On the convergence of stochasticapproximations under a subgeometricergodic Markov dynamic

Vianney Debavelaere, Stanley Durrleman, Stéphanie Allassonnière

Proof of the convergence of Stochastic Approximations under a subgeometric ergodic Markov dynamic

Vianney Debavelaere[1], **Stanley Durrleman**[2] **and Stéphanie Allassonnière**[3]

[1] *Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France, e-mail:* vianney.debavelaere@polytechnique.edu

[2] *ARAMIS Lab, Institut du Cerveau et de la Moelle épinière , Paris, France, e-mail:* stanley.durrleman@icm-institute.org

[3] *Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France, e-mail:* stephanie.allassonniere@parisdescartes.fr

Abstract: In this paper, we extend the framework of the convergence of stochastic approximations given by

$$\theta_{n+1} = \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}),$$

where $(X_n)_{n\in\mathbb{N}}$ is a Markov Chain. Such a procedure is used in many methods such as parameters estimation inside a Metropolis Hastings algorithm, stochastic gradient descent or stochastic Expectation Maximization algorithm. The convergence of such a stochastic approximation has already been proved under an assumption of geometric ergodicity of the Markov Chain. However, in many practical situations this hypothesis is not satisfied, for instance for any heavy tail target distribution in a Monte Carlo Metropolis Hastings algorithm. In this paper, we loosen this hypothesis and prove the convergence of the stochastic approximation by assuming only a subgeometric ergodicity of the Markov dynamic. This result opens up the possibility to derive more generic algorithms with proven convergence. As an example, we study an adaptative Markov Chain Monte Carlo algorithm where the proposal distribution is adapted by learning the variance of a heavy tail target distribution.

MSC 2010 subject classifications: Primary 62L20, 60J05; secondary 90C15.
Keywords and phrases: Stochastic approximation, Markovian dynamic, Subgeometric ergodicity
Acknowledgment: This work has been partly funded by the European Research Council with grant 678304..

1. Introduction

A common problem across scientific fields is to find the roots of a non-linear function $h : \Theta \to \mathbb{R}$. In statistics, the problem is further increased by the fact that h is not known, but only noisy observations of it or its gradient. This problematic can appear in a lot of different domains such as stochastic optimization [25, 30], Expectation Maximization algorithms [3, 22], reinforcement learning

[1, 10], ... In all cases, solutions to this problem often take the form of an iterative sequences $(\theta_n)_{n\in\mathbb{N}}$ that converges towards a point θ^* in the set of solutions of $h(\theta) = 0$. The general class of stochastic approximation methods, and in particular the Robbins-Monro one, falls within this framework and produces a sequence defined by:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}\zeta_{n+1}\,,$$

where ζ_{n+1} is a noisy observation of $h(\theta_n)$: $\zeta_{n+1} = h(\theta_n) + \xi_{n+1}$ with ξ_{n+1} a noise sequence of random variables. In that case, h is called the mean field. This procedure, first developed in [27], has been studied under various sets of hypotheses, see [1, 9, 10, 11, 12, 16, 23] among many other works.

In this paper, we focus on the case of a Markov state-dependent noise, which means that the noise observation of h: $(\zeta_n)_{n\in\mathbb{N}}$ takes the form $(H_{\theta_n}(X_n))_{n\in\mathbb{N}}$ where (X_n) is a Markov Chain in the state space $\mathcal{X}$, and, for all $\theta \in \Theta$, H_θ is a function from $\mathcal{X}$ to Θ:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}H_{\theta_n}(X_{n+1})\,.$$

The assumption of the state dependent Markov noise case is general, and met for instance within the framework of the stochastic gradient descent [25] or within the framework of Metropolis Hastings algorithms. In the latter, the distribution to sample from may depend on a parameter θ that is learned throughout the algorithm. This is notably the case of the Stochastic Approximation Expectation Maximization Markov Chain Monte Carlo (SAEM MCMC) algorithm [2, 3, 13]. One can also consider adaptive MCMC algorithms where the proposal distribution depends of a parameter θ. Such a procedure can be used to adapt the variance of the proposal along the algorithm [5, 6, 19, 28] and allows a better sampling. In both cases, the update of the parameter θ can be seen as a stochastic approximation.

In the framework of a state dependant noise, the authors of [5] give general hypotheses for the stochastic approximation algorithm to converge. They are based on the control of the fluctuations of the Markov Chain as well as on the regularity of the solution of a Poisson equation. In practice, these conditions can be difficult to verify and the authors show their validity when the Markov chain satisfies drift conditions implying a *geometric ergodicity* of the chain i.e. when assuming the convergence of the kernel of the Markov Chain towards its invariant distribution at a geometric rate. Under theses assumptions, we are then able to prove the convergence of the SAEM MCMC [3] and some adaptative MCMC algorithms [5].

However, in lot of practical situations, this ergodicity condition is not satisfied. If several articles study the convergence of adaptive MCMC algorithms under subgeometric ergodicity [7, 8, 29, 31], the general case of stochastic approximations with a subgeometric Markovian dynamic has not yet been proved

convergent, to the best of our knowledge. Often, authors only consider a geometric ergodic Markovian dynamic [21, 24]. This can be a problem when $(X_n)_{n\in\mathbb{N}}$ is sampled using a Metropolis Hastings algorithm targetting heavy tails distributions (Weibull ot Pareto distributions in particular) [14, 17, 18, 20], in which case the Markov Chain is not geometric ergodic. Similarly the geometric ergodicity condition has been a problem in [4] where the authors have not proved the convergence of the exponentially scaled Gaussian independent component analysis (EG-ICA) model due to the presence of a subgeometric Markov Chain. In all these cases, the theorem presented in [5] does not allow us to conclude on the convergence of the stochastic approximation algorithm using these Markovian dynamics. In the same vein, in [22], the authors prove the convergence of a mini batch SAEM algorithm using the theorem presented in [5] and hence, by assuming the geometric ergodicity of the Markov Chain.

In this paper, we propose a more general sets of hypotheses, under which we prove the convergence of stochastic approximations with subgeometric Markovian dynamics. These new conditions mainly concern the rate of convergence of the Markov Chain as well as the regularity of its kernel. In particular, most of the polynomial rates of convergence satisfy these hypotheses. Finally, we apply this new theorem to prove the convergence of a stochastic approximation used to adapt the variance of the proposal of a Metropolis Hastings algorithm. More precisely, we prove this convergence for two different classes of heavy tail target distributions including, among others, the Weibull and the Pareto distributions.

2. Stochastic approximation framework with Markovian dynamic

In this section, we summarize the stochastic approximation procedure in the case of a Markovian dynamic with adaptive truncation sets. This procedure was first described in [5]. In the following, we denote $\mathcal{X}$ the state space and Θ the parameter space that we assume to be an open subset of $\mathbb{R}^{n_\theta}$. Moreover, we suppose that both are equipped with countably generated σ-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$.

In the next subsection, we present the framework of a stochastic approximation producing a sequence of elements converging towards a solution of $h(\theta) = 0$ when there exists probability measures π_θ such that $h(\theta) = \mathbb{E}_{\pi_\theta}(H_\theta(X))$ and, for all $\theta \in \Theta$, $H_\theta : \mathcal{X} \mapsto \Theta$.

2.1. Markov Chain sequence

Let $\Delta = (\Delta_n)_{n\in\mathbb{N}}$ be a monotone non increasing sequence of positive real numbers with $\Delta_0 \leq 1$ and set $\theta_c \notin \Theta$ and $x_c \notin \mathcal{X}$ two cemetery states. We also set, for all $\theta \in \Theta$ the vector field $H_\theta : \mathcal{X} \mapsto \Theta$. We then define a Markov chain

$Y_n^\Delta = (X_n, \theta_n)$ on $\mathcal{X} \cup \{x_c\} \times \Theta \cup \{\theta_c\}$ by:

$$\theta_{n+1} = \begin{cases} \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}) & \text{and } X_{n+1} \sim P_{\theta_n}(X_n, .) \quad \text{if } \theta_n \in \Theta \\ \theta_c & \text{and } X_{n+1} = x_c \quad \text{if } \theta_n \notin \Theta\,. \end{cases} \tag{1}$$

We put the following hypothesis on the transition probabilities $(P_\theta, \theta \in \Theta)$ and on the random vector field H:

(A2) For any $\theta \in \Theta$, the Markov kernel P_θ has a single stationary distribution π_θ. In addition, $H : \Theta \times \mathcal{X} \to \Theta$ is measurable for all $\theta \in \Theta$.

The existence and uniqueness of the invariant distribution can be verified under the classical conditions of irreducibility and recurrence [26]. We also set $h(\theta) = \int_{\mathcal{X}} H_\theta(x)\pi_\theta(dx)$ the mean field of the stochastic approximation. This allows us to recognize the usual stochastic approximation procedure:

$$\theta_{n+1} = \theta_n + \Delta_{n+1}(h(\theta_n) + \xi_{n+1})$$

where $\xi_{n+1} = H_{\theta_n}(X_{n+1}) - h(\theta_n)$ is the noise sequence.

We assume the mean field h satisfies the following hypothesis that amounts to the existence of a global Lyapunov function:

(A1) $h : \Theta \to \mathbb{R}^{n_\theta}$ is continuous and there exists a continuously differentiable function $w : \Theta \to [0, +\infty[$ such that:

(i) there exists $M_0 > 0$ such that

$$\mathcal{L} := \{\theta \in \Theta, \langle \nabla w(\theta), h(\theta)\rangle = 0\} \subset \{\theta \in \Theta, w(\theta) < M_0\}\ ,$$

(ii) there exists $M_1 \in (M_0, +\infty]$ such that $\mathcal{W}_{M_1} := \{\theta \in \Theta, w(\theta) \leq M\}$ is a compact set,
(iii) for any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla w(\theta), h(\theta)\rangle < 0$,
(iv) the closure of $w(\mathcal{L})$ has an empty interior.

We denote by $\mathcal{F} = \{\mathcal{F}_n, n \geq 0\}$ the natural filtration of the Markov chain (X_n, θ_n) and by $\mathbb{P}^\Delta_{x,\theta}$ the probability measure associated to the chain (Y_n^Δ) started from the initial conditions $(x, \theta) \in \mathcal{X} \times \Theta$. Finally, we denotes by Q_{Δ_n} the sequence of transition probabilities that generates the inhomogeneous Markov chain (Y_n^Δ).

2.2. Truncation process

We introduce $(\mathcal{K}_n)_{n\in\mathbb{N}}$ a sequence of compact subsets of Θ such that

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta \qquad \text{and} \qquad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})\,.$$

Let $(\epsilon_n)_{n\in\mathbb{N}}$ be a sequence of non increasing positive numbers and K be a subset of $\mathcal{X}$. Let $\Phi : \mathcal{X} \times \Theta \to K \times \mathcal{K}_0$ be a measurable function. We then define the stochastic approximation algorithm with adaptive truncation sets as an homogeneous Markov chain on $\mathcal{X} \times \Theta \times \mathbb{N} \times \mathbb{N}$ by

$$Z_n = (X_n, \Theta_n, \kappa_n, \nu_n) \tag{2}$$

with the following transition at iteration $n+1$:

- If $\nu_n = 0$, then draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(\Phi(X_n, \theta_n), .)$. Otherwise, draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\Delta_n}(X_n, \theta_n, .)$.
- If $|\theta_{n+1} - \theta_n| \leq \epsilon_n$ and $\theta_{n+1} \in \mathcal{K}_{\kappa_n}$ then set $\kappa_{n+1} = \kappa_n$ and $\nu_{n+1} = \nu_n + 1$. Otherwise, set $\kappa_{n+1} = \kappa_n + 1$ and $\nu_{n+1} = 0$.

To summarize this process, if our parameter θ leaves the current truncation set $\mathcal{K}_{\kappa_n}$ or if the difference between two of its successive values is larger than a time dependent threshold ϵ_n, we reinitialize the Markov chain by a value inside $\mathcal{K}_0$: $\Phi(X_n, \theta_n)$ and update the truncation set to a larger one $\mathcal{K}_{\kappa_n+1}$ as well as the threshold to a smaller one: ϵ_{n+1}. Hence, κ_n represents the number of re-initialization before the step n while ν_n is the number of steps since the last re-initialization.

The idea behind this truncation process is to force the noise to be small in order for the drift $h(\theta)$ to dominate. We do so by forcing our algorithm to come back to the center of Θ whenever the parameters become too big.

2.3. Control of the fluctuations and main convergence theorem

In this section, we state two last hypothesis about the control of fluctuations before presenting the theorem proved in [5]. In this paper, the authors present several conditions (A1 to A4) that imply the convergence of the stochastic approximation algorithm. It is those conditions that we will, in the next section, verify under subgeometric ergodicity of the Markov chain.

We first define, for any compact $\mathcal{K}$ and any sequence of non increasing positive numbers $(\epsilon_k)_{k\in\mathbb{N}}$, $\sigma(\mathcal{K}) = \inf(k \geq 1, \theta_k \notin \mathcal{K})$ and $\nu_\epsilon = \inf(k \geq 1, |\theta_k - \theta_{k-1}| \geq \epsilon_k)$. Moreover, for $W : \mathcal{X} \to [1, \infty)$ and $g : \mathcal{X} \to \mathbb{R}^{n_\theta}$, we write

$$||g||_W = \sup_{x\in\mathcal{X}} \frac{|g(x)|}{W(x)}\,.$$

We can now present the hypothesis (A3):

(A3) For any $\theta \in \Theta$, the Poisson equation $g - P_\theta g = H_\theta - \pi_\theta(H_\theta)$ has a solution g_θ. Moreover, there exist a function $W : \mathcal{X} \to [1, +\infty]$ such that $\{x \in \mathcal{X}, W(x) < +\infty\} \neq \emptyset$, constants $\alpha \in (0, 1]$ and $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(i) the following holds:

$$\sup_{\theta \in \mathcal{K}} ||H_\theta||_W < \infty \tag{3}$$

$$\sup_{\theta \in \mathcal{K}} ||g_\theta||_W + ||P_\theta g_\theta||_W < \infty \tag{4}$$

$$\sup_{\theta, \theta' \in \mathcal{K}} |\theta - \theta'|^{-\alpha} \left(||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W \right) < \infty \tag{5}$$

(ii) there exist constants $\{C_k, k \geq 0\}$ such that, for any $k \in \mathbb{N}$, for any sequence Δ and for any $x \in \mathcal{X}$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k) \mathbb{1}_{\sigma(\mathcal{K}) \geq k}] \leq C_k W^p(x) \tag{6}$$

(iii) there exist ϵ and a constant C such that for any sequence Δ and for any $x \in \mathcal{X}$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k) \mathbb{1}_{\sigma(\mathcal{K}) \wedge \nu_\epsilon \geq k}] \leq C W^p(x)\,. \tag{7}$$

This assumption concerns the existence and regularity of the Poisson equation associated with each of the transition kernel P_θ. In [5], the authors show that those conditions are verified under hypothesis of geometric ergodicity of the Markov chain. In the next sections, we will relax this ergodicity conditions to be able to consider subgeometric ergodic chains.

Finally, the last condition concerns the step size sequences:

(A4) The sequences (Δ_k) and (ϵ_k) are non increasing, positive and satisfy $\sum_{k=0}^{\infty} \Delta_k = \infty$, $\lim_{k \to \infty} \epsilon_k = 0$ and

$$\sum_{k=1}^{\infty} \Delta_k^2 + \Delta_k \epsilon_k^\alpha + (\epsilon_k^{-1} \Delta_k)^p < \infty$$

where p and α are defined in (A3).

We can finally state the theorem proved in [5]:

Theorem 2.1. *Assume (A1)-(A4). Let $K \subset \mathcal{X}$ such that $\sup_{x \in K} W(x) < \infty$ and such that $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ (where M_0 and $\mathcal{W}_{M_0}$ are defined in (A1)) and let Z_n be as defined in (2). Then, for all $(x, \theta) \in \mathcal{X} \times \Theta$, we have $\lim_{k \to \infty} d(\theta_k, \mathcal{L}) = 0$, $\mathbb{P}^{\Delta}_{x,\theta}$-a.s.*

Of the four conditions (A1) to (A4), (A3) is often the most difficult to verify and we need more practical conditions to verify it. In particular, in [5], the authors give drift conditions that imply (A3). However, those drift conditions imply the geometric ergodicity of the Markov chain. In lot of cases, this ergodicity is not verified. It is the case for instance in the Metropolis Hastings algorithm when the target distribution has heavy tails. It was also a problem in the exponentially scaled Gaussian independant components analysis (EG-ICA [4]) "EM like" model where the authors could not prove the convergence of their algorithm due to a subgeometric ergodicity. To tackle this problem, we will, in the next section, state subgeometric drift conditions and hypotheses on the rate of convergence that are sufficient to insure the validity of (A3). The new theorem allows us to verify the convergence in a broader range of cases, some of them being explicited section 5.

3. Convergence of the stochastic approximation sequence under subgeometric conditions

In this section, we give the drift conditions and hypotheses under which we will work to prove the validity of (A3). Denote, for $V : \mathcal{X} \to [1, \infty)$, $\mathcal{L}_V = \{g : \mathcal{X} \to \mathbb{R}^{n_\theta}, ||g||_V < \infty\}$.

(DRI) For any $\theta \in \Theta$, P_θ is ψ-irreductible and aperiodic. In addition, there exist a function $V : \mathcal{X} \to [1, \infty)$ and a constant $p \geq 2$ such that, for any compact subset $\mathcal{K} \subset \Theta$, there exist constants b, $\delta_0 > 0$, a probability measure ν, a concave, increasing function $\phi : [1, \infty) \to (0, \infty)$, continuously differentiable such that $\lim_{v\to\infty} \phi'(v) = 0$ and a subset $\mathcal{C}$ of $\mathcal{X}$ with

$$\sup_{\theta\in\mathcal{K}} P_\theta V^p(x) + \phi \circ V^p(x) \leq V^p(x) + b\mathbb{1}_{\mathcal{C}}(x) \qquad \forall x \in \mathcal{X} \tag{8}$$

$$\inf_{\theta\in\mathcal{K}} P_\theta(x, A) \geq \delta_0 \nu(A) \qquad \forall x \in \mathcal{C}, \forall A \in \mathcal{B}(\mathcal{X}) \,. \tag{9}$$

Remark 3.1. *We could consider the following, more general, drift condition: it exists $m \in \mathbb{N}^*$ such that*

$$\sup_{\theta\in\mathcal{K}} P_\theta^m V^p(x) + \phi \circ V^p(x) \leq V^p(x) + b\mathbb{1}_{\mathcal{C}}(x) \qquad \forall x \in \mathcal{X}$$

$$\inf_{\theta\in\mathcal{K}} P_\theta^m(x, A) \geq \delta_0 \nu(A) \qquad \forall x \in \mathcal{C}, \forall A \in \mathcal{B}(\mathcal{X}) \,.$$

The results we present in the following sections would still be verified under such a drift condition. To adapt the proofs (and more precisely, the proof of the lemma 4.6), we would then need to use the lemma B.3. of [5].

Under the condition (DRI), $\mathcal{C}$ is a small set and the Markov Chain P_θ verifies a subgeometric drift condition [15]. In particular, it implies the existence of a

stationary distribution π_θ for all $\theta \in \mathcal{K}$ as well as an uniform subgeometric ergodicity on all compacts of Θ. Hence, for all $\theta \in \Theta$, there exists a constant C_θ and a sequence $(r_{\theta,k})_{k\in\mathbb{N}}$ such that, $\forall q, s > 0$ with $1/q + 1/s = 1$ and $\forall f \in \mathcal{L}_{(\phi\circ V^p)^{1/s}}$,

$$r_{\theta,k}^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi\circ V^p)^{1/s}} \leq C_\theta ||f||_{(\phi\circ V^p)^{1/s}} .$$

Moreover, it has been showed in [14] that, under a subgeometric ergodicity condition, we can choose a rate of convergence $(r_k)_{k\in\mathbb{N}}$ that only depends of the function ϕ. Hence, for $\theta \in \mathcal{K}$ a fixed compact, we can choose a rate of convergence $(r_k)_{k\in\mathbb{N}}$ that depends only of $\mathcal{K}$. Similarly, it has been proved that the constant C_θ is bounded on all compact $\mathcal{K}$. Hence, it exists a constant $C_\mathcal{K}$ and a sequence $(r_k)_{k\in\mathbb{N}}$ such that, for all $f \in \mathcal{L}_{(\phi\circ V^p)^{1/s}}$ and for all $\theta \in \mathcal{K}$,

$$\sup_{\theta\in\mathcal{K}} r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi\circ V^p)^{1/s}} \leq C_\mathcal{K} ||f||_{(\phi\circ V^p)^{1/s}} . \tag{10}$$

We will see in the following that several hypothesis must be made on that rate of convergence $(r_k)_{k\in\mathbb{N}}$ for the condition (A3) to be satisfied.

Remark 3.2. *In general, we can consider Ψ_1 and Ψ_2 a pair of inverse Young functions i.e. two strictly increasing continuous functions on $\mathbb{R}_+$ verifying $\Psi_1(x)\Psi_2(y) \leq x + y$. We then have, for all $f \in \mathcal{L}_{\Psi_2(\phi\circ V^p)}$:*

$$\Psi_1(r_k) ||P_\theta^k f - \pi_\theta(f)||_{\Psi_2(\phi\circ V^p)} \leq C_\mathcal{K} ||f||_{\Psi_2(\phi\circ V^p)} .$$

In order to simplify the notations, we will only consider in the following the pair of inverse Young functions $\Psi_1(x) = qx^{1/q}$ and $\Psi_2(x) = sx^{1/s}$. The same reasoning could be carried out for any other pair of Young functions by adapting the hypotheses (H1) and (H2).

We now state several hypothesis that we will need to prove the condition (A3). The first one concerns the choice of the inverse Young functions with respect to the rate of convergence and the regularity of H_θ. With p as defined in (DRI), we suppose:

(H1) For any compact $\mathcal{K}$, it exists $q > 0$ and $s \geq p$ with $1/q + 1/s = 1$ such that:

$$\sum_{k\geq 0} \frac{1}{r_k^{1/q}} < \infty \qquad \text{and} \qquad \sup_{\theta\in\mathcal{K}} ||H_\theta||_{(\phi\circ V^p)^{1/s}} < \infty .$$

Remark 3.3. *Most of the polynomial rates of convergence satisfy this hypothesis. The assumption $s \geq p$ is necessary to control the V-norm by the $(\phi \circ V^p)^{1/s}$-norm.*

We then need hypotheses on the regularity of H_θ and P_θ. Two of them are similar to the ones presented in [5] while the first one will help us to conclude

on the validity of Eq. (5).

(H2) For any compact $\mathcal{K}$, it exists a constant $\beta \in [0,1]$ such that

(i) there exists $T \in \mathbb{N}$ and $\alpha \in (0,1)$ such that

$$\sup_{\theta,\theta' \in \mathcal{K}} T||\theta - \theta'||^{\beta-\alpha} + ||\theta - \theta'||^{-\alpha} \sum_{k \geq T} \frac{1}{r_k^{1/q}} < \infty \,.$$

(ii) there exists C such that for all $x \in \mathcal{X}$,

$$\sup_{\theta,\theta' \in \mathcal{K}} |\theta - \theta'|^{-\beta} |H_\theta(x) - H_{\theta'}(x)| \leq C V^p(x)$$

(iii) there exists C such that for all $\theta, \theta' \in \mathcal{K}$,

$$||P_\theta g - P_{\theta'} g||_{(\phi \circ V^p)^{1/s}} \leq C||g||_{(\phi \circ V^p)^{1/s}} |\theta - \theta'|^\beta \qquad \forall g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}} \,.$$

Remark 3.4. *The condition (H2-i) can be easily verified for* $r_k^{1/q} = k^d$ *with* $d > 1$. *Indeed, we know that* $\sum_{k=T}^{\infty} \frac{1}{k^d} \sim \frac{1}{(d-1)T^{d-1}}$. *Hence, if* $0 < \alpha < 1$, *we choose* $T = \left\lfloor ||\theta - \theta'||^{-\frac{\alpha}{d-1}} \right\rfloor$ *and we have:*

$$||\theta - \theta'||^{-\alpha} \sum_{k=T}^{\infty} \frac{1}{k^d} \sim_{\theta \to \theta'} \frac{1}{d-1} \,.$$

Moreover, $T||\theta - \theta'||^{\beta-\alpha} = ||\theta - \theta'||^{\beta - \alpha - \frac{\alpha}{d-1}}$. *Choosing* α *such that* $\beta - \alpha - \frac{\alpha}{d-1} > 0$ *i.e.* $\alpha < \beta \frac{d-1}{d}$ *allows us to conclude.*

Finally, due to the subgeometric ergodicity, we are unable to iterate the drift condition without causing divergent quantities to appear. This iteration was however one of the key of the proof of the condition 7. To overcome this problem, we add one last hypothesis on the behaviour of ϕ on the petite set $\mathcal{C}$ defined by assumption (DRI):

(H3) It exists $\delta > 0$ such that, $\forall x \in \mathcal{C}$,

$$\phi \circ V^p(x) \geq \delta V^p(x) \,.$$

Remark 3.5. *It is interesting to remark that asking for this condition on the whole set* $\mathcal{X}$ *implies the geometric ergodicity of the chain. However, we only ask it on the petite set* $\mathcal{C}$ *on which we have some freedom. In fact, in most cases, this condition will be easy to verify. Indeed, according to the theorem 16.1.9. of [15], we can choose* $\mathcal{C} = \{V^p \leq d\}$ *with* $d > 0$. *Hence, if this set is compact (true if* V *is continuous and* $V(x) \longrightarrow_{x \to \infty} \infty$*) and if* $(\phi \circ V^p)^{1/s}/V^p$ *is continuous, (H3) is verified.*

We can now state our major theorem:

Theorem 3.1. *Assume (DRI) and (H1)-(H3). Then, the condition (A3) is verified. In particular, if (A1), (A2) and (A4) are also verified we can apply the theorem 2.1 to conclude that* $lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$

4. Proof of the theorem 3.1

4.1. Sketch of proof

The proof follows the principal ideas of [5]. However, due to the fact that our Markov chain is no longer supposed to be geometric ergodic, we need several new arguments.

The first important result is the fact that we are able to control the V-norm by the $(\phi \circ V^p)^{1/s}$-norm under the hypothesis (H1). This is particularly important as we need to choose $W = V^p$ in (A3) to be able to find an upper bound of the expectation of $W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\epsilon \geq k}$ (see Eq. (7)). Hence, we use this control of the V-norm to control the different quantities in Eq. (3), (4) and (5) using the rate of convergence given by Eq (10). This control is given by the lemma 4.1.

Using this lemma, we can control the norm of the solution of the Poisson equation using the subgeometric ergodicity. This is explicited lemma 4.2.

We then want to prove the condition (5) (lemma 4.5). Using once again a decomposition of the solution of the Poisson equation, we see that we need regularity conditions on $\theta \mapsto P_\theta$ and h. The regularity of $\theta \mapsto P_\theta$ is given by the condition (H2) while we prove the Hölder continuity of h in lemma 4.4.

Finally, while the condition (6) is easily proved by iterating the drift condition, we still need to prove the condition (7). In [5], the authors prove it using the same argument which does not hold anymore for us as this iteration can make appear divergent quantities. That is why we need to state the condition (H3). It is under this final condition that we are able to iterate an upper bound of the drift and to prove (7) in lemma 4.6.

After this final step, we have all the tools necessary to prove the theorem 3.1.

We will now present and prove with details the different lemmas introduced above and implying each of the conditions in (A3) before proving the theorem 3.1.

4.2. Proof of Eq. (4)

First, using (H1), we show that we can control the V-norm using the $(\phi \circ V^p)^{1/s}$-norm:

Lemma 4.1. *Assume (H1). Then, it exists $C > 0$ such that, for all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$,*

$$||g||_V \leq C||g||_{(\phi \circ V^p)^{1/s}} .$$

Proof. ϕ is concave and increasing so, $\forall v \geq 1$, $\phi(v) \leq \phi'(1)(v-1) + \phi(1) \leq cv$ with c a positive constant. Hence, for all $x \in \mathcal{X}$, since $s \geq p$ and $V(x) \geq 1$,

$$(\phi \circ V^p)^{1/s}(x) \leq c^{1/s} V^{p/s}(x) \leq c^{1/q} V(x)$$

which allows us to verify the announced inequality. □

We can now prove the equation (4).

Lemma 4.2. *Suppose (DRI). Then, the Poisson equation $g - P_\theta g = H_\theta - \pi_\theta(H_\theta)$ has a solution g_θ. Moreover, under (H1),*

$$\sup_{\theta \in \mathcal{K}} ||g_\theta||_V < \infty \qquad \textit{and} \qquad \sup_{\theta \in \mathcal{K}} ||P_\theta g_\theta||_V < \infty .$$

Proof. The proposition [21.2.4] of [15] states the existence of a solution g_θ of the Poisson equation under the subgeometric ergodicity conditions (DRI) verifying:

$$g_\theta(x) = \sum_{k \geq 0} \left(P_\theta^k H_\theta(x) - h(\theta) \right) .$$

Moreover, we know that for all compact $\mathcal{K}$, it exists a constant C and a convergence rate $(r_k)_{k \in \mathbb{N}}$ independent of $\theta \in \mathcal{K}$ such that, for all $f \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$, for all $\theta \in \mathcal{K}$,

$$r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi \circ V^p)^{1/s}} \leq C||f||_{(\phi \circ V^p)^{1/s}} .$$

Hence, using lemma 4.1,

$$\begin{aligned} r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_V &\leq r_k^{1/q} C ||P_\theta^k f - \pi_\theta(f)||_{(\phi \circ V^p)^{1/s}} \\ &\leq C||f||_{(\phi \circ V^p)^{1/s}} . \end{aligned}$$

Since $h(\theta) = \pi_\theta(H_\theta)$ and using (H1), we have that:

$$||g_\theta||_V \leq \sum_{k \geq 0} ||P_\theta^k H_\theta - \pi_\theta(H_\theta)||_V \leq C||H_\theta||_{(\phi \circ V^p)^{1/s}} \sum_{k \geq 0} \frac{1}{r_k^{1/q}} < \infty .$$

Finally, we can use the same argument for $P_\theta g_\theta$ to prove that $\sup_{\theta \in \mathcal{K}} ||P_\theta g_\theta||_V < \infty$. □

4.3. Proof of Eq. (5)

We now want to prove the condition given by Eq. (5). In particular, we need hypotheses on the regularity in θ of H_θ and P_θ presented in condition (H2). We begin by proving two lemma implying the Hölder continuity of h.

Lemma 4.3. *Assume (DRI), (H1) and (H2). Then, there exists a constant C such that, for all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$ and any $k \geq 0$,*

$$\sup_{\theta,\theta' \in \mathcal{K}} |\theta - \theta'|^{-\beta} ||P_\theta^k g - P_{\theta'}^k g||_V \leq C ||g||_{(\phi \circ V^p)^{1/s}} .$$

Proof. This result is a consequence of (H2-iii). Indeed, we can write, for all θ, θ' in $\mathcal{K}$, all $k \in \mathbb{N}$ and all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$,

$$P_\theta^k g - P_{\theta'}^k g = \sum_{j=0}^{k-1} P_\theta^j (P_\theta - P_{\theta'})(P_{\theta'}^{k-j-1} g(x) - \pi_{\theta'}(g)) .$$

But, using Eq. (10), we know that,

$$\sup_{\theta \in \mathcal{K}} ||P_\theta^l - \pi_\theta||_{(\phi \circ V^p)^{1/s}} \leq \frac{C}{r_l^{1/q}} .$$

Hence, $\sup_{l \in \mathbb{N}, \theta \in \mathcal{K}} ||P_\theta^l||_{(\phi \circ V^p)^{1/s}} < \infty$.
Finally, using this result and (H2-iii),

$$\begin{aligned} ||P_\theta^k g - P_{\theta'}^k g||_V &\leq C ||P_\theta^k g - P_{\theta'}^k g||_{(\phi \circ V^p)^{1/s}} \\ &\leq C ||\theta - \theta||^\beta \sum_{j=0}^{k-1} ||P_{\theta'}^{k-j-1} g(x) - \pi_{\theta'}(g)||_{(\phi \circ V^p)^{1/s}} \\ &\leq C ||\theta - \theta||^\beta ||g||_{(\phi \circ V^p)^{1/s}} \sum_{j=0}^{k-1} \frac{1}{r_k^{1/q}} . \end{aligned}$$

We obtain the result using the convergence of the sum of the $1/r_k^{1/q}$. □

We now prove that h is α-Hölder for any α in $(0, \beta)$. We will use this property to finally be able to prove (5).

Lemma 4.4. *Assume (DRI), (H1) and (H2). Then, for all $\alpha \in (0, \beta)$,*

$$\sup_{\theta,\theta' \in \mathcal{K}} ||\theta - \theta'||^{-\alpha} |h(\theta) - h(\theta')| < \infty .$$

Proof. We use the following decomposition of $|h(\theta) - h(\theta')|$ for $x_0 \in \mathcal{X}$ and $k \in \mathbb{N}$:

$$|h(\theta) - h(\theta')| = |A(\theta, \theta') + B(\theta, \theta') + C(\theta, \theta')|$$

with:

$$A(\theta,\theta') = h(\theta) - P_\theta^k H_\theta(x_0) + P_{\theta'}^k H_{\theta'}(x_0) - h(\theta')$$
$$B(\theta,\theta') = P_\theta^k H_\theta(x_0) - P_{\theta'}^k H_\theta(x_0)$$
$$C(\theta,\theta') = P_{\theta'}^k H_\theta(x_0) - P_{\theta'}^k H_{\theta'}(x_0)\,.$$

From lemma 4.3, hypothesis (H2-ii) and (DRI), we obtain the following inequalities:

$$|A(\theta,\theta')| \leq \frac{C}{r_k^{1/q}} ||H_\theta||_{(\phi\circ V^p)^{1/s}} (\phi\circ V^p)^{1/s}(x_0)$$

$$|B(\theta,\theta')| \leq C||H_\theta||_{(\phi\circ V^p)^{1/s}} ||\theta-\theta'||^\beta (\phi\circ V^p)^{1/s}(x_0)$$

$$\begin{aligned}|C(\theta,\theta')| &\leq \int_{\mathcal{X}} P_{\theta'}^k(x_0,dy)|H_\theta(y) - H_{\theta'}(y)| \\ &\leq C||\theta-\theta'||^\beta \int_{\mathcal{X}} P_{\theta'}^k(x_0,dy) V^p(y) \\ &\leq C||\theta-\theta'||^\beta V^p(x_0)\,.\end{aligned}$$

Hence, using the fact that $\sup_{\theta\in\mathcal{K}} ||H_\theta||_{(\phi\circ V^p)^{1/s}} < \infty$ and $(\phi\circ V^p)^{1/s} \leq cV^p$, we find

$$|h(\theta) - h(\theta')| \leq CV^p(x_0)\left(||\theta-\theta'||^\beta + \frac{1}{r_k^{1/q}}\right).$$

Finally, because $\frac{1}{r_k^{1/q}} \to 0$, it exists $k\in\mathbb{N}$ such that $\frac{1}{r_k^{1/q}} < ||\theta-\theta'||^\beta$ which concludes the proof. □

Finally, we can state the condition (5).

Lemma 4.5. *Assume (DRI), (H1) and (H2). Then,*

$$\sup_{\theta,\theta'\in\mathcal{K}} |\theta-\theta'|^{-\alpha} \left(||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W\right) < \infty\,.$$

Proof. Using (H2-iii), lemma 4.3 and 4.4, we have that, for $x\in\mathcal{X}$, $k\in\mathbb{N}$ and $\theta,\theta'\in\mathcal{K}$,

$$\begin{aligned}D_k(x,\theta,\theta') &:= ||P_\theta^k H_\theta(x) - h(\theta) - P_{\theta'}^k H_{\theta'}(x) + h(\theta')|| \\ &\leq ||P_\theta^k H_\theta(x) - P_\theta^k H_{\theta'}(x)|| + ||P_{\theta'}^k H_{\theta'}(x) - P_\theta^k H_{\theta'}(x)|| + ||h(\theta) - h(\theta')|| \\ &\leq C||\theta-\theta'||^\beta V^p(x)\,.\end{aligned}$$

On the other hand, using the ergodicity of the Markov Chain, we have that:

$$D_k(x,\theta,\theta') \leq \frac{C}{r_k^{1/q}} (\phi\circ V^p)^{1/s}(x)\,.$$

Hence for $t = 0$ or 1 and any $T \geq t$,

$$||\theta - \theta'||^{-\alpha}||P_\theta^t g_\theta - P_{\theta'}^t g_{\theta'}||_V \leq C||\theta - \theta'||^{-\alpha}||P_\theta^t g_\theta - P_{\theta'}^t g_{\theta'}||_{(\phi\circ V^p)^{1/s}}$$
$$\leq C\left((T-t)||\theta - \theta'||^{\beta-\alpha} + ||\theta - \theta'||^{-\alpha}\sum_{i\geq T}\frac{1}{r_k^{1/q}}\right).$$

Hence, we can use (H2-i) to conclude the proof. □

Finally, under (DRI), (H1) and (H2), we are able to prove the first item of (A3). We still have to prove the second and third item. The second item is easily proved using the drift condition:

$$\mathbb{E}_{x,\theta}^\Delta(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\geq k}) \leq \mathbb{E}_{x,\theta}^\Delta\left[\mathbb{E}_{x,\theta}^\Delta(PV^p(X_{k-1})|\mathcal{F}_{k-1})\right]$$
$$\leq \mathbb{E}_{x,\theta}^\Delta(V^p(X_{k-1})) + b \leq V^p(x) + kb$$

and we conclude using the fact that $V^p(x) \geq 1$.

Hence, we only need to prove the last item of (A3).

4.4. Proof of Eq. (7)

Under geometrical ergodicity, iterating the drift condition is enough to prove the necessary inequality. However, in the subgeometric case, this iteration can make appear a divergent sum. To overcome this difficulty, we will use the condition (H3).

Lemma 4.6. *Assume (DRI) and (H3). Then, there exist ϵ and a constant C such that for any sequence Δ and for any $x \in \mathcal{X}$,*

$$\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^\Delta[W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k}] \leq CW^p(X)\,.$$

Proof. Using (DRI) and (H3), we have that, for all $x \in \mathcal{X}$,

$$PV^p(X) \leq V^p(x) - \phi\circ V^p(x) + b\mathbb{1}_C(x)\,.$$

Hence, if $x \notin C$, $PV^p(x) \leq V^p(x)$ and, if $x \in C$, $PV^p(x) \leq (1-\delta)V^p(x) + b$.

We first consider the case $\delta \geq 1$. In that case, if $x \in C$, $PV^p(x) \leq b$. Hence, by induction, $\mathbb{E}_{x,\theta}^\Delta\left(V(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}\right) \leq V(x) + b$.

If $\delta < 1$, we note $\tau_k = Card(X_i|X_i \in C \text{ for } 1 \leq i \leq k)$. Then, by induction,

$$\mathbb{E}_{x,\theta}^\Delta\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}\right) = \mathbb{E}_{x,\theta}^\Delta\left(\mathbb{E}_{x,\theta}^\Delta\left(PV^p(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}\middle|\mathcal{F}_{k-1}\right)\right)$$
$$\leq \mathbb{E}_{x,\theta}^\Delta\left((1-\delta\mathbb{1}_{X_{k-1}\in C})V^p(X_{k-1}) + b\mathbb{1}_{X_{k-1}\notin C}\right)$$
$$\leq \mathbb{E}_{x,\theta}^\Delta\left((1-\delta)^{\tau_k}V^p(x) + b\sum_{i=0}^{\tau_k-1}(1-\delta)^i\right)$$
$$\leq V^p(x) + \frac{b}{1-\delta}\,.$$

Since $V^p(x) \geq 1$, we can conclude the proof. □

4.5. Proof of Theorem 3.1

We can now finalize this section by proving the theorem 3.1 using the different lemma previously presented.

Proof. Using lemma 4.1 and hypothesis (H1), we immediately obtain the first inequality in hypothesis (A3-i). The next two conditions are given respectively by 4.2 and 4.5. The last conditions are a consequence of lemma 4.6. □

5. Example: Symmetric Random Walk Metropolis Hastings (SRWMH)

5.1. Presentation of the algorithm

The SRWMH is a popular algorithm allowing for sampling from a distribution π. It consists at simulating a Markov Chain (X_n) whose stationary distribution is π. The user chooses a symmetric proposal distribution q. At each step, if the chain is currently at x, a candidate y for X_{n+1} is proposed using $q(x - .)$. This candidate is then accepted with probability:

$$\alpha(x, y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} & \text{if } \pi(x) \neq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

If the candidate is rejected, the chain stays at its current location x. The transition kernel of this Markov Chain is: $\forall x \in \mathcal{X}, \forall A \in \mathcal{B}(\mathcal{X})$,

$$P(x, A) = \int_A \alpha(x, y) q(x - y) \lambda^{Leb}(dy) + \mathbb{1}_A(x) \int_X (1 - \alpha(x, y)) q(x - y) \lambda^{Leb}(dy) \,. \tag{12}$$

The choice of the proposal distribution q is of crucial importance. In particular, proposal distributions with a too small or too large covariance matrix leads to a highly correlated Markov Chain. To overcome this difficulty, the authors of [19] have proposed to learn the covariance matrix while sampling the Markov Chain leading to adaptive MCMC samplers. We note $\theta = (\mu, \Gamma)$ and we suppose that we can choose q_θ such that $Var(q_\theta) = \Gamma$. For instance, if we choose to work with Gaussian distributions, q_θ is the density of the distribution $\mathcal{N}(0, \Gamma)$. We then write P_θ the kernel of the SRWMH when the proposal is q_θ.

The authors of [19] choose to adapt the value of Γ using the following algorithm:

$$\begin{cases} \mu_{n+1} = \mu_n + \Delta_{n+1}(X_{n+1} - \mu_n) \\ \Gamma_{n+1} = \Gamma_n + \Delta_{n+1} \left((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)^T - \Gamma_n \right) \end{cases} \tag{13}$$

with $X_{n+1} \sim P_{\theta_n}(X_n, .)$ where $\theta_n = (\mu_n, \Gamma_n)$ and with (Δ_n) a nonincreasing sequence of step sizes such that $\sum_{n=1}^{\infty} \Delta_n = \infty$ and, for some $b > 0$, $\sum_{n=1}^{\infty} \Delta_n^{1+b} < \infty$.

This procedure is in fact a stochastic approximation:

$$\theta_{n+1} = \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1})$$

with

$$H_\theta(x) = (x - \mu, (x-\mu)(x-\mu)^T - \Gamma)\,. \tag{14}$$

Moreover, assuming that $\int_{\mathcal{X}} x^2 \pi(dx) < \infty$, one can verify that:

$$h(\theta) = \left(\mu_\pi - \mu, (\mu_\pi - \mu)(\mu_\pi - \mu)^T + \Gamma_\pi - \Gamma\right)\,.$$

This algorithm has already been studied in [5]. In that paper, the authors make an hypothesis on the tail properties of the target distribution that implies the geometric ergodicity of the Markov Chain P_θ. Under this hypothesis, the authors prove that the conditions (A1)-(A4) are verified and so prove the convergence of the algorithm.
With our framework, we are able to loosen the hypothesis on π to give conditions under which we have a subgeometric ergodicity of the Markov Chain P_θ and still guarantee convergence of the algorithm.

In [5], the verification of the condition (A1) does not use the behaviour of the tail of π. Hence, it will stay true in our case and we can state it here:

Proposition 5.1. *Let*

$$w(\mu, \Gamma) = -\int_{\mathcal{X}} \log\left(\frac{\pi(x)}{\phi_{\mu,\Gamma}}\right) \pi(dx)$$

where $\phi_{\mu,\Gamma}$ is the normal density of mean μ and variance Γ. Then, this w verifies (A1). Furthermore, $\mathcal{L}$ is reduced to a single point $\theta_\pi := (\mu_\pi, \Gamma_\pi)$.

To prove (A3), we need some hypothesis on the behaviour of π. In particular, we will verify that we can apply the theorem 3.1 under two sets of hypotheses. The first contains among others the Weibull distributions while the second one includes the Pareto distributions. Those two sets of hypotheses as well as the proof of the condition (A3) are detailed in the following subsections.

5.2. *First family of distributions (including the Weibull one) satisfying our assumptions*

In [14] and [18], the authors present a set of hypotheses on the target and proposal distributions that imply the subgeometric ergodicity of the Markov Chain. The first hypothesis concerns the target distribution:

(E1) The target density π is continuous and positive on $\mathbb{R}^d$ and there exist $m \in (0,1)$, $r \in (0,1)$, positive constants d_i, D_i, $i = 0,1,2$ and $R_0 < \infty$ such that, if $|x| \geq R_0$, $x \mapsto \pi(x)$ is twice continuously differentiable and

$$\left\langle \frac{\nabla\pi(x)}{|\nabla\pi(x)|}, \frac{x}{|x|} \right\rangle \leq -r$$

$$d_0|x|^m \leq -\ln \pi(x) \leq D_0|x|^m$$

$$d_1|x|^{m-1} \leq -\ln \nabla\pi(x) \leq D_1|x|^{m-1}$$

$$d_2|x|^{m-2} \leq -\ln \nabla^2\pi(x) \leq D_2|x|^{m-2}\,.$$

Among others, the Weibull distribution on $\mathbb{R}_+$ $\pi : x \mapsto \beta\eta x^{\eta-1}\exp(-\beta x^\eta)$ with $\beta > 0$ and $\eta \in (0,1)$ verifies those conditions.
We also need some conditions on the proposal distribution:

(E2) there exists $\epsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \epsilon$. Moreover, q_θ is symmetric and bounded away from zero in a neighborhood of zero, is compactly supported i.e. it exists $c(q_\theta)$ such that, for all $|y| > c(q_\theta)$, $q(y) = 0$ and it exists $C > 0$, $\beta \in (0,1)$ such that for all $\theta, \theta' \in \Theta$,

$$\int_X |q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz) \leq C|\theta - \theta'|^\beta\,.$$

Remark 5.1. *This compactly supported condition could be relaxed with appropriate moment conditions.*

We can now prove the following theorem:

Theorem 5.1. *Let π and q_θ be distributions satisfying (E1) and (E2) and consider the processus defined in (13) with ϵ and Δ two sequences verifying (A4). Then, (A1), (A2) and (A3) are verified. Moreover, $\theta_n \to \theta_\pi$ w.p. 1 where $\theta_\pi := (\mu_\pi, \Gamma_\pi)$ is the unique stationary point of (θ_n).*

Proof. According to the theorem 3.1 of [14], if (E1) and (E2) are satisfied, it exists ξ_0 such that for all $\xi \leq \xi_0$, it exists $c > 0$, $W = \pi^{-\xi}$ and $\phi(v) = cx(1 + \ln(x))^{-2\frac{1-m}{m}}$ verifying:

$$PW + \phi \circ W \leq W + b\mathbb{1}_C\,.$$

Hence, we have a subgeometric drift condition. It is then possible to compute the associated rate of convergence: $r_k = \exp(cx^{\frac{m}{2-m}})$.

As stated in proposition 5.1, the condition (A1) is verified and (A2) is satisfied using the theorem 2.2 of [28].
We will prove (A3) using the theorem 3.1.

First, the condition (DRI) is verified with $V^2 = W$ and $p = 2$. Indeed, the drift condition is given above while the existence of small sets is insured given the continuity of π and hypothesis (E2) (see Theorem 2.2 of [28]).

We then verify the hypothesis (H1). Given the value of r_k, the sum of the $r_k^{1/q}$ will be finite for any $q > 0$. Moreover, $\sup_{\theta\in\Theta} ||H_\theta||_{(\phi\circ V^p)^{1/s}} < \infty$ if and only if $x^2\pi^{\xi/s}(x)(1 - \xi \ln \pi(x))^{\frac{2(1-m)}{sm}} < \infty$. This will be true for any $s > 0$ as $\pi(x) \leq \exp(-D_0 x^m)$.

Concerning (H2), as discussed in remark 3.4, (H2-i) is verified for polynomial rates of convergence k^d with $d > q$. Using the fact that $r_k^{1/q} > k^d$ for k big enough, we can conclude that (H2-i) is verified in this case.
To verify (H2-ii), we remark that

$$|H_\theta(x) - H_{\theta'}(x)| \leq |\mu - \mu'|(1 + |\mu + \mu'| + 2|x|) + |\Gamma - \Gamma'|\,.$$

Since $||x||_{V^p} < \infty$, we obtain the inequality (H2-ii) for any $\beta \leq 1$.
We now interest ourselves in (H2-iii). using the definition of the kernel P_θ, we have that

$$\begin{aligned}
|P_\theta g(x) - P_{\theta'} g(x)| &\leq \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|g(x+z)\lambda^{Leb}(dz) \\
&\quad + g(x)\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz) \\
&\leq ||g||_{(\phi\circ V^p)^{1/s}}(\phi \circ V^p)^{1/s}(x)\Big(\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\frac{(\phi \circ V^p)^{1/s}(x+z)}{(\phi \circ V^p)^{1/s}(x)}\lambda^{Leb}(dz) \\
&\quad + \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\Big)\,.
\end{aligned}$$

Hence, writing $W = (\phi \circ V^p)^{1/s}$, we need to study:

$$\alpha(x, x+z)\frac{W(x+z)}{W(x)} = \left(1 \wedge \frac{\pi(x+z)}{\pi(x)}\right)\frac{\pi^{-\xi}(x+z)(1 - \xi \ln \pi(x+z))^{-\frac{2(1-m)}{m}}}{\pi^{-\xi}(x)(1 - \xi \ln \pi(x))^{-\frac{2(1-m)}{m}}}\,.$$

But, if $\pi(x+z) \geq \pi(x)$, this function is always less than 1.

If $\pi(x+z) \leq \pi(x)$, we use the growth of the function $\Phi(u) = u^{1-\xi}(1-\xi \ln(u))$ for $u \leq 1$ and ξ small enough. Hence, we deduce once again that the function is less than 1.

Finally,

$$|P_\theta g(x) - P_{\theta'} g(x)| \leq 2||g||_{(\phi\circ V^p)^{1/s}}(\phi \circ V^p)^{1/s}(x)\int_X |q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\,.$$

Hence, the hypothesis (E2) allows us to conclude on the validity of (H2-iii).

Finally, we just have the hypothesis (H3) to prove. According to the theorem 16.1.9 of [15], $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. But, V^p converges towards infinity at infinity and is continuous so, $\mathcal{C}$ is compact. Hence, there exists a lower bound of $\frac{\phi \circ V^p}{V^p}$ continuous on $\mathcal{C}$ and (H3) is verified.

All the hypothesis of the theorem 3.1 are thus verified and we can apply it to conclude. □

Hence, we have proven the convergence of the Metropolis Hastings algorithm under a subgeometric ergodicity condition. In the next subsection we will interest ourselves in the case where the rate of convergence is not only subgeometric but polynomial and, once again, prove the convergence of a stochastic approximation.

5.3. Second usual family (including the Pareto distribution) covered by our framework

In [18], the authors give other conditions on the target density for the SRWMH kernel to be subgeometric ergodic when we work in $\mathbb{R}$:

(E3) π is continuous on $\mathbb{R}$ and there exist some finite constants $\alpha > 1$, $M > 0$, $C > 0$ and a function $\rho : \mathbb{R} \to [0, \infty)$ verifying $\lim_{x \to \infty} \rho(x) = 0$ such that for all $|x| > M$, π is strictly decreasing and, for all $y \in \{z \in \mathbb{R} | \pi(x + z) \leq \pi(x)\}$,

$$\left| \frac{\pi(x + y)}{\pi(x)} - 1 + \alpha y x^{-1} \right| \leq C|x|^{-1} \rho(x) y^2 \, .$$

This class of distributions contains in particular the Pareto distributions ($\pi(x) \propto x^{-\alpha}$) as well as many heavy tail distributions. We also need some hypothesis on our proposal:

(E4) there exists $\epsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \epsilon$. Moreover, q_θ is symmetric and there exists $\xi \geq 1$ such that $\int |y|^{\xi+3} q_\theta(y) dy < \infty$.

Under those conditions, we can state the following proposition, proved in [18].

Proposition 5.2. *Assume (E3) and (E4). Set $u = \xi \wedge \alpha + 1$ and $W(x) = 1 + |x|^u$. Then, it exists $c > 0$ and a small set C such that, if we set $\phi(x) = cx^{1-2/u}$,*

$$P_\theta W(x) + \phi \circ W(x) \leq W(x) + b \mathbb{1}_C \, .$$

Under such a drift condition, we are able to deduce the rate of convergence using the value of ϕ [14]: for all $k \in \mathbb{N}$, $r_k \propto k^{u/2-1}$.

Theorem 5.2. *Let π and q_θ be distributions on $\mathbb{R}$ satisfying (E3) and (E4) with $\xi \wedge \alpha > 5$ and consider the model defined in (13) with ϵ and Δ two sequences verifying (A4). Assume also that (H2-iii) is verified. Then, (A1), (A2) and (A3) are verified. Moreover, $\theta_n \to \theta_\pi$ w.p. 1 where $\theta_\pi := (\mu_\pi, \Gamma_\pi)$ is the unique stationary point of (θ_n).*

Remark 5.2. *In this theorem, we suppose that (H2-iii) is verified. This condition depends of the function π. Given the functions V and ϕ chosen here, we need, $\forall x, z \in \mathbb{R}$,*

$$\begin{cases} \pi(x+z) \leq \pi(x) & \Longrightarrow \quad \frac{\pi(x+z)}{\pi(x)} \left(\frac{1+|x+z|^u}{1+|x|^u} \right)^{\frac{u-2}{us}} \leq C \\ \pi(x+z) \geq \pi(x) & \Longrightarrow \quad |x+z| \leq C|x|\,. \end{cases} \tag{15}$$

Other conditions can appear if V or ϕ have another form. It was the case in the previous subsection when we have been able to prove this condition under the conditions (E1) and (E2). We prove this particular condition in the next section for the Pareto distribution.

Proof. (A1) is stated in proposition 5.1.
Under (E3) and (E4), P_θ is ψ-irreductible (see theorem 2.2 of [28]). Hence, we have existence and unicity of the invariant distribution π_θ. Moreover, H is measurable. Hence, (A2) is verified.
We still need to verify (A3). To do so, we will use the theorem 3.1 and prove the hypotheses (DRI) and (H1)-(H3).
The proposition 5.2 ans the theorem 2.2 of [28] give us the validity of (DRI) with $p = 2$ and $W = V^2$.

We now prove (H1). First, $\sum_{k \geq 0} \frac{1}{r_k^{1/q}}$ is finite for all $q < \frac{u-2}{2}$. Moreover, for any $\mathcal{K}$ compact of $\mathbb{R} \times \mathbb{R}_+^*$, since $(\phi \circ V^p)^{1/s} = (1 + |x|^u)^{\frac{u-2}{us}}$ and since H_θ is quadratic, $\sup_{\theta \in \mathcal{K}} ||H_\theta||_{(\phi \circ V^p)^{1/s}} < \infty$ if and only if $q > \frac{u-2}{u-4}$. Hence, we need to choose q such that:

$$\frac{u-2}{u-4} < q < \frac{u-2}{2}\,. \tag{16}$$

Since $u > 6$, such a q exists. Moreover, because $\frac{u-2}{2} > 2 = p$, we can also choose $s > p$. Hence, the condition (H1) is verified.

Concerning (H2), as discussed in remark 3.4, (H2-i) is verified if $\frac{u/2-1}{q} > 1$ which is true given Eq. (16).
Concerning (H2-ii), we have that

$$|H_\theta(x) - H_{\theta'}(x)| \leq |\mu - \mu'|(1 + |\mu + \mu'| + 2|x|) + |\Gamma - \Gamma'|\,.$$

Since $||x||_{V^p} < \infty$ because $u \geq 1$, we obtain the inequality (H2-ii) for any $\beta \leq 1$.

Hence, we only have to prove (H3) to conclude. According to the theorem 16.1.9 of [15], $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. In particular,

since $V^p(x) = 1 + |x|^u$, it exists $d_1 > 0$ such that $\{V \leq d\} = [0, d_1]$. But, $x \mapsto \frac{(\phi \circ V^p)^{1/s}(x)}{V^p(x)}$ is continuous hence, bounded on the compact $[0, d_1]$. Thus, (H3) is verified.

□

We have proved the convergence of the Metropolis Hastings algorithm under a set of hypothesis implying a polynomial rate of convergence. In the next section, we show that those hypotheses are verified for the Pareto distribution with a scale parameter more than 5.

5.4. Application to the Pareto distribution

In this application, we choose to study the case where the target distribution π is a Pareto distribution and the proposal q_θ is a normal distribution $\mathcal{N}(0, \Gamma)$. As showed in [18], the Pareto distribution $\pi(x) \propto |x|^{-\alpha}$ verifies the condition (E3). Moreover, (E4) is satisfied for any $\xi > 0$. Hence, when applying the theorem 5.2, we need $\alpha \wedge \xi > 5$ i.e. $\alpha > 5$.

We now show that the Pareto distribution verifies the condition (H2-iii):

Lemma 5.3. *Suppose that π is a Pareto distribution with shape $\alpha > 5$ and, for $\theta = (\mu, \Gamma)$, q_θ is the normal distribution $\mathcal{N}(0, \Gamma)$. Then, if P_θ is the kernel defined in (12) and $\mathcal{K}$ is a compact of $\mathbb{R}_+^*$, there exists C such that for all $\theta, \theta' \in \mathcal{K}$ and for all $g \in \mathcal{L}_{(\phi \circ V^p)^{1/s}}$*

$$||P_\theta g - P_{\theta'} g||_{(\phi \circ V^p)^{1/s}} \leq C ||g||_{(\phi \circ V^p)^{1/s}} |\theta - \theta'|^\beta .$$

Proof. As done in the proof of the theorem 5.1, writing $W = (\phi \circ V^p)^{1/s}$, we need to find an upper bound to:

$$\int_X \alpha(x, x+z) |q_\theta(z) - q_{\theta'}(z)| \frac{W(x+z)}{W(x)} \lambda^{Leb}(dz)$$
$$= \int_X \left(1 \wedge \frac{|x|^\alpha}{|x+z|^\alpha}\right) \frac{(1 + |x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1 + |x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}} |q_\theta(z) - q_{\theta'}(z)| \lambda^{Leb}(dz) .$$

But, if $|x+z|^\alpha \leq |x|^\alpha$,

$$\frac{(1 + |x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1 + |x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}} \leq 1 .$$

Similarly, if $|x+z|^\alpha \geq |x|^\alpha$, using Eq. (16), we have that $s > 1 \geq \frac{\alpha-1}{\alpha}$. Hence,

$$\frac{|x|^\alpha}{|x+z|^\alpha} \frac{(1 + |x+z|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}}{(1 + |x|^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}} \leq \left|1 + \frac{z}{x}\right|^{-\alpha} \left(1 + \left|1 + \frac{z}{x}\right|^{\alpha+1}\right)^{\frac{\alpha-1}{s(\alpha+1)}}$$

is bounded.

Finally, it exists $C > 0$ such that:

$$|P_\theta g(x) - P_{\theta'} g(x)| \leq C||g||_{(\phi \circ V^p)^{1/s}} (\phi \circ V^p)^{1/s}(x) \int_X |q_\theta(z) - q_{\theta'}(z)| dz \,.$$

But it has already been proved in [5] that, if q_θ is the normal distribution of variance Γ, for any Γ, Γ' in a compact subset $\mathcal{K}$ of $\mathbb{R}_+^*$,

$$\int_{\mathbb{R}} |q_\theta(z) - q_{\theta'}(z)| dz \leq \frac{1}{\Gamma_{\min}} |\Gamma - \Gamma'|$$

where $\Gamma_{\min}$ is the minimum value of $\mathcal{K}$ which allows us to conclude for any $\beta \leq 1$.

□

Theorem 5.4. *Suppose that π is a Pareto distribution with shape $\alpha > 5$ and, for $\theta = (\mu, \Gamma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$, q_θ is the normal distribution $\mathcal{N}(0, \Gamma)$. Let $(Z_n)_{n \in \mathbb{N}}$ be the Markov chain as described in 2 with P_θ defined in (12) and H defined in (14). Suppose that $(\Delta_n)_{n \in \mathbb{N}}$ and $(\epsilon_n)_{n \in \mathbb{N}}$ are two sequences verifying (A4). Then, $\theta_n \to \theta_\pi = (\mu_\pi, \theta_\pi)$ w.p. 1.*

Proof. It is a consequence of the theorem 5.2 and lemma 5.3. All the conditions have already been proved. □

6. Conclusion

We have been able to relax the condition of geometric ergodicity previously needed to ensure the convergence of stochastic approximations with Markovian dynamics. The new theorem implies the convergence for Markov Chains that are only subgeometric ergodic with mild hypotheses on the rate of convergence and the drift condition. In particular, this enables us to prove the convergence of a Metropolis Hastings algorithm with adapted variance, first in the case of the Weibull distribution with a shape parameter between 0 and 1 and then in the case of the Pareto distribution with a shape parameter more than 5. This new theorem should hence be applicable in a broader range of cases where the geometric ergodicity is not verified.

References

[1] Abounadi, J., Bertsekas, D. P. and Borkar, V. (2002). Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. *SIAM Journal on Control and Optimization* **41** 1–22.

[2] Allassonnière, S., Durrleman, S. and Kuhn, E. (2015). Bayesian mixed effect atlas estimation with a diffeomorphic deformation model. *SIAM Journal on Imaging Sciences* **8** 1367–1395.

[3] Allassonnière, S., Kuhn, E., Trouvé, A. et al. (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli* **16** 641–678.
[4] Allassonniere, S., Younes, L. et al. (2012). A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics* **6** 125–160.
[5] Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization* **44** 283–312.
[6] Andrieu, C. and Robert, C. P. (2001). *Controlled MCMC for optimal sampling*. INSEE.
[7] Atchadé, Y., Fort, G. et al. (2010). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli* **16** 116–154.
[8] Atchadé, Y. F., Fort, G. et al. (2012). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels: Part II. *Bernoulli* **18** 975–1001.
[9] Benveniste, A., Métivier, M. and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations* **22**. Springer Science & Business Media.
[10] Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization* **38** 447–469.
[11] Chen, H.-F. (2006). *Stochastic approximation and its applications* **64**. Springer Science & Business Media.
[12] Chen, H.-F., Guo, L. and Gao, A.-J. (1987). Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications* **27** 217–231.
[13] Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics* 94–128.
[14] Douc, R., Fort, G., Moulines, E., Soulier, P. et al. (2004). Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability* **14** 1353–1377.
[15] Douc, R., Moulines, E., Priouret, P. and Soulier, P. (2018). *Markov chains*. Springer.
[16] Duflo, M. (2013). *Random iterative models* **34**. Springer Science & Business Media.
[17] Fort, G. and Moulines, E. (2000). V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statistics & probability letters* **49** 401–410.
[18] Fort, G. and Moulines, E. (2003). Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications* **103** 57–99.
[19] Haario, H., Saksman, E., Tamminen, J. et al. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242.
[20] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications* **85** 341–361.
[21] Karimi, B., Miasojedow, B., Moulines, É. and Wai, H.-T. (2019).

Non-asymptotic analysis of biased stochastic approximation scheme. *arXiv preprint arXiv:1902.00629*.

[22] Kuhn, E., Matias, C. and Rebafka, T. (2019). Properties of the Stochastic Approximation EM Algorithm with Mini-batch Sampling. *arXiv preprint arXiv:1907.09164*.

[23] Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications* **35**. Springer Science & Business Media.

[24] Le Corff, S., Fort, G. et al. (2013). Online expectation maximization based algorithms for inference in hidden Markov models. *Electronic Journal of Statistics* **7** 763–792.

[25] Mandt, S., Hoffman, M. D. and Blei, D. M. (2017). Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research* **18** 4873–4907.

[26] Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

[27] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics* 400–407.

[28] Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

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

[30] Spall, J. C. et al. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control* **37** 332–341.

[31] Yang, C. (2008). Recurrent and ergodic properties of Adaptive MCMC. *Preprint*.