Methodol Comput Appl Probab (2006) 8: 235–254
DOI 10.1007/s11009-006-8550-0

An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift

Yves F. Atchadé

Received: 7 April 2005 / Revised: 28 December 2005 / Accepted: 18 January 2006

Abstract This paper extends some adaptive schemes that have been developed for the Random Walk Metropolis algorithm to more general versions of the Metropolis-Hastings (MH) algorithm, particularly to the Metropolis Adjusted Langevin algorithm of Roberts and Tweedie (1996). Our simulations show that the adaptation drastically improves the performance of such MH algorithms. We study the convergence of the algorithm. Our proves are based on a new approach to the analysis of stochastic approximation algorithms based on mixingales theory.

Keywords Adaptive Markov Chain Monte Carlo · Langevin algorithms · Metropolis-Hastings algorithms · Stochastic approximation algorithms

AMS 2000 Subject Classification 65C05 · 65C40 · 60J27 · 60J35

1 Introduction

Markov Chain Monte Carlo (MCMC) is a well-established probabilistic tool to sample from probability measures. A MCMC algorithm is designed by specifying a transition kernel with a predefined invariant probability measure (the target distribution). Such transition kernel typically depends on various parameters to be provided by the user. Finding the optimal values of the parameters for a given target distribution is a difficult analytical problem. As a consequence, many fine-tunings of the parameters are often necessary in practice to obtain a satisfactory implementation of a MCMC algorithm. Adaptive MCMC proposes an elegant solution to the parameter-tuning problem where the parameters (or some of the parameters) are automatically handled by the algorithm. For a general introduction to MCMC methods, see e.g., Tierney (1994). General ideas and convergence analysis of adaptive MCMC algorithms can be found in Gilks et al. (1998), Haario et al. (2001), Andrieu and Moulines (2005), Atchade and Rosenthal (2005).

Y. F. Atchadé (✉)
Department of Mathematics and Statistics, University of Ottawa,
585 King Edward St., Ottawa, ON K1N 6N5, Canada
e-mail: yatchade@uottawa.ca

Most of the existing adaptive MCMC strategies have been developed for the Independence Sampler and the Random Walk Metropolis (RWM) algorithm. In this paper, we extend some of these adaptive schemes to a more general class of Metropolis-Hastings (MH) algorithms. We consider the Metropolis Adjusted Langevin (MALA) algorithm (Roberts and Tweedie, 1996); or more generally, MH algorithms with a drift. Langevin-based MH algorithms are known to mix faster than the RWM algorithm (see e.g., Roberts and Rosenthal (2001), Breyer et al. (2002). But adaptive schemes that could facilitate their implementation by handling automatically the scaling of the algorithm are lacking. The present paper tries to fill that gap.

Our adaptive scheme is a stochastic approximation algorithm that recursively and *simultaneously* tunes the covariance matrix Λ and the scale parameter σ of the proposal kernel. The covariance matrix is tuned towards Σ_π, the covariance matrix of the target distribution in a way similar to Haario et al. (2001), Andrieu and Moulines (2005). The scale parameter is tuned as in Atchade and Rosenthal (2005) so as to achieve a prescribed acceptance rate in stationarity (approximately 0.574 for Langevin-based algorithms). Our algorithm is easy to implement and performs extremely well. Our simulations results show that the adaptive algorithm and the optimal MH algorithm (the MH algorithm where $\Lambda = \Sigma_\pi$ and σ is chosen so as to reach the acceptance rate of 0.574, say) have about the same efficiency and they both considerably outperform algorithms where the parameters are not adapted. We develop a bound on the convergence rate to stationarity and a strong law of large numbers for the adaptive algorithm. Under some additional conditions, we show that the adaptation parameters also converge to the appropriate limits. In order to study the adaptive algorithm, we derive some new results on the rate of convergence of some classes of (nonadaptive) Metropolis-Hastings algorithms. Using essentially the same argument as Jarner and Hansen (2000), we show in this paper that Metropolis-Hastings algorithms with bounded drift are geometrically ergodic when the target distribution is super-exponential with regular contours (Assumption (A1)).

Another contribution of this paper is about the convergence of stochastic approximation algorithms with Markovian dynamics. The classical approach to these limit results as initiated by Metivier and Priouret (1984) is based on the Poisson equation, see also Benveniste et al. (1990). Here, we show that the noise process of such stochastic approximation algorithm, properly centered, is a mixingale difference (see e.g., Hall and Heyde (1980) for an introduction to mixingales). This result permits a simpler analysis of adaptive chains and stochastic approximation processes. We use it to study the asymptotics of the adaptive MCMC algorithm discussed earlier. The idea has already been used in Atchade and Rosenthal (2005), although in a specific case.

The rest of the paper is organized as follows. The adaptive MH algorithm is proposed and discussed in Section 2. The convergence results on the algorithm are stated in Section 2 but the proofs are postponed to Section 5. We give some simulation examples in Section 3 to illustrate the algorithm. In Section 4, we develop a general convergence result on adaptive Markov chains and stochastic approximation algorithms. Some technical results are detailed in Section 6.

2 Adaptive Metropolis-Hastings Algorithms with Bounded Drift

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 $\mathcal{X}$) on $\mathcal{X}$. Let $D:\ \mathcal{X}\to\mathcal{X}$ be a bounded function. D is the drift function of the algorithm. To avoid a possible degeneracy in the rate of convergence of the algorithm, we restrict our analysis to algorithms with a bounded drift. But in practice, it is straightforward to truncate a drift function to obtain a bounded drift. All our theoretical results hold for any bounded drift function D but for the applications, we have a particular drift in mind, namely $D(x)=D_{MALA}(x)$ where:

$$D_{MALA}(x)=\frac{\delta}{\max(\delta,|\nabla\log\pi(x)|)}\nabla\log\pi(x), \tag{2.1}$$

where ∇ is the gradient operator, and $\delta>0$ is a fixed constant. This corresponds to the Truncated Metropolis Adjusted Langevin Algorithm (T-MALA) as proposed by Roberts and Tweedie (1996). Other choices are possible, for example a truncated version of the "self-targeting drift" proposed by Stramer and Tweedie (1999) could also be used. The RWM algorithm is also a special case with $D\equiv 0$. MH algorithms with drift like (2.1) mix faster than plain RWM algorithm ($D\equiv 0$) because typically, the drift moves the algorithm faster towards the "center" of the target distribution.

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 $\mathcal{N}\left(x+\frac{\sigma^2}{2}\Lambda D(x),\sigma^2\Lambda\right)$ the Gaussian distribution with mean $x+\frac{\sigma^2}{2}\Lambda D(x)$ and covariance matrix $\sigma^2\Lambda$. The Metropolis-Hastings algorithm with drift function D, and proposal density $Q_{\sigma,\Lambda}(x,dy)=q_{\sigma,\Lambda}(x,y)dy$ generates a Markov chain (X_n) with invariant distribution π as follows. Given X_n, a new proposal $Y_{n+1}\sim\mathcal{N}\left(X_n+\frac{\sigma^2}{2}\Lambda D(X_n),\sigma^2\Lambda\right)$ 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:

$$P_{\sigma,\Lambda}(x,A)=\int_A\alpha_{\sigma,\Lambda}(x,y)q_{\sigma,\Lambda}(x,y)dy+r_{\sigma,\Lambda}(x)\mathbf{1}_A(x),\ x\in\mathcal{X},A\in\mathcal{B}^d. \tag{2.2}$$

where

$$r_{\sigma,\Lambda}(x)=\int\left(1-\alpha_{\sigma,\Lambda}(x,y)\right)q_{\sigma,\Lambda}(x,y)dy, \tag{2.3}$$

and $\mathbf{1}_A$ is the indicator function of the set A.

In a practical implementation of this algorithm, we need to specify σ and Λ. It is well-documented in the MCMC literature that bad choice of these parameters can greatly impact the efficiency of the algorithm. How to choose the best parameter values is what we called the parameter tuning problem. It is a difficult problem unless π is well known which is rarely the case in practice. Adaptive MCMC solves this problem by running a second process (σ_n,Λ_n) that (hopefully) will converge to the best values of (σ,Λ).

2.1 An Adaptive Version of the MH with Bounded Drift

Let constants $\varepsilon_1,\varepsilon_2,A_1$ be given such that $0<\varepsilon_1<A_1<\infty$ and $\varepsilon_2>0$. Write $\Theta_\sigma=[\varepsilon_1,A_1]$ equipped with the Euclidean norm of $\mathbb{R}$. Let $B_d(0,r)$ be the ball of center 0 and radius r in $\mathbb{R}^d$. Let Θ_Γ be the convex set of all semipositive definite matrices Γ

with $|\Gamma| \leq A_1$, where for a matrix $\Gamma = (\Gamma_{i,j})$, we define $|\Gamma| := tr^{1/2}(\Gamma\Gamma') = \left\{\sum_{i,j} |\Gamma_{ij}|^2\right\}^{1/2}$, the Frobenius norm of Γ. This norm is derived from the scalar product $A \cdot B := tr(AB') := \left\{\sum_{i,j} A_{ij}B_{ij}\right\}^{1/2}$. Note that we use the same notation $|\cdot|$ to denote the norms in $\mathbb{R}^d$ and Θ_Γ. More generally we use $|\cdot|$ (resp. $\langle\cdot\rangle$) to denote the Euclidean norm (resp. inner product) in any Euclidean space. Define the set $\Theta := B_d(0, A_1) \times \Theta_\Gamma \times \Theta_\sigma$. The general approach to define our adaptive MCMC is to supplement the $\mathcal{X}$-valued process (X_n) with a Θ-valued process $(\mu_n, \Gamma_n, \sigma_n)$ that takes care of the parameter tuning problem. We refer to $(\mu_n, \Gamma_n, \sigma_n)$ as the adaptation process. Next, we introduce three projection functions p_1, p_2, p_3 that we use to contain the adaptation process inside Θ. For $\sigma \in R$, $p_1(\sigma)$ is equal to σ when $\sigma \in \Theta_\sigma$. When $\sigma < \varepsilon_1$, $p_1(\sigma) = \varepsilon_1$ and $p_1(\sigma) = A_1$ for $\sigma > A_1$. Clearly for any σ_1, $p(\sigma_1)$ is the closest point to σ in Θ_σ and we have:

$$|\sigma' - p_1(\sigma)| \leq |\sigma' - \sigma|, \ \sigma' \in \Theta_\sigma, \ \sigma \in \mathbb{R}. \tag{2.4}$$

Similarly, for a semidefinite positive matrix Σ, let $p_2(\Sigma)$ be the closest point to Σ in the convex compact cone Θ_Γ. We have $p_2(\Sigma) = \Sigma$ if $|\Sigma| \leq A_1$ and $p_2(\Sigma) = \frac{A_1}{|\Sigma|}\Sigma$ if $|\Sigma| > A_1$. Also p_2 satisfies:

$$|\Gamma' - p_2(\Gamma)| \leq |\Gamma' - \Gamma|, \ \Gamma' \in \Theta_\Gamma, \ \Gamma \tag{2.5}$$

For any $\mu \in \mathbb{R}^d$, let $p_3(\mu)$ be the closest point to μ in $B(0, A_1)$. We have, $p_3(x) = x$ if $|x| \leq A_1$ and $p_3(x) = \frac{A_1}{|x|}x$ if $|x| > A_1$ and p_3 satisfies:

$$|\mu' - p_3(\mu)| \leq |\mu' - \mu|, \ |\mu'| \leq A_1, \ \mu \in \mathbb{R}^d. \tag{2.6}$$

Let (γ_n) a sequence of positive numbers and $\overline{\tau}$ the "optimal" acceptance rate. We discuss the choice of (γ_n) and $\overline{\tau}$ in the next remark.

Algorithm 2.1 [Adaptive MH]

1. Start the algorithm at some point $x_0 \in \mathcal{X}$, with $(\mu_0, \Gamma_0, \sigma_0) \in B(0, A_1) \times \Theta_\Gamma \times \Theta_\sigma$.
2. Suppose that at time $n \geq 0$, we have $X_n \in \mathcal{X}$ and $(\mu_n, \Gamma_n, \sigma_n) \in B(0, A_1) \times \Theta_\Gamma \times \Theta_\sigma$. Set $\Lambda_n = \Gamma_n + \varepsilon_2 I_d$.
   - 2.1 Generate $Y_{n+1} \sim \mathcal{N}\left(X_n + \frac{\sigma_n^2}{2}\Lambda_n D(X_n), \sigma_n^2\Lambda_n\right)$ 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

$$\mu_{n+1} = p_3(\mu_n + \gamma_n(X_{n+1} - \mu_n)), \tag{2.7}$$

$$\Gamma_{n+1} = p_2\big(\Gamma_n + \gamma_n\big((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)' - \Gamma_n\big)\big), \tag{2.8}$$

$$\sigma_{n+1} = p_1\big(\sigma_n + \gamma_n\big(\alpha_{\sigma_n,\Lambda_n}(X_n, Y_{n+1}) - \overline{\tau}\big)\big). \tag{2.9}$$

Remark 2.1

1. At each step of the algorithm, a valid MH algorithm is used with parameters $(\sigma_n, \mu_n, \Gamma_n)$. But the parameters are recursively changed from one iteration to

another. Therefore (X_n) is no longer a Markov chain and there is, a priori, no guarantee that the distribution of X_n will converge to π. This is an example of so-called adaptive MCMC algorithms. In this paper, we show that our algorithm is indeed ergodic with stationary distribution π and also satisfies a law of large numbers (see Theorem 2.1 below). We refer the reader to Atchade and Rosenthal (2005), Andrieu and Moulines (2005), Rosenthal and Roberts (2005) for more general results on adaptive MCMC.

2. When no re-projection on the ball $B(0, A_1) \subset \mathbb{R}^d$ is inforced by time n, μ_n is nothing but the empirical mean of the sample $(X_0, \ldots, X_n)$ generated by the algorithm and we expect $\mu_n \to \int x\pi(dx) = \mu_\pi$ as $n \to \infty$. Similarly, Γ_n is approximately the empirical covariance of the sample $(X_0, \ldots, X_n)$ and we expect to have $\Gamma_n \to \left(\int xx'\pi(dx) - \mu_\pi\right)\left(\int xx'\pi(dx) - \mu_\pi\right)' = \Sigma_\pi$.
3. A similar intuition applies in the recursion on σ_n. The quantity $\alpha_{\sigma_n,\Lambda_n}(X_n, Y_{n+1})$ is an estimate of the acceptance rate in the algorithm. Some recent theoretical results (see e.g., Roberts and Rosenthal (2001)) have suggested that the best performance is obtained from the MALA algorithm when the acceptance rate is about 40–60% (about 20–30% for the RWM algorithm). Therefore in Algorithm 2.1, $\overline{\tau}$ represents this target "optimal" acceptance rate. When $\alpha_{\sigma_n,\Lambda_n}(X_n, Y_{n+1}) > \overline{\tau}$, σ_n is increased; and decreased otherwise. Therefore we should expect $\sigma_n \to \sigma_{opt}$, where σ_{opt} satisfies $\tau(\sigma_{opt}) = \overline{\tau}$, where $\tau(\sigma) = \int \pi(dx) \int \alpha_{\sigma,\Lambda_\pi}(x, y) q_{\sigma,\Lambda_\pi}(x, y) dy$ is the acceptance rate of the algorithm in stationarity and $\Lambda_\pi = \Sigma_\pi + \varepsilon_2 I_d$.
4. In Algorithm 2.1, 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.
5. The algorithm is not particularly sensible to the choice of δ, A_1, ε_1 and ε_2 as long as δ and A_1 are sufficiently large and ε_1 and ε_2 are sufficiently small. It may be safe to set ε_1 and ε_2 to some extremely small values and A_1 to some extremely large value. In the simulations below $\delta = 1000$ works very well. See the examples for some numerical values. Note that mispecifying these variables do not disturb the ergodicity of the algorithm.

2.2 Ergodicity of the Adaptive MH Algorithm

We make the following assumptions.

Assumption A1 *The density 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 exist* $\delta > 0$, $\sigma_{opt} \in \Theta_\sigma$ *such that* $\tau(\sigma_{opt}) = \overline{\tau}$ *and* $(\sigma - \sigma_{opt})(\tau(\sigma) - \overline{\tau}) < -\delta|\sigma - \sigma_{opt}|^2$, *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 *The sequence* (γ_n) *is such that* $\gamma_n > 0$, $\sum \gamma_n = \infty$ *and* $\gamma_n = O(n^{-\lambda})$, $1/2 < \lambda \leq 1$.

Remark 2.2

1. (A1) has been introduced in Jarner and Hansen (2000) to analyze the convergence rate of the RWM algorithm. These authors have shown that under (A1), the RWM algorithm is geometrically ergodic. We show a similar result in Proposition 2.1 for truncated drift Metropolis-Hastings algorithms with normally distributed proposal. Many densities of the form $e^{-p(x)}$ or $h(x)^{-p(x)}$ are known to satisfy (A1). See Jarner and Hansen (2000) for more details.
2. It is always possible to choose A_1 such that (A2)(i) hold, at least in theory. On the other hand, (A2)(ii) is difficult to check and actually may not hold. But we believe that σ_n can still converge to a solution of $\tau(\sigma) = \overline{\tau}$ even if τ is not decreasing and $\tau(\sigma) = \overline{\tau}$ has many solutions. In any case, it is worth noting that the ergodicity of the algorithm does not rely on (A2).
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 triplet* $(\Omega, \mathcal{F}, \Pr)$. *Define* $V(x) = c\pi^{1/4}(x)$ *where* c *is chosen such that* $V \geq 1$.

(i) *Assume (A1) and (A3). Then*

$$\left\|\mathcal{L}(X_n) - \pi\right\|_V = O\left(\frac{\log(n+1)}{n^\lambda}\right),\ n \geq 1 \tag{2.10}$$

where $\mathcal{L}(X_n)$ *is the distribution of* X_n *and for a signed measure* μ, $\|\mu\|_V := \sup_{|f|\leq V} |\mu(f)|$, $\mu(f) := \int f(x)\mu(dx)$.
Also, for any measurable function $f: \mathcal{X} \longrightarrow \mathbb{R}$ *with* $|f| \leq V$,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \pi(f) \text{ as } n \to \infty,\ \Pr - a.s. \tag{2.11}$$

(ii) *Assume (A1)–(A3). Then* $E\left[|\Lambda_n - \Lambda_\pi|^2\right] \longrightarrow 0$ *and* $\mathbb{E}\left[|\sigma_n - \sigma_{opt}|^2\right] \longrightarrow 0$ *as* $n \to \infty$, *where* $\Lambda_\pi = \Sigma_\pi + \varepsilon_2 I_d$.

Proof: See Section 5. ■

This theorem shows that as $n \to \infty$, the distribution of X_n converges in V-norm to π and that an estimate of $\pi(f)$ can be obtained by taking the empirical mean

$\frac{1}{n}\sum_{i=1}^{n} f(X_i)$. In (i), it is shown that the rate of convergence of the distribution of X_n to π is at least $\log(n)/n$ (assuming $\lambda = 1$). This bound is better than $\log(n)^2/n$ obtained by Atchade and Rosenthal (2005) but we suspect the true rate to be $1/n$. This rate may seem very slow compared to the geometric rate ρ^n typically enjoyed by Markov chains (see Proposition 2.1). But such comparison can be misleading unless we know ρ and the constants involved in the rates. In practice, it turns out that adaptive MCMC algorithms perform better than nonadaptive algorithms unless highly tuned.

2.3 Geometric Ergodicity of Metropolis-Hastings Algorithms with Bounded Drift

For the proof of Theorem 2.1 we need the rate of convergence of the (nonadaptive) Metropolis-Hastings transition kernels used in Algorithm 2.1. To that end, we show here that Theorem 4.3 of Jarner and Hansen (2000) on the geometric ergodicity of RMM algorithms is actually robust to the presence of a bounded drift. More precisely (see Proposition 2.1 below), we show that under (A1), a (nonadaptive) MH algorithm with a truncated drift and transition kernel (2.2) is geometrically ergodic. This result was already anticipated by Roberts and Tweedie (1996). The proof is a technical modification of the proof of Jarner and Hansen (2000). We also show in Proposition 2.2 below that the transition kernel of the (nonadaptive) MH algorithm is a smooth function of its parameters.

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 $(\sigma, \Lambda) \in \mathcal{C}$, 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* $V_\alpha(x) = c_\alpha \pi^{-\alpha}(x)$ *where* c_α *is such that* $V_\alpha \geq 1$. *There exist a set* $C \subset \mathcal{X}$, *a probability measure* ν *such that* $\nu(C) > 0$ *and constants* $\lambda_\alpha \in (0,1)$, $b_\alpha \in [0, \infty)$, $\varepsilon \in (0,1]$ *such that:*

$$\inf_{(\sigma,\Lambda)\in\mathcal{C}} P_{(\sigma,\Lambda)}(x, A) \geq \varepsilon \nu(A) \mathbf{1}_C(x), \; x \in \mathcal{X}, \; \mathcal{A} \in \mathcal{B}, \tag{2.12}$$

and

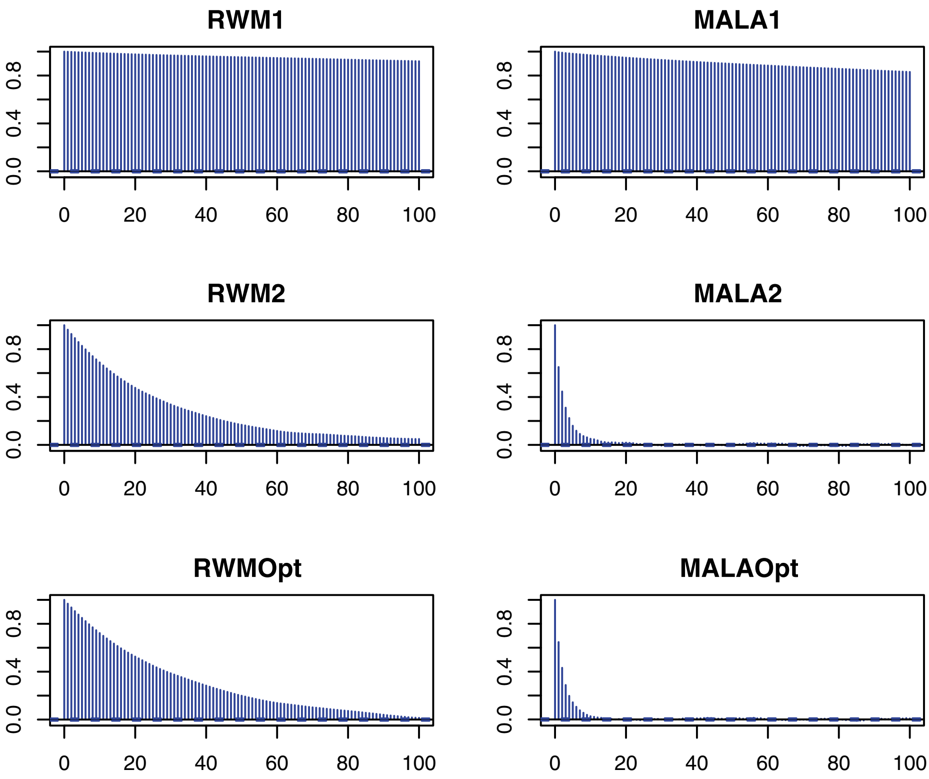
$$\sup_{(\sigma,\Lambda)\in\mathcal{C}} P_{(\sigma,\Lambda)} V_\alpha(x) \leq \lambda_\alpha V_\alpha(x) + b_\alpha \mathbf{1}_C(x) \; . \tag{2.13}$$

Proof: See Section 6. ■

A well known consequence of Proposition 2.1 is that the family of transition kernel $P_{(\sigma,\Lambda)}$ is (uniformly in (σ, Λ)) geometrically ergodic. That is: there exist $\rho < 1$, $R < \infty$ such that:

$$\sup_{(\sigma,\Lambda)\in\mathcal{C}} \left\| P^n_{\sigma,\Lambda}(x, \cdot) - \pi(\cdot) \right\|_{V_\alpha} \leq R\rho^n V_\alpha(x), \;\; n \geq 0, \; x \in \mathcal{X}. \tag{2.14}$$

See e.g., Baxendale (2005).



Graph 1 Autocorrelation functions for the different samplers for the first component of the Gaussian distribution

We can also prove that $P_{\sigma,\Lambda}f$ is a smooth function of (σ,Λ). For (σ_1,Λ_1) and (σ_2,Λ_2) elements of $\mathcal{C}$, define $\left|\!\left|\!\left| P_{(\sigma_1,\Lambda_1)} - P_{(\sigma_2,\Lambda_2)} \right|\!\right|\!\right|_{V_\alpha} := \sup_{x\in\mathcal{X}} \sup_{|f|\leq V_\alpha} \frac{\left|P_{\sigma_1,\Lambda_1}f(x) - P_{\sigma_2,\Lambda_2}f(x)\right|}{V_\alpha(x)}$.

Proposition 2.2 *Assume (A1). For* $0 < \alpha < 1$ *write* $V_\alpha(x) = c_\alpha \pi^{-\alpha}(x)$ *where* c_α *is such that* $V_\alpha \geq 1$. *Under (A1), there is a constant* $K_1 = K_1(\alpha) < \infty$ *such that for* $(\sigma_1,\Lambda_1),(\sigma_2,\Lambda_2) \in \mathcal{C}$*:*

$$\left|\!\left|\!\left| P_{(\sigma_1,\Lambda_1)} - P_{(\sigma_2,\Lambda_2)} \right|\!\right|\!\right|_{V_\alpha} \leq K_1 |(\sigma_2 - \sigma_1, \Lambda_2 - \Lambda_1)|.$$

Proof: See Section 6. ■

Table 1 Simulation results for the Gaussian example. First row: Estimates of the mean square jump in stationarity. Second row: estimation of the mean of the first component, standard errors in parenthesis. Third row: statistical efficiency (relatively to *RWM1*) in estimating μ_1

	RWM1	*RWM2*	*RWMOpt*	*MALA1*	*MALA2*	*MALAOpt*
$d = E^{1/2}\left[\lvert X_n - X_{n-1}\rvert^2\right]$	0.24	1.51	1.74	0.91	8.18	9.11
Estimation of μ_1	1.51 (1.78)	0.19 (0.17)	0.005 (0.15)	0.08 (1.22)	0.02 (0.04)	−0.008 (0.03)
Efficiency	1	10.4	12.2	1.5	47.3	56.3

Table 2 Failures and times of observation for 10 nuclear pump. See Robert and Casella (2004) for more details

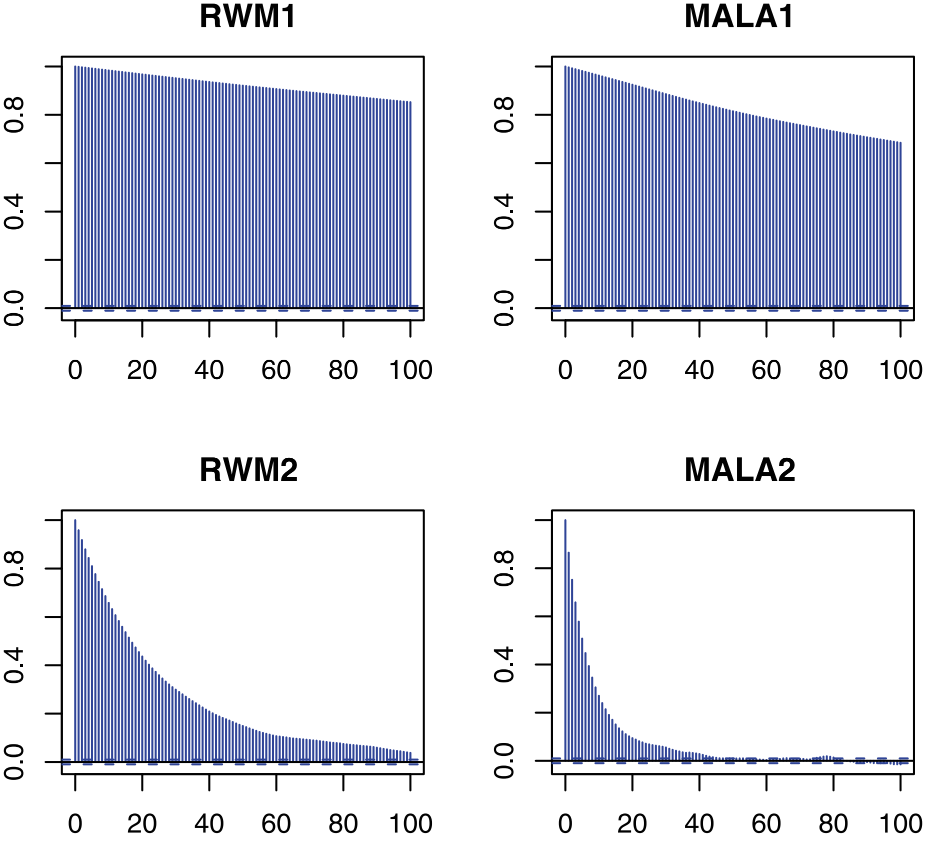
Pump	1	2	3	4	5	6	7	8	9	10
Failures (p_i)	5	1	5	14	3	19	1	1	4	22
Time (t_i)	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

3 Simulation Examples

We illustrate Algorithm 2.1 with two simulation examples.

3.1 Sampling from a 20-dimensional Gaussian Distribution

We take π to be the 20-dimensional normal distribution with mean 0 and covariance matrix Σ_π. The entries of this covariance matrix can be obtained from the supplementary file www.mathstat.uottawa.ca/∼yatch436/tmalaexcov.txt. We design Σ_π so that many of the components of the distribution are highly correlated. We compare the performances of 6 samplers in sampling from π. There are three



Graph 2 Autocorrelation functions of the samplers for the tenth component of the posterior distribution of the nuclear pump example

Table 3 Estimates of the mean square jump in stationarity for the nuclear pumps example

RWM1	*RWM2*	*MALA1*	*MALA2*
0.03	0.14	0.07	0.41

Random Walk based samplers and three Langevin based samplers. The algorithm *RWM1* corresponds to Algorithm 2.1 where $D(x) \equiv 0$ and $\Lambda_n \equiv I_{20}$, the identity matrix. So, *RWM1* is an adaptive Random Walk Metropolis amgorithm where we do not adapt the covariance matrix. The algorithm *RWM2* corresponds to the full adaptive algorithm where $D(x) \equiv 0$. The algorithm *RWMOpt* is the nonadaptive RWM algorithm where $\Lambda = \Sigma_\pi$ and $\sigma = 0.59$ the optimal value (one that gives an acceptance rate of 0.2; estimated from the adaptive algorithm *RWM2*). The Langevin based algorithms *MALA1*, *MALA2* and *MALAOpt* are defined similarly except that the drift function is $D(x) = \frac{\delta}{\max(\delta, |\nabla \log \pi(x)|)} \nabla \log \pi(x)$. For *MALAOpt*, we use $\sigma = 1.06$ (obtained from *MALA2*).

All the simulations are run for $n = 50,000$ iterations started from $X_0 = (5, 5, 5)$. For the Random Walk based algorithms, we set the target acceptance rate to $\overline{\tau}= 0.2$ and use $\overline{\tau} = 0.5$ for Langevin based algorithms. The drift is bounded by $\delta = 1,000$, an arbitrary large value. To bound the adaptation process, we use $\varepsilon_1 = 10^{-7}$, $\varepsilon_2 = 10^{-6}$ and $A_1 = 10^7$. For the step-size sequence, we choose $\gamma_n = 10/n$. Because draws from the samplers at early stage of the simulation are unreliable, when we adapt the covariance matrix, we start estimating it after $1,000$ iterations and we start using the estimates to make subsequent moves in the sampler after $5,000$ iterations.

To compare the samplers, we compare the autocorrelation functions of the first component of the random process generated. The other components show similar results. These autocorrelations are shown in Graph 1. We also compare the estimates of the mean square jump of the algorithm in stationarity defined as $d = \mathbb{E}^{1/2}\left[|X_n - X_{n-1}|^2\right]$. This indicates how fast the process is moving around the states space. Finally, we compare the statistical efficiency of the samplers by comparing how well they estimate $\mu_1 = \int x_1 \pi(dx_1, \ldots, dx_{20}) = 0$ the mean of the first component of the distribution. The standard errors of the estimates are obtained with 50 independent replications of each sampler. We give the efficiency of each sampler relatively to *RWM1* defined as the standard error of *RWM1* over the standard error of that sampler. Table 1 presents these estimates.

These simulations indicate that the fully adaptive algorithm performs almost like the optimal Markov chain and both considerably outperform the adaptive algorithm where the covariance matrix is not updated.

In practice, manual parameter tuning of MCMC algorithms is still largely used. It is now clear that manual parameter tuning of Metropolis-Hastings algorithms is inefficient and resource waisteful. For example, to muanually tune the sampler discussed in this example, one would have to run a number of preliminary simulations to obtain an estimate $\hat{\Sigma}_\pi$ of Σ_π. Given $\hat{\Sigma}_\pi$, another run of simulations will be needed to obtain $\hat{\sigma}$ an estimate of σ that gives the appropriate target acceptance rate. Only then the actual MCMC simulation starts with $\hat{\sigma}$ and $\hat{\Sigma}_\pi$; with no guarantee that these quantites have been correctly estimated. In the best case scenario, the manually tuned sampler will have similar performance as the best sampler and the adaptive sampler. So manual tuning is resource waistful.

But overall, to build confidence in adaptive MCMC, we need to know more about the random processes generated by these algorithms. Some recent progress has been made in Andrieu and Atchade (2005) where it is shown that some adaptive MCMC asymptotically behave like Markov chains and satisfy a central limit theorem.

3.2 Nuclear Pump Failures

This example is taken from Robert and Casella (2004) and model the failure of pumps at a nuclear plant. Ten pumps are observed over some period of time (t_i) and the number of failures (p_i) is recorded and is given in Table 2.

The failures of the i-th pump is assumed to occur according to a Poisson process with parameter λ_i. We assume independent prior distributions for the λ_i: $\lambda_i \underset{\sim}{iid} \mathcal{G}(\alpha, \beta)$, the Gamma distribution with parameter α, β; and $\beta \sim \mathcal{G}(\gamma, \delta)$, where $\alpha = 1.8$, $\gamma = 0.01$ and $\delta = 1$. The posterior distribution for $(\lambda_1, \dots, \lambda_{10}, \beta)$ is:

$$\pi(\lambda_1, \dots, \lambda_{10}, \beta) \propto \beta^{17.01} \exp(-\beta) \prod_{i=1}^{10} \lambda_i^{p_i+0.8} \exp(-\lambda_i(t_i + \beta)).$$

We apply the samplers *RWM1*, *RWM2*, *MALA1* and *MALA2* described above to sample from this posterior distribution and compare their performances. Graph 2 shows the autocorrelation functions of the tenth component of the random process generatd and Table 3 presents the mean square jumps in stationarity. As above we find a huge improvement in the algorithm when the covariance matrix of the proposal kernel is properly scaled as in *MALA2* and *RWM2*.

4 A Convergence Result for Stochastic Approximation Algorithms

We introduce in this section a new approach to analyze stochastic approximation algorithms with Markovian dynamics. Stochastic approximation is a well-known numerical method used to solve equations of the form $h(\theta) = 0$ when h cannot be (easily) computed and only noisy estimates are available. These recursive algorithms typically takes the general form:

$$\theta_{n+1} = \theta_n + \gamma_n h(\theta_n) + \gamma_n \varepsilon_{n+1}, \tag{4.1}$$

where (ε_n) is “the noise” process. An extensive literature exists on these algorithms (see e.g., Benveniste et al. (1990), Kushner and yin (2003) and the references therein) particularly when the noise process (ε_n) is a martingale difference. The relevance to our adaptive MCMC is that we will show (see the proof of Theorem 2.1 in Section 5) that the adaptation process $(\sigma_n, \mu_n, \Gamma_n)$ in Algorithm 2.1 follows a stochastic approximation of the form (4.1), where the noise process is fed by the adaptive chain (X_n). In that case, we show below that (ε_n) is actually a mixingale difference. Using mixingale theory (see e.g., Hall and Heyde (1980) for an introduction to mixingales), we can avoid the Poisson equation approach (Metivier and Priouret, 1984) and give an analysis similar to the case where (ε_n) is a martingale difference. This is the object of this section. The results are summarized in Theorem 4.1.

Let $(P_\theta)_{\theta\in\Theta}$ be a family of transition kernels on some probability space $(\mathcal{X}, \mathcal{B}, \pi)$, where Θ is some compact subset of $\mathbb{R}^p$, the p-dimensional Euclidean space. We assume that for $A \in \mathcal{B}$ fixed, $P_\theta(x, A)$ is a measurable function of (θ, x). We consider the $\mathcal{X}$-valued adaptive chain defined on some probability triplet $(\Omega, \mathcal{F}, \Pr)$ as follows: $X_0 = x_0 \in \mathcal{X}$, and conditional to $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$, $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$, where θ_n is some $\mathcal{F}_n$-measurable Θ-valued random variable. Let $\mathbb{E}$ denotes the expectation with respect to Pr. We are interested in the asymptotics of the random process (θ_n, X_n).

We make the following assumptions:

Assumption B1 *For any $\theta \in \Theta$, P_θ has invariant distribution π.*

Assumption B2 *There exists a measurable function $V: \mathcal{X} \longrightarrow [1, \infty)$ such that for any $\alpha \in (0, 1]$, we can find $R_\alpha < \infty$, $\rho_\alpha < 1$, $b_\alpha < \infty$ such that:*

$$\sup_{\theta \in \Theta} \left\| P_\theta^n(x, \cdot) - \pi(\cdot) \right\|_{V^\alpha} \leq R_\alpha \rho_\alpha^n V^\alpha(x), \;\; x \in \mathcal{X}, \tag{4.2}$$

and

$$\mathbb{E}(V^\alpha(X_{n+k}) | \mathcal{F}_n) \leq b_\alpha V^\alpha(X_n), \; n \geq 0, k \geq 0. \tag{4.3}$$

For transition kernels $P1$ and P_2 on $(\mathcal{X}, \mathcal{B})$ and $W: \mathcal{X} \to (0, \infty)$, define $\|P_1 - P_2\|_W := \sup_{|f| \leq W} \sup_{x \in \mathcal{X}} \frac{|P_1 f(x) - P_2 f(x)|}{W(x)}$. We assume that:

Assumption B3 *For all $\alpha \in (0, 1]$, there exists a constant $K_1 = K_1(\alpha) < \infty$ such that for all $\theta_1, \theta_2 \in \Theta$:*

$$\|P_{\theta_1} - P_{\theta_2}\|_{V^\alpha} \leq K_1 |\theta_1 - \theta_2|, \tag{4.4}$$

where V is the function introduced in Assumption (B2).

Assumption B4 *There exist $\beta \in [0, 1/2)$, $\lambda > 0$, and a sequence of positive real numbers (γ_n), $\gamma_n = O(n^{-\lambda})$ such that*

$$|\theta_n - \theta_{n-1}| \leq \gamma_{n-1} V^\beta(X_n), \;\; n \geq 1. \tag{4.5}$$

Let $G: \mathcal{X} \times \Theta \longrightarrow R^q$ be a measurable function. Assume that there exist constants $K_2, K_3 < \infty$ such that:

$$\sup_{\theta \in \Theta} |G(x, \theta)| \leq K_2 V^\beta(x), \; x \in \mathcal{X}, \tag{4.6}$$

$$|G(x, \theta_1) - G(x, \theta_2)| \leq K_3 V^\beta(x) |\theta_1 - \theta_2|, \;\; \theta_1, \theta_2 \in \Theta, x \in \mathcal{X}, \tag{4.7}$$

where β is as in (B4).

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

Theorem 4.1 *Let G be such that (4.6) and (4.7) hold and assume (B1)–(B4). For $n \geq 1$, define $Z_n = (G(X_n, \theta_{n-1}) - g(\theta_{n-1}))$. There exists a finite constant C_1 such that:*

(i)

$$|\mathbb{E}(Z_{n+k} | \mathcal{F}_n)| \leq C_1 \gamma_k \log(k+1) V^{2\beta}(X_n), \; a.s. \; n \geq 0, k \geq 1. \tag{4.8}$$

(ii) *Let (b_n) be a sequence of positive real numbers. Then the random process $(b_n[Z_n - \mathbb{E}(Z_n)])$ is a $L^{1+\varepsilon}$-mixingale with respect to $(\mathcal{F}_n)_{-\infty<n<\infty}$, with $\varepsilon = \min(1, \frac{1}{2\beta} - 1)$ and $\mathcal{F}_n = \{\emptyset, \mathcal{X}\}$ for $n \le 0$.*

(iii) *In particular if $\lambda > 1/2$, then: $\frac{1}{n}\sum_{k=1}^{n} Z_k \to 0$ a.s. as $n \to \infty$. Also, if $\varepsilon = 1$ and (b_n) is such that $\sum b_k^2 < \infty$, then $\sum b_n Z_n$ converges almost surely to a finite random variable.*

Proof:

(i) Define $f_n(x) = G(x, \theta_n) - g(\theta_n)$. Then

$$Z_{n+k} = f_n(X_{n+k}) + G(X_{n+k}, \theta_{n+k}) - G(X_{n+k}, \theta_n) + g(\theta_n) - g(\theta_{n+k}). \tag{4.9}$$

It follows from (4.7) that $|g(\theta_2) - g(\theta_1)| \le K_3 \pi(V^\beta)|\theta_2 - \theta_1|$. Thus, using (4.7) again:

$$\begin{aligned} &|G(X_{n+k}, \theta_{n+k}) - G(X_{n+k}, \theta_n)| + |g(\theta_n) - g(\theta_{n+k})| \\ &\le R_3 V^\beta(X_{n+k})|\theta_{n+k} - \theta_n|, \end{aligned} \tag{4.10}$$

for some finite constant R_3 and hence:

$$\begin{aligned} |\mathbb{E}(Z_{n+k}|\mathcal{F}_n)| &\le R_4 k \gamma_n \mathbb{E}\big(V^{2\beta}(X_{n+k})|\mathcal{F}_n\big) + |\mathbb{E}(f_n(X_{n+k})|\mathcal{F}_n)|, \\ &\le R_5 k \gamma_n V^{2\beta}(X_n) + |\mathbb{E}(f_n(X_{n+k})|\mathcal{F}_n)|, \end{aligned} \tag{4.11}$$

by (4.3) and (4.5). A similar argument to Atchade and Rosenthal (2005) (Lemma 3.1) can be used to show that there exist constants $R_6, R_7 < \infty$ and $\rho < 1$ such that

$$|\mathbb{E}(f_n(X_{n+k})|\mathcal{F}_n)| \le R_6 \rho^k V^\beta(X_n) + R_7 k \gamma_n V^{2\beta}(X_n), \tag{4.12}$$

which leads to:

$$|\mathbb{E}(Z_{n+k}|\mathcal{F}_n)| \le R_8\big(\rho^k + k\gamma_n\big)V^{2\beta}(X_n). \tag{4.13}$$

Since $(\mathcal{F}_{\mathrm{n}})$ is nondecreasing, for $n \ge 0, k \ge 1$, $0 \le j \le k$ we have

$$\begin{aligned} |\mathbb{E}(Z_{n+k}|\mathcal{F}_n)| &= \big|\mathbb{E}\big[\mathbb{E}\big(Z_{n+k}|\mathcal{F}_{n+k-j}\big)|\mathcal{F}_n\big]\big|_8\big(\rho^j + j\gamma_{n+k}\big)\mathbb{E}(V^{2\beta}(X_{n+k-j})|\mathcal{F}_n) \\ &\le R_9\big(\rho^j + j\gamma_{n+k}\big)V^{2\beta}(X_n). \end{aligned}$$

Therefore, $|\mathbb{E}(Z_{n+k}|\mathcal{F}_n)| \le \min_{0\le j\le k} R_9\big(\rho^j + j\gamma_{n+k}\big)V^{2\beta}(X_n)$ and taking $j = \frac{\lambda}{-\log(\rho)}\log(k+1)$ we obtain $|\mathbb{E}(Z_{n+k}|\mathcal{F}_n)| \le C_1 \gamma_k \log(k+1) V^{2\beta}(X_n)$ and (i) is proved.

(ii) Define $Y_n = Z_n - \mathbb{E}(Z_n)$ and $\mathcal{F}_n = \{\emptyset, \Omega\}$ if $n < 0$. For $n, k \ge 0$, if $n - k < 0$, we have $\mathbb{E}(b_n Y_n|\mathcal{F}_{n-k}) = 0$. It follows from (i) that if $n - k \ge 0$, we have $|\mathbb{E}(b_n Y_n\ |\mathcal{F}_{n-k})| \le b_n|\mathbb{E}(Z_n|\mathcal{F}_{n-k})| + b_n|\mathbb{E}(Z_n)| \le C_2\gamma_k \log(k+1) b_n V^{2\beta}(X_{n-k})$. Therefore for $p = 1 + \varepsilon$, since $\sup_n \mathbb{E}\big[V^{2\beta p}(X_n)\big] < \infty$, we have: $\{\mathbb{E}[\mathbb{E}(b_n Y_n|\mathcal{F}_{n-k})]^p\}^{1/p} \le c_n \phi_k$, where $c_n = O(b_n)$ and $\phi_n = O(\gamma_n \log(1+n))$. This implies that $(b_n Y_n, \mathcal{F}_n)$ is a L^p mixingale with mixingales sequences (c_n) and (ψ_n).

(iii) Take $b_n \equiv 1$. By Corollary 2.2 of Davidson and de Jong (1997) we conclude that $\frac{1}{n}\sum_{i=1}^{n} Y_i \to 0$ a.s. But since from (i) we have: $|\mathbb{E}(Z_n)| \le C_1\gamma_n \log(1+n)V^{2\beta}(x_0) \to 0$ as $n \to \infty$, we have $\frac{1}{n}\sum_{i=1}^{n} Z_i \to 0$ a.s.

The remaining part of the assertion is Theorem 2.7 of Hall and Heyde (1980). ■

5 Proof of Theorem 2.1

The proof is in a large part a direct application of Theorem 4.1. Let θ_n be the triplet $(\mu_n, \Gamma_n, \sigma_n)$ defined in Algorithm 2.1. Also define $\Theta = B(0, A_1) \times \Theta_\Gamma \times \Theta_\sigma$. For $\theta = (\mu, \Gamma, \sigma) \in \Theta$, define $|\theta| = \sqrt{|\sigma|^2 + |\mu|^2 + |\Gamma|^2}$. Because of the reprojection used, the adaptive chain (X_n, θ_n) generated by Algorithm 2.1 lives in $\mathcal{X} \times \Theta$. Clearly for this adaptive chain, (B1) hold. Let $\delta \in (1/2, 1)$ and define $V(x) = c\pi^\delta(x)$ and c is such that $V \ge 1$. It is well known that (B2) follows from Proposition 2.1 (see e.g., Baxendale (2005)) and (B3) is precisely Proposition 2.2. Define $\beta = 1/4\delta$ and note that $0 < \beta < 1/2$. From (2.4)–(2.6), the fact that μ_n, Γ_n, σ_n are bounded and the fact that $|x| + |x|^2 \le C_1 V^\beta(x)$, for some finite constant C_1 it follows that $|\theta_{n+1} - \theta_n| \le |\mu_{n+1} - \mu_n| + |\Gamma_{n+1} - \Gamma_n| + |\sigma_{n+1} - \sigma_n| \le R_2\gamma_n V^\beta(X_{n+1})$, which is (B4).

5.1 Proof of Theorem 2.1 (i)

It suffices to take $G(x, \theta) = f(x)$ in Theorem 4.1 and (i) gives $|\mathbb{E}(f(X_n) - \pi(f))| \le KV^{2\beta}(x_0)\gamma_n \log(1+n)$ which is (2.10). The strong law of large numbers in Theorem 4.1 (iii) yields $\frac{1}{n}\sum_{i=0}^{n-1}(f(X_i) - \pi(f)) \to 0$ which is (2.11).

Note that the adaptation process (θ_n) itself need not converge.

5.2 Proof of Theorem 2.1 (ii)

The proof combines Theorem 4.1 and a technical lemma given in Lemma 6.3. The details of the argument are similar for (μ_n), (Γ_n) and (σ_n).

Convergence of μ_n: For $n \ge 0$, we have:

$$\begin{aligned}|\mu_{n+1} - \mu_\pi|^2 &= |p_3(\mu_n + \gamma_n(X_{n+1} - \mu_n)) - \mu_\pi|^2 \\ &\le |\mu_n - \mu_\pi + \gamma_n(X_{n+1} - \mu_n)|^2 \\ &\le |\mu_n - \mu_\pi|^2 - 2\gamma_n|\mu_n - \mu_\pi|^2 + K\gamma_n^2 V^{2\beta}(X_{n+1}) + 2\gamma_n(\mu_n \\ &\quad - \mu_\pi)\cdot(X_{n+1} - \mu_\pi),\end{aligned} \tag{5.1}$$

K constant.

Since (B2) hold, the sequence $\left(E\left(V^{2\beta}(X_n)\right)\right)$ is bounded. Also, Theorem 4.1 (i) applied to $G(x, \theta) = \langle \mu - \mu_\pi, x\rangle$ yields:

$$\mathbb{E}[\langle \mu_n - \mu_\pi, X_{n+1} - \mu_\pi\rangle] = O(\gamma_n \log(1+n)). \tag{5.2}$$

For $n \geq 1$, let $U_n^{(1)} = \mathbb{E}\left[|\mu_n - \mu_\pi|^2\right]$. The inequality (5.1) implies the existence of a finite constant C_2 such that:

$$U_{n+1}^{(1)} \leq (1 - 2\gamma_n)U_n^{(1)} + C_2\gamma_n^2 \log(n+1) \tag{5.3}$$

$$\leq e^{-2\gamma_n} U_n^{(1)} + C_2\gamma_n^2 \log(n+1). \tag{5.4}$$

Lemma 6.3 then gives: $U_n^{(1)} = O(\gamma_n \log(n+1))$.

Convergence of Γ_n: The proof is similar to what is done above for (μ_n). For matrices A and B, write $A{\bullet}B = tr^{1/2}(AB)$ the inner product of A, B. In a way similar to what we did above, we have:

$$\begin{aligned}
|\Gamma_{n+1} - \Sigma_\pi|^2 &= \left|p_2\big(\Gamma_n + \gamma_n\big((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)' - \Gamma_n\big)\big) - \Sigma_\pi\right|^2 \\
&\leq \left|\Gamma_n - \Sigma_\pi + \gamma_n\big((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)' - \Gamma_n\big)\right|^2 \\
&\leq |\Gamma_n - \Sigma_\pi|^2 - 2\gamma_n|\Gamma_n - \Sigma_\pi|^2 + K\gamma_n^2 V^{2\beta}(X_{n+1}) + 2\gamma_n(\Gamma_n \\
&\quad - \Sigma_\pi){\bullet}\big((X_{n+1} - \mu_\pi)(X_{n+1} - \mu_\pi)' - \Sigma_\pi\big) + 2\gamma_n(\Gamma_n \\
&\quad - \Sigma_\pi){\bullet}\big((X_{n+1} - \mu_\pi)(\mu_\pi - \mu_n)'\big) + 2\gamma_n(\Gamma_n \\
&\quad - \Sigma_\pi){\bullet}\big((\mu_n - \mu_\pi)(X_{n+1} - \mu_n)'\big).
\end{aligned}$$

Write $U_n^{(2)} = \mathbb{E}\left[|\Gamma_n - \Sigma_\pi|^2\right]$ and

$$\begin{aligned}
W_n &= \gamma_n^2 V^{2\beta}(X_{n+1}) + 2\gamma_n(\Gamma_n - \Sigma_\pi){\bullet}\big((X_{n+1} - \mu_\pi)(X_{n+1} - \mu_\pi)' - \Sigma_\pi\big) + 2\gamma_n(\Gamma_n \\
&\quad - \Sigma_\pi){\bullet}\big((X_{n+1} - \mu_\pi)(\mu_\pi - \mu_n)'\big) + 2\gamma_n(\Gamma_n - \Sigma_\pi){\bullet}\big((\mu_n - \mu_\pi)(X_{n+1} - \mu_n)'\big).
\end{aligned}$$

Applying Theorem 4.1 (i) to the appropriate functions, we obtain that $\mathbb{E}[W_n] = O(\gamma_n \log(n+1))$ so that:

$$U_{n+1}^{(2)} \leq e^{1-2\gamma_n} U_n^{(2)} + \gamma_n^2 \log(n+1). \tag{5.5}$$

As above, we apply Lemma 6.3 to obtain that $U_n^{(2)} = O(\gamma_n \log(n+1))$ as wanted.

Convergence of σ_n: The argument procedes also as above. Consider the function $A(x, \sigma, \Gamma) = \int \alpha_{\sigma,\Lambda}(x,y) q_{\sigma,\Lambda}(x,y)dy$, $\Lambda = \Gamma + \varepsilon_2 I_d$ and $\tau(\sigma, \Gamma) = \int A(x, \sigma, \Gamma)\pi(dx)$. It can be shown that τ is Lipschitz. On the other hand, we have:

$$\begin{aligned}
\left|\sigma_{n+1} - \sigma_{opt}\right|^2 &\leq \left|\sigma_n - \sigma_{opt}\right|^2 + 2\gamma_n(\sigma_n - \sigma_{opt})(\tau(\sigma_n, \Sigma_\pi) - \overline{\tau}) + 2\gamma_n(\sigma_n \\
&\quad - \sigma_{opt})(\tau(\sigma_n, \Gamma_n) - \tau(\sigma_n, \Sigma_\pi)) + \gamma_n^2 + 2\gamma_n(\sigma_n - \sigma_{opt}) \\
&\quad \times \big(\alpha_{\sigma_n,\Lambda_n}(X_n, Y_{n+1}) - \tau(\sigma_n, \Gamma_n)\big).
\end{aligned} \tag{5.6}$$

From (A2)(ii), we have $(\sigma_n - \sigma_{opt})(\tau(\sigma_n, \Sigma_\pi) - \overline{\tau}) \leq -\delta\left|\sigma_n - \sigma_{opt}\right|^2$. From what is obtained above on (Γ_n), we have:

$$\begin{aligned}\mathbb{E}\left[\left|(\sigma_n - \sigma_{opt})(\tau(\sigma_n, \Gamma_n) - \tau(\sigma_n, \Sigma_\pi))\right|\right] &= O(\mathbb{E}[|\Gamma_n - \Sigma_\pi|]) \\ &= O\left(\gamma_n^{1/2}\log^{1/2}(n+1)\right).\end{aligned}$$

Therefore:

$$U_{n+1}^{(3)} \leq e^{-\delta\gamma_n} U_n^{(3)} + C_3 \gamma_n^{3/2} \log^{1/2}(n+1), \tag{5.7}$$

where $U_n^{(3)} = \mathbb{E}\left[\left|\sigma_n - \sigma_{opt}\right|^2\right]$ and C_3 a finite constant which imply (Lemma 6.3) that $U_n^{(3)} = O\left(\gamma_n^{1/2}\log^{1/2}(n+1)\right)$

6 Proofs of the Technical Results

We prove Proposition 2.1 in Section 6.1 and Proposition 2.2 in Section 6.2.

6.1 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 existence of a drift in the algorithm. But the fact that the drift is bounded is crucial.

Proof of Proposition 2.1: In Lemma 6.1 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 6.2 below we show that we can find $\lambda < 1$, $b < \infty$ such that

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

■

Lemma 6.1 *There is $\varepsilon > 0$, a Ball C, a nontrivial probability measure ν such that: $\inf_{(\sigma,\Lambda)\in\mathcal{C}} P_{\sigma,\Lambda}(x, A) \geq \varepsilon\nu(A), \;\; A \in \mathcal{B} \;\; x \in C$.*

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. ■

Lemma 6.2 *Assume (A1) and let α and V_α as in Proposition 2.1. There exist $\lambda = \lambda_\alpha < 1, \;\; b = b_\alpha < \infty$ 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 6.1.*

Proof: We only need to show that:

$$\sup_{x\in\mathcal{X}}\sup_{(\sigma,\Lambda)\in\mathcal{C}}\frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)}<\infty, \tag{6.1}$$

and

$$\limsup_{|x|\to\infty}\sup_{(\sigma,\Lambda)\in\mathcal{C}}\frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)}<1. \tag{6.2}$$

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:

$$k_1g_{\varepsilon_1}(y-x)\leq q_{\sigma,\Lambda}(x,y)\leq k_2g_{\varepsilon_2}(y-x), \tag{6.3}$$

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:

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

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

$$\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}q_{\sigma,\Lambda}(x,y)\leq q^{\alpha}_{\sigma,\Lambda}(y,x)q^{1-\alpha}_{\sigma,\Lambda}(x,y)\leq k_2^2g_{\varepsilon_2}(y-x), \tag{6.4}$$

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

$$\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}q_{\sigma,\Lambda}(x,y)\leq q^{1-\alpha}_{\sigma,\Lambda}(y,x)q^{\alpha}_{\sigma,\Lambda}(x,y)\leq k_2^2g_{\varepsilon_2}(y-x). \tag{6.5}$$

Hence (6.1) is satisfied.

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

$$\int_{B(x,R)}g_{\varepsilon_2}(y-x)dy\geq 1-\varepsilon. \tag{6.6}$$

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 (6.3) 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 $|x|\geq r_2$,

$$\int_{C_{\pi(x)}(u)\cap B(x,R)}g_{\varepsilon_2}(y-x)dy\leq\varepsilon. \tag{6.7}$$

Now, for $S\in\{A_{\sigma,\Lambda}(x),R_{\sigma,\Lambda}(x)\}$ and u as in (6.7), write $S=(S\cap B(x,R)^c)\bigcup\big(S\cap B(x,R)\cap C_{\pi(x)}(u)\big)\bigcup\big(S\cap B(x,R)\cap C_{\pi(x)}(u)^c\big)$. For $|x|\geq r_2$, it follows from (6.4), (6.6) and (6.7) that:

$$\begin{aligned}&\int_{A_{\sigma,\Lambda}(x)\cap B(x,R)^c} q_{\sigma,\Lambda}(x,y)\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}dy+\int_{A_{\sigma,\Lambda}(x)\cap B(x,R)\cap C_{\pi(x)}(u)} q_{\sigma,\Lambda}(x,y)\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}dy\\&\quad\leq 2k_2^2\varepsilon,\end{aligned}\tag{6.8}$$

and from (6.5), (6.6) and (6.7) we have:

$$\begin{aligned}&\int_{R_{\sigma,\Lambda}(x)\cap B(x,R)^c}\left(\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}\right)q_{\sigma,\Lambda}(x,y)dy\\&\quad+\int_{R_{\sigma,\Lambda}(x)\cap B(x,R)\cap C_{\pi(x)}(u)}\left(\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}\right)q_{\sigma,\Lambda}(x,y)dy\\&\quad\leq 2k_2^2\varepsilon.\end{aligned}\tag{6.9}$$

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 $|x|\geq r_3+R$:

$$\int_{A_{\sigma,\Lambda}(x)\cap B(x,R)\cap C_{\pi}(\delta)^c} q_{\sigma,\Lambda}(x,y)\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)}dy\leq d_{r_3}(\delta).\tag{6.10}$$

and

$$\int_{R_{\sigma,\Lambda}(x)\cap B(x,R)\cap C_{\pi(x)}(\delta)^c}\left(\frac{\pi^{1-\alpha}(y)q_{\sigma,\Lambda}(y,x)}{\pi^{1-\alpha}(x)q_{\sigma,\Lambda}(x,y)}\right)q_{\sigma,\Lambda}(x,y)dy\leq k_2d_{r_3}(u).\tag{6.11}$$

The bounds (6.8)–(6.11) implies that:

$$\begin{aligned}\limsup_{|x|\to\infty}\sup_{(\sigma,\Lambda)\in\mathcal{C}}\frac{P_{\sigma,\Lambda}V_\alpha(x)}{V_\alpha(x)}&=\limsup_{|x|\to\infty}\sup_{(\sigma,\Lambda)\in\mathcal{C}}Q\big(x,R_{\sigma,\Lambda}(x)\big)\\&=1-\liminf_{|x|\to\infty}\inf_{(\sigma,\Lambda)\in\mathcal{C}}Q_{\sigma,\Lambda}\big(x,A_{\sigma,\Lambda}(x)\big).\end{aligned}\tag{6.12}$$

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 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 (6.12) shows (6.2) 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. ■

6.2 Proof of Proposition 2.2

Proof: We only sketch the proof leaving the details to the reader. The idea is to show that for $|f| \leq V_\alpha$ and any $x \in \mathcal{X}$, there exists a finite constant K such that $\sup_{(\sigma,\Lambda)\in\mathcal{C}} \|\frac{\partial}{(\sigma,\Lambda)} P_{\sigma,\Lambda} f(x)\| \leq KV^{1/2}(x)$, 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 $\alpha_{\sigma,\Lambda}(x,y) = \min\big(1, r_{\sigma,\Lambda}(x,y)\big)$, so that

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

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: $\frac{\partial}{\partial(\sigma,\Lambda)} q_{\sigma,\Lambda}(x,y)(h,H) = q_{\sigma,\Lambda}(x,y) \times (B_1(x,y,\sigma,\Lambda,h) + B_2(x,y,\sigma,\Lambda,H))$, where the functions B_1, B_2 satisfy: $|B_1(x,y,\sigma,\Lambda,h)| + |B_2(x,y,\sigma,\Lambda,H)| \leq K_2|y-x|^2|(h,H)|$ for some finite constant K_2. And a straightforward calculus gives for any (h, H) with $|(h,H)| \leq 1$:

$$\begin{aligned}
&\left|\frac{\partial}{\partial(\sigma,\Lambda)}\left[\big(\alpha_{\sigma,\Lambda}(x,y)q_{\sigma,\Lambda}(x,y)\big)f(y)\right](h,H)\right| \\
&= \left|\frac{\pi(y)}{\pi(x)} \mathbf{1}_{\{r_{\sigma,\Lambda}(x,y)\leq 1\}} \frac{\partial}{\partial(\sigma,\Lambda)}\left[q_{\sigma,\Lambda}(y,x)f(y)\right](h,H)\right. \\
&\quad \left. + \mathbf{1}_{\{r_{\sigma,\Lambda}(x,y)>1\}} \frac{\partial}{\partial(\sigma,\Lambda)}\left[q_{\sigma,\Lambda}(y,x)f(y)\right](h,H)\right| \\
&\leq K_2|y-x|^2 V_\alpha(x) q^{\alpha}_{\sigma,\Lambda}(x,y) q^{1-\alpha}_{\sigma,\Lambda}(y,x) \leq K_3|y-x|^2 V_\alpha(x) q_{\varepsilon_2}(x,y),
\end{aligned} \tag{6.13}$$

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, $\left|\frac{\partial}{\partial(\sigma,\Lambda)}\left[(1-\alpha_{\sigma,\Lambda}(x,y))q_{\sigma,\Lambda}(x,y)\right]\right.$ $\left.(h,H)\right| \leq K_3|y-x|^2 q_{\varepsilon_2}(x,y)$.

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

$$\left|\frac{\partial}{\partial(\sigma,\Lambda)} P_{\sigma,\Lambda} f(x)(h,H)\right| \leq 2K_3 V^{1/2}(x) \int |y-x|^2 q_{\varepsilon_2}(x,y)dy, \tag{6.14}$$

and we are done. ■

6.3 A technical Lemma

We need the following technical lemma. For a proof, see Pelletier (1998). We say that a sequence (u_n) is regularly varying with exponent b if $u_n = u(n)$ and the function u satisfies $\lim_{t\to\infty}(u(tx)/u(t)) = x^b$.

LEMMA 6.3 *Let (ε_n) be a positive sequence that is decreasing for nlarge enough and regularly varying with exponent $-\nu$, $\nu \geq 0$. Let (γ_n)as in (A3). For $\lambda > 0$, let (x_n) be a non-negative sequence such that*

$$x_n \leq e^{-\lambda\gamma_n} x_{n-1} + \gamma_n \varepsilon_n. \tag{6.15}$$

Then we have:

$$\limsup_n \frac{x_n}{\varepsilon_n} \leq \frac{1}{\lambda}. \tag{6.16}$$

Acknowledgments I am grateful to an anonymous referee for very helpful comments. This work has been partly supported by NSERC Canada and University of Ottawa.

References

C. Andrieu, and Y. F. Atchade, "On the efficiency of adaptive MCMC algorithms," *Technical Report 1*, 2005.

C. Andrieu, and E. Moulines, "On the ergodicity properties of some adaptive MCMC algorithms," to appear *Annals of Applied Probability*, 2005.

Y. F. Atchade, and J. S. Rosenthal, "On adaptive Markov chain Monte Carlo algorithm," *Bernoulli* vol. 11 pp. 815–828, 2005.

P. H. Baxendale, "Renewal theory and computable convergence rates for geometrically ergodic Markov chains," *Annals of Applied Probability* vol. 15 pp. 700–738, 2005.

A. Benveniste, M. Métivier, and P. Priouret, "Adaptive algorithms and stochastic approximations," In *Applications of Mathematics*, Springer: Paris-New York, 1990.

L. Breyer, M. Piccioni, and S. Scarlatti, "Optimal scaling of MALA for nonlinear regression," *Technical Report*, 2002.

J. Davidson, and R. de Jong, "Strong laws of large numbers for dependent heteregeneous processes: a synthesis of recent and new results," *Econometric Reviews* vol. 16 pp. 251–279, 1997.

W. R. Gilks, G. O. Roberts, and S. K. Sahu, "Adaptive Markov chain Monte Carlo through regeneration," *Journal of the American Statistical Association* vol. 93 pp. 1045–1054, 1998.

H. Haario, E. Saksman, and J. Tamminen, "An adaptive metropolis algorithm," *Bernoulli* vol. 7 pp. 223–242, 2001.

P. Hall, and C. C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press: New York, 1980.

S. F. Jarner, and E. Hansen, "Geometric ergodicity of Metropolis algorithms," *Stochastic Processes and their Applications* vol. 85 pp. 341–361, 2000.

K. Kushner, and Y. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer-Verlag: New-York, 2003.

M. Metivier, and P. Priouret, "Application of Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE-IT* vol. 30, 1984.

M. Pelletier, "On the almost sure asymptotic behaviour of stochastic algorithms," *Stochastic Processes and their Applications* vol. 78 pp. 217–244, 1998.

C. P. Robert, and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, New York, 2004.

G. O. Roberts, and J. S. Rosenthal, "Optimal scaling of various Metropolis-Hastings algorithms," *Statistical Science* vol. 16, 2001.

G. Roberts, and R. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli* vol. 2 pp. 341–363, 1996.

J. S. Rosenthal, and G. O. Roberts, "Coupling and Ergodicity of adaptive MCMC," *Technical Report, MCMC preprints*, 2005.

O. Stramer, and R. L. Tweedie, "Langevin-type models (ii): Self-targeting candidates for MCMC algorithms," *Methodology and Computing in Applied Probability* vol. 1 pp. 307–328, 1999.

L. Tierney, "Markov chains for exploring posterior distributions," *The Annals of Statistics* vol. 22 pp. 1701–1762, 1994. With discussion and a rejoinder by the author.