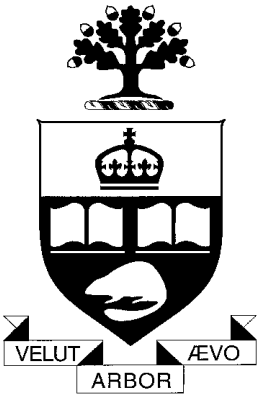


On Adaptive Markov Chair Monte Carlo Algorithms

by

Yves F. Atchadé
Department of Mathematics and Statistics
University of Ottawa

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 0304 July 5, 2003

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

On Adaptive Markov Chain Monte Carlo Algorithms

Yves F. Atchadé[1] *and Jeffrey S. Rosenthal*[2]

(July, 2003; revised March, 2005)

Abstract

We look at adaptive MCMC algorithms that generate stochastic processes based on sequences of transition kernels, where each transition kernel is allowed to depend on the past of the process. We show under certain conditions that the generated stochastic process is ergodic, with appropriate stationary distribution. We use this result to analyze an adaptive version of the Random Walk Metropolis algorithm where the scale parameter σ is sequentially adapted using a Robbins-Monro type algorithm in order to find the optimal scale parameter σ_{opt} as in Roberts et al. (1997). We close with a simulation example.

Key words: Adaptive Markov Chain Monte Carlo, Metropolis algorithm, Mixingales, Parameter Tuning, Robbins-Monro Algorithm.

MSC Numbers: 65C05, 65C40, 60J27, 60J35

1 Introduction

Markov Chain Monte Carlo (MCMC) methods have now become an important numerical tool in statistics (see e.g. Gilks et al. (1996), Liu (2001)). These methods usually require various parameters (e.g. proposal scalings) to be appropriately tuned for the algorithm to converge reasonably well. In this paper, we develop and analyze adaptive MCMC algorithms where these parameter tunings can be handled automatically.

We consider Monte Carlo algorithms based on random processes (that we shall call adaptive Markov chains) where all the past of the process is used to make the next move in the algorithm. The set up is a generalization of Haario et al. (2001). We prove two ergodicity results for such algorithms (Theorem 3.1 and Theorem 3.2). The rate of convergence obtained in Theorem 3.1 tends

[1]Department of Mathematics and Statistics, University of Ottawa, email: yatchade@uottawa.ca

[2]Department of Statistics, University of Toronto, email: jeff@math.toronto.edu

to indicate that these algorithms converge at a much slower rate. Nevertheless, their adaptability is an important attractive feature.

We apply these results to prove the convergence of a new adaptive Random Walk Metropolis (Adaptive RWM) algorithm (Algorithm 4.1) with proposal kernel $q_\sigma(x, y)$, the density of the d-dimensional Multivariate-Normal distribution $N(x, \sigma^2 I_d)$. It is well known that an effective implementation of this algorithm requires a good choice of the parameter σ^2; and this choice depends on the density π. Some theoretical and empirical results (Roberts et al. (1997), Roberts and Rosenthal (2001)) have shown that in high dimensional spaces, under various regularity conditions, it is optimal to choose σ^2 such that the asymptotic acceptance rate of the algorithm is approximately $\bar{\tau} = 0.234$. However, much trial-and-error may be required to find such value for σ^2. In Section 4, we propose an adaptive version of the RWM algorithm which sequentially adapts σ^2 so as to reach the optimal acceptance rate $\bar{\tau}$. Our adaptive algorithm is based on a stochastic approximation algorithm.

A number of interesting ideas about adaptive MCMC methodology have previously been introduced. Gilks et al. (1998) have shown that the transition kernel used in a MCMC algorithm can be updated (without damaging the ergodicity of the algorithm) at regeneration times. The problem with this approach is that regeneration times for Markov chains are difficult to identify, particularly in high dimensional spaces. Haario et al. (2001) have proposed an adaptive version of the RWM where the covariance matrix of the proposal kernel is sequentially updated. A recent paper much comparable to this work is Andrieu and Moulines (2003). These authors have simultaneously and independently developed convergence results for adaptive MCMC algorithms. Although there is not much overlap between the two papers, both have similar assumptions.

Throughout this paper, π represents the probability measure of interest defined on some measurable space $(\mathcal{X}, \mathcal{F})$. In Section 2, we provide an example of adaptive algorithm (Algorithm 2.1) that

fails to converge. A general analysis for adaptive MCMC is developed in Section 3. The main results are Theorem 3.1 and Theorem 3.2. In Section 4 (Algorithm 4.1), we introduce a new adaptive RWM algorithm that can iteratively find the optimal scale parameter (Theorem 4.1). Simulation results are presented in Section 5.

2 Cautionary Examples

We begin with a simple example due to G.O. Roberts (personal communication), where an intuitively reasonable adaptive rule fails to give the expected asymptotic distribution.

Take $\mathcal{X} = \{1, 3, 4\}$, and let π be the uniform distribution on $\mathcal{X}$. For $i = 1, 2$, and $x \in \mathcal{X}$, let $Q_i(x, \cdot)$ be the uniform distribution on $\{x - i, x + i\}$ (when $x \notin \mathcal{X}$, $Q(x, x) = 1$) and $R_i(x, \cdot) = (1 - \beta)Q_i(x, \cdot) + \beta\pi(\cdot)$, for some fixed $\beta \in [0, 1]$. Consider the following adaptive Metropolis algorithm.

Algorithm 2.1. 1. Start the algorithm at $X_0 = x_0 \in \mathcal{X}$.

2. Suppose that at some time n, $X_n = x$. If $n = 0$, sample $Y_{n+1} \sim R_2(x, \cdot)$. Otherwise:

   2.1 If the last move was a rejection, sample $Y_{n+1} \sim R_1(x, \cdot)$.

   2.2 If the last move was an acceptation, sample $Y_{n+1} \sim R_2(x, \cdot)$.

3. If $Y_{n+1} \in \mathcal{X}$, "accept" Y_{n+1} and set $X_{n+1} = Y_{n+1}$ otherwise "reject" Y_{n+1} and set $X_{n+1} = x$.

The strategy used in this algorithm is quite intuitive. Large step moves (from R_2) are proposed to help increase the mixing rate of the chain. But these moves are more likely to be rejected and when they are, the algorithm tries a smaller step move (from R_1). Each proposal R_i gives an ergodic Metropolis algorithm, but in fact, Algorithm 2.1 fails to give the right asymptotic distribution.

To see why, let (X_n) be the stochastic process resulting from algorithm 2.1 and define $Z_n := (X_n, X_{n-1}) \in \mathcal{X} \times \mathcal{X}$. It is easy to see that (Z_n) is a Markov chain. We can write the transition

matrix of (Z_n). For $m, n \in \mathcal{X}$, note $\phi(m,n) = 1$ if $m = n$ and $\phi(m,n) = 2$ otherwise. Also define $\psi(m,n) = 1 - \beta$ if $m = n = 1$ or $(m \neq n$ and $n = 4)$; and $\psi(m,n) = (1-\beta)/2$ otherwise. Then $P((m,n),(n,j))$ the probability that $Z_n = (n,j)$ given that $Z_{n-1} = (m,n)$ can be written:

$$P((m,n),(n,j)) = \begin{cases} (1-\beta)Q_{\phi(m,n)}(n,j) + \beta\pi(j) & \text{if } j \neq n \\ \beta\pi(n) + \psi(m,n) & \text{if } j = n \end{cases}$$

It can be checked that P is irreducible and aperiodic. Since $\mathcal{X} \times \mathcal{X}$ is finite, P is ergodic. Let $\nu(i,j)$ be the invariant distribution for P. Then $\{X_n = 1\} = \{X_n = 1, X_{n-1} = 1\} \cup \{X_n = 1, X_{n-1} = 3\} \cup \{X_n = 1, X_{n-1} = 4\}$ which implies that:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}_{\{X_i=1\}} &= \lim_{n\to\infty} \Pr(X_n = 1) \\ &= \nu(1,1) + \nu(1,3) + \nu(1,4). \end{aligned}$$

The computation of the matrix ν requires to solve the 9×9 linear equation $\nu P = \nu$. We do it numerically for different values of β. Table 1 summarises the results.

β	0.001	0.01	0.1	0.5	0.9	0.99
$\lim \Pr(X_n = 1)$	0.9898	0.9088	0.5589	0.3517	0.3337	0.3334

Table 1: $\lim_{n\to\infty} \Pr(X_n = 1)$ as a function of β in Algorithm 2.1.

Clearly, for all $\beta \in [0,1)$, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}_{\{X_i=1\}} = \lim_{n\to\infty} \Pr(X_n = 1) > \frac{1}{3}$. As we shall see, this adaptive MCMC algorithm fails because the successive transition kernels in use failed to stabilize as the simulation goes along, a key requirement for an adaptive MCMC algorithm. For an interactive version of a related example, see Rosenthal (2004).

3 General Ergodicity Results

Assume that we have a starting transition kernel P_0 and an initial point $x_0 \in \mathcal{X}$. Consider the following generic adaptive MCMC algorithm:

Algorithm 3.1. 1. Suppose that at some time $n \geq 0$, we have $X_n = x$ and a transition kernel $P_{n,\tilde{X}_n}$ which is allowed to depend on the path $(X_0, \ldots, X_n) = \tilde{X}_n \in \mathcal{X}^{n+1}$ of the algorithm.

2. Sample $X_{n+1} \sim P_{n,\tilde{X}_n}(x, \cdot)$.

3. Use $\tilde{X}_{n+1} = (X_0, \ldots, X_{n+1})$ to build a new transition kernel $P_{n+1,\tilde{X}_{n+1}}$ to be used at time $n+1$.

We take $P_{0,\tilde{x}_0} = P_0$ the starting transition kernel.

To run Algorithm 3.1, we assume that we have at our disposal a family $\{P_{n,\tilde{x}_n}(x, A) : n \geq 0, \tilde{x}_n \in \mathcal{X}^{n+1}, x \in \mathcal{X}, A \in \mathcal{F}\}$ which is such that for $n \geq 0$, $\tilde{x}_n \in \mathcal{X}^{n+1}$, and $x \in \mathcal{X}$ fixed, $P_{n,\tilde{x}_n}(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{F})$ and for $A \in \mathcal{F}$, $P_{n,\tilde{x}_n}(x, A)$ is a measurable function from $(\mathcal{X}^{n+1} \times \mathcal{X}, \mathcal{F}^{n+1} \times \mathcal{F})$ to $[0, 1]$.

Let (X_n) be the random process generated by Algorithm 3.1 and $\mathbb{P}_\mu$ its distribution on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ when $X_0 \sim \mu$. We shall write E_μ to denote the expectation with respect to $\mathbb{P}_\mu$. As usual, if $\mu = \delta_x$ the Dirac measure on x, we write E_x and $\mathbb{P}_x$ instead of E_μ and $\mathbb{P}_\mu$ respectively.

For a probability measure μ and a transition kernel P, the product μP defines a probability measure by $\mu P(\cdot) := \int \mu(dx) P(x, \cdot)$. And if f is a real-valued function on $\mathcal{X}$, the product Pf defines a function by $Pf(x) := \int P(x, dy) f(y)$. If P and Q are two transition kernels, the product PQ is also a transition kernel defined by $PQ(x, A) := \int P(x, dy) Q(y, A)$. This allows us to define Q^n the n times product of Q by itself, with the convention that $Q^0(x, A) = \mathbf{1}_A(x)$. Finally, for a probability measure μ and a positive function V, we define the V-norm of μ by $\|\mu\|_V := \sup_{|f| \leq V} |\mu(f)|$, where $\mu(f) := \int f(x) \mu(dx)$.

We study the ergodicity of the stochastic process generated by Algorithm 3.1.

We assume that for $n \geq 0$ and $\tilde{x}_n \in \mathcal{X}^{n+1}$, there exists a probability measure $\pi_{n,\tilde{x}_n}$ on $\mathcal{X}$ such that:

$$\pi_{n,\tilde{x}_n} P_{n,\tilde{x}_n} = \pi_{n,\tilde{x}_n}, \tag{3.1}$$

and that the function $\tilde{x}_n \longrightarrow \pi_{n,\tilde{x}_n}(A)$ is measurable for every $n \geq 0$ and $A \in \mathcal{F}$. In words, $\pi_{n,\tilde{x}_n}$ is an invariant distribution for $P_{n,\tilde{x}_n}$.

We require the following assumptions:

Assumption A1: *There exist a measurable function $V : \mathcal{X} \to [1,\infty)$ and sequences of real numbers (τ_n), (a_n), (R_n), with $\tau_n, R_n \to 0$ as $n \to \infty$ such that:*

A1.1 *for $j \geq 1$, $n \geq 0$, $x \in \mathcal{X}$ and $\tilde{x}_n \in \mathcal{X}^{n+1}$:*

$$\left\| P^j_{n,\tilde{x}_n}(x,\cdot) - \pi_{n,\tilde{x}_n}(\cdot) \right\|_V \leq R_j V(x), \tag{3.2}$$

A1.2 *for $x \in \mathcal{X}$, $\tilde{x}_n \in \mathcal{X}^{n+1}$, $\tilde{y}_k \in \mathcal{X}^{k+1}$, $\tilde{x}_{n+k} = (\tilde{x}_n, \tilde{y}_k)$,*

$$\left\| P_{n+k,\tilde{x}_{n+k}}(x,\cdot) - P_{n,\tilde{x}_n}(x,\cdot) \right\|_V \leq K_1 \tau_n a_k V(x), \tag{3.3}$$

and

$$\left\| \pi_{n+k,\tilde{x}_{n+k}} - \pi_{n,\tilde{x}_n} \right\|_V \leq K_2 \tau_n a_k, \tag{3.4}$$

A1.3 *there exists $K_3 < \infty$ such that for $n \geq 0$ and $k \geq 1$:*

$$\int P_{n,\tilde{x}_n}(x_n, dx_{n+1}) \cdots \int P_{n+k-1,\tilde{x}_{n+k-1}}(x_{n+k-1}, dx_{n+k}) V^2(x_{n+k}) \leq K_3 V^2(x_n), \tag{3.5}$$

A1.4

$$\sup_{n,\tilde{x}_n} \pi_{n,\tilde{x}_n}(V) < \infty, \tag{3.6}$$

A1.5 *for finite constants c_1, c_2, defining $B(c_1,c_2,n) := \min_{1 \leq k \leq n} (c_1 \phi_k \tau_{n-k} + c_2 R_k)$, where $\phi_n = \sum_{k=1}^n a_k$, we have $B(c_1,c_2,n) = \mathcal{O}\left(\frac{1}{n^\varepsilon}\right)$ for some $\varepsilon > 0$.*

We would like to investigate the ergodicity of (X_n) under these assumptions. In the sequel, we write $\tilde{X}_n = (X_0, \ldots, X_n)$.

Theorem 3.1. *Assume that $X_0 = x_0 \in \mathcal{X}$. Under (A1.1)-(A1.4), there are some constants $k_1, k_2 < \infty$ such that for any measurable function $f : \mathcal{X} \longrightarrow \mathbb{R}$ with $|f| \leq V$:*

$$\left|\mathbb{E}_{x_0}\left(f(X_n) - \pi_{n,\tilde{X}_n}(f)\right)\right| \leq B(k_1, k_2, n)V(x_0), \tag{3.7}$$

where V and $B(k_1, k_2, n)$ are as in (A1).

Theorem 3.2. *Under (A1.1)-(A1.5) and for any measurable function $f : \mathcal{X} \longrightarrow \mathbb{R}$ with $|f| \leq V$, where V is as in (A1), we have:*

$$\frac{1}{n}\sum_{i=0}^{n-1}\left(f(X_i) - \pi_{i,\tilde{X}_i}(f)\right) \longrightarrow 0, \ \text{as } n \to \infty, \ \mathbb{P}_{x_0} - p.s. \tag{3.8}$$

for any starting point $x_0 \in \mathcal{X}$.

Remark 3.1.

1. For most MCMC algorithms, one would have $\pi_{n,\tilde{x}_n} = \pi$ the invariant distribution of interest and in this case, theorem 3.1 gives a bound on the rate of convergence of the distribution of X_n to π and theorem 3.2 states a law of large numbers.

2. Assumption (A1.1) requires an uniform in time (geometric or subgeometric) convergence rate of $P_{n,\tilde{x}_n}$ to $\pi_{n,\tilde{x}_n}$. This may be hard to check in practice. For example, to obtain a geometric convergence rate ($R_n = R\rho^n$ for some $0 < \rho < 1$) in (A1.1), one possible way is to use quantitative bounds for Markov chains (e.g. Meyn and Tweedie (1994)) which typically requires a drift condition of the form:

$$P_{n,\tilde{x}_n}V(x) \leq \lambda V(x) + b\mathbf{1}_C(x), \tag{3.9}$$

for some $\lambda < 1$, $b < \infty$ and some small set C (for $P_{n,\tilde{x}_n}$) that do not depend on n; and a minorization condition:

$$P_{n,\tilde{x}_n}(x, \cdot) \geq \varepsilon\nu(\cdot) \quad x \in C, \tag{3.10}$$

where ε does not depend on n. It is now well known that many MCMC Markov chains satisfy a drift and a minorization condition. But the fact that the constants involved in these conditions

do not depend on n make them more difficult to establish in general. Nevertheless, there are some useful MCMC algorithms (like the Random Walk Metropolis algorithms) where (A1) can be shown to hold. We return to this point in Section 4.

3. (A1.2) requires that as $n \to \infty$ the adaptation procedure results in more and more stable transition kernels. It can be shown that the example in Algorithm 2.1 satisfies all the assumptions above but (A1.2).

4. Theorem 3.1 tells us that the rate of convergence of the adaptive MCMC will be the worst rate between the rate of convergence of the (nonadaptive) transition kernels R_n and the rate of convergence of the adaptation process τ_n as in (A1.2). For example, taking $a_n = \mathcal{O}(n^{\lambda_2})$ for some $\lambda_2 > 0$, it is easily seen that if τ_n is geometric and R_n is geometric then $B(k_1, k_2, n) := \min_{1 \le k \le n} (c_1 \phi_k \tau_{j-k} + c_2 R_k)$ is also geometric. But for most adaptive MCMC algorithms we typically have $\tau_n = \mathcal{O}(n^{-\lambda_1})$ for some $\lambda_1 > 0$ and assuming that $R_n = R\rho^n$ for some $0 < \rho < 1$, and taking $k = \alpha \log n$, $\alpha = -\lambda_1 / \log \rho$, we obtain the polynomial rate $B(k_1, k_2, n) = \mathcal{O}\left(n^{-\lambda_1} (\log n)^{\lambda_2 + 1}\right)$.

We now proceed to prove these theorems. Our proofs are based on a version of the strong law of large numbers for mixingales and closely follow Haario et al. (2001). For an introduction to mixingales see Hall and Heyde (1980). We use the following theorem, adapted from Davidson and de Jong (1997), Corollary 2.1.

Let $\mathcal{F}_n = \{\phi, \mathcal{S}\}$ be the trivial σ-algebra when $n < 0$, and $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ be the σ-algebra generated by $(X_0, \ldots, X_n)$ when $n \ge 0$.

Lemma 3.1. *Assume that (A1.1)-(A1.4) hold. Then there are constants $0 < k_1, k_2 < \infty$ such that for any $n \ge 0$, $j \ge 1$, and any measurable function f with $|f| \le V$, we have:*

$$\left| E_{x_0} \left(g_{n+j, \tilde{X}_{n+j}}(X_{n+j}) | \mathcal{F}_n \right) \right| \le B(k_1, k_2, j) V(X_n), \tag{3.11}$$

P_{x_0}-*a.s. where* $g_{k,\tilde{X}_k} = f - \pi_{k,\tilde{X}_k}(f)$.

Proof. We have $\pi_{k,\tilde{X}_k}(g_{k,\tilde{X}_k}) = 0$ $\mathbb{P}_{x_0}$-a.s. Given $(X_0, X_1, \ldots, X_{n-1}) = \tilde{x}_{n-1}$ and $X_n = x$, we have:

$$E_{x_0}\left(g_{n,\tilde{X}_n}(X_{n+j})|\tilde{X}_{n-1} = \tilde{x}_{n-1}, X_n = x\right) = \sum_{k=1}^{j-1} \eta_k(\tilde{x}_{n-1}, x) + P^j_{n,\tilde{x}_n} g_{n,\tilde{x}_n}(x), \tag{3.12}$$

where

$$\eta_k(\tilde{x}_{n-1}, x) = \int P_{n,\tilde{x}_n}(x, dx_{n+1}) \cdots$$
$$\cdots \int P_{n+k-1,\tilde{x}_{n+k-1}}(x_{n+k-1}, dx_{n+k}) \int P^{j-k-1}_{n,\tilde{x}_n} g_{n,\tilde{x}_n}(x_{n+k+1}) \left(P_{n+k,\tilde{x}_{n+k}}(x_{n+k}, dx_{n+k+1}) - P_{n,\tilde{x}_n}(x_{n+k}, dx_{n+k+1})\right).$$

Using assumption A1.1, we can bound the second term of the left hand side of (3.12) as follows:

$$\left|P^j_{n,\tilde{x}_n} g_{n,\tilde{x}_n}(x)\right| \leq R_j V(x). \tag{3.13}$$

From A1.2 and using the fact that $\sup_{n,\tilde{x}_n} \pi(V) < \infty$, we have the following bounds for some finite constant r_0:

$$\begin{aligned} |\eta_k(\tilde{x}_{n-1}, x)| &\leq r_0 \tau_n a_k \int P_{n,\tilde{x}_n}(x, dx_{n+1}) \cdots \int P_{n+k-1,\tilde{x}_{n+k-1}}(x_{n+k-1}, dx_{n+k}) V(x_{n+k}), \\ &= r_0 \tau_n a_k E_{x_0}\left(V(X_{n+k})|\tilde{X}_n = (\tilde{x}_{n-1}, x)\right). \end{aligned} \tag{3.14}$$

Putting (3.13) and (3.14) together in (3.12), we get:

$$\left|E_{x_0}\left(g_{n,\tilde{X}_n}(X_{n+j})|\mathcal{F}_n\right)\right| \leq R_j V(X_n) + r_0 \tau_n \sum_{k=1}^{j-1} a_k E_{x_0}\left(V(X_{n+k})|\mathcal{F}_n\right). \tag{3.15}$$

Taking into account (3.4) of (A1.2) leads to:

$$\begin{aligned} \left|E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_n\right)\right| &\leq R_j V(X_n) + r_0 \tau_n \sum_{k=1}^{j-1} a_k E_{x_0}\left(V(X_{n+k})|\mathcal{F}_n\right) + K_2 \tau_n a_j \quad (3.16) \\ &\leq R_j V(X_n) + \max(r_0, K_2) \tau_n \sum_{k=1}^{j} a_k V(X_n) \\ &\leq V(X_n)\left(r_3 R_j + r_2 \tau_n \phi_j\right), \quad (3.17) \end{aligned}$$

where in the last inequality, we use assumption (A1.3) and $\phi_j = \sum_{k=1}^{j} a_k$, $r_2 = \max(r_0, K_2)K_3$, $r_3 = K_3$ and K_3 is as defined in Assumption A1.4.

Since the family $(\mathcal{F}_n)_{n=-\infty}^{\infty}$ is increasing, $\mathcal{F}_n \subseteq \mathcal{F}_{n+j-k}$ for $k = 1$ to j. Therefore

$$E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_n\right) = E_{x_0}\left[E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_{n+j-k}\right)|\mathcal{F}_n\right].$$

It follows that:

$$\left|E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_n\right)\right| \leq E_{x_0}\left[\left|E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_{n+j-k}\right)\right||\mathcal{F}_n\right]. \tag{3.18}$$

Applying (3.17) to the right side of (3.18) gives:

$$\begin{aligned}\left|E_{x_0}\left(g_{n+j,\tilde{X}_{n+j}}(X_{n+j})|\mathcal{F}_n\right)\right| &\leq \min_{1\leq k\leq j}(r_2\tau_{n+j-k}\phi_k + r_3R_k)E_{x_0}\left(V(X_{n+j-k})|\mathcal{F}_n\right)\\ &\leq V(X_n)B(k_1,k_2,j)\end{aligned}$$

for some constants k_1, k_2. □

Proof of Theorem 3.2. Taking $n = 0$ in (3.11) of lemma 3.1 gives for $n \geq 1$:

$$\left|E_{x_0}\left(g_{n,\tilde{X}_n}(X_n)\right)\right| \leq B(k_1,k_2,n)V(x_0). \tag{3.19}$$

This with (A1.5) shows that:

$$E_{x_0}\left(f(X_n) - \pi_{n,\tilde{X}_n}(f)\right) \longrightarrow 0, \quad \text{as } n \longrightarrow \infty. \tag{3.20}$$

Write $Y_n = f(X_n) - \pi_{n,\tilde{X}_n}(f) - E_{x_0}\left(f(X_n) - \pi_{n,\tilde{X}_n}(f)\right)$. Given Lemma 3.1, it can be easily shown that $(Y_n, \mathcal{F}_n)$ is an L^2-mixingale with mixingale sequences $c_n \equiv c$ constant and $\psi_n = B(k_1,k_2,n)$. It follows from Corollary 2.1 of Davidson and de Jong (1997) that:

$$\frac{1}{n}\sum_{k=0}^{n-1} g_{k,\tilde{X}_k}(X_k) - E_{x_0}\left(g_{k,\tilde{X}_k}(X_k)\right) \longrightarrow 0, \quad \mathbb{P}_{x_0} - a.s. \text{ as } n \longrightarrow \infty. \tag{3.21}$$

Combining (3.20) and (3.21), we get as desired that

$$\frac{1}{n}\sum_{k=0}^{n-1}\left(f(X_k) - \pi_{k,\tilde{X}_k}(f)\right) \longrightarrow 0 \ \ \mathbb{P}_{x_0} - a.s. \text{ as } n \longrightarrow \infty. \tag{3.22}$$

□

Proof of Theorem 3.1. Taking $n = 0$ in (3.11) of Lemma 3.1, we obtain the following:

$$\left|E_{x_0}\left(f(X_n) - \pi_{n,\tilde{X}_n}(f)\right)\right| \leq B(k_1, k_2, n)V(x_0), \tag{3.23}$$

for all $|f| \leq V$, which is theorem 3.1. □

4 Application to the Random Walk Metropolis Algorithm

In this section, $\mathcal{X}$ is an open subset of $\mathbb{R}^d$, the d-dimensional Euclidean space equipped with its Borel subsets $\mathcal{B}^d$. We let π be a positive continuous density with respect to Lebesgue measure on $\mathcal{X}$. We denote by $|\cdot|$ the Euclidean norm on $\mathcal{X}$. We consider the Random Walk Metropolis (RWM) algorithm with proposal density $q_\sigma(x, y) = N(x, \sigma^2 I_d)$. This algorithm generates a Markov chain (X_n) with invariant distribution π as follows. Given X_n, a new proposal $Y_{n+1} \sim N(X_n, \sigma^2 I_d)$ is made. We then either "accept" the proposed value and set $X_{n+1} = Y_{n+1}$ with probability $\alpha(X_n, Y_{n+1})$, or we "reject" and set $X_{n+1} = X_n$ with probability $1 - \alpha(X_n, Y_{n+1})$, where $\alpha(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right)$. This algorithm always has stationary distribution π. However, the choice of the scaling parameter σ^2 has a large effect on the algorithm's mixing time. Intuitively, if σ^2 is too small, the resulting algorithm will make very small moves resulting in a poor mixing time. On the other hand, if σ^2 is too large, then large moves will usually be proposed, and these are likely to be rejected so the algorithm will again mix poorly. Here we propose an adaptive version of the RWM algorithm that can automatically find σ such that the asymptotic acceptance rate of the algorithm is approximately $\bar{\tau} = 0.234$.

4.1 The adaptive RWM algorithm

Let P_σ be the transition kernel of the RWM algorithm with proposal $q_\sigma(x, y)$. Let

$$A(\sigma, x) := \int \alpha(x, y) q_\sigma(x, y) dy \quad \text{and } \tau(\sigma) := \int A(\sigma, x)\pi(x)dx, \tag{4.1}$$

be the acceptance rate at x and in stationarity respectively. Our adaptive algorithm relies on stochastic approximation algorithms initiated by Robbins and Monro (1951). These are well-known recursive algorithms of the form $\theta_{n+1} = \theta_n + \gamma_n(h(\theta_n) + \varepsilon_{n+1})$, typically used to solve equations of the form $h(\theta) = 0$ when the function h is unknown (understand hard to compute) but can be estimated with a noise (see e.g Kushner and Yin (2003) and the references therein).

Fix $0 < \varepsilon_1 < A_1$. Define $\Delta = \{\sigma : \ \varepsilon_1 \leq \sigma \leq A_1\}$. We shall assume that there is a unique $\sigma_{opt} \in \Delta$ such that $\tau(\sigma_{opt}) = \bar{\tau}$. Next, we need a way to contain the algorithm inside Δ. We define the function $p(\sigma)$ such that $p(\sigma) = \sigma$ if $\sigma \in \Delta$, $p(\sigma) = \varepsilon_1$ if $\sigma < \varepsilon_1$ and $p(\sigma) = A_1$ if $\sigma > A_1$.

Let (γ_n) be a positive sequence of real numbers. Our adaptive algorithm is thus as follows:

Algorithm 4.1. 1. Start the algorithm at some point $x_0 \in \mathcal{X}$ and $\sigma_0 \in \Delta$.

2. Suppose that at time $n \geq 0$, we have $X_n \in \mathcal{X}$ and $\sigma_n \in \Delta$. Then:

   2.1 Generate $Y_{n+1} \sim Q_{\sigma_n}(x, \cdot)$ and generate $U \sim \mathcal{U}(0, 1)$.

   2.2 If $U \leq \alpha(X_n, Y_{n+1})$, then set $X_{n+1} = Y_{n+1}$. Otherwise, set $X_{n+1} = X_n$.

   2.3 Compute

$$\sigma_{n+1} = p\left(\sigma_n + \gamma_n\left(\alpha(X_n, Y_{n+1}) - \bar{\tau}\right)\right). \tag{4.2}$$

Remark 4.1. 1. The algorithm monitors the acceptance rate through the stochastic approximation algorithm (4.2). The algorithm lowers the scale parameter σ_n when the acceptance rate is too small and increases σ_n when the acceptance rate is too high. Instead of updating σ_n at each iteration, a more robust algorithm could be obtained by updating σ_n every w iterations. We tried various value of w in our simulations and did not find much improvement with $w > 1$. But this may not be the case with more complex examples.

2. The projection function p is used to keep σ_n inside Δ and avoid the degeneracy of the algorithm. But the drawback (as with every stochastic approximation algorithms with re-

projection on a fixed compact set) is that the optimal value cannot be found if the compact set Δ is misspecified. In most MCMC contexts though, if necessary, one may run a pilot simulation at $\sigma = \varepsilon_1$ and $\sigma = A_1$ to validate these values. Another approach dating back to Chen and Zhu (1986) has been advocated and developed by Andrieu et al. (to appear) that avoid this problem by using re-projections on a family of nested compact sets. But in MCMC settings, this approach is not necessarily better.

3. A better way to scale the RWM algorithm is to use the proposal distribution $N(x, \sigma\Sigma)$ with $\sigma = \sigma_{opt}$ and $\Sigma = \Sigma_\pi$ the covariance matrix of the distribution π. Since $(\sigma_{opt}, \Sigma_\pi)$ is not known, an adaptive algorithm can also be applied. We do not pursue this here. See Atchade (2004), Andrieu and Moulines (2003) and Haario et al. (2001).

4.2 Ergodicity of the algorithm

We assume that π is super-exponential with asymptotically regular contours (Jarner and Hansen (2000)) and that the function $\tau(\sigma)$ is decreasing on Δ. More precisely:

Assumption A2:

A2.1 *We assume that π is positive with continuous first derivative such that*

$$\lim_{|x|\to\infty} n(x) \cdot \nabla \log \pi(x) = -\infty,$$

and

$$\limsup_{|x|\to\infty} n(x) \cdot m(x) < 0,$$

where ∇ is the gradient operator, $n(x) = \frac{x}{|x|}$ and $m(x) = \frac{\nabla\pi(x)}{|\nabla\pi(x)|}$.

A2.2 *We assume that there exists $\sigma_{opt} \in \Delta$ such that $\tau(\sigma_{opt}) = 0$ and $(\sigma - \sigma_{opt})(\tau(\sigma) - \bar{\tau}) < 0$ whenever $\sigma \neq \sigma_{opt}$.*

A2.3 *(γ_n) is a positive sequence of real numbers such that $\gamma_n = \mathcal{O}(n^{-\lambda_1})$ for some constant $1/2 < \lambda_1 \leq 1$.*

Under (A2.1) it follows from Proposition 9 of Andrieu and Moulines (2003) that the family $(P_\sigma)_{\sigma\in\Delta}$ satisfies a uniform (in σ) minorization and drift condition: there exist $\varepsilon > 0$, $0 < \lambda < 1$, $b < \infty$, a compact nonempty set $C \subseteq \mathcal{X}$ and a nontrivial probability measure ν such that:

$$\inf_{\sigma\in\Delta} P_\sigma(x, A) \geq \varepsilon\nu(A)\mathbf{1}_C(x), \quad A \in \mathcal{B} \;\; x \in \mathcal{X}, \tag{4.3}$$

and

$$\sup_{\sigma\in\Delta} P_\sigma W(x) \leq \lambda W(x) + b\mathbf{1}_C(x), \quad x \in \mathcal{X}, \tag{4.4}$$

where $W(x) = c\pi(x)^{1/2}$, with c such that $W(x) \geq 1$. Moreover there exists a constant $K_1 < \infty$ such that:

$$\sup_{|f|\leq W^{1/2}} |P_{\sigma_2} f(x) - P_{\sigma_1} f(x)| \leq K_1 W^{1/2}(x)\,|\sigma_2 - \sigma_1|\,. \tag{4.5}$$

Theorem 4.1. *Let (X_n) be the stochastic process generated by algorithm 4.1. Assume Assumption (A2) and take $V = W^{1/2}$. Then:*

(i) *there is a finite constant k such that for $n \geq 2$:*

$$\|\mathcal{L}_{x_0}(X_n) - \pi\|_{TV} \leq kn^{-\lambda_1}(\log n)^2, \tag{4.6}$$

where $\mathcal{L}_{x_0}(X_n)$ is the distribution of X_n given that $X_0 = x_0$,

(ii) *for any measurable function $f : \mathcal{X} \longrightarrow \mathbb{R}$ with $|f| \leq V$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \pi(f) \;\; \mathbb{P}_{x_0} - as, \tag{4.7}$$

(iii) *$\sigma_n \longrightarrow \sigma_{opt}$ as $n \to \infty$, $\mathbb{P}_{x_0}$ almost surely.*

Proof. **(i) and (ii)** The minorization condition (4.3) and the drift condition (4.4) imply (A1.3) (with $V = W^{1/2}$), (A1.4) and (A1.1). (A1.1) actually follows from computational bound for

Markov chain in V-norm as in Meyn and Tweedie (1994). Almost surely, the sequence (σ_n) satisfies $|\sigma_{n+k} - \sigma_n| \leq A\frac{k}{n}$ for some finite constant A which together with (4.5) imply (A1.2). Therefore (i) is Theorem 3.1 and (ii) is Theorem 3.2.

(iii) We have the recursion $\sigma_{k+1} = p\left(\sigma_k - \gamma_k(\alpha(X_n, Y_{n+1}) - \bar{\tau})\right)$.

We let $\mathcal{F}_n$ be the σ-algebra generated by $(\sigma_0, X_0, \ldots, \sigma_n, X_n)$, $U_n = (\sigma_n - \sigma_{opt})^2$ and $V_n = -(\sigma_n - \sigma_{opt})(\tau(\sigma_n) - \bar{\tau})$. We recall the definition of $A(\sigma, x) = \int \alpha(x, y) q_\sigma(x, y) dy$ and $\tau(\sigma) = \int A(\sigma, x)\pi(dx)$. It can be easily shown that:

$$E_{x_0}\left(U_{n+1}|\mathcal{F}_n\right) \leq U_n - 2\gamma_n V_n + \gamma_n^2 + 2\gamma_n\varepsilon_n, \tag{4.8}$$

where $\varepsilon_n = (\sigma_n - \sigma_{opt})\left(A(\sigma_n, X_n) - \tau(\sigma_n)\right)$.

Claim: We claim that $\sum \gamma_n\varepsilon_n$ converges almost surely to a finite random variable.

We are then able to apply the Robbins-Siegmund Theorem (see e.g. Duflo (1997) Theorem 1.3.12) to obtain that $U_n = (\sigma_n - \sigma_{opt})^2$ converges (a.s.) to some finite random variable and $\sum \gamma_n V_n < \infty$ (a.s.). That is σ_n converges a.s. to some finite random variable $\sigma_\infty \in \Delta$. Now, it is clear that the function τ is continuous so that $\tau(\sigma_n) \to \tau(\sigma_\infty)$ (a.s.). Suppose that $\sigma_\infty \neq \sigma_{opt}$. Then $V_n \to -(\sigma_\infty - \sigma_{opt})(\tau(\sigma_\infty) - \bar{\tau}) > 0$ which contradicts $\sum \gamma_n V_n < \infty$ since $\sum \gamma_n = \infty$. Hence $\sigma_\infty = \sigma_{opt}$.

Proof of the claim: The proof is similar to the proof of Lemma 3.1. But first observe that we can find $k_1, k_2 < \infty$ such that $|A(\sigma_2, x) - A(\sigma_1, x)| \leq k_1 |\sigma_2 - \sigma_1| V(x)$, and $|\tau(\sigma_2) - \tau(\sigma_1)| \leq k_2 |\sigma_2 - \sigma_1|$, for every $\sigma_1, \sigma_2 \in \Delta$. The proof follows from Proposition 9 of Andrieu and Moulines (2003). It can also be shown directly using the Mean Value theorem applied to $A(\sigma, x)$, x fixed.

For $n \geq 0$ and $k \geq 1$, we have:

$$\begin{aligned}\varepsilon_{n+k} &= (\sigma_{n+k} - \sigma_n)\left(A(\sigma_{n+k}, X_{n+k}) - \tau(\sigma_{n+k})\right) + (\sigma_n - \sigma_{opt})\left(A(\sigma_{n+k}, X_{n+k}) - A(\sigma_n, X_{n+k})\right) \\ &\quad + (\sigma_n - \sigma_{opt})\left(A(\sigma_n, X_{n+k}) - \tau(\sigma_n)\right) + (\sigma_n - \sigma_{opt})\left(\tau(\sigma_n) - \tau(\sigma_{n+k})\right).\end{aligned}$$

Given the recursion on (σ_n) and the fact that the functions A and τ are Lipchitz (for x fixed), non-negative and bounded from above by 1, we can find $C_1 < \infty$ such that:

$$\left|\mathbb{E}_{x_0}\left(\varepsilon_{n+k}|\mathcal{F}_n\right)\right| \leq 3C_1 k\gamma_n V(X_n) + |\sigma_n - \sigma_{opt}|\left|\mathbb{E}_{x_0}\left(A(\sigma_n, X_{n+k}) - \tau(\sigma_n)|\mathcal{F}_n\right)\right|. \tag{4.9}$$

Now we can apply equation (3.17) to $\left|\mathbb{E}_{x_0}\left(A(\sigma_n, X_{n+k}) - \tau(\sigma_n)|\mathcal{F}_n\right)\right|$ to get for some constants $C_2, C_3 < \infty$ and $\rho < 1$:

$$\left|\mathbb{E}_{x_0}\left(\varepsilon_{n+k}|\mathcal{F}_n\right)\right| \leq V(X_n)\left(C_3\rho^k + C_2 k^2\gamma_n\right). \tag{4.10}$$

At this point the same σ-algebra trick used in the proof of Lemma 3.1 can be applied to obtain:

$$\left|\mathbb{E}_{x_0}\left(\varepsilon_{n+k}|\mathcal{F}_n\right)\right| \leq C_4 \log(k)^2\gamma_k V(X_n). \tag{4.11}$$

And if follows that $(\gamma_n(\varepsilon_n - \mathbb{E}(\varepsilon_n)), \mathcal{F}_n)$ is a mixingale with mixingale sequence $c_n \propto \gamma_n$ and $\psi_n \propto \log(n)^2\gamma_n$. Theorem 2.7 of Hall and Heyde (1980) then asserts that for such mixingale, $\sum \gamma_n(\varepsilon_n - \mathbb{E}(\varepsilon_n))$ converges a.s. to a finite random variable. And the claim is proved since $\sum \gamma_n \mathbb{E}(\varepsilon_n)$ is a convergent series.

□

5 A Simulation Example

In this section, we conduct a simulation study to illustrate the results obtained in Section 4. We take π to be the d-dimensional standard Normal distribution for $d = 10$ and 50. We use $Q_\sigma(x, \cdot) \sim N(x, \sigma^2 I_d)$, and as a function of interest, we take $f(x) = x_1$, the first coordinate of x. For all

the simulations, we start with $\sigma_0 = 10$, $a = 0.0001$, $A = 1000$, and each chain is run for 250,000 iterations. (In fact, the initial value σ_0 is not important; in any case the values of σ_n become very *low* before converging upwards to σ_{opt}.) With all the adaptive algorithms, we use $\gamma_n = \frac{\sigma_0}{n}$.

Graph 1 shows the autocorrelation functions of the Adaptive RWM (ARWM) algorithm (with $\bar{\tau} = 0.234$), and the (non-adaptive) RWM with optimal scaling σ_{opt}. Both the adaptive and the optimal non-adaptive algorithms show very comparable performances in term of mixing time as measured by the autocorrelation functions. This shows that our adaptive algorithm achieves essentially the same mixing time as the optimally scaled algorithm, but without requiring all the preliminary effort to manually tune the scaling parameter. For each value of d, we run the simulations with $w = 1$, 10 and 100 where w is the number of observations gathered before updating σ_n. The three value are quiet comparable.

Graph 2 shows the scale parameter process and the empirical acceptance rate obtained during the ARWM simulation for $w = 10$, and for a targeted acceptance rate of $\bar{\tau} = 0.234$. The empirical acceptation probability converges to 0.234, showing that we are indeed finding the optimal scaling parameter σ_{opt}. For large values of d, the value of σ_{opt} is consistent with the formula $\frac{2.38}{\sqrt{d}}$ (0.34 if $d = 50$, 0.75 if $d = 10$) given by Roberts et al. (1997).

Acknowledgement: This research was supported in part by NSERC of Canada.

References

Andrieu, C. and Moulines, E. (2003). Ergodicity of some adaptive markov chain monte carlo algorithm. *Technical Report* .

Andrieu, C., Moulines, E. and P., P. (to appear). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization* .

Andrieu, C. and Robert, C. P. (2002). Controlled mcmc for optimal sampling. *Technical report* .

Atchade, Y. F. (2004). Convergence of a class of stochastic approximations with application to mcmc. *In preparation* .

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic approximations*. Applications of Mathematics, Springer, Paris-New York.

Brockwell, A. E. and Kadane, J. B. (2002). Identification of regeneration times in mcmc simulation, with application to adaptive schemes. *Technical Report* .

Chen, H. and Zhu, Y.-M. (1986). Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica* **1** 914–926.

Davidson, J. (1994). *Stochastic limit theory*. Oxford University Press, Oxford, New-York.

Davidson, J. and de Jong, R. (1997). Strong laws of large numbers for dependent heteregeneous processes: a synthesis of recent and new results. *Econometric Reviews* **16** 251–279.

Delyon, B. (1996). General results on the convergence of stochastic algorithms. *IEEE Trans. Automat. Control* **41** 1245–1255.

Duflo, M. (1997). *Random Iterative Models*. Springer Verlag, Paris-New York.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds.) (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics, Chapman & Hall, London.

Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93** 1045–1054.

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 application*. Academic Press, New York.

JARNER, S. F. and HANSEN, E. (2000). Geometric ergodicity of metropolis algorithms. *Sto. Proc. Appl.* **85** 341–361.

KUSHNER, K. and YIN, Y. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer, Springer-Verlag, New-York.

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, New-York.

MEYN, S. P. and TWEEDIE, R. L. (1994). Computable bounds for convergence rates of markov chains. *Ann. Appl. Prob.* **4** 981–1011.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability*. Holden-Day.

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

ROBERTS, G. O., GELMAN, A. and GILKS, W. (1997). Weak convergence and optimal scaling of random walk metropolis algorithm. *Ann. Applied Prob.* **7** 110–120.

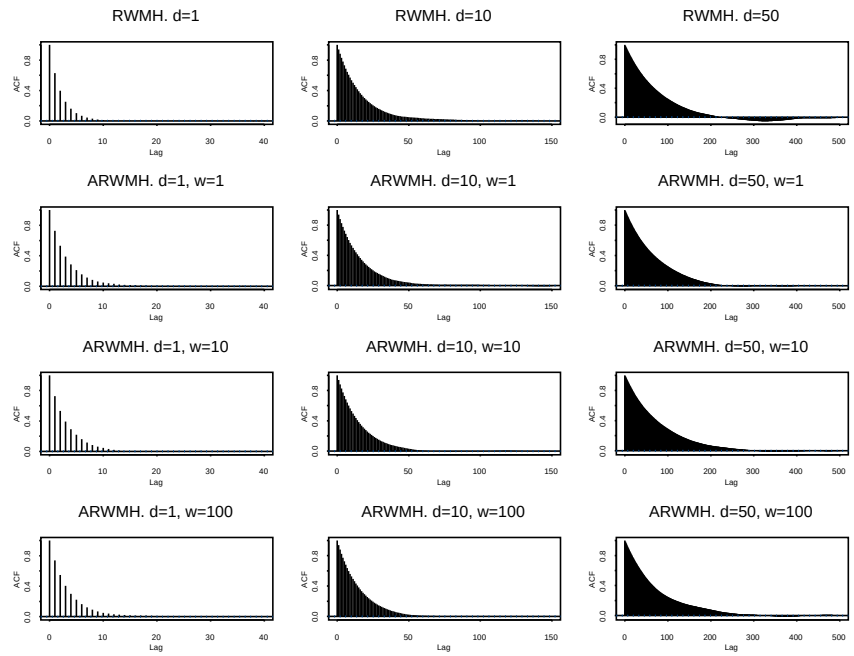
ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling of various metropolis-hastings algorithms. *Statistical Science* **16** 351–367.

ROSENTHAL, J. (2004). Adaptive mcmc java applet. *at: http://probability.ca/jeff/java/adapt.html*

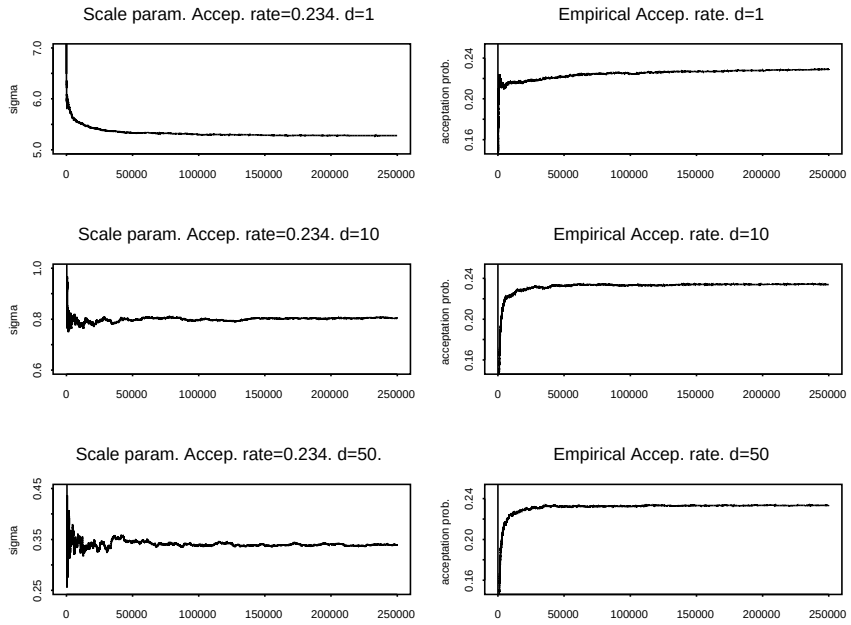
.

ROSENTHAL, J. S. (2002). Quantitative convergence rates of markov chains: a simple account. *Electronic Communications in Probability* **7** 123–128.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. With discussion and a rejoinder by the author.



Graph 1: Autocorrelations of ergodic averages of the function $f(x) = x_1$. Target density $N(0, I_d)$, proposal density $N(x, \sigma^2 I_d)$.



Graph 2: Scale parameter process and empirical acceptance probability for the ARWM with $w = 10$.