On the Ergodicity Properties of some Adaptive MCMC Algorithms

Christophe Andrieu [*], Éric Moulines [†]

Abstract

In this paper we study the ergodicity properties of some adaptive Monte Carlo Markov chain algorithms (MCMC) that have been recently proposed in the literature. We prove that under a set of verifiable conditions, ergodic averages calculated from the output of a so-called adaptive MCMC sampler converge to the required value and can even, under more stringent assumptions, satisfy a central limit theorem. We prove that the conditions required are satisfied for the Independent Metropolis-Hastings algorithm and the Random Walk Metropolis algorithm with symmetric increments. Finally we propose an application of these results to the case where the proposal distribution of the Metropolis-Hastings update is a mixture of distributions from a curved exponential family.

Keywords: Adaptive Markov chain Monte Carlo, Self-tuning algorithm, Metropolis-Hastings algorithm, Stochastic approximation, state-dependent noise, randomly varying truncation, Martingale, Poisson method.

AMS Classification: 65C05, 65C40,60J27, 60J35.

1 Introduction

Markov chain Monte Carlo (MCMC), introduced by Metropolis et al. (1953), is a popular computational method for generating samples from virtually any distribution π. In particular there is no need for the normalising constant to be known and the space $\mathsf{X} \subset \mathbb{R}^{n_x}$ (for some integer n_x) on which it is defined can be high dimensional. We will hereafter denote $\mathcal{B}(\mathsf{X})$ the associated countably generated σ-field. The method consists of simulating an ergodic Markov chain $\{X_k, k \geq 0\}$ on X with transition probability P such that π is a *stationary* distribution for this chain, *i.e* $\pi P = \pi$. Such samples can be used *e.g.* to compute integrals

$$\pi(\psi) := \int_{\mathsf{X}} \psi(x) \pi(dx),$$

[*] University of Bristol, School of Mathematics, University Walk, BS8 1TW, UK (c.andrieu@bris.ac.uk).

[†] École Nationale Supérieure des Télécommunications, URA CNRS 820, 46, rue Barrault, F 75634 PARIS Cedex 13 (moulines@tsi.enst.fr).

for some π-integrable function $\psi : \mathsf{X} \to \mathbb{R}^{n_\psi}$, for some integer n_ψ, using estimators of the type

$$S_n(\psi) = \frac{1}{n} \sum_{k=1}^{n} \psi(X_k). \tag{1}$$

In general the transition probability P of the Markov chain depends on some tuning parameter, say θ defined on some space $\Theta \subset \mathbb{R}^{n_\theta}$ for some integer n_θ, and the convergence properties of the Monte Carlo averages in Eq. (1) might highly depend on a proper choice for these parameters.

We illustrate this here with the Metropolis-Hastings (MH) update, but it should be stressed at this point that the results presented in this paper apply to much more general settings (including in particular hybrid samplers, sequential or population Monte Carlo samplers). The MH algorithm requires the choice of a *proposal distribution* q. In order to simplify the discussion, we will here assume that π and q admit densities with respect to the Lebesgue measure λ^{Leb}, denoted with an abuse of notation π and q hereafter. The rôle of the distribution q consists of proposing potential transitions for the Markov chain $\{X_k\}$. Given that the chain is currently at x, a candidate y is accepted with probability $\alpha(x, y)$ defined as

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q(y,x)}{q(x,y)} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{otherwise,} \end{cases}$$

where $a \wedge b := \min(a, b)$. Otherwise it is rejected and the Markov chain stays at its current location x. The transition kernel P of this Markov chain takes the form for $x, A \in \mathsf{X} \times \mathcal{B}(\mathsf{X})$

$$P(x,A) = \int_{A-x} \alpha(x,x+z) q(x,x+z) \lambda^{\text{Leb}}(dz) + \mathbb{1}_A(x) \int_{\mathsf{X}-x} (1-\alpha(x,x+z)) q(x,x+z) \lambda^{\text{Leb}}(dz), \tag{2}$$

where $A - x := \{z \in \mathsf{X}, x + z \in A\}$. The Markov chain P is reversible with respect to π, and therefore admits π as invariant distribution. Conditions on the proposal distribution q that guarantee irreducibility and positive recurrence are mild and many satisfactory choices are possible; for the purpose of illustration, we concentrate in this introduction on the symmetric increments random-walk MH algorithm (hereafter SRWM), which corresponds to the case where $q(x, y) = q(x - y)$ for some symmetric probability density q. The transition kernel of the Metropolis algorithm is then given for $x, A \in \mathsf{X} \times \mathcal{B}(\mathsf{X})$ by

$$\begin{aligned} P_q^{\text{SRW}}(x,A) = &\int_{A-x} (1 \wedge \pi(x+z)/\pi(x))\, q(z)\, \lambda^{\text{Leb}}(dz) + \\ &\mathbb{1}_A(x) \int_{\mathsf{X}-x} (1 - (1 \wedge \pi(x+z)/\pi(x)))\, q(z)\, \lambda^{\text{Leb}}(dz), \quad x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X}). \end{aligned} \tag{3}$$

A classical choice for the proposal distribution is $q = \phi_{0,\Gamma}$, where $\phi_{\mu,\Gamma}$ is the density of a multivariate normal distribution with mean μ and covariance matrix Γ. We will later on refer to this algorithm as the N-SRWM. It is well known that either too small or too large a covariance matrix will result in highly positively correlated Markov chains, and therefore estimators $S_n(\psi)$ with a large variance (Gelman et al. (1995) have shown that the "optimal" covariance matrix (under restrictive technical conditions not given here) for the N-SRWM is $(2.38^2/n_x)\Gamma_\pi$, where Γ_π is the true covariance matrix of the target distribution). In practice this covariance matrix Γ is determined by trial and error, using several realisations of the Markov chain. This hand-tuning requires some expertise and can be time-consuming.

In order to circumvent this problem, in the context of the N-SRWM update described above, Haario et al. (2001) have proposed to "learn Γ on the fly". The Haario et al. (2001) algorithm can be summarized as follows,

$$\begin{aligned} \mu_{k+1} &= \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k) \qquad\qquad k \geq 0 \\ \Gamma_{k+1} &= \Gamma_k + \gamma_{k+1}((X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^{\mathsf{T}} - \Gamma_k) \end{aligned} \tag{4}$$

where

- X_{k+1} is drawn from $P_{\theta_k}(X_k, \cdot)$, where for $\theta = (\mu, \Gamma)$, $P_\theta := P^{\text{SRW}}_{\phi_{0,\lambda\Gamma}}$ with $\lambda > 0$ a constant scaling factor depending only on the dimension of the state-space n_x and kept constant across the iterations,
- $\{\gamma_k\}$ is a non-increasing sequence of positive stepsizes such that $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^{1+\delta} < \infty$ for some $\delta > 0$ (Haario et al. (2001) have suggested the choice $\gamma_k = 1/k$).

It was realised in Andrieu and Robert (2001) that such a scheme is a particular case of a more general framework akin to stochastic control, combined with the use of the Robbins and Monro (1951) procedure. More precisely, let $\theta = (\mu, \Gamma) \in \Theta$, denote

$$H(x; \theta) = (x - \mu, (x - \mu)(x - \mu)^{\mathsf{T}} - \Gamma)^{\mathsf{T}}. \tag{5}$$

With this notation, the recursion in (4) may be written in the standard Robbins-Monro form as

$$\theta_{k+1} = \theta_k + \gamma_{k+1} H(X_{k+1}, \theta_k), \quad k \geq 0, \tag{6}$$

with $X_{k+1} \sim P_{\theta_k}(X_k, \cdot)$. This recursion is at the core of most of classical stochastic approximation algorithms (see *e.g.* Benveniste et al. (1990), Duflo (1997), Kushner and Yin (1997) and the references therein). This algorithm is well suited to solve the equation $h(\theta) = 0$ where, assuming that $\int_{\mathsf{X}} |H(x, \theta)| \pi(dx) < \infty$, $\theta \mapsto h(\theta)$ is the so-called mean field defined as

$$h(\theta) := \int_{\mathsf{X}} H(x, \theta) \pi(dx). \tag{7}$$

For the present example, assuming that $\int_{\mathsf{X}} |x|^2 \pi(dx) < \infty$, one can easily check that

$$h(\theta) = \int_{\mathsf{X}} H(x, \theta) \pi(dx) = (\mu_\pi - \mu, (\mu_\pi - \mu)(\mu_\pi - \mu)^{\mathsf{T}} + \Gamma_\pi - \Gamma)^{\mathsf{T}}, \tag{8}$$

with μ_π and Γ_π the mean and covariance of the target distribution. One can rewrite (6) as

$$\theta_{k+1} = \theta_k + \gamma_{k+1} h(\theta_k) + \gamma_{k+1} \xi_{k+1},$$

where $\{\xi_k = H(X_k, \theta_{k-1}) - h(\theta_{k-1}); k \geq 1\}$ is generally referred to as "the noise". The general theory of *stochastic approximation* (SA) provides us with conditions under which this recursion eventually converges to the set $\{\theta \in \Theta, h(\theta) = 0\}$. Practical conditions to prove the convergence w.p. 1 of $\{\theta_k\}$ include the stability of the "noiseless" sequence

$\bar{\theta}_{k+1} = \bar{\theta}_k + \gamma_{k+1} h(\bar{\theta}_k)$ and the ergodicity of the noise sequence $\{\xi_k\}$ (the effect of the noise sequence should eventually average out to zero in order for $\{\theta_k\}$ to follow the behaviour of $\{\bar{\theta}_k\}$). These issues are discussed in Sections 3 and 5.

In the context of adaptive MCMC, the parameter convergence is not the central issue; the focus is rather on the approximation of $\pi(\psi)$ by $S_n(\psi)$. However there is here a difficulty with the adaptive approach : as the parameter estimate $\theta_k = \theta_k(X_0, \ldots, X_k)$ depends on the whole past, the successive draws $\{X_k\}$ do not define an homogeneous Markov chain and standard arguments for the consistency and asymptotic normality of $S_n(\psi)$ do not apply in this framework. Note that this is despite the fact that for any $\theta \in \Theta$, $\pi P_\theta = \pi$. These are the problems that we address in the present paper and our main general results are in words the following:

1. In situations where $|\theta_{k+1} - \theta_k| \to 0$ as $k \to +\infty$ w.p. 1, we prove a weak law of large numbers for $S_n(\psi)$ (see Theorem 6) under mild additional conditions. Such a consistency result may arise even in situations where the parameter $\{\theta_k\}$ does not converge.

2. In situations where θ_k converges w.p. 1, we prove an invariance principle for $\sqrt{n}(S_n(\psi) - \pi(\psi))$; the limiting distribution is in general a mixture of Gaussian distributions (see Theorem 8).

Note that Haario et al. (2001) have proved the consistency of Monte Carlo averages for the specific N-SRWM algorithm. Our result applies to more general settings and rely on assumptions which are less restrictive than those used in Haario et al. (2001). The second point above, the invariance principle, has to the best of our knowledge not been addressed for adaptive MCMC algorithms.

The paper is organized as follows. In Section 2 we detail our general procedure and introduce some notation. In Section 3, we establish the consistency (*i.e.* a law of large numbers) for $S_n(\psi)$. In Section 4 we strengthen the conditions required to ensure the law of large numbers (LLN) for $S_n(\psi)$ and establish an invariance principle. In Section 5 we focus on the classical Robbins-Monro implementation of our procedure and introduce further conditions that allow us to prove that $\{\theta_k\}$ converges w.p. 1. In Section 6 we establish general properties of the generic SRWM required to ensure a LLN and an invariance principle. For pædagocical purposes we show how to apply these results to the simple N-SRWM. In Section 7 we present another application of our theory. We focus on the Independent Metropolis-Hastings algorithm (IMH) and establish general properties required for the LLN and the invariance principle. We then go on to propose and analyse an algorithm that matches the so-called proposal distribution of the IMH to the target distribution π, in the case where the proposal distribution is a mixture of distributions from the exponential family. The main result of this section is Theorem 19. We conclude with the remark that this latter result equally applies to a generalisation of the N-SRWM, where the proposal is again a mixture of distributions. Application to samplers which consist of a mixture of SRWM and IMH is straightforward.

2 Algorithm description and main definitions

Before describing the procedure under study, it is necessary to introduce some notation and definitions. Let T be a separable space and let $\mathcal{B}(\mathsf{T})$ be a countably generated σ-field on T. For a Markov chain with transition probability $\Pi : \mathsf{T} \times \mathcal{B}(\mathsf{T}) \to [0, 1]$ and any non-negative measurable function $\psi : \mathsf{T} \to [0, +\infty)$, we denote $\Pi\psi(t) = \Pi(t, \psi) := \int_{\mathsf{T}} \Pi(t, dt')\psi(t')$ and for any integer k, Π^k the k-th iterate of the kernel. For a probability measure μ we define for any $A \in \mathcal{B}(\mathsf{T})$ $\mu\Pi(A) := \int_{\mathsf{T}} \mu(dt)\Pi(t, A)$. A Markov chain on a state space T is said to be μ-irreducible if there exists a measure μ on $\mathcal{B}(\mathsf{T})$ such that, whenever $\mu(A) > 0$, $\sum_{k=0}^{\infty} \Pi^k(t, A) > 0$ for all $t \in \mathsf{T}$. Denote by μ a maximal irreducibility measure for P (see Meyn and Tweedie (1993) Chapter 4 for the definition and the construction of such a measure). If Π is μ-irreducible, aperiodic and has an invariant probability measure π, then π is unique and is a maximal irreducibility measure.

Two main ingredients are required for the definition of our adaptive MCMC algorithms:

1. A family of Markov transition kernels on X, $\{P_\theta, \theta \in \Theta\}$ indexed by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ an open set. For each θ in Θ, it is assumed that P_θ is π-irreducible and that $\pi P_\theta = \pi$, *i.e.* π is the invariant distribution for P_θ.
2. A family of *update functions* $\{H(\theta, x) : (\theta, x) \mapsto \mathbb{R}^{n_\theta}\}$, which are used to adapt the value of the tuning parameter.

In order to take into account potential jumps outside the space Θ, we extend both the parameter and state-space with two cemetery points, $\theta_c \notin \Theta$ and $x_c \notin \mathsf{X}$, and define $\bar{\Theta} = \Theta \cup \{\theta_c\}$, $\bar{\mathsf{X}} = \mathsf{X} \cup \{x_c\}$. In its general form the *basic* adaptive MCMC algorithm may be written as follows. Set $\theta_0 = \theta \in \Theta$, $X_0 = x \in \mathsf{X}$, and for $k \geq 0$ define recursively the sequence $\{X_k, \theta_k; k \geq 0\}$: if $\theta_k = \theta_c$, then set $\theta_{k+1} = \theta_c$ and $X_{k+1} = x_c$. Otherwise, draw X_{k+1} according to $P_{\theta_k}(X_k, \cdot)$ compute $\eta = \theta_k + \rho_{k+1} H(\theta_k, X_{k+1})$ and set :

$$\theta_{k+1} = \begin{cases} \eta & \text{if} \quad \eta \in \Theta, \\ \theta_c & \text{if} \quad \eta \notin \Theta. \end{cases} \tag{9}$$

where $\{\rho_k\}$, $0 \leq \rho_k \leq 1$ is a sequence of stepsizes that converges to zero. The sequence $\{(X_k, \theta_k)\}$ is a *non-homogeneous* Markov chain on the product space $\bar{\mathsf{X}} \times \bar{\Theta}$. This non-homogeneous Markov chain defines a probability measure on the canonical state space $(\bar{\mathsf{X}} \times \bar{\Theta})^{\mathbb{N}}$ equipped with the canonical product σ-algebra. We denote $\mathcal{F} = \{\mathcal{F}_k, k \geq 0\}$ the canonical filtration of this Markov chain and $\mathbb{P}^{\boldsymbol{\rho}}_{x,\theta}$ and $\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}$ the probability and the expectation associated to this Markov chain starting from $(x, \theta) \in \mathsf{X} \times \Theta$.

Because of the interaction with feedback between X_k and θ_k, the stability of this inhomogeneous Markov chain is often difficult to establish. This is a long-lasting problem in the field of stochastic optimization: known practical cures to this problem include the reprojections on a fixed set (see Kushner and Yin (1997)) or the more recent reprojection on random varying boundaries proposed in Chen and Zhu (1986), Chen et al. (1988) and generalized in Andrieu et al. (2002). In this latter case, reinitialization occurs when the current value of the parameter θ_k wanders outside a so-called *active truncation set* or when the difference between two successive values of the parameter is larger than a *time-dependent threshold*. More precisely, let $\{\mathcal{K}_q, q \geq 0\}$ be a sequence of compact subsets of

Θ such that,

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta, \quad \text{and} \quad \mathcal{K}_q \subset \operatorname{int}(\mathcal{K}_{q+1}), \quad q \geq 0, \tag{10}$$

where $\operatorname{int}(A)$ denotes the interior of set A. Let $\boldsymbol{\gamma} := \{\gamma_k\}$ and $\boldsymbol{\epsilon} := \{\epsilon_k\}$ be two monotone non-increasing sequences of positive numbers and let K be a compact subset of X. Let $\Phi : \mathsf{X} \times \Theta \to \mathsf{K} \times \mathcal{K}_0$ be a measurable function and $\phi : \mathbb{Z}^+ \to \mathbb{Z}$ be a function such that $\phi(k) > -k$ for any k. It is convenient to introduce the family of transition kernels $\{Q_\gamma, \gamma \geq 0\}$

$$Q_\gamma(x, \theta; A \times B) = \int_A P_\theta(x, dy)\delta_{\theta+\gamma H(\theta,y)}(B), \quad A \in \mathcal{B}(\mathsf{X}), B \in \mathcal{B}(\Theta), (x,\theta) \in \mathsf{X} \times \Theta$$

where δ_x is the Dirac mass at point x. Define the *homogeneous* Markov chain $\{Z_k = (X_k, \theta_k, \kappa_k, \varsigma_k, \nu_k); k \geq 0\}$ on the product space $\mathsf{Z} = \mathsf{X} \times \Theta \times (\mathbb{Z}^+)^3$ with transition probability algorithmically defined as follows: for any $(x, \theta, \kappa, \varsigma, \nu) \in \mathsf{Z}$

- If $\nu = 0$, then draw $(X', \theta') \sim Q_{\gamma_\varsigma}(\Phi(x, \theta), \cdot)$,
- If $\nu \neq 0$, then draw $(X', \theta') \sim Q_{\gamma_\varsigma}(x, \theta, \cdot)$.
- If $|\theta' - \theta| \leq \epsilon_\varsigma$ and $\theta' \in \mathcal{K}_\kappa$, then set: $\kappa' = \kappa$, $\varsigma' = \varsigma + 1$ and $\nu' = \nu + 1$; otherwise, set $\kappa' = \kappa + 1$, $\varsigma' = \varsigma + \phi(\nu)$ and $\nu' = 0$.

In words, κ, ς and ν are counters: κ is the index of the current active truncation set; ν counts the number of iterations since the last reinitialisation; ς is the current index in the sequences $\{\gamma_k\}$ and $\{\epsilon_k\}$. The event $\{\nu_k = 0\}$ means that a reinitialization occurs and the condition on ϕ ensures that the algorithm is reinitialized with a value for γ_{ς_k} smaller than that used the last time such an event occurred. This algorithm is reminiscent of the projection on random varying boundaries proposed in Chen and Zhu (1986), Chen et al. (1988). When the current iterate wanders outside the active truncation set or when the difference between two successive values of the parameter is larger than a time-dependent threshold, then the algorithm is reinitialised with a smaller initial value of the stepsize and a larger truncation set. Various choices for the function ϕ can be considered. For example, the choice $\phi(k) = 1$ for all $k \in \mathbb{N}$ coincides with the procedure proposed in Chen et al. (1988): in this case $\varsigma_k = k$. Another sensible choice consists of setting $\phi(k) = 1 - k$ for all $k \in \mathbb{N}$, in which case the number of iterations between two successive reinitialisations is not taken into account. In the latter case, we have $\varsigma_k = \kappa_k + \nu_k$.

The homogeneous Markov chain $\{Z_k, k \geq 0\}$ defines a probability measure on the canonical state space $\mathsf{Z}^{\mathbb{N}}$ equipped with the canonical product σ-algebra. We denote $\mathcal{G} = \{\mathcal{G}_k, k \geq 0\}$, $\bar{\mathbb{P}}_{x_0,\theta_0,\kappa_0,\varsigma_0,\nu_0}$ and $\bar{\mathbb{E}}_{x_0,\theta_0,\kappa_0,\varsigma_0,\nu_0}$ the filtration, probability and expectation associated to this process initialised at $(x_0, \theta_0, \kappa_0, \varsigma_0, \nu_0)$. For simplicity we will use the following short notation,

$$\bar{\mathbb{P}}_{x_0,\theta_0} = \bar{\mathbb{P}}_{x_0,\theta_0,0,0,0}. \tag{11}$$

This probability measure depends upon the deterministic sequences $\{\gamma_n\}$ and $\{\epsilon_n\}$: the dependence will be implicit here. We define recursively $\{T_n, n \geq 0\}$ the sequence of successive reinitialisation times

$$T_{n+1} = \inf\{k \geq T_n + 1,\ \nu_k = 0\}, \quad \text{with } T_0 = 0, \tag{12}$$

where by convention $\inf\{\emptyset\} = \infty$. It may be shown that under mild conditions on $\{P_\theta, \theta \in \Theta\}$, $\{H(\theta, x), (\theta, x) \in \Theta \times \mathsf{X}\}$ and the sequences $\{\gamma_k\}$ and $\{\epsilon_k\}$ then

$$\inf_{(x,\theta)\in\mathsf{X}\times\Theta} \bar{\mathbb{P}}_{x,\theta}\left(\sup_{n\geq 0} \kappa_n < \infty\right) = \inf_{(x,\theta)\in\mathsf{X}\times\Theta} \bar{\mathbb{P}}_{x,\theta}\left(\bigcup_{n=0}^{\infty}\{T_n = \infty\}\right) = 1,$$

i.e., the number of reinitialisations of the procedure described above is finite $\bar{\mathbb{P}}_{x,\theta}$-a.s., for every $(x, \theta) \in \mathsf{X} \times \Theta$. We postpone the presentation and the discussion of these conditions to Section 5. The lemma below (adapted from (Andrieu et al., 2002, Lemma 4.1)) relates the expectation of the inhomogeneous Markov chain defined by the transition in Eq. (9) to the expectation of the homogeneous Markov chain $\{Z_n\}$. Define for $\mathcal{K} \subset \Theta$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$,

$$\sigma(\boldsymbol{\epsilon}, \mathcal{K}) = \sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}), \tag{13}$$

where

$$\sigma(\mathcal{K}) = \inf\{k \geq 1, \theta_k \notin \mathcal{K}\}, \tag{14}$$

$$\nu(\boldsymbol{\epsilon}) = \inf\{k \geq 1, |\theta_k - \theta_{k-1}| \geq \epsilon_k\}. \tag{15}$$

For a sequence $\mathbf{a} = \{a_k\}$ and an integer l, we define $\mathbf{a}^{\leftarrow l} = (a_k^{\leftarrow l}, k \geq 0)$ as $a_k^{\leftarrow l} = a_{k+l}$.

Lemma 1. *For any $m \geq 1$, for any non-negative measurable function $\Psi_m : (\mathsf{X} \times \Theta)^m \to \mathbb{R}^+$, for any integers p and q, for any $x, \theta \in \mathsf{X} \times \Theta$,*

$$\bar{\mathbb{E}}_{x,\theta,p,q,0}\{\Psi_m(X_1, \theta_1, \ldots, X_m, \theta_m) I(T_1 \geq m)\} = \\ \mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}\{\Psi_m(X_1, \theta_1, \ldots, X_m, \theta_m) I(\sigma(\epsilon^{\leftarrow q}, \mathcal{K}_p) \geq m)\}.$$

3 Law of large number

As pointed out in the introduction, a LLN has been obtained for a particular adaptive MCMC algorithm by Haario et al. (2001), using mixingale theory, McLeish (1975). Our approach is more in line with the martingale proof of the LLN for Markov chains, and is based on the existence and regularity of the solutions of Poisson's equation and martingale limit theory. The existence and appropriate properties of those solutions can be easily established under the set of conditions (A1) below, see Proposition 2 and Proposition 3. These two propositions then allow us to conclude about the V-stability of the projection $\{X_k\}$ of the homogeneous Markov chain $\{Z_k\}$ described in Section 2: this is summarized in Proposition 4. We will need the following set of notation in what follows. For $W : \mathsf{X} \to [1, \infty)$ and $g : \mathsf{X} \to \mathbb{R}^{n_g}$ define

$$\|g\|_W = \sup_{x\in\mathsf{X}} \frac{|g(x)|}{W(x)} \quad \text{and} \quad \mathcal{L}_W = \{g : \|g\|_W < \infty\}. \tag{16}$$

Hereafter, for any $\psi : \Theta \times \mathsf{X} \to \mathbb{R}^{n_\psi}$ a function and any $\theta \in \Theta$ we will use the short-hand notation $\psi_\theta : \mathsf{X} \to \mathbb{R}$ for $\psi_\theta(x) = \psi(\theta, x)$ for all $x \in \mathsf{X}$.

(**A1**) For any $\theta \in \Theta$, P_θ is irreducible and aperiodic with stationary distribution π. In addition there exists a function $V : \mathsf{X} \to [1, \infty)$ such that, for any compact subset $\mathcal{K} \subset \Theta$,

(i) There exist an integer m, constants $0 < \lambda < 1$, b, κ, $\delta > 0$, a subset $\mathsf{C} \subseteq \mathsf{X}$ and a probability measure ν such that

$$\begin{aligned}
&\sup_{\theta \in \mathcal{K}} P_\theta^m V \le \lambda V + b I_{\mathsf{C}}, \\
&\sup_{\theta \in \mathcal{K}} P_\theta V \le \kappa V, \\
&\inf_{\theta \in \mathcal{K}} P_\theta^m(x, A) \ge \delta \nu(A) \qquad\qquad \forall x \in \mathsf{C}, \quad \forall A \in \mathcal{B}(\mathsf{X}).
\end{aligned}$$

(ii) For any $r \in [0,1]$, there exist a constant C and β, $0 < \beta \le 1$ such that, for any $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$\|P_\theta \psi - P_{\theta'} \psi\|_{V^r} \le C \; \|\psi\|_{V^r} \; |\theta - \theta'|^\beta \quad \forall \psi \in \mathcal{L}_{V^r}.$$

The drift and minorization conditions in (A1-i) imply that for any bounded set $\mathcal{K} \subset \Theta$, uniformly in $\theta \in \mathcal{K}$ then P_θ is geometrically ergodic. (A1-ii) is a Lipschitz continuity condition on the transition P_θ with respect to $\theta \in \mathcal{K}$.

Conditions of the type (A1-i) to establish geometric ergodicity have been extensively studied over the last decade for the Metropolis-Hastings algorithms. Typically the required drift function depends on the target distribution π, which makes our requirement of uniformity in $\theta \in \mathcal{K}$ in (A1-i) reasonable and relatively easy to establish (see Sections 6 and 7). Assumption (A1-ii) does not seem to have been studied for the Metropolis-Hastings algorithm. We establish this continuity for the SRWM algorithm and the independent MH algorithm (IMH) in Sections 6 and 7. Extension to hybrid samplers that consist of a mixture of SRWM and IMH updates is straightforward.

3.1 Stability and regularity

The theory of V-uniformly ergodic Markov chains (see *e.g.* (Meyn and Tweedie, 1993, Chapter 15,16)) shows that, under the drift and minorization conditions outlined in assumption (A1-i), for any compact set $\mathcal{K} \subset \Theta$, uniformly in $\theta \in \mathcal{K}$, the iterates of the kernel P_θ^k converge to the stationary distribution π in the V-norm at exponential rate. This automatically ensures the existence of solutions to Poisson's equation. More precisely, we have

Proposition 2. *Assume (A1-i). Then, for any compact subset $\mathcal{K} \subset \Theta$ and for any $r \in [0,1]$ there exist constants C and $\rho < 1$ such that for all $\psi \in \mathcal{L}_{V^r}$ and all $\theta \in \mathcal{K}$*

$$\|P_\theta^k \psi - \pi(\psi)\|_{V^r} \le C \rho^k \; \|\psi\|_{V^r}. \tag{17}$$

In addition, for all $\theta, x \in \Theta \times \mathsf{X}$ and $\psi \in \mathcal{L}_{V^r}$, $\sum_{k=0}^\infty |P_\theta^k \psi(x) - \pi(\psi)| < \infty$ and $u := \sum_{k=0}^\infty (P_\theta^k \psi - \pi(\psi))$ is a solution of Poisson's equation

$$u - P_\theta u = \psi_\theta - \pi(\psi_\theta). \tag{18}$$

Remark 1. For a fixed value of θ, it follows from (Meyn and Tweedie, 1993, Theorem 16.0.1) that there exists a constant C_θ, such that $\|P_\theta^k \psi - \pi(\psi)\|_{V^r} \le C_\theta \rho^k \; \|\psi\|_{V^r}$ for any $\psi \in \mathcal{L}_{V^r}$ and any $r \in [0,1]$. The fact that the constant C can be chosen uniformly for $\theta \in \mathcal{K}$ follows from recent results on computable bounds for geometrically ergodic Markov chains that show that the constant can be chosen in such a way that it depends only on the constants appearing in the Foster-Lyapunov drift condition and on the minorisation condition over small sets (see Roberts and Tweedie (1999) and Douc et al. (2002) and the references therein).

Remark 2. Poisson's equation has proven to be a fundamental tool for the analysis of additive functionals, in particular to establish limit theorems such as the (functional) central limit theorem (see *e.g.* Benveniste et al. (1990), Nummelin (1991), (Meyn and Tweedie, 1993, Chapter 17), Glynn and Meyn (1996), Duflo (1997)); The existence of solutions to Poisson's equation is well established for geometrically ergodic Markov chains (see Nummelin (1991), (Meyn and Tweedie, 1993, Chapter 17)); it has been more recently proven under assumptions weaker than geometric ergodicity (see (Glynn and Meyn, 1996, Theorem 2.3)).

We now investigate the regularity properties of the solution to Poisson's equation under (A1). Let $W \to [1, \infty)$ and $\delta \in [0, 1]$. We say that the family of functions $\{\psi_\theta : \mathsf{X} \to \mathbb{R}, \theta \in \Theta\}$ is (W, δ)-regular if for any compact subset $\mathcal{K} \subset \Theta$,

$$\sup_{\theta \in \mathcal{K}} \|\psi_\theta\|_W < \infty \quad \text{and} \quad \sup_{(\theta,\theta') \in \mathcal{K} \times \mathcal{K}, \theta \neq \theta'} |\theta - \theta'|^{-\delta} \|\psi_\theta - \psi_{\theta'}\|_W < \infty. \tag{19}$$

The following proposition shows that if $\{\psi_\theta, \theta \in \Theta\}$ is (V^r, β)-regular for some $\beta \in (0, 1]$ and $r \in [0, 1]$, then under (A1) the solutions to Poisson's equation, $\theta \to g_\theta$, are (V^r, α)-regular for any $\alpha \in (0, \beta)$.

Proposition 3. *Assume (A1). Let $\{\psi_\theta, \theta \in \Theta\}$ be (V^r, β)-regular, where β is given in (A1) and $r \in [0, 1]$. For any $(\theta, \theta') \in \Theta \times \Theta$, $\sum_{k=0}^{\infty} |P_{\theta'}^k \psi_\theta - \pi(\psi_\theta)| < \infty$. In addition, for any $\alpha \in (0, \beta)$, $\{g_\theta, \theta \in \Theta\}$ and $\{P_\theta g_\theta, \theta \in \Theta\}$ are (V^r, α)-regular.*

The proof is given in Appendix A.

Remark 3. the regularity of the solutions of Poisson's equation has been studied, under various ergodicity and regularity conditions on the mapping $\theta \mapsto P_\theta$, by Benveniste et al. (1990), and Bartusek (2000). The result of the proposition above improves upon these works.

The following proposition shows how the V-stability of the projection $\{X_k\}$ of $\{Z_k\}$ in Section 2 is implied by the V-stability of homogeneous Markov chains generated by P_θ for a fixed $\theta \in \mathcal{K}$ and $\mathcal{K}$ a compact subset of Θ, provided that one can control the magnitude of the increments $\{\theta_k - \theta_{k-1}\}$ and that $\{\theta_k\}$ stays in $\mathcal{K}$.

Proposition 4. *Assume (A1). Let $\mathcal{K}$ be a compact subset of Θ. There exists a constant C and $\epsilon > 0$ such that for any sequence $\boldsymbol{\rho} = \{\rho_k\}$ and for any $x \in \mathsf{X}$,*

$$\sup_{\theta \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \left\{ V(X_k) \mathbb{1}\{\sigma(\mathcal{K}) \wedge \nu_\epsilon \geq k\} \right\} \leq C V(x), \tag{20}$$

where for $\epsilon > 0$ $\nu_\epsilon = \inf\{k \geq 1, |\theta_k - \theta_{k-1}| > \epsilon\}$. In addition let $s \in \mathbb{N}$, then there exists a constant C_s such that for any sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ satisfying $0 < \epsilon_k \leq \epsilon$, for all $k \geq s$ and for any sequence $\boldsymbol{\rho} = \{\rho_k\}$ and $x \in \mathsf{X}$,

$$\sup_{\theta \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \left\{ V(X_k) \mathbb{1}\{\sigma(\boldsymbol{\epsilon}, \mathcal{K}) \geq k\} \right\} \leq C_s V(x), \tag{21}$$

where $\sigma(\boldsymbol{\epsilon}, \mathcal{K})$ is defined in Eq. (13).

The proof is given in Appendix A.

3.2 Law of large numbers

We prove in this section a law of large numbers (LLN) under $\bar{\mathbb{P}}^{\rho}_{x,\theta}$ for $n^{-1}\sum_{k=1}^{n}\psi_{\theta_k}(X_k)$, where $\{\psi_\theta, \theta \in \Theta\}$ is a set of sufficiently regular functions. It is worth noticing here that it is not required that the sequence $\{\theta_k\}$ converges in order to establish our result. The proof is based on the identity

$$\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \pi(dx)\psi_{\theta_k}(x) = g_{\theta_k}(X_k) - P_{\theta_k} g_{\theta_k}(X_k),$$

where $u = g_\theta$ is a solution of Poisson's equation (18). The decomposition

$$\begin{aligned} g_{\theta_k}(X_k) - P_{\theta_k} g_{\theta_k}(X_k) = &\big(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})\big) + \\ &\big(g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k)\big) + \big(P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) - P_{\theta_k} g_{\theta_k}(X_k)\big) \end{aligned} \quad (22)$$

evidences the different terms that need to be controlled to prove the LLN. The first term in the decomposition is a sequence of martingale differences under $\mathbb{P}^{\rho}_{x,\theta}$, since

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) | \mathcal{F}_{k-1}\} = 0.$$

As we shall see, this is the leading term in the decomposition and the other terms are simple remainders which are easily dealt with thanks to the regularity of the solution to Poisson's equation under (A1). We preface our main result, Theorem 6, with an intermediate proposition concerned with the control of the fluctuations of $n^{-1}\sum_{k=1}^{n}\psi_{\theta_k}(X_k)$ for the inhomogeneous chain $\{(X_k,\theta_k)\}$ under the probability $\mathbb{P}^{\boldsymbol{\rho}}_{x,\theta}$.

Proposition 5. *Assume (A1). Let $\{\psi_\theta, \theta \in \Theta\}$ be a $(V^{1/p}, \beta)$-regular family of functions, for some $p \geq 2$ and V and $\beta \in (0,1]$ defined in (A1). Let $\bar{\boldsymbol{\epsilon}} = \{\bar{\epsilon}_k\}$ be a non-increasing sequence such that $\lim_{k\to\infty}\bar{\epsilon}_k = 0$ and $\mathcal{K}$ a compact subset of Θ. Then for any $\alpha \in (0,\beta)$ there exists a constant C (depending only on α, p, $\bar{\boldsymbol{\epsilon}}$, $\mathcal{K}$ and the constants in (A1)) such that, for any non-increasing sequence $\boldsymbol{\rho} = \{\rho_k\}$ of positive numbers such that $\rho_0 \leq 1$ and any non-increasing sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ satisfying $\epsilon_k \leq \bar{\epsilon}_k$ for all $k \geq 0$, we have, for all $x \in \mathsf{X}$ and $\theta \in \mathcal{K}$,*

$$\begin{aligned} \mathbb{P}^{\boldsymbol{\rho}}_{x,\theta}\Big\{\mathbb{1}\left(\sigma(\boldsymbol{\epsilon},\mathcal{K}) > m\right) \left|\sum_{k=1}^{m}\left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}}\psi_{\theta_k}(x)\pi(dx)\right)\right| \geq \delta\Big\} \leq \\ C\ \delta^{-p} \sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_{V^{1/p}}\ V(x)\ \left\{m^{p/2} + \Big(\sum_{k=1}^{m}\epsilon_k^{\alpha}\Big)^p\right\}, \end{aligned} \quad (23)$$

where $\sigma(\boldsymbol{\epsilon},\mathcal{K})$ is given in (13).

Proof. For notational simplicity, we set $\sigma := \sigma(\boldsymbol{\epsilon},\mathcal{K})$. In this proof C is a constant which only depends upon the constants α, p, the sequence $\bar{\boldsymbol{\epsilon}}$, the compact set $\mathcal{K} \subset \Theta$, and the constants in (A1); this constant may take different values upon each appearance. Proposition 2 shows that there exists a solution g_θ to Poisson's equation $\psi_\theta - \pi(\psi_\theta) = g_\theta - P_\theta g_\theta$. Decompose the sum $\mathbb{1}(\sigma > m)\ \sum_{k=1}^{m}(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}}\psi_{\theta_k}(x)\pi(dx))$ as $\sum_{i=1}^{3} S_m^{(i)}$

where

$$\begin{aligned}
S_m^{(1)} &:= \mathbb{1}(\sigma > m)\ \sum_{k=1}^{m} (g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})) \\
S_m^{(2)} &:= \mathbb{1}(\sigma > m)\ \sum_{k=1}^{m} (g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k)) \\
S_m^{(3)} &:= \mathbb{1}(\sigma > m)\ (P_{\theta_0} g_{\theta_0}(X_0) - P_{\theta_m} g_{\theta_m}(X_m)) \,.
\end{aligned}$$

We consider these terms separately. First note that

$$\mathbb{1}(\sigma > m)\ \left| \sum_{k=1}^{m} \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) \right) \right| \leq |M_m|$$

where

$$M_m := \sum_{k=1}^{m} \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) \right)\ \mathbb{1}(\sigma \geq k).$$

Under Proposition 3, there exists a constant C such that for all $\theta \in \mathcal{K}$ and all $x \in \mathsf{X}$, $|g_\theta(x)| \leq C V^{1/p}(x)$ and $|P_\theta g_\theta(x)| \leq C V^{1/p}(x)$. Hence, by Proposition 4, there exists a constant C such that for all $k \geq 1$,

$$\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \Big\{ (|g_{\theta_{k-1}}(X_k)|^p + |P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})|^p) \mathbb{1}(\sigma \geq k) \} \leq C V(x), \quad \theta \in \mathcal{K}, x \in \mathsf{X}. \tag{24}$$

Since

$$\begin{aligned}
\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \left\{ (g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})) \mathbb{1}(\sigma \geq k) \middle| \, \mathcal{F}_{k-1} \right\} = & \\
\left(P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) \right) \mathbb{1}(\sigma \geq k) = 0, &
\end{aligned}$$

$(M_m, m \geq 0)$ is a $\mathcal{F}$-martingale under $\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}$ with increments bounded in L^p. Using the Burkholder inequality (Hall and Heyde (1980) Theorem 2.10) and Minkowski's inequality, we have

$$\begin{aligned}
\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \left\{ |M_m|^p \right\} &\leq C_p\ \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \Big\{ \Big(\sum_{k=1}^{m} |g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})|^2 \mathbb{1}(\sigma \geq k) \Big)^{p/2} \Big\} \\
&\leq C_p \Big\{ \sum_{k=1}^{m} \Big(\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \Big\{ |g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})|^p \Big)^{2/p} \mathbb{1}(\sigma \geq k) \Big\} \Big\}^{p/2},
\end{aligned} \tag{25}$$

where C_p is a universal constant. Using again Propositions 3 and 4, we obtain

$$\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} \left\{ |M_m|^p | \right\} \leq C m^{p/2} V(x), \quad \theta \in \mathcal{K}, x \in \mathsf{X}.$$

We consider now $S_m^{(2)}$. Proposition 3 shows that for any $\alpha \in (0, \beta)$, $\{g_\theta, \theta \in \Theta\}$ is $(V^{1/p}, \alpha)$-regular. Hence there exists C such that

$$\mathbb{1}(\sigma > m)\ \left| \sum_{k=1}^{m} \left\{ g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k) \right\} \right| \leq C \sum_{k=1}^{m} \epsilon_k^{\alpha} V^{1/p}\, (X_k)\, \mathbb{1}(\sigma \geq k).$$

Now Minkowski's inequality, Propositions 3 and 4 show that there exists C such that

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{|S^{(2)}_m|^p\} \leq C \left(\sum_{k=1}^m \epsilon_k^\alpha\right)^p V(x), \quad \theta \in \mathcal{K}, x \in \mathsf{X}. \tag{26}$$

Consider now $S^{(3)}_m$. From Proposition 3 and 4 there exists a constant C such that, for all $\theta \in \mathcal{K}$ and all $x \in \mathsf{X}$,

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left\{\mathbb{1}(\sigma > m)\left|P_{\theta_m} g_{\theta_m}(X_m)\right|^p\right\} \leq CV(x).$$

We can now conclude using Minkowski's and Markov's inequalities. □

We can now state our main consistency result, under the assumption that the number of reinitialization κ_n is almost surely finite.

Theorem 6. *Assume (A1) and that the sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ defined in Section 2 is non-increasing and $\lim_{k\to\infty} \epsilon_k = 0$. Let $p \geq 2$, and let $\{\psi_\theta, \theta \in \Theta\}$ be a $(V^{1/p}, \beta)$-regular family of functions where β is given in (A1). Assume in addition that for all $(x, \theta) \in \mathsf{X} \times \Theta$, $\bar{\mathbb{P}}_{x,\theta}\{\lim_{n\to\infty} \kappa_n < \infty\} = 1$. Then, for any $(x, \theta) \in \mathsf{X} \times \Theta$, we have*

$$n^{-1} \sum_{k=1}^n \left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \psi_{\theta_k}(x)\pi(dx)\right) \to_{\bar{\mathbb{P}}_{x,\theta}} 0. \tag{27}$$

If $p > 2$ and in addition $\boldsymbol{\epsilon} = \{\epsilon_k\}$ satisfies

$$\sum_{k=1}^\infty k^{-1/2} \epsilon_k^\alpha < \infty \tag{28}$$

for some $\alpha \in (0, \beta)$ where $\beta \in (0, 1]$ is defined in (A1-ii), then

$$n^{-1} \sum_{k=1}^n \left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \psi_{\theta_k}(x)\pi(dx)\right) \to 0 \quad \bar{\mathbb{P}}_{x,\theta} - a.s. \tag{29}$$

Proof. Without loss of generality, we may assume that, for any $\theta \in \Theta$, $\int_{\mathsf{X}} \psi_\theta(x)\pi(dx) = 0$. Put $S_n = \sum_{k=1}^n \psi_{\theta_k}(X_k) = S^{(1)}_n + S^{(2)}_n$, where

$$S^{(1)}_n = \sum_{k=1}^{T_{\kappa_n}} \psi_{\theta_k}(X_k) \quad \text{and} \quad S^{(2)}_n = \sum_{k=T_{\kappa_n}+1}^{n} \psi_{\theta_k}(X_k),$$

and the T_k's are defined in (12). We consider these two terms separately. Because $S^{(1)}_n$ has $\bar{\mathbb{P}}_{x,\theta}$-*a.e* finitely many terms, $n^{-1}S^{(1)}_n \to 0$, $\bar{\mathbb{P}}_{x,\theta}$-a.s. Consider now $S^{(2)}_n$. Define $\kappa_\infty = \lim_{n\to\infty} \kappa_n$ and $T_\infty = \lim_{n\to\infty} T_{\kappa_n}$. Under the stated assumption, for any $\eta > 0$, there exists K and L such that

$$\bar{\mathbb{P}}_{x,\theta}\{\kappa_\infty \geq K\} \leq \eta/3 \quad \text{and} \quad \bar{\mathbb{P}}_{x,\theta}\{T_\infty \geq L\} \leq \eta/3. \tag{30}$$

To prove (27), it is sufficient to show that for sufficiently large n,

$$\bar{\mathbb{P}}_{x,\theta}\left\{n^{-1}|S^{(2)}_n| \geq \delta, T_\infty \leq L, \kappa_\infty \leq K\right\} \leq \eta/3. \tag{31}$$

By Lemma 7 (stated and proven below) and Proposition 5, there exists a constant C such that

$$\bar{\mathbb{P}}_{x,\theta}\left\{n^{-1}|S_n^{(2)}| \geq \delta, T_{\kappa_n} \leq L, \kappa_n \leq K\right\} \leq C\delta^{-p} \sup_{\theta \in \mathcal{K}_K} \|\psi_\theta\|_{V^{1/p}} \left\{n^{-p/2} + \left(n^{-1}\sum_{k=1}^{n} \epsilon_k^\alpha\right)^p\right\},$$

which shows (31). Now, in order to prove the strong law of large numbers in (29) it is sufficient to show that

$$\bar{\mathbb{P}}_{x,\theta}\left\{\sup_{l \geq n} l^{-1}|S_l^{(2)}| \geq \delta, T_\infty \leq L, \kappa_\infty \leq K\right\} \leq \eta/3,$$

and invoke the Borel-Cantelli lemma. From Proposition 5 we have

$$\bar{\mathbb{P}}_{x,\theta}\left\{\sup_{l \geq n} l^{-1}|S_l^{(2)}| \geq \delta, T_\infty \leq L, \kappa_\infty \leq K\right\} \leq$$
$$C\delta^{-p} \sup_{\theta \in \mathcal{K}_K} \|\psi_\theta\|_{V^{1/p}} \sum_{l \geq n} \left(l^{-p/2} + l^{-p/2}\left(l^{-1/2}\sum_{k=1}^{l} \epsilon_k^\alpha\right)^p\right),$$

and the proof is concluded with Kronecker's lemma (which shows that the condition $\sum_{k=1}^\infty k^{-1/2}\epsilon_k^\alpha < \infty$ implies that $n^{-1/2}\sum_{k=1}^n \epsilon_k^\alpha \to 0$) and the fact that $p > 2$. □

Lemma 7. *Let $\{\psi_\theta, \theta \in \Theta\}$ be a family of measurable functions $\psi : \Theta \times \mathsf{X} \to \mathbb{R}$ and $K, L \in \mathbb{N}$. Then, for all $x, \theta \in \mathsf{X} \times \Theta$ and $\delta > 0$,*

$$\bar{\mathbb{P}}_{x,\theta}\left\{\left|\sum_{k=T_{\kappa_n}+1}^{n} \psi_{\theta_k}(X_k)\right| \geq \delta, T_{\kappa_n} \leq L, \kappa_n \leq K\right\} \leq$$
$$\sum_{j=0}^{K} \bar{\mathbb{E}}_{x,\theta}\left\{A_{n-T_j}(X_{T_j}, \theta_{T_j}, \kappa_{T_j}, \varsigma_{T_j})\mathbb{1}\{T_j \leq L\}\right\},$$

where for any integers κ, ς, any $x, \theta \in \mathsf{X} \times \Theta$ and $\delta > 0$,

$$A_m(x, \delta, \theta, \kappa, \varsigma) = \mathbb{P}_{\Phi(x,\theta)}^{\boldsymbol{\gamma}^{\leftarrow \varsigma}}\left\{\mathbb{1}(\sigma(\boldsymbol{\epsilon}^{\leftarrow \varsigma}, \mathcal{K}_\kappa) \geq m)\left|\sum_{k=1}^{m} \psi_{\theta_k}(X_k)\right| \geq \delta\right\}. \tag{32}$$

Proof. For any integers l and j, we have

$$\left\{\left|\sum_{k=T_{\kappa_n}+1}^{n} \psi_{\theta_k}(X_k)\right| \geq \delta, T_{\kappa_n} = l, \kappa_n = j\right\} \subset \left\{\mathbb{1}(T_{j+1} \geq n)\left|\sum_{k=T_j+1}^{n} \psi_{\theta_k}(X_k)\right| \geq \delta, T_j = l\right\}.$$

Writing $T_{j+1} = T_j + T_1 \circ \tau^{T_j}$, where τ is the shift operator on the canonical space of the chain $\{Z_n\}$, and noting that by construction $T_1 = \sigma(\epsilon^{\leftarrow \varsigma_0}, \mathcal{K}_{\kappa_0})$, we have

$$\left\{\mathbb{1}(T_{j+1} \geq n)\left|\sum_{k=T_j+1}^{n} \psi_{\theta_k}(X_k)\right| \geq \delta, T_j = l\right\} \subset$$
$$\left\{\mathbb{1}(\sigma(\epsilon^{\leftarrow \varsigma_0}, \mathcal{K}_{\kappa_0}) \geq n - l)\left|\sum_{k=1}^{n-l} \psi_{\theta_k}(X_k)\right| \circ \tau^{T_j} \geq \delta, T_j = l\right\}.$$

By Lemma 1 for any $(x,\theta) \in \mathsf{X} \times \Theta$ and any integers ς and κ, we have

$$\bar{\mathbb{P}}_{x,\theta,\kappa,\varsigma,0}\left\{\mathbb{1}(\sigma(\epsilon^{\leftarrow\varsigma},\mathcal{K}_\kappa) \geq m)\left|\sum_{k=1}^{m}\psi_{\theta_k}(X_k)\right| \geq \delta\right\} = A_m(x,\delta,\theta,\kappa,\varsigma),$$

where A_m is given in (32). The proof is concluded by applying the strong Markov property. □

Remark 4. It is worth noticing that for the above proposition to hold it is not necessary that the sequence $\{\theta_k\}$ converges, but simply that $\bar{\mathbb{P}}_{x,\theta}(\lim_{n\to\infty}\kappa_n < \infty)$ which implies that for all $k \geq T_{\kappa_\infty}$, $\theta_k \in \mathcal{K}_{\kappa_\infty}$ and $|\theta_{k+1} - \theta_k| \leq \epsilon_{\varsigma_k}$ *i.e.*, empirical averages are consistent whenever the sequence $\{\theta_k\}$ stays within a compact subset of Θ and the difference between two successive values of the parameter decreases to zero.

Remark 5. Checking $\bar{\mathbb{P}}_{x,\theta}(\lim_{n\to\infty}\kappa_n < \infty) = 1$ depends on the particular algorithm used to update the parameters. Verifiable conditions have been established in Andrieu et al. (2002) to check the stability of the algorithm; see Sections 5, 6 and 7.

4 Invariance principle

We now study the asymptotic fluctuations of $I_n(\psi)$ and prove an invariance principle. As it is the case for homogeneous Markov chains, more stringent conditions are required here than for the simple LLN. In particular we will require here that the series $\{\theta_k\}$ converges $\bar{\mathbb{P}}^{\rho}_{x,\theta}$-a.e. This is in contrast with simple consistency for which boundedness and convergence to zero of the increments of $\{\theta_k\}$ was sufficient. The main idea of the proof consists of approximating $n^{-1/2}(I_n(\psi) - \pi(\psi))$ with a triangular array of martingale differences sequence, and then apply an invariance principle for martingale differences to show the desired result.

Theorem 8. *Assume (A1) and that the sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ satisfies* (28) *for some $\alpha \in (0,\beta)$, where β is given in (A1). Let $p > 2$ and $\psi \in \mathcal{L}_{V^{1/(2p)}}$. Assume that, for all $(x,\theta) \in \mathsf{X} \times \Theta$, $\bar{\mathbb{P}}_{x,\theta}(\lim_{n\to\infty}\kappa_n < \infty)$ and that there exists a random variable θ_∞ such that, **(i)** $\bar{\mathbb{P}}_{x,\theta}(\limsup_{n\to\infty}|\theta_n - \theta_\infty| = 0) = 1$, **(ii)** $\bar{\mathbb{P}}_{x,\theta}(\theta_\infty \in \Theta) = 1$ and **(iii)** $\sup\left\{M \geq 0, \bar{\mathbb{P}}_{x,\theta}\left(V(\theta_\infty,\psi) \geq M\right) = 1\right\} > 0$, where for any $\theta \in \Theta$,*

$$V^2(\theta,\psi) = \pi\{g_\theta^2 - \pi\{(P_\theta g_\theta)^2\}\} \quad \textit{with} \quad g_\theta := \sum_{k=0}^{\infty} P_\theta^k \psi - \pi(\psi). \tag{33}$$

Then, for all $(x,\theta) \in \mathsf{K} \times \mathcal{K}_0$ and $u \in \mathbb{R}$,

$$\lim_{n\to\infty}\left|\bar{\mathbb{P}}_{x,\theta}\left(n^{-1/2}\sum_{k=1}^{n}(\psi(X_k) - \pi(\psi)) \leq u\right) - \bar{\mathbb{E}}_{x,\theta}\left\{\Phi\left(\{V(\theta_\infty,\psi)\}^{-1}u\right)\right\}\right| = 0, \tag{34}$$

where Φ is the standard normal distribution function.

Proof. Without loss of generality, assume that $\pi(\psi) = 0$. By Proposition 3, for any $\theta \in \Theta$, there exists $g_\theta \in \mathcal{L}_{V^{1/(2p)}}$ satisfying $\psi = g_\theta - P_\theta g_\theta$. Define for $k \geq 1$

$$\xi_{k,n} := n^{-1/2}\left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1})\right). \tag{35}$$

Using (22), we have $n^{-1/2}\sum_{k=1}^n \psi(X_k) - \sum_{k=1}^n \xi_{k,n} = S_n^{(1)} + S_n^{(2)}$ where

$$S_n^{(1)} = n^{-1/2}\sum_{k=1}^{T_{\kappa_n}}(g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k)) + n^{-1/2}(P_{\theta_0}g_{\theta_0}(X_0) - P_{\theta_n}g_{\theta_n}(X_n)),$$
$$S_n^{(2)} = n^{-1/2}\sum_{k=T_{\kappa_n}+1}^{n}(g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k)).$$

As above, since $\bar{\mathbb{P}}_{x,\theta}(T_{\kappa_\infty} < \infty) = 1$, $S_n^{(1)} \to 0$, $\bar{\mathbb{P}}_{x,\theta}$-a.s. Consider now $S_n^{(2)}$. Choose K and L as in (30). We only have to show that for n large enough,

$$\bar{\mathbb{P}}_{x,\theta}\left\{ n^{-1/2}\left|\sum_{k=T_{\kappa_n}+1}^{n-1} g_{\theta_k}(X_k) - g_{\theta_{k-1}}(X_k)\right| \geq \delta, \kappa_n < K, T_{\kappa_n} < L \right\} \leq \eta/3.$$

By Lemma 7 and using (26), there exists a constant C such that

$$\bar{\mathbb{P}}_{x,\theta}\left\{|S_n^{(2)}| \geq \delta, \kappa_n < K, T_{\kappa_n} < L\right\} \leq C\delta^{-p}\, n^{-p/2}\left(\sum_{k=1}^n \epsilon_k^\alpha\right)^p .$$

Since under (28) we have $n^{-1/2}\sum_{k=1}^n \epsilon_k^\alpha \to 0$, by combining the results above we obtain

$$n^{-1/2}\sum_{k=1}^n \psi(X_k) - \sum_{k=1}^n \xi_{k,n} \to_{\bar{\mathbb{P}}_{x,\theta}} 0. \tag{36}$$

Define

$$\bar{\xi}_{k,n} := n^{-1/2}\left(g_{\theta_{k-1}}(X_k) - \bar{\mathbb{E}}_{x,\theta}^{\boldsymbol{\rho}}\left\{g_{\theta_{k-1}}(X_k) \mid \mathcal{G}_{k-1}\right\}\right), \tag{37}$$

where $(\mathcal{G}_k, k \geq 0)$ is defined in Section 2. By definition, $\{\bar{\xi}_{k,n}\}$ is an array of martingale differences with respect to $\bar{\mathbb{P}}_{x,\theta}$. Because $\sum_{k=1}^n \xi_{k,n}$ and $\sum_{k=1}^n \bar{\xi}_{k,n}$ only differ by a $\bar{\mathbb{P}}_{x,\theta}$-a.s. finite number of $\bar{\mathbb{P}}_{x,\theta}$-a.s. finite terms, their difference converges to 0. Consequently from (36) we have

$$n^{-1/2}\left(\sum_{k=1}^n \psi(X_k)\right) - \sum_{k=1}^n \bar{\xi}_{k,n} \;\to_{\bar{\mathbb{P}}_{x,\theta}}\; 0, \tag{38}$$

showing that $n^{-1/2}\sum_{k=1}^n \psi(X_k)$ can be approximated by a triangular array of martingale differences.

We now apply the triangular zero-mean martingale central limit theorem (Hall and Heyde, 1980, Chapter 3) to the approximating term above in order to show (34). This requires one to establish that

$$\sum_{k=1}^n \bar{\xi}_{k,n}^2 \;\to_{\bar{\mathbb{P}}_{x,\theta}}\; V^2(\theta_\infty, \psi), \tag{39}$$

$$\text{for all}\, \delta > 0, \quad \sum_{k=1}^n \bar{\xi}_{k,n}^2\, \mathbb{1}(|\bar{\xi}_{k,n}| > \delta) \;\to_{\bar{\mathbb{P}}_{x,\theta}}\; 0. \tag{40}$$

Using the arguments to prove that $\sum_{k=1}^n \bar{\xi}_{k,n} - \sum_{k=1}^n \xi_{k,n} \to_{\bar{\mathbb{P}}_{x,\theta}} 0$ we have that $\sum_{k=1}^n \bar{\xi}_{k,n}^2 - \sum_{k=1}^n \xi_{k,n}^2 \to_{\bar{\mathbb{P}}_{x,\theta}} 0$ and a straightforward adaptation of Theorem 6 shows that

$$\sum_{k=1}^n \xi_{k,n}^2 - n^{-1} \sum_{k=1}^n \pi\{g_{\theta_{k-1}}^2 - \pi\{(P_{\theta_{k-1}} g_{\theta_{k-1}})^2\}\} \to_{\bar{\mathbb{P}}_{x,\theta}} 0.$$

The proof of (39) follows from the continuity of $\theta \to V(\theta, \psi)$. For any $\tau \in (0, p-2)$ we have

$$\begin{aligned}\sum_{k=1}^n \xi_{k,n}^2 \mathbb{1}(|\xi_{k,n}| \geq \delta) &\leq \delta^{-\tau} \sum_{k=1}^n |\xi_{k,n}|^{2+\tau} \\ &\leq 2^{1+\tau} \delta^{-\tau} n^{-\tau/2} n^{-1} \sum_{k=1}^n \left(|g_{\theta_{k-1}}(X_k)|^{2+\tau} + |P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k)|^{2+\tau} \right) \to_{\bar{\mathbb{P}}_{x,\theta}} 0\end{aligned}$$

by Theorem 6. The proof of negligibility follows since

$$\sum_{k=1}^n \left(\xi_{k,n}^2 \mathbb{1}(|\xi_{k,n}| \geq \delta) - \bar{\xi}_{k,n}^2 \mathbb{1}(|\bar{\xi}_{k,n}| \geq \delta) \right) \to_{\bar{\mathbb{P}}_{x,\theta}} 0.$$

□

Remark 6. It is still possible to obtain an invariance principle when $\{\theta_k\}$ does not converge but remains bounded and satisfies $\lim_{k\to\infty} |\theta_{k+1} - \theta_k| = 0$. In such a case, the normalization is no longer $\sqrt{n}$ but $\sqrt{\sum_{k=1}^n (g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k))^2}$ (see e.g. Hall (1977)). Moreover, by resorting to results on the rate of convergence for CLT of triangular array of martingale differences, it is possible to determine the rate of convergence in (34) (see for example Hall and Heyde (1981))

5 Stability and convergence of the stochastic approximation process

In order to conclude the part of this paper dedicated to the general theory of adaptive MCMC algorithm, we now present generally verifiable conditions under which the number of reinitializations of the algorithm that produces the Markov chain $\{Z_k\}$ described in Section 2 is $\bar{\mathbb{P}}_{x,\theta}$-a.e. finite. This is a difficult problem *per se*, which has been worked out in a companion paper, Andrieu et al. (2002). We here briefly introduce the conditions under which this key property is satisfied and give (without proof) the main stability result. The reader should refer to Andrieu et al. (2002) for more details.

As mentioned in the introduction, the convergence of the stochastic approximation procedure is closely related to the stability of the noiseless sequence $\bar{\theta}_{k+1} = \bar{\theta}_k + \gamma_{k+1} h(\bar{\theta}_k)$. A practical technique to prove the stability of the noiseless sequence consists of finding a Lyapunov function $w : \Theta \to [0, \infty)$ such that $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$, where ∇w denotes the gradient of w with respect to θ and for $u, v \in \mathbb{R}^n$, $\langle u, v \rangle$ is their Euclidian inner product (we will later on also use the notation $|v| = \sqrt{\langle v, v \rangle}$ to denote the Euclidean norm of v). This indeed shows that the noiseless sequence $\{w(\bar{\theta}_k)\}$ eventually decreases, showing that

$\lim_{k\to\infty} w(\bar{\theta}_k)$ exists. It should therefore not be surprising if such a Lyapunov function can play an important role in showing the stability of the noisy sequence $\{\theta_k\}$. With this in mind, we can now detail the conditions required to prove our convergence result.

(**A2**) Θ is an open subset of $\mathbb{R}^{n_\theta}$. The mean field $h : \Theta \to \mathbb{R}^{n_\theta}$ is continuous and there exists a continuously differentiable function $w : \Theta \to [0, \infty)$ (with the convention $w(\theta) = \infty$ when $\theta \notin \Theta$) such that:

(i) For any integer M, the level set $\mathcal{W}_M = \{\theta \in \Theta, w(\theta) \leq M\} \subset \Theta$ is compact,

(ii) The set of stationary point $\mathcal{L} = \{\theta \in \Theta, \langle \nabla w(\theta), h(\theta)\rangle = 0\}$ belongs to the interior of Θ,

(iii) For any $\theta \in \Theta$, $\langle \nabla w(\theta), h(\theta)\rangle \leq 0$ and $w(\mathcal{L})$ has an empty interior.

Some continuity conditions on the field of the algorithm are needed.

(**A3**) $\{H_\theta, \theta \in \Theta\}$ is $(V^{1/p}, \beta)$-regular for some $p \geq 2$ with V and β defined in (A1).

Finally we require some conditions on the sequence of stepsizes $\boldsymbol{\gamma} = \{\gamma_k\}$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$.

(**A4**) The sequences $\boldsymbol{\gamma} = \{\gamma_k\}$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$ are non-increasing, positive, $\lim_{k\to\infty} \epsilon_k = 0$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and there exists $\alpha \in (0, \beta)$ such that

$$\sum_{k=1}^{\infty} \left\{ \gamma_k^2 + k^{-1/2} \epsilon_k^\alpha + (\epsilon_k^{-1} \gamma_k)^p \right\} < \infty,$$

where β and p are defined in (A1) and (A3) respectively.

The following theorem (see Andrieu et al. (2002)) shows that the tail probability of the number of reinitialisations decreases faster than any exponential and that the parameter sequence $\{\theta_k\}$ converges to the stationary set $\mathcal{L}$.

Theorem 9. *Assume (A1-4). Then, for any subset* $\mathsf{K} \subset \mathsf{X}$ *such that* $\sup_{x\in\mathsf{K}} V(x) < \infty$,

$$\limsup_k k^{-1} \log \left(\sup_{(x,\theta)\in \mathsf{K}\times\mathcal{K}_0} \bar{\mathbb{P}}_{x,\theta} \left\{ \sup_n \kappa_n \geq k \right\} \right) = -\infty,$$

$$\bar{\mathbb{P}}_{x_0,\theta_0} \left\{ \lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0 \right\}, \quad \forall x_0 \in \mathsf{K}, \forall \theta_0 \in \mathcal{K}_0.$$

6 Consistency and invariance principle for the adaptive N-SRWM kernel

In this section we show how our results can be applied to the adaptive N-SRWM algorithm proposed by Haario et al. (2001) and described in Section 1. We first illustrate how the conditions required to prove the LLN in Haario et al. (2001) can be alleviated. In particular no boundedness condition is required on the parameter set Θ, but rather conditions on the tails of the target distribution π. We then extend these results further and prove a central limit theorem (Theorem 13).

In view of the results proved above it is required

(a) to prove the ergodicity and regularity conditions for the Markov kernels outlined in assumption (A1)

(b) to prove that the reinitializations occur finitely many times (stability) and that $\{\theta_k\}$ eventually converges. Note again that the convergence property is only required for the CLT.

We first focus on (a). The geometric ergodicity of the RWMH kernel has been studied by Roberts and Tweedie (1996) and refined by Jarner and Hansen (2000); the regularity of the RWMH has, to the best of our knowledge, not been considered in the literature. The geometric ergodicity of the RWMH kernel mainly depends on the tail properties of the target distribution π. We will therefore restrict our discussion to target distributions that satisfy the following set of conditions. These are not minimal but easy to check in practice (see Jarner and Hansen (2000) for details).

(M) The probability density π has the following properties:

(i) It is bounded, bounded away from zero on every compact set and continuously differentiable.

(ii) It is super-exponential, *i.e.*

$$\lim_{|x|\to+\infty} \left\langle \frac{x}{|x|}, \nabla \log \pi\left(x\right) \right\rangle = -\infty.$$

(iii) The contours $\partial \mathsf{A}\left(x\right) = \{y : \pi(y) = \pi(x)\}$ are asymptotically regular, *i.e.*

$$\lim_{|x|\to+\infty} \sup \left\langle \frac{x}{|x|}, \frac{\nabla \pi\left(x\right)}{|\nabla \pi\left(x\right)|} \right\rangle < 0.$$

We now establish uniform minorisation and drift conditions for P_q^{SRW} defined in Eq. (3). Let $\mathcal{M}(\mathsf{X})$ denote the set of probability densities w.r.t. the Lebesgue measure λ^{Leb}. Let $\varepsilon > 0$ and $\delta > 0$ and define the subset $\mathcal{K}_{\delta,\varepsilon} \subset \mathcal{M}(\mathsf{X})$,

$$\mathcal{K}_{\delta,\varepsilon} = \left\{q \in \mathcal{M}(\mathsf{X}), q(z) = q(-z) \quad \text{and} \quad |z| \leq \varepsilon \Rightarrow q(z) \geq \delta\right\}. \tag{41}$$

Proposition 10. *Assume (M). For any $\eta \in (0,1)$, set $V = \pi^{-\eta}/(\sup_{x\in\mathsf{X}} \pi(x))^{-\eta}$. Then,*

1. *Any non-empty compact set $\mathsf{C} \subset \mathsf{X}$ is a (1,δ)-small set for some $\delta > 0$ and some measure ν,*
$$\forall (x, A) \in \mathsf{C} \times \mathcal{B}(\mathsf{X}) \qquad \inf_{q\in\mathcal{K}_{\delta,\varepsilon}} P_q^{\mathrm{SRW}}(x, A) \geq \delta\nu(A). \tag{42}$$

2. *Furthermore, for any $\delta > 0$ and $\varepsilon > 0$,*

$$\begin{aligned} \sup_{q\in\mathcal{K}_{\delta,\varepsilon}} \limsup_{|x|\to+\infty} \frac{P_q^{\mathrm{SRW}} V(x)}{V(x)} &< 1, & (43) \\ \sup_{(x,q)\in\mathsf{X}\times\mathcal{K}_{\delta,\varepsilon}} \frac{P_q^{\mathrm{SRW}} V(x)}{V(x)} &< +\infty. & (44) \end{aligned}$$

3. *Let $q, q' \in \mathcal{M}(\mathsf{X})$ be two symmetric probability distributions. Then, for any $r \in [0,1]$ and any $g \in \mathcal{L}_{V^r}$ we have*

$$\left\|P_q^{\mathrm{SRW}} g - P_{q'}^{\mathrm{SRW}} g\right\|_{V^r} \leq 2\,\|g\|_{V^r} \int_{\mathsf{X}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz). \tag{45}$$

The proof is in Appendix B.

As an example of application one can again consider the adaptive N-SRWM introduced earlier in Section 1, where the proposal distribution q_Γ is zero-mean Gaussian with covariance matrix Γ. In the following lemma, we show that the mapping $\Gamma \to P_{q_\Gamma}$ is Lipschitz continuous.

Lemma 11. *Let $\mathcal{K}$ be a convex compact subset of $\mathcal{C}_+^{n_x}$ and set $V = \pi^{-\eta}/(\sup_{\mathsf{X}} \pi)^{-\eta}$ for some $\eta \in (0,1)$. For any $r \in [0,1]$, any $\Gamma, \Gamma' \in \mathcal{K} \times \mathcal{K}$, $g \in \mathcal{L}_{V^r}$, we have*

$$\left\|P_{q_\Gamma}^{\mathrm{SRW}} g - P_{q_{\Gamma'}}^{\mathrm{SRW}} g\right\|_{V^r} \leq \frac{2 n_x}{\lambda_{\min}(\mathcal{K})} \|g\|_{V^r} |\Gamma - \Gamma'|,$$

where $\lambda_{\min}(\mathcal{K})$ is the minimum possible eigenvalue for matrices in $\mathcal{K}$.

The proof is in Appendix C. We now turn on to proving that the stochastic approximation procedure outlined by Haario et al. (2001) is ultimately pathwise bounded and eventually converges. In the case of the algorithm proposed by Haario et al. (2001), the parameter estimates μ_k and Γ_k take the form of maximum likelihood estimates under the i.i.d. multivariate Gaussian model. It therefore comes as no surprise if the appropriate Lyapunov function is

$$w(\mu, \Gamma) = K(\pi, q_{\mu,\Gamma}), \tag{46}$$

the Kullback-Leibler divergence between the target density π and a normal density $q_{\mu,\Gamma}$ with mean μ and covariance Γ. Using straightforward algebra, we have

$$\begin{aligned}\langle \nabla w(\mu,\Gamma), h(\mu,\Gamma)\rangle &= -2(\mu - \mu_\pi)^{\mathrm{T}} \Gamma^{-1} (\mu - \mu_\pi) \\ &\quad - \mathrm{Tr}(\Gamma^{-1}(\Gamma - \Gamma_\pi)\Gamma^{-1}(\Gamma - \Gamma_\pi)) - \left((\mu - \mu_\pi)^{\mathrm{T}} \Gamma^{-1} (\mu - \mu_\pi)\right)^2, \end{aligned} \tag{47}$$

that is $\langle \nabla w(\theta), h(\theta)\rangle \leq 0$ for any $\theta := (\mu, \Gamma) \in \Theta$, with equality if and only if $\Gamma = \Gamma_\pi$ and $\mu = \mu_\pi$. The situation is in this case simple as the set of stationary points $\{\theta \in \Theta, h(\theta) = 0\}$ is reduced to a single point, and the Lyapunov function goes to infinity as $|\mu| \to \infty$ or Γ goes to the boundary of the cone of positive matrices.

Now it can be shown that these results lead to the following intermediate lemma, see Andrieu et al. (2002) for details.

Lemma 12. *Assume (M). Then, (A2)-(A3) are verified with H, h and w defined respectively in Eq.* (5), *Eq.* (8) *and Eq.* (46). *In addition, the set of stationary points $\mathcal{L} := \{\theta \in \Theta := \mathbb{R}^{n_x} \times \mathcal{C}_+^{n_x}, \langle \nabla w(\theta), h(\theta)\rangle = 0\}$ is reduced to a single point $\theta_\pi = (\mu_\pi, \Gamma_\pi)$ whose components are respectively the mean and the covariance of the stationary distribution π.*

From this lemma, we deduce our main theorem for this section.

Theorem 13. *Consider the adaptive N-SRWM of Haario et al. (2001) as described in Section 1, with reprojections as in Section 2. Assume (A4), (M) and let* $\psi \in \mathcal{L}(V^r)$ *where* $V = \pi^{-1}/(\sup_{x \in \mathsf{X}} \pi(x))^{-1}$*. Then,*

1. *If* $r \in [0, 1/2)$*, a strong LLN holds,* i.e.

$$n^{-1} \sum_{k=1}^{n} \left(\psi_{\theta_k}(X_k) - \int_{\mathsf{X}} \psi_{\theta_k}(x) \pi(dx) \right) \to 0 \quad \bar{\mathbb{P}}_{x,\theta} - a.s. \tag{48}$$

2. *If* $r \in [0, 1/4)$ *a CLT holds,* i.e.

$$\lim_{n \to \infty} \left| \bar{\mathbb{P}}_{x,\theta} \left(n^{-1/2} \sum_{k=1}^{n} (\psi(X_k) - \pi(\psi)) \leq u \right) - \Phi\left([V(\theta_\pi, \psi)]^{-1} u \right) \right| = 0,$$

where $\theta_\pi = (\mu_\pi, \Gamma_\pi)$ *and* $V(\theta_\pi, \psi)$ *is defined in* (33).

We refer the reader to Haario et al. (2001) for applications of this type of algorithm to various settings.

7 Application: matching π with mixtures

7.1 Setup

The independence Metropolis-Hastings algorithm (IMH) corresponds to the case where the proposal distribution used in a MH transition probability does not depend on the current state of the MCMC chain, *i.e.* $q(x, y) = q(y)$ for some density $q \in \mathcal{M}(\mathsf{X})$. The transition kernel of the Metropolis algorithm is then given for $x \in \mathsf{X}$ and $A \in \mathcal{B}(\mathsf{X})$ by

$$P_q^{\mathrm{IMH}}(x, A) = \int_A \alpha_q(x, y) q(y) \, \lambda^{\mathrm{Leb}}(dy) + \mathbb{1}_A(x) \int_{\mathsf{X}} (1 - \alpha_q(x, y)) \, q(y) \, \lambda^{\mathrm{Leb}}(dy)$$
$$\text{with} \quad \alpha_q(x, y) = 1 \wedge \frac{\pi(y) q(x)}{\pi(x) q(y)}. \tag{49}$$

Irreducibility of Markov chains built on this model naturally require that $q(x) > 0$ whenever $\pi(x) > 0$. In fact the performance of the IMH depends on how well the proposal distribution mimics the target distribution. More precisely it has been shown in Mengersen and Tweedie (1996) that the IMH sampler is geometrically ergodic if and only if there exists $\varepsilon > 0$ such that $q \in \mathcal{Q}_{\varepsilon,\pi} \subset \mathcal{M}(\mathsf{X})$, where

$$\mathcal{Q}_{\varepsilon,\pi} = \left\{ q \in \mathcal{M}(\mathsf{X}) : \lambda^{\mathrm{Leb}} \left(\{ x \in \mathsf{X} : q(x)/\pi(x) \leq \varepsilon \} \right) = 0 \right\}. \tag{50}$$

This condition implies that the whole state space X is a $(1, \varepsilon)$-small set. In practice, it can be difficult to construct a proposal that ensures efficient sampling. This has motivated several algorithms which aim at adapting the proposal distribution on the fly (see e.g. Gilks et al. (1998), Gåsemyr (2003)). The adaptive procedure we use is different from the aforementioned constructions, but shares the same ultimate goal of matching the proposal distribution π with q. As a measure of fitness, it is natural to consider the Kullback-Leibler divergence between the target distribution π and the proposal distribution q,

$$K(\pi\|q) = \int_{\mathsf{X}} \pi(x) \log \frac{\pi(x)}{q(x)} \lambda^{\mathrm{Leb}}(dx). \tag{51}$$

Realistic algorithms rely on finite dimensional parametrizations. More precisely, let $\Xi \subset \mathbb{R}^{n_\xi}$, $\mathsf{Z} \subset \mathbb{R}^{n_z}$ for some integers n_ξ and n_z and define the following family of exponential densities (defined with respect to the product measure $\lambda^{\mathrm{Leb}} \otimes \lambda$ for some measure λ on Z)

$$\mathcal{E}_c = \left\{ f : f_\xi(x,z) = \exp\left\{-\psi(\xi) + \langle T(x,z), \phi(\xi)\rangle\right\}; \xi, x, z \in \Xi \times \mathsf{X} \times \mathsf{Z} \right\}.$$

where $\psi : \Xi \to \mathbb{R}, \phi : \Xi \to \mathbb{R}^{n_\theta}$ and $T : \mathsf{X} \times \mathsf{Z} \to \mathbb{R}^{n_\theta}$. Define $\mathcal{E}$ the set of densities $\tilde{q}_\xi$ that are marginals of densities from $\mathcal{E}_c$, *i.e.* such that for any $\xi, x \in \Xi \times \mathsf{X}$ we have

$$\tilde{q}_\xi(x) = \int_{\mathsf{Z}} f_\xi(x,z) \lambda(dz). \tag{52}$$

This family of densities covers in particular finite mixtures of multivariate normal distributions, and more generally finite and infinite mixtures of distributions in the exponential family. Here, the variable z plays the role of the *label* of the class, which is not observed (see *e.g.* Titterington et al. (1985)). Using standard missing data terminology, $f_\xi(x,z)$ is the *complete data likelihood* and $\tilde{q}_\xi$ is the associated *incomplete data likelihood*, which is the marginal of the complete data likelihood with respect to the class labels. When the number of observations is fixed, a classical approach to estimate the parameters of a mixture distribution consists of using the expectation-maximisation algorithm (EM).

7.2 Classical EM algorithm

The classical EM algorithm is an iterative procedure which consists of two steps. Given n independent samples $(X_1, \ldots, X_n)$ distributed marginally according to π: (1) *Expectation step*: calculate the conditional expectation of the complete data log-likelihood given the observations and ξ_k, the estimate of ξ at iteration k,

$$\xi \mapsto Q(\xi, \xi_k) = \sum_{i=1}^{n} \mathbb{E}\{\log(f_\xi(X_i, Z_i)) | X_i, \xi_k\}.$$

(2) *Maximisation step*: maximise the function $\xi \mapsto Q(\xi, \xi_k)$ with respect to ξ. The new estimate for ξ is $\xi_{k+1} = \mathrm{argmax}_{\xi \in \Xi} Q(\xi, \xi_k)$ (provided that it exists and is unique). The key property at the core of the EM algorithm is that the incomplete data likelihood $\prod_{i=1}^n \tilde{q}_{\xi_{k+1}}(X_i) \geq \prod_{i=1}^n \tilde{q}_{\xi_k}(X_i)$ is increased as each iteration with equality if and only if ξ_k is a stationary point (*i.e.* a local or global minimum or a saddle point) : under mild additional conditions (see e.g. Wu. (1983)), the EM algorithm therefore converges to stationary points of the marginal likelihood. Note that, when $n \to \infty$, under appropriate conditions, the renormalized incomplete data log-likelihood $n^{-1} \sum_{i=1}^n \log \tilde{q}_\xi(X_i)$ converges to $\mathbb{E}_\pi[\log \tilde{q}_\xi(X)]$ which is equal, up to a constant and a sign, to the Kullback-Leibler divergence between π and q_ξ. In our particular setting the classical batch form of the algorithm is as follows. First define for $\xi \in \Xi$ the conditional distribution

$$\nu_\xi(x,z) := \frac{f_\xi(x,z)}{\tilde{q}_\xi(x)}, \tag{53}$$

where $\tilde{q}_\xi$ is given by (52). Now, assuming that $\int_{\mathsf{Z}} |T(x,z)|\nu_\xi(x,z)\lambda(dz) < \infty$, one can define for $x \in \mathsf{X}$ and $\xi \in \Xi$

$$\nu_\xi T(x) := \int_{\mathsf{Z}} T(x,z)\nu_\xi(x,z)\lambda(dz), \tag{54}$$

and check that for $f_\xi \in \mathcal{E}_c$ and any $(\xi, \xi') \in \Xi \times \Xi$

$$\mathbb{E}\{\log(f_\xi(X_i, Z_i))|X_i, \xi'\} = L(\nu_{\xi'}T(X_i); \xi),$$

where $L : \Theta \times \Xi \to \mathbb{R}$ is defined as

$$L(\theta;\xi) := -\psi(\xi) + \langle \theta, \phi(\xi) \rangle, \qquad \text{where} \quad \Theta := T(\mathsf{X}, \mathsf{Z}).$$

From this, one easily deduces that for n samples,

$$Q(\xi, \xi_k) = nL\left(\frac{1}{n}\sum_{i=1}^{n} \nu_{\xi_k}T(X_i), \xi\right).$$

Assuming now for simplicity that for all $\theta \in \Theta$, the function $\xi \to L(\theta, \xi)$ reaches its maximum at a single point denoted $\hat{\xi}(\theta)$, *i.e.* $L(\theta; \hat{\xi}(\theta)) \geq L(\theta; \xi)$ for all $\xi \in \Xi$, the EM recursion can then be simply written as

$$\xi_{k+1} = \hat{\xi}\left(\frac{1}{n}\sum_{i=1}^{n} \nu_{\xi_k}T(X_i)\right).$$

The latest condition on the existence and uniqueness of $\hat{\xi}(\theta)$ is not restrictive: it is for example satisfied for finite mixtures of normal distributions. More sophisticated generalisations of the EM algorithm have been developed in order to deal with situations where this condition is not satisfied, see *e.g.* Meng and Van Dyk (1997).

Our scenario differs from the classical setup above in two respects. First the number of samples considered evolves with time and it is required to estimate ξ on the fly. Secondly the samples $\{X_i\}$ are generated by a transition probability with invariant distribution π and are therefore not independent. We address the first problem in Subsection 7.3 and the two problems simultaneously in Subsection 7.4 and describe our particular adaptive MCMC algorithm.

7.3 Sequential EM algorithm

Sequential implementations of the EM algorithm for estimating the parameters of a mixture when the data are observed sequentially in time have been considered by several authors (see (Titterington et al., 1985, Chapter 6), Arcidiacono and Bailey Jones (2003) and the references therein). The version presented here is in many respect a standard adaptation of these algorithms and consists of recursively and jointly estimating and maximising with respect to ξ the function

$$\theta(\xi) = \mathbb{E}_\pi[\log \tilde{q}_\xi(X)] = \pi\{\nu_\xi(X)\},$$

which, as pointed out earlier, is the Kullback-Leibler divergence between π and $\tilde{q}_\xi$, up to an additive constant and a sign. At iteration $k+1$, given an estimate θ_k of θ and $\xi_k = \hat{\xi}(\theta_k)$, sample $X_{k+1} \sim \pi$ and calculate

$$\theta_{k+1} = (1-\gamma_{k+1})\theta_k + \gamma_{k+1}\nu_{\xi_k} T(X_{k+1}) = \theta_k + \gamma_{k+1}\left(\nu_{\xi_k} T(X_{k+1}) - \theta_k\right), \tag{55}$$

where $\{\gamma_k\}$ is a sequence of stepsizes and $\gamma_k \in [0,1]$. This can be interpreted as a stochastic approximation algorithm $\theta_{k+1} = \theta_k + \gamma_{k+1} H(\theta_k, X_{k+1})$ with for $\theta \in \Theta$,

$$H(\theta, x) = \nu_{\hat{\xi}(\theta)} T(x) - \theta \qquad \text{and} \qquad h(\theta) = \pi\left(\nu_{\hat{\xi}(\theta)} T\right) - \theta. \tag{56}$$

It is possible to introduce at this stage a set of simple conditions on the distributions in $\mathcal{E}_c$ that ensures the convergence of $\{\theta_k\}$. By convergence we mean here that the sequence $\{\theta_k\}$ converges to the set of stationary points of the Kullback-Leibler divergence between π and $\tilde{q}_{\hat{\xi}(\theta)}$, *i.e.*

$$\mathcal{L} := \{\theta \in \Theta : \nabla w(\theta) = 0\}.$$

where for $\theta \in \Theta$

$$w(\theta) = K(\pi \| \tilde{q}_{\hat{\xi}(\theta)}), \tag{57}$$

and K and $\tilde{q}_\xi$ are given by (51) and (52), respectively. It is worth noticing that these very same conditions will be used to prove the convergence of our adaptive MCMC algorithm.

(E1) (i) The sets Ξ and Θ are open subsets of $\mathbb{R}^{n_\xi}$ and $\mathbb{R}^{n_\theta}$ respectively. Z is a compact subset of $\mathbb{R}^{n_z}$.

(ii) For any $x \in \mathsf{X}$, $T(x) := \inf\{M : \lambda^{\mathrm{Leb}}(\{z : |T(x,z)| \geq M\}) = 0\} < \infty$.

(iii) The functions $\psi : \Xi \to \mathbb{R}$ and $\phi : \Xi \to \mathbb{R}^{n_\theta}$ are twice continuously differentiable on Ξ.

(iv) There exists a function $\hat{\xi} :\ \Theta \to \Xi$ such that,

$$\forall \xi \in \Xi, \quad \forall \theta \in \Theta, \quad L(\theta; \hat{\xi}(\theta)) \geq L(\theta; \xi).$$

Moreover, the function $\theta \mapsto \hat{\xi}(\theta)$ is continuously differentiable on Θ.

Remark 7. For many models the function $\xi \to L(\theta; \xi)$ admits a unique global maximum for any $\theta \in \Theta$ and the existence and differentiability of $\theta \to \hat{\xi}(\theta)$ follows from the implicit function theorem under mild regularity conditions.

(E2) (i) The level sets $\{\theta \in \Theta, w(\theta) \leq M\}$ are compact;

(ii) The set $\mathcal{L} := \{\theta \in \Theta, \nabla w(\theta) = 0\}$ of stationary points is included in a compact subset of Θ;

(iii) The closure of $w(\mathcal{L})$ has an empty interior.

Remark 8. Assumption (E2) depends on both the properties of π and $q_{\hat{\xi}(\theta)}$ and should therefore be checked on a case by case basis. Note however that (a) these assumptions are satisfied for finite mixtures of distributions in the exponential family under classical technical conditions on the parametrization beyond the scope of the present paper (see, among others (Titterington et al., 1985, chapter 6) and Arcidiacono and Bailey Jones (2003) for details) (b) the third assumption in (E2) can very often be checked using Sard's theorem.

The key to establish the convergence of the stochastic approximation procedure here consists of proving that $w(\theta) = K(\pi \| q_{\hat{\xi}(\theta)})$ plays the role of a Lyapunov function. This is hardly surprising as the algorithm aims at minimizing sequentially in time the incomplete data likelihood. More precisely, we have

Proposition 14. *Assume (E1). Then, for all $\theta \in \Theta$, $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$ and*

$$\mathcal{L} = \{\theta \in \Theta : \langle \nabla w(\theta), h(\theta) \rangle = 0\} = \{\theta \in \Theta : \nabla w(\theta) = 0\}\,, \tag{58}$$

$$\hat{\xi}\,(\mathcal{L}) = \{\xi \in \Xi : \nabla_\xi K(\pi \| q_\xi) = 0\}\,. \tag{59}$$

where $\theta \mapsto h(\theta)$ is given in (56)*.*

The proof is in Appendix E. Another important result to prove convergence is the regularity of the field $\theta \mapsto H_\theta$. We have

Proposition 15. *Assume (E1). Then $\{H_\theta, \theta \in \Theta\}$ is $((1+T)^2, 1)$-regular, where H_θ is defined in Eq.* (56)*.*

The proof is in Appendix E. From this and standard results on the convergence of SA, one may show that the SA procedure converges pointwise under (E1-2).

7.4 On-line EM for IMH adaptation

We now consider the combination of the sequential EM algorithm described earlier with the IMH sampler. As we shall see in Proposition 18, using $\tilde{q}_{\hat{\xi}(\theta)}$ as a proposal distribution for the IMH transition is not sufficient to ensure the convergence of the algorithm, and it will be necessary to use a mixture of a *fixed* distribution ζ (which will not be updated during the successive iterations) and an *adaptive* component, here $\tilde{q}_{\hat{\xi}(\theta)}$. More precisely we define the following family of parametrized IMH transition probabilities $\{P_\theta, \theta \in \Theta\}$. For some $\varepsilon > 0$ let $\zeta \in \mathcal{Q}_{\varepsilon,\pi}$ be a density which does not depend on $\theta \in \Theta$, let $\rho \in (0,1)$ and define the family of IMH transition probabilities

$$\{P_\theta := P^{\mathrm{IMH}}_{q_\theta}, \theta \in \Theta\} \quad \text{with} \quad \{q_\theta := (1-\rho)\tilde{q}_{\hat{\xi}(\theta)} + \rho\zeta, \theta \in \Theta\}. \tag{60}$$

The following properties on ζ and $\mathcal{E}_c$ will be required:

(E3) (i) There exist $\varepsilon > 0$ and $\zeta \in \mathcal{Q}_{\varepsilon,\pi}$ such that for any compact $\mathcal{K} \subset \Xi$

$$\sup_{\xi \in \mathcal{K}} \inf \left\{ M : \lambda^{\mathrm{Leb}} \left(\frac{\tilde{q}_\xi(x)(1+T(x))}{\zeta(x)} \geq M \right) = 0 \right\} < \infty. \tag{61}$$

(ii) There exists $W \to [1, \infty)$ such that for any compact subset $\mathcal{K} \in \Xi$,

$$\int_{\mathsf{X}} W(x)(1+T(x))\zeta(x)\lambda^{\mathrm{Leb}}(dx) + \sup_{\xi \in \mathcal{K}} \int_{\mathsf{X}} W(x)(1+T(x))\tilde{q}_\xi(x)\lambda^{\mathrm{Leb}}(dx) < \infty.$$

It is worth pointing out that the above choice for q_θ and the conditions on ζ will typically have the further benefit of ensuring better practical properties of the algorithm as they will ensure some form of uniform ergodicity.

The *basic* version (see Section 2) of our algorithm now proceeds as follows. Set $\theta_0 \in \Theta$, $\xi_0 = \hat{\xi}(\theta_0)$ and draw X_0 according to some initial distribution. At iteration $k+1$ for $k \geq 0$, draw $X_{k+1} \sim P_{\theta_k}(X_k, \cdot)$ where P_θ is given in (60). Compute $\theta_{k+1} = \theta_k + \gamma_{k+1}\left(\nu_{\xi_k} T(X_{k+1}) - \theta_k\right)$ and $\xi_{k+1} = \hat{\xi}(\theta_{k+1})$. We will study here the corresponding algorithm with reprojections which results in the homogeneous Markov chain $\{Z_k, k \geq 0\}$ as described in Section 2.

We now establish intermediate results about $\{P_\theta, \theta \in \Theta\}$ and $\{H_\theta, \theta \in \Theta\}$ which will lead to the proof that (A1-3) are satisfied. We start with a general proposition about the properties of IMH transition probabilities, relevant to check (A1).

Proposition 16. *Let $V : \mathsf{X} \to [1, +\infty)$ and let $q \in \mathcal{Q}_{\varepsilon,\pi}$ for $\varepsilon > 0$. Then,*

1. *X is a $(1, \varepsilon)$-small set and*

$$P_q^{\mathrm{IMH}} V(x) \leq (1-\varepsilon)V(x) + q(V), \quad \textit{where} \quad q(V) = \int_{\mathsf{X}} q(x)V(x)\lambda^{\mathrm{Leb}}(dx).$$

2. *For any $g \in \mathcal{L}_V$ and any proposal distributions $q, q' \in \mathcal{Q}_{\varepsilon,\pi}$*

$$(2\|g\|_V)^{-1} \left\|P_q^{\mathrm{IMH}} g - P_{q'}^{\mathrm{IMH}} g\right\|_V \leq \int_{\mathsf{X}} |q(x) - q'(x)| V(x)\lambda^{\mathrm{Leb}}(dx) + \\ [q(V) \vee q'(V)] \left((1 \wedge |1 - q^{-1}q'|_1) \vee (1 \wedge |1 - (q')^{-1}q|_1)\right). \quad (62)$$

The proof is in Appendix D. In contrast with the SRWM, the V-norm $\|P_q^{\mathrm{IMH}} g - P_{q'}^{\mathrm{IMH}} g\|_V$ can be large even in situations where $\int_{\mathsf{X}} |q(x) - q'(x)|V(x)\lambda^{\mathrm{Leb}}(dx)$ is small. This stems from the fact that the ratio of densities q/q' enters the upper bound above. As we shall see in Proposition 18, this is what motivates our definition of the proposal distributions in (60) as a mixture of $\tilde{q}_\xi \in \mathcal{E}$ and a non-adaptive distribution ζ which satisfies (E3). Before specializing the results of Proposition 16 to the family of transition probabilities defined in Eq. (60) we prove an intermediate proposition concerned with estimates of the variation $q_\xi - q_{\xi'}$ in various senses.

Proposition 17. *Let $\{\tilde{q}_\xi, \xi \in \Xi\} \subset \mathcal{E}$ be a family of distributions satisfying (E1). Then for any convex compact set $\mathcal{K} \subset \Xi$*

1. *There exists a constant $C < \infty$ such that*

$$\sup_{\xi \in \mathcal{K}} |\nabla_\xi \log \tilde{q}(x;\xi)| \leq C(1 + T(x)). \quad (63)$$

2. *For any $\xi, \xi', x \in \mathcal{K}^2 \times \mathsf{X}$ there exists a constant $C < \infty$ such that*

$$|\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)| < C|\xi - \xi'|(1 + T(x)) \sup_{\xi \in \mathcal{K}} \tilde{q}_\xi(x). \quad (64)$$

3. *For $W \to [1, \infty)$ such that $\sup_{\xi \in \mathcal{K}} \int_{\mathsf{X}} \tilde{q}_\xi(x)[1 + T(x)]W(x)\lambda^{\mathrm{Leb}}(dx) < \infty$ and any $\xi, \xi' \in \mathcal{K}$ there exists a constant $C < \infty$ such that*

$$\int_{\mathsf{X}} |\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)| W(x)\lambda^{\mathrm{Leb}}(dx) \leq C|\xi - \xi'|. \quad (65)$$

Combining Proposition 16 and Proposition 17 we obtain:

Proposition 18. *Assume that the family of distributions $\{\tilde{q}_\xi, \xi \in \Xi\} \subset \mathcal{E}$ satisfies (E1) and (E3). Then the family of transition kernels $\{P_\theta, \theta \in \Theta\}$ given in* (60) *satisfies (A1) with $V = W$, $m = 1$, $\lambda = 1 - \rho\varepsilon$, $\delta = \rho\varepsilon$ and $\beta = 1$.*

We are now in a position to present our final result:

Theorem 19. *Let $\pi \in \mathcal{M}(\mathsf{X})$. Consider the homogeneous Markov chain $\{Z_k = (X_k, \theta_k, \kappa_k, \varsigma_k, \nu_k); k \geq 0\}$ as defined in Section 2, with*

(i) *$\{P_\theta, \theta \in \Theta\}$ as in Eq.* (60) *where $\zeta \in \mathcal{Q}_{\varepsilon,\pi}$ for some $\varepsilon > 0$ and $\{\tilde{q}_\xi, \xi \in \Xi\}$ satisfying (E1), (E3) with V such that $T \in \mathcal{L}_{V^{1/4}}$, and (E2).*

(ii) *$\{H_\theta, \theta \in \Theta\}$ as in Eq.* (56).

(iii) *$\{\gamma_k, k \geq 0\}$ and $\{\epsilon_k, k \geq 0\}$ satisfying (A4).*

Then for any $(x, \theta) \in \mathsf{K} \times \mathcal{K}_0$,

1. *For $\psi \in \mathcal{L}_{V^r}$ and $r \in [0, 1/2)$*

$$n^{-1} \sum_{k=1}^{n} (\psi(X_k) - \pi(\psi)) \to 0 \quad \bar{\mathbb{P}}_{x,\theta} - a.s.$$

2. *$\bar{\mathbb{P}}_{x,\theta} - a.s.$ there exists a random variable $\theta_\infty \in \{\theta \in \Theta : \nabla_\theta K(\pi \| \tilde{q}_{\hat{\xi}(\theta)}) = 0\}$ such that for any $\psi \in \mathcal{L}_{V^r}$, $r \in [0, 1/4)$, provided that $V(\theta_\infty, \psi) > 0$, implies that for all $u \in \mathbb{R}$,*

$$\lim_{n\to\infty} \left| \bar{\mathbb{P}}_{x,\theta} \left(n^{-1/2} \sum_{k=1}^{n} (\psi(X_k) - \pi(\psi)) \leq u \right) - \bar{\mathbb{E}}_{x,\theta} \left\{ \Phi \left(\{V(\theta_\infty, \psi)\}^{-1} u \right) \right\} \right| = 0,$$

where $V(\theta, \psi)$ is given in Eq. (33) *and Φ is the standard normal distribution function.*

Proof. The application of Propositions 14-15 and 18 shows that (A1-3) are satisfied and therefore imply Theorem 9. Then we conclude with Theorems 6 and 8. □

Remark 9. It is worth noticing that provided that $\pi \in \mathcal{M}(\mathsf{X})$ satisfies (M), the results of Propositions 10, 14, 15 and 17, proved in this paper easily allow one to establish a result similar to Theorem 19 for a generalization of the N-SRWM of Haario et al. (2001) (described here in Section 1 and studied in Section 5) to the case where the proposal distribution belongs to $\mathcal{E}$, *i.e.* when the proposal is a mixture of distributions.

A Proof of propositions 3 and 4

Proof of Proposition 3. Let $\phi \in \mathcal{L}_{V^r}$ for $r \in [0,1]$. Proposition 2 shows that, for all $x \in \mathsf{X}$, $\sum_{k=0}^{\infty} |P_\theta^k \phi(x) - \pi(\phi)| < \infty$ and $g_\theta(x) = \sum_{k=0}^{\infty} P_\theta^k \phi(x) - \pi(\phi)$ is a solution to Poisson's equation. The first part of the statement follows, because by construction, for any $\theta \in \Theta$ and any $\phi \in \mathcal{L}_{V^r}$, $g_\theta \in \mathcal{L}_{V^r}$. Note also that,

$$P_\theta^n \phi(x) - P_{\theta'}^n \phi(x) = \sum_{k=0}^{n-1} P_\theta^j (P_\theta - P_{\theta'}) P_{\theta'}^{n-j-1} \phi(x) =$$

$$\sum_{k=0}^{n-1} P_\theta^j (P_\theta - P_{\theta'})(P_{\theta'}^{n-j-1} \phi(x) - \pi(\phi)).$$

Let $\mathcal{K}$ be a compact subset of Θ. Proposition 2 shows that there exists a constant C such that for any $l \geq 0$ and any $\phi \in \mathcal{L}_{V^r}$,

$$\sup_{\theta \in \mathcal{K}} \|P_\theta^l \phi - \pi(\phi)\|_{V^r} \leq C \|\phi\|_{V^r} \, \rho^l .$$

Under assumption (A1), $\sup_{j \geq 0} \sup_{\theta \in \mathcal{K}} \|P_\theta^j V^r\|_{V^r} < \infty$. Thus, for any $l \geq 0$,

$$\|(P_\theta - P_\theta')(P_{\theta'}^l \phi - \pi(\phi))\|_{V^r} \leq C|\theta - \theta'|^\beta \|P_{\theta'}^l \phi(x) - \pi(\phi)\|_{V^r} \leq C|\theta - \theta'|^\beta \|\phi\|_{V^r} \rho^l ,$$

showing that, for any $\phi \in \mathcal{L}_{V^r}$ and all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$\|P_\theta^n \phi - P_{\theta'}^n \phi\|_{V^r} \leq C|\theta - \theta'|^\beta \; \|\phi\|_{V^r} . \tag{66}$$

Let $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$ and write :

$$|(P_\theta^k \psi_\theta(x) - \pi(\psi_\theta)) - (P_{\theta'}^k (x) \psi_{\theta'}(x) - \pi(\psi_{\theta'}))| \leq$$
$$|P_\theta^k \psi_\theta(x) - P_\theta^k \psi_{\theta'}(x)| + |P_\theta^k \psi_{\theta'}(x) - P_{\theta'}^k \psi_{\theta'}(x)| + |\pi(\psi_\theta) - \pi(\psi_{\theta'})| . \tag{67}$$

Since $\{\psi_\theta, \theta \in \Theta\}$ is (V^r, β)-regular, there exists a constant C such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$, $\|\psi_\theta - \psi_{\theta'}\|_{V^r} \leq C|\theta - \theta'|^\beta$. Therefore, there exists a constant C such that, for all $x \in \mathsf{X}$, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$ and for all $k \geq 0$,

$$|P_\theta^k \psi_\theta(x) - P_\theta^k \psi_{\theta'}(x)| \leq C|\theta - \theta'|^\beta V^r(x)$$
$$|\pi(\psi_\theta) - \pi(\psi_{\theta'})| \leq C|\theta - \theta'|^\beta \pi(V^r) \leq C|\theta - \theta'|(\pi(V))^r .$$

Combining (66) and (67), there exists C such that for all $x \in \mathsf{X}$, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$, and for all $k \geq 0$,

$$|(P_\theta^k \psi_\theta(x) - \pi(\psi_\theta)) - (P_{\theta'}^k (x) \psi_{\theta'}(x) - \pi(\psi_{\theta'}))| \leq C|\theta - \theta'|^\beta .$$

On the other hand, by Proposition 2, there exist constants $\rho < 1$ and C such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$|(P_\theta^k \psi_\theta(x) - \pi(\psi_\theta)) - (P_{\theta'}^k \psi_{\theta'}(x) - \pi(\psi_\theta'))| \leq C \rho^k V^r(x).$$

Hence, for any s and $N \geq s$, we have

$$\begin{aligned}
|P^s_\theta g_\theta(x) - P^s_{\theta'} g_{\theta'}(x)| &\leq \sum_{k=s}^{\infty} |(P^k_\theta \psi_\theta(x) - \pi(\psi_\theta)) - (P^k_{\theta'}(x) - \pi(\psi_{\theta'}))| \\
&\leq C V^r(x) \left\{ N|\theta - \theta'|^\beta + \sum_{k=N+s}^{\infty} \rho^k \right\} \\
&\leq C V^r(x) \left\{ N|\theta - \theta'|^\beta + \frac{\rho^{N+s}}{1-\rho} \right\}.
\end{aligned}$$

The proof follows by setting $N = [\beta \log|\theta - \theta'| / \log \rho]$, for $|\theta - \theta'| \leq \delta < 1$, $\theta \neq \theta'$, $N = s$ otherwise, and using the fact that for any $\alpha \in (0, \beta)$, $|\theta - \theta'|^\beta \log|\theta - \theta'| = o(|\theta - \theta'|^\alpha)$. □

Proof of Proposition 4. Let $\boldsymbol{\rho} = \{\rho_k\}$ be a non-increasing sequence of positive numbers and let $\mathcal{K}$ be a compact subset of Θ. For simplicity, we denote $\sigma_\epsilon = \sigma(\mathcal{K}) \wedge \nu_\epsilon$. We first prove (20). (A1-i) shows that, for all $k \geq 0$, $l \geq 0$, all $x \in \mathsf{X}$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_{k+l}) \mathbb{1}(\sigma(\mathcal{K}) \geq k+l) | \mathcal{F}_k\} \leq \kappa^l V(X_k) \mathbb{1}(\sigma(\mathcal{K}) \geq k). \tag{68}$$

We will show that there exist constants $\epsilon > 0$, $0 < \rho < 1$ and C such that, for all k

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_{k+m}) \mathbb{1}(\sigma_\epsilon \geq k+m) | \mathcal{F}_k\} \leq \rho V(X_k) \mathbb{1}(\sigma_\epsilon \geq k) + C. \tag{69}$$

For $n \in \mathbb{N}$, write $n = um + v$, where $v \in \{0, \ldots, m-1\}$. (69) shows that

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_{um+v}) \mathbb{1}(\sigma_\epsilon \geq um+v)\} \leq \rho^u \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_v) \mathbb{1}(\sigma_\epsilon \geq v)\} + \frac{C}{1-\rho}$$

and the proof of (20) follows from (68). It remains to prove (69). We repeatedly use the following result adapted from (Benveniste et al., 1990, Lemma 3, p. 292)

Lemma 20. *Assume (A1). Let $\psi : \Theta \times \mathsf{X} \to \mathbb{R}$ be a function verifying $\sup_{\theta \in \mathcal{K}} \|\psi_\theta\|_V < \infty$. Then, for any $\epsilon > 0$, for any $l \geq 1$ there exist a constant C_l such that, for all $k \geq 0$,*

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{\psi_{\theta_k}(X_{k+l}) \mathbb{1}(\sigma_\epsilon \geq k+l) | \mathcal{F}_k\} \leq \\
&\qquad \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{P_{\theta_k} \psi_{\theta_k}(X_{k+l-1}) \mathbb{1}(\sigma_\epsilon \geq k+l-1) | \mathcal{F}_k\} + C_l \kappa^l \epsilon^\beta \sup_{\theta \in \mathcal{K}} \|\psi_\theta\|_V \, V(X_k) \mathbb{1}(\sigma_\epsilon \geq k).
\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{\psi_{\theta_k}(X_{k+l}) \mathbb{1}(\sigma_\epsilon \geq k+l) | \mathcal{F}_k\} &= \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{P_{\theta_{k+l-1}} \psi_{\theta_k}(X_{k+l-1}) \mathbb{1}(\sigma_\epsilon \geq k+l) | \mathcal{F}_k\} \\
&= \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{P_{\theta_k} \psi_{\theta_k}(X_{k+l-1}) \mathbb{1}(\sigma_\epsilon \geq k+l) | \mathcal{F}_k\} + R_{k,l},
\end{aligned}$$

where

$$R_{k,l} := \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{(P_{\theta_{k+l-1}} - P_{\theta_k}) \psi_{\theta_k}(X_{k+l-1}) \mathbb{1}(\sigma_\epsilon \geq k+l) | \mathcal{F}_k\}.$$

Under (A1-ii), there exists a constant C such that for all $x \in \mathsf{X}$

$$|(P_{\theta_{k+l-1}} - P_{\theta_k}) \psi_{\theta_k}(x) \mathbb{1}(\sigma_\epsilon \geq k+l)| \leq C \sup_{\theta \in \mathcal{K}} \|\psi_\theta\|_V \, V(x) (l\epsilon)^\beta \mathbb{1}(\sigma_\epsilon \geq k+l).$$

Finally, (A1-ii) implies that

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_{k+l-1})\mathbb{1}(\sigma_\epsilon \geq k+l)|\mathcal{F}_k\} \leq \kappa^l V(X_k)\mathbb{1}(\sigma_\epsilon \geq k),$$

which implies

$$|R_{k,l}| \leq C\kappa^l \; (l\epsilon)^\beta \; \sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_V \; V(X_k)\mathbb{1}(\sigma_\epsilon \geq k).$$

□

Using repeatedly the lemma above, we may write

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_{k+m})\mathbb{1}(\sigma_\epsilon \geq k+m)|\mathcal{F}_k\}\\
&\leq \quad \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{P_{\theta_k}V(X_{k+m-1})\mathbb{1}(\sigma_\epsilon \geq k+m-1)|\mathcal{F}_k\} + C_m\epsilon^\beta V(X_k)\mathbb{1}(\sigma_\epsilon \geq k)\\
&\leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{P^2_{\theta_k}V(X_{k+m-2})\mathbb{1}(\sigma_\epsilon \geq k+m-2)|\mathcal{F}_k\} + (C_m + C_{m-1}\kappa)\epsilon^\beta V(X_k)\mathbb{1}(\sigma_\epsilon \geq k)\\
&\vdots\\
&\leq P^m_{\theta_k}V(X_k)\mathbb{1}(\sigma_\epsilon \geq k) + \left(\sum_{i=0}^{m-1} C_{m-i}\kappa^i\right) \epsilon^\beta V(X_k)\mathbb{1}(\sigma_\epsilon \geq k).
\end{aligned}$$

The proof of (69) follows for ϵ sufficiently small. This concludes the proof of (20).

We now turn to the proof of (21). For any sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ such that $\epsilon_k \leq \epsilon$ for any $k \geq s$,

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}[V(X_{k+s})\mathbb{1}\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon})\} \geq k+s)\}\\
&\quad = \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left\{\mathbb{E}^{\boldsymbol{\rho}^{\leftarrow s}}_{X_s,\theta_s}\{V(X_k)\mathbb{1}\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}^{\leftarrow s}) \geq k\}\}\mathbb{1}(\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq s)\right\}\\
&\quad \leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left\{\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\rho}^{\leftarrow s}}_{X_s,\theta}\{V(X_k)\mathbb{1}\{\sigma_\epsilon \geq k\}\}\mathbb{1}\{\sigma(\mathcal{K}) \geq s\}\right\}\\
&\quad \leq C\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\{V(X_s)\mathbb{1}(\sigma(\mathcal{K}) \geq s)\},
\end{aligned}$$

and the proof is concluded by (A1-i). □

B Proof of Proposition 10

For any $x \in \mathsf{X}$, define the acceptance region $\mathsf{A}(x) = \{z \in \mathsf{X}; \pi(x+z) \geq \pi(x)\}$ and the rejection region $\mathsf{R}(x) = \{z \in \mathsf{X}; \pi(x+z) < \pi(x)\}$. From the definition (41) of $\mathcal{K}_{\delta,\varepsilon}$ (Roberts and Tweedie, 1996, Theorem 2.2) applies for any $q \in \mathcal{K}_{\delta,\varepsilon}$ and we can conclude that (42) is satisfied. Noting that the two sets $\mathsf{A}(x)$ and $\mathsf{R}(x)$ do not depend on the proposal distribution q and using the conclusion of the proof of Theorem 4.3 of Jarner and Hansen (2000) we have

$$\inf_{q\in\mathcal{K}_{\delta,\varepsilon}} \liminf_{|x|\to+\infty} \int_{\mathsf{A}(x)} q(z)\lambda^{\text{Leb}}(dz) > 0,$$

so that from the conclusion of the proof of Theorem 4.1 of Jarner and Hansen (2000),

$$\sup_{q\in\mathcal{K}_{\delta,\varepsilon}} \limsup_{|x|\to+\infty} \frac{P_q^{\text{SRW}}V(x)}{V(x)} = 1 - \inf_{q\in\mathcal{K}_{\delta,\varepsilon}} \liminf_{|x|\to+\infty} \int_{\mathsf{A}(x)} q(z)\lambda^{\text{Leb}}(dz) < 1,$$

which proves (43). Finally, for any $q \in \mathcal{K}_{\delta,\varepsilon}$,

$$\frac{P_q^{\text{SRW}}V(x)}{V(x)} = \int_{\mathsf{A}(x)} \frac{\pi(x+z)^{-\eta}}{\pi(x)^{-\eta}} q(z)\lambda^{\text{Leb}}(dz) + \int_{\mathsf{R}(x)} \left(1 - \frac{\pi(x+z)}{\pi(x)} + \frac{\pi(x+z)^{1-\eta}}{\pi(x)^{1-\eta}}\right) q(z)\lambda^{\text{Leb}}(dz)$$
$$\leq \sup_{0\leq u\leq 1} (1-u+u^{1-\eta}),$$

which proves (44). Now notice that

$$P_q^{\text{SRW}}g(x) - P_{q'}^{\text{SRW}}g(x) = \int_{\mathsf{X}} \alpha(x,x+z)(q(z)-q'(z))g(x+z)\lambda^{\text{Leb}}(dz) +$$
$$g(x)\int_{\mathsf{X}} \alpha(x,x+z)(q'(z)-q(z))\lambda^{\text{Leb}}(dz).$$

We therefore focus, for $r \in [0,1]$ and $g \in \mathcal{L}_{V^r}$, on the term

$$\frac{\left|\int_{\mathsf{X}} \alpha(x,x+z)(q(z)-q'(z))g(x+z)\lambda^{\text{Leb}}(dz)\right|}{\|g\|_{V^r}\, V^r(x)} \leq \frac{\int_{\mathsf{X}} \alpha(x,x+z)|q(z)-q'(z)|V^r(x+z)\lambda^{\text{Leb}}(dz)}{V^r(x)} =$$
$$= \int_{\mathsf{A}(x)} \frac{\pi(x+z)^{-r\eta}}{\pi(x)^{-r\eta}}|q(z)-q'(z)|\lambda^{\text{Leb}}(dz) + \int_{\mathsf{R}(x)} \frac{\pi(x+z)^{1-r\eta}}{\pi(x)^{1-r\eta}}|q(z)-q'(z)|\lambda^{\text{Leb}}(dz)$$
$$\leq \int_{\mathsf{X}} |q(z)-q'(z)|\lambda^{\text{Leb}}(dz).$$

We now conclude that for any $x \in \mathsf{X}$ and any $g \in \mathcal{L}_{V^r}$,

$$\frac{|P_q^{\text{SRW}}g(x) - P_{q'}^{\text{SRW}}g(x)|}{V^r(x)} \leq 2\,\|g\|_{V^r} \int_{\mathsf{X}} |q(z)-q'(z)|\lambda^{\text{Leb}}(dz).$$

C Proof of Lemma 11

We have

$$\int_{\mathsf{X}} |q_\Gamma(z) - q_{\Gamma'}(z)|dz = \int_{\mathsf{X}} \left|\int_0^1 \frac{d}{dv} q_{\Gamma+v(\Gamma'-\Gamma)}(z)dv\right| dz$$

and let $\Gamma_v = \Gamma + v(\Gamma'-\Gamma)$, so that

$$\frac{d}{dv}\log q_{\Gamma+v(\Gamma'-\Gamma)}(z) = -\frac{1}{2}\text{Tr}\left[\Gamma_v^{-1}\left(\Gamma'-\Gamma\right) + \Gamma_v^{-1}zz^{\mathsf{T}}\Gamma_v^{-1}(\Gamma'-\Gamma)\right]$$

and consequently

$$\int_{\mathsf{X}} \left|\int_0^1 \frac{d}{dv} q_{\Gamma+v(\Gamma'-\Gamma)}(z)dv\right| dz \leq |\Gamma'-\Gamma| \int_0^1 |\Gamma_v^{-1}|dv \leq \frac{n_x}{\lambda_{\min}(\mathcal{K})}|\Gamma'-\Gamma|,$$

where we have used the following inequality,

$$|\text{Tr}[\Gamma_v^{-1}zz^{\mathsf{T}}\Gamma_v^{-1}(\Gamma'-\Gamma)]| \leq |\Gamma'-\Gamma|\text{Tr}[\Gamma_v^{-1}\Gamma_v^{-1}zz^{\mathsf{T}}].$$

D Proof of Propositions 16, 18

Proof of Proposition 16. The minorization condition is a classical result, see Mengersen and Tweedie (1996). Now notice that

$$\begin{aligned}P_q^{\mathrm{IMH}}V(x) &= \int_{\mathsf{X}} \alpha_q(x,y)V(y)q(y)\lambda^{\mathrm{Leb}}(dy) + V(x)\int_{\mathsf{X}}[1-\alpha_q(x,y)]q(y)\lambda^{\mathrm{Leb}}(dy)\\ &\le \left(1 - \int_{\mathsf{X}}\left(\frac{q(x)}{\pi(x)} \wedge \frac{q(y)}{\pi(y)}\right)\pi(y)\lambda^{\mathrm{Leb}}(dy)\right)V(x) + q(V),\end{aligned}$$

where α_q is given in (49). The drift condition follows.

From the definition of the transition probability and for any $g \in \mathcal{L}_V$,

$$\begin{aligned}&\left|P_q^{\mathrm{IMH}}g(x) - P_{q'}^{\mathrm{IMH}}g(x)\right| \le \|g\|_V \times\\ &\left\{\int_{\mathsf{X}}|\alpha_q(x,y)q(y) - \alpha_{q'}(x,y)q'(y)|V(y)\lambda^{\mathrm{Leb}}(dy) + V(x)\int_{\mathsf{X}}|\alpha_{q'}(x,y)q'(y) - \alpha_q(x,y)q(y)|\lambda^{\mathrm{Leb}}(dy)\right\}\\ &\qquad\le 2\|g\|_V\, V(x)\int_{\mathsf{X}}|\alpha_q(x,y)q(y) - \alpha_{q'}(x,y)q'(y)|V(y)\lambda^{\mathrm{Leb}}(dy).\end{aligned}$$

We therefore bound

$$I = \int_{\mathsf{X}}\left|\frac{q(y)}{\pi(y)} \wedge \frac{q(x)}{\pi(x)} - \frac{q'(y)}{\pi(y)} \wedge \frac{q'(x)}{\pi(x)}\right|\pi(y)V(y)\lambda^{\mathrm{Leb}}(dy).$$

We introduce the following sets

$$\mathsf{A}_q(x) = \left\{y : \frac{q(y)}{\pi(y)} \le \frac{q(x)}{\pi(x)}\right\} \quad \text{and} \quad \mathsf{B}_q(x) = \left\{y : \frac{q(y)}{\pi(y)} \le \frac{q'(x)}{\pi(x)}\right\},$$

and notice that the following inequalities hold:

$$\begin{aligned}&\forall y \in \mathsf{A}_{q'}^c(x)\cap\mathsf{A}_q^c(x),\ \pi(y) < \frac{\pi(x)}{q(x)}q(y) \wedge \frac{\pi(x)}{q'(x)}q'(y), \quad \text{and}\\ &\qquad\qquad\forall y \in \mathsf{A}_{q'}^c(x)\cap\mathsf{B}_q^c(x),\ \pi(y) < \frac{\pi(x)}{q'(x)}(q'(y)\wedge q(y)).\end{aligned} \tag{70}$$

We now decompose I into four terms $I := \sum_{i=1}^4 I_i$, where

$$\begin{aligned}I &= \int_{\mathsf{A}_q\cap\mathsf{A}_{q'}}\left|\frac{q(y)}{\pi(y)} - \frac{q'(y)}{\pi(y)}\right|\pi(y)V(y)\lambda^{\mathrm{Leb}}(dy) + \int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}^c}\left|\frac{q(x)}{\pi(x)} - \frac{q'(x)}{\pi(x)}\right|\pi(y)V(y)\lambda^{\mathrm{Leb}}(dy)\\ &+ \int_{\mathsf{A}_q\cap\mathsf{A}_{q'}^c}\left|\frac{q(y)}{\pi(y)} - \frac{q'(x)}{\pi(x)}\right|\pi(y)V(y)\lambda^{\mathrm{Leb}}(dy) + \int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}}\left|\frac{q(x)}{\pi(x)} - \frac{q'(y)}{\pi(y)}\right|\pi(y)V(y)\lambda^{\mathrm{Leb}}(dy),\end{aligned}$$

where we have dropped x in the set notation for simplicity. We now determine bounds for I_i, $i = 2, 3$. Notice that since $y \in \mathsf{A}_q^c \cap \mathsf{A}_{q'}^c$

$$\begin{aligned}I_2 &\le \left\{\left|1 - \frac{q'(x)}{q(x)}\right|\int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}^c} V(y)q(y)\lambda^{\mathrm{Leb}}(dy)\right\} \wedge \left\{\left|1 - \frac{q(x)}{q'(x)}\right|\int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}^c} V(y)q'(y)\lambda^{\mathrm{Leb}}(dy)\right\}\\ &\le \left\{\left|1 - \frac{q'(x)}{q(x)}\right| \wedge \left|1 - \frac{q(x)}{q'(x)}\right|\right\}\left\{\int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}^c} V(y)q(y)\lambda^{\mathrm{Leb}}(dy) \vee \int_{\mathsf{A}_q^c\cap\mathsf{A}_{q'}^c} V(y)q'(y)\lambda^{\mathrm{Leb}}(dy)\right\}\end{aligned}$$

and it can easily be checked that

$$\left|1-\frac{q'}{q}\right| \wedge \left|1-\frac{q}{q'}\right| \leq \left\{1 \wedge \left|1-\frac{q'}{q}\right|\right\} \vee \left\{1 \wedge \left|1-\frac{q}{q'}\right|\right\}.$$

The term I_3 can be bounded as follows

$$I_3 \leq \left\{\int_{\mathsf{A}_q\cap\mathsf{A}^c_{q'}\cap\mathsf{B}^c_q} q(y)V(y)\lambda^{\mathrm{Leb}}(dy)\right\} \wedge \left\{\left(\frac{q(x)}{\pi(x)} - \frac{q'(x)}{\pi(x)}\right)\int_{\mathsf{A}_q\cap\mathsf{A}^c_{q'}\cap\mathsf{B}^c_q} V(y)\pi(y)\lambda^{\mathrm{Leb}}(dy)\right\} + \int_{\mathsf{A}_q\cap\mathsf{A}^c_{q'}\cap\mathsf{B}_q}\left(\frac{q'(y)}{\pi(y)} - \frac{q(y)}{\pi(y)}\right)V(y)\pi(y)\lambda^{\mathrm{Leb}}(dy),$$

and using (70) we find that

$$I_3 \leq \left\{1 \wedge \left(\frac{q(x)}{q'(x)} - 1\right)\right\}\int_{\mathsf{A}_q\cap\mathsf{A}^c_{q'}\cap\mathsf{B}^c_q} q(y)\, V(y)\lambda^{\mathrm{Leb}}(dy) + \int_{\mathsf{A}_q\cap\mathsf{A}^c_{q'}\cap\mathsf{B}_q} |q'(y) - q(y)|V(y)\lambda^{\mathrm{Leb}}(dy).$$

The bound for I_4 follows from that of I_3 by swapping q and q'. □

Proof of Proposition 17. We first note that from *Fisher's identity* we have

$$\forall \xi \in \Xi, \quad \nabla_\xi \log \tilde{q}(x;\xi) = \int_{\mathsf{Z}} \nabla_\xi \log f(x,z;\xi)\nu_\xi(x,z)\lambda(dz) = -\nabla_\xi\psi(\xi) + [\nu_\xi T(x)]^{\mathsf{T}}\nabla_\xi\phi(\xi).$$

and from (E1) we conclude that Eq. (63) holds. Eq. (64) is a direct consequence of (63). Now we prove Eq. (65).

$$\int_{\mathsf{X}} |\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)|W(x)\lambda^{\mathrm{Leb}}(dx) \tag{71}$$

$$= \int_{\mathsf{X}} \left|\int_0^1 \int_{\mathsf{Z}} [\psi(\xi') - \psi(\xi) + \langle\theta(x,z), \phi(\xi) - \phi(\xi')\rangle] f^v_\xi(x,z) f^{1-v}_{\xi'}(x,z)\lambda(dz)\lambda^{\mathrm{Leb}}(dv)\right| W(x)\lambda^{\mathrm{Leb}}(dx) \tag{72}$$

$$\leq |\psi(\xi') - \psi(\xi)| \int_{\mathsf{X}}\int_0^1\int_{\mathsf{Z}} f^v_\xi(x,z) f^{1-v}_{\xi'}(x,z) W(x)\lambda(dz)\lambda^{\mathrm{Leb}}(dvdx) \tag{73}$$

$$+ |\phi(\xi) - \phi(\xi')| \int_{\mathsf{X}}\int_0^1 T(x)\int_{\mathsf{Z}} f^v_\xi(x,z) f^{1-v}_{\xi'}(x,z) W(x)\lambda(dz)\lambda^{\mathrm{Leb}}(dvdx) \tag{74}$$

$$\leq |\psi(\xi') - \psi(\xi)| \int_0^1 \left[\int_{\mathsf{X}} W(x)\tilde{q}_\xi(x)\lambda^{\mathrm{Leb}}(dx)\right]^v \left[\int_{\mathsf{X}} W(x)\tilde{q}_{\xi'}(x)\lambda^{\mathrm{Leb}}(dx)\right]^{1-v}\lambda^{\mathrm{Leb}}(dv) \tag{75}$$

$$+ |\phi(\xi) - \phi(\xi')| \int_0^1 \left[\int_{\mathsf{X}} T(x)W(x)\tilde{q}_\xi(x)\lambda^{\mathrm{Leb}}(dx)\right]^v \left[\int_{\mathsf{X}} T(x)W(x)\tilde{q}_{\xi'}(x)\lambda^{\mathrm{Leb}}(dx)\right]^{1-v}\lambda^{\mathrm{Leb}}(dv) \tag{76}$$

and we conclude using (E1). □

Proof of Proposition 18. Denote

$$\Upsilon_{\xi,\xi',\rho}(x) := \frac{(1-\rho)\tilde{q}_\xi(x) + \rho\zeta(x)}{(1-\rho)\tilde{q}_{\xi'}(x) + \rho\zeta(x)} = 1 + \frac{\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)}{\zeta(x)[\tilde{q}_{\xi'}(x)/\zeta(x) + \frac{\rho}{1-\rho}]}.$$

Therefore, from (63), for any convex compact set $\mathcal{K} \subset \Xi$ there exists $C < \infty$ such that for any $\xi, \xi' \in \mathcal{K}^2 \times \mathsf{X}$,

$$|1 - \Upsilon_{\xi,\xi',\rho}(x)| \leq \frac{1-\rho}{\rho} \frac{|\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)|}{\zeta(x)} \leq C|\xi - \xi'| \frac{\sup_{\xi \in \mathcal{K}} \tilde{q}_\xi(x)(1 + T(x))}{\zeta(x)},$$

which with (61) implies that for all $\xi, \xi' \in \mathcal{K}$ and for λ^{Leb}-almost all x there exists $C < \infty$ such that

$$\left(1 \wedge \left|1 - \Upsilon_{\xi,\xi',\rho}(x)\right|\right) \vee \left(1 \wedge \left|1 - \Upsilon_{\xi',\xi,\rho}(x)\right|\right) \leq C\,|\xi - \xi'|.$$

Now as a direct consequence of (65) one can show that for any $r \in [0, 1]$

$$\int_{\mathsf{X}} |\tilde{q}_\xi(x) - \tilde{q}_{\xi'}(x)| V(x)^r \lambda^{\mathrm{Leb}}(dx) \leq C|\xi - \xi'| \sup_{\xi \in \mathcal{K}} \int_{\mathsf{X}} V^r(x)(1 + T(x))\tilde{q}_\xi(x)\lambda^{\mathrm{Leb}}(dx).$$

The proof is concluded by application of Proposition 16. □

E Proof of proposition 14 and 15

Proof of proposition 14. We first note that from *Fisher's identity* we have

$$\forall \xi \in \Xi, \quad \nabla_\xi \log \tilde{q}(x;\xi) = \int_{\mathsf{Z}} \nabla_\xi \log f(x, z; \xi)\nu_\xi(x, z)\lambda(dz) = -\nabla_\xi \psi(\xi) + [\nu_\xi T(x)]^{\mathsf{T}} \nabla_\xi \phi(\xi).$$

From (63) and (E1) we may derive under the sum sign to show that

$$\nabla_\xi \int_{\mathsf{X}} \pi(x) \log q_\xi(x)\lambda^{\mathrm{Leb}}(dx) = \int_{\mathsf{X}} \pi(x)\nabla_\xi \log \tilde{q}_\xi(x)\lambda^{\mathrm{Leb}}(dx) = -\nabla_\xi \psi(\xi) + [\pi(\nu_\xi T)]^{\mathsf{T}} \nabla_\xi \phi(\xi),$$

and thus by the chain rule of derivations

$$\nabla_\theta w(\theta) = -\left(-\nabla_\xi \psi(\widehat{\xi}(\theta)) + \pi\left(\nu_{\hat{\xi}(\theta)} T\right)^{\mathsf{T}} \nabla_\xi \phi(\widehat{\xi}(\theta))\right) \nabla_\theta \widehat{\xi}(\theta).$$

For any $\theta \in \Theta$, $\widehat{\xi}(\theta)$ is a stationary point of the mapping $\xi \to L(\theta, \xi)$ and thus

$$\nabla_\xi L(\theta, \widehat{\xi}(\theta)) = -\nabla_\xi \psi(\widehat{\xi}(\theta)) + \theta^{\mathsf{T}} \nabla_\xi \phi(\widehat{\xi}(\theta)) = 0.$$

Consequently (56) implies that $\nabla_\theta w(\theta) = -h(\theta)^{\mathsf{T}} \nabla_\xi \phi(\widehat{\xi}(\theta)) \nabla_\theta \widehat{\xi}(\theta)$. We also notice that $\nabla_\theta \nabla_\xi L(\theta, \xi) = \nabla_\xi \phi(\xi)^{\mathsf{T}}$. Differentiation with respect to θ of the mapping $\theta \mapsto \nabla_\xi L(\theta, \widehat{\xi}(\theta))$ yields

$$\nabla_\theta \nabla_\xi L(\theta, \widehat{\xi}(\theta)) = \nabla_\xi \phi(\widehat{\xi}(\theta)) + \nabla_\theta \hat{\xi}(\theta)^{\mathsf{T}} \nabla^2_\xi L(\theta, \widehat{\xi}(\theta)) = 0.$$

We finally have

$$\langle \nabla_\theta w(\theta), h(\theta) \rangle = h(\theta)^{\mathsf{T}} (\nabla_\theta \widehat{\xi}(\theta))^{\mathsf{T}} \nabla^2_\xi L(\theta, \widehat{\xi}(\theta)) \nabla_\theta \widehat{\xi}(\theta) h(\theta),$$

which concludes the proof, since under (E1), $\nabla^2_\xi L(\theta, \widehat{\xi}(\theta)) \leq 0$ for any $\theta \in \Theta$. □

Proof of proposition 15. For any $x \in \mathsf{X}$,

$$|H_\theta(x) - H_{\theta'}(x)| \quad \leq \quad T(x) \int_{\mathsf{Z}} |\nu_{\hat{\xi}(\theta)}(x,z) - \nu_{\hat{\xi}(\theta')}(x,z)| \lambda(dz) + |\theta' - \theta|.$$

From Proposition 17 one has that for any compact set $\mathcal{K} \in \Xi$, there exists a constant C such that, for all $\xi \in \mathcal{K}$

$$|\nabla_\xi \log f(x,z;\xi)| \leq C(1 + T(x)) \quad \text{and} \quad |\nabla_\xi \log q(x;\xi)| \leq C(1 + T(x))$$

Thus

$$|\nabla_\xi \log \nu_\xi(x,z)| \leq |\nabla_\xi \log f(x,z;\xi)| + |\nabla_\xi \log q(x;\xi)| \leq 2C(1 + T(x)).$$

Hence, for $\xi, \xi' \in \mathcal{K}$,

$$|\nu_\xi T(x) - \nu_{\xi'} T(x)| \leq 2C(1 + T(x))^2 |\xi - \xi'|$$

which concludes the proof. □

References

Andrieu, C., Moulines, E. and Priouret, P. (2002). Stability of stochastic approximation under verifiable conditions. Accepted for publication, SIAM J. on Control and Optimization.

Andrieu, C. and Robert, C. (2001). Controlled MCMC for optimal sampling. *Cahiers du Cérémade 0125* .

Arcidiacono, P. and Bailey Jones, J. (2003). Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica* **71** 933–946.

Bartusek, J. (2000). *Stochastic Approximation and Optimization of Markov Chains.* Ph.D. thesis.

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, New York.

Chen, H., Guo, L. and Gao, A. (1988). Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications* **27** 217–231.

Chen, H.-F. and Zhu, Y.-M. (1986). Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica 1* **29** 914–926.

Douc, R., Moulines, E. and Rosenthal, J. (2002). Quantitative bounds for geometric convergence rates of Markov chains. Submitted to Annals of Applied Probability.

Duflo, M. (1997). *Random Iterative Systems.* Applications of mathematics, Springer-Verlag.

Gelman, A., Roberts, G. and Gilks, W. (1995). Efficient Metropolis jumping rules. In *Bayesian Statistics* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.), vol. V. Oxford University Press.

Gilks, W., Roberts, G. and Sahu, S. (1998). Adaptive markov chain monte carlo through regeneration. *Journal American Statistical Association* **93** 1045–1054.

Glynn, P. W. and Meyn, S. P. (1996). A lyapounov bound for solutions of the poisson equation. *Annals of Probability* **24** 916–931.

Gåsemyr, J. (2003). On an adaptive version of the Metropolis-Hastings with independent proposal distribution. *Scandinavian Journal of Statistics* **30** 159–173.

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

Hall, P. (1977). Martingale invariance principles. *Annals of Probability* **5** 875–887.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and its Application.* Academic Press, New York, London.

Hall, P. and Heyde, C. C. (1981). Rates of convergence in the martingale central limit theorem. *Annals of Probability* **9** 395–404.

Jarner, S. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and Their Applications* **85** 341–361.

Kushner, H. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications.* Applications of Mathematics, Springer-Verlag, New-York.

McLeish, D. (1975). A maximal inequality and dependent strong laws. *Annals of Probability* **3** 829–839.

Meng, X. L. and Van Dyk, D. (1997). The EM algorithm–an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)* **59** 511–567.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* **24** 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, M. (1953). Equations of state calculations by fast computing machines,. *Journal of Chemical Physics* **21** 1087–1091.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability.* Communication and Control Engineering series, Springer-Verlag, London.

Nummelin, E. (1991). On the Poisson equation in the potential theory of a single kernel. *Math. Scand.* **68** 59–82.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of mathematical statistics* **22** 400–407.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorem for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

Roberts, G. and Tweedie, R. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and Their Applications* **80** 211–229.

Titterington, D., Smith, A. F. M. and Makov, U. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley and Sons, Chichester.

Wu., C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11** 95–103.