Convergence of Markovian Stochastic Approximation with discontinuous dynamics

Amandine Schreck, Gersende Fort, Eric Moulines, Matti Vihola

Convergence of Markovian Stochastic Approximation with discontinuous dynamics

A. Schreck [‡†] G. Fort [§] E. Moulines [†] M. Vihola [¶]

March 26, 2014

Abstract

This paper is devoted to the convergence analysis of stochastic approximation algorithms of the form

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1})$$

where $(\theta_n)_n$ is a $\mathbb{R}^d$-valued sequence, $(\gamma_n)_n$ is a deterministic step-size sequence and $(X_n)_n$ is a controlled Markov chain. The originality of our framework is to address the convergence under weak assumptions on the smoothness-in-θ of the function $\theta \mapsto H(\theta, x)$. It is usually assumed that this function is continuous for any x; in this work, we propose a new condition which allows to consider fields H which are not continuous in θ. Our results are illustrated by considering stochastic approximation algorithms for quantile estimation and stochastic approximation algorithms for solving a vector quantization problem.

1 Introduction

Stochastic Approximation (SA) methods have been introduced by [34] as algorithms to find the roots of $h : \mathbb{R}^d \to \mathbb{R}^d$ when only noisy measurements of h are available. It is therefore among the class of stochastic (local) optimization algorithms which solve $\min_{\theta \in \mathcal{C}} L(\theta)$ where L is the objective function (also called *loss* function), θ is the adjustable parameter and $\mathcal{C}$ is the constraint set. In this optimization framework, h is the gradient of the function L or the gradient of the sum of L and a penalty function to take into account the constraint set.

SA algorithms are iterative algorithms of the form

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \eta_{n+1} \tag{1}$$

where θ_n is the estimation of the root of h at time n, $\{\gamma_n, n \in \mathbb{N}\}$ is a sequence of deterministic nonnegative stepsizes and $\{\eta_n, n \in \mathbb{N}\}$ is a random sequence of noisy measurements of h. In the seminal work of Robbins and Monro [34], η_n is a noisy observation of $h(\theta_n)$ while in the paper by Kiefer and Wolfowitz [20], η_n is an approximation of the gradient of L based on measurements of L.

[‡]corresponding author. mail: amandine.schreck@telecom-paristech.fr
[†]Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI
[§]CNRS LTCI ; Télécom ParisTech
[¶]University of Jyväskylä ; Department of Mathematics and Statistics

There is a rich convergence theory developed for many years about stochastic approximation algorithms. Some of them are about the long-time behavior: they prove the convergence of the sequence $\{\theta_n, n \in \mathbb{N}\}$ to the set $\{\theta : h(\theta) = 0\}$ and establish asymptotic normality of the normalized deviation $\gamma_n^{-1/2}(\theta_n - \theta_\star)$ along the event $\{\lim_n \theta_n = \theta_\star\}$. Robbins and Monro in [34] provide conditions implying the mean-square convergence of the sequence $\{\theta_n, n \in \mathbb{N}\}$, while the first conditions for almost-sure convergence were given by [7]. The first results on the asymptotic normality were obtained by [37, 14, 36]. These long-time behavior analyses are derived under conditions on the set $\{\theta : h(\theta) = 0\}$, on the sequence $\{\gamma_n, n \in \mathbb{N}\}$ and on the noise sequence $\{\eta_n, n \in \mathbb{N}\}$. The limiting set $\{\theta : h(\theta) = 0\}$ has to be attractive in some sense. The stepsize sequence has to vanish at a rate such that $\sum_n \gamma_n = \infty$ in order to prevent the algorithm to have premature and false convergence, but has to decrease rapidly enough in order to control the noise sequence $\{\eta_n, n \in \mathbb{N}\}$ and allow the convergence of the iterative scheme (1). About the noise sequence $\{\eta_n, n \in \mathbb{N}\}$, the first results were obtained in the case $\eta_{n+1} = H(\theta_n, X_{n+1})$ when $\{X_n, n \in \mathbb{N}\}$ are independent and identically distributed. They were then extended to the case the conditional distribution of X_{n+1} given the past history of the algorithm depends on θ_n or on both (θ_n, X_n). The former is sometimes called Robbins-Monro algorithm (see e.g. [6]) and the latter corresponds to a controlled Markovian (also called state-dependent) dynamic. The interested reader will find in [24, 6, 40, 5, 25, 41, 8] many results on stochastic approximation algorithms.

The goal of this paper is to provide almost-sure convergence results of the sequence $\{\theta_n, n \in \mathbb{N}\}$ in the case $\eta_{n+1} = H(\theta_n, X_{n+1})$ under weaker conditions on the regularity of the field H than what is usually assumed in the literature; and in the general case when $\{X_n, n \in \mathbb{N}\}$ is a controlled Markov chain.

When the elements of the sequence $\{X_n, n \in \mathbb{N}\}$ are independent, no regularity conditions are required on the field H. Indeed, in this case, convergence results are directly obtained from supermartingale arguments [6, Section 5]. When $\{X_n, n \in \mathbb{N}\}$ is a controlled Markovian sequence, Kushner [23] and Tadić [43, Theorem 4.1] provide convergence results on SA algorithms by introducing regularity assumptions on H: for any x, $\theta \mapsto H(\theta, x)$ is assumed to be Hölder-continuous (see [23, Eq. (4.2)] and [43, Eq. (4.1)]; see also [2, assumption (DRI2) and Proposition 6.1.]). Nevertheless in many practical cases, the field H may not be continuous with respect to θ: examples are given in [9, 10] for online learning, in [21] for finding the distribution with minimal quantile of a given order among a parametric family of distributions; see also Section 4 below.

Convergence results on stochastic approximation algorithm with discontinuous dynamic H are established under restrictive assumptions. For example, in [27], the sequence $\{X_n, n \in \mathbb{N}\}$ has to satisfy a strong law of large numbers: $\lim_n n^{-1} \sum_{k=1}^n H(\theta_\star, X_n) = h(\theta_\star)$ almost surely, at some rate, for any $\theta_\star \in \{h = 0\}$. Such a strong law of large numbers with a rate of convergence may reveal to be restrictive on $\{X_n, n \in \mathbb{N}\}$.

In this paper, we aim at proving the convergence of a stochastic approximation scheme for a discontinuous dynamic H when $\{X_n, n \in \mathbb{N}\}$ is a controlled Markov chain; note that this case covers the Robbins-Monro case. A preliminary step to the proof of convergence is the proof of stability [25]. Different strategies have been proposed in the literature in order to make SA algorithms stable (see for example [44, 25, 19]). Among them, Chen and Zhu force the sequence produced by their algorithm to remain in a randomly growing compact set [12]. Below, we will address the convergence of the self-stabilized stochastic approximation algorithm proposed in [2]: this algorithm introduces

truncations on random varying sets, as Chen and Zhou do but with Markovian dynamics, and in Theorem 2.1(i) sufficient conditions implying that the number of truncations is finite almost-surely are provided. Therefore, this stabilized algorithm follows the equation (1) after some random (but almost-surely finite) time. We then provide sufficient conditions for the almost-sure convergence of $\{\theta_n, n \in \mathbb{N}\}$ (see Theorem 2.1(ii)).

These results are motivated by the following applications. Firstly, quantile and multidimensional median approximation are considered. Such approximations can be used in adaptation processes for controlled Markov chains, as suggested in [3]. Quantile stochastic approximation is for example used with Markovian dynamics in [39] as an adaptation process. The second application considered in this work is vector quantization when solved by the 0 neighbors Kohonen algorithm. This algorithm is used for example in finance, in frameworks where the observation sequence $\{X_n, n \in \mathbb{N}\}$ is Markovian [33].

The paper is organized as follows: the stabilized stochastic approximation algorithm is described in Section 2.1; the assumptions and the convergence results are resp. in Sections 2.2 and 2.3. Section 4 is devoted to some applications. Finally, the proofs are postponed in Section 3.

2 Convergence results

2.1 A stabilized Stochastic Approximation algorithm

A popular method to solve the stability problem is to force the sequence $\{\theta_n, n \in \mathbb{N}\}$ to remain in a fixed *active* compact set $\mathcal{K}$. Nevertheless, this method is not satisfactory because a good choice for $\mathcal{K}$ requires some knowledge about the location of the roots of h. The advantage of the solution proposed by Chen and Zhu [12] (see also [11, 43]) is that it does not require to fix, prior the run of the algorithm, a compact set in which all the estimates will be restricted: the active compact set is allowed to randomly grow. The proof of convergence of this stabilized algorithm relies on the key fact that the number of updates of the active set will be almost surely finite so that almost-surely, the limiting points of this stabilized sequence are the limiting points of the original algorithm. The paper by Andrieu, Moulines and Priouret [2] introduces a slightly different version of the algorithm of [11] by allowing a control on the stepsizes after each update of the active set and by modifying the variation mechanism of the active set. A last strategy for stabilization is proposed by Andrieu and Vihola [4], in which Andradottir's expanding projection approach is studied in the Markovian case. Hereafter, we will consider the stochastic approximation algorithm as proposed in [2]. We now fix some notations and describe this stabilized algorithm.

Let Θ be an open subset of $\mathbb{R}^d$ equipped with its Borel σ-field $\mathcal{B}(\Theta)$; $\mathbf{X}$ be a general space with a countably generated σ-field $\mathcal{X}$; and let $H : \Theta \times \mathbf{X} \to \mathbb{R}^d$ be a measurable function. Let $\{P_\theta, \theta \in \Theta\}$ be a family of transition kernels on $(\mathbf{X}, \mathcal{X})$. For any non negative number γ, define a transition kernel Q_γ on $(\mathbf{X} \times \Theta, \mathcal{X} \times \mathcal{B}(\Theta))$ by

$$Q_\gamma f(x, \theta) = \int P_\theta(x, dy)\, f(y, \theta + \gamma H(\theta, y)), \qquad x \in \mathbf{X}, \theta \in \Theta\,, \tag{2}$$

for any measurable function f such that the integral is well defined. Before running the stabilized algorithm, the user has to choose a sequence of increasing compact sets: let $\{\mathcal{K}_q, q \geq 0\}$ be a sequence

of compact subsets of Θ such that

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta, \quad \text{and} \quad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}), \quad q \geq 0, \tag{3}$$

where $\text{int}(A)$ denotes the interior of the set A. Roughly speaking, the algorithm runs as follows: the algorithm is initialized at (X_0, θ_0) and the active set is set to $\mathcal{K}_0$. If, at iteration n, θ_n is in the current active set then conditionally to the past, (X_{n+1}, θ_{n+1}) is sampled from $Q_{\gamma_{\varsigma_n}}(X_n, \theta_n; \cdot, \cdot)$ where $\{\varsigma_n, n \in \mathbb{N}\}$ is some (possibly random) counter. If θ_n is not in the active set, then *(i)* $(X_{n+1}, \theta_{n+1}) \sim Q_{\gamma_{\varsigma_n}}(\Phi(X_n, \theta_n); \cdot, \cdot)$ where $\Phi : \mathbf{X} \times \Theta \to \mathbf{X} \times \Theta$ is a measurable function which is usually chosen as a projection function on a bounded set of $\mathbf{X} \times \Theta$; and *(ii)* the active set is modified.

In order to give a Markovian dynamic to the above algorithm, integer-valued random variables $\kappa_n, \nu_n, \varsigma_n$ have to be introduced. κ_n is the index of the active set at the end of iteration n. ν_n is the time spent from the last update of the active set: $\nu_n = 0$ iff $\kappa_n \neq \kappa_{n-1}$. Finally, the method also allows a modification of the stepsize sequence every time the active set is modified: at time n, the step size is γ_{ς_n} where ς_n is defined by

$$\varsigma_{n+1} = \begin{cases} \varsigma_n + \phi(\nu_n) & \text{if the active compact set is modified at iteration } n \\ \varsigma_n + 1 & \text{otherwise;} \end{cases}$$

and $\phi : \mathbb{Z}^+ \to \mathbb{Z}$ is a measurable function such that $\phi(k) > -k$ for any k. As discussed in [2], different choices can be made for the function ϕ. For example, $\phi(k) = 1$ for all k in $\mathbb{N}$ implies that the sequence $\{\varsigma_n, n \in \mathbb{N}\}$ is deterministic and $\varsigma_n = n$. Another example consists in choosing $\phi(k) = 1 - k$: the number of iterations between two successive updates of the active set is not taken into account.

The above algorithm can be summarized as follows: it is initialised at $(X_0, \theta_0) \in \mathbf{X} \times \Theta$ and $\nu_0 = \kappa_0 = 0$ and $\varsigma_0 = 1$; a transition $(X_n, \theta_n, \kappa_n, \nu_n, \varsigma_n) \to (X_{n+1}, \theta_{n+1}, \kappa_{n+1}, \nu_{n+1}, \varsigma_{n+1})$ is described by Algorithm 1.

Algorithm 1

- sample

$$(X_{n+1}, \theta_{n+1}) \sim \begin{cases} Q_{\gamma_{\varsigma_n}}(\Phi(X_n, \theta_n); \cdot) & \text{if } \nu_n = 0 \\ Q_{\gamma_{\varsigma_n}}(X_n, \theta_n; \cdot) & \text{otherwise ,} \end{cases} \tag{4}$$

- set

$$(\kappa_{n+1}, \nu_{n+1}, \varsigma_{n+1}) = \begin{cases} (\kappa_n, \nu_n + 1, \varsigma_n + 1) & \text{if } \theta_{n+1} \in \mathcal{K}_{\kappa_n} \\ (\kappa_n + 1, 0, \varsigma_n + \phi(\nu_n)) & \text{otherwise .} \end{cases} \tag{5}$$

Note from (2) that, except when the active set is modified at time n, the conditional distribution of X_{n+1} given the past is a Markov transition kernel controlled by the current value of the parameter θ_n. The above framework is quite general: it covers the case when $\{X_n, n \in \mathbb{N}\}$ is a (non-controlled) Markov chain by choosing $P_\theta = P$ for any θ, the Robbins-Monro case by choosing $P_\theta(x, \cdot) = \pi_\theta$ for any x where $\{\pi_\theta, \theta \in \Theta\}$ is a class of distributions on $\mathbf{X}$, and the case when $\{X_n, n \in \mathbb{N}\}$ is an i.i.d. sequence with distribution π by choosing $P_\theta(x, dy) = \pi(dy)$ for any x, θ.

2.2 Assumptions

In the rest of the paper, for any $d \in \mathbb{N}$, let $\left\langle \cdot, \cdot \right\rangle$ denote the usual scalar product in $\mathbb{R}^d$ and $\| \cdot \|$ denote the associated norm.

We now introduce sufficient conditions for the stability and the convergence of the sequence $\{\theta_n, n \in \mathbb{N}\}$ described by Algorithm 1. By construction, a path of the sequence $\{\theta_n, n \in \mathbb{N}\}$ can be decomposed into blocks of random length such that during the block, the sequence is in a compact set. Therefore, as shown in [2], the proof essentially consists in controlling the increment $\theta_{n+1} - \theta_n$ along the event that $\{\theta_n, n \in \mathbb{N}\}$ remains in a compact. To that goal, for any $\mathcal{K} \subset \Theta$, define the stopping-time

$$\sigma(\mathcal{K}) = \inf\{n \geq 1, \theta_n \notin \mathcal{K}\} \; , \tag{6}$$

with convention $\inf \emptyset = +\infty$. For any function $W : \mathbf{X} \to [1, \infty)$, define the W-norm of a measurable function $f : \mathbf{X} \to \mathbb{R}$ and the W-norm of a signed measure μ on $(\mathbf{X}, \mathcal{X})$ by

$$|f|_W = \sup_{\mathbf{X}} \frac{|f|}{W} \; , \qquad \|\mu\|_W = \sup_{f, |f|_W \leq 1} |\mu(f)| \; .$$

Finally, for a sequence $\boldsymbol{\gamma} = \{\gamma_n, n \in \mathbb{N}\}$, denote by $\mathbb{P}^{\boldsymbol{\gamma}}_{x,\theta}$ (resp. $\mathbb{E}^{\boldsymbol{\gamma}}_{x,\theta}$) the probability (resp. the expectation) associated with the non-homogeneous Markov chain with $\delta_{(x,\theta)}$ as initial distribution and $Q_{\gamma_0}, Q_{\gamma_1}, \cdots$ as transition kernels.

It is assumed that the functions Φ and H satisfy A1; assumptions on the ergodic behavior of the transition kernels P_θ are given by A2. A4 introduce the conditions on the regularity in θ of the dynamic H and the mean field function. Finally, conditions on the stepsize sequence $\{\gamma_n, n \in \mathbb{N}\}$ are given in A5.

A1 (a) $\Phi : \mathbf{X} \times \Theta \to \mathbf{K} \times \mathcal{K}_0$ where $\mathbf{K} \in \mathcal{X}$.

(b) $H : \Theta \times \mathbf{X} \to \Theta$ is measurable and there exists a measurable function $W : \mathbf{X} \to [1, \infty)$ such that for any compact set $\mathcal{K} \subset \Theta$, $\sup_{\theta \in \mathcal{K}} |H(\theta, .)|_W < \infty$.

(c) $\sup_{\mathbf{K}} W < \infty$.

A2 The kernels $\{P_\theta, \theta \in \Theta\}$ satisfy the following conditions:

(a) For any θ in Θ, the kernel P_θ has a unique stationary distribution π_θ.

(b) For any compact $\mathcal{K} \subseteq \Theta$, there exist positive constants $C < \infty$ and $\lambda < 1$ such that for any $x \in \mathbf{X}$, $l \geq 0$,

$$\sup_{\theta \in \mathcal{K}} \|P^l_\theta(x, .) - \pi_\theta\|_W \leq C \lambda^l W(x) \; , \qquad \sup_{\theta \in \mathcal{K}} \pi_\theta(W) < \infty \; .$$

(c) There exists $p > 1$ and for any compact set $\mathcal{K} \subseteq \Theta$, there exists a constant C such that for $q \geq 0$ and any $x \in \mathbb{X}$,

$$\sup_{\theta \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[W^p(X_k) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k\}} \right] \leq C W^p(x) \; ,$$

where $\boldsymbol{\gamma}^{\leftarrow q} = \{\gamma_{q+n}, n \in \mathbb{N}\}$.

When $P_\theta = P$ for any θ, sufficient conditions for A2 are given in [30, Chapters 10 and 15]: they are mainly implied by a drift condition of the form

$$PW^p(x) \leq \lambda W^p(x) + b\mathbf{1}_{\mathcal{C}}(x) \ ,$$

where $0 < \lambda < 1$, $b > 0$ and $\mathcal{C}$ is small [30, Chapter 5]. When P_θ depends upon θ, A2(b-c) results in an homogeneous behavior of P_θ for θ being in a compact set. Sufficient conditions in terms of drift inequality and minorization conditions implying A2(b-c) can be found in [15, Lemma 2.3]. For example, a family of adaptive Metropolis-Hastings kernels with the same invariant distribution π satisfies A2 provided π is sub-exponential (see e.g. [38, Proposition 15] or [15, Example 2]). Note that in general, it is really unlikely that conditions of the form A2(b-c) when the supremum in θ is for $\theta \in \Theta$ hold. Nevertheless, the stabilization procedure only requires to control the chain when $\{\theta_n, n \in \mathbb{N}\}$ remains in a compact, thus yielding to a supremum over $\mathcal{K}$, for any compact set $\mathcal{K}$.

We now introduce the regularity conditions on H and assume that there exists a global Lyapunov function for the mean-field h defined by

$$h(\theta) = \int \pi_\theta(dx)\ H(\theta, x) \ . \tag{7}$$

A3 There exists $\alpha \in (0, 1]$ and for any compact set $\mathcal{K} \subseteq \Theta$, there exists a constant $C > 0$ such that for all $\delta > 0$,

$$\sup_{\theta \in \mathcal{K}} \int \pi_\theta(dx) \sup_{\{\theta', \|\theta' - \theta\| \leq \delta\}} \left\| H(\theta', x) - H(\theta, x) \right\| \leq C\delta^\alpha \ .$$

A3 is used to obtain some regularity for the solution to the Poisson equation associated to the field H (see details below). This solution to the Poisson equation appears when decomposing the sum of the noises obtained at each iteration in a martingale term and a remaining term. As said previously, when the sequence $\{X_n, n \in \mathbb{N}\}$ is independent, convergence results are directly obtained with martingale arguments (see [6, Section 5] for details), and there is no need to assume A3.

This assumption does not imply that $\theta \mapsto H(\theta, x)$ is continuous for any x, which is the usual framework when proving the convergence of SA algorithms [2, Section 6]; the classical assumption is of the form $\sup_{\theta_1, \theta_2 \in \mathcal{K}} \|\theta_1 - \theta_2\|^{-\beta} \, |H(\theta_1, \cdot) - H(\theta_2, \cdot)|_W < \infty$ for some $\beta \in (0, 1]$ and $\sup_{\theta \in \mathcal{K}} \pi_\theta(W) < \infty$, which implies the condition A3 with $\alpha = \beta$ so that our framework covers this usual case.

A4 h is continuous on Θ and there exists a continuously differentiable function $w : \Theta \to [0, \infty)$ such that

(a) There exists $M_0 > \sup_{\mathcal{K}_0} w$ such that

$$\mathcal{L} := \left\{ \theta \in \Theta, \left\langle \nabla w(\theta), h(\theta) \right\rangle = 0 \right\} \subset \{\theta \in \Theta, w(\theta) < M_0\} \ . \tag{8}$$

(b) There exists $M_1 \in (M_0, \infty]$ such that $\{\theta \in \Theta, w(\theta) \leq M_1\}$ is a compact set.

(c) For any $\theta \in \Theta \setminus \mathcal{L}$, $\left\langle \nabla w(\theta), h(\theta) \right\rangle < 0$.

(d) $w(\mathcal{L})$ has an empty interior.

Lemma 3.7 shows that under assumptions A2 and A3, h is continuous as soon as $\lim_{\theta\to\theta'} D_W(\theta,\theta') = 0$ where D_W is some measure of the difference between the kernels P_θ and $P_{\theta'}$ and is defined by

$$D_W(\theta,\theta') = \sup_{x\in\mathbf{X}} \frac{\|P_\theta(x,.) - P_{\theta'}(x,.)\|_W}{W(x)} \; . \tag{9}$$

A4 is a classical assumption in stochastic approximation theory (see for example [6, Part II, Section 1.6], or [8, Section 3.3]). It is known as the Robbins-Siegmund assumption [27] in reference to the Robbins-Siegmund Lemma [35]. We finally conclude this set of assumptions by conditions on the stepsize $\boldsymbol{\gamma} = \{\gamma_n, n \in \mathbb{N}\}$. To make these conditions readable, we assume that this sequence is polynomially decreasing. The proofs are nevertheless written with a generic stepsize sequence and A6 in Section 3 states the conditions on $\{\gamma_n, n \in \mathbb{N}\}$ in the general case.

A5 $\boldsymbol{\gamma} = \{\gamma_0/(n+1)^\beta, n \geq 0\}$ with β satisfying:

(a) $\beta \in (\max\left(\frac{1}{2}, \frac{1+\alpha/p}{1+\alpha}\right); 1]$, where p and α are respectively defined in A2(c) and A3.

(b) For any compact set $\mathcal{K} \subset \Theta$ and any $C > 0$, there exists $r \in (\frac{1}{\beta\alpha} - \frac{1}{\alpha}; 1 - \frac{1}{\beta p})$ such that for any $\Gamma > 0$,

$$\lim_{q\to\infty} \sum_{k:k-\lceil C\log(k+q)\rceil \geq 0} \frac{\log^2(k+q)}{(k+q)^\beta} \sum_{j=k-\lceil C\log(k+q)\rceil+1}^{k} \mathcal{D}_j(q) = 0 \; ,$$

where

$$\mathcal{D}_j(q) \quad = \quad \sup_{(x,\theta)\in\mathbf{K}\times\mathcal{K}_0} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[D_W(\theta_j,\theta_{j-1})^{p/(p-1)} \mathbf{1}_{\{\sigma(\mathcal{K})\geq j\}} \mathbf{1}_{\|\theta_j-\theta_{j-1}\|\leq\Gamma(j+q)^{-\beta r}} \right]^{(p-1)/p} \quad ,$$

$\lceil\cdot\rceil$ denotes the upper integer part and $\mathbf{K}, p$ are respectively given by A1(a) and A2(c).

In the simple case when for any $\theta \in \Theta$, $P_\theta = P$, $D_W(\theta,\theta') = 0$ for any θ, θ', and A5(b) is trivial.

When for any compact subset $\mathcal{K} \subseteq \Theta$, there exists a constant C such that $\sup_{\theta,\theta'\in\mathcal{K}} D_W(\theta,\theta') \leq C\|\theta-\theta'\|$ (this is the case for some Adaptive Metropolis kernels P_θ, see [1, Lemma 13]), then A5(b) holds if $\beta(1+r) > 1$; since $\alpha \in (0,1]$ (see A3), the condition $r > \alpha^{-1}(1/\beta - 1)$ implies $\beta(1+r) > 1$ and A5(b) holds.

2.3 Main result

Algorithm 1 defines a $\Theta \times \mathbf{X} \times \mathbb{N}^3$-valued homogeneous Markov chain $\mathbf{Z} = \{(X_n, \theta_n, \kappa_n, \nu_n, \varsigma_n),\ n \in \mathbb{N}\}$. We denote by $\overline{\mathbb{P}}_{x,\theta}$ (resp. $\overline{\mathbb{E}}_{x,\theta}$) the canonical probability (resp. the canonical expectation) associated to $\mathbf{Z}$, with initial distribution $\delta_{(x,\theta,0,0,1)}$.

The following theorem shows that the number of updates of the active set is finite almost-surely: this implies that there exists a random time N, finite almost-surely and such that for any $n > N$,

$$\theta_{n+1} = \theta_n + \gamma_{\zeta_N+n-N} H(\theta_n, X_{n+1}) \; .$$

The second statement establishes the convergence of this stabilized sequence to the set $\mathcal{L}$ defined by (8). The proof of Theorem 2.1 is given in Section 3.

Theorem 2.1. *Assume A1 to A5.*

(i) With probability one, the number of updates of the active set is finite: for any $x \in \mathbf{K}$ and $\theta \in \mathcal{K}_0$,

$$\overline{\mathbb{P}}_{x,\theta}\left(\sup_{k\geq 1} \kappa_k < \infty\right) = 1 \ .$$

(ii) With probability one, the sequence $\{\theta_n, n \in \mathbb{N}\}$ converges to the set $\mathcal{L}$ given by (8): for any $x \in \mathbf{K}$ and $\theta \in \mathcal{K}_0$,

$$\overline{\mathbb{P}}_{x,\theta}\left(\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\right) = 1 \ .$$

3 Proofs

Define the translated sequence $\gamma^{\leftarrow q} = \{\gamma_{q+n}, n \geq 0\}$ and the level set $\mathcal{W}_M = \{\theta \in \Theta, w(\theta) \leq M\}$. All the discussions below are written with a generic sequence γ_n, in order to outline the extension of our work to the case when γ_n is not polynomially decreasing. We prove Theorem 2.1 by replacing A5 with

A6 The sequence $\boldsymbol{\gamma} = \{\gamma_n, n \in \mathbb{N}\}$ is a non-increasing positive sequence such that:

(a) $\sum \gamma_k = \infty$.

(b) $\sum \left(\gamma_k^p + \gamma_k^2\right) < \infty$ where p is given by A2(c).

(c) There exists a constant $r \in (0,1)$ satisfying

(i) $\sum_k \gamma_k^{p(1-r)} < \infty$, with p defined in A2(c).

(ii) For any constant $C > 0$, $\sum_k \gamma_{1\vee(k-\lceil C|\log\gamma_k|\rceil)}^{1+r\alpha} |\log(\gamma_k)|^{1+\alpha} < \infty$, with α defined in A3.

(iii) For any constant $C > 0$, $\lim_{q\to\infty} \sup_k \psi_q(k)\gamma_{q+k-\psi_q(k)}^r = 0$ where $\psi_q(k) = (k-1) \wedge \lceil C|\log(\gamma_{k+q})|\rceil$.

(iv) For any compact subset $\mathcal{K}$ of Θ and any positive constants C, Γ,

$$\lim_{q\to\infty} \sum_k \gamma_{k+q+1}\psi_q^2(k) \sum_{j=k-\psi_q(k)+1}^{k} \sup_{(x,\theta)\in\mathbf{K}\times\mathcal{K}_0} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[D_W(\theta_j,\theta_{j-1})^{p/(p-1)}\mathbf{1}_{\{\sigma(\mathcal{K})\geq j\}}\mathbf{1}_{\|\theta_j-\theta_{j-1}\|\leq\Gamma\gamma_{j+q}^r}\right]^{(p-1)/p} = 0,$$

where $\mathbf{K}, p$ are respectively given by A1(a) and A2(c).

Note that these conditions are verified with $\gamma_n \propto n^{-\beta}$ (for all large n) with β, r satisfying A5.

3.1 Proof of Theorem 2.1.(i)

If an update of the active set occurs at time q, then until the next update of the active set, the update of $\{\theta_n, n > q\}$ is given by:

$$\theta_{n+1} = \theta_n + \gamma_{\varsigma_n} h(\theta_n) + \gamma_{\varsigma_n} \left(H(\theta_n, X_{n+1}) - h(\theta_n) \right) \ .$$

As shown in [2, Theorems 2.2. and 2.3.], it is important to control the noise between two successive updates of the active set. Note that the update of the active set mechanism described in Algorithm 1 differs from the mechanism in [2]: due to their assumptions on the m-iterated transition kernels for some $m \geq 1$ (see [2, Assumptions DRI]), the distance between two successive values of the parameters have to be controlled. To that goal, they introduce a second update of the active set every time $\|\theta_{n+1} - \theta_n\|$ is larger than a time-dependent threshold. In this paper, our assumptions on the transition kernels are in terms of geometric ergodicity (see A2) and therefore, this supplementary update of the active set is relaxed.

For a stepsize sequence $\boldsymbol{\rho} = \{\rho_n, n \in \mathbb{N}\}$, a compact subset $\mathcal{K}$ of Θ, and $l, n \geq 0$, set

$$S_{l,n}(\boldsymbol{\rho}, \mathcal{K}) = \mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} \sum_{k=l}^{n} \rho_k \left(H(\theta_{k-1}, X_k) - h(\theta_{k-1}) \right) \ ,$$

where $\sigma(\mathcal{K})$, defined by (6) denotes the first exit time from the set $\mathcal{K}$. Following the same approach as in [2, Sections 4 and 5], the proof of Theorem 2.1(i) is in two steps. The first step consists in showing that the quantity $\overline{\mathbb{P}}_{x,\theta}\left(\sup_{k\geq 1} \kappa_k \geq m\right)$ decreases at a geometric rate. This rate is an upper bound of the sum of the errors $\gamma_{\varsigma_n} \left(H(\theta_n, X_{n+1}) - h(\theta_n) \right)$ between two updates of the active set.

Proposition 3.1. *Assume A1*(a), *A1*(c) *and A4. For any $M \in (M_0, M_1]$, there exist $\delta_M > 0$ and $q_M \in \mathbb{N}$ such that for any $m \geq q_\star \geq q_M$,*

$$\sup_{x\in\mathbf{K}} \sup_{\theta\in\mathcal{K}_0} \overline{\mathbb{P}}_{x,\theta}\left(\sup_{k\geq 1} \kappa_k \geq m \right) \leq \left(\sup_{q\geq q_\star} \sup_{x\in\mathbf{K}} \sup_{\theta\in\mathcal{K}_0} \mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_{k\geq 1} |S_{1,k}(\boldsymbol{\gamma}^{\leftarrow q}, \mathcal{W}_M)| \geq \delta_M \right) \right)^m \ .$$

Proposition 3.1 is a slight adaptation of [2, Corollary 4.3] and the proof is omitted.

The second step of the proof consists in showing that there exist $M \in (M_0, M_1)$ and $q_\star \geq q_M$ large enough so that

$$\sup_{q\geq q_\star} \sup_{x\in\mathbf{K}} \sup_{\theta\in\mathcal{K}_0} \mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_{k\geq 1} |S_{1,k}(\boldsymbol{\gamma}^{\leftarrow q}, \mathcal{W}_M)| \geq \delta_M \right) < 1 \ . \tag{10}$$

To that goal, we decompose $S_{1,n}$ into a martingale and remainder terms; set

$$g_\theta = \sum_{l\geq 0} \left[P_\theta^l \left(H(\theta, .) \right) - \pi_\theta \left(H(\theta, .) \right) \right] \ . \tag{11}$$

g_θ solves the Poisson equation $g - P_\theta g = H(\theta, .) - \pi_\theta(H(\theta, .))$. Under A2(b), such a solution exists and for any compact set $\mathcal{K}$ it holds (see e.g.[30, Chapter 17.4]) $\sup_{\theta\in\mathcal{K}} \left(|g_\theta|_W + |P_\theta g_\theta|_W \right) < \infty$. This allows to write for any compact set $\mathcal{K} \subset \Theta$ and any sequence $\{\rho_n, n \in \mathbb{N}\}$

$$S_{1,n}(\rho, \mathcal{K}) = \mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} \sum_{k=1}^{n} \rho_k \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \right) = \mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} \sum_{i=1}^{5} T_{i,n} \ , \tag{12}$$

with

$$
\begin{aligned}
T_{1,n}(\mathcal{K}) &= \sum_{k=1}^{n} \rho_k \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) \right) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k\}} \ , \\
T_{2,n}(\mathcal{K}) &= \sum_{k=1}^{n-1} (\rho_{k+1} - \rho_k) P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \ , \\
T_{3,n}(\mathcal{K}) &= \rho_1 P_{\theta_0} g_{\theta_0}(X_0) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq 1\}} - \rho_n P_{\theta_{n-1}} g_{\theta_{n-1}}(X_n) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} \ , \\
T_{4,n}(\mathcal{K}) &= \sum_{k=1}^{n-1} \rho_{k+1} \left(P_{\theta_k} g_{\theta_k}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \right) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \ , \\
T_{5,n}(\mathcal{K}) &= - \sum_{k=1}^{n-1} \rho_k P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \mathbf{1}_{\{\sigma(\mathcal{K}) = k\}} \ .
\end{aligned}
$$

Note that $\mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} T_{5,n}(\mathcal{K}) = 0$. Then, by Markov's inequality, for any $p_i \geq 1$, we write

$$
\begin{aligned}
\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left(\sup_{n \geq 1} |S_{1,n}(\boldsymbol{\gamma}^{\leftarrow q}, \mathcal{K})| \geq \delta \right) \leq \sum_{i=1}^{3} \left(\frac{4}{\delta} \right)^{p_i} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 1} \mathbf{1}_{\{\sigma(\mathcal{K}) \geq n\}} |T_{i,n}(\mathcal{K})|^{p_i} \right] \\
+ \mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left(\sup_{n \geq 1} |T_{4,n}(\mathcal{K})| \geq \delta/4 \right) \ .
\end{aligned}
$$

Proposition 3.2 below controls the moments of $T_{1,n}$, $T_{2,n}$ and $T_{3,n}$. Since $\mathcal{K}_0 \subseteq \mathcal{W}_{M_0}$, $\sup_{\mathbf{K}} W < \infty$ and $\lim_q \sum_k \gamma_{k+q}^2 = \lim_q \sum_k \gamma_{k+q}^p = 0$ (see assumptions A1(c), A4(a) and A6(b)), Proposition 3.2 implies that for any $M \in (M_0, M_1]$,

$$
\lim_{q \to \infty} \sum_{i=1}^{3} \left(\frac{4}{\delta_M} \right)^{p_i} \sup_{(x,\theta) \in \mathbf{K} \times \mathcal{K}_0} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 1} \mathbf{1}_{\{\sigma(\mathcal{W}_M) \geq n\}} |T_{i,n}(\mathcal{W}_M)|^{p_i} \right] = 0 \ ,
$$

with $(p_1, p_2, p_3) = (1, 1, p)$. The originality of our work is in the control of the last term $T_{4,n}$. Note that under our assumptions on H, the condition (A3) of [2] may not hold so that the computations in the proof of [2, Proposition 5.2] can not be used for $T_{4,n}$. Choose r satisfying A6(c) and set

$$
A_{\Gamma}^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, j) = \left\{ \sup_{1 \leq k \leq j} \frac{\|\theta_k - \theta_{k-1}\|}{\gamma_{k+q}^r} \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k\}} \leq \Gamma \right\} \ . \tag{13}
$$

We write for any $\Gamma > 0$,

$$
\begin{aligned}
&\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left(\sup_{k \geq 1} |T_{4,k}(\mathcal{K})| \geq \delta/4 \right) \\
&\leq \frac{4}{\delta} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{k \geq 1} |T_{4,k}(\mathcal{K})| \mathbf{1}_{\cap_j A_{\Gamma}^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, j)} \right] + \mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left(\sup_k \frac{\|\theta_k - \theta_{k-1}\|}{\gamma_{k+q}^r} \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k\}} > \Gamma \right) \ .
\end{aligned}
$$

Lemma 3.3 combined with the assumption A1(c) shows that for any $\epsilon > 0$ and any $M \in (M_0, M_1]$, there exists $\Gamma > 0$ such that

$$\sup_{q\geq 0}\sup_{x\in\mathbf{K}}\sup_{\theta\in\mathcal{K}_0}\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_k \frac{\|\theta_k-\theta_{k-1}\|}{\gamma_{k+q}^r}\mathbf{1}_{\{\sigma(\mathcal{K})\geq k\}} > \Gamma\right) \leq \epsilon\ .$$

Proposition 3.4 and the assumption A1(c) imply that for any $\epsilon > 0$, any $M \in (M_0, M_1)$ and any $\Gamma > 0$, there exists $q_\star$ such that

$$\sup_{q\geq q_\star}\sup_{x\in\mathbf{K}}\sup_{\theta\in\mathcal{K}_0}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_{k\geq 1}|T_{4,k}(\mathcal{W}_M)|\mathbf{1}_{\cap_j A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{W}_M,j)}\right] \leq \epsilon\ .$$

This will conclude the proof of Theorem 2.1(i).

Proposition 3.2. *Assume A1*(b) *and A2. For any compact subset $\mathcal{K} \subset \Theta$, there exists a constant C such that for any $q \geq 0$ and any $x \in \mathbf{X}$,*

$$\begin{aligned}
\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_{n\geq 0}|T_{1,n}(\mathcal{K})|\right] &\leq C\left(\sum_{k=0}^{\infty}\gamma_{k+q}^2\right)^{1/2} W(x)\ ,\\
\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_{n\geq 0}|T_{2,n}(\mathcal{K})|\right] &\leq C\gamma_q W(x)\ ,\\
\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_{n\geq 0}|T_{3,n}(\mathcal{K})|^p\right] &\leq C\left(\sum_{k=0}^{\infty}\gamma_{k+q}^p\right) W^p(x)\ ,
\end{aligned}$$

where p is given in A2(c).

The proof is on the same lines as the computations in [2, Appendix A] and is omitted.

Lemma 3.3. *Assume A1(b-c), A2(c) and let $r \in (0,1)$ satisfying A6(ci). For any compact set $\mathcal{K} \subset \Theta$ and any $\epsilon > 0$, there exists $\Gamma > 0$ such that*

$$\sup_{q\geq 0}\sup_{x\in\mathbf{K}}\sup_{\theta\in\mathcal{K}_0}\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_n \frac{\|\theta_n-\theta_{n-1}\|}{\gamma_{n+q}^r}\mathbf{1}_{\{\sigma(\mathcal{K})\geq n\}} > \Gamma\right) \leq \epsilon\ .$$

Proof. Fix a compact subset $\mathcal{K} \subset \Theta$. Under $\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}$ and on the set $\{\sigma(\mathcal{K}) \geq n\}$, $\theta_n - \theta_{n-1} = \gamma_{n+q}H(\theta_{n-1}, X_n)$ for any $n \geq 0$. Then, by A1(b), there exists a constant C such that, on the set $\{\sigma(\mathcal{K}) \geq n\}$, $\|\theta_n - \theta_{n-1}\| \leq C\gamma_{n+q}W(X_n)$. This yields for any $\Gamma > 0$,

$$\begin{aligned}
\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_n \frac{\|\theta_n-\theta_{n-1}\|}{\gamma_{n+q}^r}\mathbf{1}_{\{\sigma(\mathcal{K})\geq n\}} > \Gamma\right) &\leq \frac{C^p}{\Gamma^p}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_n \gamma_{n+q}^{p(1-r)}W^p(X_n)\mathbf{1}_{\{\sigma(\mathcal{K})\geq n\}}\right]\\
&\leq \frac{C^p}{\Gamma^p}\sum_n \gamma_{n+q}^{p(1-r)}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[W^p(X_n)\mathbf{1}_{\{\sigma(\mathcal{K})\geq n\}}\right]\ .
\end{aligned}$$

By A1(c) and A2(c), there exists a constant C' such that

$$\sup_{q\geq 0}\sup_{x\in\mathbf{K}}\sup_{\theta\in\mathcal{K}_0}\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left(\sup_n \frac{\|\theta_n-\theta_{n-1}\|}{\gamma_{n+q}^r}\mathbf{1}_{\{\sigma(\mathcal{K})\geq n\}}>\Gamma\right)\leq\frac{C'}{\Gamma^p}\sum_n\gamma_n^{p(1-r)}\ ,$$

which concludes the proof by A6(ci). □

Proposition 3.4. *Assume A1 to A4 and A6(b-c). Let $M\in(M_0,M_1)$ and $\Gamma>0$. There exists $q_\star$ such that for any $q\geq q_\star$ and any $x\in\mathbf{X}$,*

$$\sup_{\theta\in\mathcal{W}_M}\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}\left[\sup_{n\geq 0}|T_{4,n}(\mathcal{W}_M)|\mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{W}_M,n)}\right]\leq\left(\sum_{k=1}^\infty C_k(\boldsymbol{\gamma}^{\leftarrow q})\right)W(x)\ , \tag{14}$$

where the $C_k(\boldsymbol{\gamma}^{\leftarrow q})$ are finite constants depending on $\boldsymbol{\gamma}^{\leftarrow q}$ and such that $\lim_{q\to\infty}\sum_{k\geq 1}C_k(\boldsymbol{\gamma}^{\leftarrow q})=0$.

Proof. Let $\Gamma>0$ and $M\in(M_0,M_1)$ be fixed. Set $\mathcal{K}=\mathcal{W}_M$ (note that by A4(a), $\mathcal{K}_0\subset\mathcal{K}$) and let $\lambda\in(0,1)$ be the ergodic rate given by A2(b) when applied with the compact $\mathcal{K}$. Set $\psi_q(k)=(k-1)\wedge\left\lceil\frac{|\log(\gamma_{k+q})|}{|\log(\lambda)|}\right\rceil$, where $\lceil\cdot\rceil$ denotes the upper integer part. Fix $M'\in(M,M_1)$ and set $\mathcal{K}'=\mathcal{W}_{M'}$.

By A6(ciii), there exists $q_\star$ such that for any $q\geq q_\star$,

$$\{\theta\in\Theta,\mathrm{d}(\theta,\mathcal{K})\leq\Gamma\ \sup_k\psi_q(k)\gamma_{q+k-\psi_q(k)}^r\}\subseteq\mathcal{K}'\ . \tag{15}$$

Hereafter, $q\geq q_\star$. By definition of g_θ (see (11)) and $h(\theta)$ (see (7)),

$$\begin{aligned}P_{\theta_k}g_{\theta_k}-P_{\theta_{k-1}}g_{\theta_{k-1}}=&\sum_{l>\psi_q(k)}\left[P_{\theta_k}^l\left(H(\theta_k,.)\right)-h(\theta_k)\right]-\sum_{l>\psi_q(k)}\left[P_{\theta_{k-1}}^l\left(H(\theta_{k-1},.)\right)-h(\theta_{k-1})\right]\\&+\psi_q(k)\left[h(\theta_k)-h(\theta_{k-1})\right]+\sum_{l=1}^{\psi_q(k)}P_{\theta_k}^l(H(\theta_k,.))-P_{\theta_{k-1}}^l(H(\theta_{k-1},.))\ .\end{aligned}$$

This implies $T_{4,n}(\mathcal{K})=\sum_{i=1}^4T_{4,n}^{(i)}$, with

$$\begin{aligned}T_{4,n}^{(1)}&=\sum_{k=1}^{n-1}\gamma_{q+k+1}\sum_{l>\psi_q(k)}\left[P_{\theta_k}^l\left(H(\theta_k,.)\right)(X_k)-h(\theta_k)\right]\mathbf{1}_{\{\sigma(\mathcal{K})\geq k+1\}}\ ,\\T_{4,n}^{(2)}&=-\sum_{k=1}^{n-1}\gamma_{q+k+1}\sum_{l>\psi_q(k)}\left[P_{\theta_{k-1}}^l\left(H(\theta_{k-1},.)\right)(X_k)-h(\theta_{k-1})\right]\mathbf{1}_{\{\sigma(\mathcal{K})\geq k+1\}}\ ,\\T_{4,n}^{(3)}&=\sum_{k=1}^{n-1}\gamma_{q+k+1}\psi_q(k)\left[h(\theta_k)-h(\theta_{k-1})\right]\mathbf{1}_{\{\sigma(\mathcal{K})\geq k+1\}}\ ,\\T_{4,n}^{(4)}&=\sum_{k=1}^{n-1}\gamma_{q+k+1}\sum_{l=1}^{\psi_q(k)}\left[P_{\theta_k}^l(H(\theta_k,.))(X_k)-P_{\theta_{k-1}}^l(H(\theta_{k-1},.))(X_k)\right]\mathbf{1}_{\{\sigma(\mathcal{K})\geq k+1\}}\ .\end{aligned}$$

Control of $T_{4,n}^{(1)}$ and $T_{4,n}^{(2)}$

By A1(b) and A2(b), there exists $C>0$ such that for any $q \geq 0$,

$$\left|T_{4,n}^{(1)}\right| \leq \frac{C}{1-\lambda} \sum_{k=1}^{n-1} \gamma_{q+k+1} \lambda^{\psi_q(k)} W(X_k) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \, .$$

Hence,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 0} \left|T_{4,n}^{(1)}\right| \right] \leq \frac{C}{1-\lambda} \left(\sum_{k=1}^{\infty} \gamma_{q+k+1} \lambda^{\psi_q(k)} \right) \sup_{\theta \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[W(X_k) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k\}} \right] \, .$$

Finally, by A2(c), there exists a constant $C>0$ (depending upon $\mathcal{K}$) such that for any $q \geq 0$,

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 0} \left|T_{4,n}^{(1)}\right| \right] \leq C \left(\sum_{k=1}^{\infty} \gamma_{q+k+1} \lambda^{\psi_q(k)} \right) W(x) \, . \tag{16}$$

Similarly, we obtain

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 0} \left|T_{4,n}^{(2)}\right| \right] \leq C \left(\sum_{k=1}^{\infty} \gamma_{q+k+1} \lambda^{\psi_q(k)} \right) W(x) \, . \tag{17}$$

Control of $T_{4,n}^{(3)}$

By Lemma 3.7, there exists $C>0$ such that for any $k \geq 1$

$$|h(\theta_k) - h(\theta_{k-1})| \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \leq C \left(D_W(\theta_k, \theta_{k-1}) + \|\theta_k - \theta_{k-1}\|^{\alpha} \right) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \, .$$

When $k \leq \sigma(\mathcal{K})$, $\theta_k - \theta_{k-1} = \gamma_{q+k} H(\theta_{k-1}, X_k)$ $\mathbb{P}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}}$-almost surely. By A1(b), this implies

$$|h(\theta_k) - h(\theta_{k-1})| \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \leq C \left(D_W(\theta_k, \theta_{k-1}) + \gamma_{q+k}^{\alpha} W^{\alpha}(X_k) \right) \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \, .$$

A2(c) finally yields

$$\begin{aligned} &\sup_{\theta \in \mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} \left[\sup_{n \geq 0} \left|T_{4,n}^{(3)}\right| \mathbf{1}_{A_{\Gamma}^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K},n)} \right] \\ &\quad \leq C \sum_{k=1}^{\infty} \gamma_{q+k+1} \psi_q(k) \left(\mathbb{E}_{x,\theta}^{\boldsymbol{\gamma}^{\leftarrow q}} [D_W(\theta_k, \theta_{k-1}) \mathbf{1}_{\{k+1 \leq \sigma(\mathcal{K})\}} \mathbf{1}_{A_{\Gamma}^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K},k)}] + \gamma_{q+k}^{\alpha} W^{\alpha}(x) \right) \, . \end{aligned} \tag{18}$$

Control of $T_{4,n}^{(4)}$

We finally consider the term $T_{4,n}^{(4)}$ along the event $A_{\Gamma}^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, n)$. We write

$$\begin{aligned} T_{4,n}^{(4)} &= \sum_{k=1}^{n-1} \gamma_{q+k+1} \sum_{l=1}^{\psi_q(k)} \left[P_{\theta_k}^l(H(\theta_k, .))(X_k) - P_{\theta_{k-\psi_q(k)}}^l(H(\theta_{k-\psi_q(k)}, .))(X_k) \right] \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \\ &\quad + \sum_{k=1}^{n-1} \gamma_{q+k+1} \sum_{l=1}^{\psi_q(k)} \left[P_{\theta_{k-\psi_q(k)}}^l(H(\theta_{k-\psi_q(k)}, .))(X_k) - P_{\theta_{k-1}}^l(H(\theta_{k-1}, .))(X_k) \right] \mathbf{1}_{\{\sigma(\mathcal{K}) \geq k+1\}} \end{aligned}$$

and consider the first term. The second term is on the same lines and will be omitted.

Note that as the sequence $\boldsymbol{\gamma}^{\leftarrow q}$ is non-increasing, on the set $A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, n)$, for any $k \leq n \wedge \sigma(\mathcal{K})$,

$$\|\theta_k - \theta_{k-\psi_q(k)}\| \leq \Gamma \ \psi_q(k) \ \gamma^r_{q+k-\psi_q(k)} \ .$$

Define the function $F_L : \Theta \times \mathbf{X} \to \Theta$ by

$$F_L(\tau, x) = \sup_{\{\theta : \|\theta - \tau\| \leq L\}} |H(\theta, x) - H(\tau, x)| \ .$$

By A4, the set $\mathcal{K}'$ is a compact set, so that by A1(b), $\sup_{\theta \in \mathcal{K}'} |H(\theta, .)|_W < \infty$. Then, (15) implies that there exists a constant $C_\star$ such that for any $k \in \mathbb{N}$ and any $q \geq q_\star$, on the set $\{\sigma(\mathcal{K}) \geq k - \psi_q(k)\}$,

$$|F_{\Gamma \sup_k \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)|_W \leq C_\star \ . \tag{19}$$

It holds on the set $A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, n) \cap \{k+1 \leq \sigma(\mathcal{K})\}$

$$\begin{aligned}
&\left| P^l_{\theta_k}(H(\theta_k, .))(X_k) - P^l_{\theta_{k-\psi_q(k)}}(H(\theta_{k-\psi_q(k)}, .))(X_k) \right| \\
&\leq P^l_{\theta_{k-\psi_q(k)}} \left| H(\theta_k, .) - H(\theta_{k-\psi_q(k)}, .) \right| (X_k) + \left| \left(P^l_{\theta_k} - P^l_{\theta_{k-\psi_q(k)}} \right) (H(\theta_k, .))(X_k) \right| \\
&\leq P^l_{\theta_{k-\psi_q(k)}} F_{\Gamma \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)(X_k) + C \sup_{\theta \in \mathcal{K}} |H(\theta, .)|_W \, D_W(\theta_k, \theta_{k-\psi_q(k)}) \, W(X_k)
\end{aligned} \tag{20}$$

where we used Lemma 3.5 in the last equality, and the constant C only depends on $\mathcal{K}$ (and not on k, q). On one hand, by the Hölder's inequality and the assumptions A2(c) and A6(civ)

$$\begin{aligned}
&\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[D_W(\theta_k, \theta_{k-\psi_q(k)}) \, W(X_k) \mathbf{1}_{k+1 \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, n)} \right] \\
&\qquad \leq \sup_{\theta \in \mathcal{K}} \left(\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[D_W(\theta_k, \theta_{k-\psi_q(k)})^{p/(p-1)} \ \ \mathbf{1}_{k+1 \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k)} \right] \right)^{(p-1)/p} W(x) \ .
\end{aligned} \tag{21}$$

On the other hand,

$$\begin{aligned}
&\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[P^l_{\theta_{k-\psi_q(k)}} F_{\Gamma \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)(X_k) \mathbf{1}_{k+1 \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k)} \right] \\
&\quad \leq \sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[P^l_{\theta_{k-\psi_q(k)}} F_{\Gamma \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)(X_k) \mathbf{1}_{k \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k)} \right] \\
&\quad \leq \sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[P_{\theta_{k-1}} P^l_{\theta_{k-\psi_q(k)}} F_{\Gamma \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)(X_{k-1}) \mathbf{1}_{k \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k-1)} \right] \\
&\quad \leq \sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[P^{l+1}_{\theta_{k-\psi_q(k)}} F_{\Gamma \psi_q(k) \gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)}, \cdot)(X_{k-1}) \mathbf{1}_{k \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k-1)} \right] \\
&\quad + C_\star \sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta} \left[D_W(\theta_{k-1}, \theta_{k-\psi_q(k)}) \ W(X_{k-1}) \ \mathbf{1}_{k \leq \sigma(\mathcal{K})} \mathbf{1}_{A_\Gamma^{\boldsymbol{\gamma}^{\leftarrow q}}(\mathcal{K}, k-1)} \right]
\end{aligned} \tag{22}$$

where we used (19) in the last inequality. By recursion, we have

$$\begin{aligned}
&\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[P^{l}_{\theta_{k-\psi_q(k)}}F_{\Gamma\psi_q(k)\gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)},\cdot)(X_k)\mathbf{1}_{k+1\le\sigma(\mathcal{K})}\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},k)}\right]\\
&\quad\le C_\star\sum_{j=1}^{\psi_q(k)-1}\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[D_W(\theta_{k-j},\theta_{k-\psi_q(k)})\;W(X_{k-j})\;\mathbf{1}_{k-j+1\le\sigma(\mathcal{K})}\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},k-j)}\right]\\
&\quad+\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[P^{l+\psi_q(k)}_{\theta_{k-\psi_q(k)}}F_{\Gamma\psi_q(k)\gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)},\cdot)(X_{k-\psi_q(k)})\mathbf{1}_{k-\psi_q(k)+1\le\sigma(\mathcal{K})}\right]\;. \qquad (23)
\end{aligned}$$

By A2(b-c), A3 and (19), there exists C such that for any $x\in\mathbf{X}$, $q,k,\ell\ge 0$

$$\begin{aligned}
&\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[P^{l+\psi_q(k)}_{\theta_{k-\psi_q(k)}}F_{\Gamma\psi_q(k)\gamma^r_{q+k-\psi_q(k)}}(\theta_{k-\psi_q(k)},\cdot)(X_{k-\psi_q(k)})\mathbf{1}_{k-\psi_q(k)+1\le\sigma(\mathcal{K})}\right]\\
&\qquad\le CC_\star\;\lambda^{\ell+\psi_q(k)}W(x)+C\;\left(\Gamma\psi_q(k)\gamma^r_{q+k-\psi_q(k)}\right)^{\alpha}\;. \qquad (24)
\end{aligned}$$

Therefore, by combining Eqs. (20) to (24), we obtain that there exists a constant C such that for any $q\ge q_\star$,

$$\begin{aligned}
&C\;\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[\sup_{n\ge0}\left|T^{(4)}_{4,n}(\mathcal{K})\right|\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},n)}\right]\\
&\quad\le\sum_k\gamma^{1+\alpha r}_{q+k-\psi_q(k)}\;\psi_q(k)^{1+\alpha}+\sum_k\gamma_{q+k+1}\;\lambda^{\psi_q(k)}\;W(x)\\
&\quad+W(x)\sum_k\gamma_{q+k+1}\psi_q(k)\sum_{j=0}^{\psi_q(k)-1}\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[D_W(\theta_{k-j},\theta_{k-\psi_q(k)})^{p/(p-1)}\;\mathbf{1}_{k-j+1\le\sigma(\mathcal{K})}\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},k-j)}\right]^{(p-1)/p}
\end{aligned}\qquad (25)$$

Conclusion

Combining the upper bounds (16), (17), (18) and (25), we obtain (14) with $C_k(\boldsymbol{\gamma}^{\leftarrow q})$ given by

$$\begin{aligned}
C_k(\boldsymbol{\gamma}^{\leftarrow q})&=\sum_k\gamma^2_{q+k}+\sum_k\gamma^{1+\alpha}_{q+k}\,\psi_q(k)+\sum_k\gamma^{1+\alpha r}_{q+k-\psi_q(k)}\;\psi_q(k)^{1+\alpha}\\
&+\sum_k\gamma_{k+q+1}\,\psi_q(k)\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[D_W(\theta_k,\theta_{k-1})\;\mathbf{1}_{k+1\le\sigma(\mathcal{K})}\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},k)}\right]\\
&+\sum_k\gamma_{q+k+1}\psi_q(k)\sum_{j=0}^{\psi_q(k)-1}\sup_{\theta\in\mathcal{K}}\mathbb{E}^{\boldsymbol{\gamma}^{\leftarrow q}}_{x,\theta}\left[D_W(\theta_{k-j},\theta_{k-\psi_q(k)})^{p/(p-1)}\;\mathbf{1}_{k-j+1\le\sigma(\mathcal{K})}\mathbf{1}_{A^{\boldsymbol{\gamma}^{\leftarrow q}}_\Gamma(\mathcal{K},k-j)}\right]^{(p-1)/p}
\end{aligned}$$

The conditions A6(b) and A6(cii-civ) imply that $\lim_q\sum_k C_k(\boldsymbol{\gamma}^{\leftarrow q})=0$. □

Lemma 3.5. *Assume A2(b). For any compact set $\mathcal{K}\subset\Theta$, there exists a constant C such that for any $\theta,\theta'\in\mathcal{K}$*

$$\sup_{n\ge0}\sup_{x\in\mathbf{X}}\frac{\|P^n_\theta(x,.)-P^n_{\theta'}(x,.)\|_W}{W(x)}\le CD_W(\theta,\theta')\;,$$

where D_W is defined by (9).

Proof. For any measurable function f such that $|f|_W \leq 1$,

$$P_\theta^n f(x) - P_{\theta'}^n f(x) = \sum_{j=0}^{n-1} P_{\theta'}^j (P_\theta - P_{\theta'}) \left(P_\theta^{n-j-1} f(x) - \pi_\theta(f) \right) .$$

Then for any $0 \leq j \leq n-1$,

$$\begin{aligned} \left| P_{\theta'}^j (P_\theta - P_{\theta'}) \left(P_\theta^{n-j-1} f(x) - \pi_\theta(f) \right) \right| &\leq P_{\theta'}^j W(x) \left| (P_\theta - P_{\theta'}) \left(P_\theta^{n-j-1} f - \pi_\theta(f) \right) \right|_W \\ &\leq D_W(\theta, \theta') \; P_{\theta'}^j W(x) \left| P_\theta^{n-j-1} f - \pi_\theta(f) \right|_W . \end{aligned}$$

By A2(b), there exist $C > 0$ and $\lambda \in (0,1)$ such that for any $\theta, \theta' \in \mathcal{K}$,

$$P_{\theta'}^j W(x) \left| P_\theta^{n-j-1} f - \pi_\theta(f) \right|_W \leq C \left(\lambda^j W(x) + \pi_{\theta'}(W) \right) \lambda^{n-j-1} .$$

This concludes the proof. □

Lemma 3.6. *Assume A2(b-c). For any compact set $\mathcal{K} \subset \Theta$, there exist $C > 0$ and $\lambda \in (0,1)$ such that for any $\theta, \theta' \in \mathcal{K}$*

$$\|\pi_\theta - \pi_{\theta'}\|_W \leq C D_W(\theta, \theta') .$$

Proof. For any $x \in \mathbf{X}, \psi \in \mathbb{N}$,

$$\|\pi_\theta - \pi_{\theta'}\|_W \leq \left\| \pi_\theta - P_\theta^\psi(x, \cdot) \right\|_W + \left\| P_\theta^\psi(x, \cdot) - P_{\theta'}^\psi(x, \cdot) \right\|_W + \left\| P_{\theta'}^\psi(x, \cdot) - \pi_{\theta'} \right\|_W .$$

By A2(b), there exist constants $C > 0$ and $\lambda \in (0,1)$ such that for any $\psi \in \mathbb{N}$ and $x \in \mathbf{X}$

$$\sup_{\theta \in \mathcal{K}} \left\| \pi_\theta - P_\theta^\psi(x, \cdot) \right\|_W \leq C \lambda^\psi W(x) .$$

Moreover, using Lemma 3.5, there exists a constant $C' > 0$ such that for any $\theta, \theta' \in \mathcal{K}$ and any $x \in \mathbf{X}$,

$$\sup_{\psi \in \mathbb{N}} \left\| P_\theta^\psi(x, \cdot) - P_{\theta'}^\psi(x, \cdot) \right\|_W \leq C' D_W(\theta, \theta') W(x) .$$

The proof follows, upon noting that x is fixed and arbitrarily chosen. □

Lemma 3.7. *Assume A1*(b)*, A2(b-c) and A3. For any compact set $\mathcal{K} \subset \Theta$, there exist $C > 0$ and $\lambda \in (0,1)$ such that for any $\theta, \theta' \in \mathcal{K}$,*

$$|h(\theta) - h(\theta')| \leq C \left(D_W(\theta, \theta') + \|\theta - \theta'\|^\alpha \right) ,$$

where D_W and α are given by (9) *and A3.*

Proof. Let $\mathcal{K}$ be a compact subset of Θ and θ and θ' be in $\mathcal{K}$. By definition of h, it holds

$$|h(\theta) - h(\theta')| = |\pi_\theta(H(\theta,.)) - \pi_{\theta'}(H(\theta',.))| \leq \pi_\theta(|H(\theta,.) - H(\theta',.)|) + |(\pi_\theta - \pi_{\theta'})(H(\theta',.))| \ .$$

Condition A3 implies that there exists a constant $C > 0$ such that for any $\theta, \theta' \in \mathcal{K}$,

$$\pi_\theta(|H(\theta,.) - H(\theta',.)|) \leq C\|\theta - \theta'\|^\alpha \ .$$

By Lemma 3.6 and condition A1(b), there exist constants $C > 0$ and $\lambda \in (0,1)$ such that for any $\theta, \theta' \in \mathcal{K}$,

$$\left|\pi_\theta(H(\theta',.)) - \pi_{\theta'}(H(\theta',.))\right| \leq C D_W(\theta, \theta') \ .$$

The proof follows. □

3.2 Proof of Theorem 2.1(ii)

By Theorem 2.1(i), there is an almost-sure finite number κ of updates of the active set. Denoting by T_κ the time when the last update occurs, the second step of the proof consists in studying the sum of the errors made from this last update to the end. Define

$$B_\kappa = \limsup_{l\to\infty} \sup_{n \geq T_\kappa + l} \left| \sum_{j=T_\kappa+l}^{n} \gamma_{\varsigma_j} \left(H(\theta_{j-1}, X_j) - h(\theta_{j-1})\right) \right| \mathbf{1}_{\{T_\kappa < \infty\}} \ .$$

Following the same lines as in [2, Theorem 5.5.], it can be shown that Propositions 3.2 and 3.4 imply $B_\kappa = 0$ almost surely. This concludes the proof of Theorem 2.1.(ii) using [2, Theorem 2.3]. Details are omitted and can be found in [2, Section 5].

4 Examples - Illustration

In all this section, for any $d \in \mathbb{N}$, any $x \in \mathbb{R}^d$ and any $r > 0$ we define

$$\mathcal{B}(x, r) = \{y \in \mathbb{R}^d, \|y - x\| \leq r\} \ .$$

4.1 Quantile approximation

The goal of this section is to estimate the quantile of order q, for a fixed $q \in (0,1)$, of a given distribution π which is assumed to satisfy the following conditions:

E1 The distribution π on $\mathbb{R}^d$ is absolutely continuous with respect to the Lebesgue measure, with bounded Radon–Nikodym derivative, and satisfies $\int \|x\|\pi(dx) < \infty$.

In particular, E1 implies that the cumulative distribution function associated with π is continuous.

E2 $\{P_\theta, \theta \in \Theta\}$ is a family of kernels satisfying A2, and such that $\pi_\theta = \pi$ for any $\theta \in \Theta$.

4.1.1 Quantile in one dimension

We focus here on the case of quantile approximation in one dimension (i.e. $d = 1$); $\Theta = \mathbb{R}$ and $\mathbf{X} = \mathbb{R}$. Let $q \in (0, 1)$. We consider the stochastic approximation procedure with field

$$H(\theta, x) = q - \mathbf{1}_{\{x \leq \theta\}} \ . \tag{26}$$

We prove that the conditions A1 (b), A3 and A4 are satisfied. Therefore, Algorithm 1 run with (π, P_θ) satisfying conditions E1 and E2, H given by (26), a truncation mapping Φ satisfying A1 (a) and a sequence $\{\gamma_n, n \in \mathbb{N}\}$ of stepsizes satisfying A6 defines a sequence $\{\theta_n, n \in \mathbb{N}\}$ converging to $\mathcal{L} = \{\theta \in \Theta, \mathbb{P}_\pi(X \leq \theta) = q\}$.

Proposition 4.1. *Assume E1 and E2. Then conditions A1* (b)*, A3 and A4 are satisfied for H given by (26), with $\mathcal{L} = \{\theta \in \Theta, \mathbb{P}_\pi(X \leq \theta) = q\}$.*

Proof. H is bounded, so A1(b) is satisfied for any function $W \geq 1$. Moreover,

$$|H(\theta_1, x) - H(\theta_2, x)| = \mathbf{1}_{\{\theta_1 \wedge \theta_2 \leq x < \theta_1 \vee \theta_2\}} \ ,$$

and by E1,

$$\sup_{\theta \in \mathbb{R}} \int \sup_{\theta_1, \theta_2 \in \mathcal{B}(\theta, \delta)} |H(\theta_1, x) - H(\theta_2, x)| \pi(dx) \leq \sup_{\theta \in \mathbb{R}} \int \mathbf{1}_{\{\theta - \delta \leq x \leq \theta + \delta\}} \pi(dx) \leq 2\delta \sup_{\mathbf{X}} \pi \ . \tag{27}$$

Therefore, A3 is satisfied with $\alpha = 1$ and $C = 2 \sup_{\mathbf{X}} \pi$.

Define

$$w(\theta) = \frac{1}{2} \mathbb{E}_\pi \left[|\theta - X| \right] + \left(\frac{1}{2} - q \right) \theta \ ,$$

where under $\mathbb{P}_\pi$ (and the associated expectation $\mathbb{E}_\pi$), $X \sim \pi$. We have for any $t \geq 0$,

$$\begin{aligned} w(\theta + t) - w(\theta) &= \frac{1}{2} \int \left(|\theta + t - x| - |\theta - x| \right) \pi(x) dx + \left(\frac{1}{2} - q \right) t \\ &= \frac{t}{2} \int_{\{x, x \leq \theta\}} \pi(x) dx - \frac{t}{2} \int_{\{x, x \geq \theta + t\}} \pi(x) dx \\ &\quad + \int_{\{x, \theta \leq x \leq \theta + t\}} (\theta - x) \pi(x) dx + \frac{t}{2} \int_{\{x, \theta \leq x \leq \theta + t\}} \pi(x) dx + \left(\frac{1}{2} - q \right) t \ . \end{aligned}$$

Therefore, w is differentiable, and

$$w'(\theta) = \frac{1}{2} \left(\mathbb{P}_\pi(X \leq \theta) - (1 - \mathbb{P}_\pi(X \leq \theta)) \right) + \left(\frac{1}{2} - q \right) = \mathbb{P}_\pi(X \leq \theta) - q \ .$$

By definition of h and H (see (7) and (26)),

$$h(\theta) = \pi(H(\theta, .)) = q - \mathbb{P}_\pi(X \leq \theta) = -w'(\theta) \ .$$

Therefore, the set $\mathcal{L}$ in A4(a) is given by $\mathcal{L} = \{\theta \in \Theta, \mathbb{P}_\pi(X \leq \theta) = q\}$, and A4(c) is satisfied. Note that w is constant on $\mathcal{L}$ since $w'(\theta) = 0$ for any $\theta \in \mathcal{L}$ and $\mathcal{L}$ is an interval. Hence, A4(a) and A4(d) hold. Moreover, by (27), there exists a constant C such that

$$\left|h(\theta) - h(\theta')\right| = \left|w'(\theta) - w'(\theta')\right| = \left|\pi\left(H(\theta, .) - H(\theta', .)\right)\right| \leq C|\theta - \theta'| \, .$$

Therefore, w' and h are continuous.

In addition, as w is continuous, A4(b) holds if $\lim_{|\theta|\to\infty} w(\theta) = \infty$. Note that this holds true since under E2,

$$w(\theta) \geq \frac{|\theta|}{2} + \left(\frac{1}{2} - q\right)\theta - \frac{1}{2}\mathbb{E}_\pi[|X|] \underset{|\theta|\to\infty}{\longrightarrow} \infty \, .$$

Finally, observe that $w(\theta)$ reaches its minimum at $\theta_\star \in \mathcal{L}$. Since the Lyapunov function w is defined up to an additive constant, we can assume with no loss of generality that w is non-negative, which concludes the proof.

□

4.1.2 Median in multi-dimensional spaces

Here, $d > 1$, $\Theta = \mathbb{R}^d$ and $\mathbf{X} = \mathbb{R}^d$. This section aims at approximating the median of a multi-dimensional distribution. To that goal, we consider the stochastic approximation procedure with field

$$H(\theta, X) = \frac{X - \theta}{\|X - \theta\|}\mathbf{1}_{X \neq \theta} \, . \tag{28}$$

Proposition 4.2. *Assume E1 and E2. Then conditions A1* (b)*, A3 and A4 are satisfied for the field H defined by (28), and $\mathcal{L}$ is the singleton $\{\theta_*\}$, where θ_* is the unique solution of $\mathbb{E}_\pi\left[\frac{X-\theta}{\|X-\theta\|}\right] = 0$.*

Proof. Throughout the proof, set $u(x) \overset{\text{def}}{=} x/\|x\|$. As $\|H\| = 1$, A1(b) is satisfied for any function $W \geq 1$. Moreover, for $x \notin \{\theta_1, \theta_2\}$,

$$\begin{aligned}
\|H(\theta_1, x) - H(\theta_2, x)\| &= \left\| \frac{(x - \theta_1)\|x - \theta_2\| - (x - \theta_2)\|x - \theta_1\|}{\|x - \theta_1\|\|x - \theta_2\|} \right\| \\
&= \left\| \frac{x - \theta_1}{\|x - \theta_1\|\|x - \theta_2\|}(\|x - \theta_2\| - \|x - \theta_1\|) + \frac{\theta_2 - \theta_1}{\|x - \theta_2\|} \right\| \\
&\leq 2\frac{\|\theta_1 - \theta_2\|}{\|x - \theta_2\|} \, .
\end{aligned}$$

Define

$$\Delta H_{\theta,\delta}(x) = \sup_{\theta_1, \theta_2 \in \mathcal{B}(\theta,\delta)} \|H(\theta_1, x) - H(\theta_2, x)\| \, .$$

Let $0 < \beta < 1/d$. Then

$$\begin{aligned}\int \pi(x)\Delta H_{\theta,\delta}(x)dx &= \int_{x\in\mathcal{B}(\theta,\delta+\delta^\beta)} \pi(x)\Delta H_{\theta,\delta}(x)dx + \int_{x\notin\mathcal{B}(\theta,\delta+\delta^\beta)} \pi(x)\Delta H_{\theta,\delta}(x)dx \\ &\le \int_{x\in\mathcal{B}(\theta,\delta+\delta^\beta)} 2\sup_{\theta,x}\|H(\theta,x)\|\pi(x)dx + \int_{x\notin\mathcal{B}(\theta,\delta+\delta^\beta)} 2\sup_{\theta_1,\theta_2\in\mathcal{B}(\theta,\delta)} \frac{\|\theta_1-\theta_2\|}{\|x-\theta_2\|}\pi(x)dx \\ &\le 2\int_{x\in\mathcal{B}(\theta,\delta+\delta^\beta)} \pi(x)dx + 4\delta^{1-\beta} \ .\end{aligned}$$

By E1, there exists a constant $C > 0$ such that for any $\delta \in (0,1)$,

$$\sup_{\theta\in\Theta}\int \pi(dx)\Delta H_{\theta,\delta}(x) \le C(\delta^{\beta d} + \delta^{1-\beta}) \ ,$$

and A3 is satisfied with $\alpha = \beta d \wedge (1-\beta) < 1$. To prove that A4 is satisfied, define

$$w(\theta) = \mathbb{E}_\pi\left[\|X-\theta\|\right] \ .$$

For any $x, \theta, t \in \mathbb{R}^d$ it holds

$$\|x-\theta+t\| = \|x-\theta\| - \Big\langle t, u(x-\theta)\Big\rangle + \frac{1}{2}t^T\left(\int_0^1 \frac{1-\lambda}{\|x-\theta+\lambda t\|}\left(I - u(x-\theta+\lambda t)u(x-\theta+\lambda t)^T\right)d\lambda\right)t \ .$$

Therefore,

$$\mathbb{E}\left[\|X-\theta+t\|\right] = \mathbb{E}\left[\|X-\theta\|\right] - \Big\langle t, \mathbb{E}\left[u(X-\theta)\right]\Big\rangle + \frac{1}{2}t^T R(\theta,t)t \ ,$$

where

$$R(\theta,t) \stackrel{\text{def}}{=} \mathbb{E}\left[\int_0^1 \frac{1-\lambda}{\|X-\theta+\lambda t\|}\left(I - u(X-\theta+\lambda t)u(X-\theta+\lambda t)^T\right)d\lambda\right] \ .$$

Lemma 4.3 and the Fubini theorem imply that

$$R(\theta,t) = \int_0^1 \mathbb{E}\left[\frac{1-\lambda}{\|X-\theta+\lambda t\|}\left(I - u(X-\theta+\lambda t)u(X-\theta+\lambda t)^T\right)\right]d\lambda \ .$$

Since $\|u(x)\| \le 1$, there exists a constant C such that for any t, θ,

$$|t^T R(\theta,t)t| \le C\sup_{\theta\in\Theta}\mathbb{E}\left[\|X-\theta\|^{-1}\right]\|t\|^2 \ .$$

This implies that $\nabla w(\theta) = -\mathbb{E}\left[u(X-\theta)\right] = -h(\theta)$ and directly gives the condition A4(c).

In addition, we can write

$$\left\|h(\theta') - h(\theta)\right\| \le \mathbb{E}_\pi\left[\left\|\frac{X-\theta}{\|X-\theta\|} - \frac{X-\theta'}{\|X-\theta'\|}\right\|\right] \ .$$

so that, by the dominated convergence theorem, ∇w and h are continuous.

Moreover, by E1, A4(b) is satisfied because

$$w(\theta) \ge \|\theta\| - \mathbb{E}_\pi[\|X\|] \underset{\|\theta\|\to\infty}{\longrightarrow} \infty \ .$$

Finally, by E1 and [31], $\mathcal{L}$ contains a single point, and A4(a) and A4(d) are satisfied. □

Lemma 4.3. *Under E1, for any* $0 \leq \kappa < d$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\pi \left[\frac{1}{\|X - \theta\|^\kappa} \right] < \infty \, .$$

Proof. Let $0 < \kappa < d$.

$$\mathbb{E}_\pi \left[\|X - \theta\|^{-\kappa} \right] = \int_0^\infty \mathbb{P}_\pi \left[\|X - \theta\|^\kappa \leq \frac{1}{t} \right] dt \leq 1 + \int_1^\infty \mathbb{P}_\pi \left[\|X - \theta\|^\kappa \leq \frac{1}{t} \right] dt \, .$$

By E1, there exists a constant C such that for any $t \geq 1$,

$$\sup_{\theta \in \Theta} \mathbb{P}_\pi \left[\|X - \theta\|^\kappa \leq \frac{1}{t} \right] = \sup_{\theta \in \Theta} \int_{x \in \mathcal{B}(\theta, t^{-1/\kappa})} \pi(dx) \leq C \, t^{-d/\kappa} \, ,$$

This concludes the proof since $d/\kappa > 1$. □

4.2 Vector quantization

4.2.1 Context

Vector quantization is a well known problem [42] which consists in approximating a random vector in $\mathbb{R}^d$ by a random vector taking at most N values in $\mathbb{R}^d$. Such a problem occurs in many mathematical fields, as for example information theory, speech coding [18], numerical integration [32] or finance [33].

For $\theta = (\theta_1, \theta_2, \ldots, \theta_N) \in (\mathbb{R}^d)^N$, and for any $1 \leq i \leq N$, define the Voronoi cells associated to the sites θ by

$$\overline{C}_i(\theta) = \left\{ u \in \mathbb{R}^d, \|u - \theta_i\| = \min_{1 \leq j \leq N} \|u - \theta_j\| \right\} \, .$$

A Voronoi partition $(C_i(\theta))_{1 \leq i \leq N}$ of $\mathbb{R}^d$ associated with $\theta \in (\mathbb{R}^d)^N$ is a collection of sets satisfying

$$\bigcup_{i=1}^N C_i(\theta) = \mathbb{R}^d \, , \quad C_i(\theta) \cap C_j(\theta) = \emptyset \quad \text{if} \quad \theta_i \neq \theta_j \quad \text{and} \quad C_i(\theta) \subset \overline{C}_i(\theta) \quad \forall \, 1 \leq i \leq N \, .$$

This partition allows to approximate a random vector X by $\widehat{X}^\theta = \sum_{i=1}^N \theta_i \mathbf{1}_{C_i(\theta)}(X)$. Denote by w the mean squared error when approximating X by $\widehat{X}^\theta$:

$$w(\theta) = \mathbb{E} \left[\|X - \widehat{X}^\theta\|^2 \right] = \sum_{i=1}^N \mathbb{E} \left[\|X - \theta_i\|^2 \mathbf{1}_{C_i(\theta)}(X) \right] \, . \tag{29}$$

w is often called the distortion. Whenever π is such that $\mathbb{E}[\|X\|^2] < \infty$, then w is guaranteed to be finite. Given the distribution of X, vector quantization consists in finding $\theta \in (\mathbb{R}^d)^N$ minimizing the distortion w.

Numerous studies of optimal quantizers and their asymptotic properties, when $N \to \infty$, have been done (see for example [17]). In practice the optimal quantizer has no explicit formulation,

and needs to be approximated. It was first proposed in the literature to retrieve optimal quantizers by deterministic methods based on a fixed point property of this optimum (see [28] and [29]). Unfortunately, these methods are in general intractable in more than one or two dimensions. Therefore, some stochastic methods with more tractable computations have been introduced by Kohonen [22]. The Kohonen algorithm (with 0 neighbors) is a stochastic approximation algorithm with field $H : (\mathbb{R}^d)^N \times \mathbb{R}^d \to (\mathbb{R}^d)^N$ given by

$$H(\theta, u) = -2\left((\theta_i - u)\mathbf{1}_{C_i(\theta)}(u)\right)_{1\leq i\leq N} . \tag{30}$$

An iteration of this algorithm is given by

$$\theta^{(n+1)} = \theta^{(n)} + \gamma_{n+1} H(\theta^{(n)}, X_{n+1}) , \tag{31}$$

where $(X_n)_{n\in\mathbb{N}}$ are random vectors with distribution related (in some sense, see below for an example) to the distribution of X.

There exist few results on the theoretical properties of the Kohonen algorithm (see [16] for a review). Indeed, the convergence of this algorithm has only been proven in one dimension for i.i.d. observations $(X_n)_{n\in\mathbb{N}}$ with the same distribution as X [13]. Nevertheless, in many applications the dimension is larger than one, and the dynamics of the observations can be Markovian (see for example the examples in finance described in [33]). The goal here is to extend these results.

4.2.2 Convergence of the Kohonen algorithm

We consider here Algorithm 1 run with H defined in (30) and a collection of kernels $\{P_\theta, \theta \in \Theta\}$ satisfying assumptions E3 and E4:

E3 The distribution of X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Denote by π its density. The density π has a bounded support, that is $\pi(x) = 0$ for any $x \in \mathcal{B}(0, \Delta)^c$ for some $\Delta > 0$.

E4 $\{P_\theta, \theta \in \Theta\}$ is a family of kernels satisfying A2, and such that $\pi_\theta = \pi$ for any $\theta \in \Theta$.

Let $\Theta = \left\{\theta = (\theta_1, \dots \theta_N) \in (\mathbb{R}^d)^N \cap (\mathcal{B}(0, \Delta))^N, \theta_i \neq \theta_j \ \forall i \neq j\right\}$. Lemma 4.4 shows that if the algorithm is initialized in Θ ($\theta^{(0)} \in \Theta$), then it remains in Θ almost surely ($\mathbb{P}(\forall n \in \mathbb{N}, \theta^{(n)} \in \Theta) = 1$).

Lemma 4.4. *For any $\gamma \leq 1/2$, $z \in \mathcal{B}(0, \Delta)$ and $\theta \in \Theta$, $\theta + \gamma H(\theta, z) \in \Theta$.*

Proof. Let $z \in \mathcal{B}(0, \Delta)$ and $\theta = (\theta_1, \cdots, \theta_N) \in (\mathcal{B}(0, \Delta))^N$. Denote by i the unique integer in $\{1, \cdots, N\}$ such that $z \in C_i(\theta)$. Set $\theta' = \theta + \gamma H(\theta, z)$. Then

$$\theta'_j = \theta_j, j \neq i \qquad\qquad \theta'_i = (1 - 2\gamma)\theta_i + 2\gamma z . \tag{32}$$

Since $2\gamma \in (0, 1)$, $\theta_k \in \mathcal{B}(0, \Delta)$ for any k and $z \in \mathcal{B}(0, \Delta)$, then $\theta' \in (\mathcal{B}(0, \Delta))^N$. Let us prove that $\theta'_j \neq \theta'_k$ for any $j \neq k$. Since $\theta \in \Theta$, this holds true by (32) for any $j, k \neq i$. When $k = i$, we have

$$\|\theta'_i - \theta'_j\| = \|\theta_i - \theta_j + 2\gamma(z - \theta_i)\| \geq |1 - 2\gamma\Pi(z)| \ \|\theta_i - \theta_j\|$$

where we wrote $z - \theta_i = \Pi(z)(\theta_j - \theta_i) + (z - \theta_i)^\perp$ for $\Pi(z) \in \mathbb{R}$ such that $\left\langle (z - \theta_i)^\perp, \theta_j - \theta_i \right\rangle = 0$. We also have $z - \theta_j = (\Pi(z) - 1)(\theta_j - \theta_i) + (z - \theta_i)^\perp$. Since $z \in C_i(\theta)$, $\|z - \theta_i\| \leq \|z - \theta_j\|$ and we have $|\Pi(z)| \leq |\Pi(z) - 1|$ which is equivalent to $\Pi(z) \leq 1/2$. Then, $1 - 2\gamma\Pi(z) \geq 1/2$ and $\|\theta'_i - \theta'_j\| > 0$. □

Lemma 4.5, which is a restatement of [32, Proposition 5 and 9] establishes that there exist optimal quantizers in Θ, which are also in the set of the zeros of the mean field associated to H.

Lemma 4.5. *Assume E3 and let h be the mean field function defined by (7) associated to the function H given by (30). Then,*

1. *w is continuous on $(\mathbb{R}^d)^N$ and differentiable on Θ; $h(\theta) = -\nabla w(\theta)$ and h is continuous on Θ.*
2. *$argmin_{\theta\in(\mathbb{R}^d)^N} w(\theta) \cap \Theta \neq \emptyset$ and $argmin_{\theta\in(\mathbb{R}^d)^N} w(\theta) \subset \left\{\theta \in (\mathbb{R}^d)^N, \theta_i \neq \theta_j \ \forall i \neq j\right\}$.*
3. *$argmin_{\theta\in\Theta} w(\theta) \subset \{\theta \in \Theta, h(\theta) = 0\}$.*

Proposition 4.6 shows that our assumptions on the function H are satisfied under E3 and E4.

Proposition 4.6. *Assume E3 and E4. Then conditions A1(b), A1(c), A3, A4(a) and A4(c) are satisfied by the field H defined in (30) and the Lyapunov function w defined in (29).*

The proof of Proposition 4.6 is postponed in Section 4.2.3. Proposition 4.6 implies the convergence of the Kohonen algorithm, as stated in Corollary 4.7.

Corollary 4.7. *Assume E3, E4 and that the density π is such that conditions A4(b) and A4(d) are satisfied. Then the 0 neighbors Kohonen algorithm converges to $\mathcal{L} = \{\theta \in \Theta, \nabla w(\theta) = 0\}$, where w is the distortion (see (29)).*

Remark 4.1. By Sard's theorem, A4(d) is satisfied if w is Nd times continuously differentiable. A sufficient condition for A4(b) and A4(d) to hold is:

$$\mathcal{L} = \text{argmin}_{\theta\in\Theta}(w(\theta)) \ . \tag{33}$$

Indeed, under (33), $w(\mathcal{L}) = \min_{\theta\in\Theta} w(\theta)$ is a singleton, so that A4(d) is satisfied. Moreover, by Lemma 4.5(2) and continuity of w, $w(\mathcal{L}) < M$, where

$$M = \inf\left\{w(\theta), \theta = (\theta_1, \dots \theta_N) \in (\mathbb{R}^d)^N | \exists i \neq j, \theta_i = \theta_j\right\} \ .$$

By choosing M_0 and M_1 such that $w(\mathcal{L}) < M_0 < M_1 < M$, we have that

$$\{\theta \in \Theta, w(\theta) \leq M_1\} = \left\{\theta \in (\mathbb{R}^d)^N \cap \mathcal{B}(0, \Delta)^N, w(\theta) \leq M_1\right\} \ ,$$

which is a compact set by Lemma 4.5(1). Therefore A4(b) is satisfied.

If $d = 1$ and π is log-concave (example: uniform distribution, Gaussian distribution), then it is proved in [26] (see also [13, Theorem 3]) that $\mathcal{L}$ is a singleton (up to a permutation of its elements), and therefore, by Lemma 4.5(3), (33) is satisfied.

As a conclusion of the above discussion, we established the convergence of the 0 neighbors Kohonen algorithm under weaker assumptions on the dynamics $(X_n)_{n\in\mathbb{N}}$ and on π than previous works (see e.g. [13]):

- Our framework addresses the case when $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with invariant distribution π, or when $(X_n)_{n\in\mathbb{N}}$ is a controlled Markov chain where each transition kernel admits π as invariant density.
- We have no condition on the dimension d. Our results apply whatever d is provided A4(b) and A4(d) are satisfied.

4.2.3 Proof of Proposition 4.6

We start with a preliminary lemma which gives a control on the intersection of two Voronoi cells associated with two different sites $\theta, \theta' \in (\mathbb{R}^d)^N$.

Lemma 4.8. *For any compact set $\mathcal{K}$ of Θ, there exists $\delta_{\mathcal{K}} > 0$ such that for any $\theta \in \mathcal{K}$ and any $i \neq j$:*

(i)

$$\sup_{\delta \leq \delta_K} \frac{1}{\sqrt{\delta}} \sup_{\theta' \in \mathcal{B}(\theta,\delta)^N \cap \Theta} \left\| \frac{\theta'_j - \theta'_i}{\|\theta'_j - \theta'_i\|} - \frac{\theta_j - \theta_i}{\|\theta_j - \theta_i\|} \right\| < \infty \, . \tag{34}$$

(ii) for any $\delta \leq \delta_{\mathcal{K}}$, there exists a measurable set $R_{i,j}(\theta, \delta)$ such that

$$\sup_{\theta' \in \mathcal{B}(\theta,\delta) \cap \Theta} \mathbf{1}_{C_i(\theta) \cap C_j(\theta') \cap \mathcal{B}(0,\Delta)} \leq \mathbf{1}_{R_{i,j}(\theta,\delta)} \, , \qquad \sup_{\delta \leq \delta_K} \frac{1}{\sqrt{\delta}} \int \mathbf{1}_{R_{i,j}(\theta,\delta)}(x) \, dx < \infty \, . \tag{35}$$

Proof. Let $\mathcal{K}$ be a compact set of Θ. The function on $(\mathbb{R}^d)^N$ given by $\theta \mapsto \min_{i \neq j} \|\theta_i - \theta_j\|$ is continuous. Since $\mathcal{K}$ is a compact subset of Θ, there exists $b_{\mathcal{K}} > 0$ such that for any $\theta \in \mathcal{K}$, $\min_{i \neq j} \|\theta_i - \theta_j\| \geq b_{\mathcal{K}}$. Choose $\delta_{\mathcal{K}} \in (0, b_{\mathcal{K}}/2 \wedge 1)$. Let $i \neq j \in \{1, \cdots, N\}$ and $\theta \in \mathcal{K}$ be fixed. For any $\delta \leq \delta_{\mathcal{K}}$ and $\theta' \in \mathcal{B}(\theta, \delta)$, it holds

$$\begin{aligned} \|\theta'_j - \theta'_i\| &\geq \|\theta_j - \theta_i\| - \|\theta'_j - \theta_j\| - \|\theta'_i - \theta_i\| \geq \|\theta_j - \theta_i\| - 2\delta \\ &\geq b_{\mathcal{K}} - 2\delta > 0 \, . \end{aligned} \tag{36}$$

Similarly,

$$\|\theta'_j - \theta'_i\| \leq \|\theta_j - \theta_i\| + 2\delta \, . \tag{37}$$

Define

$$n = \frac{\theta_j - \theta_i}{\|\theta_j - \theta_i\|} \quad \text{and} \quad n' = \frac{\theta'_j - \theta'_i}{\|\theta'_j - \theta'_i\|} \, .$$

Proof of (34) We have $\|n - n'\|^2 = 2\left(1 - \left\langle n, n' \right\rangle\right)$. In addition, for any $\delta \leq \delta_{\mathcal{K}}$ and $\theta' \in \mathcal{B}(\theta, \delta)$,

$$\begin{aligned} \left\langle n, n' \right\rangle &= \|\theta_j - \theta_i\|^{-1} \left\langle \theta_j - \theta_i, n' \right\rangle \\ &= \|\theta_j - \theta_i\|^{-1} \left\langle \|\theta'_j - \theta'_i\| n' + \theta_j - \theta'_j + \theta'_i - \theta_i, n' \right\rangle \\ &\geq \frac{\|\theta'_j - \theta'_i\|}{\|\theta_j - \theta_i\|} - \frac{2\delta}{\|\theta_j - \theta_i\|} \geq 1 - \frac{4\delta}{\|\theta_j - \theta_i\|} \geq 1 - \frac{4\delta}{b_{\mathcal{K}}} \, , \end{aligned}$$

where we used (36) in the last row. Therefore

$$\|n - n'\|^2 \leq 8\delta / b_{\mathcal{K}} \, , \tag{38}$$

which concludes the proof.

Proof of (35) Let $x \in C_i(\theta)$. We write

$$x - \theta_i = \Big\langle x - \theta_i, n \Big\rangle n + m \quad \text{where} \quad \Big\langle m, n \Big\rangle = 0 \; .$$

It stands $\|x-\theta_i\|^2 = \Big|\Big\langle x - \theta_i, n \Big\rangle\Big|^2 + \|m\|^2$. Moreover, $x-\theta_j = x-\theta_i+\theta_i-\theta_j = \Big\langle x-\theta_i, n \Big\rangle n - \|\theta_i - \theta_j\| n + m$ leading to $\|x-\theta_j\|^2 = \Big|\Big\langle x - \theta_i, n \Big\rangle - \|\theta_i - \theta_j\|\Big|^2 + \|m\|^2$. Since $x \in C_i(\theta)$, $\|x-\theta_i\| \leq \|x-\theta_j\|$ so that $\Big|\Big\langle x - \theta_i, n \Big\rangle\Big|^2 \leq \Big|\Big\langle x - \theta_i, n \Big\rangle - \|\theta_i - \theta_j\|\Big|^2$. This implies that $\Big\langle x - \theta_i, n \Big\rangle \leq \|\theta_j - \theta_i\|/2$. Therefore,

$$C_i(\theta) \subset \left\{ x \in \mathbb{R}^d, \Big\langle x - \theta_i, n \Big\rangle \leq \frac{1}{2}\|\theta_j - \theta_i\| \right\} \; .$$

Let now $x \in C_j(\theta') \cap \mathcal{B}(0,\Delta)$. Following the same lines as above and using (37)

$$\Big\langle x - \theta'_j, n' \Big\rangle \geq -\frac{1}{2}\|\theta'_j - \theta'_i\| \geq -\frac{1}{2}\|\theta_j - \theta_i\| - \delta \; . \tag{39}$$

Moreover

$$\begin{aligned}
\Big\langle x - \theta_i, n \Big\rangle &= \Big\langle x - \theta_i, n - n' \Big\rangle + \Big\langle x - \theta'_j, n' \Big\rangle + \Big\langle \theta'_j - \theta'_i, n' \Big\rangle + \Big\langle \theta'_i - \theta_i, n' \Big\rangle \\
&= \Big\langle x - \theta_i, n - n' \Big\rangle + \Big\langle x - \theta'_j, n' \Big\rangle + \|\theta'_j - \theta'_i\| + \Big\langle \theta'_i - \theta_i, n' \Big\rangle \; .
\end{aligned}$$

Since $x, \theta_i \in \mathcal{B}(0,\Delta)$, we have by (36), (38) and (39)

$$\begin{aligned}
\Big\langle x - \theta_i, n \Big\rangle &\geq -2\Delta\|n - n'\| - \frac{1}{2}\|\theta_j - \theta_i\| - \delta + \|\theta_j - \theta_i\| - 2\delta - \delta \\
&\geq \frac{1}{2}\|\theta_j - \theta_i\| - 4\delta - 4\Delta\sqrt{2/b_{\mathcal{K}}}\sqrt{\delta} \; .
\end{aligned}$$

Therefore,

$$C_j(\theta') \cap \mathcal{B}(0,\Delta) \subset \left\{ x \in \mathbb{R}^d, \Big\langle x - \theta_i, n \Big\rangle \geq \frac{1}{2}\|\theta_j - \theta_i\| - 4\delta - 4\Delta\sqrt{2/b_{\mathcal{K}}}\sqrt{\delta} \right\} \; .$$

Hence,

$$C_i(\theta) \cap C_j(\theta') \cap \mathcal{B}(0,\Delta) \subset \left\{ x \in \mathcal{B}(0,\Delta), \frac{1}{2}\|\theta_j - \theta_i\| - 4\delta - 4\Delta\sqrt{2/b_{\mathcal{K}}}\sqrt{\delta} \leq \Big\langle x - \theta_i, n \Big\rangle \leq \frac{1}{2}\|\theta_j - \theta_i\| \right\} \; .$$

Finally, since $\delta_{\mathcal{K}} < 1$, we have $\delta \leq \sqrt{\delta}$, and this concludes the proof, by noticing that this last set is independent of θ'. □

Proof of Proposition 4.6. For any compact set $\mathcal{K} \subset \Theta$, there exists C such that $\sup_{\theta \in \mathcal{K}} \|H(\theta, u)\| \leq C(\|u\| + 1)$. Therefore, A1(b) and A1(c) are satisfied with $W(u) = 1 + \|u\|$.

By Lemma 4.5(1), w is nonnegative and continuously differentiable on Θ. Moreover, as $\nabla w = -h$, A4(c) is satisfied. And A4(a) is satisfied as w is bounded on Θ.

Let us prove that A3 is satisfied. Let $\mathcal{K}$ be a compact set of Θ. For any $\theta \in \mathcal{K}$, any $\theta' \in \Theta$, and any $x \in \mathbb{R}^d$,

$$
\begin{aligned}
&1/4\|H(\theta', x) - H(\theta, x)\|^2 \\
&= \sum_{i=1}^{N} \left[\|\theta_i' - \theta_i\|^2 \mathbf{1}_{C_i(\theta) \cap C_i(\theta')}(x) + \|\theta_i - x\|^2 \mathbf{1}_{C_i(\theta) \cap C_i(\theta')^c}(x) + \|\theta_i' - x\|^2 \mathbf{1}_{C_i(\theta)^c \cap C_i(\theta')}(x) \right] .
\end{aligned}
$$

Therefore, for any $x \in \mathcal{B}(0, \Delta)$, any $\theta \in \mathcal{K}$, and any $\theta' \in \mathcal{B}(0, \delta)$,

$$
\begin{aligned}
1/2\|H(\theta', x) - H(\theta, x)\| &\leq \sqrt{\sum_{i=1}^{N} \|\theta_i' - \theta_i\|^2} + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \|\theta_i - x\| \mathbf{1}_{C_i(\theta) \cap C_j(\theta')}(x) \\
&\quad + \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \|\theta_i' - x\| \mathbf{1}_{C_j(\theta) \cap C_i(\theta')}(x) \\
&\leq \delta + 2\Delta N^2 \sup_{i \neq j} \mathbf{1}_{C_i(\theta) \cap C_j(\theta') \cap \mathcal{B}(0, \Delta)}(x) .
\end{aligned}
$$

By Lemma 4.8, there exists $\delta_{\mathcal{K}}$ such that for any $\delta \leq \delta_{\mathcal{K}}$, there exist a measurable set $R_{i,j}(\theta, \delta)$ such that

$$
\sup_{\theta' \in \mathcal{B}(0, \delta)} \mathbf{1}_{C_i(\theta) \cap C_j(\theta') \cap \mathcal{B}(0, \Delta)}(x) \leq \mathbf{1}_{R_{i,j}(\theta, \delta)}(x) .
$$

Therefore,

$$
1/2\|H(\theta', x) - H(\theta, x)\| \leq \delta + 2\Delta N^2 \sup_{i \neq j} \mathbf{1}_{R_{i,j}(\theta, \delta)}(x) .
$$

Under E3, π is bounded on Θ. In addition, Lemma 4.8 shows that

$$
\sup_{\delta \leq \delta_{\mathcal{K}}} \frac{1}{\sqrt{\delta}} \sup_{\theta \in \mathcal{K}} \sup_{i \neq j} \int \mathbf{1}_{R_{i,j}(\theta, \delta)}(x) dx < \infty .
$$

Then, there exists C' such that for any $\delta \leq \delta_{\mathcal{K}}$,

$$
\sup_{\theta \in \mathcal{K}} \int \pi(dx) \sup_{\{\theta', \|\theta' - \theta\| \leq \delta\}} \left\| H(\theta', x) - H(\theta, x) \right\| \leq C' \sqrt{\delta} .
$$

Moreover, as $\|H\|$ is bounded on $\Theta \times \mathcal{B}(0, \Delta)$, for any $\delta \geq \delta_{\mathcal{K}}$,

$$
\sup_{\theta \in \mathcal{K}} \int \pi(dx) \sup_{\{\theta', \|\theta' - \theta\| \leq \delta\}} \left\| H(\theta', x) - H(\theta, x) \right\| \leq 2 \sup_{\Theta \times \operatorname{supp}(\pi)} (\|H\|) \frac{1}{\min(1, \sqrt{\delta_{\mathcal{K}}})} \sqrt{\delta} .
$$

Therefore A3 is satisfied with $\alpha = 1/2$.

□

5 Conclusion

As briefly illustrated in Section 4, stochastic approximation procedures with discontinuous field can be found in a lot of applications, for which the independence assumption for the observation sequence $\{X_n, n \in \mathbb{N}\}$ may be unrealistic as, for example, in learning or in finance. In this paper, we have proposed a theoretical justification for the use of such procedures, in the case where the associated fields are discontinuous. This provides for example a justification to adaptation procedures using stochastic approximations of quantiles or median in Markov chain or for vector quantization in Markovian contexts that often arise in finance.

References

[1] C. Andrieu and E. Moulines. On the ergodicity property of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

[2] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312, 2005.

[3] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, 2008.

[4] C. Andrieu and M. Vihola. Markovian stochastic approximation with expanding projections. *Bernouilli*, 20(2):545–585, 2014.

[5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In J. Azéma, M. Émery, M. Ledoux, and M. Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Berlin Heidelberg, 1999.

[6] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, 1990.

[7] J.R. Blum. Approximation methods which converge with probability one. *Ann. Math. Statist.*, 25:382–386, 1954.

[8] V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[9] L. Bottou. Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-Nîmes 91*. EC2, 1991.

[10] L. Bottou. Online Algorithms and Stochastic Approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. revised, oct 2012.

[11] H. Chen, L. Guo, and A. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.*, 27:217–231, 1988.

[12] H. Chen and Y.M. Zhu. Stochastic Approximation procedures with random varying truncations. *Scientia Sinica (Series A)*, 29:914 – 926, 1986.

[13] S. Delattre, J.C. Fort, and G. Pagès. Local Distortion and μ-Mass of the Cells of One Dimensional Asymptotically Optimal Quantizers. *Communications in Statistics - Theory and Methods*, 33(5):1087–1117, 2004.

[14] V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.*, 39:1327–1332, 1968.

[15] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2012.

[16] J.C. Fort. SOM's mathematics. *Neural Netw.*, 19(6):812–816, 2006.

[17] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions Foundations of Quantization for Probability Distributions*, volume 1730. Springer Berlin Heidelberg, 2000.

[18] B. Juang, D.Y. Wong, and A.H. Gray. Recent developments in vector quantization for speech processing. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1981.

[19] S. Kamal. Stabilization of stochastic approximation by step size adaptation. *Systems and Control Letters*, 61(4):543–548, 2012.

[20] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23:462 – 466, 1952.

[21] J. Kim and W.B. Powell. Quantile optimization for heavy-tailed distributions using asymmetric signum functions. *under review*, 2011.

[22] T. Kohonen. Analysis of simple self-organising process. *Biological Cybernetics*, 44:135–140, 1982.

[23] H. J. Kushner. Stochastic approximation with discontinuous dynamics and state dependent noise: w.p. 1 and weak convergence. *J. Math. Anal. Appl.*, 81(2):524 – 542, 1981.

[24] H. J. Kushner and D. Clark. *Stochastic Approximation for constrained and unconstrained systems*. Springer-Verlag, 1978.

[25] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, 2003.

[26] D. Lamberton and G. Pagès. On the critical points of the 1-dimensional Competitive Learning Vector Quantization Algorithm. In *ESANN'1996 proceedings - European Symposium on Artificial Neural Networks*, 1996.

[27] S. Laruelle and G. Pagès. Stochastic approximation with averaging innovation applied to Finance. *Monte Carlo Methods Appl.*, 18(1):1 – 52, 2012.

[28] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE. Trans. Commun. Technol.*, COM-28:84–85, 1980.

[29] S.P. Lloyd. Least-square quantization in PCM. *IEEE Trans. Inf. Theor.*, 28:129–137, 1982.

[30] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[31] P. Milasevic and G.R. Ducharme. Uniqueness of the spatial median. *The Annals of Statistics*, 15(3):1332–1333, 1987.

[32] G. Pagès. A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89(1):1–38, 1998.

[33] G. Pagès, H. Pham, and J. Printems. Optimal quantization methods and applications to numerical problems in finance . In S.T. Rachev and G.A. Anastassiou, editors, *Handbook on Numerical Methods in Finance*, pages 253–298. Birkhäuser, Boston, MA, 2004.

[34] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400 – 407, 1951.

[35] H. Robbins and D. Siegmund. A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications. In *Herbert Robbins Selected Papers* , pages 111–135. 1985.

[36] D. Ruppert. *Handbook of Sequential Analysis*, chapter Stochastic Approximation, pages 503–529. Marcel Dekker, New-York, 1991.

[37] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29:373–405, 1958.

[38] E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Ann. Appl. Probab.*, 20(6):2178–2203, November 11 2010.

[39] A. Schreck, G. Fort, and E. Moulines. Adaptive Equi-Energy Sampler : Convergence and Illustration. *ACM Trans. Model. Comput. Simul*, 23(1):5:1–5:27, 2013.

[40] J.C. Spall. *Encyclopedia of Electrical and Electronics Engineering*, chapter Stochastic Optimization: Stochastic Approximation and Simulated Annealing, pages 529–542. Wiley, New-York, 1999.

[41] J.C. Spall. *Introduction to Stochastic Search and Optimization*. Wiley, 2003.

[42] H. Steinhaus. Sur la division des corps materiels en parties. *Bull. Acad. Polon. Sci*, 4(12):801–804, 1956.

[43] V. Tadić. Stochastic approximation with random truncations, state-dependent noise and discontinuous dynamics. *Stochastics Stochastics Rep.*, 64:283 –326, 1998.

[44] L. Younes. On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics Stochastics Rep.*, 65:177 – 228, 1999.