Bernoulli **7**(2), 2001, 223–242

An adaptive Metropolis algorithm

HEIKKI HAARIO[1*], EERO SAKSMAN[1**] and JOHANNA TAMMINEN[2]

[1]*Department of Mathematics, P.O. Box 4 (Yliopistonkatu 5), FIN-00014 University of Helsinki, Finland. E-mail: *heikki.haario@helsinki.fi; **eero.saksman@helsinki.fi.*
[2]*Finnish Meteorological Institute, Geophysics Research, P.O. Box 503, FIN-00101 Helsinki, Finland. E-mail: johanna.tamminen@fmi.fi*

A proper choice of a proposal distribution for Markov chain Monte Carlo methods, for example for the Metropolis–Hastings algorithm, is well known to be a crucial factor for the convergence of the algorithm. In this paper we introduce an adaptive Metropolis (AM) algorithm, where the Gaussian proposal distribution is updated along the process using the full information cumulated so far. Due to the adaptive nature of the process, the AM algorithm is non-Markovian, but we establish here that it has the correct ergodic properties. We also include the results of our numerical tests, which indicate that the AM algorithm competes well with traditional Metropolis–Hastings algorithms, and demonstrate that the AM algorithm is easy to use in practical computation.

Keywords: adaptive Markov chain Monte Carlo; comparison; convergence; ergodicity; Markov chain Monte Carlo; Metropolis–Hastings algorithm

1. Introduction

It is generally acknowledged that the choice of an effective proposal distribution for the random walk Metropolis algorithm, for example, is essential in order to obtain reasonable results by simulation in a limited amount of time. This choice concerns both the size and the spatial orientation of the proposal distribution, which are often very difficult to choose well since the target density is unknown (see Gelman *et al*. 1996; Gilks *et al*. 1995; 1998; Haario *et al*. 1999; Roberts *et al*. 1997). A possible remedy is provided by adaptive algorithms, which use the history of the process in order to 'tune' the proposal distribution suitably. This has previously been done (for instance) by assuming that the state space contains an atom. The adaptation is performed only at the times of recurrence to the atom in order to preserve the right ergodic properties (Gilks *et al*. 1998). The adaptation criteria are then obtained by monitoring the acceptance rate. A related and interesting self-regenerative version of adaptive Markov chain Monte Carlo (MCMC), based on introducing an auxiliary chain, is contained in the recent preprint of Sahu and Zhigljavsky (1999). For other versions of adaptive MCMC and related work, we refer to Evans (1991), Fishman (1996), Gelfand and Sahu (1994), Gilks and Roberts (1995) and Gilks *et al*. (1994), together with the references therein.

We introduce here an adaptive Metropolis (AM) algorithm which adapts continuously to the target distribution. Significantly, the adaptation affects both the size and the spatial orientation of the proposal distribution. Moreover, the new algorithm is straightforward to implement and use in practice. The definition of the AM algorithm is based on the classical

random walk Metropolis algorithm (Metropolis *et al*. 1953) and its modification, the AP algorithm, introduced in Haario *et al*. (1999). In the AP algorithm the proposal distribution is a Gaussian distribution centred on the current state, and the covariance is calculated from a fixed finite number of previous states. In the AM algorithm the covariance of the proposal distribution is calculated using *all* of the previous states. The method is easily implemented with no extra computational cost since one may apply a simple recursion formula for the covariances involved.

An important advantage of the AM algorithm is that it starts using the cumulating information right at the beginning of the simulation. The rapid start of the adaptation ensures that the search becomes more effective at an early stage of the simulation, which diminishes the number of function evaluations needed.

To be more exact, assume that at time t the already sampled states of the AM chain are $X_0, X_1, \ldots, X_t$, some of which may be multiple. The new proposal distribution for the next candidate point is then a Gaussian distribution with mean at the current point X_t and covariance given by $s_d R$, where R is the covariance matrix determined by the spatial distribution of the states $X_0, X_1, \ldots, X_t \in \mathbb{R}^d$. The scaling parameter s_d depends only on the dimension d of the vectors. This adaptation strategy forces the proposal distribution to approach an appropriately scaled Gaussian approximation of the target distribution, which increases the efficiency of the simulation. A more detailed description of the algorithm is given in Section 2 below.

One of the difficulties in constructing adaptive MCMC algorithms is to ensure that the algorithm maintains the correct ergodicity properties. We observe here (see also Haario *et al*. 1999) that the AP algorithm does not possess this property. Our main result, Theorem 1 below, verifies that *the AM process does indeed have the correct ergodicity properties*, assuming that the target density is bounded from above and has a bounded support. The AM chain is not Markovian, but we show that the asymptotic dependence between the elements of the chain is weak enough to apply known theorems of large numbers for mixingales – see McLeish (1975) and (29) below for this notion. Similar results may be proven also for variants of the algorithm, where the covariance is computed from a suitably increasing segment of the near history.

Section 3 contains a detailed description of the AM algorithm as a stochastic process and the theorem on the ergodicity of the AM. The proof is based on an auxiliary result that is proven in Section 4. Finally, in Section 5 we present results from test simulations, where the AM algorithm is compared with traditional Metropolis–Hastings algorithms (Hastings 1970) by applying both linear and nonlinear, correlated and uncorrelated unimodal target distributions. Our tests seem to imply that AM performs at least as well as the traditional algorithms for which a nearly optimal proposal distribution has been given a priori.

2. Description of the algorithm

We assume that our target distribution is supported on the subset $S \subset \mathbb{R}^d$, and that it has the (unscaled) density $\pi(x)$ with respect to the Lebesgue measure on S. With a slight abuse of notation, we shall also denote the target distribution by π.

We now explain how the AM algorithm works. Recall from Section 1 that the basic idea is to update the proposal distribution by using the knowledge we have so far acquired about the target distribution. Otherwise the definition of the algorithm is identical to the usual Metropolis process.

Suppose, therefore, that at time $t-1$ we have sampled the states $X_0, X_1, \ldots, X_{t-1}$, where X_0 is the initial state. Then a candidate point Y is sampled from the (asymptotically symmetric) proposal distribution $q_t(\cdot|X_0, \ldots, X_{t-1})$, which now may depend on the whole history $(X_0, X_1, \ldots, X_{t-1})$. The candidate point Y is accepted with probability

$$\alpha(X_{t-1}, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X_{t-1})}\right),$$

in which case we set $X_t = Y$, and otherwise $X_t = X_{t-1}$. Observe that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm. However, here the choice for the acceptance probability is not based on symmetry (reversibility) conditions since these cannot be satisfied in our case – the corresponding stochastic chain is no longer Markovian. For this reason we have to study the exactness of the simulation separately, and we do so in Section 3.

The proposal distribution $q_t(\cdot|X_0, \ldots, X_{t-1})$ employed in the AM algorithm is a Gaussian distribution with mean at the current point X_{t-1} and covariance $C_t = C_t(X_0, \ldots, X_{t-1})$. Note that in the simulation only jumps into S are accepted since we assume that the target distribution vanishes outside S.

The crucial thing regarding the adaptation is how the covariance of the proposal distribution depends on the history of the chain. In the AM algorithm this is solved by setting $C_t = s_d \operatorname{cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d$ after an initial period, where s_d is a parameter that depends only on dimension d and $\varepsilon > 0$ is a constant that we may choose very small compared to the size of S. Here I_d denotes the d-dimensional identity matrix. In order to start, we select an arbitrary, strictly positive definite, initial covariance C_0, according to our best prior knowledge (which may be quite poor). We select an index $t_0 > 0$ for the length of an initial period and define

$$C_t = \begin{cases} C_0, & t \leqslant t_0, \\ s_d \operatorname{cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d, & t > t_0. \end{cases} \tag{1}$$

The covariance C_t may be viewed as a function of t variables from $\mathbb{R}^d$ having values in uniformly positive definite matrices.

Recall the definition of the empirical covariance matrix determined by points $x_0, \ldots, x_k \in \mathbb{R}^d$:

$$\operatorname{cov}(x_0, \ldots, x_k) = \frac{1}{k}\left(\sum_{i=0}^{k} x_i x_i^{\mathrm{T}} - (k+1)\bar{x}_k \bar{x}_k^{\mathrm{T}}\right), \tag{2}$$

where $\bar{x}_k = (1/(k+1))\sum_{i=0}^{k} x_i$ and the elements $x_i \in \mathbb{R}^d$ are considered as column vectors. So one obtains that in definition (1) for $t \geqslant t_0 + 1$ the covariance C_t satisfies the recursion formula

$$C_{t+1} = \frac{t-1}{t} C_t + \frac{s_d}{t}(t\overline{X}_{t-1}\overline{X}_{t-1}^{\mathrm{T}} - (t+1)\overline{X}_t\overline{X}_t^{\mathrm{T}} + X_tX_t^{\mathrm{T}} + \varepsilon I_d). \tag{3}$$

This allows one to calculate C_t without too much computational cost since the mean $\overline{X}_t$ also satisfies an obvious recursion formula.

The choice for the length of the initial segment $t_0 > 0$ is free, but the bigger it is chosen the more slowly the effect of the adaptation is felt. In a sense the size of t_0 reflects our trust in the initial covariance C_0. The role of the parameter ε is just to ensure that C_t will not become singular (see Remark 1 below). As a basic choice for the scaling parameter we have adopted the value $s_d = (2.4)^2/d$ from Gelman *et al.* (1996), where it was shown that in a certain sense this choice optimizes the mixing properties of the Metropolis search in the case of Gaussian targets and Gaussian proposals.

Remark 1. In our test runs the covariance C_t has not had the tendency to degenerate. This has also been the case in our multimodal test examples. However, potential difficulties with $\varepsilon = 0$ (if any) are more likely to appear in the multimodal cases. In practical computations one presumably may utilize definition (1) with $\varepsilon = 0$, although the change is negligible if ε has already been chosen small enough. More importantly, we can prove the correct ergodicity property of the algorithm only under the assumption $\varepsilon > 0$; see Theorem 1 below.

Remark 2. In order to avoid the algorithm starting slowly it is possible to employ special tricks. Naturally, if a priori knowledge (such as the maximum likelihood value or approximate covariance of the target distribution) is available, it can be utilized in choosing the initial state or the initial covariance C_0. Also, in some cases it is advisable to employ the *greedy start* procedure: during a short initial period one updates the proposal using only the accepted states. Afterwards the AM is run as described above. Moreover, during the early stage of the algorithm it is natural to require it to move at least a little. If it has not moved enough in the course of a certain number of iterations, the proposal distribution could be shrunk by some constant factor.

Remark 3. It is also possible to choose an integer $n_0 > 1$ and update the covariance every n_0th step only (again using the entire history). This saves computer time when generating the candidate points. There is again a simple recursion formula for the covariances C_t.

3. Ergodicity of the AM chain

In the AP algorithm, which was briefly described in Section 1, the covariance C_t was calculated from the last H states only, where $H \geqslant 2$. This strategy has the undesirable consequence of bringing non-exactness into the simulation. There are several ways to see this. One may, for instance, study the Markov chain consisting of H-tuples of consecutive variables of the AP chain, to obtain the limit distribution for the AP by a suitable projection from the equilibrium distribution of this Markov chain. Simple examples in the case of finite state space for an analogous model show that the limiting distribution of the AP algorithm

differs slightly from the target distribution. Numerical calculations in the continuous case indicate similar behaviour. An illustrating example of this phenomenon is presented in the Appendix.

It is our aim in this section to show that the AM algorithm has the right ergodic properties and hence provides correct simulation of the target distribution.

Let us start by recalling some basic notions of the theory of stochastic processes that are needed later. We first define the set-up. Let $(S, \mathscr{B}, m)$ be a state space and denote by $\mathscr{M}(S)$ the set of finite measures on $(S, \mathscr{B})$. The norm $\|\cdot\|$ on $\mathscr{M}(S)$ denotes the total variation norm. Let $n \geqslant 1$ be a natural number. A map $K_n : S^n \times \mathscr{B} \to [0, 1]$ is a *generalized* transition probability on the set S if the map $x \mapsto K_n(x; A)$ is $\mathscr{B}^n$-measurable for each $A \subset \mathscr{B}$, where $x \in S^n$ and $K_n(x; \cdot)$ is a probability measure on $(S, \mathscr{B})$ for each $x \in S^n$. In a natural way K_n defines a positive contraction from $\mathscr{M}(S^n)$ into $\mathscr{M}(S)$. A transition probability on S corresponds to the case $n = 1$ in the above definition.

We assume that a sequence of generalized transition probabilities $(K_n)_{n=1}^{\infty}$ is given. Moreover, let μ_0 be a probability distribution (the *initial distribution*) on S. Then the sequence (K_n) and μ_0 determine uniquely the finite-dimensional distributions of the discrete-time stochastic process (chain) $(X_n)_{n=0}^{\infty}$ on S via the formula

$$P(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \int_{y_0 \in A_0} \mu_0(\mathrm{d}y_0) \left(\int_{y_1 \in A_1} K_1(y_0; \mathrm{d}y_1) \right.$$

$$\left. \times \left(\int_{y_2 \in A_2} K_2(y_0, y_1; \mathrm{d}y_2) \cdots \left(\int_{y_n \in A_n} K_n(y_0, y_1, \ldots, y_{n-1}; \mathrm{d}y_n) \right) \ldots \right) \right). \tag{4}$$

In fact, it is directly verified that these distributions are consistent and the theorem of Ionescu Tulcea (see Proposition V.1.1 of Neveu 1965) yields the existence of the chain (X_n) on S satisfying (4).

We shall now turn to the exact definition of the AM chain as a discrete-time stochastic process. We assume that *the target distribution is supported on a bounded subset* $S \subset \mathbb{R}^d$, so that $\pi(x) \equiv 0$ outside S. Thus we shall choose S to be our state space, when equipped with the Borel σ-algebra $\mathscr{B}(S)$ and choosing m to be the normalized Lebesgue measure on S. The target π has the (unscaled) density $\pi(x)$ with respect to the Lebesgue measure on S. We also assume that *the density is bounded from above on* S: for some $M < \infty$, we have that

$$\pi(x) \leqslant M \qquad \text{for } x \in S. \tag{5}$$

Let C be a symmetric and strictly positive definite matrix on $\mathbb{R}^d$ and denote by N_C the density of the mean-zero Gaussian distribution on $\mathbb{R}^n$ with covariance C. Thus

$$N_C(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|C|}} \exp\left(-\frac{1}{2} x^{\mathrm{T}} C^{-1} x \right). \tag{6}$$

The Gaussian proposal transition probability corresponding to the covariance C satisfies

$$Q_C(x;\, A) = \int_A N_C(y - x)\, \mathrm{d}y, \tag{7}$$

where $A \subset \mathbb{R}^d$ is a Borel set and $\mathrm{d}y$ is the standard Lebesgue measure on $\mathbb{R}^d$. It follows that Q_C is m-symmetric (see Haario and Saksman 1991, Definition 2.2): for $A,\ B \subset S$ one has

$$\int_B Q_C(x;\, A) m(\mathrm{d}x) = \int_A Q_C(x;\, B) m(\mathrm{d}x).$$

We next recall the definition of the transition probability M_C for the Metropolis process having the target density $\pi(x)$ and the proposal distribution Q_C:

$$M_C(x;\, A) = \int_A N_C(y - x) \min\left(1, \frac{\pi(y)}{\pi(x)}\right) m(\mathrm{d}y) + \chi_A(x) \int_{\mathbb{R}^d} N_C(y - x) \left[1 - \min\left(1, \frac{\pi(y)}{\pi(x)}\right)\right] m(\mathrm{d}y), \tag{8}$$

for $A \in \mathscr{B}(S)$, and where χ_A denotes the characteristic function of the set A. It is easily verified that M_C defines a transition probability with state space S.

The following definition of the AM chain corresponds exactly to the AM algorithm introduced in Section 2.

Definition 1. *Let S and π be as above and let the initial covariance C_0 and the constant $\varepsilon > 0$ be given. Define the functions C_n for $n \geqslant 1$ by formula* (1). *For a given initial distribution μ_0 the adaptive Metropolis (AM) chain is a stochastic chain on S defined through* (4) *by the sequence $(K_n)_{n=1}^{\infty}$ of generalized transition probabilities, where*

$$K_n(x_0, \ldots, x_{n-1};\, A) = M_{C_n(x_0, \ldots, x_{n-1})}(x_{n-1};\, A) \tag{9}$$

for all $n \geqslant 1$, $x_i \in S$ $(0 \leqslant i \leqslant n - 1)$, and for subsets $A \in \mathscr{B}(S)$.

Let us turn to the study of the ergodicity properties of the AM chain, which is more complicated than in the case of Markov chains. In order to be able to proceed we give some definitions. Recall first the definition of the coefficient of ergodicity (Dobrushin 1956). Let T be a transition probability on S and set

$$\delta(T) = \sup_{\mu_1, \mu_2} \frac{\|\mu_1 T - \mu_2 T\|}{\|\mu_1 - \mu_2\|}, \tag{10}$$

where the supremum is taken over distinct probability measures μ_1, μ_2 on $(S, \mathscr{B})$. As usual, λT denotes the measure $A \mapsto \int_S T(x;\, A) \lambda(\mathrm{d}x)$ and for bounded measurable functions we write $Tf(x) = \int_S T(x;\, \mathrm{d}y) f(y)$ as well as $\lambda f = \int_S \lambda(\mathrm{d}y) f(y)$.

Clearly $0 \leqslant \delta(T) \leqslant 1$. In the case $\delta(T) < 1$ the mapping T is a strict contraction on $\mathscr{M}(S)$ with respect to the metric defined by the total variation norm on $\mathscr{M}(S)$. From the definition it easily follows that

$$\delta(T_1 T_2 \ldots T_n) \leqslant \prod_{i=1}^{n} \delta(T_i). \tag{11}$$

The condition $\delta(T^{k_0}) < 1$ for some $k_0 \geqslant 1$ is well known to be equivalent to the uniform ergodicity (cf. Nummelin 1984, Section 6.6) of the Markov chain having transition probability T.

For our purposes it is useful to define the transition probability that is obtained from a generalized transition probability by 'freezing' the $n-1$ first variables. Hence, given a generalized transition probability K_n (where $n \geqslant 2$) and a *fixed* $(n-1)$-tuple $(y_0, y_1, \ldots, y_{n-2}) \in S^{n-1}$, we denote $\tilde{y}_{n-2} = (y_0, y_1, \ldots, y_{n-2})$ and define the transition probability $K_{n,\tilde{y}_{n-2}}$ by

$$K_{n,\tilde{y}_{n-2}}(x; A) = K_n(y_0, y_1, \ldots, y_{n-2}, x; A) \tag{12}$$

for $x \in S$ and $A \in \mathscr{B}(S)$.

We are now ready to state and prove our main theorem. The role of the assumptions on the target density is commented on in Remark 7 below.

Theorem 1. *Let π be the density of a target distribution supported on a bounded measurable subset $S \subset \mathbb{R}^d$, and assume that π is bounded from above. Let $\varepsilon > 0$ and let μ_0 be any initial distribution on S. Define the AM chain (X_n) by the generalized transition probabilities* (9) *as in Definition* 1. *Then the AM chain simulates properly the target distribution π: for any bounded and measurable function $f : S \to \mathbb{R}$, the equality*

$$\lim_{n\to\infty} \frac{1}{n+1}(f(X_0) + f(X_1) + \ldots + f(X_n)) = \int_S f(x)\pi(\mathrm{d}x).$$

holds almost surely.

The proof of the theorem is based on the following technical auxiliary result, whose proof we postpone to the next section.

Theorem 2. *Assume that the finite-dimensional distributions of the stochastic process $(X_n)_{n=0}^{\infty}$ on the state space S satisfy* (4)*, where the sequence of generalized transition probabilities (K_n) is assumed to satisfy the following three conditions:*

(i) *There are a fixed integer k_0 and a constant $\lambda \in (0, 1)$ such that*

$$\delta((K_{n,\tilde{y}_{n-2}})^{k_0}) \leqslant \lambda < 1 \qquad \text{for all } \tilde{y}_{n-2} \in S^{n-1} \text{ and } n \geqslant 2.$$

(ii) *There are a fixed probability measure π on S and a constant $c_0 > 0$ such that*

$$\|\pi K_{n,\tilde{y}_{n-2}} - \pi\| \leqslant \frac{c_0}{n} \qquad \text{for all } \tilde{y}_{n-2} \in S^{n-1} \text{ and } n \geqslant 2.$$

(iii) *We have the following estimate for the operator norm*

$$\|K_{n,\tilde{y}_{n-2}} - K_{n+k,\tilde{y}_{n+k-2}}\|_{\mathscr{M}(S)\to\mathscr{M}(S)} \leqslant c_1 \frac{k}{n},$$

where c_1 is a fixed positive constant, $n, k \geqslant 1$ and one assumes that the $(n+k-1)$-tuple $\tilde{y}_{n+k-2}$ is a direct continuation of the $(n-1)$-tuple $\tilde{y}_{n-2}$.

Then, if $f : S \to \mathbb{R}$ is bounded and measurable, then the equality

$$\lim_{n\to\infty} \frac{1}{n+1}(f(X_0) + f(X_1) + \ldots + f(X_n)) = \int_S f(x)\pi(\mathrm{d}x). \tag{13}$$

holds almost surely.

In what follows the auxiliary constants c_i, $i = 2, 3, \ldots,$ depend on S, ε or C_0, and their actual value is irrelevant for our purposes here.

Proof of Theorem 1. According to Theorem 2 it suffices to prove that the AM chain satisfies conditions (i)–(iii). In order to check condition (i) we observe that, directly from definition (1) and by the fact that S is bounded, all the covariances $C = C_n(y_0, \ldots, y_{n-1})$ satisfy the matrix inequality

$$0 < c_2 I_d \leqslant C \leqslant c_3 I_d. \tag{14}$$

Hence the corresponding normal densities $N_C(\cdot - x)$ are uniformly bounded from below on S for all $x \in S$, and (5) and (8) together trivially yield the bound

$$K_{n,\tilde{y}_{n-2}}(x;\, A) \geqslant c_4\pi(A) \qquad \text{for all } x \in S \text{ and } A \subset S,$$

with $c_4 > 0$. This easily yields (cf. Nummelin 1984, pp. 122–123) that $\delta(K_{n,\tilde{y}_{n-2}}) \leqslant 1 - c_4$, which proves (i) with $k_0 = 1$.

We next verify condition (iii). To that end we assume that $n \geqslant 2$ and observe that, for given $\tilde{y}_{n+k-2} \in S^{n+k-1}$, one has

$$\|K_{n,\tilde{y}_{n-2}} - K_{n+k,\tilde{y}_{n+k-2}}\|_{\mathscr{M}(S)\to\mathscr{M}(S)} \leqslant 2\sup_{y\in S, A\in\mathscr{B}(S)} |K_{n,\tilde{y}_{n-2}}(y;\, A) - K_{n+k,\tilde{y}_{n+k-2}}(y;\, A)|. \tag{15}$$

Fix $y \in S$ and $A \in \mathscr{B}(S)$ and introduce $R_1 = C_n(y_0, \ldots, y_{n-2}, y)$ together with $R_2 = C_{n+k}(y_0, \ldots, y_{n+k-2}, y)$. According to Definition 1 and formula (8), we see that

$$\begin{aligned}
|K_{n,\tilde{y}_{n-2}}(y;\, A) - K_{n+k,\tilde{y}_{n+k-2}}(y;\, A)| &= |M_{R_1}(y;\, A) - M_{R_2}(y;\, A)| \\
&\leqslant \left| \int_{x\in A} (N_{R_1} - N_{R_2})(x - y) \min\left(1, \frac{\pi(x)}{\pi(y)}\right) m(\mathrm{d}x) \right. \\
&\quad + \chi_A(x) \int_{x\in\mathbb{R}^d} (N_{R_1} - N_{R_2})(x - y) \\
&\quad \left. \times \left[1 - \min\left(1, \frac{\pi(x)}{\pi(y)}\right)\right] m(\mathrm{d}x) \right| \\
&\leqslant 2\int_{\mathbb{R}^d} |N_{R_1}(z) - N_{R_2}(z)|\, \mathrm{d}z \\
&\leqslant 2\int_{\mathbb{R}^d} \mathrm{d}z \int_0^1 \mathrm{d}s \left| \frac{\mathrm{d}}{\mathrm{d}s} N_{R_1 + s(R_2 - R_1)}(z) \right| \\
&\leqslant c_5 \|R_1 - R_2\|,
\end{aligned} \tag{16}$$

where at the last stage we apply (14), in order to deduce that the partial derivatives of the density $N_{R_1+s(R_2-R_1)}$ with respect to the components of the covariance are integrable over $\mathbb{R}^d$ with bounds that depend only on ε, C_0 and S. Finally, it is clear from recursion formula (3) that in general $\|C_t - C_{t+1}\| \leq c_6/t$ for $t > 1$. By applying this inductively and using the uniform boundedness from above of the covariances C_t, we easily see that

$$\|R_1 - R_2\| \leq c_7(S, C_0, \varepsilon)\frac{k}{n},$$

and hence the previous estimates yield (iii).

In order to check condition (ii), fix $\tilde{y}_{n-2} \in S^{n-1}$ and denote $C^* = C_{n-1}(y_0, \ldots y_{n-2})$. It follows that $\|C^* - C_n(y_0, \ldots, y_{n-2}, y)\| \leq c_8/n$, where c_8 does not depend on $y \in S$. We may therefore proceed exactly as in (15) and (16) to deduce that

$$\|K_{n,\tilde{y}_{n-2}} - M_{C^*}\|_{\mathcal{M}(S)\to\mathcal{M}(S)} \leq \frac{c_9}{n}.$$

Since M_{C^*} is a Metropolis transition probability we have that $\pi M_{C^*} = \pi$ (see e.g. Tierney 1994, p. 1705), and we obtain

$$\|\pi - \pi K_{n,\tilde{y}_{n-2}}\| = \|\pi(M_{C^*} - K_{n,\tilde{y}_{n-2}})\| \leq \frac{c_9}{n},$$

which completes the proof of Theorem 1. □

Let us record an expected result on the behaviour of the AM chain.

Corollory 3. *Under the assumptions of Theorem* 1 *the covariance C_t almost surely stabilizes during the algorithm. In fact, as $t \to \infty$ the covariance C_t converges to $s_d \operatorname{cov}(\pi) + \varepsilon I_d$, where* $\operatorname{cov}(\pi)$ *denotes the covariance of the target distribution π.*

Proof. The claim follows directly from the definition (1) of the covariance C_t by applying Theorem 1 with the choices $f(x) = x_i$ and $f(x) = x_i x_j$, where $1 \leq i, j \leq d$. □

We conclude this section with a number of comments on the theory presented above.

Remark 4. Our decision to use Gaussian proposal distributions is based on their tested practical applicability, even in the case of non-Gaussian targets. Gaussian proposals yield a family of proposal distributions with a natural parametrization for size and orientation and which are easy to compute with. However, in the definition of the AM chain one can easily replace the Gaussian proposals by, for example, uniform distributions in a parallelepiped. In this case the size and the orientation of the parallelepiped are guided in a natural manner by the covariance C_t that is determined by (1) as above. Our proof of Theorem 1 remains unchanged and we again obtain that the simulation is exact. The only difference is that the constant k_0 in condition (i) of Theorem 2 may now exceed 1. Naturally, here one has to add suitable assumptions on the set $A = \{x : \pi(x) > 0\}$. It is, for example, enough to assume that A is open and connected. In this connection the estimates provided by Haario and Saksman (1991, Theorem 6.5.(b)) are relevant.

Similar remarks apply to modifications where one adapts only certain parameters or some of the parameters are discrete.

Remark 5. It is clear that in the course of the AM algorithm one may also determine the covariance by using only an increasing part of the near history. For example, one may determine C_n by using only the samples $X_{[n/2]}, X_{[n/2]+1}, \ldots, X_n$. This is easily implemented in practice and in this case Theorem 1 yields that the simulation is exact with only minor changes in the proof. Similar remarks apply also to the case where one updates the covariance only every n_0th step (see Remarks 3 and 8).

Remark 6. Theorem 2 can also be used to prove the correct ergodicity for certain other variants of adaptation, as for algorithms where one suitably tunes the proposal distribution according to the acceptance rate. However, in our specific practical applications it has turned out that the tuning of the acceptance rate has yielded inferior results when compared with the AM algorithm. A similar phenomenon is demonstrated in Figure 2 below. Moreover, in high-dimensional cases with possible correlations between the parameters, it may be difficult to tune the proposal distribution effectively basing the decision on one parameter only. This is the case even if one uses the single-component Metropolis algorithm.

Remark 7. The proof of Theorem 1 requires that the target density has compact support and is bounded from above. Otherwise the uniform ergodicity (condition (i) of Theorem 2) may fail, which is important if we are to be able to control the effects of the adaptation. In the Markovian case (for example, standard Metropolis–Hastings) uniform ergodicity is, of course, not needed to ensure that the simulation is correct, although without it the theoretical convergence rate may be very slow. However, the requirements above on the target density correspond reasonably well to practical situations. We believe that one may weaken the assumptions at the cost of more elaborate proofs. We prefer to leave this topic for future research, since our main aim here is to introduce a new method and to demonstrate its usefulness. In our test runs the AM algorithm has also worked successfully with (unrestricted) Gaussian targets.

4. Proof of Theorem 2

In this section we will prove Theorem 2 by showing that a related process is a mixingale (in the sense of McLeish 1975) that satisfies an appropriate law of large numbers. The conditions of the theorem were tailored to apply to the AM chain on bounded subsets of $\mathbb{R}^n$, but they are stated in the language of a general state space. This is advantageous since one may apply them in a more general situation, especially for variants of the AM where the state space contains both discrete and continuous parts. Our proof is based on the following basic proposition.

Proposition 4. *Let the chain* (X_n) *on the state space* S *and the generalized transitition probabilities* (K_n) *fulfil the conditions of Theorem* 2. *Denote by* $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$

the σ-algebra generated by the chain up to time n and write $\lambda' = \lambda^{1/k_0}$. *Let* $n \geqslant 1$ *and* $k \geqslant 2$. *Then for all initial distributions and for any bounded measurable function f on S, the inequality*

$$\left\| \mathrm{E}(f(X_{n+k})|\mathscr{F}_n) - \int_S f(y)\pi(\mathrm{d}y) \right\|_\infty \leqslant c(c_0, c_1, \lambda) \inf_{1 \leqslant j \leqslant k} \left(\frac{j^2}{n+k-j} + \lambda'^{\,j} \right) \|f\|_\infty \tag{17}$$

holds.

Proof. We may clearly assume that $\pi f = \int_S f(y)\pi(\mathrm{d}y) = 0$ since the general case is then obtained by applying the proposition to the function $f - \pi f$. Let $n \geqslant 1$, $k \geqslant 2$ and note that from the definition of the conditional expectation and (4) it follows that (almost surely)

$$\begin{aligned} &\mathrm{E}(f(X_{n+k})|\mathscr{F}_n) \\ &= \int_{y_{n+1} \in S} K_{n+1}(X_0, X_2, \ldots, X_n; \mathrm{d}y_{n+1}) \left(\int_{y_{n+2} \in S} K_{n+2}(X_0, X_2, \ldots, X_n, y_{n+1}; \mathrm{d}y_{n+2}) \right. \\ &\quad \left. \cdots \left(\int_{y_{n+k} \in S} K_{n+k}(X_0, X_2, \ldots, X_n, y_{n+1}, \ldots, y_{n+k-1}; \mathrm{d}y_{n+k}) f(y_{n+k}) \right) \cdots \right). \end{aligned} \tag{18}$$

Let us denote $(X_0, \ldots, X_n) = \tilde{X}_n$. In what follows $\tilde{X}_n$ does not interfere with the integrations and hence it may be thought as a free variable (or constant). We also introduce the transition probability Q, where $Q(y; \mathrm{d}z) = K_{n+2}(\tilde{X}_n, y; \mathrm{d}z)$. Condition (iii) yields for arbitrary values of $\tilde{X}_n$ and $y_{n+1}, \ldots, y_{n+k-1}$ that

$$\begin{aligned} &\left| \int_{y_{n+k} \in S} (K_{n+k}(\tilde{X}_n, y_{n+1}, \ldots, y_{n+k-1}; \mathrm{d}y_{n+k}) - K_{n+2}(\tilde{X}_n, y_{n+k+1}; \mathrm{d}y_{n+k})) f(y_{n+k}) \right| \\ &\qquad \leqslant c_1 \|f\|_\infty \frac{k-2}{n+2}. \end{aligned} \tag{19}$$

This estimate enables us to write (18) in the form

$$\begin{aligned} \mathrm{E}(f(X_{n+k})|\mathscr{F}_n) = g_k(\tilde{X}_n) + \int_{y_{n+1} \in S} K_{n+1}(\tilde{X}_n; \mathrm{d}y_{n+1}) \left(\int_{y_{n+2} \in S} K_{n+2}(\tilde{X}_n, y_{n+1}; \mathrm{d}y_{n+2}) \right. \\ \cdots \left(\int_{y_{n+k-1} \in S} K_{n+k-1}(\tilde{X}_n, y_{n+1}, \ldots, y_{n+k-2}; \mathrm{d}y_{n+k-1}) \right. \\ \left. \left. \cdots \left(\int_{y_{n+k} \in S} K_{n+2}(\tilde{X}_n, y_{n+k-1}; \mathrm{d}y_{n+k}) f(y_{n+k}) \right) \cdots \right) \right), \end{aligned} \tag{20}$$

where $g_k = g_k(\tilde{X}_n)$ satisfies

$$|g_k(\tilde{X}_n)| \leqslant \int_{y_{n+1}\in S} K_{n+1}(\tilde{X}_n;\, \mathrm{d}y_{n+1})\Bigg(\int_{y_{n+2}\in S} K_{n+2}(\tilde{X}_n,\, y_{n+1};\, \mathrm{d}y_{n+2})$$

$$\cdots \left(\int_{y_{n+k-1}\in S} K_{n+k-1}(\tilde{X}_n,\, y_{n+1},\, \ldots,\, y_{n+k-2};\, \mathrm{d}y_{n+k-1}) f(y_{n+k}) c_1 \|f\|_\infty \frac{k-2}{n+2}\right) \cdots \Bigg)$$

$$\leqslant c_1 \|f\|_\infty \frac{k-2}{n+2}.$$

In the next step we iterate the procedure by replacing the generalized transition probability $K_{n+k-1}(\tilde{X}_n, y_{n+1}, \ldots, y_{n+k-2};\, \mathrm{d}y_{n+k-1})$ by the transition probability Q in formula (20). By continuing in this manner we obtain

$$\mathrm{E}(f(X_{n+k})|\mathscr{F}_n) = \int_{y_{n+1}\in S} K_{n+1}(\tilde{X}_n;\, \mathrm{d}y_{n+1})\Bigg(\int_{y_{n+2}\in S} Q(y_{n+1};\, \mathrm{d}y_{n+2})$$

$$\cdots \left(\int_{y_{n+k}\in S} Q(y_{n+k-1};\, \mathrm{d}y_{n+k}) f(y_{n+k})\right) \cdots \Bigg)$$

$$+\, g_2(\tilde{X}_n) + g_3(\tilde{X}_n) + \cdots + g_k(\tilde{X}_n),$$

where

$$g_j(\tilde{X}_n) = \int_{y_{n+1}\in S} K_{n+1}(\tilde{X}_n;\, \mathrm{d}y_{n+1})\Bigg(\int_{y_{n+2}\in S} K_{n+2}(\tilde{X}_n,\, y_{n+1};\, \mathrm{d}y_{n+2})$$

$$\cdots \left(\int_{y_{n+j}\in S} (K_{n+j}(\tilde{X}_n,\, y_{n+1},\, \ldots,\, y_{n+j-1};\, \mathrm{d}y_{n+j})\right.$$

$$\left. - K_{n+2}(\tilde{X}_n,\, y_{n+j-1};\, \mathrm{d}y_{n+j}))Q^{k-j} f(y_{n+j})\right) \cdots \Bigg). \tag{21}$$

Recall here that Q^{k-j} denotes the $(k-j)$th iterate of the transition probability Q and we apply the standard notation $(Q^{k-j}f)(x) = \int_S Q^{k-j}(x;\, \mathrm{d}y) f(y)$.

Since $\|Q^{k-j}f\|_\infty \leqslant \|f\|_\infty$ we obtain as before from condition (iii) that

$$|g_j| \leqslant c_1 \frac{j-2}{n+2} \|f\|_\infty.$$

Summing up, we have shown that

$$\mathrm{E}(f(X_{n+k})|\mathscr{F}_n) = \varepsilon_{n,k} + \int_{y_{n+1}\in S} K_{n+1}(X_0, \ldots, X_n,\, \mathrm{d}y_{n+1}) Q^{k-1} f(y_{n+k}), \tag{22}$$

where $\varepsilon_{n,k} = \varepsilon_{n,k}(X_0, \ldots, X_n)$ satisfies

$$|\varepsilon_{n,k}| \leqslant \sum_{j=2}^{k} c_1 \frac{j-2}{n+2} \|f\|_\infty \leqslant \frac{c_1 k^2}{n} \|f\|_\infty. \tag{23}$$

Next write $[(k-1)/k_0] = k'$ and notice that $\delta(Q^{k-1}) \leqslant \lambda^{k'}$ according to (i). By (ii) and the definition of Q, we have

$$\|\pi Q^{k-1} - \pi\| \leqslant \sum_{j=0}^{k-2} \|\pi Q^{j+1} - \pi Q^j\| \leqslant \sum_{j=0}^{k-2} \frac{c_0}{n+2} \leqslant \frac{c_0(k-1)}{n+2},$$

and hence, using the assumption $\pi f = 0$, we may estimate

$$\begin{aligned} \|Q^{k-1} f\|_\infty &= \sup_{x \in S} |\delta_x Q^{k-1} f| \leqslant \sup_{x \in S} |(\delta_x - \pi) Q^{k-1} f| + |\pi Q^{k-1} f| \\ &\leqslant 2\lambda^{k'} \|f\|_\infty + |(\pi Q^{k-1} - \pi) f| \leqslant \left(\frac{c_0(k-1)}{n+2} + 2\lambda^{k'} \right) \|f\|_\infty. \end{aligned} \tag{24}$$

Combining this with (22) and (23), it follows that

$$\|\mathrm{E}(f(X_{n+k}) | \mathscr{F}_n)\|_\infty \leqslant \tilde{c}(c_0, c_1, \lambda) \left(\frac{k^2}{n} + \lambda^{[(k-1)/k_0]} \right) \|f\|_\infty, \tag{25}$$

which is valid for all $n,\ k \geqslant 2$.

In order to deduce the proposition, we first observe that for any index j between 1 and k the standard properties of the conditional expectation yield that

$$\|\mathrm{E}(f(X_{n+k}) | \mathscr{F}_n)\|_\infty \leqslant \|\mathrm{E}(f(X_{n+k}) | \mathscr{F}_{n+k-j})\|_\infty.$$

Hence, by replacing n by $n+k-j$ and k by j in the estimate (25), we finally deduce that

$$\|\mathrm{E}(f(X_{n+k}) | \mathscr{F}_n)\|_\infty \leqslant \inf_{1 \leqslant j \leqslant k} \tilde{c}(c_0, c_1, \lambda) \left(\frac{j^2}{n+k-j} + \lambda^{[(j-1)/k_0]} \right) \|f\|_\infty. \tag{26}$$

The claim of the proposition follows immediately from this estimate. □

Proof of Theorem 2. From Proposition 4 we obtain, for $n \geqslant 1$ and $k \geqslant 0$, that

$$\|\mathrm{E}(f(X_{n+k}) - \int_S f(y) \pi(\mathrm{d}y) | \mathscr{F}_n)\|_\infty \leqslant \psi(k), \tag{27}$$

where $\psi(0) = \psi(1) = 2\|f\|_\infty$, and for $k \geqslant 2$ we have

$$\psi(k) \equiv c(c_0, c_1, \lambda) \inf_{1 \leqslant j \leqslant k} \left(\frac{j^2}{k-j} + \lambda'^j \right) \|f\|_\infty \leqslant c'(c_0, c_1, f, \lambda) \frac{\log^2 k}{k}, \tag{28}$$

where the last estimate is obtained by choosing $j = \log k / \log(1/\lambda')$ for $k \geqslant k_1(\lambda')$.

At this stage the estimate (28) for the asymptotic independence, together with the definition of the σ-algebra $\mathscr{F}_n$, makes it clear that $f(X_n) - \mathrm{E}f(X_n)$ is a mixingale in the sense of McLeish – see McLeish (1975) or Hall and Heyde (1980, p. 19). For the convenience of the reader, let us recall here the definition of mixingales. Let $(\mathscr{F}_n)_{n=-\infty}^{\infty}$ be

an increasing sequence of sub-σ-algebras on a probability space. A sequence $(Y_n)_{n=1}^{\infty}$ of square-integrable random variables is a mixingale (difference) sequence if there are real sequences $(r_m)_{m=0}^{\infty}$ and $(a_n)_{n=1}^{\infty}$ such that $r_m \to 0$ as $m \to \infty$, and

$$\|\mathrm{E}(Y_n|\mathscr{F}_{n-m})\|_2 \leqslant r_m a_n \quad \text{and} \quad \|Y_n - \mathrm{E}(Y_n|\mathscr{F}_{n+m})\|_2 \leqslant r_{m+1} a_n \tag{29}$$

for all $n \geqslant 1$ and $m \geqslant 0$. In our case, where $Y_n = f(X_n) - \mathrm{E}f(X_n)$, we take (a_n) to be a constant sequence and let $\mathscr{F}_n$ be the trivial σ-algebra for $n < 0$. The right-hand side condition in (29) is automatically satisfied. Moreover, we may choose $r_k = \psi(k)$, and it follows that $r_k \leqslant C(\varepsilon)k^{\varepsilon-1}$ for every $\varepsilon > 0$. Hence we may apply directly the well-known laws of large numbers for mixingales in the form of Hall and Heyde (1980, Theorem 2.21, p. 41) to the sequence $f(X_n) - \mathrm{E}f(X_n)$. The desired conclusion is obtained by observing that (27) yields $\lim_{n\to\infty} \mathrm{E}f(X_n) = \int_S f(y)\pi(\mathrm{d}y)$. □

Remark 8. We refer to the original article (McLeish 1975) or to the recent review article (Davidson and de Jong 1997) for basic properties of mixingales. However, we point out that the proof of Theorem 2 could be concluded by elementary means, without referring to the theory of mixingales, by applying Proposition 4 to estimate the variance of the sum $S_n = (1/n)\sum_{k=1}^{n} f(X_n) - \int_S f(y)\pi(\mathrm{d}y)$ and utilizing the boundedness of the function f. Nevertheless, the reference to mixingales is useful since it is possible to weaken condition (iii) and still obtain Theorem 2. In this manner one obtains Theorem 1 also in the case where the covariance is calculated from a relatively slowly increasing segment of the near history only (cf. Remark 5). For instance, this is the case if at time t this segment has length $\sim t^{\alpha}$, where $\alpha \in (\frac{1}{2}, 1)$.

Finally, we note that in this paper we have left open the question whether the convergence of the algorithm (as established in Theorem 1) satisfies a central limit theorem.

5. Testing AM in practice and comparison with traditional methods

In this section we present results obtained from testing the AM algorithm numerically. From the practical point of view, it is important to know how accurate the simulations of the target distribution will be that one can expect to get from finite MCMC runs. In Haario *et al*. (1999) we compared three different methods: the random walk Metropolis algorithm (M), the single-component Metropolis algorithm (SC), and the adaptive proposal algorithm (AP) – see Section 1 or Haario *et al*. (1999) for the exact definition and more details. Recall again that the difference between the AP and AM algorithms was simply that in AP the covariance for the proposal distribution was computed only from a fixed number of previous states. Here we have done similar tests to those in Haario *et al*. (1999) and included the AM algorithm in the comparison.

We have tested the AM algorithm for various dimensions up to $d = 200$. The algorithm appears to work successfully. Naturally the adaptation becomes slower as the dimension increases and becomes more sensitive to a very bad choice of the initial covariance. Here we present the results of extensive tests in dimension $d = 8$. We used two restricted

Gaussian distributions as the target distributions – uncorrelated (π_1) and correlated (π_2) – and two nonlinear 'banana'-shaped distributions with compact supports – moderately 'twisted' (π_3) and strongly 'twisted' (π_4). The supports of the test distributions are compact in order to satisfy the assumptions of our theoretical result (Theorem 1).

Our test distributions are obtained by those used in Haario *et al*. (1999) by setting the densities to zero outside a compact set. Hence, the density of our first test distribution π_1 is an uncorrelated Gaussian density, which is centred and has covariance diag(100, 1, ..., 1) and is restricted to a parallelpiped with corners at points ($\pm 35, \pm 3.5, \ldots, \pm 3.5$). In this set-up about 99.6% of the probability mass of the unrestricted Gaussian is contained in the parallelpiped. The correlated restricted Gaussian distribution π_2 is obtained from π_1 simply by rotating the distribution so that the main axis corresponds to the direction (1, ..., 1). The twisted (and restricted) Gaussian test distributions π_3 and π_4 are similarly obtained from π_1 by applying the same measure-preserving transformations as are used in Haario *et al*. (1999, p. 381). We refer to Haario *et al*. (1999) for a more detailed explanation of the test procedure – see especially p. 382 for pictures of the corresponding unrestricted target distributions.

The number of function evaluations varied depending on the target distribution: 20 000 for π_1 and π_2, 40 000 for π_3 and 80 000 for π_4. The starting values were sampled relatively close to the peak values of the target densities. The burn-in period was chosen to be half of the chain length. Each test case was run 100 times in order to retrieve statistically relevant information. Hence, each accuracy criterion number is an average value over 100 repetitions.

We have tried to be fair in choosing the proposal distributions for the random walk Metropolis and the single-component Metropolis algorithms. For example, in the case of the restricted Gaussian target distributions we used for the Metropolis algorithm covariances corresponding to the unrestricted targets and normalized them with the heuristic optimal scaling from Gelman *et al*. (1996).

In Figure 1 the test results in dimension 8 are summarized in graphical form. We present the mean and the error bars giving the standard deviations corresponding to the 68.3% probability region. The results expressed in the figure indicate that the AM algorithm simulates the target distribution most accurately in these tests. With the restricted Gaussian target distributions the results obtained using the AM algorithm are equally as good as those using the Metropolis algorithm with an optimal proposal distribution. Moreover, in the case of nonlinear distributions the AM algorithm seems to be superior.

In Figure 2 we compare the performance of the AM algorithm with the Metropolis algorithm. The proposal distribution for the Metropolis algorithm was symmetric and the size was selected so that the acceptance rate becomes quite optimal. We used π_1 as the target distribution. In Figure 2 the autocorrelation functions of the AM and the Metropolis algorithm are drawn for two projections. In the direction of the largest width of the target distribution the autocorrelation of the Metropolis algorithm indicates weaker convergence. This example demonstrates how the tuning of the proposal distribution according to the acceptance rate only may lead to difficulties.

Finally, we point out that according to our tests there was no essential difference in the performance of the AM algorithm between the restricted and unrestricted target

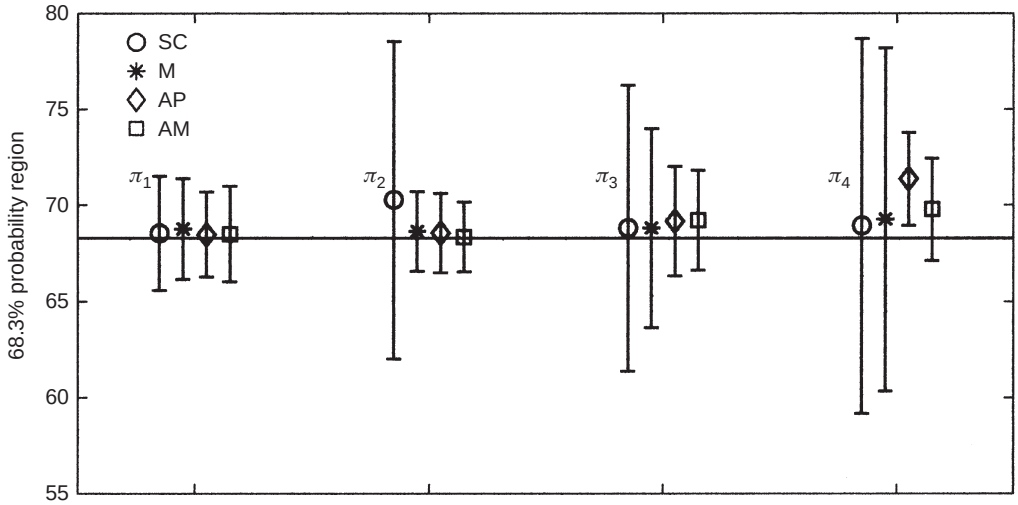

Figure 1. Comparison of the performance of the single-component Metropolis algorithm (SC), Metropolis algorithm (M), adaptive proposal algorithm (AP) and adaptive Metropolis algorithm (AM) with different eight-dimensional target distributions $\pi_1 - \pi_4$. The symbols correspond to the mean frequency of hits in the 68.3% probability region of 100 simulations and the error bar around the symbol corresponds to the standard deviation of the hits. The true value (68.3%) is indicated by a horizontal line.

distributions. Thus, it is reasonable to expect that an analogue of Theorem 1 also holds for non-compactly supported distributions whose densities decay rapidly enough.

Appendix

We present here an illustrative two-dimensional example also considered in Haario *et al*. (1999). There the target distribution was tested with the AP algorithm, where the covariance C_t was calculated from the last 200 states – see Section 1 or Haario *et al*. (1999) for the definition of the AP algorithm. In the example the AP algorithm produced considerable error in the simulation. This phenomenon underlines the importance of calculating the covariance from an increasing segment of the history, as is done in the AM algorithm. When the AM algorithm was applied to the same example it produced, as expected, simulation that was free of bias. For many practical applications the error produced by the AP algorithm is, however, ignorable (see Haario *et al*. 1999).

Example 1. Let us define the density π on the rectangle $R = [-18, 18] \times [-3, 3] \subset \mathbb{R}^2$. Let $S = [-0.5, 0.5] \times [-3, 3]$ and set

$$\pi(x) = \begin{cases} 36 & \text{if } x \in S, \\ 1 & \text{if } x \in R \backslash S. \end{cases}$$

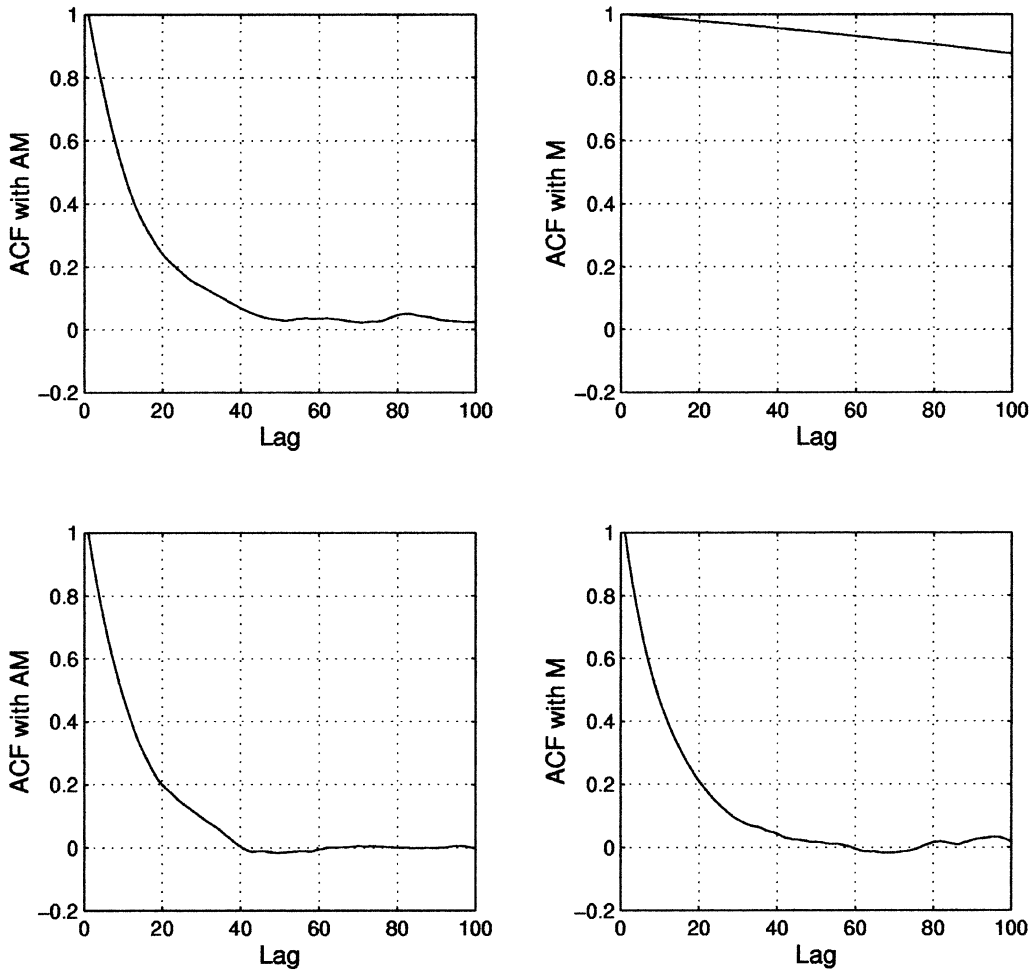

Figure 2. Comparison of the performance of the AM algorithm (left column) and the Metropolis algorithm with a fairly optimal acceptance rate (right column). The target distribution was π_1. On the top line the autocorrelation function in the direction of the largest eigenvalue of target's covariance matrix is shown. The bottom line corresponds to an orthogonal direction. The acceptance rate with the AM algorithm was 27% and with the Metropolis algorithm 26%.

(See Figure 3 for the one-dimensional projection of the density function.)

With this choice $\pi(S) : \pi(R \backslash S) = 36 : 35$ and hence about half of the mass is concentrated on the middle strip S. Thus an ergodic MCMC algorithm should stay for about the same amount of time on S and on R. However, S and R are thin rectangles with opposite orientations. This forces the AP algorithm to regularly turn the direction of the proposal distribution. This causes notable bias in the simulation on S (see Figure 4(a),

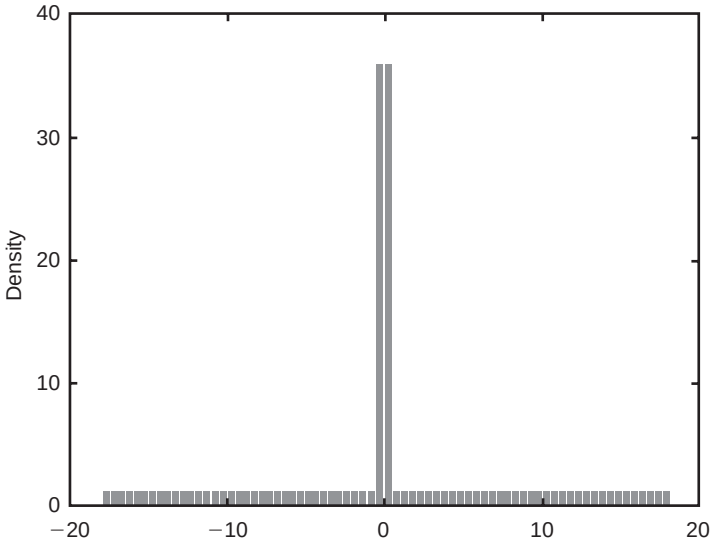

Figure 3. The (unscaled) one-dimensional projection of true target distribution of Example 1.

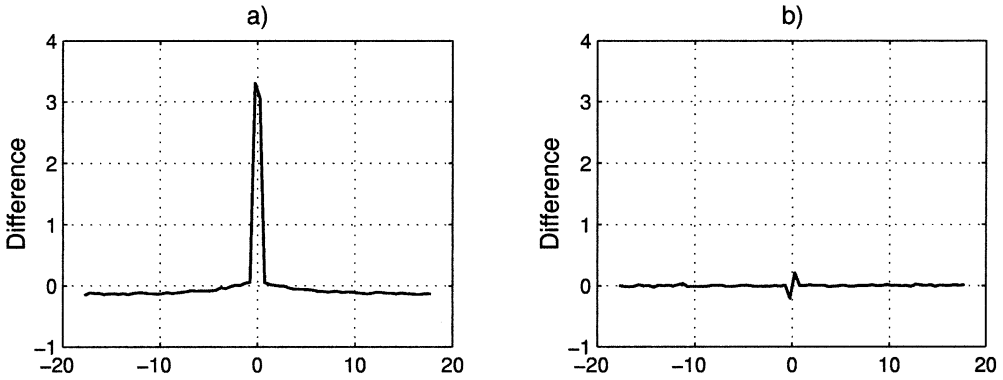

Figure 4. The difference between the real target distribution of Example 1 and the sampled distributions. In (a) the sampling was done using the AP algorithm. The curve represents the mean values of 100 runs with 100 000 states. In (b) the sampling method was the AM algorithm.

where the difference between the true target distribution and the one simulated by AP is presented). In fact, the relative error in the simulation on S is about 10%. There is also a slight error in the simulation near the far ends of the rectangle R. The corresponding unbiased results of the AM algorithm are presented in Figure 4(b).

Acknowledgements

We thank Elja Arjas, Kari Auranen, Esa Nummelin and Antti Penttinen for useful discussions on the topics of the paper. The second author (ES) was supported by the Academy of Finland, Project 32837.

References

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

Dobrushin, R. (1956) Central limit theorems for non-stationary Markov chains II. *Theory Probab. Appl.*, **1**, 329–383.

Evans, M. (1991) Chaining via annealing. *Ann. Statist.*, **19**, 382–393.

Fishman, G.S. (1996) *Monte Carlo: Concepts, Algorithms and Applications*. New York: Springer-Verlag.

Gelfand, A.E. and Sahu, S.K. (1994) On Markov chain Monte Carlo acceleration. *J. Comput. Graph. Statist.*, **3**, 261–276.

Gelman, A.G., Roberts, G.O. and Gilks, W.R. (1996) Efficient Metropolis jumping rules. In J.M. Bernardo, J.O. Berger, A.F. David and A.F.M. Smith (eds), *Bayesian Statistics V*, pp. 599–608. Oxford: Oxford University Press.

Gilks, W.R. and Roberts, G.O. (1995) Strategies for improving MCMC. In W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, pp. 75–88. London: Chapman & Hall.

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

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995) Introducing Markov chain Monte Carlo. In W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, pp. 1–19. London: Chapman & Hall.

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

Haario, H. and Saksman, E. (1991) Simulated annealing process in general state space. *Adv. Appl. Probab.*, **23**, 866–893.

Haario, H., Saksman, E. and Tamminen, J. (1999) Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.*, **14**, 375–395.

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

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

McLeish, D.L. (1975) A maximal inequality and dependent strong laws. *Ann. Probab.*, **3**, 829–839.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Neveu, J. (1965) *Mathematical Foundations of the Calculus of Probability*. San Francisco: Holden-Day.

Nummelin, E. (1984) *General Irreducible Markov Chains and Non-negative Operators*. Cambridge: Cambridge University Press.

Roberts, G.O., Gelman, A. and Gilks, W.R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.

Sahu, S.K. and Zhigljavsky, A.A. (1999) Self regenerative Markov chain Monte Carlo with adaptation. Preprint. *http://www.statslab.cam.ac.uk/∼mcmc.*

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Received June 1998 and revised February 2000