Applied Probability Trust (21 October 2013)

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 MCMC 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

[*] Postal address: Department of Statistics, Texas A&M University, College Station, TX 77840, US.
[**] Postal address: Shell Technology Center Houston, 3333 Highway 6 South, Houston, TX 77082, US.
[***] Postal address: Department of Statistics, Texas A&M University, College Station, TX 77840, US.

Email: fliang@stat.tamu.edu

1. Introduction

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

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

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 dependent on θ. The stochastic approximation algorithm is a recursive algorithm which proceeds as follows:

Stochastic Approximation Algorithm

(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 developed into an important area in systems control, and has also served as a prototype for development of recursive algorithms for on-line estimation and control of stochastic systems. Recently, the stochastic approximation algorithm has been used with Markov chain Monte Carlo (MCMC), which replaces the step (a) by a MCMC sampling step:

(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 (Younes, 1989; Gu and Kong, 1998), marginal density estimation (Liang, 2007), and adaptive MCMC (Haario *et al.*, 2001; Andrieu and Moulines, 2006; Roberts and Rosenthal, 2009; Atchadé and Fort, 2009).

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 two-fold. 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, P.332, Theorem 13), 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 by Benveniste *et al.* (1990), and this motivates our design of the population SAMCMC algorithm. Second, we propose a general population SAMCMC algorithm, and contrasts 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 contrasts 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, which compares the performance of SAMC and Pop-SAMC on 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 $\mathcal{X}^\kappa$ where $\boldsymbol{x}_0$ is drawn from. 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:

Population Stochastic Approximation MCMC Algorithm

(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 $f_{\theta_t}(\boldsymbol{x}) = f_{\theta_t}(x^{(1)}) \ldots f_{\theta_t}(x^{(\kappa)})$ as the invariant distribution.

(b) Set $\theta_{t+1} = \theta_t + \gamma_{t+1}\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1})$, where $\boldsymbol{x}_{t+1} = (x_{t+1}^{(1)}, \ldots, x_{t+1}^{(\kappa)})$, 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 a SAMCMC algorithm with the mean field function specified by

$$\begin{aligned} h(\theta) &= \int \boldsymbol{H}(\theta, \boldsymbol{x})\boldsymbol{f}_\theta(\boldsymbol{x})d\boldsymbol{x} \\ &= \int \cdots \int \left[\frac{1}{\kappa}\sum_{i=1}^{\kappa} H(\theta, x^{(i)})\right] f_\theta(x^{(1)}) \ldots f_\theta(x^{(\kappa)}) dx^{(1)} \ldots dx^{(\kappa)} = 0, \end{aligned} \tag{2}$$

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

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 are run in parallel, the population SAMCMC algorithm is able to incorporate some advanced multiple chain operators, such as the crossover operator (Liang and Wong, 2000, 2001), the snooker operator (Gilks *et al.*, 1994) and the gradient operator (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 purpose, we consider in this paper primarily the single-chain operator, for which we have

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

Extension of our convergence result to the general population SAMCMC algorithm which consist 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 that $\Theta = \mathbb{R}^{d_\theta}$ is trivial with 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 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: $\mathbb{X} \to \mathbb{R}^d$, $\boldsymbol{P}_\theta f(X) = \int_{\mathbb{X}} \boldsymbol{P}_\theta(X, y) f(y) dy$.

Lyapunov condition on $h(\theta)$. Let $\mathcal{L} = \{\theta \in \Theta : h(\theta) = \mathbf{0}\}$.

(A_1) The function $h : \Theta \to \mathbb{R}^d$ is continuous, and there exists a continuously differentiable function $v : \Theta \to [0, \infty)$ such that $v_h(\theta) = \nabla^T v(\theta) h(\theta) < 0$ 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.

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.

Stability Condition on $h(\theta)$.

(A_2) The mean field function $h(\theta)$ is measurable and locally bounded on Θ. There exist a stable matrix F (i.e., all eigenvalues of F are with negative real parts), $\rho > 0$, and a constant c such that, for any $\theta_* \in \mathcal{L}$ (defined in A_1),

$$\|h(\theta) - F(\theta - \theta_*)\| \leq c\|\theta - \theta_*\|^2, \quad \forall \theta \in \{\theta : \|\theta - \theta_*\| \leq \rho\}.$$

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, certain approximation may be needed.

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

(A_3) For any given $\theta \in \Theta$, the transition kernel $\boldsymbol{P}_\theta$ is irreducible and aperiodic. In addition,

(i) [Doeblin condition] There exist a constants $\delta > 0$, an integer $l > 0$ and a probability measure ν such that

- $$\inf_{\theta\in\Theta} \boldsymbol{P}^l_\theta(X, A) \geq \delta\nu(A), \quad \forall X \in \mathbb{X},\ \forall A \in \mathcal{B}_{\mathbb{X}}, \tag{4}$$

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 exist a constant $c > 0$ such that for all $X \in \mathbb{X}$,

- $$\sup_{\theta\in\Theta} \|\boldsymbol{H}(\theta, \cdot)\| \leq c. \tag{5}$$
- $$\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-1}\|\boldsymbol{H}(\theta, \cdot) - \boldsymbol{H}(\theta', \cdot)\| \leq c. \tag{6}$$

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

- $$\sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-1}\|\boldsymbol{P}_\theta g - \boldsymbol{P}_{\theta'} g\| \leq c\|g\|. \tag{7}$$

The Doeblin condition of Assumption (A_3)-(i) is equivalent to assuming that the resulting Markov chain has an unique stationary distribution and is uniformly ergodic (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 $\boldsymbol{P}_\theta$. Assumption (A_3)-(ii) gives conditions on $\boldsymbol{H}(\theta, X)$, which directly lead to the boundedness of the observation noise. It is also worthy to note that the property that $\boldsymbol{P}_\theta$ satisfies the condition (A_3)-(i) and (A_3)-(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 $\boldsymbol{P}_\theta$. One can refer to the arguments used in the proof of Theorem 4 in the supplementary material of this paper (Song et al., 2013).

Conditions on step-sizes.

(A_4) It consists of two parts:

(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 non-increasing and satisfies the following conditions:

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \quad \frac{\gamma_{t+1} - \gamma_t}{\gamma_t} = O(\gamma_{t+1}^{\tau}), \quad \sum_{t=1}^{\infty} \frac{\gamma_t^{(1+\tau')/2}}{\sqrt{t}} < \infty, \tag{8}$$

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 eigenvalue of the matrix F (defined in condition A_2) with $\lambda_F < 0$.

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

$$\sum_{t=1}^{\infty} \gamma_t^{1+\tau'} = \sum_{t} (\sqrt{t}\gamma_t^{(1+\tau')/2})(\frac{\gamma_t^{(1+\tau')/2}}{\sqrt{t}}) < \infty, \tag{9}$$

which is often assumed in studying the convergence of stochastic approximations. While condition (8) is often assumed in studying the weak convergence of the trajectory averaging estimator of θ_t (see, e.g., Chen, 2002). (A_4)-(ii) can be applied to the usual gains $\gamma_t = t_0/t^{\beta}$, $1/2 < \beta \leq 1$. Following Pelletier (1998), we deduce that

$$\left(\frac{\gamma_t}{\gamma_{t+1}}\right)^{1/2} = 1 + \frac{\beta}{2t} + o(\frac{1}{t}). \tag{10}$$

In terms of γ_t, (10) can be rewritten as

$$\left(\frac{\gamma_t}{\gamma_{t+1}}\right)^{1/2} = 1 + \zeta\gamma_t + o(\gamma_t), \tag{11}$$

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

Theorem 1 concerns the convergence of the general stochastic approximation MCMC algorithm, whose proof can be found in Appendix A.

Theorem 1. *Assume that Θ is compact and the conditions (A_1), (A_3) and (A_4)-(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 $t \to \infty$,*

$$d(\theta_t, \mathcal{L}) \to 0, \quad a.s.,$$

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

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

$$\theta_{t+1} = \theta_t + \gamma_t [h(\theta_t) + \xi_{t+1}], \tag{12}$$

where $h(\theta_t) = \int_{\mathbb{X}} \boldsymbol{H}(\theta_t, X) f_{\theta_t}(X) dX$, and $\xi_{t+1} = \boldsymbol{H}(\theta_t, X_{t+1}) - h(\theta_t)$ is called the observation noise. Lemma 1 concerns the decomposition of the observation noise, whose parts (i) and (iv) are partial restatement of Lemma A.5 of Liang (2010). The proof can be found in Appendix B.

Lemma 1. *Assume 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 A_2),

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

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 $t \geq 0$ and $\alpha \geq 2$ is a constant.

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

$$\frac{1}{n} \sum_{t=1}^{n} E(e_{t+1} e'_{t+1} | \mathcal{F}_t) \to \Gamma, \quad a.s., \tag{13}$$

where Γ is a positive definite matrix.

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

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

This lemma plays a key role in the proof of Theorem 2, which concerns the asymptotic normality of θ_t. The proof of Theorem 2 can be found in Appendix B.

Theorem 2. *Assume that* Θ *is compact and the conditions* (A_1)*–*(A_4) *hold. Conditioned on* $\Lambda(\theta_*) = \{\theta_t \to \theta_*\}$,

$$\frac{\theta_t - \theta_*}{\sqrt{\gamma_t}} \Longrightarrow \mathbb{N}(0, \Sigma), \tag{14}$$

with $\Longrightarrow$ *denoting the weak convergence,* $\mathbb{N}$ *the Gaussian distribution and*

$$\Sigma = \int_0^\infty e^{(F'+\zeta I)t}\Gamma e^{(F+\zeta I)t}dt, \tag{15}$$

where F *is defined in* (A_2), ζ *is defined in (11), and* Γ *is defined in Lemma 1.*

Remarks

1. The same result 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 (A_3) that $\boldsymbol{P}_\theta$ satisfies a minorization 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 impose a moment condition on X_t (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) present a central limit theorem for the average of $\phi(X_t)$, where $\phi(\cdot)$ is a V^r-Lipschitz function for some $r \in [0, 1/2)$ and $V(\cdot)$ is the drift function. Unlike Andrieu and Moulines (2006), we here 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

$$dX_t = (F + \zeta I)X_t + \Gamma^{1/2} dB_t,$$

where B_t stands for 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 complication.

3. From equation (42) in the proof of Lemma 1, it is not difficult to derive that

$$\Gamma = \sum_{k=-\infty}^{\infty} \int H(\theta_*, x)[P_{\theta_*}^k H(\theta_*, x)]^T d\pi_{\theta_*}(dx), \tag{16}$$

where π_{θ_*} denotes the invariant distribution of the transition kernel P_{θ_*}. This is the same expression of Γ as given in Benveniste *et al.* (1990; equation 4.4.6, p.321). Compared to equation (16), our expression of Γ, given in equation (13), is more interpretable, which corresponds to the asymptotic covariance matrix of e_t. Given the gain factor sequence $\{\gamma_k\}$, the efficiency of a SAMCMC algorithm is determined by Γ. Based on this observation, we show in Theorem 3 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 thus is asymptotically more efficient.

4. The condition "Conditioned on $\Lambda(\theta_*)$" accommodates the case that there exist multiple solutions for the equation $h(\theta) = 0$.

Theorem 3 compares the efficiency of the population SAMCMC and the single-chain SAMCMC, whose proof can be found in Appendix A.

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 $(\theta_t^p - \theta_*)/\sqrt{\gamma_t}$ and $(\theta_{\kappa t}^s - \theta_*)/\sqrt{\kappa\gamma_{\kappa t}}$*

have the same asymptotic distribution with the convergence rate ratio

$$\frac{\gamma_t}{\kappa \gamma_{\kappa t}} = \kappa^{\beta - 1}, \tag{17}$$

where κ denotes the population size, and β is defined in (A_4). [Note: $1/2 < \beta < 1$ for the case A_4-(ii.1) and $\beta = 1$ for the case A_4-(ii.2).]

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 by Figure 1(a) and Figure 2, 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 provides a more accurate estimate of $h(\theta_t)$ than the single-chain SAMCMC, 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. This is illustrated by Figure 1(b).
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 for 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,

$$f(x) = c\psi(x), \quad x \in \mathcal{X}, \tag{18}$$

where $\mathcal{X}$ 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 (Metropolis *et al.*, 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984), are prone to get trapped into local modes in simulations from such a 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 (Geyer, 1991), simulated tempering (Marinari and Parisi, 1992), dynamic weighting (Wong and Liang, 1997), Wang-Landau algorithm (Wang and Landau, 2001), SAMC algorithm (Liang *et al.*, 2007), among others. The SAMC algorithm can be described as follows.

Let $E_1, ..., E_m$ 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 following subregions: $E_1 = \{x : U(x) \leq u_1\}$, $E_2 = \{x : u_1 < U(x) \leq u_2\}$, $\ldots$, $E_{m-1} = \{x : u_{m-2} < U(x) \leq u_{m-1}\}$ and $E_m = \{x : U(x) \geq u_m\}$, where $u_1 < u_2 < \ldots < u_{m-1}$ are user-specified numbers. If $\int_{E_i} \psi(x)dx = 0$, then E_i is called an empty subregion. Refer to Liang *et al.* (2007) for more discussions on sample space partitioning. For the time being, we assume that all the subregions are non-empty; that is, $\int_{E_i} \psi(x)dx > 0$ for all $i = 1, \ldots, m$. Given the partition, SAMC seeks to draw samples from the distribution

$$f_w(x) \propto \sum_{i=1}^{m} \frac{\pi_i \psi(x)}{w_i} I(x \in E_i) \tag{19}$$

where $w_i = \int_{E_i} \psi(x)dx$, and π_i's define the desired sampling frequency for each of the subregions and they satisfy the constraints: $\pi_i > 0$ for all i and $\sum_{i=1}^{m} \pi_i = 1$. If $w_1, ..., w_m$ 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 $w_1, \ldots, w_m$ are generally unknown, SAMC employs the stochastic approximation algorithm to estimate their values. This leads to the following iterative procedure:
The SAMC algorithm

1. (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:

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{e^{\theta_{t,i}}} I(x \in E_i), \tag{20}$$

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

2. (Weight updating) Set

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, x_{t+1}), \tag{21}$$

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

A remarkable feature of SAMC is that it possesses the self-adjusting mechanism, which operates based on the past samples. This mechanism penalizes the over-visited subregions and rewards the under-visited subregions, and thus enables 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 to be visited at the next iteration. On the other hand, for those 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 to visit these regions will increase at the next iteration. SAMC has been successfully applied to many different problems for which the energy landscape is rugged, such as phylogeny inference (Cheon and Liang, 2009) and Bayesian network learning (Liang and Zhang, 2009).

3.2. The Population SAMC Algorithm

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

The Pop-SAMC algorithm:

1. (Population sampling) For $i = 1, \ldots, \kappa$, simulate a sample $x_{t+1}^{(i)}$ by running, for one step, the Metropolis-Hasting algorithm which starts with $x_t^{(i)}$ and admits (20) as the invariant distribution. Denote the population of samples by $\boldsymbol{x}_{t+1} = (x_{t+1}^{(1)}, \ldots, x_{t+1}^{(\kappa)})$.

2. (Weight updating) Set

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}), \tag{22}$$

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, Pop-SAMC can converge asymptotically faster than SAMC. 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 $f_{\theta_t}(x^{(1)}) \ldots f_{\theta_t}(x^{(\kappa)})$ 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 SAMC, if multiple MH updates were allowed at each iteration of SAMC.

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 infinity. Therefore, without loss of generality, we show in the supplementary material (Song et al., 2013) only the convergence of the algorithm for the case that all subregions are non-empty; that is, $\int_{E_i} \psi(x)dx > 0$ for all $i = 1, \ldots, m$. Extending the proof to the general case is trivial, since replacing (22) by (23) (given below) will not change the process of Pop-SAMC simulation:

$$\theta'_{t+1} = \theta_t + \gamma_{t+1}(\boldsymbol{H}(\theta_t, \boldsymbol{x}_{t+1}) - \boldsymbol{\nu}), \tag{23}$$

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

In our proof, we assume that Θ is a compact set. As aforementioned for the general SAMCMC algorithms, this assumption is made only for the reason of mathematical simplicity. Extension of our results to the case that $\Theta = \mathbb{R}^m$ is trivial with the technique of varying truncations (Chen, 2002; Andrieu *et al.*, 2005; Liang, 2010). Interested readers can refer to Liang (2010) for the details, where the convergence of SAMC is studied with $\Theta = \mathbb{R}^m$. In the simulations of this paper, we set $\Theta = [-10^{100}, 10^{100}]^m$, as a practical matter, this is equivalent to setting $\Theta = \mathbb{R}^m$.

Under the above assumptions, we have the following theorem concerning the convergence of the Pop-SAMC algorithm, whose proof can be found in the supplementary material (Song et al., 2013).

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 (A_4). If Θ is compact, all subregions are nonempty, and each of the transition kernels satisfies (A_3)-(i), then, as $t \to \infty$,*

$$\theta_t \to \theta_*, \quad a.s., \tag{24}$$

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

$$\theta_*^{(i)} = C + \log\left(\int_{E_i} \psi(x)dx\right) - \log(\pi_i), \quad i = 1, \ldots, m, \tag{25}$$

with C being a constant.

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

Remark As aforementioned, 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 material (Song et al., 2013), we have

$$\theta_*^{(i)} = \begin{cases} C + \log\left(\int_{E_i} \psi(x)dx\right) - \log(\pi_i + \nu), & \text{if } E_i \neq \emptyset, \\ -\infty, & \text{if } E_i = \emptyset. \end{cases} \tag{26}$$

where C is a constant, $\nu = \sum_{j \in \{i: E_i = \emptyset\}} \pi_j/(m - m_0)$, and m_0 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. To have this condition 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 $\{x : \psi(x) \geq \psi_{min}\}$, where ψ_{min} is sufficiently small such that the region $\{x : \psi(x) < \psi_{min}\}$ is not of interest. For the proposal distribution used in the paper, we assume that it satisfies the local positive condition; that is, there exists two quantities $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $q(x, y) \geq \epsilon_2$ if $|x - y| \leq \epsilon_1$, where $q(x, y)$ denotes the proposal mass/density function. In the supplementary material (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 concerns the asymptotic normality of θ_t, whose proof can be found in the supplementary material (Song et al., 2013).

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

$$\frac{\theta_t - \theta_*}{\sqrt{\gamma_t}} \Longrightarrow \mathbb{N}(0, \Sigma), \tag{27}$$

where θ_* *is as defined in (25), and*

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

with F *being defined in* (A_2), ζ *defined in (11), and* Γ *defined in Lemma 1.*

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 $\gamma_t = O(1/t)$, Pop-SAMC still has some numerical advantages over SAMC in convergence due to the population effect. This will be illustrated by Figure 2.

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, into 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

$$\boldsymbol{P}_{\theta_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = P_{\theta_t\times\theta_t}\{(x_t^{(1)}, x_t^{(2)}), (x_{t+1}^{(1)}, x_{t+1}^{(2)})\}\prod_{i=3}^{\kappa} P_{\theta_t}(x_t^{(i)}, x_{t+1}^{(i)}), \tag{28}$$

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

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

where r_{co} is the probability to apply crossover kernel $P_{\theta_t,co}$. Following the proof in the supplementary material (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 $r_{co} < 1$. As long as $\mathcal{X}$ is compact, and $f(x)$ is bounded away from 0 and ∞, (A_3)-(i) is satisfied by $P_{\theta_t,co}$. The condition A_3-(ii) is satisfied because it is independent of the kernel used. The condition (A_3)-(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 $\boldsymbol{x} = (x^{(1)}, x^{(2)})$ and $\boldsymbol{y} = (y^{(1)}, y^{(2)})$, 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

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

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

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

Following the same argument as in Liang *et al.* (2007), (A_3)-(iii) is satisfied by $P_{\theta_t,co}$. This concludes that each kernel in the right of (28) satisfies the drift condition (A_3).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 3.1 and Theorem 3.2) still hold for this general Pop-SAMC algorithm with crossover operators. We conjecture that the incorporation of crossover operators will bring Pop-SAMC more efficiency. How these advanced operators improve the performance of Pop-SAMC will be explored elsewhere.

4. An Illustrative Example

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

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2}\sum_{i=1}^{20}\alpha_i \exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big\}, \tag{29}$$

where each component has an equal variance $\sigma^2 = 0.01$ and an equal weight $\alpha_1 = ... = \alpha_{20} = 0.05$, and the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{20}$ 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 puts a great challenge on the existing MCMC algorithms.

We set the sample space $\mathcal{X} = [-10^{100}, 10^{100}]^2$, and then partitioned it according to the energy function $U(x) = -\log\{p(x)\}$ with an equal energy bandwidth $\Delta u = 0.5$ into the following subregions: $E_1 = \{x : U(x) \le 0\}, E_2 = \{x : 0 < U(x) \le 0.5\}, ..., E_{20} = \{x : U(x) > 9.0\}$. Pop-SAMC 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., $\pi_1 = \cdots = \pi_{20} = 1/20$. 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. To have a fair comparison with SAMC, we initialize the population in a small region $[0, 1] \times [0, 1]$, which is far from the separated components. Tables 1 and 2 show the resulting estimates of $P(E_i)$ (i.e. $w_i = \int_{E_i} p(x)dx$), for $i = 2, \ldots, 11$, based on 100 independent runs. The computation was done on on a Intel Core 2 Duo 3.0 GHz computer. As shown by the true values of $P(E_i)$'s, which are calculated with a total of 2×10^9 samples drawn equally from each of the 20 components of $p(x)$, the subregions $E_2, \ldots, E_{11}$ have covered more than 99% of the total mass of the distribution. For comparison, SAMC 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 Pop-SAMC and SAMC consists of the same number of energy evaluations and thus costs about the same CPU

TABLE 1: Comparison of efficiency of Pop-SAMC and SAMC for the multimodal example with $\gamma_t = t_0/\max\{t_0, t\}$. The number in the parentheses shows the standard error of the estimate of $P(E_i)$.

Setting	True	Pop-SAMC	SAMC	SAMC
		$(t_0, \tau, N) =$ $(100, 10, 10^6)$	$(t_0, N) =$ $(100, 10^7)$	$(t_0, N) =$ $(1000, 10^7)$
$P(E_2)$	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

time. The resulting estimates of $P(E_i)$'s are summarized 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 estimates of $P(E_i)$ should follow the same limiting rule (17) as θ_t does. For this example, when the same gain factor sequence $\gamma_t = 100/\max(100, t)$ is used, SAMC is as efficient as Pop-SAMC 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 SAMC and Pop-SAMC 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, Pop-SAMC is more efficient than SAMC. Table 2 shows

Table 2: Comparison of efficiency of Pop-SAMC and SAMC for the multimodal example with $\gamma_t = t_0/\max\{t_0, t^{0.6}\}$. The number in the parentheses shows the standard error of the estimate of $P(E_i)$.

Setting	True	Pop-SAMC	SAMC	SAMC
		$(t_0, \tau, N) =$	$(t_0, N) =$	$(t_0, N) =$
		$(100, 10, 10^6)$	$(100, 10^7)$	$(1000, 10^7)$
$P(E_2)$	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

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 that 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 $10^4 \sim 10^6$ for Pop-SAMC and $10^5 \sim 10^7$ for SAMC. The results are shown in Figure 1. The plots indicate that Pop-SAMC can converge much faster than SAMC, even when the gain factor sequence $\gamma_t = t_0/\max(t_0, t)$ is used. As discussed previously, this is due to the population effect: Pop-SAMC 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 Pop-SAMC, both Pop-SAMC and SAMC were re-run 100 times with a smaller gain factor sequence $\gamma_t = 50/\max(50, t)$. Figure 2 shows that under this setting, SAMC converges very slowly, while Pop-SAMC still converges very fast. This experiment shows that Pop-SAMC is more robust to the choice of gain factor sequence, and it can work with a smaller gain factor sequence than can SAMC.

5. Conclusion

In this paper, we have proposed a population SAMCMC algorithm and contrasted its convergence rate with that of the single-chain SAMCMC algorithm. As the main theoretical result, we establish the limiting ratio between the L_2 rates of convergence of the two types of SAMCMC algorithms. Our result provides 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 has been confirmed with a numerical example.

In this paper, we have also proved the asymptotic normality of SAMCMC estimators under mild conditions. As mentioned previously, the major difference between this work and Benveniste *et al.* (1990) are the assumptions on Markov transition kernels. Our assumptions are easier to verify than those by Benveniste *et al.* (1990). We note that the work by Chen (2002) and Pelletier (1998) can potentially be extended to SAMCMC algorithms. The major differences between their work and ours are the assumptions on 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

$$\epsilon_t = e_t + \nu_t,$$

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 the 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 π_θ. In addition, $H : \Theta \times \mathbb{X} \to \Theta$ is measurable and for all $\theta \in \Theta$, $\int_{\mathbb{X}} \|H(\theta, x)\| \pi_\theta(x) dx < \infty$.*

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

$$\begin{aligned} (i) \quad & \sup_{\theta\in\Theta} (\|u_\theta(\cdot)\| + \|P_\theta u_\theta(\cdot)\|) < \infty, \\ (ii) \quad & \sup_{(\theta,\theta')\in\Theta\times\Theta} \|\theta - \theta'\|^{-\eta} \{\|u_\theta(\cdot) - u_{\theta'}(\cdot)\| + \|P_\theta u_\theta(\cdot) - P_{\theta'} u_{\theta'}(\cdot)\|\} < \infty. \end{aligned} \tag{30}$$

(B_3) *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 stochastic approximation MCMC algorithm under different conditions from those given in Andrieu, Moulines and Priouret (2005) and Andrieu and Moulines (2006). We combined some results of the three papers and got the following lemma, which corresponds to Theorem 4.1 and Lemma 2.2 of Tadić (1997).

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

(C_1) *There exist $\mathbb{R}^{d_\theta}$-valued random processes $\{\epsilon_t\}_{t\geq 0}$, $\{\epsilon_t^{'}\}_{t\geq 0}$ and $\{\epsilon_t^{''}\}_{t\geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that*

$$\gamma_{t+1}\xi_{t+1} = \epsilon_{t+1} + \epsilon_{t+1}^{'} + \epsilon_{t+1}^{''} - \epsilon_t^{''}, \quad t \geq 0, \tag{31}$$

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

(C_2) *The series $\sum_{t=0}^\infty \|\epsilon_t^{'}\|$, $\sum_{t=0}^\infty \|\epsilon_t^{''}\|^2$ and $\sum_{t=0}^\infty \|\epsilon_{t+1}\|^2$ all converge a.s. and*

$$E(\epsilon_{t+1}|\mathcal{F}_t) = 0, \quad a.s., \quad n \geq 0, \tag{32}$$

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* $\sigma\{\epsilon_t, \epsilon_t', \epsilon_t^{''}\} \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, $t \geq 0$.

(C_3) *Let* $R_t = R_t' + R_t^{''}$, $t \geq 1$, *where* $R_t' = \gamma_{t+1}\nabla^T v(\theta_t)\xi_{t+1}$, *and*

$$R_{t+1}^{''} = \int_0^1 \big[\nabla v(\theta_t + s(\theta_{t+1} - \theta_t)) - \nabla v(\theta_t)\big]^T (\theta_{t+1} - \theta_t) ds.$$

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

Proof. (C_1) Let $\epsilon_0 = \epsilon_0' = 0$, and

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

It is easy to verify that (31) is satisfied.

(C_2) Since

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

which concludes (32). It follows from (B_2), (A_3) and (A_4) that there exist constants $c_3, c_4, c_5, c_6, c_7 \in \mathbb{R}^+$ such that

$$\begin{aligned}
&\|\epsilon_{t+1}\|^2 \leq 2c_3\gamma_{t+1}^2, \quad \|\epsilon_{t+1}^{''}\|^2 \leq c_4\gamma_{t+1}^2, \\
&\|\epsilon_{t+1}'\| \leq c_5\gamma_{t+1}\|\theta_{t+1} - \theta_t\|^\eta + c_6\gamma_{t+1}^{1+\tau} \leq c_7\gamma_{t+1}^{1+\eta},
\end{aligned}$$

for any $\eta \in (0,1)$. Following from (9) and setting $\eta \geq \tau'$ (τ' is defined in A_4), we have

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

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

(C_3) Let $M = \sup_{\theta\in\Theta} \max\{\|h(\theta)\|, \|\nabla v(\theta)\|\}$, and L is the Lipschitz constant of $\nabla v(\cdot)$. Since $\sigma\{\theta_t\} \subset \mathcal{F}_t$, it follows from (C_2) that $E(\nabla^T v(\theta_t)\epsilon_{t+1}|\mathcal{F}_t) = 0$. In addition, we have

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

It follows from the martingale convergence theorem (Hall and Heyde, 1980; Theorem 2.15) that both $\sum_{t=0}^{\infty} \epsilon_{t+1}$ and $\sum_{t=0}^{\infty} \nabla^T v(\theta_t)\epsilon_{t+1}$ converge almost surely. Since

$$\sum_{t=0}^{\infty} |\nabla^T v(\theta_t)\epsilon'_{t+1}| \le M \sum_{t=1}^{\infty} \|\epsilon'_t\|,$$
$$\sum_{t=1}^{\infty} \gamma_t^2 \|\xi_t\|^2 \le C \left(\sum_{t=1}^{\infty} \|\epsilon_t\|^2 + \sum_{t=1}^{\infty} \|\epsilon'_t\|^2 + \sum_{t=0}^{\infty} \|\epsilon''_t\|^2 \right),$$

for some constant C. It follows from (C_2) that both $\sum_{t=0}^{\infty} |\nabla^T v(\theta_t)\epsilon'_{t+1}|$ and $\sum_{t=1}^{\infty} \gamma_t^2 \|\xi_t\|^2$ converge. In addition,

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

for all $t \ge 0$. Consequently,

$$\begin{aligned} \sum_{t=1}^{\infty} |R''_t| &\le 2LM^2 \sum_{t=1}^{\infty} \gamma_t^2 + 2L \sum_{t=1}^{\infty} \gamma_t^2 \|\xi_t\|^2 < \infty, \\ \sum_{t=0}^{\infty} \left| (\nabla v(\theta_{t+1}) - \nabla v(\theta_t))^T \epsilon''_{t+1} \right| &\le \left(2L^2M^2 \sum_{t=1}^{\infty} \gamma_t^2 + 2L^2 \sum_{t=1}^{\infty} \gamma_t^2 \|\xi_t\|^2 \right)^{1/2} \\ &\quad \times \left(\sum_{t=1}^{\infty} \|\epsilon''_t\|^2 \right)^{1/2} < \infty. \end{aligned}$$

Since

$$\begin{aligned} &\sum_{t=1}^{n} \gamma_t \xi_t = \sum_{t=1}^{n} \epsilon_t + \sum_{t=1}^{n} \epsilon'_t + \epsilon''_n - \epsilon''_0, \\ &\sum_{t=0}^{n} R'_{t+1} = \sum_{t=0}^{n} \nabla^T v(\theta_t)\epsilon_{t+1} + \sum_{t=0}^{n} \nabla^T v(\theta_t)\epsilon'_{t+1} - \sum_{t=0}^{n} (\nabla v(\theta_{t+1}) - \nabla v(\theta_t))^T \epsilon''_{t+1} \\ &\qquad + \nabla^T v(\theta_{n+1})\epsilon''_{n+1} - \nabla^T v(\theta_0)\epsilon''_0, \end{aligned}$$

and ϵ''_n convergent to zero by (C_2), it is obvious that $\sum_{t=1}^{\infty} \gamma_t \xi_t$ and $\sum_{t=1}^{\infty} R_t$ converge almost surely.

The proof for Lemma 3 is completed.

Based on Lemma 3, Theorem 1 can be proved in a similar way to Theorem 2.2 of Tadić (1997). Since Tadić (1997) is not available publicly, we reproduce the proof for Theorem 1 in Supplemental Materials.

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

B.1. Proof of Lemma 1.

Lemma 4 is a restatement of Proposition 6.1 of Andrieu *et al.* (2005). It has a little overlap with (B_2).

Lemma 4. *Assume A_3-(i) and A_3-(iii) hold. Suppose that the family of functions $\{g_\theta, \theta \in \Theta\}$ satisfies the condition: For any compact subset $\mathcal{K} \subset \Theta$,*

$$\sup_{\theta \in \mathcal{K}} \|g_\theta(\cdot)\| < \infty, \qquad \sup_{(\theta,\theta') \in \mathcal{K} \times \mathcal{K}} |\theta - \theta'|^{-\iota} \|g_\theta(\cdot) - g_{\theta'}(\cdot)\| < \infty, \tag{33}$$

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 $\pi_\theta(g_\theta(x)) = \int_{\mathbb{X}} g_\theta(x)\pi_\theta(x)dx$. Then, for any compact set $\mathcal{K}$ and 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 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 the conditions (A_3) and (A_4)-(i) hold. Let $\{g_\theta, \theta \in \Theta\}$ be a family of functions satisfying (33) with $\iota \in ((1+\tau')/2, 1)$, where τ' is defined in condition A_4. Then*

$$n^{-1} \sum_{k=1}^{n} \left(g_{\theta_k}(X_k) - \int_{\mathbb{X}} g_{\theta_k}(x) d\pi_{\theta_k}(x) \right) \to 0, \qquad a.s.$$

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 vales on $\mathbb{R}^d$, the proof can be done elementwisely.) Let $S_n = \sum_{k=1}^n [g_{\theta_k}(X_k) - \pi_{\theta_k}(g_{\theta_k}(X_k))]$, where $\pi_{\theta_k}(g_{\theta_k}(X_k)) = \int_{\mathbb{X}} g_{\theta_k}(x))\pi_{\theta_k}(x)dx$. 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

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

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

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

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 $(1+\tau')/2 \le \iota' < \iota$ (where τ' is defined in (A_4)). Thus, there exists a constant c such that

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

Since

$$\begin{aligned} E\left[u_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) | \mathcal{F}_{k-1} \right] &= P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) - P_{\theta_{k-1}} u_{\theta_{k-1}}(X_{k-1}) \\ &= 0, \end{aligned}$$

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

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

Now we consider $S_n^{(2)}$. By Lemma 4, the fact $\|\theta_k - \theta_{k-1}\| = \gamma_k \|H(\theta_{k-1}, X_k)\|$ and (A_3)-(ii),

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

Hence, $E(|S_n^{(2)}|^p) \le c'^p (\sum_{k=1}^{n} \gamma_k^{\iota'})^p$. Also, the third term is also bounded by some constant c, $E(|S_n^{(3)}|^p) < c$.

Hence, by Minkowski's inequality, Markov's inequality,we can conclude

$$P\{n^{-1}|S_n| \geq \delta\} \leq C\delta^{-p} \left\{ n^{-p/2} + (n^{-1}\sum_{k=1}^{n} \gamma_k^{\iota'})^p + n^{-p} \right\}, \tag{34}$$

where C denotes a constant. By (34) and the Borel-Cantelli lemma, we have

$$P\{\sup_{n\geq 1} n^{-1}|S_n| \geq \delta\} \leq C\delta^{-p} \sum_{n\geq 1} \left\{ n^{-p/2} + n^{-p/2}(n^{-1/2}\sum_{k=1}^{n} \gamma_k^{\iota'})^p + n^{-p} \right\}.$$

Then, the SLLN is concluded with Kronecker's lemma, condition (8) and the condition $p > 2$.

Proof of Lemma 1

(i) Define

$$\begin{aligned}
e_{k+1} &= u_{\theta_k}(X_{k+1}) - P_{\theta_k} u_{\theta_k}(X_k),\\
\nu_{k+1} &= \left[P_{\theta_{k+1}} u_{\theta_{k+1}}(X_{k+1}) - P_{\theta_k} u_{\theta_k}(X_{k+1})\right] + \frac{\gamma_{k+2}-\gamma_{k+1}}{\gamma_{k+1}} P_{\theta_{k+1}} u_{\theta_{k+1}}(X_{k+1}),\\
\tilde{\varsigma}_{k+1} &= \gamma_{k+1} P_{\theta_k} u_{\theta_k}(X_k),\\
\varsigma_{k+1} &= \frac{1}{\gamma_{k+1}}(\tilde{\varsigma}_{k+1} - \tilde{\varsigma}_{k+2}),
\end{aligned} \tag{35}$$

where $u_{\cdot}(\cdot)$ is the solution of the Poisson equation (Refer to 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 (35), we have

$$E(e_{k+1}|\mathcal{F}_k) = E(u_{\theta_k}(X_{k+1})|\mathcal{F}_k) - P_{\theta_k} u_{\theta_k}(X_k) = 0, \tag{36}$$

Hence, $\{e_k\}$ forms a martingale difference sequence. Following from Lemma 2-(B_2), we have

$$\sup_{k\geq 0} E(\|e_{k+1}\|^{\alpha}|\mathcal{F}_k) 1_{\{\|\theta_k-\theta_*\|\leq\rho\}} < \infty. \tag{37}$$

This concludes part (ii).

(iii) By (35), we have

$$\begin{aligned}
E(e_{k+1}e_{k+1}^T|\mathcal{F}_k) &= E\left[u_{\theta_k}(X_{k+1}) u_{\theta_k}(X_{k+1})^T|\mathcal{F}_k\right] - P_{\theta_k} u_{\theta_k}(X_k) P_{\theta_k} u_{\theta_k}(X_k)^T\\
&\triangleq l(\theta_k, X_k).
\end{aligned} \tag{38}$$

By (B_2) and (A_3)-(i), there exist constants c_1, c_2, c_3 and M such that

$$\|l(\theta_k, X_k)\| \leq E(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^T\||\mathcal{F}_k) + \|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^T\| < c_1$$

For any θ_k, $\theta'_k \in \Theta$,

$$\begin{aligned}
&\|l(\theta_k, X_k) - l(\theta'_k, X_k)\| && (39)\\
\leq &E(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^T - u_{\theta'_k}(X_{k+1})u_{\theta'_k}(X_{k+1})^T\||\mathcal{F}_k)\\
&+ \|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^T - P_{\theta'_k}u_{\theta'_k}(X_k)P_{\theta'_k}u_{\theta'_k}, X_k)^T\|. && (40)
\end{aligned}$$

By lemma 2 (B_2)-(ii), we have, for any $\eta \in (0, 1)$

$$\begin{aligned}
&\|P_{\theta_k}u_{\theta_k}(X_k)P_{\theta_k}u_{\theta_k}(X_k)^T - P_{\theta'_k}u_{\theta'_k}(X_k)P_{\theta'_k}u_{\theta'_k}(X_k)^T\|\\
\leq &\|(P_{\theta_k}u_{\theta_k}(X_k) - P_{\theta'_k}u_{\theta'_k}(X_k))P_{\theta_k}u_{\theta_k}(X_k)^T\|\\
&+ \|P_{\theta'_k}u_{\theta'_k}(X_k)(P_{\theta_k}u_{\theta_k}(X_k)^T - P_{\theta'_k}u_{\theta'_k}(X_k)^T)\|\\
\leq &c_2\|\theta_k - \theta'_k\|^\eta,
\end{aligned}$$

and

$$\begin{aligned}
&E(\|u_{\theta_k}(X_{k+1})u_{\theta_k}(X_{k+1})^T - u_{\theta'_k}(X_{k+1})u_{\theta'_k}(X_{k+1})^T\||\mathcal{F}_k)\\
\leq &E(\|(u_{\theta_k}(X_{k+1}) - u_{\theta'_k}(X_{k+1}))u_{\theta_k}(X_{k+1})^T\||\mathcal{F}_k)\\
&+ E(\|u_{\theta'_k}(X_{k+1})(u_{\theta_k}(X_{k+1})^T - u_{\theta'_k}(X_{k+1})^T)\||\mathcal{F}_k)\\
\leq &c_3\|\theta_k - \theta'_k\|^\eta.
\end{aligned}$$

Plug into equation (40), we have $\|l(\theta_k, X_k) - l(\theta'_k, X_k)\| \leq M\|\theta_k - \theta'_k\|^\eta$ 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

$$\frac{1}{n}\sum_{k=1}^{n}[l(\theta_k, X_k) - \pi_{\theta_k}(l(\theta_k, X))] \to 0, \quad a.s. \tag{41}$$

where $\pi_{\theta_k}(l(\theta_k, X)) = \int_{\mathbb{X}} l(\theta_k, x)\pi_{\theta_k}(x)dx$.

On the other hand, we have

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

Given $\theta_k \to \theta_*$ a.s., the first term goes to 0 almost surely as $k \to \infty$. By condition (A_3), which implies the conditions of proposition 1.3.6 of Atchadé *et al.* (2011) holds, therefore $\pi_{\theta_k}(l(\theta_*, X)) - \pi_{\theta_*}(l(\theta_*, X)) \to 0$ almost surely. Thus, $\| \int_{\mathbb{X}} l(\theta_k, x) d\pi_{\theta_k}(x) - \int_{\mathbb{X}} l(\theta_*, x) d\pi_{\theta_*}(x) \| \to 0$ almost surely and

$$\frac{1}{n} \sum_{k=1}^{n} l(\theta_k, X_k) \to \int_{\mathbb{X}} l(\theta_*, x) d\pi_{\theta_*}(x) = \Gamma, \quad a.s. \tag{42}$$

for some positive definite matrix Γ. This concludes part (iii).

(iv) By condition (A_4), 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 (35) and (30), there exists a constant c_1 such that the following inequality holds,

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

which implies, by (5), that there exists a constant c_2 such that

$$\|\nu_{k+1}\| \leq c_2 \gamma_{k+1}. \tag{43}$$

therefore,

$$E(\|\nu_k\|^2/\gamma_k) 1_{\{\|\theta_k - \theta_*\| \leq \rho\}} \to 0.$$

This concludes 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 (B_2), $E\left[\|P_{\theta_k} u_{\theta_k}(X_k)\|\right]$ 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 a combined restatement of Theorem D.6.4 (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 $\{e_k\}$ satisfies (ii)-(iv) given in Lemma 1, and the conditions (A_2) and (A_4) are satisfied. On the set $\Lambda(z^*) = \{Z_k \to z^*\}$,

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

with $\Longrightarrow$ denoting the weak convergence, $\mathbb{N}$ the Gaussian distribution and

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

where F is defined in (A_2), ζ is defined in (11), and Γ is defined in Lemma 1.

Proof of Theorem 2 Rewrite the SAMCMC algorithm in the form

$$\theta_{k+1} - \theta_* = (\theta_k - \theta_*) + \gamma_{k+1} h(\theta_k) + \gamma_{k+1}\xi_{k+1}. \tag{44}$$

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

$$\tilde{\theta}_k = \theta_k + \tilde{\varsigma}_{k+1}, \tag{45}$$

where $\tilde{\varsigma}_{k+1}$ is as defined in equation (35) in the proof of Lemma 1. Then, for the SAMCMC algorithm, we have

$$\begin{aligned}
\tilde{\theta}_{k+1} - \theta_* &= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\theta_k) + \gamma_{k+1}\xi_{k+1} + \tilde{\varsigma}_{k+2} - \tilde{\varsigma}_{k+1} \\
&= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(h(\theta_k) - h(\tilde{\theta}_k) + \xi_{k+1} - \varsigma_{k+1}) \\
&= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(h(\theta_k) - h(\tilde{\theta}_k) + \upsilon_{k+1} + e_{k+1}) \\
&= (\tilde{\theta}_k - \theta_*) + \gamma_{k+1} h(\tilde{\theta}_k) + \gamma_{k+1}(\tilde{\nu}_{k+1} + e_{k+1}),
\end{aligned} \tag{46}$$

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 equation (35) in the proof of Lemma 1 as well. Since $h(\cdot)$ is Hölder continuous on Θ (by the result B_3 of Lemma 2) and Θ is compact, there exists a constant M such that $\|h(\theta_k) - h(\tilde{\theta}_k)\| \leq M\|\tilde{\theta}_k - \theta_k\|^\eta = M\|\tilde{\varsigma}_{k+1}\|^\eta$ for any $\eta \in (0.5, 1)$. Thus, by (35), there exists a constant c such that

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

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 the property (iv) of Lemma 1.

By Lemma 1 and Lemma 6, we have

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

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

$$\frac{\tilde{\varsigma}_{k+1}}{\sqrt{\gamma_k}} \to 0, \qquad \text{in probability.} \tag{47}$$

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

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

which concludes Theorem 2.

B.3. Proof of Theorem 3

Proof of Theorem 3 Let $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(\kappa)})$ denote the samples drawn at an iteration of population SAMCMC. 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 $\boldsymbol{u} = \sum_{n \geq 0}(\boldsymbol{P}^n \boldsymbol{H} - h)$ be a solution of Poisson equation $\boldsymbol{u} - \boldsymbol{P}\boldsymbol{u} = \boldsymbol{H} - h$, and let $u = \sum_{n \geq 0}(P^n H - h)$ 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 $\boldsymbol{u}_\theta(\boldsymbol{x}) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} u_\theta(x^{(i)})$. By (35), we further have

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

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

$$\boldsymbol{\Gamma} = \Gamma / \kappa,$$

which, by Theorem 2, further implies

$$\Sigma_p = \Sigma_s / \kappa,$$

where Σ_p and Σ_s denote the limiting covariance matrices of 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 $N(0, \Sigma_p)$. By condition (A_4), $\gamma_t/(\kappa\gamma_{\kappa t}) = \kappa^{\beta-1}$, which concludes the proof.

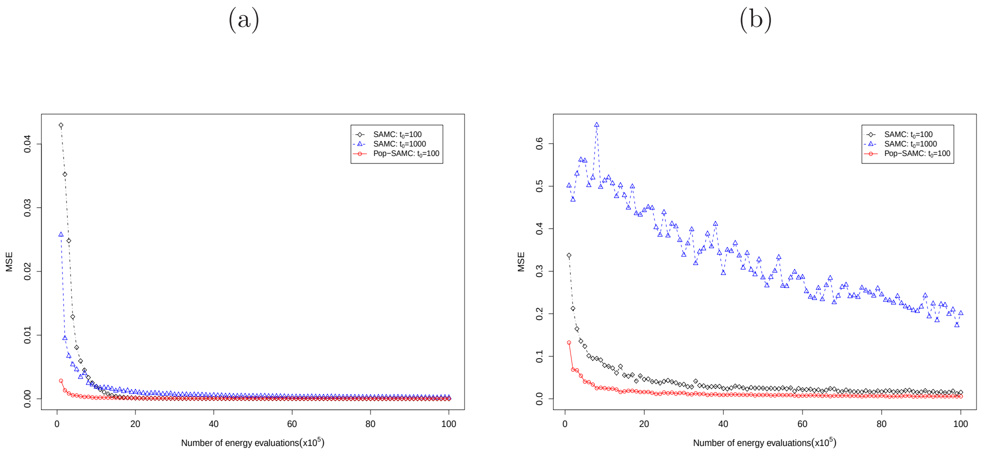

FIGURE 1: Mean square errors (MSEs) produced by Pop-SAMC and SAMC at different iterations. The upper plot is produced with $\gamma_t = t_0/\max(t_0, t)$, and the lower plot is produced with $\gamma_t = t_0/\max(t_0, t^{0.6})$.

Acknowledgements

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

References

ALDOUS, D., LOVÁSZ, L., AND WINKLER, P. (1997). Mixing times for uniformly ergodic Markov chains. *Stoch. Proc. 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.

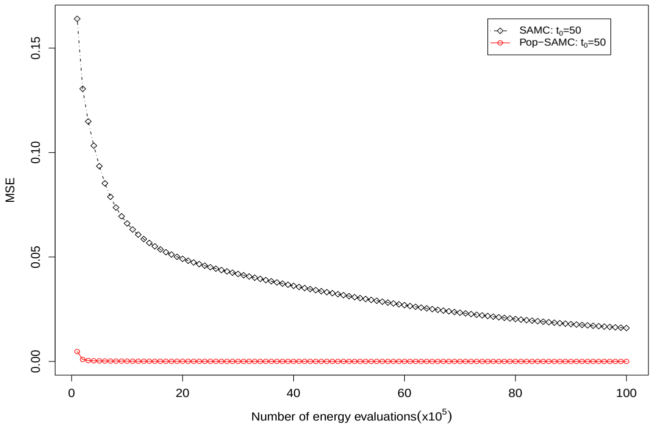

FIGURE 2: Mean square errors (MSEs) produced by Pop-SAMC and SAMC at different iterations with the gain factor sequence $\gamma_t = 50/\max(50, t)$.

ANDRIEU, C., MOULINES, É, AND PRIOURET, P. (2005). Stability of Stochastic Approximation Under Verifiable Conditions. *SIAM J. Control Optim.*, **44**, 283-312.

ATCHADÉ, Y. AND FORT, G. (2009). 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, Oxford, UK.

BENVENISTE, A., MÉTIVIER, M., AND PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* New York: Springer-Verlag.

BILLINGSLEY, P. (1986). *Probability and Measure* (2nd edition). New York: John Wiley & Sons.

BLUM, J.R. (1954). Approximation Methods which Converge with Probability one. *Ann. Math. Statist.***25**, 382-386.

CASELLA, G. AND BERGER, R.L. (2002). *Statistical Inference* (second edition). Duxbury Thomson Learning.

CHAUVEAU, D. AND DIEBOLT, J. (2000). Stability properties for a product Markov chain. Preprint No 06/2000, Université Marne-la-Vallée.

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

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

Duan, G.-R. and Patton, R.J. (1998). A Note on Hurwitz Stability of Matrices. *Automatica*, **34**, 509-511.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal.*, **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* (ed. E.M. Keramigas), pp.153-163.

Gilks, W.R., Roberts, G.O., and George, E.I. (1994). Adaptive Direction Sampling, *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. Natl. 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, London.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chain 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. Stat.*, **16**, 608-632

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., Liu, C. and Carroll, R. J. (2007) Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Soc.*, **102**, 305-320.

Liang, F., and Wong, W.H. (2000). Evolutionary Monte Carlo: Application to C_p model sampling and change point problem. *Stat. Sinica.*, **10**, 317-342.

Liang, F., and Wong, W.H. (2001). Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. *J. Amer. Statist. Soc.*, **96**, 653-666.

Liang, F. and Zhang, J. (2009). Learning Bayesian Networks for Discrete Data. *Comput. Stat. Data. An.*, **53**, 865-876.

Liu, J.S., Liang, F., and Wong, W.H. (2000). The use of multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Soc.*, **94**, 121-134.

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.

Meyn, S. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability* (second edition). Cambridge University Press.

Nummelin, E. (1984), *General Irreducible Markov Chains and Nonnegative Operators.* Cambridge: 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. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.*, **44**, 458-475.

Roberts, G.O., and Rosenthal, J.S.(2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat.*, **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". *arXiv:submit/0828780* (also available at http://www.stat.tamu.edu/∼fliang).

Tadić, V. (1997). On the convergence of stochastic iterative algorithms and their applications to machine learning. A short version of this paper was published in *Proc. 36th Conf. on Decision & Control* 2281-2286. San Diego, USA.

Younes, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Relat. Field*, **82** 625-645.

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

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.

Ziedan, I.E. (1972). Explicit solution of the Lyapunov-matrix equation. *IEEE Trans. Automat. Contr.*, **17**, 379-381.