

On the efficiency of adaptive MCMC algorithms

Christophe Andrieu¹ and Yves F. Atchadé²

(October 2005)

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 process and the “optimal” (nonadaptive) MCMC process share identical 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). We also propose a new class of adaptive MCMC algorithms, called *quasi-perfect adaptive MCMC* which possesses appealing theoretical and practical properties, as demonstrated through numerical simulations.

Key words: Adaptive Markov chains, Coupling, Markov Chain Monte Carlo, Metropolis Algorithm, Stochastic Approximation, Rate of convergence.

MSC Numbers: 60J35, 60J22, 65C40

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 choice 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).

An attractive solution to this problem, which has recently received attention, consists of using so-called adaptive MCMC 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

¹Department of Mathematics, University of Bristol, email: c.andrieu@bris.ac.uk

²Department of Mathematics and Statistics, University of Ottawa, email: yatchade@uottawa.ca

and Robert (2001), Andrieu and Moulines (2005) 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 $\{\rho_n : \Theta \times \mathcal{X}^n \rightarrow \Theta, \text{ for } n \geq 0\}$ 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 $(\theta_0, X_0, \dots, X_{n-1})$, and $\theta_{n-1} = \rho_{n-1}(\theta_0, X_0, \dots, X_{n-1})$ (with the convention that $\rho_0(\theta, X) = \theta$), X_n is generated according to $P_{\theta_{n-1}}(X_{n-1}, \cdot)$ and $\theta_n = \rho_n(\theta_0, X_0, \dots, X_n)$. 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 $\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\|_{TV} \rightarrow 0$ as $n \rightarrow \infty$? The answer is “no” in general and counter-examples abound (see *e.g.* Andrieu and Moulines (2005), 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 $|\theta_n - \theta_{n-1}| \rightarrow 0$, as $n \rightarrow \infty$) then ergodicity is also preserved under mild additional assumptions (see *e.g.* Andrieu and Moulines (2005), Atchade and Rosenthal (2005), Rosenthal and Roberts (2005)). However, beyond ergodicity, it is still unclear how efficient adaptive MCMC are.

This note 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 . We show that the adaptive chain $\{X_n\}$ and the optimal Markov chain $\{Y_n\}$ share identical asymptotic properties. For some class of measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$, we show (Theorem 2.3) that $\sum_{k=1}^n f(X_k)$ satisfies a central limit theorem (CLT) with asymptotic variance equal to the asymptotic variance in the CLT of $\sum_{k=1}^n f(Y_k)$. In words, the Monte Carlo estimates from the adaptive MCMC algorithm and the “optimal” MCMC algorithm have the same statistical efficiency. We also show that the process $\{X_n\}$ is asymptotically stationary (in the weak convergence sense) with stationary distribution given by the distribution of $\{Y_n\}$. We obtain this result by showing

that for any finite integer $p \geq 0$ the distribution of (X_n, \dots, X_{n+p}) converges to that of (Y_0, \dots, Y_p) in total variation norm as $n \rightarrow \infty$, with a rate of convergence that depends explicitly on the rate of convergence of θ_n to θ^* (Theorem 2.1). Theorem 2.1 also implies that if θ_n converges to θ^* fast enough, then $\{X_n\}$ is asymptotically stationary in the total variation norm sense and 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 $n \geq T$, $\hat{X}_n = \hat{Y}_n$. As a result the asymptotic properties of the optimal process $\{Y_n\}$ are in this case automatically inherited by the apparently more complicated adaptive process $\{X_n\}$. 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 (2005), 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.

In order to address the limitations of current adaptive MCMC algorithms, we introduce a new scheme, called *quasi-perfect adaptive sampler*, for which the conditions for the existence of the aforementioned coupling of $\{X_n, Y_n\}$ are satisfied. We demonstrate through numerical simulations the interest of our approach.

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 5.1. 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 3.2). In Section 4, we introduce our new quasi perfect adaptive sampler and an application to the case of the adaptive Metropolis algorithm of Haario et al. (2001), together with numerical simulations.

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 $P_\theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \rightarrow [0, 1]$. 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 $f : E \rightarrow \mathbb{R}$ 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 $n \rightarrow \infty$ and write $\mu_n \xrightarrow{w} \mu$ if $\int_E f(y)\mu_n(dy) \rightarrow \int_E f(y)\mu(dy)$ for any real-valued bounded continuous function f on E .

For any function $f : \mathcal{X} \rightarrow \mathbb{R}$ and $W : \mathcal{X} \rightarrow [1, \infty)$ we denote $|f|_W := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{W(x)}$, and define the set $\mathcal{L}_W := \{f, f : \mathcal{X} \rightarrow \mathbb{R}, |f|_W < \infty\}$. 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 $\rho_n : \Theta \times \mathcal{X}^{n+1} \rightarrow \Theta$ 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 $(\theta_0, X_0, \dots, X_n)$, $\theta_n = \rho_n(\theta_0, X_0, \dots, X_n)$ 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}_θ 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 $V : \mathcal{X} \rightarrow [1, \infty)$, 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 $A \in \mathcal{B}$,

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

and

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

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 (2005), (A1) is sufficient to ensure that V -geometric ergodicity of the Markov chain holds uniformly in θ , *i.e.* there exist a measurable function $V : \mathcal{X} \rightarrow [1, \infty)$, $\rho \in [0, 1)$ and $C \in [0, \infty)$ such that for any $\theta \in \Theta$ and $x \in \mathcal{X}$,

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

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 $\alpha \in [0, 1]$,

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

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 $\{\alpha_n\}$, $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$ and a fixed $\theta^* \in \Theta$ such that

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

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 $s_{n,p} : \mathcal{X}^\infty \rightarrow \mathcal{X}^{p+1}$ with $s_{n,p}(w_0, w_1, \dots) = (w_n, \dots, w_{n+p})$. 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 $\{X_n \dots, X_{n+p}\}$ and $\{Y_n \dots, Y_{n+p}\}$.

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 $n \geq 1$, $p \geq 0$,*

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

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

Proof. See Section 5.1. □

2.1 Asymptotic stationarity

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,*

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

When $\sum_{i \geq 1} \alpha_i < \infty$, we can strengthen 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 for $n \geq T$, $\hat{X}_n = \hat{Y}_n$.*

Proof. $\sum_{i \geq 1} \alpha_i < \infty$ implies from Theorem 2.1 that $\|\mathbb{P}^{(n)} - \mathbb{Q}_{\theta^*}\|_{TV} \rightarrow 0$ 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. One can make the following comments:

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. In this case, we say that $\{X_n\}$ is *weakly efficient*. When $\sum \alpha_n < \infty$, Corollary 2.2 asserts that the adaptive chain is asymptotically stationary in the total variation norm. In that case, we say that $\{X_n\}$ is *strongly efficient*.

2. When strong efficiency holds, $\{X_n\}$ and $\{Y_n\}$ have the same asymptotics and any limit result which holds for the Markov chain $\{Y_n\}$ will also be valid for the adaptive chain $\{X_n\}$. For example, if a central limit theorem (CLT) holds for the sequence $\{f(Y_n)\}$ for some function f , then $\{f(X_n)\}$ will also satisfy a CLT with the same asymptotic variance.

The next result shows that the condition $\sum_{n \geq 1} \alpha_n < \infty$ cannot be removed from the assumptions that lead to the conclusion of Corollary 2.2.

Theorem 2.2. *Assume that (A1-3) hold. If $\sum_{n \geq 1} \alpha_n = \infty$ then $\{X_n\}$ may fail to be strongly efficient.*

Proof. We give an example where \mathcal{X} is Polish, (A1-3) hold, $\sum_{n \geq 1} \alpha_n = \infty$ and $\{X_n\}$ is weakly efficient but not strongly efficient. Let us consider $\mathcal{X} = [0, 1]$ equipped with its Borel σ -algebra and the Lebesgue measure. Let π be the uniform distribution on \mathcal{X} . For any $a > 0$ define $q_a(x) = \frac{1}{a} \mathbf{1}_{(-a/2, a/2)}(x)$. Fix $\delta \in (0, 1/2)$ and let $\{\delta_n\}$ be such that $\delta_n \rightarrow \delta$ and $\delta_n > \delta$. Consider the following nonhomogeneous Markov chain: $X_0 \in \mathcal{X}$ and for $n \geq 1$, $X_n \sim P_n(X_{n-1}, \cdot)$, where P_n is the transition kernel of the Metropolis algorithm with invariant density π and proposal probability density $q_{\delta_n}(x, y) = q_{\delta_n}(y - x)$.

Clearly assumptions (A1-3) are all satisfied here: P_n is uniformly (both in n and the initial condition) geometrically ergodic; for $\delta \geq \delta'$, $\sup_{x,A} |P_\delta(x, A) - P_{\delta'}(x, A)| \leq 2(1 - \delta'/\delta)$ and naturally $\delta_n \rightarrow \delta$. As a consequence $\mathbb{P}(X_n \in A) \rightarrow \pi(A)$ as $n \rightarrow \infty$ for any measurable A . Let $\{Y_n\}$ be the stationary optimal homogeneous Markov chain with proposal density q_δ .

For $\alpha > 0$ and $x \in \mathbb{R}$, write $A_\alpha(x) = (x - \alpha/2, x + \alpha/2)$ and define for $\delta > 0$ as above and any integer $k \geq 1$ the set

$$A_k := \{(x_1, \dots, x_k) \in \mathcal{X}^k : x_1 \in A_\delta(1/2), x_2 \in A_\delta(x_1), \dots, x_k \in A_\delta(x_{k-1})\}. \quad (9)$$

Then for any $n, k \geq 1$ we have:

$$\Pr((Y_n, \dots, Y_{n+k}) \in A_{k+1}) = \delta. \quad (10)$$

However

$$\begin{aligned} \Pr((X_n, \dots, X_{n+k}) \in A_{k+1}) &\leq \Pr(X_n \in A_\delta(1/2)) \prod_{i=n}^{n+k} \left(\frac{1}{2} + \frac{\delta}{2\delta_i} \right) \\ &\leq \prod_{i=n}^{n+k} (1 - a_i), \end{aligned} \quad (11)$$

with $\delta_i = \frac{\delta}{1-2a_i}$, where $\{a_n\}$ converges to zero and satisfies $0 < a_n < 1/2$. It is easily seen that $\{|\delta_n - \delta|\}$ and $\{a_n\}$ are equivalent sequences.

Now, if $\sum_{i \geq 1} a_i = \infty$, then for any n , we can always make (11) arbitrarily small by taking k sufficiently large, while (10) remains equal to δ . Consequently Corollary 2.2 cannot hold. \square

2.2 Central limit theorem

There is yet another way in which the adaptive chain $\{X_n\}$ can be efficient: a Monte Carlo estimate from the adaptive algorithm may be asymptotically as efficiency as an estimate from the “optimal” Markov chain. More precisely, we show here that additive functionals of $\{X_n\}$ satisfy a central limit theorem (CLT) with asymptotic variance equal to the asymptotic variance in the “optimal” Markov chain. We derive this result under the additional assumption that the adaptation is diminishing ($|\theta_n - \theta_{n-1}| \rightarrow 0$ in the appropriate sense) but with no particular assumption on the dynamics of $\{\theta_n\}$. The proof follows Andrieu and Moulines (2005) where a CLT is obtained for an adaptive chain based on stochastic approximation algorithm with reprojction on random boundaries. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be some measurable function. Without any loss of generality, we assume that $\pi(f) = 0$. Define $S_n(f) = \sum_{k=1}^n f(X_k)$ and $\sigma_*^2(f) = \pi(f^2) + 2 \sum_{k=1}^{\infty} \pi(f P_{\theta_*^k} f)$. As a result of the drift and minorization conditions on P_{θ_*} (assumption (A1)), it is well known that $\sigma_*^2(f) < \infty$ and if $\sigma_*^2(f) > 0$ then $\{f(Y_n)\}$ satisfies a CLT: $\frac{1}{\sqrt{n}} \sum_{k=1}^n f(Y_k) \xrightarrow{w} Z$ where $Z \sim \mathcal{N}(0, \sigma_*^2(f))$ (see e.g. Galin (2004) and the references therein). Such CLT can be extended to the adaptive chain $\{X_n\}$.

We assume that:

(A4) There exist $\beta \in [0, 1/4)$, $\lambda \in (1/2, 1]$, $C_1 < \infty$ and a sequence of positive number $\{\gamma_n\}$, $\gamma_n = O(n^{-\lambda})$ such that:

$$|\theta_n - \theta_{n-1}| \leq C_1 \gamma_n V^\beta(X_n), \mathbb{P} - a.s. \quad (12)$$

where V is defined in (A1).

Theorem 2.3. *Assume that (A1-3) and (A4) hold. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function such that $|f| \leq V^\beta$ and $\pi(f) = 0$. If $\sigma_*^2(f) > 0$ then*

$$\frac{1}{\sqrt{n}} S_n(f) \xrightarrow{w} Z \quad \text{as } n \rightarrow \infty, \quad (13)$$

where Z is a random variable with distribution $\mathcal{N}(0, \sigma_*^2(f))$.

Proof. See Section 5.2. □

Remark 2.2. It is known (see Andrieu and Moulines (2005), Atchade (2005)) that under (A1-2) and (A4) $\{f(X_n)\}$ satisfies a strong law of large numbers (SLLN) hold for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ with $|f| \leq V^{2\beta}$:

$$\frac{1}{n} \sum_{k=1}^n f(X_k) \longrightarrow \pi(f), \mathbb{P} - a.s., \text{ as } n \rightarrow \infty. \quad (14)$$

2.3 Comments and Discussions

In conclusion, we have shown in this section that if an adaptive MCMC is such that $\theta_n \rightarrow \theta^*$, then it is weakly efficient and its Monte Carlo estimates are statistically efficient in estimating $\pi(f)$ for a large class of functions $f : \mathcal{X} \rightarrow \mathbb{R}$. If in addition, the convergence of θ_n to θ^* is fast enough then we obtain a strong efficiency and the two chains share the same asymptotics. In Section 3, we develop the L^2 rate of convergence of θ_n to θ^* in the stochastic approximation framework. The result suggests that adaptive MCMC algorithms governed by stochastic approximation are weakly efficient and satisfy a CLT but are not strongly efficient.

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 $H : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^p$ 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 $\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n)$ 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. Define the function $h(\theta) = \int_{\mathcal{X}} H(\theta, x)\pi(dx)$. Stochastic approxima-

tion 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 $H(\theta, x)$ are available. This is better seen if, defining $\varepsilon_{n+1} = H(\theta_n, X_{n+1}) - h(\theta_n)$, we rewrite θ_n as:

$$\theta_{n+1} = \theta_n + \gamma_{n+1}h(\theta_n) + \gamma_{n+1}\varepsilon_{n+1} . \quad (15)$$

An extensive literature exists on these 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 $f : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^q$ for some positive integer q and let $W : \mathcal{X} \rightarrow [1, \infty)$ be two functions. We say that f is W -bounded in its first argument if*

$$\sup_{\theta \in \Theta} |f(\theta, \cdot)|_W < \infty , \quad (16)$$

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

$$\sup_{\substack{\theta, \theta' \in \Theta \\ \theta \neq \theta'}} \frac{|f(\theta, \cdot) - f(\theta', \cdot)|_W}{|\theta - \theta'|} < \infty . \quad (17)$$

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

(A5) 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 $\theta \in \Theta$,

$$\langle \theta - \theta^*, h(\theta) \rangle \leq -\delta |\theta - \theta^*|^2 . \quad (18)$$

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

Theorem 3.1. *Assume (A1-2) and (A5), that $\{\gamma_n\}$ is non-increasing and such that there exists $\bar{\gamma} \in (\gamma_1, +\infty)$ such that*

$$\limsup_{k \rightarrow \infty} \gamma_k^{-1} \gamma_{k - \lfloor \log(\bar{\gamma}^{-1} \gamma_k) / \log(\rho) \rfloor - 1} < +\infty , \quad (19)$$

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

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

where δ is as in Eq. (4). Then there exists a constant $C < +\infty$ such that for any $n \in \mathbb{N}$,

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

Remark 3.1. It can be checked that any sequence $\gamma_k = \frac{A}{n^{\alpha+B}}$ with $0 \leq \alpha \leq 1$, satisfies (19). If $0 \leq \alpha < 1$ or $\alpha = 1$ and $2\delta A > 1$ then $\liminf_{k \rightarrow \infty} \frac{1}{\gamma_k} - \frac{1}{\gamma_{k+1}} > -2\delta$.

Proof of Theorem 3.1. 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

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

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

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

From assumptions (A1) and (A5), and *e.g.* Lemma 5.1 in Andrieu et al. (2005) we deduce that,

$$\sup_{n \geq 0} \mathbb{E} [|H(\theta_n, X_{n+1})|^2 \mathbf{1}(n < \tau)] \leq C \sup_{n \geq 0} \mathbb{E} [V(X_{n+1}) \mathbf{1}(n < \tau)] < +\infty , \quad (20)$$

$$\mathbb{E} [\langle \Delta_n, h(\theta_n) \rangle \mathbf{1}(n < \tau)] \leq -\delta \mathbb{E} [|\Delta_n|^2 \mathbf{1}(n < \tau)] . \quad (21)$$

From Proposition 3.1 we have that

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

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

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

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

We first recall the following fundamental lemma, which can be found in the proof of Proposition 3 in Andrieu and Moulines (2005).

Lemma 3.1. *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 $[x]$ the largest integer such that $[x] \leq x$. For any $g_\theta(x) : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^d$ denote for any $\theta \in \Theta$, $\bar{g}_\theta := \pi(g_\theta)$.

Proposition 3.1. *Assume that $\{\gamma_k\}$ is nonincreasing, such that $\lim_{k \rightarrow \infty} \gamma_k = 0$ and that there exists $\bar{\gamma} \in (0, +\infty)$ such that*

$$\limsup_{k \rightarrow \infty} \gamma_k^{-1} \gamma_{k - \lfloor \log(\bar{\gamma}^{-1} \gamma_k) / \log(\rho) \rfloor - 1} < +\infty, \quad (22)$$

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 $g_\theta(x) : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^d$ such that $\sup_{\theta \in \Theta} |g_\theta|_{V^{1/2}} < \infty$ any $x \in \mathcal{X}$ and any $k \geq 1$,

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

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

$$\begin{aligned} & \mathbb{E} [(g_{\theta_{k-1}}(X_k) - \bar{g}_{\theta_{k-1}}) \mathbf{1}(\tau > k)] \\ &= \mathbb{E} [(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i} g_{\theta_{k-1}}(X_i)) \mathbf{1}(\tau > k)] + \mathbb{E} [(P_{\theta_i}^{k-i} g_{\theta_{k-1}}(X_i) - \bar{g}_{\theta_{k-1}}) \mathbf{1}(\tau > k)]. \end{aligned} \quad (23)$$

We consider the first term and use the following decomposition,

$$\begin{aligned} & |\mathbb{E} [(g_{\theta_{k-1}}(X_k) - P_{\theta_i}^{k-i} g_{\theta_{k-1}}(X_i)) \mathbf{1}(\tau > k)]| \\ & \leq \sum_{j=1}^{k-i} \left| \mathbb{E} [(P_{\theta_{k-j+1}}^{j-1} g_{\theta_{k-1}}(X_{k-j+1}) - P_{\theta_{k-j}}^j g_{\theta_{k-1}}(X_{k-j})) \mathbf{1}(\tau > k - j + 1)] \right| \\ & \leq \sum_{j=1}^{k-i} \left| \mathbb{E} \left[\mathbb{E} [(P_{\theta_{k-j+1}}^{j-1} g_{\theta_{k-1}}(X_{k-j+1}) - P_{\theta_{k-j}}^{j-1} g_{\theta_{k-1}}(X_{k-j+1})) \mathbf{1}(\tau > k - j + 1)] | \mathcal{F}_{k-j} \right] \right|. \end{aligned} \quad (24)$$

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

$$\begin{aligned} & \left| \mathbb{E} [(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1}) g_{\theta_{k-1}}(X_{k-j+1}) \mathbf{1}(\tau > k - j + 1)] \right| \\ &= \left| \mathbb{E} \left[\mathbb{E} [(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1}) g_{\theta_{k-1}}(X_{k-j+1}) | \mathcal{F}_{k-j+1}] \mathbf{1}(\tau > k - j + 1) \right] \right| \\ &= \left| \mathbb{E} [(P_{\theta_{k-j+1}}^{j-1} - P_{\theta_{k-j}}^{j-1}) \{ \mathbb{E} [g_{\theta_{k-1}}(\cdot) | \mathcal{F}_{k-j+1}] \} (X_{k-j}) \mathbf{1}(\tau > k - j + 1)] \right|. \end{aligned} \quad (25)$$

Consequently we apply Lemma 3.1 to each term in the sum in Eq. (24), which for $0 \leq i < k$ leads to,

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

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

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

which combined with Eq. (23) gives

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

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. (19). \square

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 ,

$$q_{\Sigma}(x) = \det(2\pi\Sigma)^{-1/2} \exp\left(-\frac{1}{2}x^T \Sigma^{-1}x\right), \quad (26)$$

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

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right). \quad (27)$$

Define $\mu^* := \int_{\mathcal{X}} x \pi(dx)$ the mean of π and $\Lambda^* := \int_{\mathcal{X}} x x^T \pi(dx)$ 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

$$\mu_{n+1} = \mu_n + \frac{1}{n+1} (X_{n+1} - \mu_n) , \quad (28)$$

$$\Lambda_{n+1} = \Lambda_n + \frac{1}{n+1} (X_{n+1}X_{n+1}^T - \Lambda_n) . \quad (29)$$

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 $P_{\Sigma^* + \varepsilon I_p}$. For a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$, define $\sigma_*^2(f) = \pi(f^2) + 2 \sum_{i=1}^{\infty} \pi [f P_*^i f]$, where $P_* = P_{\Sigma^* + \varepsilon I_p}$. We have:

Proposition 3.2. *The adaptive RWM algorithm described above is such that:*

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

$$\|\mathbb{P}(X_n) - \pi\|_{TV} \leq C/n, \quad \mathbb{E} [|\theta_n - \theta^*|^2] \leq C/n . \quad (30)$$

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

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{a.s.} \pi(f), \quad \text{and if } \sigma_*^2(f) > 0, \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n [f(X_i) - \pi(f)] \xrightarrow{w} Z, \quad (31)$$

where $Z \sim \mathcal{N}(0, \sigma_*^2(f))$.

(iii) *the process is weakly consistent,*

$$\mathbb{P}^{(n)} \xrightarrow{w} \mathbb{Q}_{\theta^*} \quad \text{as } n \rightarrow \infty , \quad (32)$$

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

$$\left\| \mathbb{P}^{(n,p)} - \mathbb{Q}_{\theta^*}^{(0,p)} \right\|_{TV} \leq C \log \left(1 + \frac{p}{n} \right) . \quad (33)$$

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

$$\left\| \mathbb{P}^{(n,p_n)} - \mathbb{Q}_{\theta^*}^{(0,p_n)} \right\|_{TV} \rightarrow 0. \quad (34)$$

Proof. For two $p \times q$ matrices A, B we define the inner product $\langle A, B \rangle := \text{Tr}(A^T B)$ and define the norm $|A| := \sqrt{\langle A, A \rangle}$ where $\text{Tr}(A)$ is the trace of A . It is clear from the set-up of the algorithm that there exists $r > 0$ such that $|\mu^*| \leq r$, $|\Lambda^*| \leq r$ and for all $n \geq 1$, $|\mu_n| \leq r$, $|\Lambda_n| \leq r$ (take for example $r = \max[|\mu^*|, |\Lambda^*|, 4 \sup_{x \in \mathcal{X}} |x| (1 + \sup_{x \in \mathcal{X}} |x|) + \varepsilon]$). Let $B_p(r)$ be the ball of \mathbb{R}^p with center 0 and radius r and let $\mathcal{C}_p(r, \varepsilon)$ be the set of all $p \times p$ (symmetric positive definite) matrices of the form $\Sigma = \Lambda + \varepsilon I_p$ where Λ is symmetric $p \times p$ positive matrix with $|\Lambda| \leq r$. Define $\Theta = B_p(r) \times \mathcal{C}_p(r, \varepsilon)$. We equip Θ with the inner product $\langle (\mu_1, \Lambda_1), (\mu_2, \Lambda_2) \rangle := \langle \mu_1, \mu_2 \rangle + \langle \Lambda_1, \Lambda_2 \rangle$ and the norm $|\theta| := \sqrt{|\mu|^2 + |\Lambda|^2}$, $\theta = (\mu, \Lambda)$.

The adapting process (μ_n, Λ_n) lives in Θ (and as a result for any $n \geq 1$, $\mathbf{1}(n < \tau) = 1$) and is governed by the recursion

$$(\mu_n, \Lambda_n) = (\mu_{n-1}, \Lambda_{n-1}) + \frac{1}{n} H(\mu_{n-1}, \Lambda_{n-1}; X_n),$$

where

$$H(\mu, \Lambda; x) = (x - \mu, xx^T - \Lambda),$$

$$h(\mu, \Lambda) := \int_{\mathcal{X}} H(\mu, \Lambda; x) \pi(dx) = (\mu^* - \mu, \Lambda^* - \Lambda).$$

We thus clearly have $\langle \theta - \theta^*, h(\theta) \rangle = \langle \mu - \mu^*, \mu^* - \mu \rangle + \langle \Lambda - \Lambda^*, \Lambda^* - \Lambda \rangle = -|\theta - \theta^*|^2$ and (A5) holds with $\delta = 1$.

It is shown in Haario et al. (2001) (see also Proposition 9 and Lemma 10 of Andrieu and Moulines (2005) for a generalization) that assumptions (A1) and (A2) hold for the family $\{P_\theta : \theta = (\mu, \Lambda) \in \Theta\}$, with $C = \mathcal{X}$ and $V(x) \equiv 1$. The results follow from the development in Section 2. \square

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 (2005), see also Andrieu (2004) for a discussion.

4 A Quasi-perfect adaptive sampling algorithm

In this section, we propose a new framework for adaptive MCMC algorithms, which has the advantage to make statistical inference from such algorithms particularly easy. We call this type of

algorithms *quasi-perfect adaptive samplers*. These algorithms are based on subsampling. The idea consists of replacing P_{θ_n} at iteration $n \geq 1$ with $P_{\theta_n}^{a_n}$ for some predefined integer sequence $\{a_n\}$ in the generic algorithm of Section 2. Under appropriate mixing conditions, as $a_n \rightarrow \infty$ the resulting chain can be seen as a sequence of i.i.d. samples from π . In fact we find that a rate as slow as $a_n \propto \log(n)$ is sufficient to produce quasi-perfect samples.

4.1 Algorithm and properties

For the details, we assume the set-up of Section 2. Given a sequence of integers $\{a_n\}$, the algorithm proceeds as follows:

Algorithm 4.1. (Initialization) *Initialize the algorithm with $(\theta_0, X_0) \in \Theta \times \mathcal{X}$.*

(Iteration) *For $n \geq 0$, given (X_n, θ_n) : set $W_{n+1}^0 = X_n$ and for $i = 1, \dots, a_n$ generate $W_{n+1}^i \sim P_{\theta_n}(W_{n+1}^{i-1}, \cdot)$; set $X_{n+1} = W_{n+1}^{a_n}$ and compute θ_{n+1} as any measurable function of $\{W_k^i, k = 1, \dots, n+1 \text{ and } i = 0, \dots, a_k\}$.*

Let \mathbb{T} be the distribution on $(\mathcal{X}^\infty, \mathcal{B}(\mathcal{X})^\infty)$ of the i.i.d. sequence $Z = \{Z_n\}$ where each Z_n has distribution π . The following result is immediate.

Theorem 4.1. *Assume (A1). Then there exists a constant $C \in (0, \infty)$ such that for any $n, p \geq 1$,*

$$\|\mathbb{P}^{(n,p)} - \mathbb{T}^{(0,p)}\|_{TV} \leq C \sum_{i=n}^{\infty} \rho^{a_i}, \quad (35)$$

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

Proof. The proof follows easily from the fact that for any integer $p \geq 0$ and any $(p+1)$ -uple of measurable functions (f_0, \dots, f_p) , $f_i : \mathcal{X} \rightarrow [-1, 1]$ for $i = 0, \dots, p$, we have the bound

$$\left| \mathbb{E} \left[\prod_{i=0}^p f_i(X_{n+i}) - \prod_{i=0}^p \pi(f_i) \mid (X_0, \dots, X_{n-1}) \right] \right| \leq V(X_{n-1}) \sum_{j=n-1}^{n+p-1} \rho^{a_j}.$$

□

We have the following immediate corollary.

Corollary 4.1. *Assume (A1) and that \mathcal{X} is Polish. If $\sum_{n \geq 1} \rho^{a_n} < \infty$, then there exists a finite coupling time T such that for $n \geq T$, (X_n, X_{n+1}, \dots) and Z have the same distribution. As a consequence, for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$,*

- (i) If $\pi(|f|) < \infty$ then $\frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{as} \pi(f)$ as $n \rightarrow \infty$;
- (ii) If $\pi(f^2) < \infty$ then $\frac{1}{\sqrt{n}} \sum_{i=1}^n [f(X_i) - \pi(f)] \xrightarrow{w} \mathcal{N}(0, \sigma^2(f))$ as $n \rightarrow \infty$; where $\sigma^2(f) = \pi(f^2) - \pi(f)^2$.

Remark 4.1. • It follows from this corollary that the output from a quasi-perfect sampler can be analyzed as the output from an i.i.d. Monte Carlo sampler.

- The condition $\sum_{n \geq 1} \rho^{a_n} < \infty$ is satisfied for example with $a_n = c \log(n)$ and $-c \log(\rho) > 1$. In practice, since ρ is not known, we suggest $a_n = \lceil \log(1 + \log(n + 1)) \log(n) \rceil$, where $\lceil x \rceil$ is the closest integer larger or equal than x .
- Strictly speaking, it is not necessary to adapt θ_n in Algorithm 4.1. But in practice, the adaptation is vital as it allows the algorithm to move its sampling kernel towards P_{θ^*} . This can significantly improve its coupling time.

4.2 A simulation example

We illustrate Algorithm 4.1 with the RWM algorithm presented in Section 3.2. Define $\mathcal{X} = B_3(0, 10^3)$, the \mathbb{R}^3 -ball centered at 0 with radius 10^3 . The probability measure of interest π is the Gaussian distribution with mean 0 and covariance matrix

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

We compare a plain RWM algorithm and its quasi-perfect adaptive version. For the plain RWM, we use a Gaussian proposal kernel $\mathcal{N}(x, \sigma I_3)$ with $\sigma = 0.56$. The value $\sigma = 0.56$ is used to obtain an acceptance rate of about 30%. For the quasi-perfect adaptive sampler, the proposal kernel is $\mathcal{N}(x, \Sigma)$ where the matrix Σ is adapted as in Section 3.2. We use $a_n = \lceil \log(1 + \log(n + 1)) \log(n) \rceil$. For the quasi-perfect adaptive algorithm, we made $n = 5,000$ iterations, which correspond to about 79,000 iterations of the plain RWM algorithm. Suppose we are interested in $\pi(f)$ where $f(x_1, x_2, x_3) = x_1$. Let $\hat{\pi}_{QPS}(f)$ (resp. $\hat{\pi}_{RWM}(f)$) be the estimate of $\pi(f)$ given by the quasi-perfect sampler (resp. the plain RWM sampler). In order to compare the finite sample performance of the algorithms, we generated 100 independent chains of each. Let $\hat{\pi}_{QPS}^{(i)}(f)$ (resp. $\hat{\pi}_{RWM}^{(i)}(f)$) be the estimate of $\pi(f)$ from the i -th chain of the quasi-perfect sampler (resp. the plain RWM sampler). Based on

these samples, we estimate the densities of $\hat{\pi}_{QPS}(f)$ and $\hat{\pi}_{RWM}(f)$. The results are presented in Fig. 1. Fig. 1(a) shows the autocorrelation function of the last 2,000 iterations from one run of the quasi-perfect adaptive sampler. This autocorrelation function is virtually identical to the autocorrelation function of an i.i.d. sample. Fig. 1(b) presents the density estimates of the samples $(\hat{\pi}_{QPS}^{(i)}(f))_{1 \leq i \leq 100}$ (dashed line) and $(\hat{\pi}_{RWM}^{(i)}(f))_{1 \leq i \leq 100}$. Clearly, the quasi-perfect sampler (dashed line) is more precise than the plain RWM algorithm with an estimated efficiency of $\hat{e}_n(f) = 2.73$.

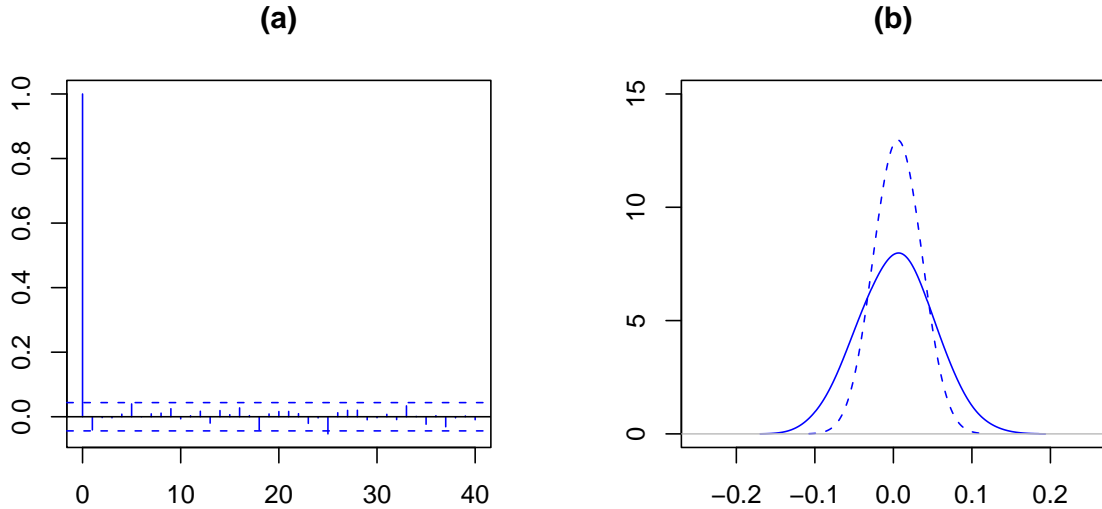


Fig. 1: (a): Autocorrelation function of the quasi-perfect adaptive sampler. (b): density estimates from 100 realizations of each sampler (quasi-perfect adaptive sampler shown in dashed line).

5 Proofs

5.1 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, $\mathcal{F}_n = \sigma(X_0, Y_0, \dots, X_n, Y_n)$. It is sufficient to work with functions of the form $f := \prod_{i=0}^p f_i$ for $\{f_i : \mathcal{X} \rightarrow \mathbb{R}, |f_i| \leq 1, i = 0, \dots, p\}$ a family of measurable functions and any $p > 1$. The proof relies on the following decomposition

$$\mathbb{E} \left[\prod_{i=0}^p f_i(X_{n+i}) - \prod_{i=0}^p f_i(Y_{n+i}) \right] = \mathbb{E} \left[\mathbb{E} \left[\prod_{i=0}^p f_i(X_{n+i}) - \prod_{i=0}^p f_i(Y_{n+i}) \middle| \mathcal{F}_{n-1} \right] \right]. \quad (36)$$

An estimate of the inner conditional expectation term is given in Proposition 5.2 below and the outer expectation operator is studied in Proposition 5.1 below as well. The combination of these

results leads to

$$|\mathbb{E} [\prod_{i=0}^p f_i(X_{n+i}) - \prod_{i=0}^p f_i(Y_{n+i})]| \leq C \left\{ \rho^{n-i_n} + \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} + \sum_{j=n-1}^{n+p-1} \alpha_j \right\} \quad (37)$$

$$\leq C \sum_{j=n}^{n+p-1} \alpha_j \text{ when } \alpha_j \propto j^{-\gamma} \text{ for some } \gamma > 0, \quad (38)$$

hence the result. \square

Proposition 5.1. *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 $n \geq 1$,*

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \mathbb{E} [P_{\theta^*}^{n-i_n} g(X_{i_n}) - g(X_n)] &= \mathbb{E} \left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - P_{\theta^*}^{n-(j+1)} g(X_{j+1}) \right] \\ &= \mathbb{E} \left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - \mathbb{E}_x \left[P_{\theta^*}^{n-(j+1)} g(X_{j+1}) | \mathcal{F}_j \right] \right] \\ &= \mathbb{E} \left[\sum_{j=i_n}^{n-1} P_{\theta^*}^{n-j} g(X_j) - P_{\theta_j} P_{\theta^*}^{n-(j+1)} g(X_j) \right] \\ &= \mathbb{E} \left[\sum_{j=i_n}^{n-1} (P_{\theta^*} - P_{\theta_j}) P_{\theta^*}^{n-(j+1)} g(X_j) \right] \\ &= \mathbb{E} \left[\sum_{j=i_n}^{n-1} (P_{\theta^*} - P_{\theta_j}) P_{\theta^*}^{n-(j+1)} (g(X_j) - \pi(g)) \right] \end{aligned} \quad (39)$$

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

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

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

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

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,

$$n^{-\gamma} I_n \leq \sum_{j=i_n}^{n-1} \alpha_j \rho^{n-(j+1)} \leq i_n^{-\gamma} I_n \text{ with } I_n \frac{1 - \rho^{n-i_n}}{1 - \rho}, \quad (40)$$

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. \square

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

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

We have the proposition,

Proposition 5.2. *Assume (A1-3). Let $\{X_i\}$ be the adaptive chain and Let $\{f_i : \mathcal{X} \rightarrow \mathbb{R}, |f_i| \leq 1, i = 0, \dots, p\}$ be a family of measurable functions. Then there exists a constant $C \in (0, \infty)$ such that for any $n, p \in \mathbb{Z}^+$,*

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

Proof. We have the following telescoping sum decomposition,

$$\begin{aligned} &\mathbb{E} [\prod_{i=0}^p f_i(X_{n+i}) - \phi_0(X_n) | \mathcal{F}_{n-1}] \\ &= \mathbb{E} \left[\sum_{k=0}^p \left(\phi_{k+1}(X_{n+k}) \prod_{i=0}^k f_i(X_{n+i}) - \phi_k(X_{n+k-1}) \prod_{i=0}^{k-1} f_i(X_{n+i}) \right) \middle| \mathcal{F}_{n-1} \right]. \quad (41) \end{aligned}$$

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

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

Finally,

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

□

5.2 Proof of Theorem 2.3

Proof. Throughout, fix $f : \mathcal{X} \rightarrow \mathbb{R}$ measurable with $|f| \leq V^\beta$ and $\pi(f) = 0$.

For a sequence of random variables X, X_1, X_2, \dots , we say that X_n converge in probability to X and write $X_n \xrightarrow{Prob.} X$ if for all $\varepsilon > 0$, $\Pr [|X_n - X| > \varepsilon] \rightarrow 0$ as $n \rightarrow \infty$. Denote $\hat{f}_\theta = \sum_{k=0}^{\infty} P_\theta^k f$ the solution of the Poisson equation $f = \hat{f}_\theta - P_\theta \hat{f}_\theta$. As shown by Andrieu and Moulines (2005), Proposition 3, under (A1-2) such solutions exist and satisfy:

$$\begin{aligned}
& \sup_{\theta \in \Theta} \left(|P_\theta \hat{f}_\theta|_{V^\beta} + |\hat{f}_\theta|_{V^\beta} \right) < +\infty \\
& \sup_{\theta, \theta' \in \Theta, \theta \neq \theta'} |\theta - \theta'|^{-1} \left[|\hat{f}_\theta - \hat{f}_{\theta'}|_{V^\beta} + |P_\theta \hat{f}_\theta - P_{\theta'} \hat{f}_{\theta'}|_{V^\beta} \right] < +\infty.
\end{aligned}$$

Therefore we can decompose $f(X_n)$ as:

$$\begin{aligned}
f(X_n) &= \hat{f}_{\theta_n}(X_n) - P_{\theta_n} \hat{f}_{\theta_n}(X_n) \\
&= \delta M_n + T_n^{(1)} + T_n^{(2)},
\end{aligned} \tag{42}$$

where

$$\begin{aligned}
\delta M_n &= \hat{f}_{\theta_{n-1}}(X_n) - P_{\theta_{n-1}} \hat{f}_{\theta_{n-1}}(X_{n-1}), \\
T_n^{(1)} &= \hat{f}_{\theta_n}(X_n) - \hat{f}_{\theta_{n-1}}(X_n),
\end{aligned}$$

$$T_n^{(2)} = P_{\theta_{n-1}} \hat{f}_{\theta_{n-1}}(X_{n-1}) - P_{\theta_n} \hat{f}_{\theta_n}(X_n).$$

To prove that $\frac{1}{\sqrt{n}} S_n(f)$ satisfies a CLT, we shall prove $\frac{1}{\sqrt{n}} \sum_{k=1}^n T_k^{(i)} \xrightarrow{Prob.} 0$ ($i = 1, 2$) and $\frac{1}{\sqrt{n}} \sum_{k=1}^n \delta M_k$ satisfies a CLT with asymptotic variance $\sigma_*^2(f) = \pi \left[P_{\theta^*} \hat{f}_{\theta^*}^2 - \left(P_{\theta^*} \hat{f}_{\theta^*} \right)^2 \right] < \infty$. One can verify that this asymptotic variance can also be written $\sigma_*^2(f) = \pi(f^2) + 2 \sum_{k=1}^{\infty} \pi [f P_{\theta^*}^k f]$.

The terms $T_k^{(2)}$ telescope: $\sum_{k=1}^n T_k^{(2)} = P_{\theta_0} \hat{f}_{\theta_0}(X_0) - P_{\theta_n} \hat{f}_{\theta_n}(X_n)$. From Lemma 5.1 in Andrieu et al. (2005), there exists $C < \infty$ such that $\mathbb{E} \left| P_{\theta_0} \hat{f}_{\theta_0}(X_0) - P_{\theta_n} \hat{f}_{\theta_n}(X_n) \right| \leq C \sup_n V^\beta(X_n) < \infty$. Therefore $\frac{1}{\sqrt{n}} \sum_{k=1}^n T_k^{(2)} \xrightarrow{Prob.} 0$.

By Markov's inequality and the fact that $\theta \rightarrow \hat{f}_\theta$ is V^β -Lipschitz, we have for some finite constant C, C' :

$$\begin{aligned} \Pr \left[\frac{1}{\sqrt{n}} \left| \sum_{k=1}^n T_k^{(1)} \right| > \varepsilon \right] &\leq \frac{1}{\varepsilon \sqrt{n}} \sum_{k=1}^n \mathbb{E} \left[\left| T_k^{(1)} \right| \right] \\ &\leq \frac{C}{\varepsilon \sqrt{n}} \sum_{k=1}^n \mathbb{E} \left[|\theta_k - \theta_{k-1}| V^\beta(X_k) \right] \\ &\leq \frac{C'}{\varepsilon} \sup_n \mathbb{E} \left[V^{2\beta}(X_n) \right] \frac{1}{\sqrt{n}} \sum_{k=1}^n \gamma_k \rightarrow 0, \end{aligned}$$

hence $\frac{1}{\sqrt{n}} \sum_{k=1}^n T_k^{(1)} \xrightarrow{Prob.} 0$.

Define $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$. We have $\mathbb{E}[\delta M_n | \mathcal{F}_{n-1}] = 0$. Thus the sequence $\{\delta M_n; \mathcal{F}_n\}$ is a difference martingale. With some minor modifications to the arguments in Andrieu and Moulines (2005), we can show that:

- (a) $V_n^2 = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[\delta M_k^2 | \mathcal{F}_{k-1}] \xrightarrow{Prob.} \sigma_*^2(f)$,
- (b) for all $\varepsilon > 0$, $\frac{1}{n} \sum_{k=1}^n \mathbb{E}[\delta M_k^2 \mathbf{1}_{\{\delta M_k > \varepsilon \sqrt{n}\}} | \mathcal{F}_{k-1}] \xrightarrow{Prob.} 0$.

Applying Corollary 3.1 of Hall and Heyde (1980) we conclude that $\frac{1}{\sigma_*(f)\sqrt{n}} \sum_{k=1}^n \delta M_k \xrightarrow{w} Z$, where $Z \sim \mathcal{N}(0, 1)$ and the theorem is proved. □

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. (2005). On the ergodicity Properties of some Adaptive MCMC Algorithms. *to appear Ann. Appl. Probab.* .

- ANDRIEU, C., MOULINES, E. and PRIOURET, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization* **44** 283–312.
- ANDRIEU, C. and ROBERT, C. P. (2001). Controlled mcmc for optimal sampling. *Technical report, Université Paris Dauphine, Ceremade 0125* .
- ATCHADE, Y. F. (2005). An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Technical Report* .
- ATCHADE, Y. F. and ROSENTHAL, J. S. (2005). On adaptive markov chain monte carlo algorithm. *Bernoulli* **11** 815–828.
- BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic markov chains. *Annals of Applied Probability* **15** 700–738.
- BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic approximations*. Applications of Mathematics, Springer, Paris-New York.
- GALIN, L. J. (2004). On the markov chain central limit theorem. *Probability surveys* **1** 299–320.
- GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (eds.) (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary Statistics, Chapman & Hall, London.
- 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.
- GOLDSTEIN, S. (1979). Maximal coupling. *Z. Wahrsch. Verw. Gebiete* **46** 193–204.
- HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7** 223–242.
- HALL, P. and HEYDE, C. C. (1980). *Martingale Limit theory and its application*. Academic Press, New York.
- KUSHNER, K. and YIN, Y. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer, Springer-Verlag, New-York.
- ROSENTHAL, J. S. and ROBERTS, G. O. (2005). Coupling and ergodicity of adaptive mcmc. *Technical Report, MCMC preprints* .