On the stability and convergence of adaptive MCMC

Matti Vihola
matti.vihola@iki.fi

University of Jyväskylä, Department of mathematics and statistics

joint work with Eero Saksman (University of Helsinki)

AdapskIII, Park City, Utah, January 4th 2011

The Gaussian random walk Metropolis algorithm I

- The starting point $X_1 \equiv x_1 \in \mathbb{R}^d$ is in the support, $\pi(x_1) > 0$.
- The parameter $\Sigma \in \mathbb{R}^{d \times d}$ is some symmetric and positive definite matrix.

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

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

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

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

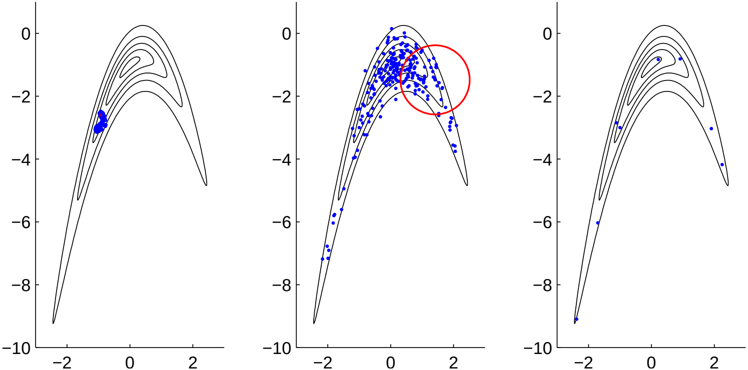
The Gaussian random walk Metropolis algorithm II

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

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

- Here, the proposal distribution is Gaussian and determined by the covariance parameter Σ.
- How to choose Σ in practice?

The effect of proposal covariance



$\Sigma = (0.01)^2 \times I$ $\Sigma = (1.1)^2 \times I$ $\Sigma = (10)^2 \times I$

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.

Choosing proposal covariance

- There are some rules of thumb on choosing Σ:
  - If the target distribution π has finite second moments, then Σ should be proportional to the covariance of π.
  - The acceptance rate should be around 44% $(d = 1)$ and 23.4% $(d \geq 2)$.
  - (These rules are based on theoretical findings [e.g. Gelman, Roberts, and Gilks, 1996, Roberts, Gelman, and Gilks, 1997, Roberts and Rosenthal, 2001].)
- In practice:
  - The 'classical' solution is to perform one or more 'pilot runs' and determine Σ based on these pilot runs.
  - The 'modern' approach is to use adaptation...

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 $t > 0$ and $\epsilon \geq 0$.

For $n = 2, 3, \ldots$

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

where $\Sigma_{n-1} := t^2 \operatorname{Cov}(X_1, \ldots, X_{n-1}) + \epsilon I$, 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

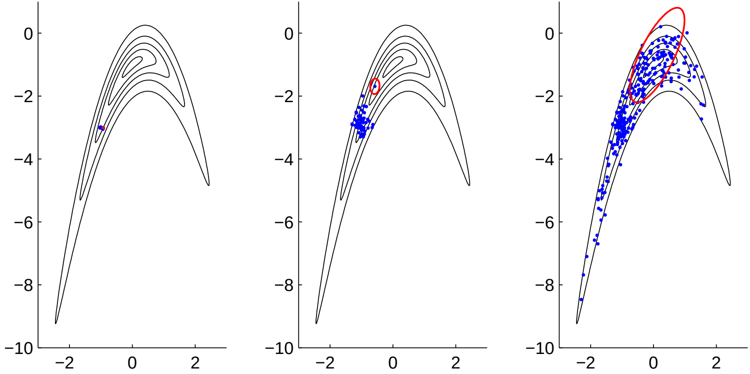


Figure: The first 10, 100 and 1000 samples of the AM algorithm started with $s_1 = (0.01)^2 I$, $t = 2.38/\sqrt{2}$ and $\epsilon = 0$.

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 $\alpha^* \in (0,1)$. (Usually $\alpha^* = 0.44$ or $\alpha^* = 0.234\ldots$)

Adaptive scaling Metropolis algorithm II

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

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

Example run of ASM

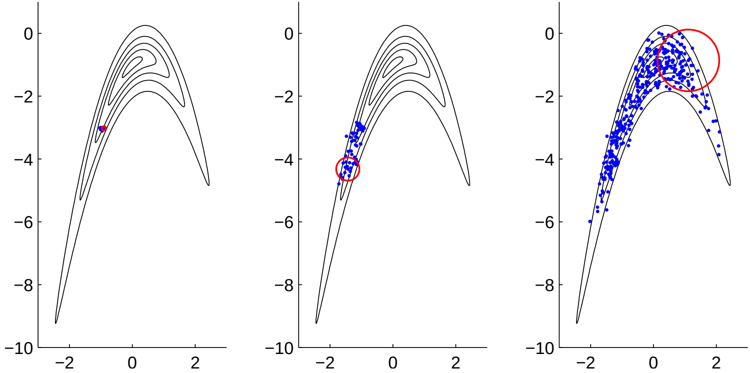


Figure: The first 10, 100 and 1000 samples of the ASM algorithm started with $\theta_1 = 0.01$ and using $\eta_n = n^{-3/4}$.

Robust adaptive Metropolis algorithm I

The RAM algorithm is a multidimensional 'extension' of the ASM adaptation mechanism [Vihola, 2010b].

- Define $X_1 \equiv x_1 \in \mathbb{R}^d$ such that $\pi(x_1) > 0$.
- Let $S_1 \equiv s_1 \in \mathbb{R}^{d \times d}$ lower diagonal with positive diagonal components.
- Define a sequence of positive adaptation step sizes $(\eta_n)_{n \geq 2}$ decaying to zero.
- Define the desired mean acceptance rate $\alpha^* \in (0, 1)$.

Robust adaptive Metropolis algorithm II

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

$$\begin{aligned} Y_n &= X_{n-1} + S_{n-1} W_n, \qquad \text{where } W_n \sim N(0, I), \text{ and} \\ X_n &= \begin{cases} Y_n, & \text{with probability } \alpha(X_{n-1}, Y_n) \text{ and} \\ X_{n-1}, & \text{otherwise.} \end{cases} \\ S_n S_n^T &= S_{n-1} \left(I + \eta_n \left[\alpha(X_{n-1}, Y_n) - \alpha^* \right] \frac{W_n W_n^T}{\|W_n\|^2} \right) S_{n-1}^T \end{aligned}$$

Robust adaptive Metropolis algorithm III

Figure: The RAM update with $\eta_n\left[\alpha(X_{n-1}, Y_n) - \alpha_*\right] = \pm 0.8$, resp. The ellipsoids defined by $S_{n-1}S_{n-1}^T$ ($S_n S_n^T$) are drawn in solid (dashed), and the vector $S_{n-1}W_n/\|W_n\|$ as a dot.

Example run of RAM

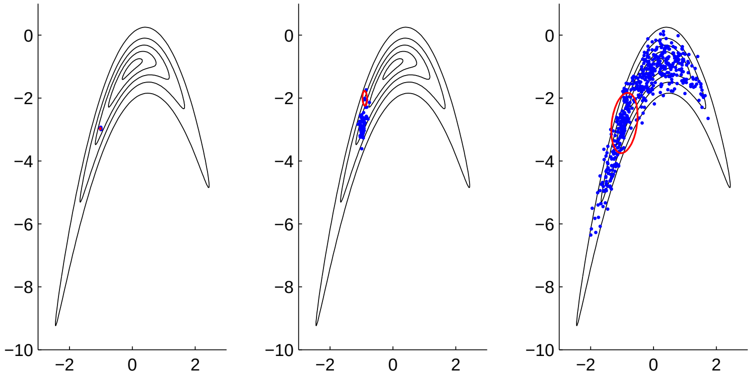


Figure: The first 10, 100 and 1000 samples of the RAM algorithm started with $S_1 = 0.01 \times I$ and using $\eta_n = \min\{1, 2 \cdot n^{-3/4}\}$.

Stochastic Approximation Framework I

- Andrieu and Robert [2001] observed the connection between Robbins-Monro stochastic approximation and adaptive MCMC.
- The algorithm has the following form (essentially)

$$\begin{aligned} X_n &\sim P_{\Theta_{n-1}}(X_{n-1}, \cdot\,) \\ \Theta_n &= \Theta_{n-1} + \eta_n H(\Theta_{n-1}, X_n). \end{aligned}$$

The recursion for $(\Theta_n)_{n\geq 1}$ can be considered as an attempt to find a root of the *mean field*

$$h(\theta) = \int_{\mathbb{R}^d} H(\theta, x)\pi(x)\mathrm{d}x.$$

Stochastic Approximation Framework II

- The AM, ASM and RAM algorithms can be formulated as Robbins-Monro stochastic approximation.
- In the case of AM $\Theta_n = (M_n, S_n)$, and whenever π has second moments, the mean field h has a unique root at (m_π, s_π), the mean and covariance of π, respectively.
- In case of ASM and RAM, the root of h may be non-unique, but there are no assumptions on the moments of π.

Previous conditions on validity I

- What conditions are sufficient to guarantee that a law of large numbers (LLN) holds: $\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{n\to\infty} \int f(x)\pi(x)\mathrm{d}x$?
- Key conditions due to Roberts and Rosenthal [2007]:

Diminishing adaptation (DA) The effect of the adaptation becomes smaller and smaller,

$$\text{``}\|P_{\theta_n} - P_{\theta_{n-1}}\| \xrightarrow{n\to\infty} 0.\text{''}$$

Containment (C) The kernels P_{θ_n} have all the time sufficiently good mixing properties:

$$\left(\sup_{n\geq 1}\sup_{k\geq m} \|P_{\theta_n}^k(X_n,\,\cdot\,) - \pi\|\right) \xrightarrow{m\to\infty} 0.$$

Previous conditions on validity II

- DA is usually easier to verify.
- Containment is often tightly related to the stability of the process $(\theta_n)_{n\geq 1}$.
- Establishing containment (or stability) is often difficult before showing that a LLN holds...
- Typical solution is to enforce containment.
  - In the case of AM, ASM and RAM algorithms, this means that the eigenvalues of Σ_n and θ_n are constrained to be within $[a, b]$ for some constants $0 < a \leq b < \infty$.

Previous conditions on validity III

There are also other results in the literature on the ergodicity of adaptive MCMC given conditions similar to DA and C.

- 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 recent paper by Atchadé and Fort [2010] is based on a resolvent kernel approach.

New results I

- Strong belief (based on overwhelming empirical evidence) that many algorithms (incl. AM, ASM and RAM) are intrinsically stable.

$\Longrightarrow$ Containment is satisfied automatically and does not need not be enforced.

- Key ideas in the new results:
  - Containment is, in fact, not a necessary condition for LLN.
  - That is, the ergodic properties of P_{θ_n} can become more and more unfavourable, and still a LLN can hold.
  - The adaptation mechanism may imply a ‘drift’ away from ‘bad values’ of θ.

General ergodicity result I

- Assume $K_1 \subset K_2 \subset \cdots \subset \Theta$ are increasing subsets of the adaptation space. Assume that the adaptation follows the stochastic approximation dynamic

$$\begin{aligned} X_n &\sim P_{\Theta_{n-1}}(X_{n-1}, \cdot\,) \\ \Theta_n &= \Theta_{n-1} + \eta_n H(\Theta_{n-1}, X_n). \end{aligned}$$

  and the adaptation parameter $\Theta_n \in K_n$ for all $n \geq 1$.
  - One may also enforce $\Theta_n \in K_n$...
- Consider the following conditions for constants $c < \infty$ and $0 \leq \epsilon \ll 1$:

General ergodicity result II

(A1) Drift and minorisation:

There is a drift function $V : \mathbb{X} \to [1, \infty)$ such that for all $n \geq 1$ and all $s \in K_n$

$$P_s V(x) \leq \lambda_n V(x) + \mathbb{1}_{C_n}(x) b_n \quad \text{and} \tag{1}$$

$$P_s(x, A) \geq \mathbb{1}_{C_n}(x)\delta_n \nu_n(A) \tag{2}$$

where $C_n \subset \mathbb{R}^d$ are Borel sets, $\delta_n, \lambda_n \in (0,1)$ and $b_n < \infty$ are constants and ν_n is concentrated on C_n. Furthermore, the constants are polynomially bounded so that

$$(1 - \lambda_n)^{-1} \vee \delta_n^{-1} \vee b_n \leq cn^{\epsilon}.$$

General ergodicity result III

(A2) Continuity:

For all $n \geq 1$ and any $r \in (0,1]$, there is $c' = c'(r) \geq 1$ such that for all $s, s' \in K_n$,

$$\|P_s f - P_{s'} f\|_{V^r} \leq c' n^{\epsilon} \|f\|_{V^r} |s - s'|.$$

(A3) Bound for adaptation function:

There is a $\beta \in [0, 1/2]$ such that for all $n \geq 1$, $s \in K_n$ and $x \in \mathbb{R}^d$

$$|H(s,x)| \leq c n^{\epsilon} V^{\beta}(x).$$

General ergodicity result IV

Theorem (Saksman and Vihola [2010])

Assume (A1)–(A3) hold and let f be a function with $\|f\|_{V^\alpha} < \infty$ for some $\alpha \in (0, 1-\beta)$. Assume $\epsilon < \kappa_*^{-1}[(1/2) \wedge (1-\alpha-\beta)]$, where $\kappa_* \gg 1$ is an independent constant, and that $\sum_{k=1}^\infty k^{\kappa_*\epsilon - 1}\eta_k < \infty$. Then,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n f(X_k) = \int f(x)\pi(x)\mathrm{d}x \qquad \textit{almost surely.}$$

Verifiable assumptions I

Definition (Strongly super-exponential target)

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

$$\limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0 \qquad \text{and}$$
$$\lim_{|x|\to\infty} \frac{x}{|x|^{\rho}} \cdot \nabla \log \pi(x) = -\infty,$$

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.)

Verifiable assumptions II

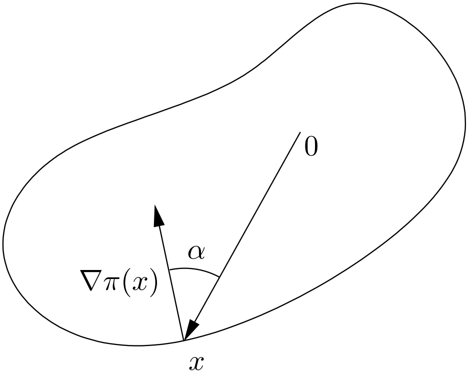


Figure: The condition $\limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0$ implies that there is an $\varepsilon > 0$ such that for any sufficiently large $|x|$, the angle $\alpha < \pi/2 - \varepsilon$.

Verifiable assumptions III

Definition

The function $f : \mathbb{R}^d \to \mathbb{R}$ has at most exponential tails if there are constants $M, \xi < \infty$ such that

$$|f(x)| \leq M e^{\xi |x|} \qquad \text{for all } x \in \mathbb{R}^d.$$

AM when π has unbounded support

- Saksman and Vihola [2010]: First ergodicity results for the original AM algorithm, without upper bounding the eigenvalues $\lambda(\Sigma_n)$.
- Use the proposal covariance $\Sigma_n = t^2 \operatorname{Cov}(X_1, \ldots, X_n) + \epsilon I$, 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

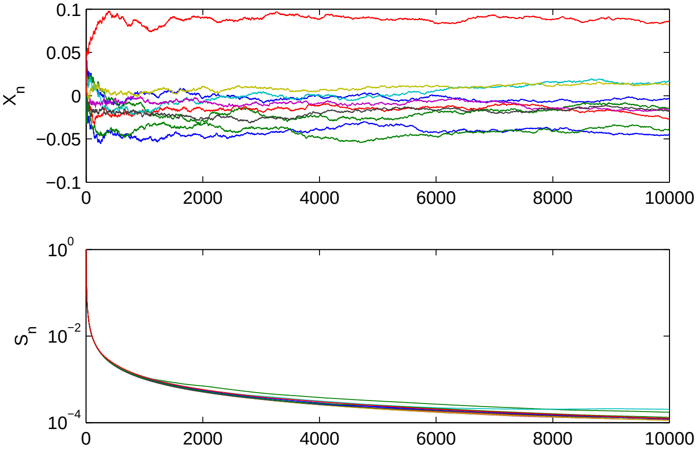
Vihola [2011] contains partial results on the case $\Sigma_n = t^2 S_n$, that is, without lower bounding the eigenvalues $\lambda(\Sigma_n)$.

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

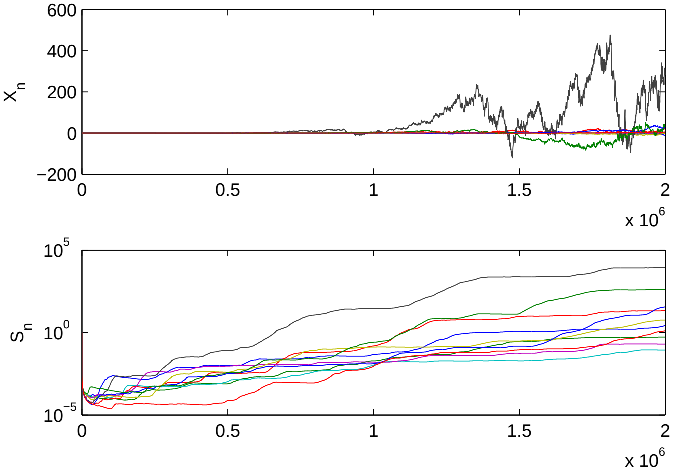
$$\begin{aligned} X_{n+1} &= X_n + t S_n^{1/2} W_{n+1} \\ S_{n+1} &= \frac{n}{n+1} S_n + \frac{1}{n+1} (X_{n+1} - \overline{X}_n)^2. \end{aligned}$$

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

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 $\mathbb{E}[S_n] \sim e^{2t\sqrt{n}}$.
- 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 $\lambda(\Sigma_n)$, 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

$$Y_n = X_{n-1} + \begin{cases} \Sigma_0^{1/2} W_n, & \text{with probability } \beta, \\ \Sigma_{n-1}^{1/2} W_n, & \text{otherwise.} \end{cases}$$

  with some fixed symm.pos.def. Σ_0.
- In other words, one employs a mixture of 'adaptive' and 'nonadaptive' Markov kernels:

$$\mathbb{P}\left(X_n \in A \mid X_1, \ldots, X_{n-1} \right) = (1-\beta) P_{\Sigma_{n-1}}(A) + \beta P_{\Sigma_0}(A)$$

AM with a fixed proposal component II

Vihola [2011] shows that, for example with a bounded and compactly supported π or with a super-exponential π, the eigenvalues $\lambda(\Sigma_n)$ 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 Σ_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.
- Also exponentially decaying π are considered; in this case the fixed covariance Σ_0 must be large enough.

Ergodicity of unconstrained ASM algorithm

Vihola [2009] 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, 2009] applies also in this case, provided that the eigenvalues of the covariance part are bounded within $0 < a \leq b < \infty$.
- The new RAM algorithm may be sometimes better than ASM within AM...

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 $p > 0$ [Vihola, 2009].

- 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 free software available for testing several adaptive random-walk MCMC algorithms, including the new RAM approach [Vihola, 2010a]:
`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.

References II

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

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.

References III

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.

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

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.

References IV

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.*, 20(6):2178–2203, Dec. 2010.

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

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

M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. Preprint, arXiv:1011.4381v1, Nov. 2010b.

References V

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