On the stability of adaptive random walk Metropolis algorithms

Matti Vihola
matti.vihola@iki.fi

University of Jyväskylä, Department of mathematics and statistics

joint work with Eero Saksman (University of Helsinki)

MCQMC, Warsaw, August 20th 2010

The problem

- Compute numerically the integral

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

- Assumptions:
  - π probability density
  - $\pi(|f|) < \infty$
  - possible to evaluate $\pi(x)$ at each $x \in \mathbb{R}^d$ (up to a normalising constant)
  - possible to evaluate $f(x)$ at any $x \in \mathbb{R}^d$

Markov chain Monte Carlo

- Construct a Markov chain $(X_k)_{k\geq 1}$ having π as the unique invariant probability distribution (the 'target distribution').
- With some mild assumptions, the strong law of large numbers (SLLN) can be shown to hold:

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{n\to\infty} \pi(f) \qquad \text{almost surely.}$$

- In some cases, one may establish a central limit theorem (CLT) too:

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}[f(X_k) - \pi(f)] \xrightarrow{n\to\infty} N(0, \sigma_f^2) \qquad \text{in distribution.}$$

The Gaussian random walk Metropolis algorithm I

- Define $X_1 \equiv x_1 \in \mathbb{R}^d$ such that $\pi(x_1) > 0$.
- Let $c \in \mathbb{R}^{d \times d}$ be a symmetric and positive definite matrix.

For $n = 2, 3, \ldots$, set recursively

$$\begin{aligned} Y_n &= X_{n-1} + c^{1/2} W_n, \qquad \text{where } W_n \sim N(0, I), \text{ and} \\ X_n &= \begin{cases} Y_n, & \text{with probability } \alpha(X_{n-1}, Y_n) \text{ and} \\ X_{n-1}, & \text{otherwise.} \end{cases} \end{aligned}$$

The probability $\alpha(x, y)$ of accepting a proposal y at x is defined as

$$\alpha(x, y) := \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\}.$$

The Gaussian random walk Metropolis algorithm II

The variables $(X_k)_{k\geq 1}$ form a Markov chain, with the transition probability

$$\mathbb{P}\left(X_n \in A \mid X_1, \ldots, X_{n-1}\right) = P_c(X_{n-1}, A).$$

Example I

- In the Gaussian random walk Metropolis algorithm, the only user-defined parameter is the covariance c of the proposal increment.
- Theoretically, any symmetric and positive definite c is valid and yields a SLLN (and a CLT with some additional conditions on π and f).
- Some choices of c are better than others...

Example II

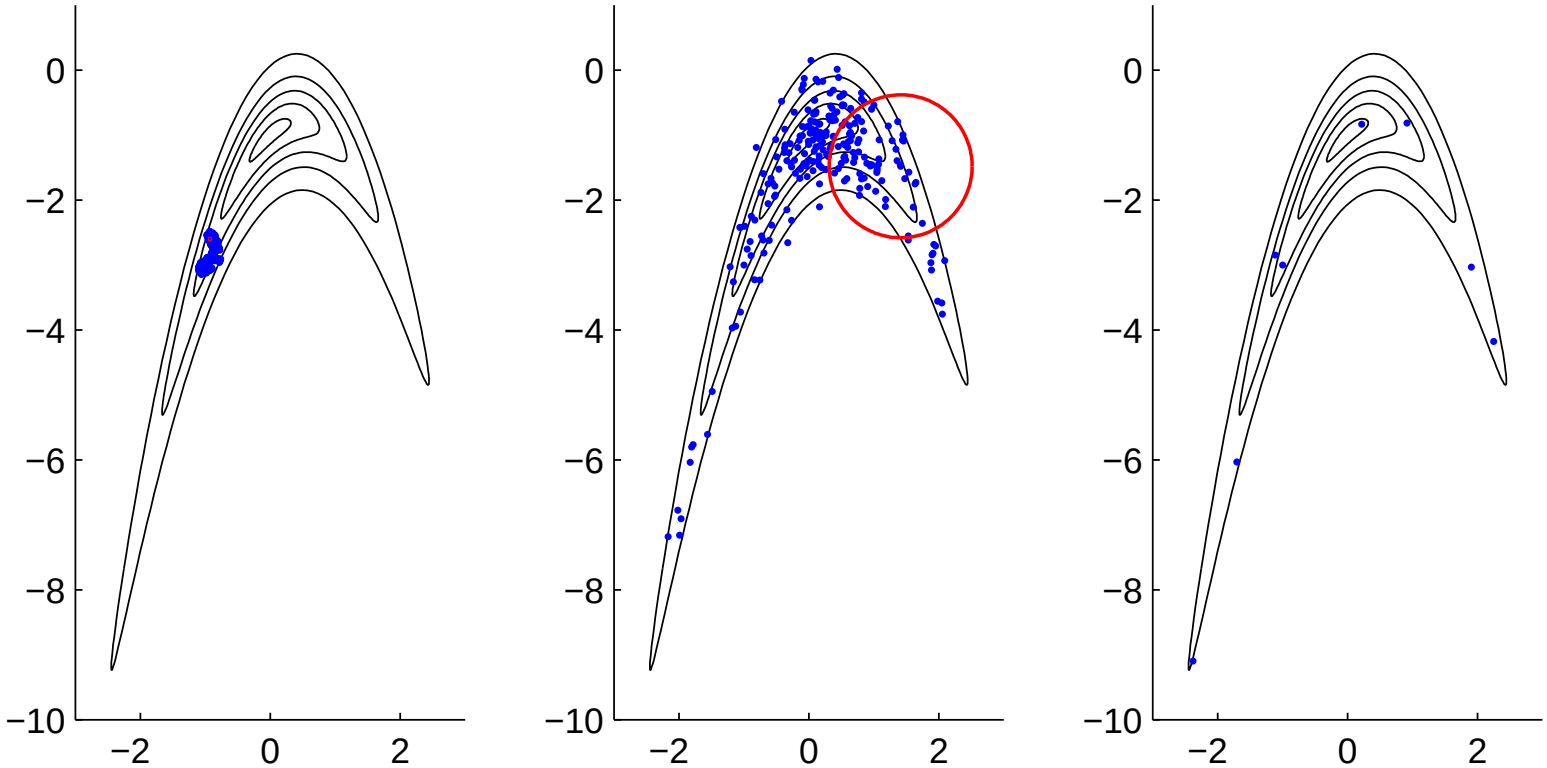


Figure: First 1000 samples of the Gaussian random walk Metropolis chain in $\mathbb{R}^2$. The black solid lines show the contours of the 'banana-shaped' π.

Adaptive MCMC I

- Adaptive MCMC methods, in general, are based on a collection of Markov kernels $\{P_c\}$, each having π as the unique invariant distribution.
- The aim is to find a good kernel (i.e. the value of the parameter c) automatically.
- In practice, one sets

$$\mathbb{P}\left(X_n \in A \mid X_1, \ldots, X_{n-1}, C_1, \ldots, C_{n-1}\right) = P_{C_{n-1}}(X_{n-1}, A)$$

where C_k are random variables that typically depend on the history $X_1, \ldots, X_k$.

Adaptive MCMC II

- The chain $(X_n, C_n)_{n\geq 1}$ is often inhomogeneous Markov, but $(X_n)_{n\geq 1}$ alone is *non-Markovian*.
- Without further assumptions, one can easily construct examples where the correct ergodic properties are destroyed and $\frac{1}{n}\sum_{k=1}^{n} f(X_k) \not\to \pi(f)$.
- Typical conditions to ensure correct ergodicity:
  - the effect of the adaptation 'diminishes,' so that "$\|P_{C_n} - P_{C_{n-1}}\| \to 0$."
  - the ergodic properties of P_{C_n} are suitably good. This is often strongly related to the *stability* of the process $(C_n)_{n\geq 1}$.

Adaptive Metropolis algorithm I

The AM algorithm [Haario, Saksman, and Tamminen, 2001]:

- Define $X_1 \equiv x_1 \in \mathbb{R}^d$ such that $\pi(x_1) > 0$.
- Choose parameters $\theta > 0$ and $\epsilon \geq 0$.

For $n = 2, 3, \ldots$

$$
\begin{aligned}
Y_n &= X_{n-1} + C_{n-1}^{1/2} W_n, \qquad \text{where } W_n \sim N(0, I), \text{ and} \\
X_n &= \begin{cases} Y_n, & \text{with probability } \alpha(X_{n-1}, Y_n) \text{ and} \\ X_{n-1}, & \text{otherwise.} \end{cases}
\end{aligned}
$$

where $C_{n-1} := \theta^2 \operatorname{Cov}(X_1, \ldots, X_{n-1}) + \epsilon I$, and $I \in \mathbb{R}^{d \times d}$ stands for the identity matrix.

Adaptive Metropolis algorithm II

$\mathrm{Cov}(\cdots)$ is *some* consistent covariance estimator (not necessarily the standard sample covariance!).

- In what follows, consider the definition $\mathrm{Cov}(X_1, \ldots, X_n) := S_n$, where $S_1 \equiv s_1 \in \mathbb{R}^{d \times d}$ is symmetric and positive definite, and

$$S_n = \frac{n-1}{n} S_{n-1} + \frac{1}{n}(X_n - \overline{X}_{n-1})(X_n - \overline{X}_{n-1})^T$$

where $\overline{X}_{n-1}$ stands for the average of $X_1, \ldots, X_{n-1}$.

Example run of AM

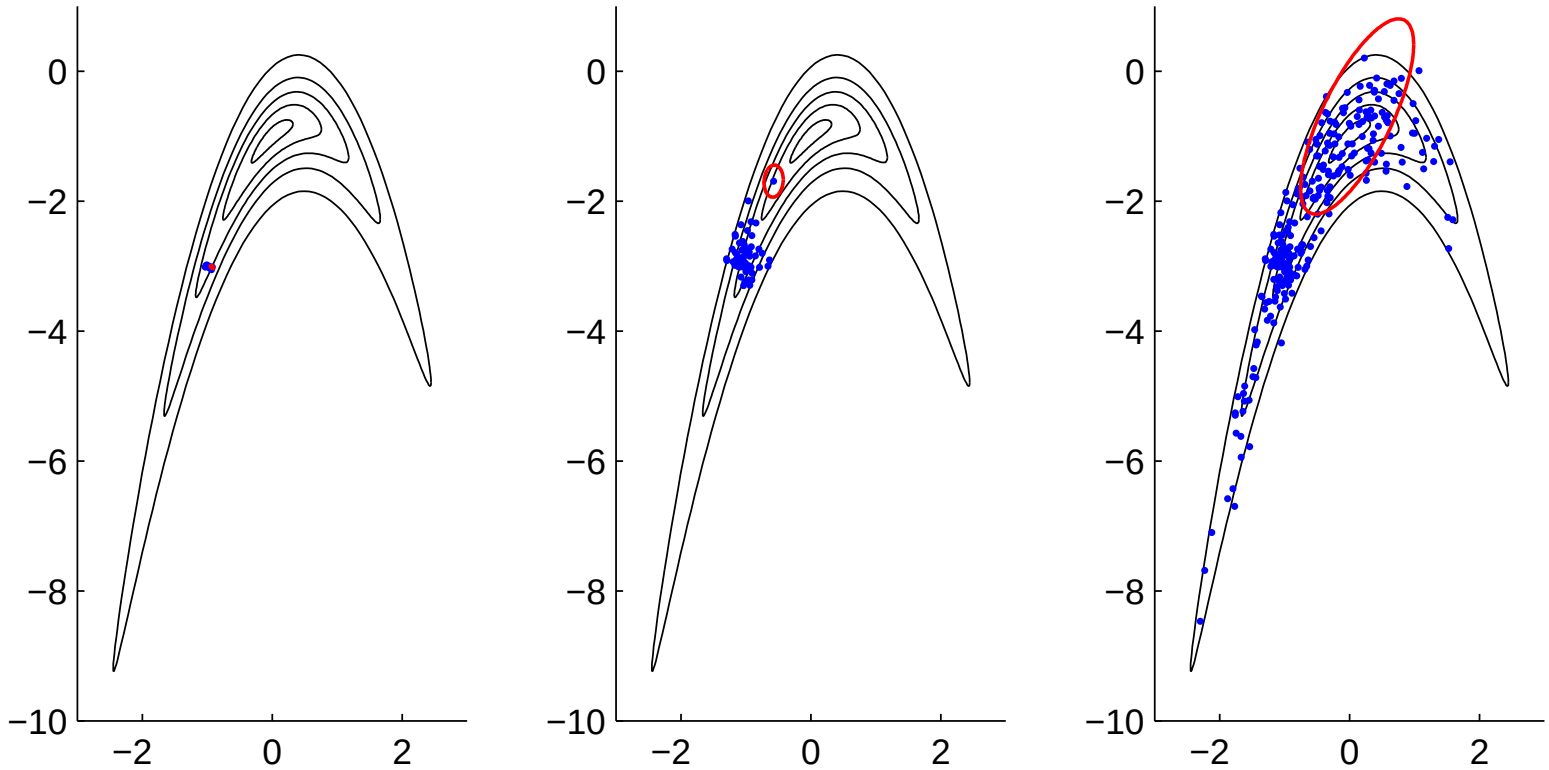


Figure: The first 10, 100 and 1000 samples of the AM algorithm started with $s_1 = (0.01)^2 I$, $\theta = 2.38/\sqrt{2}$ and $\epsilon = 0$.

Adaptive scaling Metropolis algorithm I

This algorithm, essentially proposed by [Gilks, Roberts, and Sahu, 1998, Andrieu and Robert, 2001], adjusts the size of the proposal jumps, and tries to attain a given mean acceptance rate.

- Define $X_1 \equiv x_1 \in \mathbb{R}^d$ such that $\pi(x_1) > 0$.
- Let $\Theta_1 \equiv \theta_1 > 0$.
- Define a sequence of positive adaptation step sizes $(\eta_n)_{n \geq 2}$ decaying to zero.
- Define the desired mean acceptance rate $\alpha^* \in (0, 1)$.

Adaptive scaling Metropolis algorithm II

For $n = 2, 3, \ldots$, iterate

$$
\begin{aligned}
Y_n &= X_{n-1} + \Theta_{n-1} W_n, \qquad \text{where } W_n \sim N(0, I), \text{ and} \\
X_n &= \begin{cases} Y_n, & \text{with probability } \alpha(X_{n-1}, Y_n) \text{ and} \\ X_{n-1}, & \text{otherwise.} \end{cases} \\
\log \Theta_n &= \log \Theta_{n-1} + \eta_n \big[\alpha(X_{n-1}, Y_n) - \alpha^* \big].
\end{aligned}
$$

Adaptive scaling Metropolis algorithm III

- The ASM algorithm relates to *optimal scaling* of the random walk Metropolis proposal distribution [e.g. Gelman, Roberts, and Gilks, 1996, Roberts, Gelman, and Gilks, 1997].
- The 'rule of thumb' is to let $\alpha^* = 0.44$ in dimension one, and $\alpha^* = 0.234$ if $d \geq 2$.
- These choices are not always optimal, but should often work well in practice.

Example run of ASM

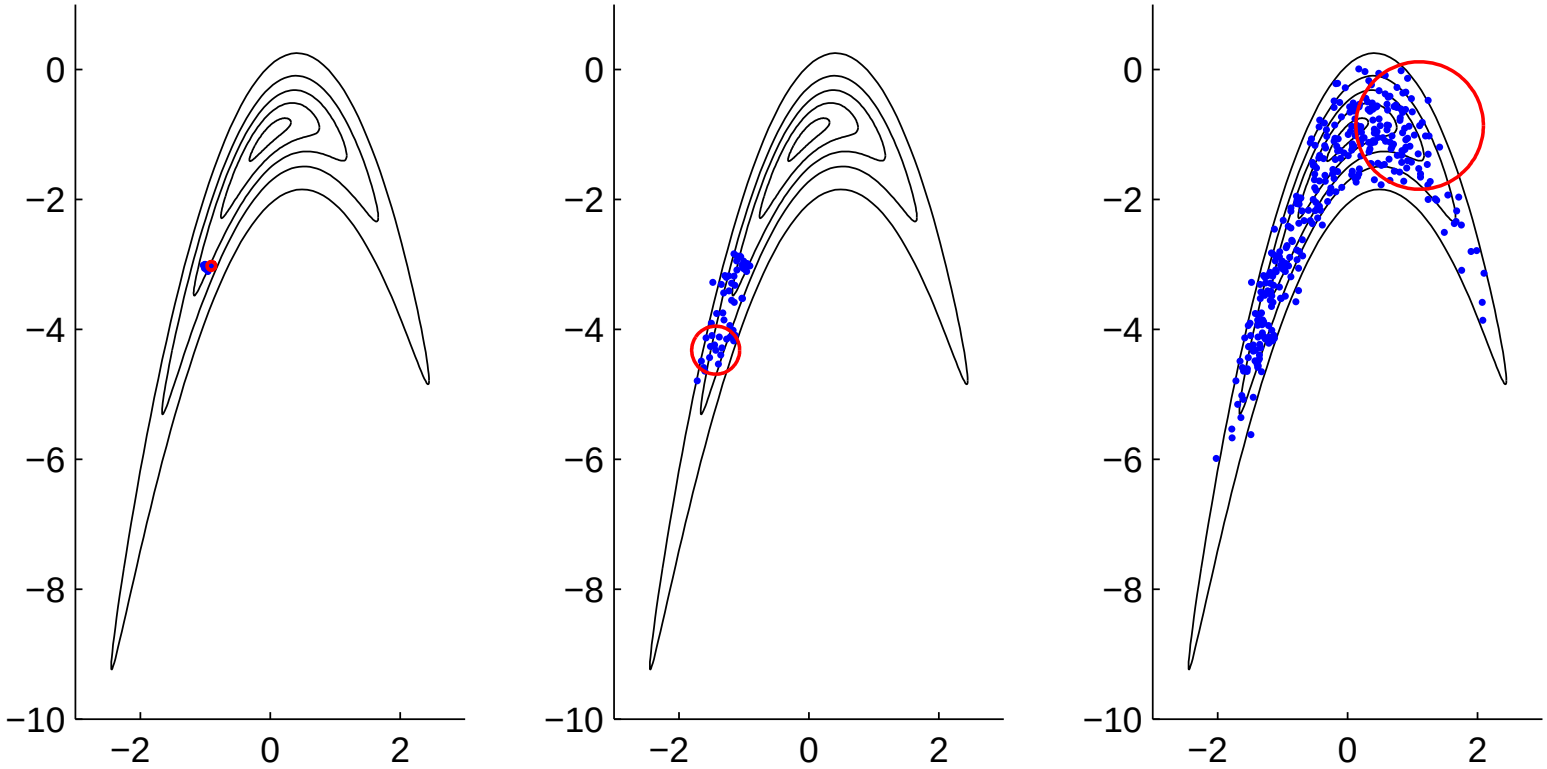


Figure: The first 10, 100 and 1000 samples of the ASM algorithm started with $\theta_1 = 0.01$ and using $\eta_n = n^{-3/4}$.

Results in the literature I

There are various results in the literature on the ergodicity of the AM and the ASM algorithms.

- The original work on AM [Haario, Saksman, and Tamminen, 2001].
- Atchadé and Rosenthal [2005] analyse the ASM algorithm following the original mixingale approach.
- The stochastic approximation formulation [Andrieu and Robert, 2001] and the ergodicity results [Andrieu and Moulines, 2006].
- The coupling approach of Roberts and Rosenthal [2007].
- The recent paper by Atchadé and Fort [2010] allows also heavy-tailed target distributions.

Results in the literature II

- All the results require one to truncate the eigenvalues of the covariance $0 < a \leq \lambda(C_n) \leq b < \infty$.
- A notable exception is [Andrieu and Moulines, 2006], who consider a reprojection approach with increasing sequence of reprojection sets $[a_1, b_2]$, $[a_2, b_2] \ldots$
- Strong belief (based on empirical evidence): truncation (or reprojection) is unnecessary!

Assumptions I

Definition (Strongly super-exponential target)

The target density π is continuously differentiable and has regular tails that decay super-exponentially,

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

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

with some $\rho > 1$.

(The "super-exponential case", $\rho = 1$, is due to Jarner and Hansen [2000] who ensure the geometric ergodicity of a nonadaptive random-walk Metropolis process.)

Assumptions II

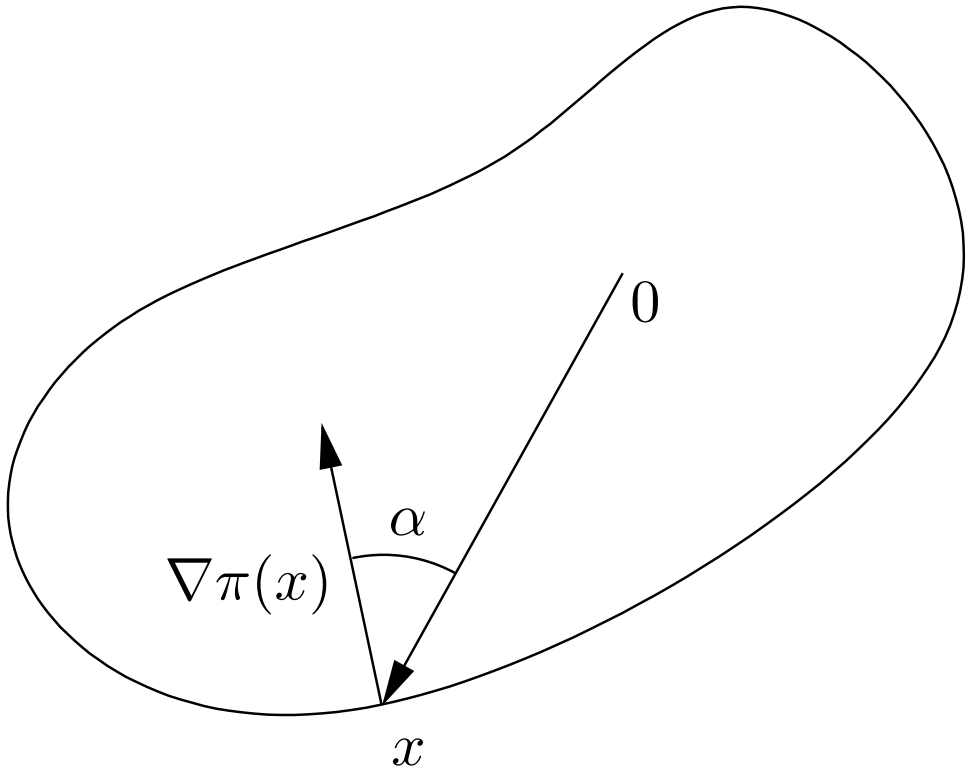


Figure: The condition $\limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0$ implies that there is an $\varepsilon > 0$ such that for any sufficiently large $|x|$, the angle $\alpha < \pi/2 - \varepsilon$.

Assumptions III

Definition

The function $f : \mathbb{R}^d \to \mathbb{R}$ has at most exponential tails if there is a constant $M < \infty$ such that

$$|f(x)| \leq M \max\{1, e^{|x|}\}.$$

AM when π has unbounded support

- Saksman and Vihola [to appear]: First ergodicity results for the original AM algorithm, without upper bounding the eigenvalues $\lambda(C_n)$.
  - Our approach is based on the technique by Andrieu and Moulines [2006].
- Use the proposal covariance $C_n = \theta^2 \operatorname{Cov}(X_1, \ldots, X_n) + \epsilon I$, with $\epsilon > 0$.
- SLLN and CLT for strongly super-exponential π and functions with at most exponential tails.

AM without the covariance lower bound I

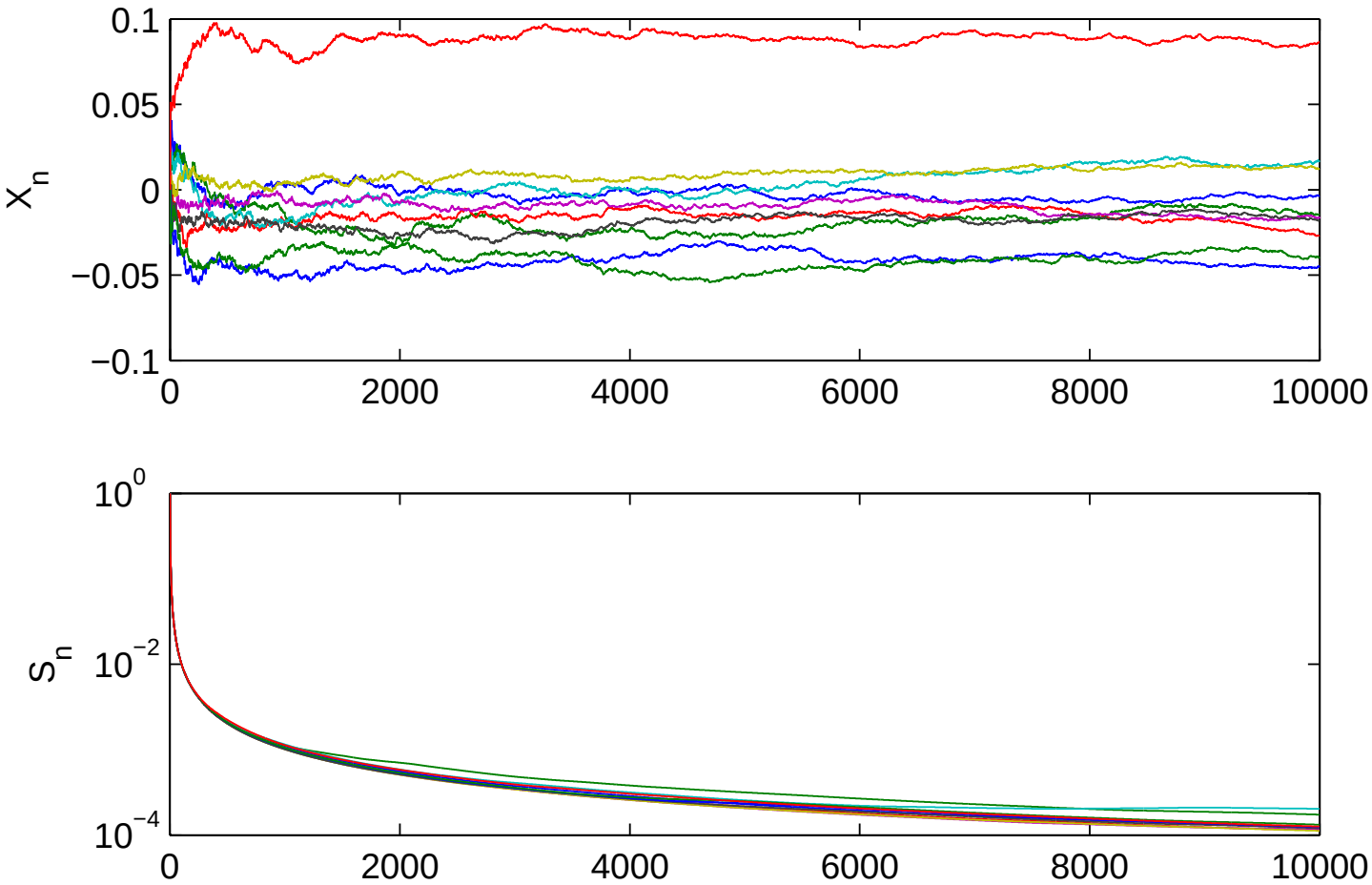
Vihola [2009b] contains partial results on the case $C_n = \theta^2 S_n$, that is, without lower bounding the eigenvalues $\lambda(C_n)$.

- Analysed first an 'adaptive random walk': AM run with 'flat target $\pi \equiv 1$',

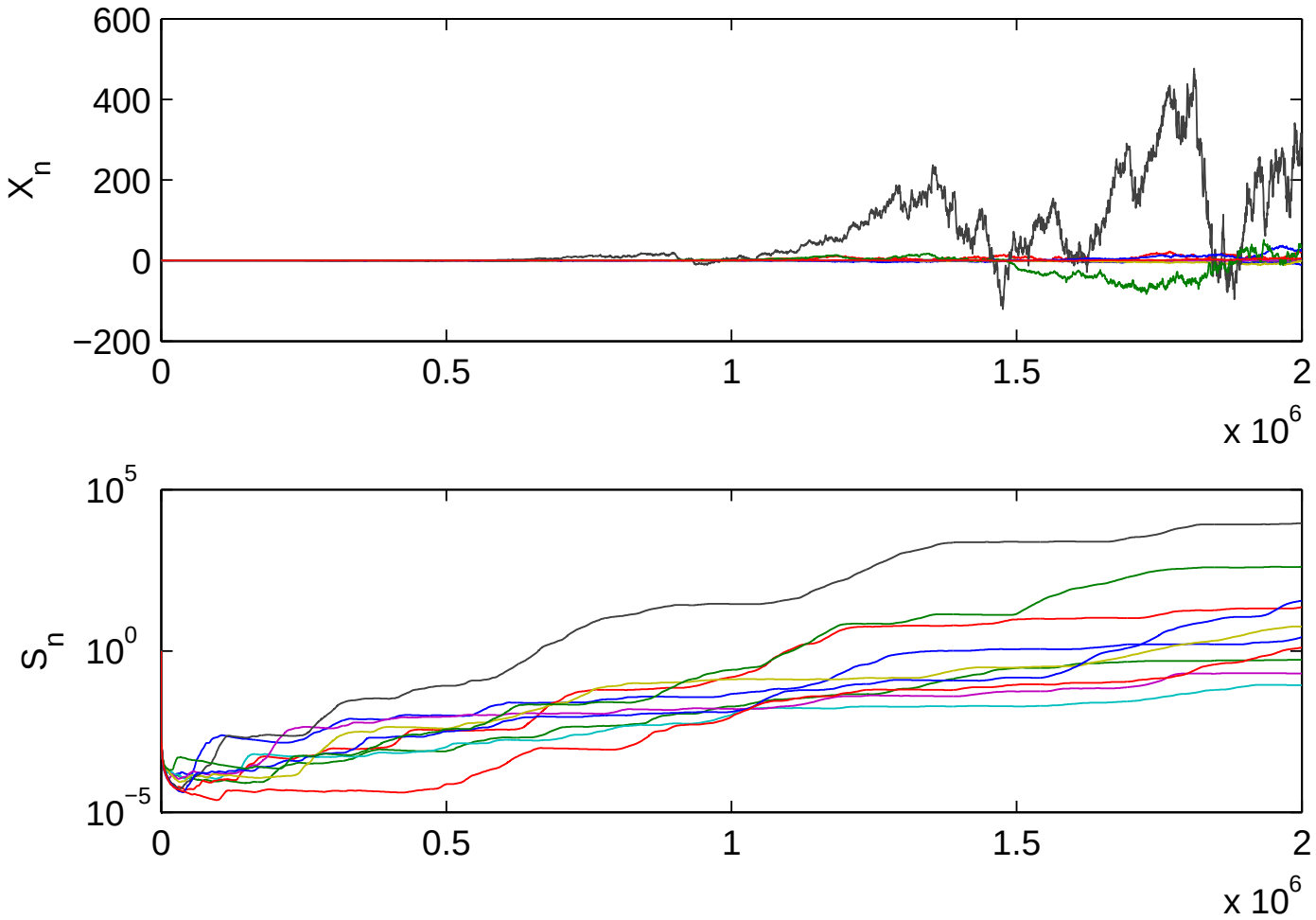
$$\begin{aligned} X_{n+1} &= X_n + \theta S_n^{1/2} W_{n+1} \\ S_{n+1} &= \frac{n}{n+1} S_n + \frac{1}{n+1} (X_{n+1} - \overline{X}_n)^2. \end{aligned}$$

- Ten sample paths of this process (univariate) started at $x_0 = 0$, $s_0 = 1$ and with the constant $\theta = 0.01$:

AM without the covariance lower bound II



AM without the covariance lower bound III



AM without the covariance lower bound IV

- It is shown that $S_n \to \infty$ almost surely.
- The speed of growth is $\mathbb{E}[S_n] \sim e^{2\theta\sqrt{n}}$.
- Using the same techniques, one can show the stability (and ergodicity) of AM run with a univariate Laplace target π.
- These results have little direct practical value, but they indicate that the AM covariance parameter S_n does not tend to collapse.

AM with a fixed proposal component I

- Instead of lower bounding the eigenvalues $\lambda(C_n)$, employ a fixed proposal component with a probability $\beta \in (0,1)$ [Roberts and Rosenthal, 2009].
- This corresponds to an algorithm where the proposals Y_n are generated by

$$Y_n = X_{n-1} + \begin{cases} c_0^{1/2} W_n, & \text{with probability } \beta, \\ C_{n-1}^{1/2} W_n, & \text{otherwise.} \end{cases}$$

  with some fixed symm.pos.def. c_0.
- In other words, one employs a mixture of 'adaptive' and 'nonadaptive' Markov kernels:

$$\mathbb{P}\left(X_n \in A \mid X_1, \ldots, X_{n-1}\right) = (1-\beta) P_{C_{n-1}}(A) + \beta P_{c_0}(A)$$

AM with a fixed proposal component II

Vihola [2009b] shows that, for example with a bounded and compactly supported π or with a super-exponential π, the eigenvalues $\lambda(C_n)$ are bounded away from zero.

- Having a strongly super-exponential target, SLLN and CLT hold for functions with at most exponential tails.
- Explicit upper and lower bounds for C_n unnecessary.
- Mixture proposal may be better in practice than the lower bound ϵI.

Also Bai, Roberts, and Rosenthal [2008] analyse this algorithm.

- A completely different approach, with different assumptions.
- (Weak) ergodicity shown for (at least) exponential π if the fixed covariance is "large enough".

Ergodicity of unconstrained ASM algorithm

Vihola [2009a] shows two results for the unmodified adaptation, without any (upper or lower) bounds for Θ_n. Assume:

- the desired acceptance rate $\alpha^* \in (0, 1/2)$.
- the adaptation weights satisfy $\sum \eta_n^2 < \infty$ (e.g. $\eta_n = n^{-\gamma}$ with $\gamma \in (1/2, 1]$).

Two cases:

1. π is bounded, bounded away from zero on the support, and the support $\mathbb{X} = \{x : \pi(x) > 0\}$ is compact and has a smooth boundary.
   Then, SLLN holds for bounded functions.
2. Suppose a strongly super-exponential target having tails with uniformly smooth contours.
   Then, SLLN holds for functions with at most exponential tails.

Ergodicity of ASM within AM

The AM and ASM algorithms can be naturally used simultaneously [Atchadé and Fort, 2010, Andrieu and Thoms, 2008].

- Define proposal covariance $C_{n-1} := \Theta_{n-1} \operatorname{Cov}(X_1, \ldots, X_n)$.
- Coerced acceptance rate *and* target covariance structure in the adaptation.
- The technique in [Vihola, 2009a] applies also in this case, provided that the eigenvalues of the covariance part are bounded within $0 < a \leq b < \infty$.

Final remarks I

- Current results show that some adaptive MCMC algorithms are *intrinsically stable*, requiring no additional stabilisation structures.
  - Easier for practitioners; less parameters to 'tune.'
  - Showing that the methods are fairly 'safe' to apply.
- The results are related to the more general question of the stability of the Robbins-Monro stochastic approximation with Markovian noise.

Final remarks II

- It is not necessary to use Gaussian proposal distributions. All the above results apply to elliptical proposals satisfying a particular tail decay condition. For example, the results apply with multivariate Student distributions having the form

$$q_c(z) \propto (1 + \|c^{-1/2} z\|^2)^{-\frac{d+p}{2}}$$

  where $p > 0$.
- Current results apply only for targets with rapidly decaying tails. It is important to establish similar results with heavy-tailed targets.
- Overall, there is a need for more general yet practically verifiable conditions to check the validity of the methods.

Final remarks III

- There is also a software available for testing several adaptive random-walk MCMC algorithms [Vihola, 2010]: `http://iki.fi/mvihola/grapham/`

References I

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

C. Andrieu and C. P. Robert. Controlled MCMC for optimal sampling. Technical Report Ceremade 0125, Université Paris Dauphine, 2001.

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

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

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

References II

Y. Bai, G. O. Roberts, and J. S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. Preprint, July 2008. URL `http://probability.ca/jeff/research.html`.

A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, pages 599–607. Oxford University Press, 1996.

W. R. Gilks, G. O. Roberts, and S. K. Sahu. Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.*, 93(443):1045–1054, 1998.

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

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

References III

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

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

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

E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Ann. Appl. Probab.*, to appear. Preprint, arXiv:0806.2933v4.

M. Vihola. On the stability and ergodicity of an adaptive scaling Metropolis algorithm. Preprint, arXiv:0903.4061v2, Mar. 2009a.

References IV

M. Vihola. Can the adaptive Metropolis algorithm collapse without the covariance lower bound? Preprint, arXiv:0911.0522v1, Nov. 2009b.

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