An Adaptive Metropolis algorithm

Heikki Haario, Eero Saksman* and Johanna Tamminen

Abstract

A proper choice of a proposal distribution for MCMC methods, e.g. 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 Algorithm (AM), 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 AM provides an easy to use algorithm for practical computation.

1991 Mathematics Subject Classification: 65C05, 65U05.
Keywords: adaptive MCMC, 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 e.g. the random walk Metropolis algorithm is essential in order to obtain reasonable results by the simulation in a limited amount of time. This 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 e.g. [Gelman *et al.* 1996], [Gilks *et al.* 1995], [Gilks *et al.* 1998],[Haario *et al.* 1998], and [Roberts *et al.* 1994]). 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 recurrence times 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 interesting self regenerative version of adaptive MCMC, based on introducing an auxiliary chain, is contained in the

*Supported by the Academy of Finland, Project 32837

recent preprints [Sahu and Zhigljavsky 1998a] and [Sahu and Zhigljavsky 1998b]. For other versions of adaptive MCMC and related work we refer to e.g. [Evans 1991], [Fishman 1996] [Gelfand and Sahu 1994], and [Marinari and Parisi 1992] together with the references therein.

We introduce here an adaptive Metropolis algorithm (AM), which adapts continuously to the target distribution. What is important, 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.* 1998]). In the AP algorithm the proposal distribution is a Gaussian distribution centered at the current state, and the covariance is calculated from a fixed finite number of the 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 uses the cumulating information right from the very begin of the simulation. The rapid start of the adaptation ensures that the search becomes more effective already 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 $X_0, X_1, \ldots, X_t \in \mathbb{R}^{\mathrm{d}}$. 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.* 1998]) that the AP algorithm fails this property. Our main result, Theorem 1 below, verifies that *the AM process indeed has the correct ergodicity properties.* 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. 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 by applying both linear and non-linear, 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 shall assume for simplicity that our target distribution is supported on a *bounded* subset $S \subset \mathbb{R}^{d}$, and that it has the (unscaled) density $\pi(x)$ with respect to the Lebesgue measure on S. We shall also denote the target distribution by π with a slight abuse of the notation.

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

Suppose thus 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 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 the probability

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

in which case of we set $X_t = Y$, and otherwise $X_t = X_{t-1}$.

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 assumed 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 $C_t = s_d \mathrm{Cov}(X_0, \ldots X_{t-1}) + s_d \varepsilon I_d$ 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. In order to start we select an arbitrary strictly positive definite initial covariance C_0, which of course is chosen according to our best priori knowledge (which may be quite poor). We select an index $t_0 > 0$ for the length of an initial period and define

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

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 $x_0, \ldots, x_k \in \mathbb{R}^{d}$:

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

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

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

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 slower the effect of the adaptation starts to take place. In a sense the size of t_0 reflects our trust in the initial covariance C_0. As S was assumed to be bounded, the covariance C_t will never blow up. 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. Hence, in practical computations one may safely utilize the definition (1) with $\varepsilon = 0$, although the change is negligible if ε was already chosen small enough. However, for the purposes of our theoretical considerations it is convenient to assume that $\varepsilon > 0$.

Remark 2 In order to avoid a slow start of the algorithm it is possible to employ special tricks. Naturally, if a priori knowledge (like 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 that it moves at least a little. If it has not moved enough during some 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 (using again the whole 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 the introduction, the covariance C_t was calculated from the last H states only, where $H \geq 2$. This strategy brings non-exactness to the simulation as an undesirable consequence. 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. Moreover, 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 setup. Let $(S, \mathcal{B}, m)$ be a state space. For natural numbers $n \geq 1$ the symbol K_n will always denote a *generalized* transition probability from the product $S^n \to S$. Thus K_n is a map $S^n \times \mathcal{B} \to [0,1]$ for which $x \to K_n(x; A)$ is $\mathcal{B}^n$-measurable for each $A \subset \mathcal{B}$, and $K_n(x; \cdot)$ is a probability measure on $(S, \mathcal{B})$ for each $x \in S^n$. In a natural way K_n defines a positive contraction from $\mathcal{M}(S^n)$ into $\mathcal{M}(S)$. Here $\mathcal{M}(S)$ denotes the set of finite measures on $(S, \mathcal{B})$ and the norm $\|\cdot\|$ on $\mathcal{M}(S)$ denotes the total variation norm. 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

$$\begin{aligned}\mathbf{P}(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \quad &= \quad \int_{y_0 \in A_0} \mu_0(dy_0) \Big(\int_{y_1 \in A_1} K_1(y_0; dy_1) \\ \Big(\int_{y_2 \in A_1} K_2(y_0, y_1; dy_2) \quad \cdots \quad &\Big(\int_{y_n \in A_n} K_n(y_0, y_1, \ldots, y_{n-1}; dy_n) \Big) \ldots \Big).\end{aligned} \tag{4}$$

In fact, it is directly verified that these distributions are consistent and the theorem of Ionescu-Tulcea (see e.g. 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 recall, that our 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 sigma algebra $\mathcal{B}(S)$ 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. For simplicity we shall assume here that *the density is bounded from above on* S : for some $M < \infty$ it holds that

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

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

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

The Gaussian proposal transition probability corresponding to the covariance C satisfies

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

where $A \subset \mathbb{R}^d$ is a Borel set and dy is the standard Lebesgue measure on $\mathbb{R}^d$. It follows that Q_C is m-symmetric (see e.g. [Haario and Saksman 1991, Definition 2.2]).

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

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

for $A \in \mathcal{B}(S)$, 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 AM chain, that we now define, is a Markov chain with infinite memory. The definition 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 ε be given. Define the functions C_n for $n \geq 1$ by the formula (1). For a given initial distribution μ_0 the Adaptive Metropolis chain (the 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 $n \geq 1$, $x_i \in S$ $(0 \leq i \leq n-1)$, and for subsets $A \in \mathcal{B}(S)$.

Let us turn to the study of the ergodicity properties of the AM chain. 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 an 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 $(S, \mathcal{B})$. As usual, λT denotes the measure $A \mapsto \int_S T(x; A) \lambda(dx)$.

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

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

The condition $\delta(T^{k_0}) < 1$ for some $k_0 \geq 1$ is well known to be equivalent to the uniform ergodicity (compare [Nummelin 1984, Section 6.6.]) of the Markov chain having the 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 $n \geq 2$) and a *fixed* $(n-1)$-tuple $(y_0, y_1, \ldots y_{n-2}) \in S^{n-1}$, we denote $\widetilde{y}_{n-2} = (y_0, y_1, \ldots y_{n-2})$ and define the transition probability $K_{n,\widetilde{y}_{n-2}}$ by

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

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

We are now ready to state and prove our main theorem.

Theorem 1 *The AM chain (X_n) defined by the generalized transition probabilities* (9) *simulates properly the target distribution π regardless of the initial distribution: for any bounded and measurable function $f : S \to \mathbb{R}$ it holds almoast surely that*

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

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) – (iii):*

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

$$\delta((K_{n,\widetilde{y}_{n-2}})^{k_0}) \leq \lambda < 1 \quad \text{for all } \widetilde{y}_{n-2} \in S^{n-1} \text{ and } n \geq 2.$$

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

$$\|\pi K_{n,\widetilde{y}_{n-2}} - \pi\| \leq \frac{c_0}{n} \quad \text{for all } \widetilde{y}_{n-2} \in S^{n-1} \text{ and } n \geq 2.$$

(iii) The estimate for the operator norm

$$\|K_{n,\widetilde{y}_{n-2}} - K_{n+k,\widetilde{y}_{n+k-2}}\|_{\mathcal{M}(S)\to\mathcal{M}(S)} \leq c_1 \frac{k}{n},$$

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

Then, if $f : S \to \mathbb{R}$ is bounded and measurable, it holds almost surely that

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

In what follows the auxiliary constants c_i, $i = 1, 2, \ldots$, depend on S, ε, or C_0, and their actual value is irrelevant for our purposes here.

Proof of Theorem 1. According to the above theorem it suffices to prove that the AM chain satisfies the 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

$$0 < c_1 I_d \leq C \leq c_2 I_d. \tag{14}$$

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

$$K_{n,\widetilde{y}_{n-2}}(x; A) \geq c_3 \pi(A) \quad \text{for all } x \in S \text{ and } A \subset S,$$

with $c_3 > 0$. This easily yields (compare e.g. [Nummelin 1984, pp. 122–123]) that $\delta(K_{n,\widetilde{y}_{n-2}}) \leq 1 - c_3$, which proves (i) with $k_0 = 1$.

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

$$\|K_{n,\widetilde{y}_{n-2}} - K_{n+k,\widetilde{y}_{n+k-2}}\|_{\mathcal{M}(S)\to\mathcal{M}(S)} \leq 2\, \sup_{y\in S, A\in\mathcal{B}(S)} |K_{n,\widetilde{y}_{n-2}}(y; A) - K_{n+k,\widetilde{y}_{n+k-2}}(y; A)|.$$

Fix $y \in S$ and $A \in \mathcal{B}(S)$ 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

$$\begin{aligned}
&|K_{n,\widetilde{y}_{n-2}}(y; A) - K_{n+k,\widetilde{y}_{n+k-2}}(y; A)| = |M_{R_1}(y; A) - M_{R_2}(y; A)| \\
\leq\ & \Big|\int_{x\in A} (N_{R_1} - N_{R_2})(x - y) \min\left(1, \frac{\pi(x)}{\pi(y)}\right) m(dx) \\
&\quad + \chi_A(x) \int_{x\in\mathbb{R}^d} (N_{R_1} - N_{R_2})(x - y) \left[1 - \min\left(1, \frac{\pi(x)}{\pi(y)}\right)\right] m(dx)\Big| \\
\leq\ & 2\int_{\mathbb{R}^d} |N_{R_1}(z) - N_{R_2}(z)| dz \\
\leq\ & 2\int_{\mathbb{R}^d} dz \int_0^1 ds \left|\frac{d}{ds} N_{R_1+s(R_2-R_1)}(z)\right| ds \\
\leq\ & c_4 \|R_1 - R_2\|,
\end{aligned}$$

where at the last stage we applied the knowledge (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 $\|C_t - C_{t+1}\| \leq c/t$ for $t > 1$. By applying this inductively and using the uniform boundedness from above of the covariances C_t we easily see that

$$\|R_1 - R_2\| \leq c(S, C_0, \varepsilon)k/n,$$

and hence the previous estimates yield (iii).

In order to check condition (ii) fix $\widetilde{y}_{n-2} \in S^{n-1}$ and denote $C^* = C_{n-1}(y_0, \ldots y_{n-2})$. It follows that $\|C^* - C_n(y_0, \ldots y_{n-2}, y)\| \leq c_5/n$, where c_5 does not depend on $y \in S$. We may hence estimate as before that

$$\|K_{n,\widetilde{y}_{n-2}} - M_{C^*}\|_{\mathcal{M}(S)\to\mathcal{M}(S)} \leq \frac{c_6}{n}.$$

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

$$\|\pi - \pi K_{n,\widetilde{y}_{n-2}}\| = \|\pi(M_{C^*} - K_{n,\widetilde{y}_{n-2}})\| \leq \frac{c_6}{n},$$

which completes the proof of Theorem 1. □

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

Corollary 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 $s_d\mathrm{Cov}(\pi) + \varepsilon I_d$, where $\mathrm{Cov}(\pi)$ 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 by choosing repeatedly $f(x) = x_i$ and $f(x) = x_i x_j$, where $1 \leq i,\ j \leq d$. □

Remark 4 In the definition of the AM chain one can replace the Gaussian proposals by e.g. uniform distributions in a parallelepiped. In this case the size and orientation of the parallelepiped is 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 the exactness of the simulation. The only difference is that in this case the constant k_0 in condition (i) of Theorem 2 may exceed 1. Naturally, here one has to make the necessary assumption that the set $\{x : \pi(x) > 0\}$ does not contain components that are too much separated. 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 exactness of the simulation with only minor changes in the proof. Similar remarks apply also to the case, where one updates the covariance only every n_0:th step (see Remarks 3 and 6). 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.

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 transition probabilities (K_n) fulfill the conditions of Theorem 2. Denote by $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ the sigma algebra generated by the chain up to time n and write $\lambda' = \lambda^{1/k_0}$. Then for all initial distributions and for any measurable and bounded function f on S it holds that*

$$\|\mathbf{E}(f(X_{n+k})|\mathcal{F}_n) - \int_S f(y)\pi(dy)\|_\infty \leq c_2(c_0, \lambda) \inf_{1 \leq j \leq k} \left(\frac{j^2}{n+k-j} + \lambda'^j \right) \|f\|_\infty. \tag{15}$$

Proof. We may clearly assume that $\int_S f(y)\pi(dy) = 0$ since the claim holds trivially if the function f is a constant. Let $n \geq 2, k \geq 2$ and note that from the definition of the conditional expectation and (4) it follows that (almost surely)

$$\begin{aligned} & \mathbf{E}(f(X_{n+k})|\mathcal{F}_n) && (16) \\ = & \int_{y_{n+1}\in S} K_{n+1}(X_0, X_2, \ldots, X_n; dy_{n+1}) \left(\int_{y_{n+2}\in S} K_{n+2}(X_0, X_2, \ldots, X_n, y_{n+1}; dy_{n+2}) \right. \\ \cdots & \left(\int_{y_{n+k}\in S} K_{n+k}(X_0, X_2, \ldots, X_n, y_{n+1}, \ldots, y_{n+k-1}; dy_{n+k}) f(y_{n+k}) \right) \ldots \Bigg). \end{aligned}$$

Let us denote $(X_0, \ldots, X_n) = \widetilde{X}_n$. In what follows $\widetilde{X}_n$ 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 $Q(y; dz) = K_{n+2}(\widetilde{X}_n, y; dz)$. The assumption

(iii) yields for arbitrary values of $\widetilde{X}_n$ and $y_{n+1}, \dots y_{n+k-1}$ that

$$\begin{aligned}|\int_{y_{n+k}\in S}\Big(&K_{n+k}(\widetilde{X}_n, y_{n+1}, \dots, y_{n+k-1}; dy_{n+k}) \\ &- K_{n+2}(\widetilde{X}_n, y_{n+k+1}; dy_{n+k})\Big) f(y_{n+k})| \le c_0 \|f\|_\infty \frac{k-2}{n+2}.\end{aligned} \tag{17}$$

This estimate enables us write (16) in the form

$$\begin{aligned}\mathbf{E}(f(X_{n+k})|\mathcal{F}_n) &= g_k(\widetilde{X}_n) + \int_{y_{n+1}\in S} K_{n+1}(\widetilde{X}_n; dy_{n+1}) \left(\int_{y_{n+2}\in S} K_{n+2}(\widetilde{X}_n, y_{n+1}; dy_{n+2})\right. \\ &\cdots \left(\int_{y_{n+k-1}\in S} K_{n+k-1}(\widetilde{X}_n, y_{n+1}, \dots, y_{n+k-2}; dy_{n+k-1})\right. \\ &\cdots \left.\left.\left(\int_{y_{n+k}\in S} K_{n+2}(\widetilde{X}_n, y_{n+k-1}; dy_{n+k}) f(y_{n+k})\right) \dots\right)\right),\end{aligned} \tag{18}$$

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

$$\begin{aligned}&|g_k(\widetilde{X}_n)| \le \int_{y_{n+1}\in S} K_{n+1}(\widetilde{X}_n; dy_{n+1}) \left(\int_{y_{n+2}\in S} K_{n+2}(\widetilde{X}_n, y_{n+1}; dy_{n+2})\right. \\ \cdots \quad &\left.\left(\int_{y_{n+k-1}\in S} K_{n+k-1}(\widetilde{X}_n, y_{n+1}, \dots y_{n+k-2}; dy_{n+k-1}) f(y_{n+k}) c_0 \|f\|_\infty \frac{k-2}{n+2}\right) \dots\right) \\ &\le c_0 \|f\|_\infty \frac{k-2}{n+2}.\end{aligned}$$

In the next step we iterate the procedure by replacing the generalized transition probability $K_{n+k-1}(\widetilde{X}_n, y_{n+1}, \dots, y_{n+k-2}; dy_{n+k-1})$ by the transition probability Q in formula (18). By continuing in this manner we obtain

$$\begin{aligned}\mathbf{E}(f(X_{n+k})|\mathcal{F}_n) &= \int_{y_{n+1}\in S} K_{n+1}(\widetilde{X}_n; dy_{n+1}) \left(\int_{y_{n+2}\in S} Q(y_{n+1}; dy_{n+2})\right. \\ &\cdots \left.\left(\int_{y_{n+k}\in S} Q(y_{n+k-1}; dy_{n+k}) f(y_{n+k})\right) \dots\right) \\ &+ g_2(\widetilde{X}_n) + g_2(\widetilde{X}_n) + \dots + g_k(\widetilde{X}_n),\end{aligned}$$

where

$$\begin{aligned}g_j &= \int_{y_{n+1}\in S} K_{n+1}(\widetilde{X}_n; dy_{n+1}) \left(\int_{y_{n+2}\in S} K_{n+2}(\widetilde{X}_n, y_{n+1}; dy_{n+2})\right. \\ &\cdots \left(\int_{y_{n+j}\in S}\Big(K_{n+j}(\widetilde{X}_n, y_{n+1}, \dots, y_{n+j-1}; dy_{n+j})\right. \\ &- \left.\left. K_{n+2}(\widetilde{X}_n, y_{n+j-1}; dy_{n+j})\Big) Q^{k-j} f(y_{n+j})\right) \dots\right).\end{aligned} \tag{19}$$

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

Since $\|Q^{k-j}f\|_\infty \leq \|f\|_\infty$ we obtain as before from assumption (i) that

$$|g_j| \leq c_0 \frac{j-2}{n+2}\|f\|_\infty.$$

Summing up, we have shown that

$$\mathbf{E}(f(X_{n+k})|\mathcal{F}_n) = \varepsilon_{n,k} + \int_{y_{n+1}\in S} K_{n+1}(X_0,\ldots,X_n,dy_{n+1})Q^{k-1}f(y_{n+k}), \tag{20}$$

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

$$|\varepsilon_{n,k}| \leq \sum_{j=2}^{k} c_0 \frac{j-2}{n+2}\|f\|_\infty \leq \frac{c_0 k^2}{n}\|f\|_\infty. \tag{21}$$

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

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

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

$$\begin{aligned}\|Q^{k-1}f\|_\infty &= \sup_{x\in S}|\delta_x Q^{k-1}f| \leq \sup_{x\in S}|(\delta_x - \pi)Q^{k-1}f| + |\pi Q^{k-1}f| \\ &\leq 2\lambda^{k'}\|f\|_\infty + |(\pi Q^{k-1} - \pi)f| \leq (\frac{c_0(k-1)}{n+2} + 2\lambda^{k'})\|f\|_\infty. \end{aligned} \tag{22}$$

Combining this with (20) and (21), it follows that

$$\|\mathbf{E}(f(X_{n+k})|\mathcal{F}_n)\|_\infty \leq c_1(c_0,\lambda)\left(\frac{k^2}{n} + \lambda^{[(k-1)/k_0]}\right)\|f\|_\infty, \tag{23}$$

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

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

$$\|\mathbf{E}(f(X_{n+k})|\mathcal{F}_n)\|_\infty \leq \|\mathbf{E}(f(X_{n+k})|\mathcal{F}_{n+k-j})\|_\infty.$$

Hence, by replacing n by $n+k-j$ and k by j in the estimate (23) we finally deduce that

$$\|\mathbf{E}(f(X_{n+k})|\mathcal{F}_n)\|_\infty \leq \inf_{1\leq j\leq k} c_1(c_0,\lambda)\left(\frac{j^2}{n+k-j} + \lambda^{[(j-1)/k_0]}\right)\|f\|_\infty. \tag{24}$$

The claim of the proposition follows immediately from this estimate. $\square$

Proof of Theorem 2. From Proposition 4 we obtain that

$$\|\mathbf{E}(f(X_{n+k}) - \int_S f(y)\pi(dy))|\mathcal{F}_n)\|_\infty \leq \psi(k), \tag{25}$$

where

$$\psi(k) \equiv c_2(c_0, \lambda) \inf_{1\leq j\leq k}(\frac{j^2}{k-j} + \lambda'^j)\|f\|_\infty \leq c_2(c_0, f, \lambda)\frac{\log^2 k}{k}, \tag{26}$$

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

At this stage the estimate (26) for the asymptotic independence together with the definition of the sigma-algebra $\mathcal{F}_n$ make it clear that $f(X_n) - \int_S f(y)\pi(dy)$ is a mixingale in the sense of McLeish (see [McLeish 1975] or [Hall and Heyde 1980, p. 19]). Moreover, $\psi(k) \leq C(\varepsilon)k^{\varepsilon-1}$ 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 $f(X_n)$ in order to obtain the desired conclusion. □

Remark 6 We refer to the original article [McLeish 1975] or to the recent review article [Davidson and 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 $S_n = (1/n)\sum_{k=1}^n f(X_n) - \int_S f(y)\pi(dy)$ 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. By 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 (compare Remark 5). For instance, this is the case if at time t this segment has the length $\sim t^\alpha$, where $\alpha \in (1/2, 1)$.

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 simulations of the target distribution one can expect to get from finite MCMC runs. In [Haario *et al.* 1998] we compared three different methods: random walk Metropolis algorithm (M), single component Metropolis algorithm (SC), and the adaptive proposal algorithm (AP) (see the introduction or [Haario *et al.* 1998] for the exact definition and more detailes). Recall again that the difference between 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 made similar tests as in [Haario *et al.* 1998] and included the AM algorithm in the comparison.

We have applied the AM algorithm succesfully in various dimensions up to $d = 200$. Here we present the results of extensive tests in dimension $d = 8$. As target disributions

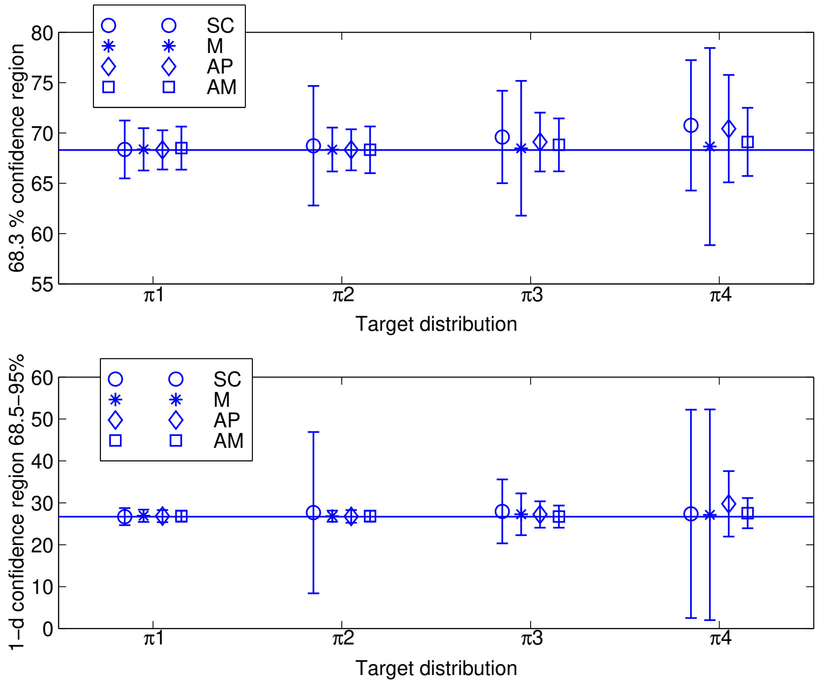

Figure 1: *The result with different 8-dimensional target distributions. The top figure corresponds to the frequency of hits to the 68.3% confidence region. The lower figure corresponds to the frequency of hits to the 1-dimensional confidence region between 68.3% and 95%. The true value is denoted with a horizontal line.*

we used two Gaussian distributions: correlated (π_1) and uncorrelated (π_2), and two non-linear 'banana' shaped distributions: moderately 'twisted' (π_3) and strongly 'twisted' (π_4). The number of function evaluations varied depending on the target distribution: 20000 for π_1 and π_2, 40000 for π_3 and 80000 for π_4. 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 refer to [Haario *et al.* 1998] for a more detailed explanation of the test procedure.

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 Gaussian target distributions we used covariances corresponding to the 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 a graphical form. In the top picture we present the mean and the error bars given by the standard deviations for the 68.3% confidence region. In the lower picture the mean and the error bars for the 1-dimensional confidence region between 68.3% and 95% are given. The 1-dimensional confidence region contains all those points whose projections to the direction of the main semi-axis of the target distribution belong to the corresponding 1-dimensional confidence interval. In the case of the non-linear target distributions π_3 and π_4 this value reflects the accuracy of the simulation on the tails of the 'banana'.

Table 1: *Results obtained from 8-dimensional non-linear distribution π_4. The number of function evaluations was 80000 and the run was performed* 100 *times. Mean*$(||E||)$ *denotes the mean distance of the expectation values of the chains from the true value at origo calculated over 100 repetitions. Thus mean*$(||E||) = \frac{1}{100}\sum_{j=1}^{100}\left(\sum_{i=1}^{d}(E_j^i)^2\right)^{\frac{1}{2}}$, *where E_j is the mean vector of chain j. Std*$(||E||)$ *denotes the standard deviation of the distances from the true value at origo.*

	Non-linear (π_4)			
Method	**SC**	**M**	**AP**	**AM**
Mean(||E||)	6.53	7.89	4.85	2.41
Std(||E||)	4.79	7.54	4.20	1.15
Acceptance rate (%)	70.39	7.84	9.60	4.59

We present in the Table 1 some standard performance criteria in the case of the most nonlinear target distribution π_4 (again in dimension 8). The rows of the table give the performance criteria while the columns represent different methods compared.

The results expressed in the figure and in the table indicate that the AM algorithm simulates the target distribution most accurately in these tests. With the Gaussian target distributions the results obtained using the AM algorithm are equally good as those using the Metropolis algorithm with an optimal proposal distribution. Moreover, in the case of non-linear distributions the AM algorithm seems to be superior.

Acknowledgements: We thank Elja Arjas, Esa Nummelin and Antti Penttinen for useful discussions on the topics of the paper.

References

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

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

[Evans 1991] M. Evans: *Chaining via annealing,* Ann. Statistics 19 (1991), 382–393.

[Fishman 1996] G. S. Fishman: *Monte Carlo. Concepts, algorithms and applications.* Springer (1996).

[Gelfand and Sahu 1994] A. E. Gelfand and S. K. Sahu: *On Markov chain Monte Carlo acceleration,* J. Computational and Graphical Stat. 3 (1994), 261–276.

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

[Gilks *et al.* 1994] W. R. Gilks, G.O. Roberts, E.I. George: *Adaptive direction sampling,* The Statistician, 43, 179–189.

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

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

[Gilks *et al.* 1998] W. R. Gilks, G.O. Roberts and S.K. Sahu: *Adaptive Markov Chain Monte Carlo,* J. Amer. Statist. Assoc. (to appear).

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

[Haario *et al.* 1998] H. Haario, E. Saksman and J. Tamminen: *Adaptive proposal distribution for random walk Metropolis algorithm,* Reports of the Department of Mathematics, University of Helsinki. Preprint 176 (1998).

[Hall and Heyde 1980] P. Hall and C.C. Heyde: *Martingale limit theory and its application.* Academic Press, 1980.

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

[Marinari and Parisi 1992] E. Marinari and G. Parisi: *Simulated tempering: a new Monte Carlo Scheme,* Europhys. Lett. 19 (1992), 451–458.

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

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

[Neveu 1965] J. Neveu: *Mathematical foundations of the calculus of probability.* Holden-Day 1965.

[Nummelin 1984] E. Nummelin: *General irreducible Markov chains and non-negative operators.* Cambridge University Press, 1984.

[Roberts *et al.* 1994] G. O. Roberts, A. Gelman and W. R. Gilks: *Weak convergence and optimal scaling of random walk metropolis algorithms.* Preprint. (`http://www.stats.bris.ac.uk/MCMC/`)

[Sahu and Zhigljavsky 1998a] S. K. Sahu and A. A. Zhigljavsky: *Self regenerative Markov chain Monte Carlo.* Preprint. (`http://www.stats.bris.ac.uk/MCMC/`)

[Sahu and Zhigljavsky 1998b] S. K. Sahu and A. A. Zhigljavsky: *Adaption for self regenerative MCMC,* Preprint. (`http://www.stats.bris.ac.uk/MCMC/`)

[Tierney 1994] L. Tierney: *Markov chains for exploring posterior distributions, with discussion*, Ann. Statistics (1994), 1701–1762.

Appendix

We present here an illustrative two-dimensional example from [Haario *et al.* 1998], where in the AP algorithm the covariance C_t was calculated from last 200 states and 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.* 1998].

Example. Let us consider a density f defined on the rectangle $R = [-18, 18] \times [-3, 3] \subset \mathbb{R}^2$ containing the middle strip $S = [-0.5, 0.5] \times [-3, 3]$. We set

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

With this choice $\pi(S) : \pi(R \setminus S) = 36 : 35$ and hence about half of the mass is concentrated on S. Thus the AP algorithm stays about the same time on S and on R. However, S and R are thin rectangles with opposite direction which forces the AP to regularly turn the direction of the proposal distribution. This causes notable bias in the simulation on S (see Figure 2). In fact, the relative error in the simulation on S is 13%. 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 3.

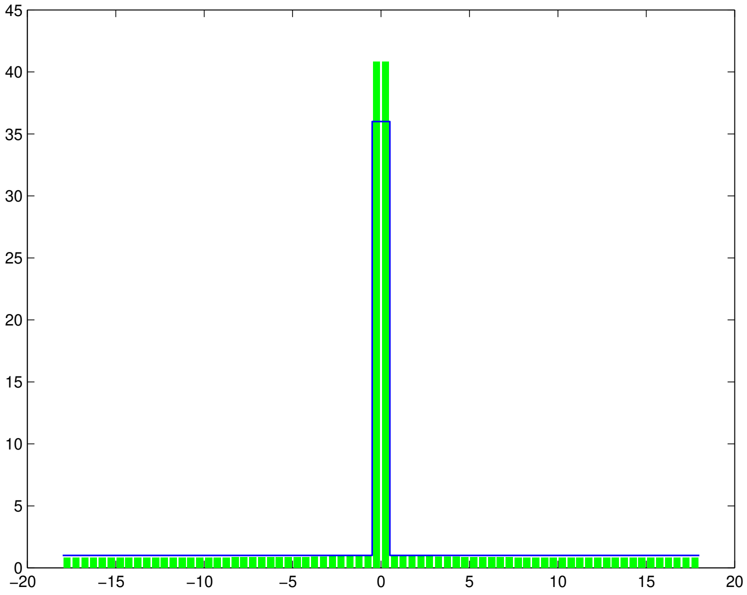

Figure 2: *The solid line represents the 1-dimensional projection of true target distribution of Example 1 simulated with the AP algorithm calculating the covariance from 200 previous states. The histogram shows the mean on 100 runs with 100000 states.*

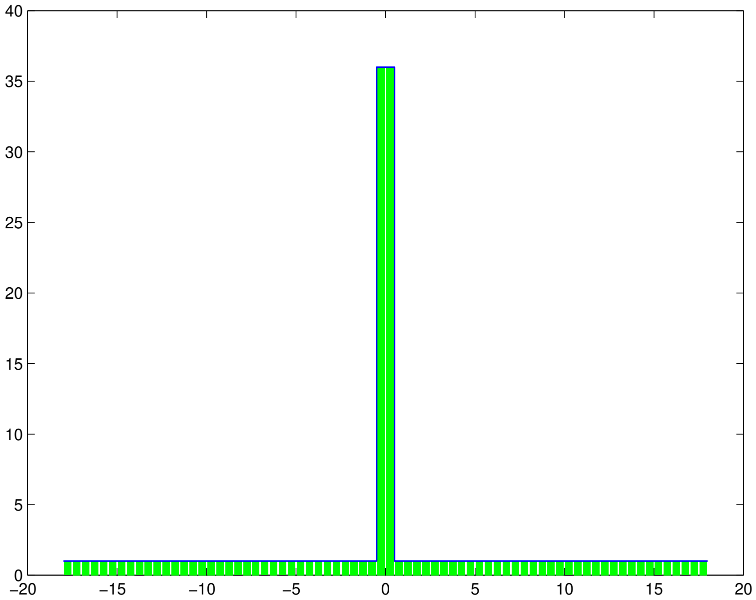

Figure 3: *The solid line represents the 1-dimensional projection of true target distribution of Example 1 simulated with the AM algorithm. The histogram shows the mean on 100 runs with 100000 states.*

Addresses of the authors:

H. HAARIO: Department of Mathematics, P.O. Box 4 (Yliopistonkatu 5), FIN-00014 University of Helsinki, Finland. E-mail : *heikki.haario@helsinki.fi*

E. SAKSMAN: Department of Mathematics, P.O. Box 4 (Yliopistonkatu 5), FIN-00014 University of Helsinki, Finland. E-mail : *eero.saksman@helsinki.fi*

J. TAMMINEN: Finnish Meteorological Institute, Geophysics Research, P.O. Box 503, FIN-00101 Helsinki, Finland. E-mail : *johanna.tamminen@fmi.fi*