Some Results on the Ergodicity of Adaptive MCMC Algorithms

Omar Khalil

Supervisor: Jeffrey Rosenthal

September 2, 2011

Contents

Introduction

Markov Chain Monte Carlo (MCMC) is a computational method for approximately sampling from distributions of interest. The method consists of starting with some measure space $(X, B(X))$, where $X \subset \mathbb{R}^n$ for some n and $B(X)$ is a countably generated sigma field, and then simulating an ergodic Markov chain $\{X_k, k \geq 0\}$ on X , with transition probability P such that π is a stationary distribution for the chain. An important usage of the sampling process is to compute integrals of the form

$$\pi(f) \triangleq \int_X f(x)\pi(dx)$$

where $f : X \to \mathbb{R}$ is a π-integrable function, by using estimators of the type

$$S_n(f) = \frac{1}{n}\sum_{k=1}^{n} f(X_k)$$

which motivates the concern for law of large number results for MCMC. Typically the transition probability P depends on some tuning parameter θ which affects the convergence properties of the algorithm. This tuning process is often done manually done by trial and error, but this can be time consuming and increasingly difficult in higher dimensions. An alternative approach is using adaptive MCMC algorithms which attempt to learn the best parameter values on the fly while they run. However, adaptive MCMC algorithms may not always preserve the stationarity of π, as we will demonstrate in the examples section. This failure at ergodicity can sometimes occur very counter-intuitively. For example, in cases where the tuning parameter is the index of a family of Markov transition kernels for the chain, adaptive MCMC algorithms can ruin ergodicity even if each individual kernel converges to π. The goal of this write up is to summarize the results found in different papers which deal with conditions that ensure ergodicity of such algorithms, along with considering the relationship between them. We will focus mainly on the strong law of large numbers results found in [1],[4] and [9], as well as the ergodicity result found in [2].

A popular MCMC sampling technique is the Metropolis Hastings (MH) algorithm which requires the choice of a proposal distribution q, and for our discussion we will assume the target distribution π and q both admit densities which we will also denote π and q respectively with a slight abuse of notation. The distribution q is used to generate a potential transition for the Markov chain $\{X_k\}$. If the chain is currently at x and the proposed candidate is y, the proposal is accepted with probability $\alpha(x, y)$ given by

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, & \text{if } \pi(x)q(x,y) > 0, \\ 1 & \text{otherwise} \end{cases}$$

otherwise the proposal is rejected and the chain stays at x. It is clear that α is chosen such that the Markov chain is reversible with respect to π. An algorithm proposed by Haario, Saksman and Tammien in [6] deals with the case where the space is $\mathbb{R}^n$ for some n and the proposal distribution q being multivariate normal with zero mean and covariance matrix Γ. Gelman, Roberts and Gilks have shown in [5] that the optimal covariance matrix is $(2.38^2/n)\Gamma_\pi$ where Γ_π is the true covariance matrix of the proposal distribution. They propose to learn Γ_π on the fly with an algorithm which can be summarized by

$$\begin{aligned} \mu_{k+1} &= \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k), \quad k \geq 0, \\ \Gamma_{k+1} &= \Gamma_k + \gamma_{k+1}\big((X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^T - \Gamma_k\big) \end{aligned}$$

where $\theta_k = (\mu_k, \Gamma_k)$ is the index for a Metropolis Hastings transition kernel with multivariate normal increment distribution having mean μ_k and covariance matrix $\lambda\Gamma_k$, where λ is a positive constant dependent only on the dimension of the space and kept constant throughout the iterations.
It was realized that such a scheme is a particular case of the Robins-Monro stochastic control algorithm which takes the form

$$\theta_{k+1} = \theta_k + \gamma_k H(\theta_k, X_{k+1})$$

where γ_k is a sequence of step sizes. In our case $H(\theta, x)$ is given by

$$H(\theta, x) \triangleq (x - \mu, (x - \mu)(x - \mu)^T - \Gamma)^T \tag{1}$$

The first result we will mention deals with this recursion which is quite important in stochastic approximation algorithms.

1 Andrieu-Moulines

The result we will focus on in this paper is a strong law of large numbers dealing with the case when the index parameter updates according to the Robins-Monro algorithm described above. The assumptions will require geometric drift uniformly on compact sets, and will utilize stochastic approximation to cover the space using a (possibly infinite) union of compact sets embedded in each other, and re-initializing the chain once it leaves the current compact set and wanders off to another one.

Definition of the chain

We will be working with an underlying measure space $(X, B(X))$, where $X \subset \mathbb{R}^{n_x}$, n_x being some integer, and $B(X)$ is a countably generated σ-field. We also introduce:
1. A family Markov transition kernels on X, $\{P_\theta, \theta \in \Theta\}$, indexed by a finite-dimensional parameter θ belonging to some open set $\Theta \subset \mathbb{R}^{n_\theta}$. We assume that for each $\theta \in \Theta$, P_θ is π-irreducible and that $\pi P_\theta = \pi$.
2. A family of update functions $\{H(\theta, x) : \Theta \times X \mapsto \mathbb{R}^{n_\theta}\}$ used to adapt the value of the tuning parameter.
We extend the parameter space with a cemetery point, $\theta_c \notin \Theta$ and define $\bar{\Theta} \triangleq \Theta \cup \{\theta_c\}$. We also introduce the family of transition kernels $\{Q_{\tilde{\gamma}}, \tilde{\gamma} \geq 0\}$ such that for any $\tilde{\gamma} \geq 0$, $(x, \theta) \in X \times \Theta, A \in \mathcal{F}$ and $B \in B(\bar{\Theta})$ (where $B(\bar{\Theta})$ denotes a countably generated σ-field of subsets of $\bar{\Theta}$),

$$\begin{aligned} Q_{\tilde{\gamma}}(x, \theta; A \times B) &= \int_A P_\theta(x, dy) \mathbb{I}\{\theta + \tilde{\gamma} H(\theta, y) \in B\} \\ &+ \delta_{\theta_c}(B) \int_A P_\theta(x, dy) \mathbb{I}\{\theta + \tilde{\gamma} H(\theta, y) \notin \Theta\} \end{aligned}$$

where δ_θ denotes the Dirac delta function at θ. Set $\theta_0 = \theta \in \Theta$, $X_0 = x \in X$ and for $k \geq 0$ define the sequence $\{(X_k, \theta_k), k \geq 0\}$: if $\theta_k = \theta_c$, then set $\theta_{k+1} = \theta_c$ and $X_{k+1} = x$, otherwise $(X_{k+1}, \theta_{k+1}) \sim Q_{\rho_{k+1}}(X_k, \theta_k; \cdot)$, where ρ is a sequence of step sizes. We will use $\tilde{\mathbb{P}}^\rho_{x,\theta}$ and $\tilde{\mathbb{E}}^\rho_{x,\theta}$

to denote respectively the distribution and expectation of the stopped process with step size sequence ρ . This corresponds to the algorithm which updates

$$\theta_{k+1} = \theta_k + \rho_{k+1} H(\theta_k, X_{k+1})$$

The chain described above is the basic stopped form of the algorithm. Due to the interaction between X_k and θ_k the stabilization of this inhomogeneous Markov chain is often difficult to achieve and so we consider the new chain Z defined below. First we need to introduce some notions. We say a family $\{\mathcal{K}_q, q \geq 0\}$ of compact subsets of Θ is a compact coverage of Θ if

$$\bigcup_{q \geq 0} \mathcal{K}_q = \Theta \quad \text{and} \quad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}), \quad q \geq 0$$

where $\text{int}(A)$ denotes the interior of the set A. Let $\gamma \triangleq \{\gamma_k\}$ be a monotone non-increasing sequence of positive numbers and let K be a subset of X. For a sequence $a = \{a_k\}$ and an integer l we define the shifted sequence $a^{\leftarrow l}$ as follows : for any $k \geq 1$, $a_k^{\leftarrow l} \triangleq a_{k+l}$. Let $\Pi : X \times \bar{\Theta} \to K \times \mathcal{K}_0$ be a measurable function. We now define the homogeneous Markov chain $Z_k = \{(X_k, \theta_k, \kappa_k, \nu_k), k \geq 0\}$ on the product space $Z \triangleq X \times \bar{\Theta} \times \mathbb{N} \times \mathbb{N}$, with transition probability $R : Z \times B(Z) \to [0, 1]$ algorithmically defined as follows. For any $(x, \theta, \kappa, \nu) \in Z$:
1. If $\nu = 0$ then draw $(X', \theta') \sim Q_{\gamma_\kappa}(\Pi(x, \theta); \cdot)$; otherwise draw $(X', \theta) \sim Q_{\gamma_{\kappa+\nu}}(x, \theta; \cdot)$.
2. If $\theta' \in \mathcal{K}_\kappa$, then set $\kappa' = \kappa$ and $\nu' = \nu + 1$; otherwise, set $\kappa' = \kappa + 1$ and $\nu' = 0$.

Assumptions

For $W : X \to [1, \infty)$ and $f : X \to \mathbb{R}$ a measurable function, define

$$||f||_W = \sup_{x \in X} \frac{|f(x)|}{W(x)} \quad \text{and} \quad \mathcal{L}_W = \{f : ||f||_W < \infty\}$$

We say a family of functions $\{f_\theta : X \to \mathbb{R}, \theta \in \Theta\}$ is $W-$Lipschitz if, for any compact subset $\mathcal{K} \subset \Theta$,

$$\sup_{\theta \in \mathcal{K}} ||f_\theta||_W < \infty \quad \text{and} \quad \sup_{(\theta, \theta') \in \mathcal{K} \times \mathcal{K}, \theta \neq \theta'} \frac{||f_\theta - f_{\theta'}||_W}{|\theta - \theta'|} < \infty$$

Conditions:
(A1) For any $\theta \in \Theta$), P_θ has π as its stationary distribution. In addition there exists a function $V : X \to [1, \infty)$ such that $\sup_{x \in K} V(x) < \infty$ with K defined earlier and such that, for any compact subset $\mathcal{K} \subset \Theta$:

(i) Minorization condition. There exists $C \in B(X), \epsilon > 0$ and a probability measure φ such that $\varphi(C) > 0$ and for all $A \in B(X)$ and $(x, \theta) \in C \times \mathcal{K}$,

$$P_\theta(x, A) \geq \epsilon \varphi(A)$$

.

(ii) Geometric Drift condition. There exist constants $\lambda \in [0, 1)$, $b \in (0, \infty)$ satisfying

$$P_\theta V(x) \leq \begin{cases} \lambda V(x) & \text{if } x \in C^c, \\ b & \text{if } x \in C \end{cases}$$

for all $\theta \in \mathcal{K}$.

(A2) For any compact subset $\mathcal{K} \subset \Theta$ and any $r \in [0,1]$, there exists a constant C such that for any $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$ and $f \in \mathcal{L}_{V^r}$,

$$||P_\theta f - P_{\theta'} f||_{V^r} \leq C||f||_{V^r}|\theta - \theta'|$$

where V is given in (A1).
(A3) $\{H_\theta : \theta \in \Theta\}$ is V^β-Lipschitz for some $\beta \in [0, 1/2]$ with V defined in (A1).

Result

Theorem 1. *Let $\{\mathcal{K}_q, q \geq 0\}$ be a compact coverage of Θ and let $\gamma = \{\gamma_k\}$ be a nonincreaasing positive sequence such that $\sum_{k=1}^{\infty} k^{-1}\gamma_k < \infty$. Consider the time homogeneous Markov chain $\{Z_k\}$ on Z with transition probability R defined earlier. Assume (A1) - (A3) and let $f : X \to \mathbb{R}$ be a function such that $||f||_{V^\alpha} < \infty$ for some $\alpha \in [0, 1-\beta)$, with β as in (A3) and V as in (A1). Assume, in addition that $\bar{\mathbb{P}}_*\{\lim_{n\to\infty} \kappa_n < \infty\} = 1$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} [f(X_k) - \pi(f)] = 0$$

almost surely - $\bar{\mathbb{P}}_*$ where $\bar{\mathbb{P}}_*$ corresponds to the Markov chain started at $(x, \theta, 0, 0)$.

Example 1. We return to the algorithm of Haario et. al, and state a result due to theorem 1 above, for when the target distribution is super-exponential in the tails, and the step size sequence γ satisfies certain conditions. More precisely we will assume:

(M)

(i) π is bounded away from zero on every compact set and continuously differentiable.

(ii)

$$\lim_{|x|\to\infty} \left\langle \frac{x}{|x|}, \nabla \log \pi(x) \right\rangle = -\infty$$

(iii)

$$\lim_{|x|\to\infty} \sup \left\langle \frac{x}{|x|}, \nabla \log \pi(x) \right\rangle$$

A discussion of the case when π is sub-exponential in the tails comes in the relationship between AF and AM section.
We will also assume the sequence of step sizes γ is non-increasing, positive and

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \quad \sum_{k=1}^{\infty} \{\gamma_k^2 + k^{-1/2}\gamma_k\} < \infty \tag{2}$$

Now consider the process $\{Z_k\}$ with $\{P_\theta, \theta = (\mu, \Gamma) \in \Theta \triangleq \mathbb{R}^{n_x} \times \mathcal{C}_+^{n_x}$ where n_x is some integer and $\mathcal{C}_+^{n_x}$ is the cone of positive $n_x \times n_x$ matrices. We will let the proposal distribution q_θ be normal with covariance matrix $\lambda\Gamma$ where λ is a constant depending on the dimension of the space only. Let $\{H_\theta, \theta \in \Theta\}$ be as in (1), π satisfy (M) and γ satisfy (2). If we let $W \triangleq \pi^{-1}/(\sup \pi)^{-1}$, then for any $\alpha \in [0, 1)$, for any $f \in \mathcal{L}(W^\alpha)$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\infty} f(X_k) \to \pi(f)$$

almost surely $\bar{\mathbb{P}}_*$.

2 Roberts-Rosenthal

The following is another ergodicity result due to Roberts and Rosenthal which is stated in terms of distributional convergence rather than the strong law of large number format in which the previous result was stated. We define a new chain on X such that letting $\mathcal{G}_n = \sigma(X_0, \ldots, X_n, \theta_0, \ldots, \theta_n)$ we have that

$$P[X_{n+1} \in B | X_n = x, \theta_n = \theta, \mathcal{G}_{n-1}] = P_\theta(x, B), \quad x \in X,\, \theta \in \Theta,\, B \in B(X)$$

We will use $||\cdot||$ to denote the total variation distance.

Theorem 2. *Suppose that* $\lim_{n\to\infty} ||P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot)|| = 0$ *in probability. Let* $x_* \in X$ *and* $\theta_* \in \Theta$*. Suppose further that for every* $\epsilon > 0$ *the sequence* $\{M_\epsilon(X_n, \theta_n)\}_{n=0}^\infty$ *is bounded in probability given* $X_0 = x_*$ *and* $\theta_0 = \theta_*$*, where*

$$M_\epsilon(x, \theta) = \inf\{n \geq 1 : ||P_\theta^n(x,\cdot) - \pi(\cdot)|| \leq \epsilon\}$$

Then

$$\lim_{n\to\infty} ||\mathbf{P}[X_n \in \cdot | X_0 = x_*, \theta_0 = \theta_*] - \pi(\cdot)|| = 0 \tag{3}$$

This motivates the following open problem posed in [2], which attempts to weaken the so called *containment condition* that $\{M_\epsilon(X_n, \theta_n)\}_{n=0}^\infty$ is bounded in probability.

Open Problem 1. *Suppose that* $\lim_{n\to\infty} ||P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot)|| = 0$ *in probability. Let* $x_* \in X$ *and* $\theta_* \in \Theta$*. Suppose further that for every* $\epsilon > 0$ *there is* $m \in \mathbb{N}$ *such that* $\mathbb{P}[M_\epsilon(X_n, \theta_n) < m,$ *infinitely often* $|X_0 = x_*, \theta_0 = \theta_*] = 1$*. Does this imply (3) ?*

Example 2. An interesting example demonstrating the limitations of adaptive MCMC is the following running example discussed in [7] which is also available in an animated Java applet (See [8]). Let $K \geq 4$ be an integer and let $X = \{1, \ldots, K\}$. Let $\pi\{2\} = b > 0$ be very small, $\pi\{1\} = a > 0$, $\pi\{3\} = \pi\{4\} = \cdots = \pi\{K\} = (1 - a - b)/(K - 2) > 0$ and $\pi(x) = 0$ for $x \notin X$. Let $\Theta = \mathbb{N}$. For $\theta \in \Theta$, let P_θ be the kernel corresponding to a random walk Metropolis algorithm for $\pi(\cdot)$, with proposal distribution

$$q_\theta(x, \cdot) = \text{Uniform}\{x - \theta, x - \theta + 1, \ldots, x - 1, x + 1, \ldots, x + \theta\}$$

The adaptive scheme is defined as follows. Begin with $\theta_0 = 1$. Let $M \in \mathbb{N} \cup \{\infty\}$ and let $p : \mathbb{N} \to [0, 1]$. For $n = 0, 1, 2, \ldots$, let

$$Z_n = \begin{cases} 1 & \text{if } X_n \neq X_{n+1} \\ -1 & \text{if } X_n = X_{n+1} \end{cases}$$

If $Z_n = 1$ and $\theta_n = M$ then $\theta_{n+1} = \theta_n$. If $Z_n = -1$ and $\theta_n = 1$ then $\theta_{n+1} = \theta_n$. Otherwise with probability $p(n)$ let $\theta_{n+1} = \theta_n + Z_n$ and with probability $1 - p(n)$ let $\theta_{n+1} = \theta_n$.
By changing M, K and $p(n)$ and by manipulating a and b we can obtain a number of variations for this algorithm, some of which surprisingly fail to be ergodic despite seeming fairly simple. For example it is shown in [2] that for $M = 2$, $K = 4$ and $p(n) \equiv 1$ if we let $a = \epsilon$ and $b = \epsilon^2$ for some $\epsilon > 0$ the chain gets stuck at $\{x = \theta = 1\}$ and has a disproportionate probability of entering to that of leaving, which leads to the chain's failure to be ergodic. More precisely it is shown that $\lim_{\epsilon\to 0} \lim_{n\to\infty} P(X_n = \theta_n = 1) = 1$. However it is also shown in [2] that as long as $p(n) \to 0$ the example will be ergodic as a result of theorem 2 above.

3 Atchadé and Fort

In this result due to Atchadeé and Fort, the geometric drift condition is relaxed to a polynomial drift condition which allows consideration of a wider class of chains. For example if the target distribution of interest has heavy tails, then the Random Walk Metropolis Algorithm and the Metropolis Adjusted Langevin algorithm result in sub geometric kernels. We consider a chain $\{Y_k = (X_k, \theta_k),\, k \geq 0\}$ on $X \times \Theta$ with transition kernels $\{\bar{P}(n;\cdot)\, n \geq 0\}$ satisfying that for any $A \in B(X)$,

$$\int_{A\times\Theta} \bar{P}\big(n;(x,\theta);(dx,d\theta')\big) = P_\theta(x,A)$$

We also introduce, for any integer $l \geq 0$, a family of sequences of transition kernels $\{\bar{P}_l(n;\cdot)\, n \geq 0\}$ where $\bar{P}_l(n;\cdot) \triangleq \bar{P}(l+n;\cdot)$ and let $\mathbb{P}^{(l)}_{x,\theta}$ denote the probability of of the Markov chain $\{Y_k = (X_k, \theta_k),\, k \geq 0\}$ with transition kernels $\{\bar{P}_l(n;\cdot)\, n \geq 0\}$, using $\mathbb{P}_{x,\theta}$ as shorthand notation for $\mathbb{P}^{(0)}_{x,\theta}$. The conditions needed for the strong law of large numbers result are as follows:

(B1) There exists $C \in B(X)$, $\epsilon > 0$ and a probability measure φ such that $\varphi(C) > 0$ and for all $A \in B(X)$ and $(\theta, x) \in \Theta \times C$,

$$P_\theta(x,A) \geq \epsilon\varphi(A)$$

(B2) There exists a measurable function $V : X \to [1,\infty)$, constants $\alpha \in (0,1)$ and $b, c \in (0,\infty)$ satisfying

$$P_\theta V(x) \leq \begin{cases} V(x) - cV^{1-\alpha}(x) & \text{if } x \in C^c, \\ b & \text{if } x \in C \end{cases}$$

(B3) There exists a probability measure π and some constant $\beta \in [0, 1-\alpha)$ such that for any level set $\mathcal{D} \triangleq \{x \in X,\, V(x) \leq d\}$ of V,

$$\lim_{n\to\infty} \sup_{\mathcal{D}\times\Theta} ||P^n_\theta(x,\cdot) - \pi||_{V^\beta} = 0$$

(B4) For any level set $\mathcal{D}$ of V and any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{l\geq 0} \sup_{\mathcal{D}\times\Theta} \mathbb{P}^{(l)}_{x,\theta}(D(\theta_n, \theta_{n-1}) \geq \epsilon) = 0$$

where $D(\theta,\theta') \triangleq \sup_x ||P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)||$.

Theorem 3. *Assume B1-B4. Then for any measurable function $f : X \to \mathbb{R}$ in $\mathcal{L}_{V^\beta}$ and any $(x,\theta) \in X \times \Theta$,*

$$\lim_n \frac{1}{n}\sum_{k=1}^n f(X_k) = \pi(f), \quad \mathbb{P}_{x,\theta} - a.s$$

The following is the special case of the above result which we will focus on:

Proposition 1. *Assume that B3 and B4 hold with $\mathcal{D} = X$ and $\beta = 0$. Then for any measurable bounded function $f : X \to \mathbb{R}$ and any $(x,\theta) \in X \times \Theta$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n f(X_k) = \pi(f)$$

almost surely $\mathbb{P}_{x,\theta}$.

Example 3. The following example mentioned in [4] demonstrates two points. The first being that the conditions in the AF paper holding uniformly over compact sets is not sufficient to guarantee that their result holds, and the second is that it displays that even if the marginals converge the law of large numbers might not hold for bounded functions.
Let $X = \{0,1\}$ and $\{P_\theta, \theta \in (0,1)\}$ be a family of transition matrices with $P_\theta(0,0) = P_\theta(1,1) = 1-\theta$. Let $\{\theta_n, n \geq 0\}$, $\theta_n \in (0,1)$, be a deterministic sequence decreasing to 0, and $\{X_n,\, n \geq 0\}$ be a Markov chain on $\{0,1\}$ with transition matrices $\{P_{\theta_n},\, n \geq 0\}$. One can check that the conditions mentioned in the AF section hold uniformly over compact sets, and that $\mathbb{P}_{x,\theta_0}(X_n \in \cdot) \to \pi(\cdot)$ as $n \to \infty$ but that, however, even the weak law of large numbers fails to hold for bounded functions f. For this example in which the marginals do converge to π, neither the conditions for AF nor the ones for RR hold.

In the next result found in [9] Atchade and Fort deal specifically with the Robins Monro stochastic control algorithm dealt with in the AM paper. However the assumptions again require only a sub-geometric drift instead of a geometric one. Consider the chain $\{Z_k; k \geq 0\}$ introduced in the AM section. We will require the following conditions:

(C1) There exists $\alpha \in (0,1]$, and a measurable bounded function $V : X \to [1,\infty)$ such that for any compact subset $\mathcal{K}$ of Θ, there exists $b, c \in (0,\infty)$ such that for any $(x,\theta) \in X \times \mathcal{K}$,

$$P_\theta V(x) \leq V(x) - cV^{1-\alpha}(x) + b$$

(C2) For any $\beta \in [0, 1-\alpha]$, $\xi \in [0, \frac{1-\beta}{\alpha} - 1]$ there exists a constant C such that

$$\sup_{\{g:\, ||g||_{V^\beta} \leq 1\}} |P_\theta^n g(x) - \pi(g)|(n+1)^\xi \leq CV^{\beta+\alpha\xi}(x), \quad n \geq 0$$

(C3)

$$\bar{\mathbb{P}}_*\{\lim_{n\to\infty} \kappa_n < \infty\} = 1$$

(C4) This conditions holds for $\beta \in [0, 1-\alpha)$, if there exist $\epsilon > 0, \xi > 0, \beta + \alpha\xi < 1-\alpha$ such that for any $(x,\theta,l) \in K \times \mathcal{K}_0 \times \mathbb{N}$,

$$\tilde{\mathbb{E}}_{x,\theta}^{\gamma^{\leftarrow l}}\left[\sum_{k=1}^{\infty} \frac{1}{k^{1-\epsilon}} D_\beta(\theta_k, \theta_{k+1}) \mathbb{I}\{\sigma_{\mathcal{K}_l} > k\} V^{\beta+\alpha\xi}(X_k)\right] < \infty$$

where $\tilde{\mathbb{E}}_{x,\theta}^{\gamma^{\leftarrow l}}$ as in the AM section is the expectation with respect to the stopped process $\{(X_k, \theta_k); k \geq 0\}$ with step size sequence $\gamma^{\leftarrow l}$, and

$$\sigma_{\mathcal{K}_l} \triangleq \{\inf k \geq 1 :\, \theta_k \notin \mathcal{K}_\updownarrow\}$$

Now for $\beta \in [0,1]$, $\theta, \theta' \in \Theta$ define

$$D_\beta(\theta,\theta') \triangleq \sup_{||f||_{V^\beta} \leq 1} ||P_\theta f - P_{\theta'} f||_{V^\beta}$$

Theorem 4. *Assume (C1)-(C3) and (C4) with some $\beta \in [0, 1-\alpha)$. Let $f : X \to \mathbb{R} \in \mathcal{L}_{V^\beta}$ be such that $\sup_{\theta\in\mathcal{K}} ||f||_{V^\beta} < \infty$ for any compact subset $\mathcal{K}$ of Θ. Then*

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{n} f(X_k) = 0$$

in $\bar{\mathbb{P}}_*$ probability.

This brings us to another open problem.

Open Problem 2. *Assume (C1) -(C4). Does this imply a weak law of large numbers for measurable functions f without the additional assumptions that $f \in \mathcal{L}_{V^\beta}$, $0 \leq \beta < 1 - \alpha$?*

4 Relationship between RR and AM

Theorem 5. *If the conditions in Theorem 1 hold, then the conditions in Theorem 2 hold.*

Proof. Suppose the conditions in theorem 1 hold. We first show that the diminishing adaptation condition must hold as well. Since we have that $\lim_{n\to\infty} \kappa_n < \infty$ almost surely, we can work on the set where κ_∞ is finite. We can write the said set as the union, $\bigcup_n \{\kappa_\infty = n\}$, and hence it suffices to prove that the diminishing adaptation condition holds almost surely on each set in the union. Suppose $\kappa_\infty = m$. Then our kernel indices θ_k eventually live in the compact set $\mathcal{K}_m$. Let $\tau = \inf_k\{\theta_n \in \mathcal{K}_m \,\forall\, n > k\}$. Since our chain lives in $\bigcup_n \{\tau = n\}$ we need only prove that the desired result holds on each set of the form $\{\tau = k\}$. Now suppose $\tau = k$. Then by (A2) we have that for all $n > k$ and any $f \in \mathcal{L}_V$ there exists a constant C such that

$$||P_{\theta_{n+1}} f - P_{\theta_n} f||_V \leq C||f||_V |\theta_{n+1} - \theta_n|$$

where V is given in (A1). Since, in particular, this holds for all indicator functions we deduce that the total variation distance between $P_{\theta_{n+1}}(x, \cdot)$ and $P_{\theta_n}(x, \cdot)$ is bounded above by $C|\theta_{n+1} - \theta_n|$, for some constant C, uniformly over all x. On the other hand, by construction, $\theta_{n+1} - \theta_n = \gamma_{n+1} H(\theta_n, X_{n+1})$ and, under assumption (A3) since V is bounded there exists a constant C such that, for any $x, \theta \in X \times \mathcal{K}_m$, $|H(\theta, x)| \leq C$. Hence

$$\lim_{n\to\infty} |\theta_{n+1} - \theta_n| \leq \lim_{n\to\infty} \gamma_{n+1} C = 0$$

since $\gamma_n \to 0$ by assumption. Hence $||P_{\theta_{n+1}}(x, \cdot) - P_{\theta_n}(x, \cdot)||$ goes to 0 almost surely as n goes to infinity.

We now have to show that the containment condition holds. We know by theorem 2.3 in [3] that (A1) implies that there exists positive constants $C < \infty$ and $\rho < 1$ such that for any $f \in \mathcal{L}_V$, all $\theta \in \mathcal{K}_m$ and any $j \geq 0$,

$$||P_\theta^j - \pi(f)||_V \leq C||f||_V \rho^j$$

and again, by the same logic used earlier, the total variation distance between $P_\theta^j(x, \cdot)$ and π is bounded above by $C\rho^j$ uniformly over all x and all $\theta \in \mathcal{K}_m$. Given $\epsilon > 0$ let $N_1 = \inf_{j\in\mathbb{N}} C\rho^j < \epsilon$. Recalling that we are working on the set $\{\tau = k\}$, for $n \leq k$ $M_\epsilon(X_n, \theta_n)$ is bounded almost surely by some random integer, $\mathcal{N}$. Since our chain lives in the set $\bigcup_n \{\mathcal{N} = n\}$, we need only prove that the containment condition holds almost surely on each set in the union. Suppose $\mathcal{N} = N_2$. For $n > k$, $M_\epsilon(X_n, \theta_n)$ is bounded above by N_1 almost surely. Letting $N = \max(N_1, N_2)$, we have that for all $x \in X$, and $\theta \in \Theta$,

$$P[M_\epsilon(X_n, \theta_n) \leq N | X_0 = x, \theta_0 = \theta] = 1, \quad \forall n \in \mathbb{N}$$

This completes the proof. For an example where the ergodicity result holds but the convergence is not geometric (and therefore the AM conditions do not hold) consult example 4 in the relationship between AF and AM section. □

5 Relationship between AF and AM

If the conditions for theorem 1 hold, the the conditions for theorem 4 hold as discussed in section 2.5.1 in [4]. However the following example allows us to mention a case where the convergence is sub-geometric and therefore the AM conditions do not hold, but where the AF conditions do hold. More precisely we will consider a specific and slightly modified version of the algorithm of Haario et. al.

Example 4. Consider the sequence $\{Z_k : k \geq 0\}$ with $\Theta = K \times \Theta_+ \times [a, b]$ where K is a compact subset of $\mathbb{R}^{n_x}$, Θ_+ is a convex compact subset of the cone of positive $n_x \times n_x$ matrices, and $-\infty < a < b < \infty$. We will use the sequence of step sizes $\{1/(k+1); k \geq 0\}$. The algorithm updates according to the Haario et. al recursion discussed earlier, with the additional parameter c updating according to

$$c_{n+1} = \frac{1}{n+1}(\alpha(X_n, Y_{n+1}) - \bar{\alpha})$$

where Y_{n+1} is the candidate generated from the proposal distribution q_{θ_n}, and $\bar{\alpha}$ is some constant. We will also assume π is sub-exponential in the tails. More specifically, assume

(N1) π is positive, continuous and twice continuously differentiable in the tails.

(N2) There exists $m \in (0, 1)$, positive constants $d_i < D_i$, $i = 0, 1, 2$ and $r, R > 0$ such that

(i) $\left\langle \frac{\nabla\pi(x)}{|\Delta\pi(x)|}, \frac{x}{|x|} \right\rangle \leq -r$

(ii) $d_0|x|^m \leq -\log \pi(x) \leq D_0|x|^m$

(iii) $d_1|x|^{m-1} \leq |\nabla \log \pi(x)| \leq D_1|x|^{m-1}$

(iv) $d_2|x|^{m-2} \leq |\nabla^2 \log \pi(x)| \leq D_2|x|^{m-2}$

(N3) There exists $s_* > 0$, $0 < v < 1 - m$ and $0 < \eta < 1$ such that

$$\lim_{|x|\to\infty} \sup_{\theta\in\Theta} \int_{\{z,\, |z|\geq\eta|x|^v\}} \left(1 \vee \frac{\pi(x)}{\pi(x+z)}\right)^{s_*} q_\theta(z) \mathrm{d}\mu_{Leb}(z) = o(|x|^{2(m-1)})$$

Assuming (N1) - (N3) then the algorithm described above satisfies the property that there exists $0 < s \leq s_*$ such that for any function $f \in \mathcal{L}_{\pi^{-r}+1}$, $0 \leq r < s$,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \pi(f)$$

almost surely $\bar{\mathbb{P}}_*$. The result above is due to the AF theorem, since the algorithm converges sub-geometrically and the AM conditions are not satisfied.

6 Relationship between RR and AF

It is clear that if the conditions for proposition 1 hold then the conditions for theorem 2 must hold immediately. However example 4 will demonstrate that the conditions for theorem 2 being satisfied does not guarantee that the AF result will hold.

Example 5. We begin with the special case of example 1 discussed in the RR section, with $M = 2$, $p(n) \equiv 1$, $K = 4$ and parameters a and b chosen so that $\lim_{n\to\infty} P(X_n = 1) = p$ for some $p > \pi(1)$. (We can do that since $\lim_{\epsilon\to 0} \lim_{n\to\infty} P(X_n = \theta_n = 1) = 1$). We will call this the "One-Two" adaptation scheme. Let $\{C_k\}_{k=0}^{\infty}$ be a sequence of independent Bernoulli random variables with $P(C_k = 1) = 1/k$, and let $q_k(\cdot)$ denote the distribution of C_k.
We consider the chain $\{(X_n, C_{\psi(n)}, \theta_n) : n \geq 0\}$, where $\psi : \mathbb{N} \to \mathbb{N}$ maps n to k such that $2^{k^2} + 1 \leq n \leq 2^{(k+1)^2}$. The adaptation scheme sets $\theta_0 = \theta_1 = \theta_2 = 1$, and for $n \geq 3$ proceeds with the One-Two adaptation scheme if $C_{\psi(n)} = 1$ and sets $\theta_n = 1$ otherwise. The transitions of this Markov process are given by the family of transition kernels $\{\bar{P}(n; (x, c, \theta), (dx', dc', d\theta'), n \geq 0\}$ where

$$\begin{aligned}
\bar{P}(n; (x, c, \theta), (dx', dc', d\theta')) &= \Big(\mathbb{I}\{c = 1\} P_\theta(x, dx') \big(\mathbb{I}\{x = x'\} \delta_{1\vee(\theta-1)}(d\theta') + \mathbb{I}\{x \neq x'\} \delta_{M\wedge(\theta+1)}(d\theta')\big) \\
&+ \mathbb{I}\{c = 0\} P_\theta(x, dx') \delta_\theta(d\theta')\Big) \\
&\times \Big(\mathbb{I}\{\psi(n) \neq \psi(n+1)\} q_{\psi(n+1)}(dc') + \mathbb{I}\{\psi(n) = \psi(n+1)\} \delta_c(dc')\Big)
\end{aligned}$$

It is shown in [2] that the above chain satisfies the conditions for theorem 2, however the strong law of large numbers fails for the function $g(x) = \mathbb{I}\{x = 1\}$.

References

[1] Andrieu, C. and Moulines, E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. 16 14621505.

[2] Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive MCMC. Journal of Applied Probablity 44 458475.

[3] Meyn, S. and Tweedie, R. (1994). Computable bounds for convergence rates of Markov chains. Ann. Appl. Probab. 4 9811011.

[4] Atchade, Y. and Fort, G. (2008). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. Bernoulli 16 116-154.

[5] Gelman, A., Roberts, G. and Gilks, W. (1995). Efficient Metropolis jumping rules. In Bayesian Statistics 5 (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 599608. Oxford University Press, New York.

[6] Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. Bernoulli 7 223242.

[7] Y.F. Atchade and J.S. Rosenthal (2005), On Adaptive Markov Chain Monte Carlo Algorithms. Bernoulli 11(5), 815828.

[8] J.S. Rosenthal (2004), Adaptive MCMC Java Applet. Available at:
http://probability.ca/jeff/java/adapt.html

[9] Atchade, Y. and Fort, G. (2008). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels: Part II. Bernoulli, (to appear).