Stat Comput (2012) 22:997–1008
DOI 10.1007/s11222-011-9269-5

Robust adaptive Metropolis algorithm with coerced acceptance rate

Matti Vihola

Received: 19 November 2010 / Accepted: 1 July 2011 / Published online: 4 August 2011

Abstract The adaptive Metropolis (AM) algorithm of Haario, Saksman and Tamminen (Bernoulli 7(2):223–242, 2001) uses the estimated covariance of the target distribution in the proposal distribution. This paper introduces a new robust adaptive Metropolis algorithm estimating the shape of the target distribution and simultaneously coercing the acceptance rate. The adaptation rule is computationally simple adding no extra cost compared with the AM algorithm. The adaptation strategy can be seen as a multidimensional extension of the previously proposed method adapting the scale of the proposal distribution in order to attain a given acceptance rate. The empirical results show promising behaviour of the new algorithm in an example with Student target distribution having no finite second moment, where the AM covariance estimate is unstable. In the examples with finite second moments, the performance of the new approach seems to be competitive with the AM algorithm combined with scale adaptation.

Keywords Acceptance rate · Adaptive Markov chain Monte Carlo · Ergodicity · Metropolis algorithm · Robustness

1 Introduction

Markov chain Monte Carlo (MCMC) is a general method to approximate integrals of the form

$$I := \int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x < \infty$$

where π is a probability density function which can be evaluated point-wise up to a normalising constant. Such an integral occurs frequently when computing Bayesian posterior expectations (e.g., Robert and Casella 1999; Gilks et al. 1998; Roberts and Rosenthal 2004). The MCMC method is based on a Markov chain $(X_n)_{n\ge 1}$ that is easy to simulate in practice, and for which the ergodic averages $I_n := n^{-1}\sum_{k=1}^{n} f(X_k)$ converge to the integral I as the number of samples n tends to infinity.

One of the most generally applicable MCMC method is the random walk Metropolis (RWM) algorithm. Suppose q is a symmetric probability density supported on $\mathbb{R}^d$ (for example the standard Gaussian density) and let $S \in \mathbb{R}^{d\times d}$ be a non-singular matrix. Set $X_1 \equiv x_1$, where $x_1 \in \mathbb{R}^d$ is a given starting point in the support; $\pi(x_1) > 0$. For $n \ge 2$ apply recursively the following two steps:

(M1) simulate $Y_n = X_{n-1} + SU_n$, where $U_n \sim q$ is a independent random vector, and
(M2) with probability $\alpha_n := \alpha(X_{n-1}, Y_n) := \min\{1, \pi(Y_n)/\pi(X_{n-1})\}$ the proposal is accepted, and $X_n = Y_n$; otherwise the proposal is rejected and $X_n = X_{n-1}$.

This algorithm will produce a valid chain, that is, $I_n \to I$ almost surely as $n \to \infty$ (e.g. Nummelin 2002, Theorem 1). However, the efficiency of the method, that is, the speed of the convergence $I_n \to I$, is crucially affected by the choice of the shape matrix S.

Recently, there has been an increasing interest on adaptive MCMC algorithms that try to learn some properties of the target distribution π on-the-fly, and use this information to facilitate more efficient sampling (Haario et al. 2001; Andrieu and Robert 2001; Atchadé and Rosenthal 2005; Andrieu and Moulines 2006; Roberts and Rosenthal 2007, 2009); see also the recent review by Andrieu and Thoms (2008). In the context of the RWM algorithm, this is typically implemented by replacing the constant shape S in (M1)

M. Vihola (✉)
Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box 35, 40014 University of Jyväskylä, Finland
e-mail: matti.vihola@iki.fi

with a random matrix S_{n-1} that depends on the past (on the random variables U_k, X_k, and Y_k for $1 \le k \le n-1$).

Different strategies have been proposed to compute the matrix S_{n-1}. The seminal Adaptive Metropolis (AM) algorithm (Haario et al. 2001) uses $S_{n-1} = \theta L_{n-1}$ where L_{n-1} is the Cholesky factor of the (possibly modified) empirical covariance matrix $C_{n-1} = \mathrm{Cov}(X_1, \ldots, X_{n-1})$. Under certain assumptions, the empirical covariance converges to the true covariance of the target distribution π (see, e.g., Haario et al. 2001; Andrieu and Moulines 2006; Saksman and Vihola 2010; Vihola 2011a). The constant scaling parameter $\theta > 0$ is a tuning parameter chosen by the user; the value $\theta = 2.4/\sqrt{d}$ proposed in the original paper is widely used, as it is asymptotically optimal under certain theoretical setting (Gelman et al. 1996).

In fact, the theory behind the value $\theta = 2.4/\sqrt{d}$ connects the mean acceptance rate to the efficiency of the Metropolis algorithm in more general settings. Therefore, it is sensible to try to find such a scaling factor θ that yields a desired mean acceptance rate; typically 23.4% in multidimensional settings (Roberts et al. 1997). The first algorithms coercing the acceptance rate did not adapt the shape factor at all, but only the scale of the proposal distribution. That is, $S_{n-1} = \theta_{n-1} I$, a multiple of a constant matrix, where the factor $\theta_{n-1} \in (0, \infty)$ is adapted roughly by increasing the value of the acceptance probability is too low, and vice versa (Atchadé and Rosenthal 2005; Andrieu and Thoms 2008; Roberts and Rosenthal 2009; Atchadé and Fort 2010). This adaptive scaling Metropolis (ASM) algorithm has some nice properties, and it has been shown that the algorithm is stable under quite a general setting (Vihola 2011b). It is, however, a ‘one-dimensional’ scheme, in the sense that it is unable to adapt to the shape of the target distribution like the AM algorithm. This can result in slow mixing with certain target distributions π having a strong correlation structure.

The scale adaptation in the ASM approach has been proposed to be used within the AM algorithm (Atchadé and Fort 2010; Andrieu and Thoms 2008). This algorithm, which shall be referred here to as the adaptive scaling within AM (ASWAM), combines the shape adaptation of AM and the acceptance probability optimisation. Namely, $S_{n-1} = \theta_{n-1} L_{n-1}$, where θ_{n-1} is computed from the observed acceptance probabilities $\alpha_2, \ldots, \alpha_{n-1}$ and L_{n-1} is the Cholesky factor of $\mathrm{Cov}(X_1, \ldots, X_{n-1})$. This multi-criteria adaptation framework provides a coerced acceptance probability, and at the same time captures the covariance shape information of π. Empirical findings indicate this algorithm can overcome some difficulties encountered with the AM method (Andrieu and Thoms 2008).

The present paper introduces a new algorithm alternative to the ASWAM approach. The aim is to seek a matrix factor S_* that captures the shape of π and at the same time allows to attain a given mean acceptance rate. Unlike the multi-criteria adaptation in ASWAM, the new approach is based on a single matrix update formula that is computationally equivalent to the covariance factor update in AM. The algorithm, called here the robust adaptive Metropolis (RAM), differs from the ASWAM approach by avoiding the use of the empirical covariance, which can be problematic in some settings, especially if π has no finite second moment. The proposed approach is reminiscent, yet not equivalent, with robust pseudo-covariance estimation, which has also been proposed to be used in place of the AM approach (Andrieu and Thoms 2008).

The RAM algorithm is described in detail in the next section. Section 3 provides analysis on the stable points of the adaptation rule, that is, where the sequence of matrices S_n is supposed to converge. In Sect. 4, the validity of the algorithm is verified under certain sufficient conditions. It is also shown that the adaptation converges to a shape of an elliptically symmetric target distribution. The RAM algorithm was empirically tested in some example settings and compared with the AM and the ASWAM approaches. Section 5 summarises the encouraging findings. The final section concludes with some discussion on the approach as well as directions of further research.

2 Algorithm

In what follows, suppose that the proposal density q is spherically symmetric: there exists a function $\hat{q} : \mathbb{R} \to [0, \infty)$ such that $q(x) = \hat{q}(\|x\|)$ for all $x \in \mathbb{R}^d$. Let $s_1 \in \mathbb{R}^{d \times d}$ be a lower-diagonal matrix with positive diagonal elements, and suppose $\{\eta_n\}_{n \ge 1} \subset (0, 1]$ is a step size sequence decaying to zero. Furthermore, let $x_1 \in \mathbb{R}^d$ be some point in the support of the target distribution, $\pi(x_1) > 0$, and let $\alpha_* \in (0, 1)$ stand for the target mean acceptance probability of the algorithm.

The robust adaptive Metropolis process is defined recursively through

(R1) compute $Y_n := X_{n-1} + S_{n-1} U_n$, where $U_n \sim q$ is an independent random vector,

(R2) with probability $\alpha_n := \min\{1, \pi(Y_n)/\pi(X_{n-1})\}$ the proposal is accepted, and $X_n = Y_n$; otherwise the proposal is rejected and $X_n = X_{n-1}$, and

(R3) compute the lower-diagonal matrix S_n with positive diagonal elements satisfying the equation

$$S_n S_n^T = S_{n-1} \left(I + \eta_n (\alpha_n - \alpha_*) \frac{U_n U_n^T}{\|U_n\|^2} \right) S_{n-1}^T \qquad (1)$$

where $I \in \mathbb{R}^{d \times d}$ stands for the identity matrix.

The steps (R1) and (R2) implement one iteration of the RWM algorithm, but with a random matrix S_{n-1} in (R1). In the adaptation step (R3) the unique S_n satisfying (1) always exists, since it is the Cholesky factor of the matrix in the right hand side, which is verified below to be symmetric and positive definite.

Proposition 1 *Suppose $S \in \mathbb{R}^{d\times d}$ is a non-singular matrix, $u \in \mathbb{R}^d$ is a non-zero vector and $a \in (-1,\infty)$ is a scalar. Then, the matrix $M := S(I + a\frac{uu^T}{\|u\|^2})S^T$ is symmetric and positive definite.*

Proof The symmetricity is obvious. Let $x \in \mathbb{R}^d \setminus \{0\}$, denote $\tilde{u} := u/\|u\|$ and define $z := S\tilde{u}$. We may write $M = SS^T + azz^T$, whence

$$x^T Mx = \|x^T S\|^2 + a(x^T z)^2 = \|x^T S\|^2 \left(1 + a\frac{(x^T z)^2}{\|x^T S\|^2}\right).$$

This already establishes the claim in the case $a \geq 0$. Suppose then $a \in (-1,0)$. Clearly $(x^T z)^2 = \|x^T S\tilde{u}\|^2 \leq \|x^T S\|^2$ and so $x^T Mx \geq \|x^T S\|^2(1-|a|) > 0$. □

Let us then see what happens in the adaptation in intuitive terms. Observe first that in (R1) the proposal Y_n is formed by adding an increment $W_n := S_{n-1}U_n$ to the previous point X_{n-1}. Since U_n is distributed according to the spherically symmetric q, the random variable W_n is distributed according to the elliptically symmetric density $q_{S_{n-1}}(w) := \det(S_{n-1})^{-1} q(S_{n-1}^{-1}w)$ with the main axes defined by the eigenvectors and the corresponding eigenvalues of the matrix $S_{n-1}S_{n-1}^T$.

To illustrate the behaviour of the RAM update (R3), Fig. 1 shows two examples how the contours of the proposal change in the update. The example on the left shows how the contour ellipsoid expands to the direction of W_n when $\eta_n(\alpha_n - \alpha_*) = 0.8 > 0$. Similarly, the example on the right shows how the ellipsoid shrinks when $\eta_n(\alpha_n - \alpha_*) = -0.8 < 0$. These examples reflect the basic idea behind the approach. If the acceptance probability is smaller than desired, $\alpha_n < \alpha_*$ (or more than desired, $\alpha_n > \alpha_*$) the proposal distribution is shrunk (or expanded) with respect to the direction of the current proposal increment.

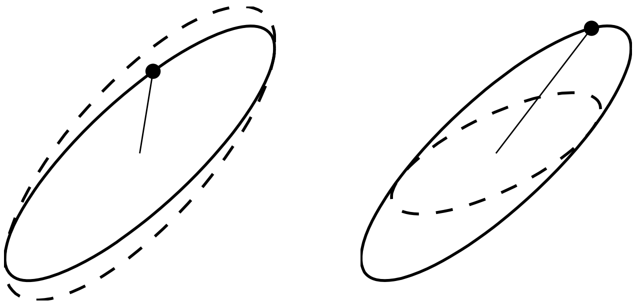

Fig. 1 Two examples of the RAM update (R3). The *solid line* represents the contour ellipsoid defined by $S_{n-1}S_{n-1}^T$, and the vector $S_{n-1}U_n/\|U_n\|$ is drawn as a *dot*. The contours defined by $S_nS_n^T$ are *dashed*

We can also see this behaviour from the update equation by considering the radius of the contour ellipsoid defined by $S_nS_n^T$ with respect to different directions. Let $v \in \mathbb{R}^d$ be a unit vector. As in the proof of Proposition 1, we may write

$$\|S_n^T v\|^2 = \|S_{n-1}^T v\|^2 + \eta_n(\alpha_n - \alpha_*)(Z_n^T v)^2$$

where $Z_n = S_nU_n/\|U_n\|$. If Z_n and v are orthogonal, the latter term vanishes and $\|S_n^T v\| = \|S_{n-1}^T v\|$. If they are parallel, that is, $v = \pm Z_n/\|Z_n\|$, then the factor $(Z_n^T v)^2$ equals $\|S_{n-1}^T v\|^2$, and so $\|S_n^T v\| = \sqrt{1 + \eta_n(\alpha_n - \alpha_*)}\|S_{n-1}^T v\|$. Any other choices of the unit vector v fall in between these two extremes.

Remark 1 In dimension one, the value of S_n can be computed directly by

$$\log S_n = \log S_{n-1} + \frac{1}{2}\log\bigl(1 + \eta_n(\alpha_n - \alpha_*)\bigr).$$

When η_n is small, this is almost equivalent to the update

$$\log S_n = \log S_{n-1} + \frac{\eta_n}{2}(\alpha_n - \alpha_*)$$

implying that the RAM algorithm will exhibit a similar behaviour with the ASM algorithm as proposed by Atchadé and Fort (2010), Andrieu and Thoms (2008) and analysed in Vihola (2011b). Therefore, it is justified to consider RAM as a multidimensional generalisation of the ASM adaptation rule.

Remark 2 In practice, the matrix S_n in (R3) can be computed as a rank one Cholesky update or downdate of S_{n-1} when $\alpha_n - \alpha_* > 0$ and $\alpha_n - \alpha_* < 0$, respectively (Dongarra et al. 1979). Therefore, the algorithm is computationally efficient up to a relatively high dimension. In fact, the full d-dimensional matrix multiplication required when generating the proposal in (R1) has the same $O(d^2)$ complexity as the Cholesky update or downdate, rendering the adaptation to only add a constant factor to the complexity of the RWM algorithm.

Remark 3 While the step size sequence η_n can be chosen quite freely, in practice it is often defined as $\eta_n = n^{-\gamma}$ with an exponent $\gamma \in (1/2, 1]$. The choice $\gamma = 1$, which is employed in the original setting of the AM algorithm (Haario et al. 2001) is not advisable for the RAM algorithm. For simplicity, consider a one-dimensional setting like in Remark 1. Then, if $\eta_n = n^{-1}$ the logarithm of S_n can increase or decrease only at the speed $\pm\sum_{k=1}^n \eta_k \approx \log(n)$. Therefore, S_n can grow or shrink only linearly or at the speed $1/n$, respectively. This renders the adaptation inefficient, if the initial value s_1 differs significantly from the scale and shape of π.

3 Stable points

The RAM algorithm introduced in the previous section has, under suitable conditions, a *stable point*, that is, a matrix $S_* \in \mathbb{R}^{d\times d}$, where the adaptation process S_n should converge as n increases. Before considering the convergence, we shall study the stable points of the algorithm in certain settings.

One can write the update equation (1) in the following form

$$S_n S_n^T = S_{n-1}S_{n-1}^T + \eta_n H(S_{n-1}, X_{n-1}, U_n) \tag{2}$$

where

$$H(S,x,u) = S\left(\min\left\{1, \frac{\pi(x+Su)}{\pi(x)}\right\} - \alpha_*\right)\frac{uu^T}{\|u\|^2}S^T.$$

The recursion (2) implements a so called Robbins-Monro stochastic approximation algorithm on $(S_nS_n^T)_{n\geq 1}$ (e.g. Benveniste et al. 1990; Kushner and Yin 2003; Borkar 2008). Such an algorithm seeks the root of the so called mean field h_π defined as

$$h_\pi(S) := S\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\min\left\{1, \frac{\pi(x+Su)}{\pi(x)}\right\} - \alpha_*\right) \\ \times \frac{uu^T}{\|u\|^2} q(u)\mathrm{d}u\pi(x)\mathrm{d}x S^T.$$

We shall see that under some sufficient conditions, there exists a stable point, that is, $h_\pi(S) = 0$.

First, we shall observe a fundamental property of the RAM algorithm; that it is invariant under affine transformations.

Theorem 1 *Let π be a probability density and let $(X_n, S_n)_{n\geq 1}$ be the RAM process* (R1)–(R3) *targeting π and started from (x_1, s_1). Suppose $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix, $b \in \mathbb{R}^d$ and define $\tilde{\pi}(x) := |\det(A)|^{-1}\pi(A^{-1}x - b)$. Let $(\hat{X}_n, \hat{S}_n)_{n\geq 1}$ be the RAM process targeting $\tilde{\pi}$ and started from $(Ax_1 + b, As_1)$. Then, the processes $(AX_n + b, (AS_n)(AS_n)^T)_{n\geq 1}$ and $(\hat{X}_n, \hat{S}_n\hat{S}_n^T)_{n\geq 1}$ have identical distributions.*

Proof Let $U_n \sim q$ and $W_n \sim U(0,1)$ be the independent sequences that drive the RAM process $(X_n, S_n)_{n\geq 1}$ targeting π; that is

$$Y_n = X_{n-1} + S_{n-1}U_n \tag{3}$$

$$X_n = Y_n \mathbb{1}_{\{W_n \leq \alpha_n\}} + X_n \mathbb{1}_{\{W_n > \alpha_n\}}. \tag{4}$$

The proof proceeds by constructing an independent sequence $\hat{U}_n \sim q$, so that the RAM process $(\tilde{X}_n, \tilde{S}_n)_{n\geq 1}$ targeting $\tilde{\pi}$ and driven by $(\tilde{U}_n)_{n\geq 1}$ and $(W_n)_{n\geq 1}$ will satisfy the claim path-wise: $AX_n = \hat{X}_n$ and $AS_n(AS_n)^T = \hat{S}_n\hat{S}_n^T$ for all $n \geq 1$.

Write the QR decomposition $(AS_n)^T = Q_nR_n$ where Q_n is orthogonal and where $\hat{S}_n := R_n^T$ is lower-diagonal and chosen so that it has a positive diagonal. We observe that $AS_n(AS_n)^T = \hat{S}_n\hat{S}_n^T$ and defining $\hat{U}_{n+1} := Q_n^T U_{n+1}$ we have also $AS_nU_{n+1} = \hat{S}_n\hat{U}_{n+1}$. Since the distribution of U_{n+1} is spherically symmetric and U_{n+1} is independent of orthogonal Q_n, the sequence $(\tilde{U}_n)_{n\geq 1}$ is i.i.d. with distribution q.

Now, we may verify inductively using (3) and (4) that $\hat{X}_n = AX_n$ can be computed through

$$\hat{Y}_n = \hat{X}_{n-1} + \hat{S}_{n-1}\hat{U}_n$$

$$\hat{X}_n = \hat{Y}_n\mathbb{1}_{\{W_n \leq \hat{\alpha}_n\}} + X_{n-1}\mathbb{1}_{\{W_n > \hat{\alpha}_n\}}$$

where

$$\hat{\alpha}_n = \min\left\{1, \frac{\hat{\pi}(\hat{Y}_n)}{\hat{\pi}(\hat{X}_{n-1})}\right\} = \min\left\{1, \frac{\pi(Y_n)}{\pi(X_{n-1})}\right\} = \alpha_n.$$ □

After Theorem 1, it is no surprise that the mean field of the algorithm satisfies similar invariance properties.

Theorem 2 *Suppose π is a probability density.*

(i) *Let $\hat{\pi}$ be an affine transformation of π, that is, $\hat{\pi}(x) = |\det(A)|^{-1}\pi(A^{-1}x - b)$ for some non-singular matrix $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. Then, $Ah_\pi(S)A^T = h_{\hat{\pi}}(AS)$ for all $S \in \mathbb{R}^{d\times d}$.*

(ii) *For any orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ and for all $S \in \mathbb{R}^{d\times d}$, $h_\pi(S) = h_\pi(SQ)$.*

(iii) *Suppose that S is a unique lower-diagonal matrix with positive diagonal satisfying $h_\pi(S) = 0$. Then, restricted to such matrices, the solution of $h_{\hat{\pi}}(\hat{S}) = 0$ is also unique, and of the form $\hat{S} = ASQ$ for some orthogonal $Q \in \mathbb{R}^{d\times d}$.*

Proof The claim (i) follows by a change of variable $x = A^{-1}z - b$,

$$\begin{aligned} h_\pi(S) &= S\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\min\left\{1, \frac{\pi(x+Su)}{\pi(x)}\right\} - \alpha_*\right) \\ &\quad\times \pi(x)\mathrm{d}x\frac{uu^T}{\|u\|^2}q(u)\mathrm{d}uS^T \\ &= S\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\min\left\{1, \frac{\hat{\pi}(z+ASu)}{\hat{\pi}(z)}\right\} - \alpha_*\right) \\ &\quad\times \hat{\pi}(z)\mathrm{d}z\frac{uu^T}{\|u\|^2}q(u)\mathrm{d}uS^T = A^{-1}h_{\hat{\pi}}(AS)A^{-T}. \end{aligned}$$

The claim (ii) follows similarly, by a change of variable $u = Qv$ and due to the spherical symmetry of q. The uniqueness up to rotations, that is, only the matrices of the form $\hat{S} = ASQ$ satisfy $h_{\hat{\pi}}(\hat{S}) = 0$ follows directly as above. The claim (iii) is completed by writing the QR-decomposition $(AS)^T = QR$ and by observing that the upper-triangular R can be chosen to have positive diagonal elements. □

Theorem 2 verifies that the stable points of the algorithm are affinely invariant like the covariance (or more generally robust pseudo-covariance) matrices (Huber 1981). Theorem 3 below verifies that in the case of a suitable elliptically symmetric target distribution π, the stable points of the RAM algorithm in fact coincide with the (pseudo-)covariance of π. This is an interesting connection, but in general the fixed points of the RAM algorithm do not coincide with the pseudo-covariance.

Theorem 3 *Assume $\alpha_* \in (0,1)$ and π is elliptically symmetric, that is, $\pi(x) \equiv \det(\Sigma)^{-1} p(\|\Sigma^{-1}x\|)$ for some $p : [0,\infty) \to [0,\infty)$ and for some symmetric and positive definite $\Sigma \in \mathbb{R}^{d\times d}$. Then,*

(i) *there exists a lower-diagonal matrix with positive diagonal $S_* \in \mathbb{R}^{d\times d}$ such that $h_\pi(S_*) = 0$ and such that $S_* S_*^T$ is proportional to Σ^2.*
(ii) *assuming the function p is non-increasing, the solution S_* is additionally unique.*

Proof In light of Theorem 2, it is sufficient to consider any spherically symmetric π, that is, the case Σ is an identity matrix.

Let S be a lower-diagonal matrix with positive diagonal. Observe that since S is non-singular, $h_\pi(S) = 0$ is equivalent to $S^{-1} h_\pi(S) S^{-T} = 0$, that is

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\min\left\{1, \frac{\pi(x+Su)}{\pi(x)}\right\} - \alpha_*\right) \times \frac{uu^T}{\|u\|^2} q(u)\mathrm{d}u\pi(x)\mathrm{d}x = 0. \tag{5}$$

Define the function

$$\bar{h}(S) := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\min\left\{1, \frac{\pi(x+Su)}{\pi(x)}\right\}\right) \times \frac{uu^T}{\|u\|^2} q(u)\mathrm{d}u\pi(x)\mathrm{d}x.$$

It is easy to see by symmetry and taking traces that (5) is equivalent to $\bar{h}(S) = \frac{\alpha_*}{d} I$, where $I \in \mathbb{R}^{d\times d}$ stands for the identity matrix.

We can write $\bar{h}(S)$ in a more convenient form by using the polar coordinate representation $u = rv$, where $v \in \mathcal{S}^d := \{v \in \mathbb{R}^d : \|v\| = 1\}$ is a unit vector in the unit sphere, and $r = \|u\|$ is the length of u. Then, by Fubini's theorem

$$\bar{h}(S) = \int_{\mathcal{S}^d}\left[\int_0^\infty \int_{\mathbb{R}^d} \min\{\pi(x), \pi(x + rSv)\}\mathrm{d}x\tilde{q}(r)\mathrm{d}r\right] \times vv^T \mu(\mathrm{d}v)$$

where μ stands for the uniform distribution on the unit sphere $\mathcal{S}^d$ and the proposal is written in terms of the probability density $\tilde{q}(\|u\|) \propto \|u\|^{d-1} q(u)$.

By applying the representation of π by the radial function p one can write the term above in brackets as

$$g(\|Sv\|) := \int_0^\infty \int_{\mathbb{R}^d} \min\{p(\|x\|), p(\|x + rSv\|)\}\mathrm{d}x\tilde{q}(r)\mathrm{d}r,$$

since due to symmetry, the value of the integral depends only on the norm $\|Sv\|$.

For any $\theta \in \mathbb{R}_+$, one can now write

$$\bar{h}(\theta I) = \int_{\mathcal{S}^d} g(\theta) vv^T \mu(\mathrm{d}v) = \frac{g(\theta)}{d} I,$$

since $\operatorname{trace}(\bar{h}(\theta I)) = g(\theta)$ and by symmetry. Proposition 2 in Appendix A shows that $g : (0,\infty) \to (0,\infty)$ is continuous, that $\lim_{\theta\to\infty} g(\theta) = 0$ and that $\lim_{\theta\to 0+} g(\theta) = \int_0^\infty \tilde{q}(r)\mathrm{d}r = 1$. Therefore, there exists a $\theta_* > 0$ such that $g(\theta_*) = \alpha_*$ so that $\bar{h}(\theta_* I) = \frac{\alpha_*}{d} I$, establishing (i).

For (ii), let us first show that g is in this case strictly decreasing, at least before hitting zero. Observe that since p is non-increasing, one can write

$$\begin{aligned} g(\theta) &= \int_0^\infty \left(\int_{\|x\|>\|x+r\theta v\|} p(\|x\|)\mathrm{d}x \right. \\ &\qquad \left. + \int_{\|x\|\le\|x+r\theta v\|} p(\|x+r\theta v\|)\mathrm{d}x\right)\tilde{q}(r)\mathrm{d}r \\ &= \int_0^\infty \left(1 - \int_{A_{r\theta v}} \pi(x)\mathrm{d}x\right)\tilde{q}(r)\mathrm{d}r. \end{aligned}$$

It is easy to see that the width of the strip $A_{r\theta v} := \{\|x\| \le \|x + r\theta v\|\} \cap \{\|x\| < \|x - r\theta v\|\}$ is increasing with respect to θ. Therefore, for any fixed r and v, the term $b_{rv}(\theta) := 1 - \int_{A_{r\theta v}} \pi(x)\mathrm{d}x$ is strictly decreasing with respect to θ as long as the support of π is not completely covered by $A_{r\theta v}$, in which case $b_{rv}(\theta) = 0$. This implies that $g(\theta)$ is strictly decreasing with respect to θ, until possibly $g(\theta) = 0$. Therefore, there is a unique $\theta_* > 0$ for which $g(\theta_*) = \alpha_*$.

Let us assume that $S \in \mathbb{R}^{d\times d}$ is a matrix satisfying $\bar{h}(S) = \frac{\alpha_*}{d} I$. By symmetry, we can assume S to be diagonal, with positive diagonal elements $s_1, \ldots, s_d > 0$. Let $e_1, \ldots, e_d$ stand for the standard basis vectors of $\mathbb{R}^d$. The diagonal element $[\bar{h}(S)]_{ii} = \frac{\alpha_*}{d}$ is equivalent to

$$\int_{\mathcal{S}^d} \left[g(\|Sv\|) - \alpha_*\right](v^T e_i)^2 \mu(\mathrm{d}v) = 0,$$

since $\int_{\mathcal{S}^d} (v^T e_i)^2 \mu(\mathrm{d}v) = d^{-1}$. Denoting $\bar{g}(\|Sv\|) := g(\|Sv\|) - \alpha_*$, this implies

$$\int_{\mathcal{S}^d} \bar{g}\left(\left(\sum_{i=1}^d s_i^2 v_i^2\right)^{1/2}\right)\left(\sum_{i=1}^d \lambda_i v_i^2\right)\mu(\mathrm{d}v) = 0 \tag{6}$$

for any choice of the constants $\lambda_i \in \mathbb{R}$. Particularly, choosing $\lambda_i = 1$ for $i = 1, \ldots, d$ implies that for any constant $c \in \mathbb{R}$

we have

$$\int_{\mathcal{S}^d} \bar{g}\left(\left(\sum_{i=1}^{d} s_i^2 v_i^2\right)^{1/2}\right) c\mu(\mathrm{d}v) = 0. \tag{7}$$

Now, summing (6) and (7) with a specific choice of constants $c = \theta_*^2$ and $\lambda_i = -s_i^2$, we obtain

$$\int_{\mathcal{S}^d} \bar{g}\left(\left(\sum_{i=1}^{d} s_i^2 v_i^2\right)^{1/2}\right)\left(\theta_*^2 - \sum_{i=1}^{d} s_i^2 v_i^2\right)\mu(\mathrm{d}v) = 0.$$

But now, $\bar{g}((\sum_{i=1}^d s_i^2 v_i^2)^{1/2}) \geq 0$ exactly when $\sum_{i=1}^d s_i^2 v_i^2 \leq \theta_*^2$, so the integrand is always non-negative. Moreover, if any $s_i \neq \theta_*$, then by continuity there is a neighbourhood $U_i \subset \mathcal{S}^d$ of e_i such that the integrand is strictly positive, implying that the integral is strictly positive. This concludes the proof of the uniqueness (ii). □

The following theorem shows that when π is the joint density of d independent and identically distributed random variables, the RAM algorithm has, as expected, a stable point proportional to the identity matrix.

Theorem 4 *Assume $\alpha_* \in (0,1)$ and $\pi(x) = \prod_{i=1}^d p(x_i)$ for some one-dimensional density p. Then, there exists a $\theta > 0$ such that $\hat{h}(\theta I) = 0$.*

Proof Let $e_1, \ldots, e_d$ stand for the coordinate vectors of $\mathbb{R}^d$, and let $\tilde{q}$ and μ be defined as in the proof of Theorem 3. Consider the functions

$$a_i(\theta) := \int_{\mathcal{S}^d}\int_0^\infty \left(\int_{\mathbb{R}^d} \min\{\pi(x), \pi(x + r\theta u)\}\,\mathrm{d}x\right) \times \tilde{q}(r)\mathrm{d}r(u^T e_i)^2 \mu(\mathrm{d}u).$$

Let P be a permutation matrix. It is easy to see that $\pi(x + r\theta u) = \pi(P(x + r\theta u))$ by the i.i.d. product form of π. Therefore, by the change of variable $Px = z$ and $Pu = v$, one obtains that

$$a_i(\theta) = \int_{\mathcal{S}^d}\int_0^\infty \left(\int_{\mathbb{R}^d} \min\{\pi(z), \pi(z + r\theta v)\}\,\mathrm{d}x\right) \times \tilde{q}(r)\mathrm{d}r(v^T P^T e_i)^2 \mu(\mathrm{d}v) = a_j(\theta)$$

by a suitable choice of P. Moreover, $\lim_{\theta\to\infty} a_i(\theta) = 0$ and $\lim_{\theta\to 0+} a_i(\theta) = d^{-1}$ and a_i are continuous. Therefore, there exists a $\theta_* > 0$ such that $a_i(\theta_*) = \alpha_* d^{-1}$, and so $e_i^T h(\theta_* I) e_i = 0$.

It remains to show that $e_i h(\theta_* I) e_j = 0$ for all $i \neq j$. But for this, it is enough to show that the integrals of the form

$$\int_{E_{i,j}^*}\int_0^\infty \left(\int_{\mathbb{R}^d} \min\{\pi(z), \pi(z + r\theta v)\}\,\mathrm{d}x\right) \times \tilde{q}(r)\mathrm{d}r|(v^T e_i)(v^T e_j)|\mu(\mathrm{d}v)$$

have the same value for both $E_{i,j}^+ := \{v \in \mathcal{S}^d : (v^T e_i)(v^T e_j) > 0\}$ and $E_{i,j}^- := \{v \in \mathcal{S}^d : (v^T e_i)(v^T e_j) < 0\}$. But this is obtained due to the symmetry of the sets $E_{i,j}^+$ and $E_{i,j}^-$ and the product form of π, since

$$\int_{\mathbb{R}^d} \min\{\pi(z), \pi(z + r\theta v)\}\,\mathrm{d}x = \int_{\mathbb{R}^d} \min\left\{\pi\left(z - \frac{1}{2}r\theta v\right), \pi\left(z + \frac{1}{2}r\theta v\right)\right\}\mathrm{d}x$$

so one can change the sign of any coordinate of v without affecting this integral. □

Remark 4 Checking the existence and uniqueness of the stable point in a more general setting is out of the scope of this paper. It is believed that there always exists at least one solution $S_* \in \mathbb{R}^{d\times d}$ such that $h(S_*) = 0$. Notice, however, that the fixed point may not be always unique; see an example of such a situation for one-dimensional adaptation (the ASM algorithm) in Hastie (2005, Sect. 4.4).

Remark 5 It is not very difficult to show that for any given target π and proposal q, there exist some constants $0 < \theta_1 < \theta_2 < \infty$ such that the matrices $h_\pi(\theta_1 I)$ and $h_\pi(\theta_2 I)$ are positive definite and negative definite, respectively. This indicates that, on average, S_n should shrink whenever it is 'too big' and expand whenever it is 'too small,' so the algorithm should admit a stable behaviour. The empirical results in Sect. 5 support the hypothesis of general stability.

To be more precise, we can identify a Lyapunov function w_π for h_π in the case π is elliptically symmetric with a non-increasing tail. This will allow us to establish the convergence of the sequence $(S_n S_n^T)_{n\geq 1}$ in Theorem 7.

Theorem 5 *Assume the conditions of Theorem 3(ii) and denote $R_* := S_* S_*^T$. Define a function $w_\pi : \mathbb{R}^{d\times d} \to [0,\infty)$ by*

$$w_\pi(R) := \mathrm{trace}(R_*^{-1}R) - \log\left(\frac{\det R}{\det R_*}\right) - d.$$

Then, for any non-singular $S \in \mathbb{R}^{d\times d}$ it holds that $\langle\nabla w_\pi(SS^T), h_\pi(S)\rangle \leq 0$ with equality only if $SS^T = R_*$.

Proof Denote $\hat{\pi}(x) := \det(R_*)^{1/2}\pi(R_*^{1/2}x)$, then by Theorem 2(i) $h_\pi(S) = R_*^{1/2} h_{\hat{\pi}}(R_*^{-1/2}S)R_*^{1/2}$. Moreover, Theorem 3(ii) together with Theorem 2(iii) imply that $\hat{\pi}$ is spherically symmetric and $S = I$ is the unique solution of $h_{\hat{\pi}}(S) = 0$ (up to orthogonal transformations).

We can write

$$\nabla w_\pi\big(R_*^{1/2}S(R_*^{1/2}S)^T\big) = R_*^{-1/2}\big(I - (SS^T)^{-1}\big)R_*^{-1/2} = R_*^{-1/2}\nabla w_{\hat{\pi}}(S)R_*^{-1/2},$$

so we obtain

$$\begin{aligned}&\big\langle \nabla w_\pi\big(R_*^{1/2}S(R_*^{1/2}S)^T\big), h_\pi(R_*^{1/2}S\big\rangle\\ &\quad=\mathrm{trace}\big[\nabla w_\pi\big(R_*^{1/2}S(R_*^{1/2}S)^T\big)^T h_\pi(R_*^{1/2})\big]\\ &\quad=\big\langle \nabla w_{\hat\pi}(S), h_{\hat\pi}(S)\big\rangle.\end{aligned}$$

Therefore, it is sufficient to check that the claim holds for spherically symmetric $\hat\pi$ with $R_* = I$.

Let S be non-singular and write the singular value decomposition $S = U\bar S V^T$ where U and V are orthogonal and $\bar S = \mathrm{diag}(\bar s_1,\ldots,\bar s_d)$ with positive diagonal entries. By Theorem 2(ii) we have $h_{\hat\pi}(S) = h_{\hat\pi}(SV) = h_{\hat\pi}(U\bar S)$. We may write, using the notation in Theorem 3,

$$\begin{aligned}\mathrm{trace}\big(h_{\hat\pi}(S)\big) &= \mathrm{trace}\big(U^T h_{\hat\pi}(U\bar S)U\big)\\ &= \int_{\mathcal{S}^d}\bar g\big(\|\bar S w\|\big)\Bigg(\sum_{i=1}^d \bar s_i^2 w_i^2\Bigg)\mu(\mathrm{d}w).\end{aligned}$$

We have $SS^T = U\bar S^2 U^T$, so we obtain similarly

$$\begin{aligned}\mathrm{trace}\big((SS^T)^{-1}h_{\hat\pi}(S)\big) &= \mathrm{trace}\big(\bar S^{-1}U^T h_{\hat\pi}(SV)U\bar S^{-1}\big)\\ &= \int_{\mathcal{S}^d}\bar g\big(\|\bar S w\|\big)\mu(\mathrm{d}w).\end{aligned}$$

Putting everything together,

$$\begin{aligned}&\big\langle \nabla w_{\hat\pi}(SS^T), h_{\hat\pi}(S)\big\rangle\\ &\quad= \int_{\mathcal{S}^d}\bar g\Bigg(\Bigg(\sum_{i=1}^d \bar s_i^2 w_i^2\Bigg)^{1/2}\Bigg)\Bigg(\sum_{i=1}^d \bar s_i^2 w_i^2 - 1\Bigg)\mu(\mathrm{d}w).\end{aligned}$$

As in the proof of Theorem 3, $\bar g((\sum_{i=1}^d \bar s_i^2 w_i^2)^{1/2}) > 0$ exactly when $\sum_{i=1}^d \bar s_i^2 w_i^2 < 1$ and vice versa. The integral can equal zero only if all $\bar s_i = 1$. □

4 Validity

This section describes some sufficient conditions under which the RAM algorithm is valid; that is, when the empirical averages converge to the integral

$$I_n = \frac{1}{n}\sum_{k=1}^n f(X_k) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x =: I \tag{8}$$

almost surely.

Let us start by introducing assumptions on the forms of the proposal density q and the target density π.

Assumption 1 The proposal density q is either a Gaussian or a Student distribution, that is,

$$q(z) \propto e^{-\frac{1}{2}\|z\|^2} \quad\text{or}\quad q(z)\propto (1+\|z\|^2)^{-\frac{d+p}{2}}$$

for some constant $p > 0$.

Assumption 2 The target density π satisfies either of the following assumptions.

(i) The density π is bounded and supported on a bounded set: there exists a constant $m < \infty$ such that $\pi(x) = 0$ for all $\|x\| \ge m$.

(ii) The density π is positive everywhere in $\mathbb{R}^d$ and continuously differentiable. The tails of π are super-exponentially decaying and have regular contours, that is, respectively

$$\lim_{\|x\|\to\infty}\frac{x}{\|x\|}\cdot\nabla\log\pi(x) = -\infty \quad\text{and}$$

$$\limsup_{\|x\|\to\infty}\frac{x}{\|x\|}\cdot\frac{\nabla\pi(x)}{\|\nabla\pi(x)\|} < 0.$$

Remark 6 Assumption 2 ensures the geometric ergodicity of the RWM algorithm under fairly general settings; Jarner and Hansen (2000) discuss the limitations of (ii) and give several examples.

Before stating the theorem, consider the following conditions on the adaptation step size sequence $(\eta_n)_{n\ge1}$ and on the stability of the process $(S_n)_{n\ge1}$.

Assumption 3 The adaptation step sizes $\eta_n \in [0,1]$ are non-increasing and satisfy $\sum_{n=1}^\infty n^{-1}\eta_n < \infty$.

Assumption 4 There exist random variables $0 \le A \le B \le \infty$ such that all the eigenvalues $\lambda_n^{(i)}$ of the random matrices $S_n S_n^T$ are almost surely bounded by $A \le \lambda_n^{(i)} \le B$, for all $n = 1, 2, \ldots$ and all $i = 1, \ldots, d$.

Theorem 6 *Suppose Assumptions 1–4 hold and denote $\Omega_0 := \{A > 0, B < \infty\}$. Suppose also that the function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies for some $p \in [0, 1)$*

$$\sup_{x\in\mathbb{R}^d:\pi(x)>0} |f(x)|\pi^{-p}(x) < \infty.$$

Then, for almost every $\omega \in \Omega_0$, the strong law of large numbers (8) holds.

The proof follows by existing results in the literature; the details are given in Appendix B.

The convergence of the adaptation can be established as well in case π is elliptically symmetric.

Theorem 7 *If the conditions of Theorem 3(ii) and Theorem 6 hold and additionally $\sum_{n=1}^\infty \eta_n = \infty$, then $S_n S_n^T \to S_* S_*^T$ for almost every $\omega \in \Omega_0$.*

The proof follows by Theorem 5 and results in the literature; see Appendix B.

Remark 7 Assumptions 1–3 are common when verifying the ergodicity of an adaptive MCMC algorithm. Assumption 4 on stability is natural but it can be difficult to check with $\mathbb{P}(A > 0,\ B < \infty) = 1$ in practice. The empirical evidence supports this hypothesis under a very general setting; see also Remark 5 in Sect. 3. Similar stability results have been established only for few adaptive MCMC algorithms, including the AM and the ASM algorithms (Saksman and Vihola 2010; Vihola 2011a, 2011b). The precise stability analysis is beyond the scope of this paper. Instead, the stability can be enforced as described below.

Let $0 < a \le b < \infty$ be some constants so that the eigenvalues of $s_1 s_1^T$ are within $[a, b]$. Then, replace the step (R3) in the RAM algorithm with the following:

(R3') compute the lower-diagonal matrix $\hat{S}_n$ with positive diagonal so that $\hat{S}_n \hat{S}_n^T$ equals the right hand side of (1). If the eigenvalues of $\hat{S}_n \hat{S}_n^T$ are within $[a, b]$, then set $S_n = \hat{S}_n$, otherwise set $S_n = S_{n-1}$.

While this modification ensures stability, it may change the stable points of the algorithm and the conclusion of Theorem 7 may not hold. This could possibly be avoided, for example, by considering an adaptive reprojections approach (Andrieu et al. 2005; Andrieu and Moulines 2006), but we do not pursue this here.

5 Experiments

The RAM algorithm was tested with three types of target distributions: heavy-tailed Student, Gaussian and a mixture of Gaussians. The performance of RAM was compared against the seminal adaptive Metropolis (AM) algorithm (Haario et al. 2001) and an adaptive scaling within adaptive Metropolis (ASWAM) algorithms (Andrieu and Thoms 2008; Atchadé and Fort 2010). Especially the comparison against ASWAM is of interest, since it attains a given acceptance rate like the RAM algorithm.

There are several parameters that are fixed throughout the experiments. The adaptation step size sequence was set to $\eta_n = n^{-2/3}$ for the AM and the ASWAM algorithms. For the RAM approach, the weight sequence was modified slightly so that $\eta_n = \min\{1, d \cdot n^{-2/3}\}$. The extra factor was added to compensate the expected growth or shrinkage of the eigenvalues being of the order d^{-1}; see the proof of Theorem 3. The target mean acceptance rate was $\alpha_* = 0.234$. In all the experiments, the Student proposal distribution of the form $q(z) = (1 + \|z\|^2)^{-\frac{d+1}{2}}$ was used. Such a heavy-tailed proposal was employed in order to have good convergence properties in case of heavy-tailed target densities (Jarner and Roberts 2007).

All the tests were performed using the publicly available Grapham software (Vihola 2010); the latest version of the software includes an implementation of the RAM algorithm.

5.1 Multivariate Student distribution

The first example is a bivariate Student distribution with $n = 1$ degrees of freedom and the following location and pseudo-covariance matrix

$$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{bmatrix},$$

respectively. That is, the target density $\pi(x) \propto (1 + x^T \Sigma^{-1} x)^{-3/2}$. Clearly, π has no second moments and thereby the empirical covariance estimate used by AM and ASWAM is deemed to be unstable in this example.

Figure 2 shows the results for one hundred runs of the algorithms. The grey area indicates the interval between the 10% and the 90% percentiles, and the black line shows the median. The top row shows the logarithm of the first diagonal element of the matrix S_n. The AM covariance grows without an upper bound as expected. When the scale adaptation is added, the ASWAM approach manages to keep the factor $S_n = \theta_n L_n$ within certain bounds, but there is a considerable variation that does not seem to vanish. This is due to the fact that L_n, the Cholesky factor of $\mathrm{Cov}(X_1, \ldots, X_n)$, grows without an upper bound but at the same time the scaling factor θ_n decays to keep the acceptance rate around the desired 23.4%. The RAM algorithm seems to converge nicely to a limiting value.

Such undesired behaviour of the AM and the ASWAM algorithms may also have an effect on the validity of their simulation. Indeed, let us consider the 90% highest probability density (HPD) set of the target, that is, the set $A := \{x \in \mathbb{R}^2 : (x - \mu)^T \Sigma^{-1} (x - \mu)^T > 99\}$. Figure 2 (bottom) shows the percentage of X_n outside the 90% HPD computed after a 100,000 sample burn-in period. The AM algorithm tends to overestimate the ratio slightly, with more variation than the ASWAM and the RAM approaches. The estimate produced by the ASWAM algorithm has approximately the same variation as RAM, but there is a tendency to underestimate the ratio. The RAM estimates are centred around the true value.

To check how the RAM algorithm copes with higher dimensions, let us follow Roberts and Rosenthal (2009) and consider a matrix $\Sigma = MM^T$, where $M \in \mathbb{R}^{d \times d}$ is randomly generated with i.i.d. standard Gaussian elements. Such a matrix Σ is used as the pseudo-covariance of a Student distribution, so that $\pi(x) \propto (1 + x^T \Sigma^{-1} x)^{-\frac{d+1}{2}}$. Roberts and Rosenthal (2001) showed that in the case of Gaussian target and proposal distributions, one can measure the 'suboptimality' by the factor $b := d(\sum_{i=1}^d \lambda_i^{-2})(\sum_{i=1}^d \lambda_i^{-1})^{-2}$ where λ_i are the eigenvalues of the matrix $(S_n S_n^T)^{1/2} \Sigma^{-1/2}$. The factor equals one if the matrices are proportional to each other, and is larger otherwise. While the factor may not have the same interpretation in the present setting involving Student distributions, it serves as a good measure of mismatch between

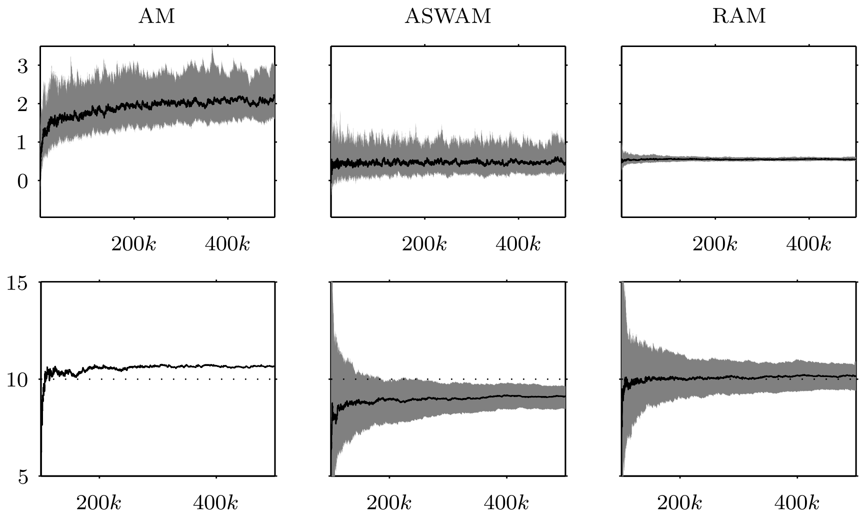

Fig. 2 Bivariate Student example: logarithm of the first diagonal component of the matrix S_n (*top*) and the proportion of X_n in the set A after 100,000 burn-in iterations (*bottom*)

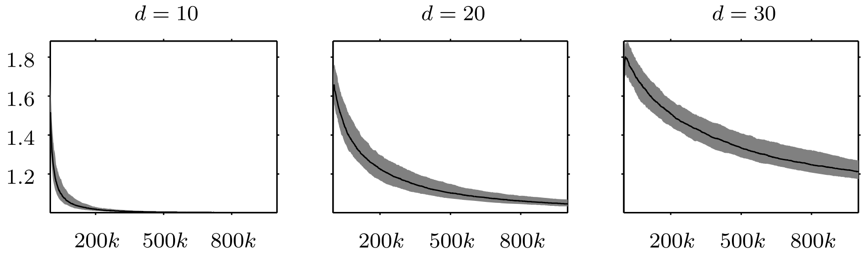

Fig. 3 Suboptimality factor b over one million iterations of the RAM algorithm with a different dimensional Student target

$S_n S_n^T$ and Σ. Figure 3 shows the factor b in increasing dimensions each based on 100 runs of the RAM algorithm. The convergence of $S_n S_n^T \to \Sigma$ is slower in higher dimensions, but the algorithm seems to find a fairly good approximation already with a moderate number of samples.

5.2 Gaussian distribution

The multivariate Gaussian target $\pi(x) = \mathcal{N}(0, \Sigma)$ serves as a baseline comparison for the algorithms, as they should converge to the same matrix factor[1] $S_n S_n^T \to \theta_* \Sigma$.

The algorithms were tested in different dimensions, for one thousand covariance matrices randomly generated as described in Sect. 5.1. The algorithms were always started in 'steady state' so that $X_1 \sim N(0, \Sigma)$. The algorithms were run half a million iterations: 100,000 burn-in and 400,000 to estimate the proportions of the samples X_n in the 10%, 25%, 50%, 75% and 90% HPD of the distribution. Table 1 shows the overall root mean square error. For dimension two, the results are comparable. Surprisingly, when the dimension increases the RAM approach provides more accurate results than the AM and the AMS algorithms.

One possible explanation is that in order to approximate the sample covariance, the covariance adaptation in AM and ASWAM should be done using the weight sequence $\eta_n = n^{-1}$ as this corresponds almost exactly to the usual sample covariance estimator. This setting was tried also; the results appear also in Table 1. It seems that using such a sequence will indeed imply better results, when starting from $s_1 \equiv I$ or $s_1 \equiv 10^{-4} \cdot I$. However, when the initial factor $s_1 = 10^4 \cdot I$ was 'too large', this approach failed. This is probably due to the fact that in this case the eigenvalues of the covariance estimate can decay only slowly, at the speed n^{-1}.

Another explanation for the unsatisfactory performance of the AM and ASWAM approaches is that in the experiments the adaptation was started right away, not after a burn-in phase run with a fixed proposal covariance as suggested in the original work (Haario et al. 2001). It is expected that the AM and the ASWAM algorithms would perform better by a suitable fixed proposal burn-in and perhaps with yet another step size sequences. In any case, this experiment demonstrates one strength of the RAM adaptation mechanism, namely that it does not require such a burn-in period.

5.3 Mixture of separate Gaussians

The last example concerns a mixture of two Gaussians distributions in $\mathbb{R}^d$ with mean vectors $m_1 := [4, 0, \ldots, 0]^T$ and $m_2 := -m_1$ and with a common diagonal covariance matrix

[1] For the AM algorithm, the constant θ_* is slightly different, but approximately equal in higher dimensions.

Table 1 Errors in Gaussian quantiles in different dimensions. The step sizes $\eta_n = n^{-1}$ were used for covariance estimation for AM† and ASWAM†

	$s_1 \equiv I$					$s_1 \equiv 10^{-4} \cdot I$					$s_1 \equiv 10^{4} \cdot I$				
d	2	4	8	16	32	2	4	8	16	32	2	4	8	16	32
AM	0.21	0.33	1.25	6.83	33.87	0.20	0.33	1.26	6.79	35.73	0.21	0.33	1.24	6.83	32.49
ASWAM	0.22	0.32	1.23	6.67	33.78	0.21	0.34	1.25	6.67	35.77	0.21	0.33	1.23	6.63	32.11
AM†	0.21	0.27	0.41	0.70	1.70	0.20	0.28	0.39	0.55	2.90	6.22	27.54	53.21	57.69	58.20
ASWAM†	0.22	0.36	0.37	0.53	1.05	0.22	0.28	0.37	0.53	3.03	0.88	1.94	3.17	5.34	8.48
RAM	0.21	0.27	0.37	0.52	1.03	0.22	0.27	0.38	0.62	2.51	0.22	0.28	0.45	0.75	1.61

Table 2 Errors of the expectations of the first and the other coordinates in the mixture example

	$X^{(1)}$					$X^{(2)}, \ldots, X^{(d)}$				
d	2	4	8	16	32	2	4	8	16	32
AM	0.04	0.05	0.08	1.69	3.87	0.08	0.11	0.15	0.19	0.27
ASWAM	0.04	0.06	0.10	1.82	3.86	0.08	0.11	0.14	0.18	0.27
RAM	0.07	0.21	0.66	1.34	1.77	0.05	0.08	0.11	0.16	0.29

$\Sigma := \mathrm{diag}(1, 100, \ldots, 100)$. In such a case, the mixing will be especially problematic with respect to the first coordinate.

Table 2 shows the root mean square error of the expectation of the first coordinate $X^{(1)}$ and the overall error for the rest $X^{(2)}, \ldots, X^{(d)}$. The errors in the first coordinate for the RAM are significantly higher than for the AM and the ASWAM for dimensions 2, 4 and 8. The estimates from all the algorithms are already quite unreliable in dimension 16. For the latter coordinates, the RAM approach seems to provide better estimates. Observe also that when comparing ASWAM with AM, the results are also worse in the first coordinate and better in the rest, like in the RAM approach. This indicates that the true optimal acceptance rate is here probably slightly less than the enforced 23.4%.

The example shows how the RAM approach finds the 'local shape' of the distribution. In fact, it is quite easy to see what happens if the means of the mixture components would be made further and further apart: there would be a stable point of the RAM algorithm that would approach the common covariance of the mixture components. Such a behaviour of the RAM approach is certainly a weakness in certain settings, as this example, but it can be also advantageous. Notice also that even such a simple multimodal setting poses a challenge for the random walk based approaches.

6 Discussion

A new robust adaptive Metropolis (RAM) algorithm was presented. The algorithm attains a given acceptance probability, and at the same time finds an estimate of the shape of the target distribution. The algorithm can cope with targets having arbitrarily heavy tails unlike the AM and ASWAM algorithms based on the covariance estimate. The RAM algorithm has some obvious limitations. It is not suitable for strongly multi-modal targets, but this is the case for any random walk based approach. For sufficiently regular targets, it seems to work well and the experiments indicate that RAM is competitive with the AM and ASWAM algorithms also in case of light-tailed targets having second moments.

There are several interesting directions of further research that were not covered in the present work. The RAM algorithm can be used also within Gibbs sampling, that is, when updating a block of coordinate variables at a time instead of the whole vector. This approach is often very useful especially when the target distribution π consists of a product of conditional densities, which is often the case with Bayesian hierarchical models. In such a setting, the computational cost of evaluating the ratio $\pi(y)/\pi(x)$ after updating one coordinate block can be significantly less than the full evaluation of $\pi(y)$. It would also be worth investigating the effect of different adaptation step sizes, perhaps even adaptive ones as suggested by Andrieu and Thoms (2008).

Regarding theoretical questions, the existence and uniqueness of the fixed points of the approach could be verified in a more general setting; the present work only covers elliptically symmetric and product type target densities, which are too restrictive in practice. The experiments indicate the overall stability of the RAM algorithm; see also Remark 5. However, proving the stability of RAM without prior bounds is directly related to the more general open question on the stability of adaptive MCMC algorithms, or even more generally to the stability of stochastic approximation. Having the stability and more general conditions on the fixed points, one could also prove the convergence of S_n in a more general setting.

Acknowledgements The author was supported by the Finnish Centre of Excellence in Analysis and Dynamics Research and by the Finnish Graduate School in Stochastics and Statistics. The author thanks Professor Christophe Andrieu and Professor Eric Moulines for useful discussions on the behaviour of the algorithm, as well as the anonymous reviewer.

Appendix A: Regularity of directional mean acceptance probability

Proposition 2 *Let π and q be probability densities on $\mathbb{R}^d$ and on $(0,\infty)$, respectively, and let $v \in \mathbb{R}^d$ be a unit vector. The function $g : (0,\infty) \to (0,\infty)$ defined by*

$$g(\theta) := \int_0^\infty \int_{\mathbb{R}^d} \min\{\pi(x), \pi(x + r\theta v)\} \mathrm{d}x q(r)\mathrm{d}r$$

is continuous, $\lim_{\theta\to\infty} g(\theta) = 0$ *and* $\lim_{\theta\to 0+} g(\theta) = 1$.

Proof Suppose first that π is a continuous probability density on $\mathbb{R}^d$. Then, write

$$g(\theta) = \int_0^\infty \int_A \min\left\{1, \frac{\pi(x + r\theta v)}{\pi(x)}\right\} \pi(x)\mathrm{d}x q(r)\mathrm{d}r$$

where $A := \{x \in \mathbb{R}^d : \pi(x) > 0\}$ stands for the support of π. Let $(\theta_n)_{n\ge 1} \subset (0,\infty)$ be any sequence and define $f_\theta(x, r) := \min\{1, \frac{\pi(x+r\theta v)}{\pi(x)}\}$. Clearly, whenever θ_n converges to some θ, then $f_{\theta_n}(x, r) \to f_\theta(x, r)$ pointwise on $A \times (0,\infty)$ by the continuity of π. Since $|f_n(x, r)| \le 1$, the dominated convergence theorem yields that $|g(\theta_n) - g(\theta)| \to 0$, establishing the continuity. For any sequence $\theta_n \to 0+$ one clearly has $f_{\theta_n}(x, r) \to 1$, and for any sequence $\theta_n \to \infty$ one obtains $f_{\theta_n}(x, r) \to 0$, establishing the claim.

Let us then proceed to the general case. Let $\epsilon > 0$ be arbitrary. We shall show that there exists a continuous probability density $\tilde{\pi}$ on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} |\tilde{\pi}(x) - \pi(x)|\mathrm{d}x < \epsilon.$$

Having such $\tilde{\pi}$, one can bound the difference

$$\left| g(\theta) - \int_0^\infty \int_A \min\left\{1, \frac{\hat{\pi}(x + r\theta v)}{\hat{\pi}(x)}\right\} \hat{\pi}(x)\mathrm{d}x q(r)\mathrm{d}r \right|$$
$$\le \int_{\mathbb{R}^d} |\pi(x) - \tilde{\pi}(x)|\mathrm{d}x < \epsilon.$$

It remains to verify that such a continuous probability density $\tilde{\pi}$ exists. Approximate π first by smooth non-negative functions π_n such that $\int_{\mathbb{R}^d} |\pi(x) - \pi_n(x)|\mathrm{d}x \to 0$, and then normalise them to probability densities $\tilde{\pi}_n(x) := c_n\pi_n(x)$. Clearly, the constants $c_n := (\int_{\mathbb{R}^d} \pi_n(z)\mathrm{d}z)^{-1} \to 1$, and so $\int_{\mathbb{R}^d} |\pi(x) - \tilde{\pi}_n(x)|\mathrm{d}x \le \int_{\mathbb{R}^d} |\pi(x) - \pi_n(x)|\mathrm{d}x + |1 - c_n| \to 0$. □

Appendix B: Proofs of convergence

Proof of Theorem 6 Let $0 < a \le b < \infty$ be arbitrary constants and denote by $\mathbb{S}_{a,b} \subset \mathbb{R}^{d\times d}$ the set of all lower triangular matrices with positive diagonal, such that the eigenvalues of ss^T are within $[a, b]$. Let P_s stand for the random walk Metropolis kernel with a proposal density $q_s(z) := \det(s)^{-1}q(s^{-1}z)$, that is, for any $x \in \mathbb{R}^d$ and any Borel set $A \subset \mathbb{R}^d$

$$P_s(x, A) := \mathbb{1}_A(x)\left(1 - \int_{\mathbb{R}^d} \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} q_s(y - x)\mathrm{d}y\right)$$
$$+ \int_A \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} q_s(y - x)\mathrm{d}y.$$

We shall use the notation $P_s f(x) := \int_{\mathbb{R}^d} f(y)P_s(x, \mathrm{d}y)$ to denote the integration of a function with respect to the kernel P_s.

Let us check that the following assumptions are satisfied by the RAM algorithm.

(A1) For all possible $s \in \mathbb{S}_{a,b}$, the kernels P_s have a unique invariant probability distribution π for which $\int_{\mathbb{R}^d} P(x, A)\pi(\mathrm{d}x) = \pi(A)$ for any Borel set $A \subset \mathbb{R}^d$.

(A2) There exist a Borel set $C \subset \mathbb{R}^d$, a function $V : \mathbb{R}^d \to [1,\infty)$, constants $\delta, \lambda \in (0, 1)$ and $b < \infty$, and a probability measure ν concentrated on C such that

$$P_s V(x) \le \lambda V(x) + \mathbb{1}_C(x)b \quad \text{and}$$
$$P_s(x, A) \ge \mathbb{1}_C(x)\delta\nu(A)$$

for all possible $x \in \mathbb{R}^d$, $s \in \mathbb{S}_{a,b}$ and all Borel sets $A \subset \mathbb{R}^d$.

(A3) For all $n \ge 1$ and any $r \in (0, 1]$, there is a constant $c' = c'(r) \ge 1$ such that for all $s, s' \in \mathbb{S}_{a,b}$,

$$\sup_{x\in\mathbb{R}^d} \frac{|P_s f(x) - P_{s'} f(x)|}{V^r(x)} \le c'|s - s'| \sup_{x\in\mathbb{R}^d} \frac{|f(x)|}{V^r(x)}.$$

(A4) There is a constant $c < \infty$ such that for all $n \ge 1$, $s \in \mathbb{S}_{a,b}$, $x \in \mathbb{R}^d$ and $u \in \mathbb{R}^d$ the bound $|H(s, x, u)| \le c$ holds.

The uniqueness of the invariant distribution (A1) follows by observing that the kernels P_s are irreducible, aperiodic and reversible with respect to π (see, e.g. Nummelin 2002). The simultaneous drift and minorisation condition (A2) was established by Andrieu and Moulines (2006). The continuity condition (A3) was established by Andrieu and Moulines (2006) for Gaussian proposal distributions and was extended to cover the Student proposal in Vihola (2011b). The bound (A4) is easy to verify.

Assumption 4 ensures that for any $\epsilon > 0$ there exist constants $0 < a_\epsilon \le b_\epsilon < \infty$ such that all the eigenvalues of

$S_n S_n^T$ stay within the interval $[a_\epsilon, b_\epsilon]$ at least with probability $\mathbb{P}(\Omega_0) - \epsilon$. This is enough to ensure that the strong law of large numbers holds by Andrieu and Moulines (2006, Proposition 6). For details, see also Saksman and Vihola (2010, Theorem 2) and Vihola (2011b, Theorem 20). □

Proof of Theorem 7 The proof follows by Andrieu et al. (2004, Theorem 5) by using a similar technique as in the proof of Theorem 6. Consider the Lyapunov function $w_\pi(R)$ defined in Theorem 5. It is straightforward to verify items 1–4 of Andrieu et al. (2004, Condition 1) when we take Θ to be the space of symmetric positive definite matrices and consider $S_n S_n^T \in \Theta$. The compact sets are of the form $K = \{ss^T : s \in \mathbb{S}_{a_\epsilon, b_\epsilon}\}$ with a_ϵ, b_ϵ as in the proof of Theorem 6. Item 5 follows by invoking Saksman and Vihola (2010, Proposition 6) with $f_\theta((x,u)) = H(\theta, x, u)$. □

References

Andrieu, C., Moulines, É.: On the ergodicity properties of some adaptive MCMC algorithms. Ann. Appl. Probab. **16**(3), 1462–1505 (2006)

Andrieu, C., Robert, C.P.: Controlled MCMC for optimal sampling. Tech. Rep. Ceremade 0125, Université Paris Dauphine (2001)

Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**(4), 343–373 (2008)

Andrieu, C., Moulines, É., Volkov, S.: Convergence of stochastic approximation for Lyapunov stable dynamics: a proof from first principles. Technical report (2004)

Andrieu, C., Moulines, É., Priouret, P.: Stability of stochastic approximation under verifiable conditions. SIAM J. Control Optim. **44**(1), 283–312 (2005)

Atchadé, Y., Fort, G.: Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. Bernoulli **16**(1), 116–154 (2010)

Atchadé, Y.F., Rosenthal, J.S.: On adaptive Markov chain Monte Carlo algorithms. Bernoulli **11**(5), 815–828 (2005)

Benveniste, A., Métivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Applications of Mathematics, vol. 22. Springer, Berlin (1990)

Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, Cambridge (2008)

Dongarra, J.J., Bunch, J.R., Moler, C.B., Stewart, G.W.: LINPACK Users' Guide. Society for Industrial and Applied Mathematics (1979)

Gelman, A., Roberts, G.O., Gilks, W.R.: Efficient Metropolis jumping rules. In: Bayesian Statistics 5, pp. 599–607. Oxford University Press, Oxford (1996)

Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton (1998)

Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**(2), 223–242 (2001)

Hastie, D.: Toward automatic reversible jump Markov chain Monte Carlo. PhD thesis, University of Bristol (2005)

Huber, P.J.: Robust Statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1981)

Jarner, S.F., Hansen, E.: Geometric ergodicity of Metropolis algorithms. Stoch. Process. Appl. **85**, 341–361 (2000)

Jarner, S.F., Roberts, G.O.: Convergence of heavy-tailed Monte Carlo Markov chain algorithms. Scand. J. Stat. **34**(4), 781–815 (2007)

Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, 2nd edn. Applications of Mathematics: Stochastic Modelling and Applied Probability, vol. 35. Springer, Berlin (2003)

Nummelin, E.: MC's for MCMC'ists. Int. Stat. Rev. **70**(2), 215–240 (2002)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (1999)

Roberts, G.O., Rosenthal, J.S.: Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**(4), 351–367 (2001)

Roberts, G.O., Rosenthal, J.S.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004)

Roberts, G.O., Rosenthal, J.S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. **44**(2), 458–475 (2007)

Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. J. Comput. Graph. Stat. **18**(2), 349–367 (2009)

Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**(1), 110–120 (1997)

Saksman, E., Vihola, M.: On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. Ann. Appl. Probab. **20**(6), 2178–2203 (2010)

Vihola, M.: Grapham: Graphical models with adaptive random walk Metropolis algorithms. Comput. Stat. Data Anal. **54**(1), 49–54 (2010)

Vihola, M.: Can the adaptive Metropolis algorithm collapse without the covariance lower bound? Electron. J. Probab. **16**, 45–75 (2011a)

Vihola, M.: On the stability and ergodicity of adaptive scaling Metropolis algorithms. Preprint (2011b). arXiv:0903.4061v3