Elect. Comm. in Probab. **12** (2007), 336–349

ELECTRONIC COMMUNICATIONS in PROBABILITY

ON THE EFFICIENCY OF ADAPTIVE MCMC ALGORITHMS

CHRISTOPHE ANDRIEU
Department of Mathematics, University of Bristol, Bristol, England
email: c.andrieu@bristol.ac.uk

YVES F. ATCHADÉ[1]
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
email: yvesa@umich.edu

Submitted February 21, 2007, *accepted in final form* August 13, 2007

AMS 2000 Subject classification: 60J35, 60J22, 65C40
Keywords: Adaptive Markov chains, Coupling, Markov Chain Monte Carlo, Metropolis Algorithm, Stochastic Approximation, Rate of convergence

Abstract

We study a class of adaptive Markov Chain Monte Carlo (MCMC) processes which aim at behaving as an "optimal" target process via a learning procedure. We show, under appropriate conditions, that the adaptive MCMC chain and the "optimal" (nonadaptive) MCMC process share many asymptotic properties. The special case of adaptive MCMC algorithms governed by stochastic approximation is considered in details and we apply our results to the adaptive Metropolis algorithm of Haario et al. (2001)

1 Introduction

Markov chain Monte Carlo (MCMC) is a popular computational method for generating samples from virtually any distribution π defined on a space $\mathcal{X}$. These samples are often used to efficiently compute expectations with respect to π by invoking some form of the law of large numbers. The method consists of simulating an ergodic Markov chain $\{X_n, n \geq 0\}$ on $\mathcal{X}$ with transition probability P such that π is a *stationary* distribution for this chain. In practice the choice of P is not unique, and instead it is required to choose among a family of transition probabilities $\{P_\theta,\ \theta \in \Theta\}$ for some set Θ. The problem is then that of choosing the "best" transition probability P_θ from this set, according to some well defined criterion. For example, the efficiency of a Metropolis-Hastings algorithm highly depends on the scaling of its proposal transition probability. In this case, the *optimal scaling* depends on π, the actual target distribution, and cannot be set once for all. For more details on MCMC methods, see e.g. Gilks et al. (1996).

[1]RESEARCH SUPPORTED IN PART BY NSERC CANADA

An attractive solution to this problem, which has recently received attention, consists of using so-called adaptive MCMC methods where the transition kernel of the algorithm is sequentially tuned during the simulation in order to find the "best" θ (see e.g. Gilks et al. (1998), Haario et al. (2001), Andrieu and Robert (2001), Andrieu and Moulines (2006) and Atchade and Rosenthal (2005)). These algorithms share more or less the same structure and fit, as pointed out in Andrieu and Robert (2001), in the general framework of controlled Markov chains. More precisely one first defines a sequence of measurable functions $\{\rho_n : \Theta \times \mathcal{X}^{n+1} \longrightarrow \Theta, \text{ for } n \geq 0\}$ which encodes what is meant by "best". The adaptive chain is initialized with some arbitrary but fixed values $(\theta_0, X_0) \in \Theta \times \mathcal{X}$. At iteration $n \geq 1$, given $(\theta_0, X_0, \ldots, X_{n-1})$, and $\theta_{n-1} = \rho_{n-1}(\theta_0, X_0, \ldots, X_{n-1})$ (with the convention that $\rho_0(\theta, X) = \theta$), X_n is generated according to $P_{\theta_{n-1}}(X_{n-1}, \cdot)$ and $\theta_n = \rho_n(\theta_0, X_0, \ldots, X_n)$. Most examples currently developed in the literature rely on stochastic approximation type recursions *e.g.* Haario et al. (2001), Andrieu and Robert (2001) and Atchade and Rosenthal (2005). Clearly, the random process $\{X_n\}$ is in general not a Markov chain. However, with an abuse of language, we will refer here to this type of process as an adaptive MCMC algorithm. Given the non-standard nature of adaptive MCMC algorithms and the given aim of sampling from a given distribution π, it is natural to ask if adaptation preserves the desired ergodicity of classical MCMC algorithms. For example, denoting $\|\cdot\|_{TV}$ the total variation norm, do we have $\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\|_{TV} \to 0$ as $n \to \infty$? The answer is "no" in general and counter-examples abound (see e.g. Andrieu and Moulines (2006), Atchade and Rosenthal (2005)). However positive ergodicity results do also exist. For example if adaptation of θ occurs at regeneration times, then ergodicity is preserved for almost any adaptation scheme (Gilks et al. (1998)). It is also now well established that if adaptation is diminishing (for example in the sense that $|\theta_n - \theta_{n-1}| \to 0$, as $n \to \infty$) then ergodicity is also preserved under mild additional assumptions (see e.g. Andrieu and Moulines (2006), Atchade and Rosenthal (2005), Rosenthal and Roberts (2007)). However, beyond ergodicity, it is still unclear how efficient adaptive MCMC are.

This paper addresses the problem of efficiency of adaptive MCMC. We consider the case where the adaptation process $\{\theta_n\}$ converges (in the mean square sense for example) to a unique nonrandom limit θ^*. Let $\{Y_n\}$ be the stationary Markov chain with transition kernel P_{θ^*} and invariant distribution π. Under some standard assumptions, we obtain a bound on the rate of convergence in total variation norm of the distribution of $(X_n, \ldots, X_{n+p})$ towards the distribution of $(Y_0, \ldots, Y_p)$ as $n \to \infty$ for any finite integer $p \geq 0$ (Theorem 2.1). This bound, which depends explicitly on the rate of convergence of θ_n to θ^*, shed some new light on adaptive MCMC processes. Theorem 2.1 implies that the process $\{X_n\}$ is asymptotically stationary (in the weak convergence sense) with stationary distribution given by the distribution of $\{Y_n\}$. If θ_n converges to θ^* fast enough, it follows as well from Theorem 2.1 that $\{X_n\}$ is asymptotically stationary in the total variation norm sense and as a result, there exists a coupling $\{\hat{X}_n, \hat{Y}_n\}$ of $\{X_n, Y_n\}$ and a finite coupling time T such that for any $n \geq T$, $\hat{X}_n = \hat{Y}_n$. Unfortunately, as we shall see, the rates required for the convergence of $\{\theta_n\}$ towards θ^* for this latter result to hold are not realistic for current stochastic approximation based implementations of adaptive MCMC.

More precisely, we pay particular attention to the case where $\{\theta_n\}$ is constructed through a stochastic approximation recursion: most existing adaptive MCMC algorithms rely on this mechanism (Haario et al. (2001), Andrieu and Moulines (2006), Atchade and Rosenthal (2005)). In particular we derive some verifiable conditions that ensure the mean square convergence of θ_n to a unique limit point θ^* and prove a bound on this rate of convergence (Theorem 3.1). These results apply for example to the adaptive Metropolis algorithm of Haario et al.

(2001) and show that the stochastic process generated by this algorithm is asymptotically stationary in the weak convergence sense with a limit distribution that is (almost) optimal.
The paper is organized as follows. In the next section we state our main result (Theorem 2.1) and briefly discuss some of its implications. The proof of Theorem 2.1 is postponed to Section 4.1. Section 3.1 is devoted to the special case of stochastic approximation updates. We first establish a Theorem 3.1 which establishes the mean square error convergence of θ_n to some θ^* under verifiable conditions. We then apply our results to the adaptive Metropolis algorithm of Haario et al. (2001) (Proposition 3.1).

2 Statement and discussion of the results

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \pi)$ be a probability space, $(\Theta, |\cdot|)$ a normed space and $\{P_\theta : \theta \in \Theta\}$ a family of transition kernels $P_\theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$. We assume that for any $A \in \mathcal{B}(\mathcal{X})$, $P_\theta(x, A)$ is measurable as a function of (θ, x). We introduce the following classical notation. If P is a transition kernel on some measure space $(E, \mathcal{E})$, for $n \geq 0$, we write P^n for the transition kernel defined recursively as $P^n(x, A) = \int_E P(x, dy) P^{n-1}(y, A)$; $P^0(x, A) = \mathbf{1}_A(x)$ where $\mathbf{1}_A(x)$ is the indicator function of set A (which we might denote $\mathbf{1}$(A) at times). Also if ν is a probability measure on $(E, \mathcal{E})$ and $f : E \to \mathbb{R}$ is a measurable function, we define $\nu P(\cdot) := \int_E \nu(dx) P(x, \cdot)$, $Pf(x) := \int_E P^n(x, dy) f(y)$ and $\nu(f) := \int_E f(y) \nu(dy)$ whenever these integrals exist. If E is a topological space, we say that E is Polish if the topology on E arises from a metric with respect to which E is complete and separable. In this case E is equipped with its Borel σ-algebra. For μ a probability measure and $\{\mu_n\}$ a sequence of probability measures on $(E, \mathcal{E})$ with E a Polish space, we say that μ_n converges weakly to μ as $n \to \infty$ and write $\mu_n \stackrel{w}{\to} \mu$ if $\int_E f(y) \mu_n(dy) \to \int_E f(y) \mu(dy)$ for any real-valued bounded continuous function f on E.
For any function $f : \mathcal{X} \to \mathbb{R}$ and $W : \mathcal{X} \to [1, \infty)$ we denote $|f|_W := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{W(x)}$, and define the set $\mathcal{L}_W := \{f,\ f : \mathcal{X} \to \mathbb{R}, |f|_W < \infty\}$. When no ambiguity is possible, we will use the piece of notation $|\cdot|$ to denote the norm on Θ and the Euclidean norm. A signed measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ can be seen as a linear functional on $\mathcal{L}_W$ with norm $\|\nu\|_W := \sup_{|f|_W \leq 1} |\nu(f)|$. For $W \equiv 1$, we obtain the total variation norm, denoted $\|\nu\|_{TV}$ hereafter. Similarly, for two transition kernels P_1 and P_2 we define $\|P_1 - P_2\|_W$ as

$$\|P_1 - P_2\|_W := \sup_{|f|_W \leq 1} |P_1 f - P_2 f|_W \ .$$

Let $\rho_n : \Theta \times \mathcal{X}^{n+1} \longrightarrow \Theta$ be a sequence of measurable functions and define the adaptive chain $\{X_n\}$ as follows: $\theta_0 = \theta \in \Theta$, $X_0 = x \in X$ and for $n \geq 1$, given $(\theta_0, X_0, \ldots, X_n)$, $\theta_n = \rho_n(\theta_0, X_0, \ldots, X_n)$ and X_{n+1} is generated from $P_{\theta_n}(X_n, \cdot)$. Without any loss of generality, we shall work with the canonical version of the process $\{X_n\}$ defined on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$ and write $\mathbb{P}$ for its distribution and $\mathbb{E}$ the expectation with respect to $\mathbb{P}$. We omit the dependence of $\mathbb{P}$ on θ_0, X_0 and $\{\rho_n\}$. Let $\mathbb{Q}_\theta$ be the distribution on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$ of a Markov chain with initial distribution π and transition kernel P_θ. When convenient, we shall write Z to denote the random process $\{Z_n\}$.
We assume the following:

(**A1**) We assume that for any $\theta \in \Theta$, P_θ has invariant distribution π and there exist a measurable function $V : \mathcal{X} \longrightarrow [1, \infty)$, a set $C \subset \mathcal{X}$, a probability measure ν such that $\nu(C) > 0$ and constants $\lambda \in (0,1)$, $b \in [0, \infty)$, $\varepsilon \in (0,1]$ such that for any $\theta \in \Theta$, $x \in \mathcal{X}$

and $A \in \mathcal{B}$,

$$P_\theta(x, A) \geq \varepsilon\nu(A)\mathbf{1}_C(x) \;, \tag{1}$$

and

$$P_\theta V(x) \leq \lambda V(x) + b\mathbf{1}_C(x) \;. \tag{2}$$

The inequality (2) of (A1) is the so-called drift condition and (1) is the so-called $(1, \varepsilon, \nu)$-minorization condition. These conditions have proved very effective in analyzing Markov chains. As pointed out in Andrieu and Moulines (2006), (A1) is sufficient to ensure that V-geometric ergodicity of the Markov chain holds uniformly in θ, *i.e.* there exist a measurable function $V : \mathcal{X} \to [1, \infty)$, $\rho \in [0, 1)$ and $C \in [0, \infty)$ such that for any $\theta \in \Theta$ and $x \in \mathcal{X}$,

$$\|P_\theta^n(x, \cdot) - \pi(\cdot)\|_V \leq CV(x)\rho^n. \tag{3}$$

For a proof, see e.g. Baxendale (2005) and the references therein.
Next, we assume that P_θ is Lipschitz (in $|||\cdot|||$-norm) as a function of θ.

(**A2**) For all $\alpha \in [0, 1]$,

$$\sup_{\substack{\theta,\theta' \in \Theta \\ \theta \neq \theta'}} \frac{|||P_\theta - P_{\theta'}|||_{V^\alpha}}{|\theta - \theta'|} < \infty, \tag{4}$$

where V is defined in (A1).

We assume that θ_n converges to some fixed element $\theta^* \in \Theta$ in the mean square sense,

(**A3**) There exist a deterministic sequence of positive real numbers $\{\alpha_n\}$, $\alpha_n \to 0$ as $n \to \infty$ and a fixed $\theta^* \in \Theta$ such that

$$\sqrt{\mathbb{E}\left[|\theta_n - \theta^*|^2\right]} = O\left(\alpha_n\right). \tag{5}$$

We assume that an optimality criterion has been defined with respect to which P_θ^* is the best possible transition kernel. Of course, in general θ^* is not known and our objective here is to investigate how well the adaptive chain $\{X_n\}$ performs with respect to the optimal chain. Let $Y = \{Y_n\}$ be the stationary Markov chain on $\mathcal{X}$ with transition kernel P_{θ^*} and initial distribution π.
For $n, p \geq 0$, n finite, we introduce the projection $s_{n,p} : \mathcal{X}^\infty \to \mathcal{X}^{p+1}$ with $s_{n,p}(w_0, w_1, \ldots) = (w_n, \ldots, w_{n+p})$. For $p = \infty$, we write s_n. If μ is a probability measure on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$, define $\mu^{(n,p)} := \mu \circ s_{n,p}^{-1}$, the image of μ by $s_{n,p}$. If $p = \infty$, we simply write $\mu^{(n)}$. The following result is fundamental. It provides us with a comparison of the distributions of $\{X_n \ldots, X_{n+p}\}$ and $\{Y_n \ldots, Y_{n+p}\}$.

Theorem 2.1. *Assume (A1-3). Let $\{i_n\} \subset \mathbb{Z}^+$ be such that for all $n \in \mathbb{Z}$, $i_n < n$. Then there exists $C \in (0, \infty)$ such that with $\rho \in (0, 1)$ as in Eq. (3) and $\{\alpha_k\}$ as in (A3) then for any $n \geq 1$, $p \geq 0$,*

$$\|\mathbb{P}^{(n,p)} - \mathbb{Q}_{\theta^*}^{(0,p)}\|_{TV} \leq C \left\{\rho^{n-i_n} + \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} + \sum_{j=n-1}^{n+p-1} \alpha_j \right\} \tag{6}$$

$$\leq C \sum_{j=n}^{n+p} \alpha_j \text{ when } \alpha_j \propto j^{-\gamma} \text{ for some } \gamma > 0 \;. \tag{7}$$

Proof. See Section 4.1. □

The bound in Theorem 2.1 implies that under suitable conditions on $\{\alpha_k\}$ any finite dimensional distribution of $\{s_n(X)\}$ converges weakly to the corresponding finite dimensional distribution of Y. As a result if $\mathcal{X}$ is Polish, and since weak convergence of finite dimensional marginals implies weak convergence of measures, we conclude that:

Corollary 2.1. *Assume that $\mathcal{X}$ is Polish. Under the assumptions of Theorem 2.1,*

$$\mathbb{P}^{(n)} \xrightarrow{w} \mathbb{Q}_{\theta^*}, \quad \text{as } n \to \infty. \tag{8}$$

When $\sum_{i\geq 1} \alpha_i < \infty$, we can strenghten the conclusion of Corollary 2.1 as follows.

Corollary 2.2. *In addition to the assumptions of Theorem 2.1, assume that $\mathcal{X}$ is Polish and that $\sum \alpha_i < \infty$. Then there exist a coupling $(\hat{X}, \hat{Y})$ of (X, Y) and a finite coupling time T such that $\hat{X}_{T+n} = \hat{Y}_{T+n}$, $n \geq 0$.*

Proof. $\sum_{i\geq 1} \alpha_i < \infty$ implies from Theorem 2.1 that $\left\|\mathbb{P}^{(n)} - \mathbb{Q}_{\theta^*}\right\|_{TV} \to 0$ and according to Theorem 2.1 of Goldstein (1979) on maximal coupling of random processes, this is equivalent to the existence of the asserted coupling . □

Remark 2.1.

1. The conclusion of Corollary 2.1 is that the adaptive chain is asymptotically stationary in the weak convergence sense with limiting distribution equal to the distribution of the "optimal" chain. When $\sum \alpha_n < \infty$, Corollary 2.2 asserts that the adaptive chain is asymptotically stationary in the total variation norm. In the latter case, the existence of the coupling is interesting as it suggests that the adaptive chain and the optimal Markov chain are essentially the same process.

2. The condition $\sum_{n\geq 1} \alpha_n < \infty$ cannot be removed in general from Corollary 2.2.

3 Adaptive chains governed by stochastic approximation

3.1 Validity of (A3) for the stochastic approximation recursion

Our main objective here is to show that (A3) holds when the family of updating equations $\{\rho_n\}$ corresponds to the popular stochastic approximation procedure. We will assume for simplicity that Θ is a compact subset of the Euclidean space $\mathbb{R}^p$ for some positive integer p and denote by $\langle\cdot,\cdot\rangle$ the inner product on $\mathbb{R}^p$. We assume that $\{\theta_n\}$ is a stochastic approximation sequence, defined as follows. Let $H : \Theta\times\mathcal{X} \to \mathbb{R}^p$ and let $\{\gamma_n\}$ be some sequence of positive real numbers. For $n \geq 0$ we recursively define the sequence, $\{(d_n, \theta_n, X_n), n \geq 0\} \in \{0, 1\}^{\mathbb{N}} \times \Theta^{\mathbb{N}} \times \mathcal{X}^{\mathbb{N}}$ as follows. Set $\theta_0 = \theta \in \Theta$, $X_0 = x \in \mathcal{X}$ and $d_0 = 0$. Given θ_n and X_n, sample $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$. If $d_n = 1$, then set $\theta_{n+1} = \theta_n$ and $d_{n+1} = 1$. Otherwise if $\theta := \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1}) \in \Theta$ then $\theta_{n+1} = \theta$ and $d_{n+1} = 0$, otherwise $\theta_{n+1} = \theta_n$ and $d_{n+1} = 1$. We define $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ and will denote $\mathbb{P}$ and $\mathbb{E}$ the probability and expectation of this process, omitting again the dependence of these probability and expectation on θ, x and $\{\gamma_n\}$.
This set-up is particularly relevant as many recently proposed adaptive MCMC algorithms rely on stochastic approximation. An extensive literature exists on stochastic approximation algorithms (see e.g. Benveniste et al. (1990), Kushner and Yin (2003) and the references therein). In order to establish our result, we will need the following definitions and assumption.

Definition 3.1. *Let $f : \Theta \times \mathcal{X} \to \mathbb{R}^q$ for some positive integer q and let $W : \mathcal{X} \to [1, \infty)$ be two functions. We say that f is W-bounded in its first argument if*

$$\sup_{\theta \in \Theta} |f(\theta, \cdot)|_W < \infty \ , \tag{9}$$

and we say that f is W-Lipschitz in its first argument if

$$\sup_{\substack{\theta, \theta' \in \Theta \\ \theta \neq \theta'}} \frac{|f(\theta, \cdot) - f(\theta', \cdot)|_W}{|\theta - \theta'|} < \infty \ . \tag{10}$$

Also define

$$h(\theta) = \int_{\mathcal{X}} H(\theta, x) \pi(dx).$$

In this section we will require the following additional assumption, specific to the stochastic approximation framework.

(**A4**) Let the function V be as in (A1). Assume that H is $V^{1/2}$-bounded and $V^{1/2}$-Lipschitz in its first argument. Assume that the equation $h(\theta) = 0$ has a unique solution $\theta^* \in \Theta$ and that there exists $\delta > 0$ such that for all $\theta \in \Theta$,

$$\langle \theta - \theta^*, h(\theta) \rangle \leq -\delta \left| \theta - \theta^* \right|^2 . \tag{11}$$

Let $\tau := \inf\{n \geq 1 : d_n = 1\}$ be the exit time from Θ, with the usual convention that $\inf\{\varnothing\} = +\infty$. The main result of this section is:

Theorem 3.1. *Assume (A1-2) and (A4), that $\{\gamma_n\}$ is non-increasing and such that there exists $\bar{\gamma} \in (\gamma_1, +\infty)$ such that*

$$\limsup_{k \to \infty} \gamma_k^{-1} \gamma_{k - \lfloor \log(\bar{\gamma}^{-1} \gamma_k) / \log(\rho) \rfloor - 1} < +\infty \ , \tag{12}$$

(where $\rho \in (0,1)$ is as in Eq. (3)) and

$$\liminf_{k \to \infty} \frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}} > -2\delta \, ,$$

where δ is as in Eq. (11). Then there exists a constant $C < +\infty$ such that for any $n \in \mathbb{N}$,

$$\mathbb{E}\left[|\theta_n - \theta^*|^2 \, \mathbf{1}(n < \tau) \right] \leq C \ \gamma_n \ .$$

Proof. See Section 4.2. □

Remark 3.1. It can be checked that any sequence $\gamma_k = \frac{A}{n^\alpha + B}$ with $0 \leq \alpha \leq 1$, satisfies (12). If $0 \leq \alpha < 1$ or $\alpha = 1$ and $2\delta A > 1$ then $\liminf_{k \to \infty} \frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}} > -2\delta$.

3.2 Application to the Adaptive Metropolis algorithm

In this section, we apply our result to the adaptive version of the Random Walk Metropolis (RWM) algorithm of Haario et al. (2001). We assume here that $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$ the p-dimensional ($p \geq 1$) Euclidian space equipped with the Euclidean topology and the associated σ-algebra $\mathcal{B}(\mathcal{X})$. Let π be the probability measure of interest and assume that π has a bounded density (also denoted π) with respect to the Lebesgue measure on $\mathcal{X}$. Let q_Σ be the density of the 0 mean Normal distribution with covariance matrix Σ,

$$q_\Sigma(x) = \det(2\pi\Sigma)^{-1/2} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right) \,, \tag{13}$$

where x^T is the transpose of x.
The RWM algorithm with target density π and proposal density q_Σ is the following. Given X_n, a 'proposal' Y is generated from $q_\Sigma(X_n, \cdot)$. Then we either 'accept' Y and set $X_{n+1} = Y$ with probability $\alpha(X_n, Y)$ or 'reject' Y and set $X_{n+1} = X_n$ with probability $1 - \alpha(X_n, Y)$ where

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right) . \tag{14}$$

Define $\mu^* := \int_{\mathcal{X}} x\pi(dx)$ the mean of π and $\Lambda^* := \int_{\mathcal{X}} xx^T \pi(dx)$ and $\Sigma^* := \Lambda^* - \mu^* (\mu^*)^T$ its covariance matrix. It is intuitively clear that the best performance should be obtained when Σ is proportional to Σ^*. In Haario et al. (2001), an adaptive algorithm has been proposed to learn Σ^* on the fly. As pointed out in Andrieu and Robert (2001), their algorithm is a particular instance of the Robbins-Monro algorithm with Markovian dynamic. We present here an equivalent alternative Robbins-Monro recursion which naturally lends itself to the application of Theorem 3.1. Let I_p be the $p \times p$ identity matrix, the algorithm we study is as follows:

Algorithm 3.1. Initialization *Choose $X_0 = x_0 \in \mathcal{X}$ the initial point. Choose $\mu_0 \in \mathcal{X}$ an initial estimate of μ^* and Λ_0 a symmetric positive matrix, an initial estimate of Λ^*, such that $\Lambda_0 - \mu_0 \mu_0^T$ is positive. Let $\varepsilon > 0$.*

Iteration *At time $n+1$ for $n \geq 0$, given $X_n \in \mathcal{X}$, $\mu_n \in \mathcal{X}$ and Λ_n a symmetric positive matrix:*

1 *Let $\Sigma_n := \Lambda_n - \mu_n \mu_n^T + \varepsilon I_p$. Generate $Y_{n+1} \sim q_{\Sigma_n}(X_n, \cdot)$;*

2 *With probability $\alpha(X_n, Y_{n+1})$ set $X_{n+1} = Y_{n+1}$; otherwise, set $X_{n+1} = X_n$;*

3 *Set*

$$\mu_{n+1} = \mu_n + \frac{1}{n+1}\left(X_{n+1} - \mu_n\right) \,, \tag{15}$$

$$\Lambda_{n+1} = \Lambda_n + \frac{1}{n+1}\left(X_{n+1} X_{n+1}^T - \Lambda_n\right) \,. \tag{16}$$

The small matrix εI_p ensures that the covariance matrix Σ_n remains positive definite, Haario et al. (2001). We write $\theta_n := (\mu_n, \Lambda_n)$, $\theta^* := (\mu^*, \Lambda^*)$ and $\Sigma^* := \Lambda^* - \mu^* (\mu^*)^T$. Let $\mathbb{P}$ be the distribution of the process (X_n) on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$ and $\mathbb{E}$ its associated expectation. As before, we omit the dependence of $\mathbb{P}$ on the initial values and other parameters of the algorithm x_0, θ_0 etc... Let also $\mathbb{Q}$ denote the distribution on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$ of the stationary Markov chain with initial distribution π and transition kernel $P_{\Sigma^* + \varepsilon I_p}$. An application of Theorems 3.1 and 2.1 give the following proposition. We omit the details.

Proposition 3.1. *The adaptive RWM algorithm described above is such that:*

(i) *there exists a constant $C \in (0,\infty)$ such that for any $n \geq 1$*

$$\|\Pr(X_n \in \cdot) - \pi\|_{TV} \leq C/n, \qquad \mathbb{E}\left[|\theta_n - \theta^*|^2\right] \leq C/n . \tag{17}$$

(ii) *for any bounded measurable $f : \mathcal{X} \to \mathbb{R}$, as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} \pi(f), \tag{18}$$

(iii) *the process is weakly consistent in the sense that,*

$$\mathbb{P}^{(n)} \xrightarrow{w} \mathbb{Q}_{\theta^*} \quad \text{as } n \to \infty , \tag{19}$$

and there exist $C \in (0,\infty)$ such that for any finite $n, p \geq 1$:

$$\left\|\mathbb{P}^{(n,p)} - \mathbb{Q}_{\theta^*}^{(0,p)}\right\|_{TV} \leq C \log\left(1 + \frac{p}{n}\right) . \tag{20}$$

Furthermore for any integer sequence $\{p_n\}$ such that $p_n = o(n)$,

$$\left\|\mathbb{P}^{(n,p_n)} - \mathbb{Q}_{\theta^*}^{(0,p_n)}\right\|_{TV} \to 0 . \tag{21}$$

Remark 3.2. Note that in the case of this linear Robbins-Monro recursion, a more precise L^2 result can also be directly obtained from the martingale decomposition in Andrieu and Moulines (2006), see also Andrieu (2004) for a discussion.

4 Proofs

4.1 Proof of Theorem 2.1

Proof. Let $\{X_i\}$ be our adaptive process and $\{Y_i\}$ the homogeneous Markov chain with transition probability P_{θ^*}. Throughout this section, $\mathcal{F}_n = \sigma(X_0, Y_0, \ldots, X_n, Y_n)$. It is sufficient to work with functions of the form $f := \prod_{i=0}^{p} f_i$ for $\{f_i : \mathcal{X} \to \mathbb{R}, |f_i| \leq 1, i = 0, \ldots, p\}$ a family of measurable functions and any $p > 1$. The proof relies on the following decomposition

$$\mathbb{E}\left[\prod_{i=0}^{p} f_i(X_{n+i}) - \prod_{i=0}^{p} f_i(Y_{n+i})\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{p} f_i(X_{n+i}) - \prod_{i=0}^{p} f_i(Y_{n+i})\middle| \mathcal{F}_{n-1}\right]\right] . \tag{22}$$

An estimate of the inner conditional expectation term is given in Proposition 4.2 below and the outer expectation operator is studied in Proposition 4.1 below as well. The combination of these results leads to

$$\left|\mathbb{E}\left[\textstyle\prod_{i=0}^{p} f_i(X_{n+i}) - \prod_{i=0}^{p} f_i(Y_{n+i})\right]\right| \leq C \left\{\rho^{n-i_n} + \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} + \sum_{j=n-1}^{n+p-1} \alpha_j\right\} \tag{23}$$

$$\leq C \sum_{j=n}^{n+p-1} \alpha_j \text{ when } \alpha_j \propto j^{-\gamma} \text{ for some } \gamma > 0 , \tag{24}$$

hence the result. □

Proposition 4.1. *Assume (A1-3). Let $g \in \mathcal{L}_{V^{1/2}}$ and $\{i_n\} \subset \mathbb{Z}^+$ be such that for all $n \in \mathbb{Z}$, $i_n < n$. Then there exists $\rho \in (0,1)$ and $C \in (0,\infty)$ such that for any $n \geq 1$,*

$$|\mathbb{E}\left[g(X_n) - g(Y_n)\right]| \leq C \ \|g\|_{V_{1/2}} \ \left\{\rho^{n-i_n} + \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)}\right\} V^{1/2}(x) \ .$$

If $\alpha_n \propto n^{-\gamma}$ for $\gamma > 0$, then there exists $C \in (0,\infty)$ such that

$$|\mathbb{E}\left[g(X_n) - g(Y_n)\right]| \leq \frac{C \ \|g\|_{V_{1/2}} \ V^{1/2}(x)}{n^\gamma} \ .$$

Proof. Let $\{X_n\}$ be our adaptive process and $\{Y_n\}$ be the time-homogeneous Markov chain with transition probability P_{θ^*}. First we have the following decomposition

$$\mathbb{E}\left[g(X_n) - g(Y_n)\right] = \mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - g(Y_n)\right] - \mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - g(X_n)\right].$$

The first term is easily dealt with since from the Markov property

$$\mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - g(Y_n)\right] = \mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - P_{\theta^*}^{n-i_n} g(Y_{i_n})\right]$$

and by Lemma 5.1 Andrieu et al. (2005),

$$\begin{aligned}\left|\mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - P_{\theta^*}^{n-i_n} g(Y_{i_n})\right]\right| &\leq C \ \|g\|_{V_{1/2}} \ \rho^{n-i_n} \ \mathbb{E}\left[V^{1/2}(X_{i_n}) + V^{1/2}(Y_{i_n})\right] \\ &\leq C \ \|g\|_{V_{1/2}} \ \rho^{n-i_n} \ V^{1/2}(x) \ ,\end{aligned}$$

For the second term we introduce the following telescoping sum decomposition,

$$\begin{aligned}\mathbb{E}\left[P_{\theta^*}^{n-i_n} g(X_{i_n}) - g(X_n)\right] &= \mathbb{E}\left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - P_{\theta^*}^{n-(j+1)} g(X_{j+1})\right] \\ &= \mathbb{E}\left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - \mathbb{E}_x\left[P_{\theta^*}^{n-(j+1)} g(X_{j+1}) | \mathcal{F}_j\right]\right] \\ &= \mathbb{E}\left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - P_{\theta_j} P_{\theta^*}^{n-(j+1)} g(X_j)\right] \\ &= \mathbb{E}\left[\sum_{j=i_n}^{n-1} \left(P_{\theta^*} - P_{\theta_j}\right) P_{\theta^*}^{n-(j+1)} g(X_j)\right] \\ &= \mathbb{E}\left[\sum_{j=i_n}^{n-1} \left(P_{\theta^*} - P_{\theta_j}\right) P_{\theta^*}^{n-(j+1)} \left(g(X_j) - \pi(g)\right).\right] \qquad (25)\end{aligned}$$

Now, for $j \in \{i_n, \ldots, n-1\}$ from Cauchy-Schwartz's inequality,

$$\begin{aligned}\left|\mathbb{E}\left[\left(P_{\theta^*} - P_{\theta_j}\right) P_{\theta^*}^{n-(j+1)} \left(g(X_j) - \pi(g)\right)\right]\right| &\leq C \ \|g\|_{V_{1/2}} \ \mathbb{E}\left[|\theta^* - \theta_j| \, \rho^{n-(j+1)} V^{1/2}(X_j)\right] \\ &\leq C \ \|g\|_{V_{1/2}} \ \rho^{n-(j+1)} \ \sqrt[2]{\mathbb{E}\left[|\theta^* - \theta_j|^2\right]} \sqrt[2]{\mathbb{E}\left[V(X_j)\right]},\end{aligned}$$

and consequently, using Lemma 5.1 Andrieu et al. (2005), we first conclude that

$$|\mathbb{E}\left[g(X_n)-g(Y_n)\right]| \leq C \;\; \|g\|_{V^{1/2}} \left\{ \rho^{n-i_n} + \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} \right\} V^{1/2}(x) \ .$$

Now in the case where $\alpha_j \propto j^{-\gamma}$ we will choose i_n in order to balance the two terms depending on n on the right hand side. To that purpose we note that,

$$n^{-\gamma} I_n \leq \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} \leq i_n^{-\gamma} I_n \text{ with } I_n \frac{1-\rho^{n-i_n}}{1-\rho} \ , \tag{26}$$

and check that the choice

$$n - i_n = \left\lfloor \frac{-\gamma \log(n)}{\log(\rho)} \right\rfloor ,$$

leads to $\sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} \sim \frac{n^{-\gamma}}{1-\rho}$ and $\rho^{n-i_n} \sim n^{-\gamma}$, which concludes the proof. □

Let $\{\phi_k : \mathcal{X} \to [-1,1],\ k = 0,\ldots,p+1\}$ be a family of functions defined as $\phi_{p+1}(x) = 1$ and for $k = p,\ldots,0$

$$\phi_k(x) = P_{\theta^*}\{\phi_{k+1} f_k\}(x) = \int_{\mathcal{X}} P_{\theta^*}(x,dy)\phi_{k+1}(y) f_k(y) \ .$$

We have the proposition,

Proposition 4.2. *Assume (A1-3). Let $\{X_i\}$ be the adaptive chain and Let $\{f_i : \mathcal{X} \to \mathbb{R}, |f_i| \leq 1, i = 0,\ldots,p\}$ be a family of measurable functions. Then there exists a constant $C \in (0,\infty)$ such that for any $n,p \in \mathbb{Z}^+$,*

$$\mathbb{E}\left[\prod_{i=0}^p f_i(X_{n+i}) - \phi_0(X_n) \middle| \mathcal{F}_{n-1}\right] \leq C \sum_{k=0}^{p} \alpha_{n-1+k} \ .$$

Proof. We have the following telescoping sum decomposition,

$$\begin{aligned} &\mathbb{E}\left[\prod_{i=0}^p f_i(X_{n+i}) - \phi_0(X_n) \middle| \mathcal{F}_{n-1}\right] \\ &\quad = \mathbb{E}\left[\sum_{k=0}^{p} \left(\phi_{k+1}(X_{n+k}) \prod_{i=0}^{k} f_i(X_{n+i}) - \phi_k(X_{n+k-1}) \prod_{i=0}^{k-1} f_i(X_{n+i}) \right) \middle| \mathcal{F}_{n-1} \right] \ . \end{aligned} \tag{27}$$

For any $k = 0,\ldots,p$, using the Markov property, one has

$$\begin{aligned} &\mathbb{E}\left[\phi_{k+1}(X_{n+k}) \prod_{i=0}^{k} f_i(X_{n+i}) - \phi_k(X_{n+k-1}) \prod_{i=0}^{k-1} f_i(X_{n+i}) \middle| \mathcal{F}_{n-1} \right] \\ &= \mathbb{E}\left[\left(\prod_{i=0}^{k-1} f_i(X_{n+i}) \right) \left(\mathbb{E}\left[\phi_{k+1}(X_{n+k}) f_k(X_{n+k}) | \mathcal{F}_{n+k-1}\right] - P_{\theta^*}\left\{\phi_{k+1} f_k\right\}(X_{n+k-1}) \right) \middle| \mathcal{F}_{n-1} \right] \\ &= \mathbb{E}\left[\left(\prod_{i=0}^{k-1} f_i(X_{n+i}) \right) \left(P_{\theta_{n+k-1}}\left(\phi_{k+1} f_k\right)(X_{n+k-1}) - P_{\theta^*}\left(\phi_{k+1} f_k\right)(X_{n+k-1}) \right) \middle| \mathcal{F}_{n-1} \right] \\ &= \mathbb{E}\left[\left(\prod_{i=0}^{k-1} f_i(X_{n+i}) \right) \left(P_{\theta_{n+k-1}} - P_{\theta^*} \right) \left(\phi_{k+1} f_k\right)(X_{n+k-1}) \middle| \mathcal{F}_{n-1} \right] \ . \end{aligned}$$

Finally,

$$\left|\sum_{k=0}^{p}\mathbb{E}\left[\left(\prod_{i=0}^{k-1}f_i(X_{n+i})\right)\ \left(P_{\theta_{n+k-1}}-P_{\theta^*}\right)(\phi_{k+1}f_k)\,(X_{n+k-1})\middle|\,\mathcal{F}_{n-1}\right]\right|$$
$$\leq \sum_{k=0}^{p}\left|\mathbb{E}\left[\left(\prod_{i=0}^{k-1}f_i(X_{n+i})\right)\ \left(P_{\theta_{n+k-1}}-P_{\theta^*}\right)(\phi_{k+1}f_k)\,(X_{n+k-1})\middle|\,\mathcal{F}_{n-1}\right]\right|$$
$$\leq C\sum_{k=0}^{p}\mathbb{E}\left[\ |\theta_{n+k-1}-\theta^*|\right]\leq C\sum_{k=0}^{p}\alpha_{n+k-1}=C\sum_{k=n-1}^{n+p-1}\alpha_k\ .$$

□

4.2 Proof of Theorem 3.1

Proof. In what follows C is a finite universal constant, whose value might change upon each appearance. With for any $n\geq 0$ $\Delta_n:=\theta_n-\theta^*$ we have

$$\Delta_{n+1}\mathbf{1}(n+1<\tau)=\left[\Delta_n+\gamma_{n+1}H(\theta_n,X_{n+1})\right]\mathbf{1}(n+1<\tau)\ .$$

First, since $\mathbf{1}(n+1<\tau)\leq\mathbf{1}(n<\tau)$ we have for any $n\geq 0$,

$$\begin{aligned}
&|\Delta_{n+1}|^2\,\mathbf{1}(n+1<\tau)\\
&\leq|\Delta_n|^2\,\mathbf{1}(n<\tau)+\gamma_{n+1}^2\,|H(\theta_n,X_{n+1})|^2\,\mathbf{1}(n<\tau)+2\gamma_{n+1}\left\langle\Delta_n,H(\theta_n,X_{n+1})\right\rangle\mathbf{1}(n<\tau)\\
&\quad\leq|\Delta_n|^2\,\mathbf{1}(n<\tau)+\gamma_{n+1}^2\,|H(\theta_n,X_{n+1})|^2\,\mathbf{1}(n<\tau)+2\gamma_{n+1}\left\langle\Delta_n,h(\theta_n)\right\rangle\mathbf{1}(n<\tau)\\
&\qquad+2\gamma_{n+1}\left\langle\Delta_n,H(\theta_n,X_{n+1})-h(\theta_n)\right\rangle\mathbf{1}(n<\tau)\,.
\end{aligned}$$

From assumptions (A1) and (A4), and *e.g.* Lemma 5.1 in Andrieu et al. (2005) we deduce that,

$$\sup_{n\geq 0}\mathbb{E}\left[|H(\theta_n,X_{n+1})|^2\,\mathbf{1}(n<\tau)\right]\leq C\sup_{n\geq 0}\mathbb{E}\left[V(X_{n+1})\mathbf{1}(n<\tau)\right]<+\infty\ ,\tag{28}$$
$$\mathbb{E}\left[\left\langle\Delta_n,h(\theta_n)\right\rangle\mathbf{1}(n<\tau)\right]\leq-\delta\mathbb{E}\left[|\Delta_n|^2\,\mathbf{1}(n<\tau)\right]\ .\tag{29}$$

From Proposition 4.3 we have that

$$\left|\mathbb{E}\left[\left\langle\Delta_n,H(\theta_n,X_{n+1})-h(\theta_n)\right\rangle\mathbf{1}(n<\tau)\right]\right|\leq C\gamma_{n+1}V^{1/2}(x)\,.$$

Consequently there exists a constant C_1 such that for $n\geq 1$,

$$\mathbb{E}\left[|\Delta_{n+1}|^2\,\mathbf{1}(n+1<\tau)\right]\leq(1-2\delta\gamma_{n+1})\,\mathbb{E}\left[|\Delta_n|^2\,\mathbf{1}(n<\tau)\right]+C_1\gamma_{n+1}^2\ ,$$

and we conclude using Lemma 23 p. 245 in Benveniste et al. (1990). □

We first recall the following fundamental lemma, which can be found in the proof of Proposition 3 in Andrieu and Moulines (2006).

Lemma 4.1. *Assume (A1-2). Then there exists $C \in (0,+\infty)$ such that for any $\theta,\theta' \in \Theta$, $n \geq 1$ and any $g \in \mathcal{L}_{V^r}$ for any $r \in [0,1]$,*

$$|P_\theta^n g - P_{\theta'}^n g|_{V^r} \leq C\,|g|_{V^r}\;\; n\rho^{n-1}\;|\theta-\theta'|\;.$$

For any $x \in \mathbb{R}$, let us denote $\lfloor x \rfloor$ the largest integer such that $\lfloor x \rfloor \leq x$. For any $g_\theta(x) : \Theta \times \mathcal{X} \to \mathbb{R}^d$ denote for any $\theta \in \Theta$, $\bar{g}_\theta := \pi\,(g_\theta)$.

Proposition 4.3. *Assume that $\{\gamma_k\}$ is nonincreasing, such that $\lim_{k\to\infty}\gamma_k = 0$ and that there exists $\bar\gamma \in (0,+\infty)$ such that*

$$\limsup_{k\to\infty}\;\gamma_k^{-1}\gamma_{k-\lfloor \log(\bar\gamma^{-1}\gamma_k)/\log(\rho)\rfloor-1} < +\infty\;, \tag{30}$$

where $\rho \in (0,1)$ is as in Eq. (3). Assume that $\sup_{\theta\in\Theta}|H(\theta,\cdot)|_{V^{1/2}} < \infty$. Then there exists a constant $C \in (0,+\infty)$ such that for any $g_\theta(x) : \Theta \times \mathcal{X} \to \mathbb{R}^d$ such that $\sup_{\theta\in\Theta}|g_\theta|_{V^{1/2}} < \infty$ any $x \in \mathcal{X}$ and any $k \geq 1$,

$$\left|\mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - \bar{g}_{\theta_{k-1}})\mathbf{1}(\tau > k)\right]\right| \leq C\;\sup_{\theta\in\Theta}|g_\theta|_{V^{1/2}}\;\;\gamma_k\;V(x)\;.$$

Proof. We introduce for integers i and k such that $0 \leq i < k$ the following decomposition,

$$\begin{aligned}&\mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - \bar{g}_{\theta_{k-1}})\mathbf{1}(\tau > k)\right]\\ &= \mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i}g_{\theta_{k-1}}(X_i))\mathbf{1}(\tau > k)\right] + \mathbb{E}\left[(P_{\theta_i}^{k-i}g_{\theta_{k-1}}(X_i) - \bar{g}_{\theta_{k-1}})\mathbf{1}(\tau > k)\right]\;.\end{aligned} \tag{31}$$

We consider the first term and use the following decomposition,

$$\begin{aligned}&|\mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i}g_{\theta_{k-1}}(X_i))\mathbf{1}(\tau > k)\right]|\\ &\qquad\leq \sum_{j=1}^{k-i}\left|\mathbb{E}\left[(P_{\theta_{k-j+1}}^{j-1}g_{\theta_{k-1}}(X_{k-j+1}) - P_{\theta_{k-j}}^{j}g_{\theta_{k-1}}(X_{k-j}))\mathbf{1}(\tau > k-j+1)\right]\right|\\ &\leq \sum_{j=1}^{k-i}\left|\mathbb{E}\left[\mathbb{E}\left[(P_{\theta_{k-j+1}}^{j-1}g_{\theta_{k-1}}(X_{k-j+1}) - P_{\theta_{k-j}}^{j-1}g_{\theta_{k-1}}(X_{k-j+1}))\mathbf{1}(\tau > k-j+1)\right]|\mathcal{F}_{k-j}\right]\right|\;.\end{aligned} \tag{32}$$

Now for $j = 1,\ldots,k-i$,

$$\begin{aligned}&\left|\mathbb{E}\left[(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1})g_{\theta_{k-1}}(X_{k-j+1})\mathbf{1}(\tau > k-j+1)\right]\right|\\ &\qquad= \left|\mathbb{E}\left[\mathbb{E}\left[(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1})g_{\theta_{k-1}}(X_{k-j+1})|\mathcal{F}_{k-j+1}\right]\mathbf{1}(\tau > k-j+1)\right]\right|\\ &\qquad= \left|\mathbb{E}\left[(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1})\left\{\mathbb{E}\left[g_{\theta_{k-1}}(\cdot)|\mathcal{F}_{k-j+1}\right]\right\}(X_{k-j})\,\mathbf{1}(\tau > k-j+1)\right]\right|\;.\end{aligned} \tag{33}$$

Consequently we apply Lemma 4.1 to each term in the sum in Eq. (32), which for $0 \leq i < k$ leads to,

$$\left|\mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i}g(X_i))\mathbf{1}(\tau > k)\right]\right| \leq C\;\sup_{\theta\in\Theta}|g_\theta|_{V^{1/2}}\;\sum_{j=1}^{k-i-1}\;j\rho^j\gamma_{k-j}\mathbb{E}\left[V(X_{k-j})\mathbf{1}(\tau > k-j)\right]\;.$$

This, together with Lemma 4.1 in Andrieu et al. (2005), implies that

$$\left|\mathbb{E}\left[(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i} g(X_i))\mathbf{1}(\tau > k)\right]\right| \leq C \sup_{\theta\in\Theta} |g_\theta|_{V^{1/2}} \; V(x) \sum_{j=1}^{k-i-1} j\rho^j \gamma_{k-j} \; ,$$

which combined with Eq. (31) gives

$$\left|\mathbb{E}\left[g_{\theta_{k-1}}(X_k) - \bar{g}_{\theta_{k-1}}\right]\right| \leq C \sup_{\theta\in\Theta} |g_\theta|_{V^{1/2}} \left[\rho^{k-i} + \sum_{j=1}^{k-i-1} j\rho^j \gamma_{k-j}\right] V(x) \; .$$

Let $k_0 := \inf\{k : \gamma_k < \rho\bar{\gamma}\} < +\infty$ where ρ is as in Eq. (3), and for $k \geq k_0$ let

$$i_k := k - \left\lfloor \frac{\log(\bar{\gamma}^{-1}\gamma_k)}{\log(\rho)} \right\rfloor \; ,$$

and $i_k := 0$ for $k < k_0$. Then, since $\{\gamma_k\}$ is non increasing,

$$\rho^{k-i_k} + \sum_{j=1}^{k-i_k-1} j\rho^j \gamma_{k-j} \leq \bar{\gamma}^{-1}\gamma_k + \gamma_{k+1-\lfloor \log(\bar{\gamma}^{-1}\gamma_k)/\log(\rho)\rfloor} \sum_{j=1}^{+\infty} j\rho^j \; ,$$

and the result follows from Eq. (12). $\square$

References

Andrieu, C. (2004). Discussion, ordinary meeting on inverse problems, wednesday 10th december, 2003, london. *Journal of the Royal Statistical Society B* **66** 627–652.

Andrieu, C. and Moulines, E. (2006). On the ergodicity Properties of some Adaptive MCMC Algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070

Andrieu, C., Moulines, E. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization* **44** 283–312. MR2177157

Andrieu, C. and Robert, C. P. (2001). Controlled mcmc for optimal sampling. *Technical report, Université Paris Dauphine, Ceremade 0125* .

Atchade, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithm. *Bernoulli* **11** 815–828. MR2172842

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic markov chains. *Annals of Applied Probability* **15** 700–738. MR2114987

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic approximations*. Applications of Mathematics, Springer, Paris-New York.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds.) (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics, Chapman & Hall, London. MR1397966

Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93** 1045–1054. MR1649199

Goldstein, S. (1979). Maximal coupling. *Z. Wahrsch. Verw. Gebiete* **46** 193–204. MR0516740

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

Kushner, K. and Yin, Y. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer, Springer-Verlag, New-York. MR1993642

Rosenthal, J. S. and Roberts, G. O. (2007). Coupling and ergodicity of adaptive mcmc. *Journal of Applied Probablity* **44** 458–475.