On the convergence of stochastic approximations under a subgeometric ergodic Markov dynamic

Vianney Debavelaere, Stanley Durrleman, Stéphanie Allassonnière

Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

On 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] *Inria Center of Paris, Sorbonne Université, CNRS UMR 7225, Inserm U 1127, Institut du Cerveau (ICM), Paris, France, e-mail:* stanley.durrleman@inria.fr

[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. 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. It is 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 sequence of random variables following a parametric distribution which depends on $(\theta_n)_{n\in\mathbb{N}}$, and $(\Delta_n)_{n\in\mathbb{N}}$ is a step sequence. The convergence of such a stochastic approximation has already been proved under an assumption of geometric ergodicity of the Markov dynamic. 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 relax this hypothesis and prove the convergence of the stochastic approximation by only assuming 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 first study an adaptive Markov Chain Monte Carlo algorithm where the proposal distribution is adapted by learning the variance of a heavy tail target distribution. We then apply our work to the Independent Component Analysis when a positive heavy tail noise leads to a subgeometric dynamic in an Expectation Maximization algorithm.

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 $h : \Theta \to \mathbb{R}$. Numerical schemes such as Newton's methods have been developed to provide a numerical solution to this equation. In statistics, the

problem is further increased by the fact that h is not known, but only noisy values of it, or of its gradient. This problem appears across different domains such as stochastic optimization [32, 37], Expectation Maximization algorithms [3, 27] or reinforcement learning [1, 12] for instance. In all cases, solutions to this problem often take the form of an iterative sequence $(\theta_n)_{n\in\mathbb{N}}$ that converges towards a point θ^* in the set of solutions of $h(\theta) = 0$. The general class of stochastic approximation methods, such as Robbins-Monro methods, falls within this framework. These methods produce a sequence of the form:

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

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 sequence of random variables. In that case, h is called the mean field. This procedure, first developed in [34], has been studied under various sets of hypotheses, see [1, 11, 12, 15, 16, 20, 28] among many other works.

In this paper, we focus on the case of a state-dependent noise with a Markovian dynamic. The sequence $(\zeta_n)_{n\in\mathbb{N}}$ takes the form of $(H_{\theta_n}(X_n))_{n\in\mathbb{N}}$, with $h(\theta_n)$ being the expectation of H_{θ_n}:

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

The sequence $(X_n, \theta_n)_{n\in\mathbb{N}}$ is a Markov chain on $\mathcal{X}\times\Theta$. For all $\theta\in\Theta$, H_θ is a function from the state space $\mathcal{X}$ to the parameter space Θ.

The assumption of state-dependent noise is met for instance in stochastic gradient descent or Metropolis Hastings algorithms. Eq. (1) is also used as a step in stochastic optimization algorithms where the parameter to estimate is a function of θ_n. These algorithms include the Stochastic Approximation Expectation Maximization Markov Chain Monte Carlo (SAEM MCMC) algorithm [2, 3, 17]. Eq. (1) also appears in some adaptive MCMC algorithms where the proposal distribution depends on a parameter θ. They are used to adapt the variance of the proposal across iterations for better sampling [5, 6, 23, 35].

The convergence of stochastic approximation algorithms has been studied in [5] for state-dependent noise. Conditions to ensure convergence include control of the fluctuations of the Markov Chain and of the regularity of the solution of a Poisson equation. These conditions are difficult to verify in practice. Authors introduce then a more restrictive, but more practical condition: the Markov chain must satisfy drift conditions implying a *geometric ergodicity* of the chain. This condition amounts to assuming the convergence of the kernel of the Markov Chain towards its invariant distribution at a geometric rate. Further developments lead to prove the convergence of the SAEM MCMC algorithm [3], some adaptive MCMC algorithms [5] and mini-batch MCMC [27] under the same conditions.

Nevertheless, the ergodicity condition is a limiting factor in practice. For instance, the sequence $(X_n)_{n\in\mathbb{N}}$ is often sampled using a Metropolis Hastings algorithm. The ergodic condition is not met if one targets heavy tail distributions

such as Weibull or Pareto distribution [18, 21, 22, 24]. The models for independent component analysis presented in [4] with non-Gaussian distributions of the sources or the noise do not meet the condition either. These examples show that these methods may be used in practice without any theoretical guarantee of convergence.

This situation leads us to study the convergence of such stochastic algorithms for Markov chains with a relaxed assumption of subgeometric ergodicity. The convergence of adaptive MCMC algorithms under subgeometric constraints has been studied in [7, 8, 36, 38]. To the best of our knowledge, there are no results on the convergence for subgeometric Markovian dynamic in the general case.

In this paper, we propose a general set of hypotheses, under which we prove the convergence of stochastic approximations with subgeometric Markovian dynamics. Our hypotheses are essentially about the rate of convergence of the Markov Chain and the regularity of its kernel. Most of the polynomial rates of convergence satisfy these hypotheses. Furthermore, the proof shows the regularity of the solution of the Poisson equation under the same subgeometric conditions. We use this result to prove two corollaries. The first corollary proves the convergence of a stochastic approximation used to adapt the variance of the proposal within a Metropolis Hastings algorithm. We prove this convergence for two different classes of heavy tail target distributions including the Weibull and the Pareto distributions among others. The second corollary is about the independent component analysis model where distributions with positive heavy tails lead to a subgeometric ergodic Markov Chain in a Stochastic Approximation Expectation Maximization Monte Carlo Markov Chain (SAEM MCMC) algorithm.

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 exist probability measures π_θ such that, for any $\theta \in \Theta$, $h(\theta) = \mathbb{E}_{\pi_\theta}(H_\theta(X))$ with $H_\theta : \mathcal{X} \mapsto \Theta$.

2.1. Markovian dynamic

Let $\Delta = (\Delta_n)_{n\in\mathbb{N}}$ be a 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:

$$\theta_{n+1} = \begin{cases} \theta_n + \Delta_{n+1} H_{\theta_n}(X_{n+1}) & \text{and } X_{n+1} \sim P_{\theta_n}(X_n, .) \quad \text{if } \theta_n \in \Theta \\ \theta_c & \text{and } X_{n+1} = x_c \quad \text{if } \theta_n \notin \Theta \,. \end{cases} \tag{2}$$

Keeping notations and hypotheses labels from [5], 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 $(\theta, x) \in \Theta \times \mathcal{X}$.

The existence and uniqueness of the invariant distribution can be verified under the classical conditions of irreducibility and recurrence [33]. 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 $\mathcal{W}_{M_1} := \{\theta \in \Theta, w(\theta) \leq M_1\}$ 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 denote by Q_{Δ_n} the sequence of transition probabilities that generate the inhomogeneous Markov chain (Y_n^Δ).

2.2. Truncation process

To ensure convergence of the sequence towards a root of h, the sequence $(\theta_n)_{n\in\mathbb{N}}$ is required to remain in a given compact set. This assumption is rarely satisfied. To alleviate this constraint, we introduce the usual trick which consists in re-projecting on increasing compact sets. It is then proved that the sequence will be projected only a finite number of times along the algorithm. Using this trick, the sequence $(\theta_n)_{n\in\mathbb{N}}$ now remains in a compact set of Θ. We detail this process below.

We assume that there exists $(\mathcal{K}_n)_{n\in\mathbb{N}}$ 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 $(\varepsilon_n)_{n\in\mathbb{N}}$ 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 a homogeneous Markov chain on $\mathcal{X}\times\Theta\times\mathbb{N}\times\mathbb{N}$ by

$$Z_n = (X_n, \Theta_n, \kappa_n, \nu_n) \tag{3}$$

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 $|\theta_{n+1} - \theta_n| \leq \varepsilon_n$ 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-initializations 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 large.

2.3. Control of the fluctuations and main convergence theorem

In this section, we state two last hypotheses about the control of fluctuations before presenting the theorem proved in [5]. In that 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 $(\varepsilon_k)_{k\in\mathbb{N}}$, $\sigma(\mathcal{K}) = \inf(k \geq 1, \theta_k \notin \mathcal{K})$ and $\nu_\varepsilon = \inf(k \geq 1, |\theta_k - \theta_{k-1}| \geq \varepsilon_k)$. 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 $g - P_\theta g = H_\theta - h(\theta)$ 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:

$$\sup_{\theta\in\mathcal{K}} ||H_\theta||_W < \infty \tag{4}$$

$$\sup_{\theta\in\mathcal{K}} ||g_\theta||_W + ||P_\theta g_\theta||_W < \infty \tag{5}$$

$$\sup_{\theta,\theta'\in\mathcal{K}} ||\theta - \theta'||^{-\alpha} \left(||g_\theta - g_{\theta'}||_W + ||P_\theta g_\theta - P_{\theta'} g_{\theta'}||_W\right) < \infty \tag{6}$$

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

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

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

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}^{\Delta}_{x,\theta}[W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon\geq k}] \leq C W^p(x)\,. \tag{8}$$

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 the hypothesis of geometric ergodicity of the Markov chain. In the next sections, we will relax this ergodicity condition to be able to consider subgeometric ergodic chains.

Finally, the last condition concerns the step size sequences:

(A4) The sequences $(\Delta_k)_{k\in\mathbb{N}}$ and $(\varepsilon_k)_{k\in\mathbb{N}}$ are non-increasing, positive and satisfy $\sum_{k=0}^{\infty}\Delta_k=\infty$, $\lim_{k\to\infty}\varepsilon_k=0$ and

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

where p and α are defined in (A3).

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

Theorem 2.1. *[5] 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 (3). 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. where $\mathcal{L}$ is defined in (A1).*

Of the four conditions (A1) to (A4), (A3) is often the most difficult to verify and we need more practical conditions. In particular, in [5], the authors show that drift conditions imply (A3). However, those drift conditions are only true for geometric ergodic Markov chains. In a lot of cases, this ergodicity is not satisfied. 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 ensure the validity of (A3). The new theorem then allows us to verify the convergence in a broader range of cases, some of them being presented in sections 5 and 6.

3. Convergence of the stochastic approximation sequence under subgeometric conditions

In this section, we state 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-irreducible 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

$$\sup_{\theta\in\mathcal{K}}P_\theta V^p(x)+\phi\circ V^p(x)\leq V^p(x)+b\mathbb{1}_{\mathcal{C}}(x)\qquad\forall x\in\mathcal{X}\tag{9}$$

$$\inf_{\theta\in\mathcal{K}}P_\theta(x,A)\geq\delta_0\nu(A)\qquad\forall x\in\mathcal{C},\forall A\in\mathcal{B}(\mathcal{X})\,.\tag{10}$$

Remark 3.1. *We could consider the following, more general, drift condition:*

there 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 kernel P_θ verifies a subgeometric drift condition [19]. In particular, it implies the existence of a stationary distribution π_θ for all $\theta \in \mathcal{K}$ as well as a uniform subgeometric ergodicity on all compacts of Θ. Hence, for all $\theta \in \Theta$, there exist 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 [18] that, under such a subgeometric ergodicity condition, we can choose a rate of convergence $(r_k)_{k\in\mathbb{N}}$ that only depends on the function ϕ and so only on the fixed compact $\mathcal{K}$. Similarly, it has been proved that the constant C_θ is bounded on all compact $\mathcal{K}$. Hence, there exist 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 $\theta \in \mathcal{K}$,

$$\sup_{\theta\in\mathcal{K}} r_k^{1/q} ||P_\theta^k f - \pi_\theta(f)||_{(\phi\circ V^p)^{1/s}} \leq C_\mathcal{K} ||f||_{(\phi\circ V^p)^{1/s}}\,. \tag{11}$$

We will see in the following that several hypotheses 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 any pair* Ψ_1 *and* Ψ_2 *of inverse Young functions i.e. two strictly increasing continuous functions on* $\mathbb{R}_+$ *verifying for all* x, y *in* $\mathbb{R}_+$, $\Psi_1(x)\Psi_2(y) \leq x + y$. *Under the subgeometric drift condition, 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 hypotheses that we will need in order 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 $\mathcal{K}$, there exist $q > 0$ 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. *We will show in section 5.3 that this hypothesis can be verified even for polynomial rates of convergence ($r_k = k^d$ with $d > 2$ in that example). This hypothesis can be seen as a compromise in the choice of q and s between the rate of convergence r_k and the regularity of H_θ. 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. (6).

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

(i) there exist $T_{\theta,\theta'} \in \mathbb{N}^*$ and $\alpha \in (0, 1)$ such that

$$\sup_{\theta,\theta' \in \mathcal{K}} T_{\theta,\theta'} ||\theta - \theta'||^{\beta - \alpha} + ||\theta - \theta'||^{-\alpha} \sum_{k \geq T_{\theta,\theta'}} \frac{1}{r_k^{1/q}} < \infty \,.$$

(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. *In the condition (H2-i), $T_{\theta,\theta'}$ is a positive integer. It implies in particular $\beta \geq \alpha$.*
This condition 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 $T_{\theta,\theta'} = 1 \vee \left\lfloor ||\theta - \theta'||^{-\frac{\alpha}{d-1}} \right\rfloor$ and we have:

$$||\theta - \theta'||^{-\alpha} \sum_{k=T_{\theta,\theta'}}^{\infty} \frac{1}{k^d} \sim_{\theta \to \theta'} \frac{1}{d-1} \,.$$

Moreover, if $||\theta - \theta'|| \leq 1$, $T_{\theta,\theta'} ||\theta - \theta'||^{\beta - \alpha} = ||\theta - \theta'||^{\beta - \alpha - \frac{\alpha}{d-1}}$. 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 making divergent quantities appear. This iteration was however one of the keys of the proof of the condition 8. 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) there 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 [19], 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. In particular, the behaviour of ϕ on the petite set $\mathcal{C}$ and the hypotheses on the rate of convergence $(r_k)_{k\in\mathbb{N}}$ will be of the upmost importance.

The first important result is the fact that we are able to dominate 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 $W = V$ in (A3) to be able to find an upper bound of the expectation of $W^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon \geq k}$ (see Eq. (8)). Hence, we use this control of the V-norm to control the different quantities in Eq. (4), (5) and (6) using the rate of convergence given by Eq (11). 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 explained lemma 4.2.

We then want to prove the condition (6) (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 (7) is easily proved by iterating the drift condition, we still need to prove the condition (8). 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 prove (8) 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. (5)

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, there 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 (5).

Lemma 4.2. *Suppose (DRI). Then, the Poisson equation $g - P_\theta g = H_\theta - h(\theta)$ 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 [19] 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 any compact $\mathcal{K}$, there exist 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,

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

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

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

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. (6)

We now want to prove the condition given by Eq. (6). In particular, we need the hypotheses on the regularity in θ of H_θ and P_θ presented in condition (H2). We begin by proving two lemmas 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 $k \geq 0$,*

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

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. (11), we know that, for any $l \geq 0$,

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

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),

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

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

We now prove that h is β-Hölder. We will use this property to finally be able to prove (6).

Lemma 4.4. *Assume (DRI), (H1) and (H2). Then,*

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

Proof. We use the following decomposition of $|h(\theta) - h(\theta')|$ for $x_0 \in \mathcal{X}$, $(\theta, \theta') \in \mathcal{K}^2$ and $k \in \mathbb{N}$:

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

with:

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

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

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

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

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

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

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

Finally, because $\frac{1}{r_k^{1/q}} \to 0$, there 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 (6).

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), lemmas 4.3 and 4.4, we have that, for $x \in \mathcal{X}$, $k \in \mathbb{N}$ and $\theta, \theta' \in \mathcal{K}$,

$$\begin{aligned} D_k(x,\theta,\theta') &:= |P_\theta^k H_\theta(x) - h(\theta) - P_{\theta'}^k H_{\theta'}(x) + h(\theta')| \\ &\leq |P_\theta^k H_\theta(x) - P_\theta^k H_{\theta'}(x)| + |P_\theta^k H_{\theta'}(x) - P_{\theta'}^k H_{\theta'}(x)| + |h(\theta) - h(\theta')| \\ &\leq C||\theta - \theta'||^\beta (\phi \circ V^p)^{1/s}(x) \end{aligned}$$

where we have used the fact that $(\phi \circ V^p)^{1/s}(x) \geq \phi(1) > 0$.

On the other hand, using the ergodicity of the Markov Chain (11) and (H1), there exists $c > 0$ such 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 $T \geq t$ by splitting the sum at $k = T$ and using the two upper bounds found above, we have:

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

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

Remark 4.1. *Here, we have in fact proved that, under the hypotheses (DRI), (H1) and (H2), the solution of the Poisson equation is α-Hölder.*

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:

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

and we conclude using the fact that for any $x \in \mathcal{X}$, $V^p(x) \geq 1$.

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

4.4. Proof of Eq. (8)

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 a sequence* $(\varepsilon_k)_{k\in\mathbb{N}}$ *and a constant* C *such that for any sequence* Δ *and for any* $x \in \mathcal{X}$,

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

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

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

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

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

If $\delta < 1$, we note $\tau_k = Card(X_i | X_i \in \mathcal{C} \text{ for } 1 \leq i \leq k)$ the number of elements $(X_i)_{1\leq i\leq k}$ belonging to $\mathcal{C}$. Then,

$$\begin{aligned}\mathbb{E}^{\Delta}_{x,\theta}\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\right) &= \mathbb{E}^{\Delta}_{x,\theta}\left(\mathbb{E}^{\Delta}_{x,\theta}\left(PV^p(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\middle|\mathcal{F}_{k-1}\right)\right)\\ &\leq \mathbb{E}^{\Delta}_{x,\theta}\left((1-\delta\mathbb{1}_{X_{k-1}\in\mathcal{C}})V^p(X_{k-1}) + b\mathbb{1}_{X_{k-1}\in\mathcal{C}}\right)\end{aligned}$$

Hence, at each iteration $i \leq k-1$, if $X_i \in \mathcal{C}$, we multiply the expression by $(1-\delta)$ and add b. Such a case happens τ_{k-1} times. Otherwise, we keep the same expression as before, but at the rank $i-1$. By iterating, we have:

$$\begin{aligned}\mathbb{E}^{\Delta}_{x,\theta}\left(V^p(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq k}\right) &\leq \mathbb{E}^{\Delta}_{x,\theta}\left((1-\delta)^{\tau_{k-1}}V^p(x) + b\sum_{i=0}^{\tau_{k-1}-1}(1-\delta)^i\right)\\ &\leq V^p(x) + \frac{b}{\delta}\,.\end{aligned}$$

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 lemmas 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 in simulating a Markov Chain $(X_n)_{n\in\mathbb{N}}$ 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:

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} & \text{if } \pi(x) \neq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

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

$$P(x,A) = \int_A \alpha(x,y)q(x-y)\lambda^{Leb}(dy) + \mathbb{1}_A(x)\int_X (1-\alpha(x,y))q(x-y)\lambda^{Leb}(dy)\,. \tag{13}$$

The choice of the proposal distribution q is of crucial importance. In particular, proposal distributions with a too small or too large covariance matrix lead to a highly correlated Markov Chain. To overcome this difficulty, the authors of [23] 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.

We can then adapt the value of Γ using the following algorithm:

$$\begin{cases} \mu_{n+1} = \mu_n + \Delta_{n+1}(X_{n+1} - \mu_n) \\ \Gamma_{n+1} = \Gamma_n + \Delta_{n+1}\left((X_{n+1}-\mu_n)(X_{n+1}-\mu_n)^T - \Gamma_n\right) \end{cases} \tag{14}$$

with $X_{n+1} \sim P_{\theta_n}(X_n, .)$ where $\theta_n = (\mu_n, \Gamma_n)$ and with $(\Delta_n)_{n\in\mathbb{N}}$ a non-increasing 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

$$H_\theta(x) = (x-\mu, (x-\mu)(x-\mu)^T - \Gamma)\,. \tag{15}$$

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

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

with μ_π and Γ_π respectively the mean and variance of π.

This algorithm has already been studied in [5]. In that paper, the authors make a 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.
Within 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 while still guaranteeing 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*

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

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

To prove (A3), we need some hypotheses on the behaviour of π. In particular, we will verify that we can apply the theorem 3.1 under two different 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 [18] and [22], 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

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

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

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

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

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 exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. Moreover, q_θ is symmetric, bounded away from zero in a neighborhood of zero, and is compactly supported. We also assume that there exist $C > 0$ and $\beta \in (0, 1)$ such that for all $(\theta, \theta') \in \Theta^2$,

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

Remark 5.1. *The 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 process defined in (14) 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 $(\theta_n)_{n\in\mathbb{N}}$.*

Proof. According to the theorem 3.1 of [18], if (E1) and (E2) are satisfied, there exists ξ_0 such that for all $\xi \leq \xi_0$, there exist $c > 0$, $W := \pi^{-\xi}$ and $\phi : x \mapsto cx(1 + \ln(x))^{-2\frac{1-m}{m}}$ 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: $r_k = \exp(ck^{\frac{m}{2-m}})$.

As stated in proposition 5.1, the condition (A1) is verified and (A2) is satisfied using the theorem 2.2 of [35].
We will prove (A3) using the theorem 3.1.

First, the condition (DRI) is verified with $V^2 = \pi^{-\xi}$ and $p = 2$. Indeed, the drift condition is given above while the existence of small sets is ensured given the continuity of π and hypothesis (E2) (see Theorem 2.2 of [35]).

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, $\sup_{\theta\in\Theta} ||H_\theta||_{(\phi\circ V^2)^{1/s}} < \infty$ 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 $||x||_{V^2} < \infty$, 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

$$\begin{aligned}|P_\theta g(x) - P_{\theta'} g(x)| &\leq \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|g(x+z)\lambda^{Leb}(dz) \\ &\quad + g(x)\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz) \\ &\leq ||g||_{(\phi\circ V^2)^{1/s}}(\phi\circ V^2)^{1/s}(x)\Big(\int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\frac{(\phi\circ V^2)^{1/s}(x+z)}{(\phi\circ V^2)^{1/s}(x)}\lambda^{Leb}(dz) \\ &\quad + \int_X \alpha(x, x+z)|q_\theta(z) - q_{\theta'}(z)|\lambda^{Leb}(dz)\Big)\,.\end{aligned}$$

Hence, writing $\Psi := (\phi\circ V^2)^{1/s}$, we need to study:

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

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 $\Phi(u) = u^{1-\xi}(1 - \xi\ln(u))^{-\frac{2(1-m)}{m}}$ for u in a compact and ξ small enough. Hence, we deduce once again that the function is less than 1.

Finally,

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

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 [19], $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. But, V^2 converges towards infinity at infinity and is continuous so, $\mathcal{C}$ is compact. Hence, because $\frac{\phi\circ V^2}{V^2}$ is continuous, there exists a lower bound of $\frac{\phi\circ V^2}{V^2}$ on $\mathcal{C}$ and (H3) is verified.

All the hypotheses 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 [22], 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 hypotheses on our proposal:

(E4) There exist $\varepsilon > 0$ and $r < \infty$ such that $y < r \implies q_\theta(y) \geq \varepsilon$. 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 [22].

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

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

Under such a drift condition, we are able to deduce the rate of convergence using the value of ϕ [18]: 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 (14) 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 $(\theta_n)_{n\in\mathbb{N}}$.*

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

$$\begin{cases} \pi(x+z) \leq \pi(x) & \implies \quad \frac{\pi(x+z)}{\pi(x)} \left(\frac{1+|x+z|^u}{1+|x|^u} \right)^{\frac{u-2}{us}} \leq C \\ \pi(x+z) \geq \pi(x) & \implies \quad |x+z| \leq C|x| \,. \end{cases} \tag{16}$$

Other conditions can appear if V or ϕ have another form. It was the case in the previous subsection where we have been able to prove this condition under

the hypotheses (E1) and (E2). We prove this particular condition (H2-iii) in the next section in the case of the Pareto distribution.

Proof. (A1) is stated in proposition 5.1.
Under (E3) and (E4), P_θ is ψ-irreducible (see theorem 2.2 of [35]). Hence, we have existence and uniqueness 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 and the theorem 2.2 of [35] 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 any $q < \frac{u-2}{2}$. Moreover, recalling that $1/s + 1/q = 1$, that $(\phi \circ V^p)^{1/s} = (1 + |x|^u)^{\frac{u-2}{us}}$ and that H_θ is quadratic, for any $\mathcal{K}$ compact of $\mathbb{R} \times \mathbb{R}_+^*$, $\sup_{\theta\in\mathcal{K}} ||H_\theta||_{(\phi\circ V^2)^{1/s}} < \infty$ if and only if $q > \frac{u-2}{u-4}$. Hence, we need to choose q such that:

$$\frac{u-2}{u-4} < q < \frac{u-2}{2}\,. \tag{17}$$

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. (17).
Concerning (H2-ii), we have that

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

Since $||x||_{V^2} < \infty$ 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 [19], $\mathcal{C}$ can be chosen as $\{V \leq d\}$ with $d \in [0, \infty)$. In particular, since $V^2(x) = 1 + |x|^u$, there exists $d_1 > 0$ such that $\{V \leq d\} = [0, d_1]$. But, $x \mapsto \frac{(\phi\circ V^2)^{1/s}(x)}{V^2(x)}$ 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 hypotheses 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 [22], 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 only need to assume $\alpha \wedge \xi > 5$ 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 (13) 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 $\Psi = (\phi \circ V^p)^{1/s}$, we need to find an upper bound to:

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

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. (17), 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 since $u \mapsto u^{-\alpha}(1+u^{\alpha+1})^{\frac{\alpha-1}{s(\alpha+1)}}$ is bounded on $[1, +\infty)$.

Finally, there 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 then, 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 3 with P_θ defined in (13) and H defined in (15). Suppose that $(\Delta_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ 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. □

Thus, we have been able to prove the convergence of an adaptive MCMC algorithm targeting distributions for which the theorem proved in [5] was not enough to conclude.

6. Application to Independent Component Analysis

Independent component analysis (ICA) is a method which aims at representing a data set of random vectors as linear combinations of a fixed family of vectors with independent random weights. ICA follows somehow the same goal as the Principal Component Analysis (PCA). However, PCA imposes orthogonality between principal components which amounts to supposing that the observed vectors follow a Normal distribution. As for the ICA, it assumes a more general statistical model where the observations are decomposed on components weighted by independent random coefficients. It is sometimes called source separation. ICA has a large range of applications in medical image analysis [13, 14], computer vision [9, 10, 30], computational biology [29, 31], etc.. This method is also used to map the data set onto a smaller space (not orthogonal) as one can choose the number of components in the linear combination.

This method writes an observation $X \in \mathbb{R}^d$ as:

$$X = \sum_{j=1}^{p} \beta_j a_j + \varepsilon = A\beta + \varepsilon\,, \tag{18}$$

where $A := (a_1, ..., a_p) \in \mathbb{R}^{d\times p}$ is a parameter, $(\beta_1, ..., \beta_p)$ are independent scalars whose law q_m must be specified and ε is the additive noise.

In a lot of cases, ε is supposed to follow a normal distribution. This approximation enables to develop easily many estimation algorithms. However, numerical images are rather affected by a positive valued noise (MRI images for instance). Moreover, the Gaussian assumption reduces the study to very rapidly decreasing noise. In this example, to take into account these two bottlenecks of the Gaussian noise, we choose to model our data with a positive noise with heavy tail: the Weibull distribution.

We suppose that each coordinate of ε satisfies: $\varepsilon_j \sim \mathcal{W}(\lambda_0, \eta_0)$ with $\lambda_0 \in \mathbb{R}_+^*$ and $\eta_0 \in (0, 1)$.

To estimate A, we will use a Monte Carlo Markov Chain - Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm introduced in [25]. For this algorithm to converge, we need our joint distribution to belong to the curved exponential family i.e. to be of the following form:

$$q(X, \beta, A) = \phi(A) + \langle S(X, \beta), \psi(A) \rangle \,,$$

where $S(x, \beta)$ is called the sufficient statistic of the model.

However, it can be seen that the joint likelihood does not verify this hypothesis here. A usual work around, first introduced in [26], is to consider that all vectors of A: $(a_j)_{1 \leq j \leq p}$ are random vectors following a Gaussian prior. The goal is then to estimate the mean of this prior. This writes, for each vector a_j: $a_j \sim \mathcal{N}(a_{0,j}, \sigma_A^2 Id)$.

If X is a data set of n observations $(X^1, ..., X^n)$, we finally have, writing $A_0 = (a_{0,1}, ..., a_{0,p}) \in \mathbb{R}^{d \times p}$:

$$\begin{aligned} \log q(X, \beta, A, A_0) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left((\eta_0 - 1) \log(X_j^i - (A\beta^i)_j) - \left(\frac{X_j^i - (A\beta^i)_j}{\lambda_0} \right)^{\eta_0} \right) \\ + \sum_{i=1}^{n} q_m(\beta^i) - \frac{||A - A_0||^2}{\sigma_A^2} + C \end{aligned} \tag{19}$$

where $||A - A_0||^2 = \sum_{j=1}^{p} ||a_j - a_{0,j}||_2^2$.

The joint distribution now belongs to the curved exponential family. Indeed, it can be written as:

$$\log q(X, \beta, A, A_0) = \phi(A_0) + < S(X, \beta, A), \psi(A_0) > + \tilde{S}(X, \beta, A) \tilde{\phi}(A_0)$$

with:

$$\begin{cases} \phi(A_0) = \frac{||A_0||^2}{\sigma_A^2} + C \\ S(X, \beta, A) = A \\ \psi(A_0) = -2A_0 \\ \tilde{S}(X, \beta, A) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left((\eta_0 - 1) \log(X_j^i - (A\beta^i)_j) - \left(\frac{X_j^i - (A\beta^i)_j}{\lambda_0} \right)^{\eta_0} \right) \\ \qquad\qquad + \sum_{i=1}^{n} q_m(\beta^i) + \frac{||A||^2}{\sigma_A^2} \\ \tilde{\psi}(A_0) = 1 \end{cases}$$

The maximum of the log-likelihood can then be expressed as a function of the sufficient statistics: the maximum of $q(X, \beta, A, \theta)$ is reached for $A_0 = \hat{\theta}(S(X, \beta, A)) = A$.

Then, the MCMC-SAEM algorithm consists in the following steps:
(i) Simulation of β, A using a Metropolis Hastings algorithm targeting the conditional distribution $q(\beta, A|X, \theta_{k-1})$.
(ii) Stochastic approximation of the sufficient statistics:

$$S_k = S_{k-1} + \Delta_k(A - S_{k-1})\,.$$

(iii) Maximization of the conditional distribution using the sufficient statistics: $\theta_k = \hat{\theta}(S_k)$.

Remark 6.1. *A fourth step not indicated above for clarity is the truncation process executed as described in section 2.2 and allowing our parameters to stay on compact sets.*

We can easily see that the described procedure is a particular case of the theorem 2.1 with P_s the kernel of the Metropolis Hastings algorithm targeting $q(\beta, A|X, \hat{\theta}(s))$ and with

$$H_s(\beta, A) = S(X, \beta, A) - s\,.$$

This problem has been tackled for instance in [4]. In that paper, the authors propose several distributions for β leading to geometrically ergodic Markov Chains. Using theorem 3.1, we are now able to tackle distributions leading to subgeometric ergodic chains which enables to introduce models with higher variability. We provide here an example of such a chain and prove convergence of the associated ICA parameters.

In the following, we suppose that all coordinates of β follow a Weibull distribution: $\forall i \in [|1, n|], \forall j \in [|1, p|], \beta_j^i \sim \mathcal{W}(\lambda_1, \eta_1)$ with $\lambda_1 \in \mathbb{R}_+^*$ and $\eta_1 \in (0, 1)$. Other distributions with heavy tails such as the Pareto distribution would yield to similar results.

Theorem 6.1. *Assume (A4), (A1i) and that the proposal distribution in the Metropolis Hastings algorithm verifies (E2). Define* $l(\theta) = \log \int q(X, \beta, A, \theta) d\beta dA$ *and* $\mathcal{L}' = \{\theta \in \hat{\theta}(\mathcal{S}) | \partial_\theta l(\theta)) = 0\}$. *We then have* $d(\theta_k, \mathcal{L}') \to 0$.

Remark 6.2. *Most of the work has in fact already been done in section 5.2. Indeed, the proof of the hypothesis (A3) follows the exact same steps as in 5.2 and thus will not be detailed here.*
Note that Condition (A1i) remains an assumption of the theorem as in many cases.

Proof. We first check the conditions (A1) (ii), (iii) and (iv). Let $w(s) = -l(\hat{\theta}(s))$. As showed in [17], this function verifies (A1) (iii) and (iv). Moreover, the authors prove that $\mathcal{L} = \mathcal{L}'$.
It is then easy to verify (A1)(ii) by remarking that $w(s) \to_{||s||\to\infty} \infty$. Since w is continuous, (A1)(ii) is verified for any $M_1 > 0$.

Concerning (A2), the theorem 2.2 of [35] gives the ψ-irreducibility of the Markov Chain and thus the existence of the unique stationary distribution π_θ. The measurability of H_θ is immediate.

We can easily verify that (E1) is true for $m = \eta_0 \vee \eta_1$. Hence, we can follow the exact same proof as in theorem 5.1 to prove that (H1), (H2) and (H3) are verified and thus the condition (A3) by theorem 3.1.

(A4) being supposed, we can apply the theorem 2.1 to conclude the proof. □

Hence, this simple example shows that the algorithm can be applied not only on simulation algorithms but also on optimization algorithms such as Expectation Maximization or stochastic gradient which are involved in many machine learning and deep learning methods.

7. Conclusion

In this paper, we relaxed the condition of geometric ergodicity previously needed to ensure the convergence of stochastic approximations with Markovian dynamics. We provide therefore theoretical guarantees for a wider class of algorithms that are used in practice.

Our main result proves the convergence of these stochastic approximations for Markov Chains which are only subgeometric ergodic assuming hypotheses on the rate of convergence and the drift condition. A corollary is 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. Another corollary applies to the convergence of a Stochastic Approximation Expectation Maximization algorithm when subgeometric Markov Chains appear. These results suggest that the main theorem could be used to show the convergence of a broader range of algorithms for which 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] Bartlett, M. S., Movellan, J. R. and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *IEEE Transactions on neural networks* **13** 1450–1464.

[10] Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation* **7** 1129–1159.

[11] Benveniste, A., Métivier, M. and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations* **22**. Springer Science & Business Media.

[12] 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.

[13] Calhoun, V. D., Adali, T., McGinty, V., Pekar, J. J., Watson, T. and Pearlson, G. (2001). fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage* **14** 1080–1088.

[14] Calhoun, V. D., Adali, T., Pearlson, G. D. and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping* **14** 140–151.

[15] Chen, H.-F. (2006). *Stochastic approximation and its applications* **64**. Springer Science & Business Media.

[16] 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.

[17] Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics* 94–128.

[18] 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.
[19] Douc, R., Moulines, E., Priouret, P. and Soulier, P. (2018). *Markov chains*. Springer.
[20] Duflo, M. (2013). *Random iterative models* **34**. Springer Science & Business Media.
[21] Fort, G. and Moulines, E. (2000). V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statistics & probability letters* **49** 401–410.
[22] Fort, G. and Moulines, E. (2003). Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications* **103** 57–99.
[23] Haario, H., Saksman, E., Tamminen, J. et al. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242.
[24] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications* **85** 341–361.
[25] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics* **8** 115–131.
[26] Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis* **49** 1020–1038.
[27] Kuhn, E., Matias, C. and Rebafka, T. (2019). Properties of the Stochastic Approximation EM Algorithm with Mini-batch Sampling. *arXiv preprint arXiv:1907.09164*.
[28] Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications* **35**. Springer Science & Business Media.
[29] Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18** 51–60.
[30] Liu, C. and Wechsler, H. (2003). Independent component analysis of Gabor features for face recognition. *IEEE transactions on Neural Networks* **14** 919–928.
[31] Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D. and Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences* **94** 10979–10984.
[32] 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.
[33] Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.
[34] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics* 400–407.
[35] Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.
[36] Rosenthal, J. and Roberts, G. (2007). Coupling and ergodicity of adap-

tive mcmc. *Journal of Applied Probablity* **44** 458–475.
[37] Spall, J. C. et al. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control* **37** 332–341.
[38] Yang, C. (2008). Recurrent and ergodic properties of Adaptive MCMC. *Preprint*.