An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift

Yves F. Atchadé[1]

(March, 2005)

Abstract

This paper proposes an adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift (T-MALA). The scale parameter and the covariance matrix of the proposal kernel of the algorithm are simultaneously and recursively updated in order to reach the optimal acceptance rate of 0.574 (see Roberts and Rosenthal (2001)) and to estimate and use the correlation structure of the target distribution. We develop some convergence results for the algorithm. A simulation example is presented.

Key words: Markov Chain Monte Carlo, Stochastic approximation algorithms, Metropolis Adjusted Langevin algorithm, geometric rate of convergence.

MSC Numbers: 65C05, 65C40, 60J27, 60J35

1   Introduction

Markov Chain Monte Carlo (MCMC) is a well-established probabilistic tool to sample from probability measures known only up to a normalizing constant. A MCMC algorithm is designed by specifying a transition kernel with a predefined invariant probability measure. Such transition kernel typically depends on various parameters to be provided by the user. This is a strength of the method as it allows the user to possibly run the best algorithm for its problem by providing the appropriate parameter value. But finding the best value of the parameters for a given target distribution is a difficult analytical problem. This problem needs to be solve in a satisfactory way for MCMC to become routinely used by non-experts. Adaptive MCMC is a possible solution. The idea is to solve both problems (the sampling problem and the optimal parameter value finding problem) simultaneously by updating the transition kernel in the course of the simulation given the sample generated so far. Recently an approach based on stochastic approximation and recursive estimation has been developed and applied to the Independent Metropolis algorithm and to the Random Walk Metropolis (RWM) algorithm (Haario et al. (2001), Andrieu and Moulines (2003), Atchade and Rosenthal (2003)). In Haario et al. (2001) and Andrieu and Moulines (2003) the covariance matrix of the RWM algorithm is sequentially updated to find the correlation structure of the target distribution. In Atchade and Rosenthal (2003) the covariance matrix is fixed and the scale parameter of the RWM algorithm is sequentially updated to find the one that gives the optimal acceptance rate. The main objective of this paper is to extend this methodology to the Metropolis adjusted Langevin algorithm with a truncated drift (denoted T-MALA in the sequel). We update both the covariance matrix and the scale parameter simultaneously. It is worth noting that the algorithm and the results developed in the paper actually apply to any random walk Metropolis type algorithm with bounded drift; so they also apply to the RWM algorithm.

The adaptive T-MALA is proposed and analyzed in Section 2. A simulation example is presented in Section 3 to illustrate the algorithm. The proof are postponed to Section 4.

---

[1]Department of Mathematics and Statistics, University of Ottawa, email: yatchade@uottawa.ca

2 Adapting the T-MALA

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, the d-dimensional Euclidean space (equipped with its Borel subsets $\mathcal{B}^d$) and π a positive and continuously differentiable density (with respect to Lebesgue measure on $\mathcal{X}$). For $\delta > 0$, Define the drift function of the algorithm by $D(x) = \frac{\delta}{\max(\delta, |\nabla \log \pi(x)|)} \nabla \log \pi(x)$, where ∇ is the gradient operator. For a positive definite matrix Λ and a scale parameter $\sigma > 0$, let $q_{\sigma,\Lambda}(x,y)$ be the density (with respect to Lebesgue measure on $\mathcal{X}$) of $\mathcal{N}\left(x + \frac{\sigma^2}{2}D(x), \sigma^2\Lambda\right)$ the Gaussian distribution with mean $x + \frac{\sigma^2}{2}D(x)$ and covariance matrix $\sigma^2\Lambda$. The Truncated Metropolis Adjusted Langevin Algorithm (T-MALA) with proposal density $Q_{\sigma,\Lambda}(x, dy) = q_{\sigma,\Lambda}(x,y)dy$ has been introduced in Roberts and Tweedie (1996). This algorithm generates a Markov chain (X_n) with invariant distribution π as follows. Given X_n, a new proposal $Y_{n+1} \sim \mathcal{N}\left(X_n + \frac{\sigma^2}{2}D(X_n), \sigma^2\Lambda\right)$ is made. We then either "accept" the proposed value and set $X_{n+1} = Y_{n+1}$ with probability $\alpha_{\sigma,\Lambda}(X_n, Y_{n+1})$, or we "reject" it and set $X_{n+1} = X_n$ with probability $1 - \alpha_{\sigma,\Lambda}(X_n, Y_{n+1})$, where $\alpha_{\sigma,\Lambda}(x, y) = \min\left(1, \frac{\pi(y)q_{\sigma,\Lambda}(y,x)}{\pi(x)q_{\sigma,\Lambda}(x,y)}\right)$. Let $P_{\sigma,\Lambda}$ be the transition kernel of the Markov chain generated by such algorithm. We have:

$$P_{\sigma,\Lambda}(x, A) = \int_A \alpha_{\sigma,\Lambda}(x,y)q_{\sigma,\Lambda}(x,y)dy + r_{\sigma,\Lambda}(x)\mathbf{1}_A(x), \tag{2.1}$$

where

$$r_{\sigma,\Lambda}(x) = \int \left(1 - \alpha_{\sigma,\Lambda}(x,y)\right)q_{\sigma,\Lambda}(x,y)dy. \tag{2.2}$$

As its name indicates, the T-MALA is a truncated drift version of the Metropolis Adjusted Langevin algorithm (MALA) whose drift function is $D(x) = \nabla \log \pi(x)$. The MALA has better mixing properties than the Random Walk Metropolis algorithm, but its rate of convergence is often unstable due to the unbounded drift (see Roberts and Tweedie (1996)). Here we show (see Proposition 2.1) that the T-MALA has similar geometric convergence property as the Random Walk Metropolis algorithm.

2.1 An adaptive version of the T-MALA

The choice of the scaling parameters (σ, Λ) has a large effect on the mixing time of the T-MALA. It is believed that the strategy that works best is to take $\Lambda = \Sigma_\pi$ the covariance matrix of the distribution π and to choose σ so as to achieve a prescribed global acceptance rate in stationarity, (approximately 0.574 for Langevin type algorithms). Many theoretical works have been done that support this strategy (see e.g. Roberts and Rosenthal (2001), Breyer et al. (2004)). Clearly those optimal values are not known in general. Often in practice, tedious pilot simulations are necessary to first estimate those parameters. We propose an adaptive T-MALA that generates an inhomogeneous Markov chain $(X_n, \Lambda_n, \sigma_n)$ where no pilot simulation and parameter tuning is necessary.

Fix $0 < \varepsilon_1 < A_1 < \infty$ and $\varepsilon_2 > 0$. Write $\Theta_\sigma = [\varepsilon_1, A_1]$ equipped with the Euclidean norm of $\mathbb{R}$. Let Θ_Γ be the convex set of all semipositive definite matrices Γ with $|\Gamma| \le A_1$, where $|\Gamma| := tr^{1/2}(\Gamma\Gamma') = \left\{\sum_{ij} |\gamma_{ij}|^2\right\}^{1/2}$ is the Frobenius norm. This norm is derived from the scalar product $A \cdot B := tr(AB')$. We introduction three projection functions p_1, p_2, p_3 to contain the algorithm. $p_1(\sigma) = \sigma$ if $\sigma \in \Theta_\sigma$, $p_1(\sigma) = \varepsilon_1$ if $\sigma < \varepsilon_1$ and $p_1(\sigma) = A_1$ if $\sigma > A_1$. For a semidefinite positive matrix Σ, define $p_2(\Sigma) = \Sigma$ if $|\Sigma| \le A_1$ and $p_2(\Sigma) = \frac{A_1}{|\Sigma|}\Sigma$ if $|\Sigma| > A_1$. For $x \in \mathbb{R}^d$, $p_3(x) = x$ if $|x| \le A_1$ and $p_3(x) = \frac{A_1}{|x|}x$ if $|x| > A_1$. The p_i are orthogonal projections and satisfy:

$$|\sigma' - p_1(\sigma)| \le |\sigma' - \sigma|, \ \sigma' \in \Theta_\sigma, \ \sigma \in \mathbb{R}, \tag{2.3}$$

$$|\Gamma' - p_2(\Gamma)| \leq |\Gamma' - \Gamma|, \ \Gamma' \in \Theta_\Gamma, \ \Gamma \text{ semidefinite positive}, \tag{2.4}$$

and

$$|\mu' - p_3(\mu)| \leq |\mu' - \mu|, \ |\mu'| \leq A_1, \ \mu \in \mathbb{R}^d. \tag{2.5}$$

Let (γ_n) a sequence of positive numbers and $\bar{\tau}$ the optimal acceptance rate; (here $\bar{\tau} = 0.574$).

Algorithm 2.1. [Adaptive T-MALA]

1. Start the algorithm at some point $x_0 \in \mathcal{X}$, with $\mu_0 \in \mathbb{R}^d$, $\sigma_0 > 0$, Γ_0 semidefinite positive matrix.

2. Suppose that at time $n \geq 0$, we have $X_n \in \mathcal{X}$, μ_n, σ_n and Γ_n. Set $\Lambda_n = \Gamma_n + \varepsilon_2 I_d$.

   2.1 Generate $Y_{n+1} \sim \mathcal{N}\left(X_n + \frac{\sigma_n^2}{2} R(X_n), \sigma_n^2 \Lambda_n\right)$ and generate $U \sim \mathcal{U}(0,1)$.

   2.2 If $U \leq \alpha_{\sigma_n, \Lambda_n}(X_n, Y_{n+1})$, then set $X_{n+1} = Y_{n+1}$. Otherwise, set $X_{n+1} = X_n$.

   2.3 Set
   $$\sigma_{n+1} = p_1\left(\sigma_n + \gamma_n\left(\alpha_{\sigma_n, \Lambda_n}(X_n, Y_{n+1}) - \bar{\tau}\right)\right). \tag{2.6}$$
   $$\mu_{n+1} = p_3\left(\mu_n + \gamma_n\left(X_{n+1} - \mu_n\right)\right), \tag{2.7}$$
   $$\Gamma_{n+1} = p_2\left(\Gamma_n + \gamma_n\left((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)' - \Gamma_n\right)\right). \tag{2.8}$$

Remark 2.1.   1. At each step of the algorithm, a valid T-MALA in used. A small diagonal matrix is added to the current estimate of Σ_π. This improves the numerical stability of the algorithm (particularly if Σ_π is not positive definite) and is also crucial in proving the ergodicity of the algorithm.

2. The parameter $(\mu_n, \sigma_n, \Gamma_n)$ is sequentially updated with a stochastic approximation algorithm with re-projection on a fixed compact set. Stochastic approximations are well-known random iterative algorithms of the form $\theta_{n+1} = \theta_n + \gamma_n(h(\theta_n) + \varepsilon_{n+1})$ initiated by Robbins and Monro (1951) and used to find solutions of equations of the form $h(\theta) = 0$ when the function h is unknown and/or hard to compute; see Kushner and Yin (2003) and the references therein. This type of algorithms has been introduced in MCMC by Haario et al. (2001) in a restricted setting and by Andrieu and Robert (2002). By adapting simultaneously σ and Λ, the algorithm proposed here is sensibly better than (see the simulations below) the Random Walk Metropolis type algorithms proposed in Andrieu and Moulines (2003) and Atchade and Rosenthal (2003).

2.2   Ergodicity of the algorithm

We make the following assumptions.
Assumption A1: *Assume that π is positive with continuous first derivative such that*

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

and

$$\limsup_{|x| \to \infty} n(x) \cdot m(x) < 0,$$

where ∇ is the gradient operator, $n(x) = \frac{x}{|x|}$ and $m(x) = \frac{\nabla \pi(x)}{|\nabla \pi(x)|}$.
Assumption A2:

(i) $|\mu_\pi| \leq A_1$ and $|\Sigma_\pi| \leq A_1$, where $\mu_\pi = \int x\pi(dx)$ and $\Sigma_\pi = \int xx'\pi(dx) - \mu_\pi\mu_\pi'$.

(ii) There exists $\sigma_{opt} \in \Theta_\sigma$ such that $\tau(\sigma_{opt}) = \bar{\tau}$ and $(\sigma - \sigma_{opt})(\tau(\sigma) - \bar{\tau}) < 0$ whenever $\sigma \neq \sigma_{opt}$, where the acceptance rate in stationarity function τ is defined as

$$\tau(\sigma) = \int \pi(dx) \int \alpha_{\sigma,\Lambda_\pi}(x,y)q_{\sigma,\Lambda_\pi}(x,y)dy,$$

where $\Lambda_\pi = \Sigma_\pi + \varepsilon_2 I_d$.

Assumption A3: $\gamma_n = \mathcal{O}\left(n^{-\lambda}\right)$, $1/2 < \lambda \leq 1$.

Remark 2.2. 1. (A1) has been introduced in Jarner and Hansen (2000) to analyze the convergence rate of the RWM algorithm. Many densities of the form $e^{-p(x)}$ or $h(x)^{-p(x)}$, where p is polynomial are known to satisfy (A1). See Jarner and Hansen (2000) for more details.

2. It is always possible to choose A_1 such (A2)(i) hold, at least in theory. (A2)(ii) is difficult to check and actually may not hold. But we believe that σ_n may still converge to a solution of $\tau(\sigma) = \bar{\tau}$ even if τ is not decreasing and $\tau(\sigma) = \bar{\tau}$ has many solutions. In any case, it is worth noting that the ergodicity of the algorithm does not rely on (A2)(ii).

3. We recommend $\gamma_n = \frac{c_0}{n}$ for some constant c_0.

Theorem 2.1. Let (X_n) be the stochastic process generated by algorithm 2.1 on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define $V(x) = c\pi^{1/2}(x)$, where c is any constant such that $V \geq 1$.

(i) Assume (A1), (A3) and (A2)(i). There exists a finite constant K such that

$$\|\mathcal{L}_{x_0}(X_n)(\cdot) - \pi(\cdot)\|_{V^{1/2}} \leq Kn^{-\lambda}\log(n)V(x_0), \quad n \geq 2 \tag{2.9}$$

where $\mathcal{L}_{x_0}(X_n)$ is the distribution of X_n given that $X_0 = x_0$ and for a signed measure μ, $\|\mu\|_{V^{1/2}} := \sup_{|f|\leq V^{1/2}} |\mu(f)|$, $\mu(f) := \int f(x)\mu(dx)$.

Also, for any measurable function $f : \mathcal{X} \longrightarrow \mathbb{R}$ with $|f| \leq V^{1/2}$,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \pi(f) \quad \mathbb{P} - a.s. \tag{2.10}$$

(ii) Assume (A1)-(A3). Then $\Lambda_n \longrightarrow \Lambda_\pi = \Sigma_\pi + \varepsilon_2 I_d$ and $\sigma_n \longrightarrow \sigma_{opt}$ as $n \to \infty$, $\mathbb{P}$ a.s.

Proof. **(i)** Take $G(x,\sigma,\Lambda) = f(x)$ and apply Lemma 4.5 and Lemma 4.8.

(ii) is proved in Theorem 4.1.

$\square$

As part of the proof of Theorem 2.1 we will see that the transition kernel of the (nonadaptive) T-MALA has a geometric rate of convergence and is a smooth function of its parameters. These results are interesting on their own and are stated here.

For $0 < b_1 < b_2 < \infty$, let $\mathcal{C} = \mathcal{C}(b_1, b_2)$ be the set of all couples (σ, Λ) where $\sigma \in [b_1, b_2]$ and Λ is a positive definite matrix such that $|\Lambda| \leq b_2$ and such that the smallest eigenvalue of Λ is greater or equal to b_1. For $(\sigma, \Lambda) \in \mathbb{R} \times \mathbb{R}^{d^2}$, define the norm $|(\sigma, \Lambda)| := \left(|\sigma|^2 + |\Lambda|^2\right)^{1/2}$. $\mathcal{C}$ is convex and compact.

Proposition 2.1. *Assume (A1). For $0 < \alpha < 1$ write $V_\alpha(x) = c\pi^{-\alpha}(x)$ where c is such that $V_\alpha \geq 1$. Then there exist $\rho < 1$, $R < \infty$ such that:*

$$\sup_{(\sigma, \Lambda) \in \mathcal{C}} \left| P_{\sigma, \Lambda}^n(x, \cdot) - \pi(\cdot) \right| \leq R\rho^n V_\alpha(x), \quad n \geq 0, \ x \in \mathcal{X}. \tag{2.11}$$

Proof. See Section 4. □

We can also prove that $P_{\sigma, \Lambda} f(x)$ is a smooth function of (σ, Λ).

Proposition 2.2. *Under (A1), there is a constant $K_1 < \infty$ such that for $(\sigma_1, \Lambda_1), (\sigma_2, \Lambda_2) \in \mathcal{C}$:*

$$\sup_{|f| \leq V^{1/2}} |P_{\sigma_2, \Lambda_2} f(x) - P_{\sigma_1, \Lambda_1} f(x)| \leq K_1 V^{1/2}(x) |(\sigma_2 - \sigma_1, \Lambda_2 - \Lambda_1)|, \tag{2.12}$$

where $V(x) = c\pi^{-1/2}(x)$ with c chosen such that $V(x) \geq 1$.
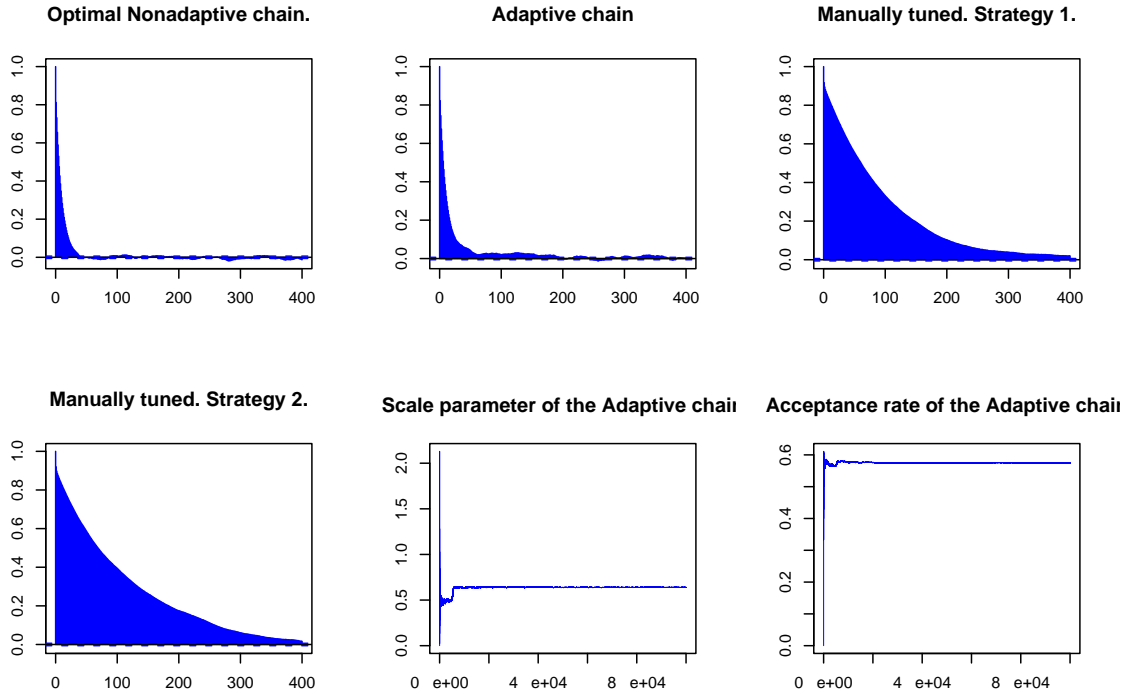
Proof. See Section 4. □

3 Simulation Example

We take π to be the 3-dimensional normal distribution with mean 0 and covariance matrix

$$\Sigma_\pi = \begin{pmatrix} 0.9575 & 2.4384 & -0.3741 \\ 2.4384 & 7.0338 & -1.0638 \\ -0.3741 & -1.0638 & 0.2632 \end{pmatrix}.$$

We compare 4 strategies to sample from π using the T-MALA. All the simulations are run for $n = 100,000$ iterations started from $X_0 = (5, 5, 5)$ and the drift is bounded by $\delta = 1,000$. We compare four different strategies.

1. A fully adaptive version of the T-MALA as presented in Algorithm 2.1. We use $\gamma_n = \frac{10}{n}$. $\varepsilon_1 = 10^{-4}$, $A_1 = 10^5$, $\varepsilon_2 = 0.01$. We start using the estimated covariance matrix only after $5,000$ iterations.

2. A nonadaptive T-MALA with the optimal values of the parameters σ_{opt} and Σ_π. $\sigma_{opt} = 0.6395$ (estimated from the adaptive chain).

3. A nonadaptive chain manually tuned. We take the proposal covariance matrix to be I_3 and estimate the value of σ that gives an acceptance rate of 0.574. We find $\sigma = 0.49$. Many trial-and-errors were required. We call this Strategy 1.

4. Finally we also try a nonadaptive chain where the covariance matrix has been manually tuned first and given the estimate of Σ_π obtained, σ was tuned to reach the 0.574 acceptance rate. Even more trial-and-errors were required. We call this Strategy 2.

We only look at the first component of the chains. Graph 1 displays the correlation functions of the output of the 4 strategies. We see that the adaptive chain is almost optimal and clearly outperforms the two strategies where the parameters are manually tuned. We also show the scale parameter σ_n and the acceptance rate of the adaptive chain.

Graph 1: Autocorrelation functions of the four strategies implemented, and scale parameter and acceptance rate of the adaptive chain.

4   Proofs of the results

We prove Proposition 2.1 in Section 4.1 and Proposition 2.2 in Section 4.2. The main theorem (Theorem 2.1) is proved in Section 4.3.

4.1   Proof of Proposition 2.1

Essentially, the idea of the proof is the same as the proof of the geometric ergodicity of the RWM algorithm developed by Jarner and Hansen (2000). There are some additional technicalities due to the drift of the algorithm. But the fact that the drift is bounded is crucial.

Proof of Proposition 2.1. In Lemma 4.1 below we show that there are $\varepsilon > 0$, a Ball C, a nontrivial probability measure ν such that:

$$\inf_{(\sigma,\Lambda)\in\mathcal{C}} P_{\sigma,\Lambda}(x, A) \geq \varepsilon\nu(A), \;\; A \in \mathcal{B}, \;\; x \in C,$$

and in Lemma 4.2 below we show that we can find $\lambda < 1$, $b < \infty$ such that

$$\sup_{(\sigma,\Lambda)\in\mathcal{C}} P_{\sigma,\Lambda}V_\alpha(x) \leq \lambda V_\alpha(x) + b\mathbf{1}_C(x), \;\; x \in \mathcal{X},$$

where C is as above and $V_\alpha(x) = c\pi^{-\alpha}(x)$, $0 < \alpha < 1$ and c is such that $V_\alpha(x) \geq 1$.

The theorem then follows from the fact that geometric bound for Markov chain can be obtained from these two inequalities based solely on the constants $\varepsilon > 0$, C, ν, $\lambda < 1$, $b < \infty$ and V_α. See e.g. Meyn and Tweedie (1994). □

Lemma 4.1. *There is $\varepsilon > 0$, a Ball C, a nontrivial probability measure ν such that:*
$$\inf_{(\sigma,\Lambda)\in\mathcal{C}} P_{\sigma,\Lambda}(x, A) \geq \varepsilon\nu(A), \quad A \in \mathcal{B} \; x \in C.$$

Proof. For $a > 0$, let g_a be the density of the d-dimensional normal distribution with zero mean and covariance matrix aI_d. Because the drift of the algorithm is bounded by δ and $(\sigma, \Lambda) \in \mathcal{C}$, we can find $\varepsilon_1 > 0$ and $k_1 > 0$ such that $\inf_{(\sigma,\Lambda)\in\mathcal{C}} q_{\sigma,\Lambda}(x, y) \geq k_1 g_{\varepsilon_1}(y - x)$. Take $R > 0$ and $C = B(0, R)$. Define $\tau = \min_{(\sigma,\Lambda)\in\mathcal{C}} \min_{y-x,x\in C} \frac{\pi(y)q_{\sigma,\Lambda}(y,x)}{\pi(x)q_{\sigma,\Lambda}(x,y)}$. $\tau > 0$. Write $\varepsilon = \tau k_1$ and $\nu(A) = \frac{\int_{A\cap C} g_{\varepsilon_1}(z)dz}{\int_C g_{\varepsilon_1}(z)dz}$. We have $\inf_{(\sigma,\Lambda)\in\mathcal{C}} P_{\Lambda}(x, A) \geq \varepsilon\nu(A)\mathbf{1}_C(x)$ as needed. $\quad\square$

Lemma 4.2. *Assume (A1) and let α and V_α as in Proposition 2.1. There exist $\lambda < 1$, $b < \infty$ such that $\sup_{(\sigma,\Lambda)\in\mathcal{C}} P_{\sigma,\Lambda}V_\alpha(x) \leq \lambda V_\alpha(x) + b\mathbf{1}_C(x)$, $x \in \mathcal{X}$, where C can be chosen as in Lemma 4.1.*

Proof. We only need to show that:

$$\sup_{x\in\mathcal{X}} \sup_{(\sigma,\Lambda)\in\mathcal{C}} \frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)} < \infty, \tag{4.1}$$

and

$$\limsup_{|x|\to\infty} \sup_{(\sigma,\Lambda)\in\mathcal{C}} \frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)} < 1. \tag{4.2}$$

See Jarner and Hansen (2000) Lemma 3.5.

For $x \in \mathcal{X}$, note $A_{\sigma,\Lambda}(x) = \{y : \frac{\pi(y)q_{\sigma,\Lambda}(y,x)}{\pi(x)q_{\sigma,\Lambda}(x,y)} \geq 1\}$ and $R_{\sigma,\Lambda}(x) = A_{\sigma,\Lambda}(x)^c$ the complement of $A_{\sigma,\Lambda}(x)$. Because the drift of the algorithm is bounded and $(\sigma, \Lambda) \in \mathcal{C}$, we can find $0 < \varepsilon_1 < \varepsilon_2 < \infty$, $0 < k_1 < k_2 < \infty$ such that:

$$k_1 g_{\varepsilon_1}(y - x) \leq q_{\sigma,\Lambda}(x, y) \leq k_2 g_{\varepsilon_2}(y - x), \tag{4.3}$$

where for a positive number a, g_a is the density of the d-dimensional normal distribution with mean 0 and covariance matrix aI_d. We have:

$$
\begin{aligned}
\frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)} &= \int_{A_{\sigma,\Lambda}(x)} q_{\sigma,\Lambda}(x, y)\frac{V_\alpha(y)}{V_\alpha(x)}dy + \int_{R_{\sigma,\Lambda}(x)} \frac{\pi(y)q_{\sigma,\Lambda}(y,x)V_\alpha(y)}{\pi(x)q_{\sigma,\Lambda}(x,y)V_\alpha(x)}q_{\sigma,\Lambda}(x,y)dy \\
&\quad + \int_{R_{\sigma,\Lambda}(x)} \left(1 - \frac{\pi(y)q_{\sigma,\Lambda}(y,x)}{\pi(x)q_{\sigma,\Lambda}(x,y)}\right)q_{\sigma,\Lambda}(x,y)dy \\
&\leq Q_{\sigma,\Lambda}(x, R_{\sigma,\Lambda}(x)) + \int_{A_{\sigma,\Lambda}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}q_{\sigma,\Lambda}(x,y)dy \\
&\quad + \int_{R_{\sigma,\Lambda}(x)} \left(\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}\right)q_{\sigma,\Lambda}(x,y)dy.
\end{aligned}
$$

On $A_{\sigma,\Lambda}(x)$,

$$\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}q_{\sigma,\Lambda}(x,y) \leq q_{\sigma,\Lambda}^\alpha(y,x)q_{\sigma,\Lambda}^{1-\alpha}(x,y) \leq k_2^2 g_{\varepsilon_2}(y-x), \tag{4.4}$$

and on $R_{\sigma,\Lambda}(x)$,

$$\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}q_{\sigma,\Lambda}(x,y) \leq q_{\sigma,\Lambda}^{1-\alpha}(y,x)q_{\sigma,\Lambda}^\alpha(x,y) \leq k_2^2 g_{\varepsilon_2}(y-x). \tag{4.5}$$

Hence (4.1) is satisfied.

Let $\varepsilon > 0$. we can find $R < \infty$ such that:

$$\int_{B(x,R)} g_{\varepsilon_2}(y-x)dy \geq 1 - \varepsilon. \tag{4.6}$$

Define $C_{\pi(x)} = \{y: \ \pi(y) = \pi(x)\}$ and for $u > 0$, $C_{\pi(x)}(u) = \{y + sn(y): \ y \in C_{\pi(x)}, \ -u \leq s \leq u\}$. Because π super-exponential, we can find r_1 such that for $|x| \geq r_1$, any point $y \in \mathcal{X}$ can be written $y = x_1 + sn(x_1)$ for $s \in \mathbb{R}$ and $x_1 \in C_{\pi(x)}$.

From (4.3) and the proof of Theorem 4.1 of Jarner and Hansen (2000), it follows that we can find $u > 0$ and $r_2 > r_1$ such that for $|x| \geq r_2$,

$$\int_{C_{\pi(x)}(u) \cap B(x,R)} g_{\varepsilon_2}(y-x)dy \leq \varepsilon. \tag{4.7}$$

Now, for $S \in \{A_{\sigma,\Lambda}(x), R_{\sigma,\Lambda}(x)\}$ and u as in (4.7), write $S = (S \cap B(x,R)^c) \bigcup (S \cap B(x,R) \cap C_{\pi(x)}(u)) \bigcup (S \cap B(x,R) \cap C_{\pi(x)}(u)^c)$. For $|x| \geq r_2$, it follows from (4.4), (4.6) and (4.7) that:

$$\int_{A_{\sigma,\Lambda}(x) \cap B(x,R)^c} q_{\sigma,\Lambda}(x,y) \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} dy + \int_{A_{\sigma,\Lambda}(x) \cap B(x,R) \cap C_{\pi(x)}(u)} q_{\sigma,\Lambda}(x,y) \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} dy \leq 2k_2^2 \varepsilon, \tag{4.8}$$

and from (4.5), (4.6) and (4.7) we have:

$$\int_{R_{\sigma,\Lambda}(x) \cap B(x,R)^c} \left(\frac{\pi^{1-\alpha}(y) q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x) q_{\sigma,\Lambda}(x,y)}\right) q_{\sigma,\Lambda}(x,y)dy \ + \int_{R_{\sigma,\Lambda}(x) \cap B(x,R) \cap C_{\pi(x)}(u)} \left(\frac{\pi^{1-\alpha}(y) q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x) q_{\sigma,\Lambda}(x,y)}\right) q_{\sigma,\Lambda}(x,y)dy$$
$$\leq \ 2k_2^2 \varepsilon. \tag{4.9}$$

For $r > 0$ and $a > 0$, write $d_r(a) = \sup_{|x| \geq r} \frac{\pi(x + an(x))}{\pi(x)}$. That π is super-exponential implies that $d_r(a) \to 0$ as $r \to \infty$. From this we can show that $r_3 < \infty$ exists such that for $|x| \geq r_3 + R$:

$$\int_{A_{\sigma,\Lambda}(x) \cap B(x,R) \cap C_\pi(\delta)^c} q_{\sigma,\Lambda}(x,y) \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} dy \leq d_{r_3}(\delta). \tag{4.10}$$

and

$$\int_{R_{\sigma,\Lambda}(x) \cap B(x,R) \cap C_{\pi(x)}(\delta)^c} \left(\frac{\pi^{1-\alpha}(y) q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x) q_{\sigma,\Lambda}(x,y)}\right) q_{\sigma,\Lambda}(x,y)dy \leq k_2 d_{r_3}(u). \tag{4.11}$$

The bounds (4.8), (4.9), (4.10), (4.11) implies that:

$$\limsup_{|x| \to \infty} \sup_{(\sigma,\Lambda) \in \mathcal{C}} \frac{P_{\sigma,\Lambda} V_\alpha(x)}{V_\alpha(x)} = \limsup_{|x| \to \infty} \sup_{(\sigma,\Lambda) \in \mathcal{C}} Q(x, R_{\sigma,\Lambda}(x))$$
$$= 1 - \liminf_{|x| \to \infty} \inf_{(\sigma,\Lambda) \in \mathcal{C}} Q_{\sigma,\Lambda}(x, A_{\sigma,\Lambda}(x)). \tag{4.12}$$

For $R > 0$, we can find $c_0 > 0$ such that $\inf_{y \in B(x,R)} \inf_{(\sigma,\Lambda) \in \mathcal{C}} \frac{q_{\sigma,\Lambda}(y,x)}{q_{\sigma,\Lambda}(x,y)} \geq c_0$.Take $u > 0$. Because π is super-exponential, $\pi(x - un(x)) \geq \frac{\pi(x)}{c_0}$ for any x such that $|x|$ is sufficiently large. Thus, for $|x|$ sufficiently large and $u < R$, $x_1 = x - un(x) \in A_{\sigma,\Lambda}(x)$. For $\varepsilon > 0$ arbitrary small define $W(x) = \{x_1 - a\zeta, \ 0 < a < R - u, \ \zeta \in S^{d-1}, \ |\zeta - n(x_1)| < \varepsilon/2\}$, where $\mathcal{S}^{d-1}$ is the unit-sphere in $\mathbb{R}^d$. We show that for $|x|$ sufficiently large, $W(x) \subset A_{\sigma,\Lambda}(x)$ for all $(\sigma,\Lambda) \in \mathcal{C}$. therefore $Q_{\sigma,\Lambda}(x, A_{\sigma,\Lambda}(x)) \geq k_2 \int_{W(x)} q_{\varepsilon_1}(y-x)dy = c > 0$. This together with (4.12) shows (4.2) and the Proposition will be proved.

Assumption (A1) implies that for $|x|$ sufficiently large, $m(x) \cdot n(x) < -\varepsilon$. Also for $|x|$ sufficiently large, $|n(y) - n(x)| < \varepsilon/2$ for any $y \in W(x)$. For any $y \in W(x)$, $m(y) \cdot \zeta = m(y) \cdot (\zeta - n(x_1) + n(x_1) - n(y) + n(y)) < \varepsilon/2 + \varepsilon/2 - \varepsilon = 0$, for $|x|$ sufficiently large. For $y = x_1 - a\zeta \in W(x)$, consider the function $f(t) = \pi(x_1 - t\zeta)$. $f(0) = \pi(x_1)$, $f(a) = \pi(y)$ and f is differentiable. Therefore there is $\tau \in (0, a)$ such that $f(a) - f(0) = -a\tau\zeta \cdot \nabla\pi(x_1 - \tau\zeta) > 0$ as seen above. Therefore $\pi(y) > \pi(x_1)$ which implies that $y \in A_{\sigma,\Lambda}(x)$ for $|x|$ sufficiently large. $\qquad\square$

4.2 Proof of Proposition 2.2

Proof. We only sketch the proof leaving the details to the reader. The idea is to show that for any $x \in \mathcal{X}$, there exists a finite constant K such that $\sup_{(\sigma,\Lambda)\in\mathcal{C}} \left\| \frac{\partial}{\partial(\sigma,\Lambda)} P_{\sigma,\Lambda} f(x) \right\| \le KV^{1/2}(x)$, where $\left\| \frac{\partial}{\partial(\sigma,\Lambda)} P_{\sigma,\Lambda} f(x) \right\|$ is the norm of the differential of $P_{\sigma,\Lambda} f(x)$ (x fixed) seen as a linear functional on $\mathbb{R} \times \mathbb{R}^{d^2}$. Since $\mathcal{C}$ is convex, the result follows from mean value theorem.

Write $r_{\sigma,\Lambda}(x, y) = \frac{\pi(y)q_{\sigma,\Lambda}(y,x)}{\pi(x)q_{\sigma,\Lambda}(x,y)}$ and $\alpha_{\sigma,\Lambda}(x, y) = \min(1, r_{\sigma,\Lambda}(x, y))$, so that

$$P_{\sigma,\Lambda} f(x) = \int \min(1, r_{\sigma,\Lambda}(x, y)) f(y) q_{\sigma,\Lambda}(x, y) dy + f(x) \int (1 - \alpha_{\sigma,\Lambda}(x, y)) q_{\sigma,\Lambda}(x, y) dy.$$

It is not hard to show that for $(h, H) \in \mathbb{R}\times\mathbb{R}^{d^2}$, the derivative of $q_{\sigma,\Lambda}(x, y)$ with respect to (σ, Λ) evaluated at (h, H) can be written: $\frac{\partial}{\partial(\sigma,\Lambda)} q_{\sigma,\Lambda}(x, y)(h, H) = q_{\sigma,\Lambda}(x, y)(B_1(x, y, \sigma, \Lambda, h) + B_2(x, y, \sigma, \Lambda, H))$, where the functions B_1, B_2 satisfy: $|B_1(x, y, \sigma, \Lambda, h)| + |B_2(x, y, \sigma, \Lambda, H)| \le K_2 |y - x|^2 |(h, H)|$ for some finite constant K_2. And a straightforward calculus gives for any (h, H) with $|(h, H)| \le 1$:

$$\begin{aligned}
\left| \frac{\partial}{\partial(\sigma,\Lambda)} \left[(\alpha_{\sigma,\Lambda}(x, y) q_{\sigma,\Lambda}(x, y)) f(y) \right] (h, H) \right| &= \left| \frac{\pi(y)}{\pi(x)} \mathbf{1}_{\{r_{\sigma,\Lambda}(x,y)\le 1\}} \frac{\partial}{\partial(\sigma,\Lambda)} \left[q_{\sigma,\Lambda}(y, x) f(y) \right] (h, H) \right. \\
&\qquad \left. + \mathbf{1}_{\{r_{\sigma,\Lambda}(x,y)>1\}} \frac{\partial}{\partial(\sigma,\Lambda)} \left[q_{\sigma,\Lambda}(y, x) f(y) \right] (h, H) \right| \\
&\le K_2 |y - x|^2 V^{1/2}(x) q_{\varepsilon_2}(x, y), \qquad\qquad (4.13)
\end{aligned}$$

for some finite constant $\varepsilon_2 > 0$ where q_{ε_2} is the density of the d-dimensional normal distribution with mean 0 and covariance $\varepsilon_2 I_d$. Similarly, $\left| \frac{\partial}{\partial(\sigma,\Lambda)} \left[(1 - \alpha_{\sigma,\Lambda}(x, y)) q_{\sigma,\Lambda}(x, y) \right] (h, H) \right| \le K_2 |y - x|^2 q_{\varepsilon_2}(x, y)$.

Thus $P_{\sigma,\Lambda} f(x)$ is differentiable under the integral and:

$$\left| \frac{\partial}{\partial(\sigma,\Lambda)} P_{\sigma,\Lambda} f(x)(h, H) \right| \le K_2 V^{1/2}(x) \int |y - x|^2 q_{\varepsilon_2}(x, y) dy, \qquad\qquad (4.14)$$

and we are done. $\qquad\square$

4.3 Proof of Theorem 2.1

Showing the convergence of the stochastic approximation processes (point (ii) of the theorem) is slightly harder than showing that the algorithm is ergodic (point (i) of the Theorem). The ergodicity of the algorithm is proved as in Atchade and Rosenthal (2003) through Lemma 4.5 and Lemma 4.8. Essentially, we need the uniform (in (σ, Λ)) rate of convergence as shown in Proposition 2.1 and the fact that the adaptation is diminishing (Proposition 2.2 and Lemma 4.4). The adapting process $(\mu_n, \sigma_n, \Lambda_n)$ need not converge. To prove the convergence of the stochastic approximation algorithm, we use an improved version of the Robbins-Siegmund Theorem (Lemma 4.7).

Unless otherwise stated, (X_n) refers to the random process generated by Algorithm 2.1 on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $b_1 = \min(\varepsilon_1, \varepsilon_2)$ and $b_2 = A_1 + \varepsilon_2$. Let $\Theta = \mathcal{C}(b_1, b_2)$ be the set of all couples (σ, Λ) where $\sigma \in [b_1, b_2]$ and Λ is a positive definite matrix such that $|\Lambda| \leq b_2$ and the smallest eigenvalue of Λ is greater or equal to b_1. For $n \geq 1$, $\theta_n = (\sigma_n, \Lambda_n) \in \Theta$ and the minorization and drift conditions established in Lemma 4.1 and Lemma 4.2 are readily available and the constants involved are independent from n. A repetitive application of this uniform drift condition yields the following simple Lemma.

Lemma 4.3. *Assume (A1). For any $\alpha \in (0, 1]$ there is a constant $R_1 = R_1(\alpha) < \infty$ such that for $n \geq 0$, $j \geq 0$*

$$\mathbb{E}\left(V^\alpha(X_{n+j})|\mathcal{F}_n\right) \leq R_1 V^\alpha(X_n). \tag{4.15}$$

The next lemma will allow us to control the variations in the stochastic approximation algorithms.

Lemma 4.4. *Assume (A1). There exists $R_2 < \infty$ such that for $n \geq 0$:*

$$|\mu_{n+1} - \mu_n| + |\Gamma_{n+1} - \Gamma_n| + |\sigma_{n+1} - \sigma_n| \leq R_2 \gamma_n V^{1/2}(X_{n+1}). \tag{4.16}$$

Proof. Follows from (2.3)-(2.5), the fact that μ_n, Γ_n, σ_n are bounded and the fact that $|x| + |x|^2 \leq V^{1/2}(x)$. $\qquad\square$

The following lemma is adapted from Atchade and Rosenthal (2003). Let $G : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}$ be a measurable function, where $\Theta = \mathcal{C}(b_1, b_2)$. Assume that there exist constants $K_2, K_3 < \infty$ such that:

$$\sup_{\theta \in \Theta} |G(x, \theta)| \leq K_2 V^{1/2}(x), \ x \in \mathcal{X}, \tag{4.17}$$

$$|G(x, \theta_1) - G(x, \theta_2)| \leq K_3 V^{1/2}(x) |\theta_1 - \theta_2|, \ \ \theta_1, \theta_2 \in \Theta, x \in \mathcal{X}. \tag{4.18}$$

Define $g(\theta) := \int G(x, \theta) \pi(dx)$.

Lemma 4.5. *Assume (A1). Let G and g as defined above. Then there exist constants $C_1 < \infty$ $0 < \rho < 1$ (that depend on G only through K_2 and K_3) such that for $n \geq 0$, $k \geq 0$,*

$$|\mathbb{E}\left(G(X_{n+k}, \theta_{n+k}) - g(\theta_{n+k})|\mathcal{F}_n\right)| \leq C_1 \left(\rho^k + \gamma_n k\right) V(X_n), \ \ \mathbb{P} - a.s. \tag{4.19}$$

Proof. Define $f_n(x) = G(x, \theta_n) - g(\theta_n)$. Then
$G(X_{n+k}, \theta_{n+k}) - g(\theta_{n+k}) = f_n(X_{n+k}) + G(X_{n+k}, \theta_{n+k}) - G(X_{n+k}, \theta_n) + g(\theta_n) - g(\theta_{n+k})$.
From (4.18) it follows that $|g(\theta_2) - g(\theta_1)| \leq K_3 \pi(V^{1/2}) |\theta_2 - \theta_1|$. Thus:

$$|G(X_{n+k}, \theta_{n+k}) - G(X_{n+k}, \theta_n)| + |g(\theta_n) - g(\theta_{n+k})| \leq R_3 V^{1/2}(X_{n+k}) |\theta_{n+k} - \theta_n|, \tag{4.20}$$

for some finite constant R_3. Therefore:

$$|\mathbb{E}\left(G(X_{n+k}, \theta_{n+k}) - g(\theta_{n+k})|\mathcal{F}_n\right)| \leq |\mathbb{E}\left(f_n(X_{n+k})|\mathcal{F}_n\right)| + R_3 \mathbb{E}\left(V^{1/2}(X_{n+k}) |\theta_{n+k} - \theta_n| \,|\mathcal{F}_n\right). \tag{4.21}$$

An argument similar to the one used in Atchade and Rosenthal (2003) (Lemma 3.1) can be used here to show that

$$|\mathbb{E}\left(f_n(X_{n+k})|\mathcal{F}_n\right)| \leq R_2 \rho^k V^{1/2}(X_n) + R_4 \sum_{j=1}^{k-1} \rho^{k-1-j} \mathbb{E}\left(V^{1/2}(X_{n+j}) |\theta_{n+j} - \theta_n| \,|\mathcal{F}_n\right). \tag{4.22}$$

Back to (4.21), (4.22) gives:

$$\left|\mathbb{E}\left(G(X_{n+k}, \theta_{n+k}) - g(\theta_{n+k})|\mathcal{F}_n\right)\right| \leq R_2 \rho^k V^{1/2}(X_n) + R_5 \sum_{j=1}^{k} \rho^{k-j} \mathbb{E}\left(V^{1/2}(X_{n+j})\left|\theta_{n+j} - \theta_n\right||\mathcal{F}_n\right)$$

$$\leq C_1\left(\rho^k + \gamma_n k\right) V(X_n), \tag{4.23}$$

using Lemma 4.4 and Lemma 4.3 for some finite constant C_1. $\qquad\square$

We introduce the functions $e_1(x) = x$, $e_2(x) = (x - \mu_\pi)(x - \mu_\pi)'$, $A(x, \sigma, \Lambda) = \int \alpha_{\sigma,\Lambda}(x, y) q_{\sigma,\Lambda}(x, y) dy$, and $\tau(\sigma, \Lambda) = \int A(x, \sigma, \Lambda)\pi(dx)$. For two matrices A and B recall that the scalar product of A and B is $A \cdot B = tr(AB')$.

Lemma 4.6. *Assume (A1) and (A3). Then a.s. we have:*

(i) $\sum \gamma_n^2 V^{1/2}(X_n) < \infty.$

(ii) $\sum \gamma_n(\mu_n - \mu_\pi) \cdot (P_{\sigma_n,\Lambda_n} e_1(X_n) - \mu_\pi) < \infty.$

(iii) $\sum \gamma_n(\Gamma_n - \Sigma_\pi) \cdot (P_{\sigma_n,\Lambda_n} e_2(X_n) - \Sigma_\pi) < \infty.$

(iv) $\sum \gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((\mu_n - \mu_\pi)(P_{\sigma_n,\Lambda_n} e_1(X_n) - \mu_\pi))' < \infty.$

(v) $\sum \gamma_n(\sigma_n - \sigma_{opt}) \cdot (A(X_n, \sigma_n, \Lambda_n) - \tau(\sigma_n, \Lambda_n)) < \infty.$

Proof. The idea is to choose the appropriate function and to apply Lemma 4.5 and Lemma 4.8 with $\mathcal{F}_n = \sigma(X_0, \sigma_0, \ldots, X_n, \sigma_n).$

(i) Take $G(x, \theta) = V^{1/2}(x)$. Recall that $\theta = (\sigma, \Lambda)$. Lemma 4.5 implies that $\left|\mathbb{E}(V^{1/2}(X_{n+k}) - \pi(V^{1/2})|\mathcal{F}_n)\right| \leq C_1\left(\rho^k + \gamma_n k\right) V(X_n)$. Then Lemma 4.8 below implies that $\sum \gamma_n^2(V^{1/2}(X_n) - \pi(V^{1/2})) < \infty$ and since $\sum \gamma_n^2 < \infty$, $\sum \gamma_n^2 V^{1/2}(X_n) < \infty.$

(ii) Take $G(x, \theta) = P_\theta e_1(x)$. Then $\int G(x, \theta)\pi(dx) = \mu_\pi$, $|G(x, \theta)| \leq K P_\theta V^{1/2}(x) \leq K V^{1/2}(x)$ for some finite constant K and from Proposition 2.2 we have $|P_{\theta_2} e_1(x) - P_{\theta_1} e_1(x)| \leq K_3 |\theta_2 - \theta_1| V^{1/2}(x)$. Therefore by writing

$$(\mu_{n+k} - \mu_\pi) \cdot \left(P_{\theta_{n+k}} e_1(X_{n+k}) - \mu_\pi\right) = (\mu_{n+k} - \mu_n) \cdot \left(P_{\theta_{n+k}} e_1(X_{n+k}) - \mu_\pi\right)$$
$$+ (\mu_n - \mu_\pi) \cdot \left(P_{\theta_{n+k}} e_1(X_{n+k}) - \mu_\pi\right),$$

and applying Lemma 4.5 we get:

$$\left|\mathbb{E}\left((\mu_{n+k} - \mu_\pi) \cdot \left(P_{\theta_{n+k}} e_1(X_{n+k}) - \mu_\pi\right)|\mathcal{F}_n\right)\right| \leq K_1 k \gamma_n V^{1/2}(X_n) + C_1(\rho^n + k\gamma_n)V(X_n)$$
$$\leq C_2(\rho^n + k\gamma_n)V(X_n).$$

Lemma 4.8 then implies $\sum \gamma_n(\mu_n - \mu_\pi) \cdot (P_{\theta_n} e_1(X_n) - \mu_\pi)$ converge to a finite random variable.

(iii) Similar to (ii) with $G(x, \theta) = e_2(x).$

(iv) Similar arguments as in (ii) and (iii).

(v) Take $G(x, \theta) = A(x, \theta) = A(x, \sigma, \Lambda)$. As one can see by applying the mean value theorem from Equation (4.13) in the proof of Proposition 2.2, $A(x, \cdot)$ is Liptschitz. For $n \geq 0$, $k \geq 0$:

$(\sigma_{n+k} - \sigma_{opt}) \cdot (A(X_{n+k}, \sigma_{n+k}, \Lambda_{n+k}) - \tau(\sigma_{n+k}, \Lambda_{n+k})) \leq k\gamma_n$
$+ (\sigma_n - \sigma_{opt}) \cdot (A(X_{n+k}, \sigma_{n+k}, \Lambda_{n+k}) - \tau(\sigma_{n+k}, \Lambda_{n+k})).$

Then Lemma 4.5 gives: $|\mathbb{E}((\sigma_{n+k} - \sigma_{opt}) \cdot (A(X_{n+k}, \sigma_{n+k}, \Lambda_{n+k}) - \tau(\sigma_{n+k}, \Lambda_{n+k}))|\mathcal{F}_n)| \leq k\gamma_n$
$+ C_1\left(\rho^k + \gamma_n k\right) V(X_n)$ and applying Lemma 4.8 once more yields (v).

$\square$

We are now ready to prove the convergence of the stochastic approximation processes.

Theorem 4.1. *Assume (A1)-(A3). Then:*

(i) $\mu_n \longrightarrow \mu_\pi$ *a.s., as* $n \to \infty$.

(ii) $\Gamma_n \to \Sigma_\pi$ *a.s. as* $n \to \infty$.

(iii) $\sigma_n \to \sigma_{opt}$ *a.s. as* $n \to \infty$.

Proof. **(i)** For $n \geq 0$, define $\mathcal{F}_n = \sigma(X_0, \sigma_0, \ldots, X_n, \sigma_n)$. We have:

$$
\begin{aligned}
|\mu_{n+1} - \mu_\pi|^2 &= |p_3\left(\mu_n + \gamma_n(X_{n+1} - \mu_n)\right) - \mu_\pi|^2 \\
&\leq |\mu_n - \mu_\pi + \gamma_n(X_{n+1} - \mu_n)|^2 \\
&\leq |\mu_n - \mu_\pi|^2 - 2\gamma_n|\mu_n - \mu_\pi|^2 + K\gamma_n^2 V^{1/2}(X_{n+1}) + 2\gamma_n(\mu_n - \mu_\pi) \cdot (X_{n+1} - \mu_\pi),
\end{aligned}
$$

K constant. Therefore writing $U_n = V_n = |\mu_n - \mu_\pi|^2$ and $W_n = KR_1\gamma_n^2 V^{1/2}(X_n) + 2\gamma_n(\mu_n - \mu_\pi) \cdot (P_{\sigma_n,\Lambda_n} e_1(X_n) - \mu_\pi)$ we get:

$$\mathbb{E}\left(U_{n+1}|\mathcal{F}_n\right) \leq U_n - 2V_n + W_n. \tag{4.24}$$

From Lemma 4.6, $\sum W_n < \infty$ a.s. and we can apply Lemma 4.7 to obtain that $\sum V_n < \infty$ which implies that $\mu_n \to \mu_\pi$.

(ii) Similarly, we have:

$$
\begin{aligned}
|\Gamma_{n+1} - \Sigma_\pi|^2 &\leq |\Gamma_n - \Sigma_\pi|^2 - 2\gamma_n|\Gamma_n - \Sigma_\pi|^2 + K\gamma_n^2 V^{1/2}(X_{n+1}) \\
&\quad + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((X_{n+1} - \mu_\pi)(X_{n+1} - \mu_\pi)' - \Sigma_\pi) \\
&\quad + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((X_{n+1} - \mu_\pi)(\mu_\pi - \mu_n)') \\
&\quad + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((\mu_n - \mu_\pi)(X_{n+1} - \mu_n)').
\end{aligned}
$$

Write $U_n = V_n = |\Gamma_n - \Sigma_\pi|^2$ and

$$
\begin{aligned}
W_n &= R_1\gamma_n^2 V^{1/2}(X_n) + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot (P_{\sigma_n,\Lambda_n} e_2(X_n) - \Sigma_\pi) \\
&\quad + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((P_{\sigma_n,\Lambda_n} e_1(X_n) - \mu_\pi)(\mu_\pi - \mu_n)') \\
&\quad + 2\gamma_n(\Gamma_n - \Sigma_\pi) \cdot ((\mu_n - \mu_\pi)(P_{\sigma_n,\Lambda_n} e_1(X_n) - \mu_\pi))'.
\end{aligned}
$$

We get:

$$\mathbb{E}\left(U_{n+1}|\mathcal{F}_n\right) \leq U_n - 2V_n + W_n. \tag{4.25}$$

It is easily seen from Lemma 4.6 that $\sum W_n < \infty$ and (ii) follows from Lemma 4.7.

(iii) We have:

$$
\begin{aligned}
|\sigma_{n+1} - \sigma_{opt}|^2 &\leq |\sigma_n - \sigma_{opt}|^2 + 2\gamma_n(\sigma_n - \sigma_{opt})(\tau(\sigma_n, \Lambda_n) - \bar{\tau}) \\
&\quad + \gamma_n^2 + 2\gamma_n(\sigma_n - \sigma_{opt})\left(\alpha_{\sigma_n,\Lambda}(X_n, Y_{n+1}) - \tau(\sigma_n, \Lambda_n)\right).
\end{aligned}
$$

Therefore:

$$\mathbb{E}(U_{n+1}|\mathcal{F}_n) \leq U_n - 2\gamma_n V_n + W_n, \tag{4.26}$$

where $U_n = |\sigma_n - \sigma_{opt}|^2$, $V_n = -(\sigma_n - \sigma_{opt})(\tau(\sigma_n, \Lambda_n) - \bar{\tau})$ and $W_n = \gamma_n^2 + 2\gamma_n(\sigma_n - \sigma_{opt})\left(A(X_n, \sigma_n, \Lambda_n) - \tau(\sigma_n, \Lambda_n)\right)$. From Lemma 4.6, $\sum \gamma_n(\sigma_n - \sigma_{opt})\left(A(\sigma_n, \Lambda_n, X_n) - \tau(\sigma_n, \Lambda_n)\right) < \infty$ and since $\Lambda_n \to \Sigma_\pi + \varepsilon I_d$, it follows from point (ii) of (A2) that $V_n \geq 0$ for n sufficiently large. From Lemma 4.7 we conclude that σ_n converges to a finite random variable and $\sum \gamma_n V_n$ is finite a.s. which implies (iii) since if σ_n converges to a limit that is not σ_{opt}, $\sum \gamma_n V_n = \infty$ because of (A2)(ii) and (A3), leading to a contradiction.

$\square$

4.4 Some useful technical lemmas

Lemma 4.7 (Robbins-Siegmund Theorem). *Let $(U_n)_{n\geq 0}$, $(V_n)_{n\geq 0}$ and $(W_n)_{n\geq 0}$ be three random processes defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted with respect to a filtration $(\mathcal{F}_n)$ such that:*

(i) *$U_n \geq 0$, $\sum W_n < \infty$ and for $\mathbb{P}$-almost any every $\omega \in \Omega$ there exists $n_0(\omega)$ such that $V_n(\omega) \geq 0$, $n \geq n_0(\omega)$,*

(ii) *$\mathbb{E}(U_{n+1}|\mathcal{F}_n) \leq U_n - V_n + W_n$.*

Then U_n and $\sum V_n$ converge almost surely to finite random variables.

Proof. Let $Y_n = U_n - \sum_{i=0}^{n-1}(W_i - V_i)$, $n \geq 1$. Then (Y_n) is a supermartingale. For a positive integer $N > 0$ define $\mathcal{S}_N = \{\omega : \sum_{i=0}^n (V_i - W_i) < N, \ n \geq 0\}$. Then on $\mathcal{S}_N$, (Y_n) is a supermartingale bounded from below (by $-N$) therefore converges a.s. to a finite random variable. We have $\sum_{i=0}^{n-1} V_i = Y_n - U_n + \sum_{i=0}^{n-1} W_i \leq Y_n + \sum_{i=0}^{n-1} W_i$. Since $\sum W_n < \infty$ and V_n is nonegative for n sufficiently large, $\sum V_n < \infty$ on $\mathcal{S}_N$ and U_n also converges on $\mathcal{S}_N$ to a finite random variable.

Let $\omega \in \Omega$ be such that $\sum W_n(\omega) < \infty$ and $V_n(\omega) \geq 0$ for $n \geq n_0(\omega)$. $\sum_{i=0}^n (W_i(\omega) - V_i(\omega)) \leq \sup_n \sum_{i=0}^n W_i(\omega) + \sum_{i=0}^{n_0(\omega)} |V_i(\omega)| < \infty$. Then taking $N > \sup_n \sum_{i=0}^n W_i(\omega) + \sum_{i=0}^{n_0(\omega)} |V_i(\omega)|$, we get $\omega \in \mathcal{S}_N$. In conclusion $\Omega = \cup \mathcal{S}_N$ and we are done. $\qquad \square$

Lemma 4.8. *Let $(X_n)_{n\geq 0}$ be a random sequence on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ adapted with respect to a nondecreasing filtration $(\mathcal{F}_n)$. Assume that there exist constants $K_1, K_2 < \infty$, $0 < \rho < 1$, a sequence of positive numbers $\gamma_n = \mathcal{O}\left(n^{-\lambda}\right)$, $\lambda \in (\frac{1}{2}, 1]$ and an adapted positive random sequence V_n such that $|\mathbb{E}(X_{n+k}|\mathcal{F}_n)| \leq K_1(\rho^n + k\gamma_n)V_n$, $\sup_n \mathbb{E}(V_n^2) < \infty$ and $\mathbb{E}(V_{n+k}|\mathcal{F}_n) \leq K_2 V_n$.*

Then there exists a constant $K < \infty$ (that depends only on K_1, K_2, ρ and (γ_n)) such that

$$|\mathbb{E}(X_{n+k}|\mathcal{F}_n)| \leq K\gamma_k \log(k)V_n, \quad \mathbb{P}-a.s. \tag{4.27}$$

$$\frac{1}{n}\sum_{k=1}^n X_k \to 0, \ a.s., \ as \ n \to \infty, \tag{4.28}$$

and

$$\sum \gamma_n X_n \quad converges \ a.s. \ to \ a \ finite \ random \ variable. \tag{4.29}$$

Proof. Since $(\mathcal{F}_n)$ is nondecreasing, for $n, k \geq 0$, $0 \leq j \leq k$:

$$
\begin{aligned}
|\mathbb{E}(X_{n+k}|\mathcal{F}_n)| &= |\mathbb{E}[\mathbb{E}(X_{n+k}|\mathcal{F}_{n+j})|\mathcal{F}_n]| \\
&\leq K_1\left(\rho^{k-j} + (k-j)\gamma_{n+j}\right)\mathbb{E}(V_{n+j}|\mathcal{F}_n) \\
&\leq K_1 K_2\left(\rho^{k-j} + (k-j)\gamma_{n+j}\right)V_n.
\end{aligned}
$$

Therefore, $|\mathbb{E}(X_{n+k}|\mathcal{F}_n)| \leq \min_{0\leq j\leq k} K_1 K_2\left(\rho^{k-j} + (k-j)\gamma_{n+j}\right)V_n \leq K_3\gamma_k \log k V_n$.

Define $Y_n = X_n - \mathbb{E}(X_n)$ and $\mathcal{F}_n = \{\emptyset, \Omega\}$ if $n < 0$. Then it is easily seen that $(Y_n, \mathcal{F}_n)$ is a mixingale with mixingales sequences $c_n \equiv const.$ and $\rho_n = \gamma_n \log(2 + n)$. We apply Corollary 2.1 of Davidson and de Jong (1997) to obtain that $\frac{1}{n}\sum_{i=1}^n Y_i \to 0$ a.s. But since $|\mathbb{E}(X_n)| \leq K_3\gamma_n \log(2 + n)\mathbb{E}(V_0) \to 0$ as $n \to \infty$, (i) follows.

Similarly, $(\gamma_n Y_n, \mathcal{F}_n)$ is a mixingale with mixingale sequence $c_n \propto \gamma_n$ and $\rho_n = \gamma_n \log(2 + n)$. From Theorem 2.7 of Hall and Heyde (1980), we have $\sum \gamma_n Y_n$ converges a.s. to a finite random variable. (ii) follows since $\sum \gamma_n \mathbb{E}(X_n)$ is a convergent series. $\qquad \square$

References

ANDRIEU, C. and MOULINES, E. (2003). Ergodicity of some adaptive markov chain monte carlo algorithm. *Technical Report* .

ANDRIEU, C. and ROBERT, C. P. (2002). Controlled mcmc for optimal sampling. *Technical report* .

ATCHADE, Y. F. and ROSENTHAL, J. S. (2003). On adaptive markov chain monte carlo algorithm. *Technical Report* .

BREYER, L., PICCIONI, M. and S, S. (2004). Optimal scaling of mala for nonlinear regression. *Annals of Applied Probability* **14** 1479–1505.

DAVIDSON, J. and DE JONG, R. (1997). Strong laws of large numbers for dependent heteregeneous processes: a synthesis of recent and new results. *Econometric Reviews* **16** 251–279.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7** 223–242.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit theory and its application.* Academic Press, New York.

JARNER, S. F. and HANSEN, E. (2000). Geometric ergodicity of metropolis algorithms. *Sto. Proc. Appl.* **85** 341–361.

KUSHNER, K. and YIN, Y. (2003). *Stochastic approximation and recursive algorithms and applications.* Springer, Springer-Verlag, New-York.

MEYN, S. P. and TWEEDIE, R. L. (1994). Computable bounds for convergence rates of markov chains. *Ann. Appl. Prob.* **4** 981–1011.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407.

ROBERTS, G. and TWEEDIE, R. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363.

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling of various metropolis-hastings algorithms. *Statistical Science* **16** 351–367.