k. Hence,

$$\frac{\tilde{\varsigma}_{k+1}}{\sqrt{\gamma_k}} \to 0 \quad \text{in probability}.$$

It follows from Slutsky's theorem (see, e.g. Casella and Berger (2002)) that

$$\frac{\theta_k - \theta_*}{\sqrt{\gamma_k}} \Rightarrow \mathbb{N}(0, \Sigma),$$

which completes the proof.

B.3. Proof of Theorem 3

Let $\boldsymbol{x} = (x^{(1)}, \dots, x^{(\kappa)})$ denote the samples drawn at an iteration of the population SAMCMC algorithm. Let $\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{y})$ and $P(x, y)$ denote the Markovian transition kernels used in the population and single-chain SAMCMC algorithms, respectively. Let $\boldsymbol{H}(\theta, \boldsymbol{x})$ and $H(\theta, x)$ be the parameter updating function associated with the population and single-chain SAMCMC algorithms, respectively. Let $\boldsymbol{u} = \sum_{n\ge0}(\boldsymbol{P}^n\boldsymbol{H} - h)$ be a solution of Poisson equation $\boldsymbol{u} - \boldsymbol{P}\boldsymbol{u} = \boldsymbol{H} - h$, and let $u = \sum_{n\ge0}(P^n H - h)$ be a solution of Poisson equation $u - Pu = H - h$. Since

$$\boldsymbol{H}(\theta, \boldsymbol{x}) = \frac{1}{\kappa}\sum_{i=1}^{\kappa} H(\theta, x^{(i)}),$$

we have $\boldsymbol{u}_\theta(\boldsymbol{x}) = (1/\kappa)\sum_{i=1}^{\kappa} u_\theta(x^{(i)})$. By (18a), we further have

$$\boldsymbol{e}_{t+1} = \frac{1}{\kappa}\sum_{i=1}^{\kappa} e_{t+1}^{(i)}.$$

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

Since $x_{t+1}^{(1)}, \ldots, x_{t+1}^{(\kappa)}$ are mutually independent conditional on $\mathcal{F}_t$, $e_{t+1}^{(1)}, \ldots, e_{t+1}^{(\kappa)}$ are also independent conditional on $\mathcal{F}_t$ and, thus,

$$\mathbf{\Gamma} = \frac{\Gamma}{\kappa},$$

which, by Theorem 2, further implies that

$$\Sigma_p = \frac{\Sigma_s}{\kappa},$$

where Σ_p and Σ_s denote the limiting covariance matrices of the population SAMCMC and single-chain SAMCMC algorithms, respectively. Therefore, $(\theta_t^p - \theta_*)/\sqrt{\gamma_t}$ and $(\theta_{\kappa t}^s - \theta_*)/\sqrt{\kappa\gamma_{\kappa t}}$ both converge in distribution to $\mathbb{N}(0, \Sigma_p)$. By condition (A4), $\gamma_t/(\kappa\gamma_{\kappa t}) = \kappa^{\beta-1}$, which completes the proof.

Acknowledgements

Liang's research was supported in part by the National Science Foundation grants DMS-1106494 and DMS-1317131, and the King Abdullah University of Science and Technology (KAUST) award KUS-C1-016-04. The authors thank the editor, the associate editor, and the anonymous referee for constructive comments that led to a significant improvement of this paper.

References

Aldous, D., Lovász, L. and Winkler, P. (1997). Mixing times for uniformly ergodic Markov chains. *Stoch. Process. Appl.* **71,** 165–185.

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

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

Atchadé, Y. and Fort, G. (2010). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli* **16,** 116–154.

Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2011) Adaptive Markov chain Monte Carlo: theory and methods. In *Bayesian Time Series Models*, Cambridge University Press, pp. 32–51.

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

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edn. Thomson Learning, Pacific Grove, CA.

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

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

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

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation, Fairfax Station, VA, pp. 153–163.

Gilks, W. R., Roberts, G. O., and George, E. I. (1994). Adaptive direction sampling. *J. R. Statist. Soc. Ser. D (The Statistician)* **43,** 179–189.

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

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7,** 223–242.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

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

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

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press

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

LIANG, F. (2010). Trajectory averaging for stochastic approximation MCMC algorithms. *Ann. Statist.* **38,** 2823–2856.

LIANG, F., AND WONG, W. H. (2000). Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statistica Sinica* **10,** 317–342.

LIANG, F., AND WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96,** 653–666.

LIANG, F. AND ZHANG, J. (2009). Learning Bayesian networks for discrete data. *Comput. Statist. Data. Anal.* **53,** 865–876.

LIANG, F., LIU, C. AND CARROLL, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102,** 305–320.

LIU, J. S., LIANG, F. AND WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95,** 121–134.

MARINARI, E. AND PARISI, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19,** 451–458.

METROPOLIS, N. *et al.* (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1092.

MEYN, S. AND TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge University Press.

NUMMELIN, E. (1984), *General Irreducible Markov Chains and Nonnegative Operators.* Cambridge University Press.

PELLETIER, M. (1998). Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Prob.* **8,** 10–44.

ROBBINS, H. AND MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22,** 400–407.

ROBERTS, G. O. AND ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18,** 349–367.

ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83,** 95–110.

SONG, Q., WU, M. AND LIANG, F. (2013). Supplementary material for 'Weak convergence rates of population versus single-chain stochastic approximation MCMC algorithms'. Available at http://www.stat.tamu.edu/∼fliang.

TADIĆ, V. (1997). On the convergence of stochastic iterative algorithms and their applications to machine learning. Technical report, Mihajlo Pupin Institute, Serbia, Yugoslavia. A short version of this paper was published in *Proc. 36th Conf. Decision & Control*, San Diego, CA, pp. 2281–2286.

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.

WONG, W. H. AND LIANG, F. (1997). Dynamic weighting in Monte Carlo and optimization. *Proc. Nat. Acad. Sci. USA* **94,** 14220–14224.

YOUNES, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Prob. Theory Relat. Fields* **82,** 625–645.

ZIEDAN, I. E. (1972). Explicit solution of the Lyapunov-matrix equation. *IEEE Trans. Automatic Control* **17,** 379–381.

https://doi.org/10.1239/aap/1418396243 Published online by Cambridge University Press