General Results on the Convergence of Stochastic Algorithms

Bernard Delyon
IRISA/INRIA, campus de Beaulieu, 35042 RENNES Cedex, France *

Abstract

A deterministic approach is proposed for proving the convergence of stochastic algorithms of the most general form, under necessary conditions on the input noise, and reasonable conditions on the (non-necessarily continuous) mean field. Emphasis is made on the case where more than one stationary point exist. We use also this approach to prove the convergence of stochastic algorithm with Markovian dynamics.

1 Introduction

The general structure of stochastic algorithms is the following :

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) \tag{1}$$

γ_n is a non-negative decreasing sequence, typically $1/n$ (or $1/n^{2/3}$ when an averaging technique is used, cf [18]), X_n is a "somehow stationary" sequence and θ_n is at step n the estimated solution of

$$E[H(\theta, X)] = 0$$

where the expectation is taken over the distribution of X.

Stochastic algorithms have a wide range of application in recursive system identification, adaptive filtering, pattern recognition, adaptive learning [20], sequential change detection ... A typical list of interesting examples may be found in [1]. Study of stochastic algorithms is generally restricted to those which have the markovian structure introduced in [2] and considered in section 3, where X_n is a random variable depending only on θ_{n-1} and on X_{n-1} (i.e. typically $X_n = f(\theta_{n-1}, X_{n-1}, Y_n)$ for some function f and i.i.d variables Y_n; cf equation (15)); in this case, the expectation above has to be taken with respect to the stationary distribution of the Markov chain X_n when θ is fixed.

As a simple typical example, consider the Robbins-Monro procedure for dosage : for each dosage θ of some chemical product, one makes an experiment that delivers some effect X depending on θ, i.e. X is a random variable whose distribution depends on θ; you want to find the value of θ for which the mean effect is α :

$$E_\theta[X] = \alpha.$$

Robbins-Monro procedure consists in making a series of experiments with changing values of θ according to

$$\theta_n = \theta_{n-1} + \frac{1}{n}(\alpha - X_n)$$

where X_n is the result of the experiment made with θ_{n-1}. One can prove that under reasonable assumptions, θ_n converges to the desired solution (cf [8]).

*delyon@irisa.fr

Following Kushner and Clark, the study of stochastic algorithms is made by rewriting equation (1) as

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n \tag{2}$$

where $h(\theta)$ is the deterministic function of θ we want to cancel ($\alpha - E_\theta[X]$ in the example above) and $\eta_n = H(\theta_{n-1}, X_n) - h(\theta_{n-1})$ is a perturbation (the martingale increment $\eta_n = E_{\theta_{n-1}}[X_n] - X_n$ in the example above). The proof of convergence is then made in two steps

1/ find general conditions on a deterministic vector-valued sequence (η_n) and on h which ensure the convergence of (θ_n)

2/ show that these conditions are satisfied with probability 1 (i.e. for almost each sample path $(\eta_0(\omega), \eta_1(\omega), ...)$) of the stochastic system which is considered.

The conditions given by Kushner and Clark for convergence of (θ_n) when (η_n) is a given sequence are essentially (cf [10, 2, 3])

- the sequence (θ_n) is bounded and returns infinitely often to a compact domain of attraction of one stationary point of the continuous vector field h
- $$\lim_{n \longrightarrow \infty} \left(\sup_{n < k \leq m(n)} \left| \sum_{i=n+1}^{k} \gamma_i \eta_i \right| \right) = 0 \tag{3}$$
  where
  $$m(n) = \max\{k : \ \gamma_n + ... + \gamma_k \leq 1\}.$$
- $\gamma_n \longrightarrow 0$, $\sum \gamma_n = +\infty$

(cf. assumption A2.2.4. p. 28 of [10]). The basic difficulty with these conditions lies generally with the first one (the recurrence of (θ_n)); in most case, the second one is proved using the standard sample path properties of martingales.

If h has only one stationary point, the recurrence condition on (θ_n) generically reduces to the boundedness of (θ_n).

If there is more than one stationary point, it is very difficult to prove the convergence : the requirement of returning infinitely often to a compact set contained in the domain of attraction of one stationary point is strictly stronger than boundedness since one has to prove that (θ_n) does not go infinitely often from one domain of attraction to another. For instance, up to our knowledge, the blind equalizer problem (cf section 4.1) has never been solved completely.

In the case where stationary points are not isolated (this happens, for instance if the algorithm researches eigenvectors of a matrix with multiple eigenvalues as in section section 4.2 below (cf [5, 19, 16])), there exist results about the convergence to the *set* of stationary points (cf [15]) but there does not exist, up to our knowledge, a general theory of convergence to a point within this set.

We should also refer to [15] where convergence to the set $\{h = 0\}$ is proved under stringent assumptions on the random sequence η_n (theorem V.2.3). However, the convergence to a particular point of $\{h = 0\}$ is not generally proved. Moreover, the Lyapunov function V is required to have bounded continuous second derivatives, a condition which helps for the boundedness of θ_n but may be quite difficult to check in particular practical cases.

The aim of this paper is to give reasonable conditions on the field h under which the condition that θ_n *does not tend to infinity* will guarantee the convergence to some limit $\theta^* \in \{h = 0\}$ *even if this set contains several or non-isolated stationary points, stable or unstable.* The identification of this limit is not in the scope of this paper (especially checking that θ^* is not an unstable point of h if η_n brings in enough excitation; cf [13, 17, 16, 4]; see also [15] section V.4). Note that the vector field h is allowed here to have discontinuities. In the case where stationary points are not isolated, an extra condition will be required on h (assumption (B)).

Our basic convergence result (theorem 2) is completely deterministic and works in all the cases cited above. The condition (3) on η_n is changed into an equivalent one which is much easier to handle and gives the key to the treatment of difficult cases (cf the introduction of

§2). An asymptotic formula for the estimation error is given which may directly lead, in a stochastic setting, to the central limit-theorem for the estimation error as well as large deviation estimates. In section 2.2, we extend our results to the case where a projection is used (in this situation, the boundedness condition on θ_n is automatically satisfied). In section 3, we show how our convergence result can easily be encapsulated into the general markovian setting for these algorithms: in other words, we give conditions under which assumptions of theorem 2 are satisfied with probability one and we obtain almost sure convergence results. Application examples are given in section 4.

2 Deterministic convergence results

In this section we give convergence results in a deterministic context, i.e. (η_n) is a sequence of vectors of $\mathbf{R}^d$. They will be applied with probability one to stochastic algorithms in next section.

In what follows, condition (3) will be replaced by the following condition: The perturbation η_n may be written in the form $\eta_n = e_n + r_n$ where

- the series $\sum \gamma_n e_n$ converges
- r_n tends to zero.

Note that, if θ_n converges to a stationary point of h, we can write $\eta_n = e_n + r_n$, with

$$\begin{aligned} r_n &= -h(\theta_{n-1}) \\ e_n &= \eta_n + h(\theta_{n-1}) = (\theta_n - \theta_{n-1})/\gamma_n. \end{aligned}$$

Thus this assumption on η_n is also necessary, at least from a theoretical point of view (since this decomposition is non-unique and not necessarily obvious); one can show that the existence of such a decomposition is equivalent to (3). It turns out that in practice e_n appears typically as a martingale increment and r_n as a small remainder (cf the example of section 4.2).

2.1 General convergence results

The vector field h is defined on an open set $\mathcal{O} \subset \mathbf{R}^d$; the boundary of $\mathcal{O}$ is denoted by $\partial\mathcal{O}$. The following assumptions will be considered :

(A) *h is a vector field defined on $\mathcal{O}$, bounded on compact subsets of $\mathcal{O}$, such that*

(a) *there exist a non-negative C^1 function V_1 and a compact set $\mathcal{K} \subset \mathcal{O}$ such that*

- *$V_1(x)$ tends to $+\infty$ if $x \to \partial\mathcal{O}$ or $|x| \to \infty$*
- *one of those two conditions is satisfied*
  - *h is continuous, and $\langle \nabla V_1(x), h(x)\rangle < 0$ if $x \notin \mathcal{K}$*
  - *for any compact subset $\mathcal{C}$ of $\mathcal{O}\backslash\mathcal{K}$, $\max_{x\in\mathcal{C}}\langle \nabla V_1(x), h(x)\rangle < 0$*

(b) *there exist a non-negative C^1 function V_2 and a closed set $\mathcal{S} \subset \mathcal{O}$ such that*

- *one of these two conditions is satisfied*
  - *h is continuous, and $\langle \nabla V_2(x), h(x)\rangle < 0$ if $x \notin \mathcal{S}$*
  - *for any compact subset $\mathcal{C}$ of $\mathcal{O}\backslash\mathcal{S}$, $\max_{x\in\mathcal{C}}\langle \nabla V_2(x), h(x)\rangle < 0$*
- *$V_2(\mathcal{S})$ is a finite set.*

Assumption (A) is based on the existence of Lyapunov functions (possibly the same) V_1 and V_2 for the field h; this assumption is theoretically non-restrictive because it is strongly related to the convergence of the trajectories of h (see [11], theorem 5.3). It is more general than conditions on the convergence of trajectories of h because it does not require the existence of such trajectories. If h is a gradient field, $h = -\nabla Q$, the choice $V_1 = V_2 = Q$ will work if Q is

C^1 and tends to ∞ in the vicinity of $\partial\mathcal{O}$. Assumption (Aa) will guarantee the boundedness of (θ_n) and assumption (Ab) will ensure to convergence to $\mathcal{S}$ (typically $\mathcal{S} = \{x : h(x) = 0\}$). Note that $\mathcal{S}$ is not required to be bounded. A stronger but simpler version of (A) is

(A') *h is a continuous vector field defined on $\mathcal{O}$, such that $\mathcal{S} = \{x : h(x) = 0\}$ is a compact subset of $\mathcal{O}$, and there exist a non-negative C^1 function V such that*

- *$V(x)$ tends to $+\infty$ if $x \to \partial\mathcal{O}$ or $|x| \to \infty$*
- *$\langle \nabla V(x), h(x)\rangle < 0$ $x \notin \mathcal{S}$.*

Next assumption is used only in the cases where $\mathcal{S}$ is not a set of isolated points:

(B) *$\mathcal{S}$ is a closed set which has a neighbourhood $\mathcal{N}$ where h is uniformly Lipschitz and there exist two uniformly Lipschitz functions π, $\mathbf{R}^d$-valued, and W, $\mathbf{R}$-valued, defined in $\mathcal{N}$, which satisfy*

- $|\pi(x_t) - \pi(x_s)| \le W(x_s) - W(x_t)$ *for any trajectory of h, $(x_u)_{s\le u\le t}$, contained in $\mathcal{N}$,*
- $\pi(x) = x$ *if* $x \in \mathcal{S}$.

The choice $W = 0$ and $\pi(x) = x_\infty$ (limit point of the trajectory sarting from x) may work; this is the case in section 4.2. Note that equation satisfied by W implies that this function is non-increasing on the trajectories of h.

Assumption (B) forces the stationary set $\mathcal{S}$ to be attractive enough. For instance, if the field tends to be tangential to $\mathcal{S}$ in its neighbourhood, an arbitrarily small perturbation η_n in (2) may lead to infinitely long trajectories spiralling closer and closer around $\mathcal{S}$. A typical example where (B) is not satisfied is the two dimensional case where the field h has the form $h(x_1, x_2) = (x_2, 0)^T$ on $[-1, 1]^2$; in this case, taking $\gamma_n = 1/n$, in equation (2), θ_n can move from $(-1, \epsilon)$ to $(1, \epsilon)$ without any perturbation ($\eta_n = 0$), then move in one step to $(1, -\epsilon/2)$ with a perturbation $\gamma_n\eta_n$ of amplitude $3\epsilon/2$ then go back to $(-1, -\epsilon/2)$ without any perturbation, then cross the x-axis to $(-1, -\epsilon/4)$ with a $\gamma_n\eta_n$ of amplitude $3\epsilon/4$ etc... This algorithm does not converge.

We shall consider here deterministic algorithms of the form

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1} + \delta_n) + \gamma_n e_n + \gamma_n r_n, \qquad \gamma_n \ge 0, \qquad \sum_{p=1}^{\infty} \gamma_p = \infty, \qquad \lim_{n\longrightarrow\infty} \gamma_n = 0, \tag{4}$$

where $n \ge 1$, θ_0 is given, h is a vector field on $\mathcal{O} \subset \mathbf{R}^d$, e_n, r_n and δ_n are perturbations, and γ_n is a non-negative scalar gain sequence. δ_n is such that $\theta_n + \delta_{n+1} \in \mathcal{O}$ and may, for instance, take into account the fact that some algorithms use $h(\theta_{n-p})$, where p varies between two constants, rather than $h(\theta_n)$ (in the sequential form of block algorithms for example); in most cases however $\delta_n = 0$.

Before stating the theorem, we need a definition

Definition 1 *We say that the algorithm is A-stable if*

$$\theta_n \quad \textit{remains in a compact subset of} \quad \mathcal{O} \tag{5}$$

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n \quad \textit{exists} \quad \textit{and} \quad \lim_{n\to\infty} |r_n| + |\delta_n| = 0 \tag{6}$$

A-stability is the basic assumption for convergence of θ_n (cf theorem 2); the following theorem says that if θ_n does not tend to infinity and if e_n, r_n, δ_n are well controlled when θ_n is in a prescribed compact set, then A-stability occurs :

Theorem 1 *We assume that assumption (Aa) (or (A')) is satisfied and the existence of a compact subset $\mathcal{K}_0 \subset \mathcal{O}$ such that*

$$\theta_n \in \mathcal{K}_0 \quad \textit{infinitely often} \tag{7}$$

We assume that for any $M \in \mathbf{N}$,

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n \quad \textit{exists} \quad \textit{or} \quad \lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n 1_{V_1(\theta_{n-1})\le M} \quad \textit{exists} \tag{8}$$

$$\lim_{n\to\infty} (|r_n| + |\delta_n|) 1_{V_1(\theta_{n-1})\le M} = 0. \tag{9}$$

Then the algorithm (4) is A-stable.

Condition (7) is generally the most difficult one to check; this problem is considered in next section. In a stochastic context, it is almost surely satisfied as soon as the sequence (θ_n) has a uniformly bounded moment.

The second form of condition (8) is particularly adapted to the case where e_n is some sort of martingale increment whose variance depends on θ_n (for example in the case of the Robbins-Monro procedure (cf introduction and [15, 8, 10, 1, 2])).

Next result gives convergence properties:

Theorem 2 *We assume that the algorithm is A-stable.*

- *Under assumption (Ab) (or (A')), the distance of θ_n to the set $\mathcal{S}$, $d(\theta_n, \mathcal{S})$, converges to 0; in particular, if $\mathcal{S}$ is a finite set, (θ_n) converges to some point of $\mathcal{S}$.*
- *If $\mathcal{S}$ is a set satisfying (B), such that $d(\theta_n, \mathcal{S})$ tends to zero, and if*

$$\sum_p \gamma_p |\delta_p - \sum_{i=p}^{\infty} \gamma_i e_i| + \gamma_p^2 + \gamma_p |r_p| < \infty, \tag{10}$$

then (θ_n) converges to some point of $\mathcal{S}$.

- *If (θ_n) converges to some limit θ_∞, if h is C^1 in some neighbourhood of θ_∞ with first derivatives Lipschitz, and if the derivative of h at θ_∞, say h'_∞, is a matrix with all eigenvalues with negative real part then*

$$\theta_n - \theta_\infty = \mu_n + O\left(e^{-\lambda s_n} + \sum_1^n \gamma_i e^{-\lambda(s_n - s_i)} (\gamma_i^2 + |\delta_i|^2 + |\mu_i|^2)\right)$$

where

$$\mu_n = \sum_1^n e^{(s_n - s_i)h'_\infty} \gamma_i (e_i + h'_\infty \delta_i + r_i),$$

$$-\lambda \quad \textit{is the largest real part of the eigenvalues of } h'_\infty,$$

$$s_n = \gamma_1 + \gamma_2 + .. + \gamma_n$$

and μ_n tends to zero.

This theorem contains three independent results and the only common assumption is the A-stability; in particular, the set $\mathcal{S}$ of assumption (A) is not necessarily the same as the one of assumption (B).

We give in Appendix 1 a result which allows to obtain bounds for those series of exponentials.

We do not have space here to prove central-limit results but we just would like to point out that this can be done without much difficulty by using the last statement of the theorem 2: setting $A_n = E[\mu_n \mu_n^T]^{-1/2}$, one has to prove that

- the term $A_n \mu_n$ tends in distribution to the normal variable (cf [8] theorem 3.2)
- the inside of the $O()$ term is $o(\|A_n\|^{-1})$ in mean (the tools of Appendix 1 may be useful).

These two facts imply the convergence in distribution of $A_n(\theta_n - \theta_\infty)$ to the normal law. We won't carry over the calculations which are quite standard and depend on each particular cases (see also [15] theorem 6.1 of chapter VI). In a similar way, large deviation results can be inferred.

2.2 Stabilization and projection

It turns out that, when an algorithm is given, the most difficult point to check is that θ_n does not tend to infinity (in order to have the A-stability). It would be very helpful to have a general stability result for dealing with this particular point; this seems rather difficult without drastic assumptions on the Lyapunov function V_1. On the other hand, practical implementations are generally performed with some projection which brings back θ at each step into a safe region where we know the true parameter lies. We show here that this can be included into our formalism, by remarking that a projection is nothing but a modification of the field H outside of the projection set.

Suppose we start with an algorithm written in the form

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n)$$

where H is a function $\mathbf{R}^d \times \mathbf{R}^l \longrightarrow \mathbf{R}^d$. If we know that $\mathcal{S}$ (or part of it) is included in some bounded set $\mathcal{Q}$, we can modify the algorithm by using a projection π on $\mathcal{Q}$ and consider the new algorithm

$$\theta_n = \pi(\theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n)) \tag{11}$$

hoping that this algorithm will be more stable and will converge faster. We shall show that under suitable conditions the algorithm converges. Let us rewrite the algorithm (11); the sequence $(\bar{\theta}_p)$, consisting of (θ_n) with insertion of the unprojected estimates, satisfies the equations

$$\bar{\theta}_p = \bar{\theta}_{p-1} + \gamma_{n_p} H(\bar{\theta}_{p-1}, X_{n_p}), \qquad n_p = n_{p-1} + 1, \text{ if } \bar{\theta}_{p-1} \in \mathcal{Q} \tag{12}$$

$$\bar{\theta}_p = \bar{\theta}_{p-1} + \gamma'_p \frac{\pi(\bar{\theta}_{p-1}) - \bar{\theta}_{p-1}}{|\pi(\bar{\theta}_{p-1}) - \bar{\theta}_{p-1}|}, \qquad \gamma'_p = |\pi(\bar{\theta}_{p-1}) - \bar{\theta}_{p-1}| \text{ if } \bar{\theta}_{p-1} \notin \mathcal{Q}. \tag{13}$$

The gain sequence of this algorithm is the orginal sequence γ_n with some insertions of γ'_p; it tends to zero if π is continuous and $\gamma_n H(\theta_{n-1}, X_n)$ tends to zero (in most situations, this condition is easily checked because θ_{n-1} is bounded). The corresponding mean vector field is

$$\bar{h}(x) = h(x) 1_{x \in \mathcal{Q}} + \frac{\pi(x) - x}{|\pi(x) - x|} 1_{x \notin \mathcal{Q}}.$$

This change implies that we have to modify accordingly the assumption (Ab); since we shall assume that $\gamma_{p_n} H(\theta_{n-1}, X_n)$ tends to 0, θ_n will tend to $\mathcal{Q}$, and one can take $\mathcal{O}$ as an arbitrary neighbourhood of $\mathcal{Q}$; the new assumption is thus

(PJ) *$\mathcal{Q}$ is a bounded set, π is a continuous function such that $\pi(x) \in \mathcal{Q}$ for all x and $\pi(x) = x$ if $x \in \mathcal{Q}$. There exists a neighbourhood $\mathcal{O}$ of $\mathcal{Q}$ where (Ab) is satifsfied and for some $\delta > 0$*

$$\langle \nabla V_2, \pi(x) - x \rangle < -\delta |\pi(x) - x|, \quad \textit{for} \ \ x \in \mathcal{O} \backslash \mathcal{Q}.$$

Rephrasing theorem 2 in this context leads to the

Theorem 3 *Consider θ_n given by algorithm (11) and assume that*

- $\sum \gamma_n = \infty$ *and* $\gamma_n \longrightarrow 0$
- $\gamma_n H(\theta_{n-1}, X_n)$ *tends to zero*
- $H(\theta_{n-1}, X_n)$ *may be rewritten as* $h(\theta_{n-1} + \delta_n) + e_n + r_n$ *where* e_n, r_n *and* δ_n *satisfy the condition*

$$\lim_{p \to \infty} \sum_{n=1}^{p} \gamma_n e_n \quad \textit{exists} \quad \textit{and} \quad \lim_{n \to \infty} |r_n| + |\delta_n| = 0$$

- *assumption (PJ) is satisfied*

then θ_n converges to $\mathcal{S}$.

Remarks: The set $\mathcal{Q}$ is not required to have non-empty interior, for instance it may reduce to the unit sphere. Generally however, $\mathcal{Q}$ will be more like some large ball and projection π will be done a finite number of times (because of the convergence of θ_n inside). In this case, once convergence is obtained, one can apply theorem 2 for getting estimates of the rate of convergence.

3 Application to algorithms with Markovian dynamics

We consider the model introduced in of [2, 1]; examples are given in these references (see also the stochastic EM algorithm of [12]). The stochastic algorithm is defined by the following equation

$$\theta_n = \pi\left(\theta_{n-1} + \gamma_n \psi_n(H(\theta_{n-1}, X_n))\right) \tag{14}$$

where $H(\theta, X)$ is a function $\mathbf{R}^d \times \mathbf{R}^l \longrightarrow \mathbf{R}^d$, π is a projection on some compact set $\mathcal{Q}$ of $\mathbf{R}^d$, and X_{n+1} is randomly chosen from X_n through the transition probability $\Pi_{\theta_n}(x, dy)$:

$$P(X_{n+1} \in B | \theta_0, ..\theta_n, X_0, ..X_n) = \Pi_{\theta_n}(X_n, B). \tag{15}$$

θ_0 and X_0 are arbitrary; ψ_n is a projection on a ball of increasing radius which is necessary only if Π_θ depends effectively on θ and X, and H is unbounded (cf the assumptions in the theorem below). This form encompasses most algorithms (cf [1]).

We shall show in the proof that the projection ψ_n is made only a finite number of times. We see many advantages to these projections: the proof is much simpler with less stringent assumptions (compare with [1]) and the algorithm is more robust and closer to what is done in practice.

The proof requires some assumptions on the Markov chains parametrized with θ which are not very restrictive but not always easy to check (examples are given in [2]); they are essentially borrowed to [2]:

(MK1) *for every $\theta \in \mathcal{Q}$, the Markov chain has a unique invariant probability μ_θ and there exist $\alpha < 1$, $q \geq 1$, $n_0 > 0$ and C such that for all $x \in \mathbf{R}^l$, $\theta, \theta' \in \mathcal{Q}$*

$$\begin{aligned}
\int \Pi_\theta^{n_0}(x, dy)|y|^q &\leq \alpha |x|^q + C \\
\int \Pi_\theta(x, dy)|y|^q &\leq C(1 + |x|^q) \\
\int |y|^q \left|\Pi_\theta(x, dy) - \Pi_{\theta'}(x, dy)\right| &\leq C|\theta - \theta'|(1 + |x|^q).
\end{aligned}$$

(MK2) *H satisfies for some $p \leq q/2$*

$$|H(\theta, x)| \leq C(1 + |x|^p), \quad x \in \mathbf{R}^l, \theta \in \mathcal{Q}$$

and we set $h(\theta) = \int H(\theta, x)\mu_\theta(dx)$.

(MK3) *for every $\theta \in \mathcal{Q}$, the Poisson equation*

$$(I - \Pi_\theta)v_\theta = H(\theta, .) - h(\theta)$$

has a solution which satisfies for some $C > 0$ and the same p as in assumption (MK2)

$$\begin{aligned}
&|v_\theta(x)| \leq C(1 + |x|^p) \\
&|\Pi_\theta v_\theta(x) - \Pi_{\theta'} v_{\theta'}(x)| \leq C|\theta - \theta'|(1 + |x|^p).
\end{aligned}$$

For each θ, the existence of v_θ and the first bound of (MK3) is generally a consequence of the first two assumptions (cf [14] section 17.4). The most efficient tool for verifying last equation of (MK3) is geometric ergodicity (cf theorems 15.0.1(iii), 6.0.1(iii), 6.2.5(ii) of [14]); this theory allows to write v_θ as a rapidly converging series

$$v_\theta(.) = \tilde{H}(\theta,.) + \Pi_\theta \tilde{H}(\theta,.) + \Pi_\theta^2 \tilde{H}(\theta,.) + ... \qquad \tilde{H} = H - h.$$

Indications for a direct proof of these conditions (i.e. without explicit computation of v_θ) are given in [7] section 5. We have the following result

Theorem 4 *Assume that θ_n is given by algorithm (14) and assume that*

(i) *assumptions (MK1-3) and (PJ) are satisfied (h is given by (MK2)),*

(ii) $\gamma_n \geq 0$, $\sum \gamma_n = \infty$, $\sum \gamma_n^2 < \infty$, *and* $\sum |\gamma_n - \gamma_{n+1}| < \infty$,

(iii) *if the kernel $\Pi_\theta(x, dy)$ does not depends on θ or $n_0 = 1$ or H is bounded, then $\psi_n(x) = x$, else*

- *for some $C > 0$ and $\eta < 1$: $\sum \gamma_n^{\eta q/p} < \infty$ and $\psi_n(x) = x$ if $|x| < C\gamma_n^{-\eta}$*
- $\lim_{n\longrightarrow\infty} \gamma_n \|\psi_n\|_\infty = 0$,

where p and q are the exponents of assumptions (MK1-3),

(iv) *π is Lipschitz,*

then θ_n converges almost surely to $\mathcal{S}$. Furthermore, the sequence X_n is bounded in L_q and the projections ψ_n are made a finite number of times.

Remarks: If Π_θ does not depend on θ, X_n is a Markov chain in the usual sense ; if Π_θ does not depend on x, then $n_0 = 1$ and we are in the Robins-Monro case ; in these two situations, the projections ψ_n are unnecessary.

Typical values are for instance $p = 1$, $q = 2$, $\eta = 3/4$, and ψ_n is the projection on the ball of radius $\gamma_n^{3/4}$. One can show that the projection ψ_n can be replaced by a freezing of θ_n in H :

$$\theta_n = \pi\left(\theta_{n-1} + \gamma_n(H(\theta_{[n/n_0]n_0-1}, X_n))\right)$$

This last algorithm has the drawback of requiring the knowledge of n_0. The convergence may be proved by using the general form of the algorithm (4) with $\delta_n = \theta_{[n/n_0]n_0-1} - \theta_{n-1}$; we shall not do it here.

Proof The proof reuses the ideas of [2], but here, the use of the projections greatly simplifies the task. The basic idea to show that the projection ψ_n is made only a finite number of times and then to decompose $H(\theta_n, X_{n+1})$ by using Poisson equation and making appear four terms, the deterministic drift, one martingale increment and two remainders :

$$\begin{aligned} H(\theta_n, X_{n+1}) &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_n) \\ &\quad + \Pi_{\theta_n} v_{\theta_n}(X_n) - \Pi_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) \\ &\quad + \Pi_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= h(\theta_n) + e_{n+1}^{(1)} + e_{n+1}^{(2)} + e_{n+1}^{(3)}. \end{aligned}$$

We shall finally prove that the series $\sum \gamma_n(e_n^{(1)} + e_n^{(2)} + e_n^{(3)})$ is a.s. convergent.

Let us prove first that the sequence X_n is bounded in L_q. If the projections ψ_n are used (or if H is bounded), we set

$$\begin{aligned} u_n &= \gamma_n \sup_{\theta, X} |\psi_n(H(\theta, X))| \\ v_n &= u_{n-1} + u_{n-2} + .. + u_{n-n_0}. \end{aligned}$$

and note that since π is Lipschitz :

$$|\theta_n - \theta_{n-1}| = |\pi\left(\theta_{n-1} + \gamma_n \psi_n(H(\theta_{n-1}, X_n))\right) - \pi(\theta_{n-1})| \leq C u_n.$$

Then

$$
\begin{aligned}
E[|X_n|^q|\mathcal{F}_{n-1}] &= \int |y|^q \Pi_{\theta_{n-1}}(X_{n-1},dy) \\
&= \int |y|^q \Pi_{\theta_{n-n_0}}(X_{n-1},dy) + \int |y|^q (\Pi_{\theta_{n-1}}(X_{n-1},dy) - \Pi_{\theta_{n-n_0}}(X_{n-1},dy)) \\
&\le \int |y|^q \Pi_{\theta_{n-n_0}}(X_{n-1},dy) + C(1+|X_{n-1}|^q)|\theta_{n-1}-\theta_{n-n_0}| \\
&\le \int |y|^q \Pi_{\theta_{n-n_0}}(X_{n-1},dy) + C' v_n (1+|X_{n-1}|^q) \\
E[|X_n|^q|\mathcal{F}_{n-2}] &\le \int\int |y|^q \Pi_{\theta_{n-n_0}}(z,dy)\Pi_{\theta_{n-2}}(X_{n-2},dz) + v_n C' \int (1+|z|^q)\Pi_{\theta_{n-2}}(X_{n-2},dz) \\
&\le \int |y|^q \Pi^2_{\theta_{n-n_0}}(X_{n-2},dy) + v_n C''(1+|X_{n-2}|^q)
\end{aligned}
$$

and after n_0 iterations, we obtain

$$
\begin{aligned}
E[|X_n|^q|\mathcal{F}_{n-n_0}] &\le \int |y|^q \Pi^{n_0}_{\theta_{n-n_0}}(X_{n-n_0},dy) + v_n C'''(1+|X_{n-n_0}|^q) \\
&\le \alpha |X_{n-n_0}|^q + v_n C'''(1+|X_{n-n_0}|^q) + M \\
E[|X_n|^q] &\le (\alpha + v_n C''')E[|X_{n-n_0}|^q] + v_n C''' + M.
\end{aligned}
$$

Since v_n tends to zero, this inequality implies that (X_n) is bounded in L_q. If $\psi_n(x) = x$ (no projection), then $n_0 = 1$ or Π_θ does not depend on θ and the same conclusion is immediate.

Now assumption (MK2) implies that $\sum \gamma_n^{\eta q/p}|H(\theta_{n-1},X_n)|^{q/p}$ is a.s. finite (since the expectation is finite), hence $\gamma_n^\eta H(\theta_{n-1},X_n)$ tends a.s. to zero, and the projection ψ_n is almost surely made only a finite number of times. It remains to prove that the series $\sum \gamma_n(e_n^{(1)}+e_n^{(2)}+e_n^{(3)})$ converges a.s.

Assumptions (MK1-3) imply that for some C'

$$|\Pi_\theta v_\theta(x)| \le |v_\theta(x)| + |\Pi_\theta H(\theta,x)| + |h(\theta)| \le C'(1+|x|^p).$$

Thus, since $2p \le q$, we deduce that $(e_n^{(1)})$ is an L_2-bounded martingale-increment sequence and that the martingale $\sum \gamma_i e_i^{(1)}$ converges. We have

$$
\begin{aligned}
\sum_{n=1}^{l} \gamma_n e_n^{(2)} &= \sum_{n=1}^{l} \gamma_n (\Pi_{\theta_{n-1}} v_{\theta_{n-1}}(X_{n-1}) - \Pi_{\theta_n} v_{\theta_n}(X_n)) \\
&= \sum_{n=1}^{l-1} (\gamma_{n+1}-\gamma_n)\Pi_{\theta_n} v_{\theta_n}(X_n) + \gamma_1 \Pi_{\theta_0} v_{\theta_0}(X_0) - \gamma_l \Pi_{\theta_l} v_{\theta_l}(X_l)
\end{aligned}
$$

and for proving the convergence of this series, note that

$$\sum_{n=1}^{\infty} |(\gamma_{n+1}-\gamma_n)\Pi_{\theta_n} v_{\theta_n}(X_n)| \le C\sum_{n=1}^{\infty}(1+|X_n|^p)|\gamma_n - \gamma_{n+1}| < \infty \quad a.s.$$

(since the expectation is finite) and that

$$\sum_{l=1}^{\infty} \gamma_l^2 |\Pi_{\theta_l} v_{\theta_l}(X_l)|^2 \le C \sum_{l=1}^{\infty} \gamma_l^2 (1+|(X_l)|^p)^2 < \infty \quad a.s.$$

For the last remainder we obtain

$$\sum_{n=1}^{\infty} \gamma_n |e_n^{(3)}| = \sum_{n=1}^{\infty} \gamma_n |\Pi_{\theta_n} v_{\theta_n}(X_n) - \Pi_{\theta_{n-1}} v_{\theta_{n-1}}(X_n)|$$

$$
\begin{aligned}
&\le C' \sum_{n=1}^{\infty} \gamma_n |\theta_n - \theta_{n-1}|(1+|X_n|^p) \\
&\le C'' \sum_{n=1}^{\infty} \gamma_n^2 |H(\theta_{n-1}, X_n)|(1+|X_n|^p) \\
&\le C''' \sum_{n=1}^{\infty} \gamma_n^2 (1+|X_n|^q) \\
&< \infty \ a.s.
\end{aligned}
$$

Finally we have written

$$H(\theta_n, X_{n+1}) = h(\theta_n) + e_{n+1}$$

as required in theorem 3 and the convergence is proved. ∎

We could as well have treated the more general case

$$\theta_n = \pi\left(\theta_{n-1} + \gamma_n \psi_n (H(\theta_{n-1}, X_n) + e_n + r_n)\right)$$

where e_n and r_n satisfy with probability one equation (6) and are such that the Markov property is preserved :

$$P(X_{n+1} \in B | \mathcal{F}_n) = \Pi_{\theta_n}(X_n, B)$$

where

$$\mathcal{F}_n = \sigma(\theta_0, X_0, e_1, r_1, ...\theta_n, X_n, e_n, r_n).$$

4 Examples

We give in this section two examples where the vector field h has more than one stationary point.

4.1 Blind equalizers

The algorithm has the form (cf [1] p.64)

$$\theta_n = \pi(\theta_{n-1} + \gamma_n Y_n(\alpha \ \mathrm{sign}(Y_n^T \theta_{n-1}) - Y_n^T \theta_{n-1})) \tag{16}$$

where $Y_n = (y_n, ..y_{n-p})$ the received signal :

$$y_n = \sum_{i=0}^{q} c_i a_{n-i} + \xi_n,$$

(c_i) is the channel filter, (a_i) is the emitted symbol sequence (a zero-mean i.i.d. sequence with finite number of possible values), ξ_n an i.i.d. gaussian white noise sequence, independent of the a_n's, and

$$\alpha = \frac{E[|a_i|^2]}{E[|a_i|]}.$$

The projection π is made on a ball of radius A large enough : $\pi(\theta) = A\theta/|\theta|$ if $|\theta| > A$.

The aim of the algorithm is to find a vector θ^* which minimizes

$$V(\theta) = E[|Y_n^T \theta - \alpha \ \mathrm{sign}(Y_n^T \theta)|^2]/2.$$

Local minima generally exist (cf [1] figure 13 and 14). One can show (cf [1]) that such a vector will be a good candidate for the inverse filter (i.e. $Y_n^T \theta^* \simeq a_{n-p}$ for some p). We are in the situation of section 3 with

$$X_n = (a_n, a_{n-1}, ..., a_{n-p-q}, \xi_n, .., \xi_{n-p})$$

and $\Pi_\theta = \Pi$ is independent of θ and satisfies

$$\Pi^{p+q} = \mu.$$

Hence

$$v_\theta(.) = \tilde{H}(\theta, .) + \Pi\tilde{H}(\theta, .) + ... + \Pi^{p+q}\tilde{H}(\theta, .), \qquad \tilde{H} = H - h.$$

Assumptions (MK1-3) are straightforwardly checked ($n_0 = p+q$, $\alpha = 0$, the regularity of $\Pi v_\theta(.)$ comes from the smoothing effect of the noise ξ). Since $h(\theta) = -\nabla V(\theta)$, V is a convenient Lyapunov function for assumption (Ab). It remains to show that (PJ) is satisfied:

$$\begin{aligned}
\langle \nabla V(\theta), \pi(\theta) - \theta \rangle &= E[\langle Y(Y^T\theta) - \alpha\ \mathrm{sign}(Y^T\theta)), (A/|\theta| - 1)\theta \rangle] \\
&= -E[|Y^T\theta|^2 - \alpha|Y^T\theta|](1 - A/|\theta|) \\
&\leq -E[|Y^T\theta|^2 - \alpha^2](1 - A/|\theta|)/2
\end{aligned}$$

and one checks easily that

$$E[|Y^T\theta|^2] = |\theta|^2(\sigma_\xi^2 + \sigma_a^2 \sum c_i^2) = |\theta|^2\sigma_y^2$$

(σ_y^2 is the variance of y_n). Hence, (PJ) is satisfied if for instance A is chosen larger than α/σ_y and we have the

Theorem 5 *Under the assumptions above, if $A > \alpha/\sigma_y$, then the algorithm (16) converges almost surely to some solution θ^* of the equation*

$$E[Y(Y^T\theta)] = \alpha E[Y\, sign(Y^T\theta)].$$

Since the distribution of a_i is generally known and the variance of Y is easily estimated, this theorem gives a simple way to obtain a reasonnable value of A.

4.2 Estimation of eigenvectors

We are given a sequence of random matrices A_k with common expectation A, a symmetric matrix; an algorithm for estimating the first eigenvector of A is studied in [16]:

$$\theta_n = \frac{\theta_{n-1} + \gamma_n A_n \theta_{n-1}}{|\theta_{n-1} + \gamma_n A_n \theta_{n-1}|}. \tag{17}$$

We shall only prove the convergence of the algorithm to some eigenvector; this is actually a first step for proving the convergence to some eigenvector associated to the largest eigenvalue of A, under more restrictive assumptions on the sequence A_k (the noise $A_k - A$ has to be rich enough; see [4, 16]). Note that in [16], it is assumed that the largest eigenvalue of has unit multiplicity.

Theorem 6 *We consider the algorithm (17) where*

$$\begin{aligned}
&\sum \gamma_i = \infty \\
&\sum \gamma_i^2 < \infty \\
&\sup_k E[||A_k||^4] < \infty \\
&E[A_k] = A.
\end{aligned}$$

If A has at least one multiple eigenvalue, we require in addition that

$$\sum_p \gamma_p (\sum_{i=p}^{\infty} \gamma_i^2)^{1/2} < \infty.$$

Then the sequence θ_n converges almost surely to some eigenvector of the matrix A.

Proof

Step1 : Rewriting equation (17) into the form (4). The assumptions imply that

$$\sum \gamma_k^2 \|A_k\|^4$$

is finite a.s. and $\gamma_k \|A_k\|^2$ tends to zero; so does $\gamma_k \|A_k\|$. Using the relation true for $|\theta| = 1$ and $|\epsilon| < 1$

$$\frac{\theta + \epsilon}{|\theta + \epsilon|} = \theta + \epsilon - (\theta^T \epsilon)\theta + O(|\epsilon|^2)$$

we obtain

$$\begin{aligned}
\theta_n &= \frac{\theta_{n-1} + \gamma_n A_n \theta_{n-1}}{|\theta_{n-1} + \gamma_n A_n \theta_{n-1}|} \\
&= \theta_{n-1} + \gamma_n A_n \theta_{n-1} - \gamma_n (\theta_{n-1}^T A_n \theta_{n-1})\theta_{n-1} + O(\gamma_n^2 \|A_n\|^2) \\
&= \theta_{n-1} + \gamma_n (A\theta_{n-1} - (\theta_{n-1}^T A \theta_{n-1})\theta_{n-1}) \\
&\quad + \gamma_n ((A_n - A)\theta_{n-1} - \theta_{n-1}^T (A_n - A)\theta_{n-1}\theta_{n-1}) \\
&\quad + \gamma_n O(\gamma_n \|A_n\|^2).
\end{aligned}$$

The correspondence with equation (4) is

$$\begin{aligned}
\mathcal{O} &= \mathbf{R}^d \backslash \{0\} \\
h(x) &= (A + \lambda I)x - (x^T (A + \lambda I)x)x \\
e_n &= (A_n - A)\theta_{n-1} - (\theta_{n-1}^T (A_n - A)\theta_{n-1})\theta_{n-1} && (18) \\
r_n &= O(\gamma_n \|A_n\|^2) && (19) \\
\delta_n &= 0.
\end{aligned}$$

The λ-term in the expression of h is added for technical reasons (it is allowed because $|\theta_n| = 1$); λ is any fixed real number such that $A + \lambda I > 0$, so that the stationary points of h are the unit eigenvectors of A.

Step2 : Checking the A-stability and (10) Since r_n tends to zero and e_n is an L_2-bounded martingale increment, the A-stability is satisfied and it remains to check equation (10). Note that

$$\begin{aligned}
E[\sum_p \gamma_p |\sum_{i=p}^{\infty} \gamma_i e_i|] &= \sum_p \gamma_p E[|\sum_{i=p}^{\infty} \gamma_i e_i|] \\
&\le \sum_p \gamma_p E[|\sum_{i=p}^{\infty} \gamma_i e_i|^2]^{1/2} \\
&= \sum_p \gamma_p (\sum_{i=p}^{\infty} \gamma_i^2 E[|e_i|^2])^{1/2} \\
&\le C \sum_p \gamma_p (\sum_{i=p}^{\infty} \gamma_i^2)^{1/2} < \infty
\end{aligned}$$

and

$$E[\sum_p \gamma_p |r_p|] \le C \sum_p \gamma_p^2 < \infty$$

Step3 : Checking (A)

Set

$$B = A + \lambda I$$

and take

$$V(x) = \frac{e^{|x|^2}}{x^T B x}.$$

Then

$$\nabla V(x) = 2V(x)x - \frac{e^{|x|^2}}{(x^T Bx)^2} 2Bx = \frac{2V(x)}{x^T Bx}((x^T Bx)x - Bx) = -\frac{2V(x)}{x^T Bx} h(x).$$

Step4: Checking (B)

We shall find a function π which is constant on the trajectories of h, in which case the choice $W = 0$ will be convenient.

Consider a neighbourhood of $\mathcal{S}$, say $\mathcal{N} = \{x, d(x, \mathcal{S}) < \epsilon\}$; if ϵ is small enough, this neighbourhood is a union of separate open neighbourhoods of each connected component of $\mathcal{S}$. Thus, each $x \in \mathcal{N}$ is associated to a specific eigenspace of A. We define

$$\pi(x) = \frac{Px}{|Px|}$$

where P is the orthogonal projection on the eigenspace associated to x. We have

$$\pi'(x) = \frac{P}{|Px|} - \frac{Pxx^T P^T P}{|Px|^3};$$

thus

$$\pi'(x)x = 0$$

and, since $BPx = \lambda Px$,

$$\pi'(x)Bx = \frac{\lambda Px}{|Px|} - \frac{Px\lambda |Px|^2}{|Px|^3} = 0.$$

Finally $\pi'(x)h(x) = 0$, π is constant on the trajectories of h and assumption (B) is satisfied. ∎

5 Proof of theorem 1

We assume first that the series $\sum \gamma_n e_n$ does not converge.

Choose $M \in \mathbf{N}$ such that the sets $\mathcal{K}_0$ and $\mathcal{K}$ (from assumption (Aa)) satisfy

$$\mathcal{K} \cup \mathcal{K}_0 \subset \{V_1 < M - 1\},$$

and set

$$\mathcal{K}_1 = \{V_1 \leq M + 1\}.$$

Because of the assumptions we can define

$$\begin{aligned} \theta'_n &= \theta_n + \sum_{i=n+1}^{\infty} \gamma_i e_i 1_{\mathcal{K}_1}(\theta_{i-1}) \\ \delta'_n &= \delta_n - \sum_{i=n+1}^{\infty} \gamma_i e_i 1_{\mathcal{K}_1}(\theta_{i-1}). \end{aligned}$$

Note that

$$\lim_{n \longrightarrow \infty} |\theta'_n - \theta_n| = 0.$$

Since θ_n returns infinitely often to $\mathcal{K}_0$, the continuity of V_1 implies that $V_1(\theta'_n) < M$ infinitely often. On the other hand, for an n such that $V_1(\theta'_{n-1}) < M$, we have, if n is large enough, $\theta_{n-1} \in \mathcal{K}_1$.

The sequence θ'_n satisfies

$$\theta'_n = \theta'_{n-1} + \gamma_n h(\theta'_{n-1} + \delta'_n) + \gamma_n e_n (1 - 1_{\mathcal{K}_1}(\theta_{n-1})) + \gamma_n r_n$$

and if $V_1(\theta'_{n-1}) < M$:

$$V_1(\theta'_n) = V_1(\theta'_{n-1}) + \gamma_n \langle \nabla V_1(\theta'_{n-1}), h(\theta'_{n-1} + \delta'_n) \rangle + \gamma_n r'_n$$

where r'_n tends to zero as n tends to infinity, because V_1 is C^1 and h is bounded on compact sets. Thus

$$V_1(\theta'_n) = V_1(\theta'_{n-1}) + \gamma_n \langle \nabla V_1(\theta'_{n-1} + \delta'_n), h(\theta'_{n-1} + \delta'_n) \rangle + \gamma_n r''_n$$

and r''_n tends also to zero. From this equation, we can conclude that, for n large enough,

- if $V_1(\theta'_{n-1}) \in [0, M - 1/2]$, then $V_1(\theta'_n) < M$
- if $V_1(\theta'_{n-1}) \in [M - 1/2, M]$, then $V_1(\theta'_{n-1} + \delta'_n) \in [M - 3/4, M + 1]$ and using assumption (Aa) with $\mathcal{C} = \{M - 3/4 \leq V_1 \leq M + 1\}$, we conclude that $V_1(\theta'_n) < V_1(\theta'_{n-1})$.

Since $V_1(\theta'_{n-1})$ is smaller than M infinitely often, this proves the boundedness of θ'_n, which in turn proves the boundeness of θ_n.

In the case where the series $\sum \gamma_n e_n$ converges, we just carry on the same proof with $\mathcal{K}_1 = \mathcal{O}$, $\theta'_n = \theta_n$, $\delta'_n = \delta_n$.

6 Proof of theorem 2

6.1 Proof of the first assertion

We consider

$$\theta'_n = \theta_n + \sum_{i=n+1}^{\infty} \gamma_i e_i \tag{20}$$

$$\delta'_n = \delta_n - \sum_{i=n+1}^{\infty} \gamma_i e_i \tag{21}$$

leading to

$$\theta'_n = \theta'_{n-1} + \gamma_n h(\theta'_{n-1} + \delta'_n) + \gamma_n r_n.$$

Recall that V_2 is C^1, h is bounded on compact subsets of $\mathcal{O}$ and $\theta'_{n-1} + \delta'_n$ remains in a compact subset of $\mathcal{O}$; hence we have

$$V_2(\theta'_n) = V_2(\theta'_{n-1}) + \gamma_n \langle \nabla V_2(\theta'_{n-1} + \delta'_n), h(\theta'_{n-1} + \delta'_n) \rangle + \gamma_n r'_n$$

where r'_n tends to zero as n tends to infinity. Consider an open neighbourhood $\mathcal{N}$ of $\mathcal{S}$, and set $\mathcal{E} = \mathcal{O} \backslash \mathcal{N}$ one can find $\epsilon, C > 0$ such that

$$V_2(\theta'_n) \leq V_2(\theta'_{n-1}) - \gamma_n \epsilon 1_{\mathcal{E}}(\theta'_{n-1}) + \gamma_n C 1_{\mathcal{N}}(\theta'_{n-1}) \tag{22}$$

for n large enough.

We shall end the proof by demonstrating first that $V_2(\theta'_n)$ converges to some point of $V_2(\mathcal{S})$, and second, that θ'_n converges to some point of $\mathcal{S}$.

Consider the neighbourhood of the finite set $V_2(\mathcal{S})$ (cf assumption (Ab)) :

$$A = \{x \in \mathbf{R} : \ d(x, V_2(\mathcal{S})) < \alpha\}.$$

If α is small enough, A is simply a union of disjoint intervals of length exactly 2α. We fix a small $\alpha > 0$, and choose $\mathcal{N}$ such that $V_2(\mathcal{N}) \subset A$. In that case, from equation (22), the sequence $u_n = V_2(\theta'_n)$ satisfies

$$u_n \leq u_{n-1} - \gamma_n \epsilon (1 - 1_A(u_{n-1})) + \gamma_n C 1_A(u_{n-1}).$$

Each time that u_n is out of A, it diminishes of $-\gamma_n\epsilon$; since $\sum\gamma_i$ is infinite, this implies that u_n will reach another interval of A corresponding to smaller values of u_n; since u_n cannot jump of more than $\gamma_n C$, this implies that u_n tends to A. Since α is arbitrary, $V_2(\theta'_n)$ converges to some point of $V_2(\mathcal{S})$.

The convergence of $V_2(\theta'_n)$ implies that for any $\tau > 0$, $|V_2(\theta'_n) - V_2(\theta'_p)| > \tau$ happens finitely often; on the other hand, if $\theta'_n \notin \mathcal{N}$, $V_2(\theta'_n)$ diminishes of $-\gamma_n\epsilon$. Thus, for n large enough there exists p such that $\epsilon(\gamma_n + ... + \gamma_p) < \tau$ and $\theta'_p \in \mathcal{N}$; since on the interval $[n,p]$ the variation of θ'_k is arbitrary small (of order $O(\tau)$), θ'_n is itself at distance $O(\tau)$ from $\mathcal{N}$. Since τ is arbitrary, this implies that $d(\theta'_n, \mathcal{N})$ tends to zero. Since $\mathcal{N}$ is arbitrary, θ'_n tends to $\mathcal{S}$, and so does θ_n.

6.2 Proof of second assertion

Let x be a point of $\mathcal{N}$ (cf assumption(B)) and denote by x_t the trajectory of h starting at x:

$$\frac{dx_t}{dt} = h(x_t), \qquad x_0 = x.$$

Consider ϵ such that $x + \epsilon h(x) \in \mathcal{N}$, using Lipschitz property of h, π, and W, we obtain

$$\begin{aligned}
|\pi(x+\epsilon h(x)) - \pi(x)| &\le |\pi(x+\epsilon h(x)) - \pi(x_\epsilon)| + |\pi(x_\epsilon) - \pi(x)| \\
&\le C|x + \epsilon h(x) - x_\epsilon| + W(x_\epsilon) - W(x) \\
&\le C'|x + \epsilon h(x) - x_\epsilon| + W(x + \epsilon h(x)) - W(x) \\
&\le C''\epsilon^2 + W(x+\epsilon h(x)) - W(x).
\end{aligned}$$

Using this inequality with $x = \theta'_{n-1}$ (cf equation (20)) we obtain, thanks again to the regularity of h and π,

$$\begin{aligned}
|\pi(\theta'_n) - \pi(\theta'_{n-1})| &\le |\pi(\theta'_{n-1} + \gamma_n h(\theta'_{n-1})) - \pi(\theta'_{n-1})| + |\pi(\theta'_n) - \pi(\theta'_{n-1} + \gamma_n h(\theta'_{n-1}))| \\
&\le W(\theta'_n) - W(\theta'_{n-1}) + C''\gamma_n^2 + C'''\gamma_n(|r_n| + |\delta'_n|).
\end{aligned}$$

Thus

$$\sum_{n=1}^{\infty} |\pi(\theta'_n) - \pi(\theta'_{n-1})| < \infty$$

and $\pi(\theta'_n)$ converges. Since $|\pi(x) - x|$ tends to zero as x tends to $\mathcal{S}$, θ'_n converges to some limit and so does θ_n.

6.3 Proof of third assertion

We need first to prove the μ_n tends to zero. In this section, H will denote h'_∞. We have

$$\mu_n = e^{\gamma_n H}\mu_{n-1} + \gamma_n e_n + \gamma_n r_n + \gamma_n H\delta_n, \qquad \mu_0 = 0. \tag{23}$$

The sequence

$$\mu'_n = \mu_n + \sum_{i=n+1}^{\infty} \gamma_i e_i$$

satisfies

$$\begin{aligned}
\mu'_n &= e^{\gamma_n H}(\mu'_{n-1} - \sum_{i=n}^{\infty}\gamma_i e_i) + \sum_{i=n}^{\infty}\gamma_i e_i + \gamma_n r_n + \gamma_n H\delta_n \\
&= e^{\gamma_n H}\mu'_{n-1} + \gamma_n r'_n
\end{aligned}$$

and r'_n tends to zero. μ'_n rewrites

$$\mu'_n = \sum_{1}^{n} e^{(s_n - s_i)H}\gamma_i r'_i + e^{s_n H}\mu'_0$$

which leads to the bound

$$|\mu'_n| \leq C\sum_1^n e^{-\lambda(s_n-s_i)}\gamma_i + Ce^{-\lambda s_n}|\mu'_0| \leq C'\sum_1^n e^{-\lambda(s_n-s_{i-1})}\gamma_i + Ce^{-\lambda s_n}|\mu'_0|.$$

If we use in this last bound the inequality

$$\gamma_i \leq C''(e^{\lambda\gamma_i}-1)$$

we finally obtain

$$|\mu'_n| \leq C'C''\sum_1^n (e^{-\lambda(s_n-s_i)} - e^{-\lambda(s_n-s_{i-1})}) + Ce^{-\lambda s_n}|\mu'_0| = C'C''(1-e^{-\lambda s_n}) + Ce^{-\lambda s_n}|\mu'_0|$$

Thus μ'_n is bounded and μ_n also. Since

$$\mu_n = \mu_{n-1} + \gamma_n H\mu_{n-1} + \gamma_n e_n + \gamma_n(r_n + O(\gamma_n\mu_n) + H\delta_n),$$

the stability of the matrix H and the A-stability of this algorithm implies that μ_n tends to zero (first assertion of the theorem).

Now we study $z_n = \theta_n - \theta_\infty - \mu_n$. We have

$$\begin{aligned}
\theta_n &= \theta_{n-1} + \gamma_n h(\theta_{n-1}+\delta_n) + \gamma_n e_n + \gamma_n r_n \\
&= \theta_{n-1} + \gamma_n H(\theta_{n-1} - \theta_\infty + \delta_n) + \gamma_n O(|\theta_{n-1}-\theta_\infty+\delta_n|^2) + \gamma_n e_n + \gamma_n r_n \\
\theta_n - \theta_\infty &= e^{\gamma_n H}(\theta_{n-1}-\theta_\infty) + \gamma_n H\delta_n \\
&\quad + \gamma_n O(\gamma_n|\theta_{n-1}-\theta_\infty| + |\theta_{n-1}-\theta_\infty+\delta_n|^2) + \gamma_n e_n + \gamma_n r_n \\
&= e^{\gamma_n H}(\theta_{n-1}-\theta_\infty) + \gamma_n O(\gamma_n^2 + |\theta_{n-1}-\theta_\infty|^2 + |\delta_n|^2) + \gamma_n H\delta_n + \gamma_n e_n + \gamma_n r_n.
\end{aligned}$$

Subtracting (23) to this equation, we obtain

$$z_n = e^{\gamma_n H} z_{n-1} + \gamma_n O(\gamma_n^2 + |\delta_n|^2 + |\mu_n|^2 + |z_{n-1}|^2).$$

z_n rewrites

$$z_n = \sum_1^n e^{(s_n-s_i)H}\gamma_i O(\gamma_i^2 + |\delta_i|^2 + |\mu_i|^2 + |z_{i-1}|^2) + e^{-s_n H}z_0$$

which leads to the bound

$$|z_n| \leq C\sum_1^n e^{-\lambda(s_n-s_i)}\gamma_i(\tau_i + |z_{i-1}|^2) + Ce^{-\lambda s_n}|z_0|$$

where

$$\tau_n = \gamma_n^2 + |\delta_n|^2 + |\mu_n|^2.$$

The only difference with the anounced bound is the term $|z_{i-1}|^2$. $|z_n|$ is smaller than

$$x_n = C\sum_1^n e^{-\lambda(s_n-s_i)}\gamma_i(\tau_i + |z_{i-1}|^2) + Ce^{-\lambda s_n}|z_0|$$

which is bounded (like μ'_n) and satisfies

$$x_n = e^{-\gamma_n\lambda}x_{n-1} + C\gamma_n|z_{n-1}|^2 + C\gamma_n\tau_n.$$

Note that

$$x_n^2 = e^{-2\gamma_n\lambda}x_{n-1}^2(1+\gamma_n o(1)) + \gamma_n\tau_n O(1)$$

and for any positive constant D

$$\begin{aligned}
x_n + Dx_n^2 &= e^{-\gamma_n\lambda}x_{n-1} + C\gamma_n|z_{n-1}|^2 + De^{-2\gamma_n\lambda}x_{n-1}^2(1+\gamma_n o(1)) + O(1)\gamma_n\tau_n \\
&\leq e^{-\gamma_n\lambda}(x_{n-1} + e^{\gamma_n\lambda}C\gamma_n x_{n-1}^2 + De^{-\gamma_n\lambda}x_{n-1}^2(1+\gamma_n o(1))) + O(1)\gamma_n\tau_n \\
&= e^{-\gamma_n\lambda}(x_{n-1} + (D + \gamma_n(C - \lambda D + o(1)))x_{n-1}^2) + O(1)\gamma_n\tau_n.
\end{aligned}$$

If $D = 2C/\lambda$, we have for n large enough

$$x_n + Dx_n^2 \leq e^{-\gamma_n \lambda}(x_{n-1} + Dx_{n-1}^2) + C' \gamma_n \tau_n$$

which proves that for some constant C''

$$x_n + Dx_n^2 \leq C'' \sum_1^n e^{-\lambda(s_n - s_i)} \gamma_i \tau_i + C'' e^{-\lambda s_n}$$

and

$$|z_n| \leq C'' \sum_1^n e^{-\lambda(s_n - s_i)} \gamma_i \tau_i + C'' e^{-\lambda s_n}.$$

∎

7 Appendix 1

This appendix is devoted to the statement and the proof of a result concerning summation of series:

Theorem 7 *Let γ_i be a non-negative sequence, $s_n = \gamma_1 + ... + \gamma_n$ and b and p be two real numbers, $b > 0$, $p \geq 1$ and consider*

$$\sigma_n = \sum_{i=1}^n i^{-p} e^{-b(s_n - s_i)}.$$

The following is true

- *if* $\gamma_n n^\alpha \longrightarrow 1$ *for some* $0 < \alpha < 1$ *then*

$$\sigma_n = \frac{n^{-p}}{b\gamma_n}(1 + o(1))$$

- *if* $\gamma_n n \longrightarrow 1$ *then*

$$\begin{aligned} \sigma_n &= \frac{n^{-p}}{(1 + b - p)\gamma_n}(1 + o(1)) \qquad & b > p - 1 \\ \sigma_n &= n^{-b} O(1) & b < p - 1 \end{aligned}$$

Proof In the case $\gamma_n n^\alpha \longrightarrow 1$ for some $0 < \alpha < 1$ we write

$$\sigma_n = a_n + b_n$$

with

$$a_n = \sum_{i=n-n^{2\alpha}}^n i^{-p} e^{-b(s_n - s_i)}, \qquad b_n = \sigma_n - a_n.$$

Then

$$\begin{aligned} a_n &= (1 + o(1)) n^{-p} \sum_{i=n-n^{2\alpha}}^n e^{-b(s_n - s_i)} \\ &= (1 + o(1)) n^{-p} \sum_{i=n-n^{2\alpha}}^n e^{-b\gamma_n (n-i)(1+o(1))} \\ &= (1 + o(1)) n^{-p} \frac{1 + o(1)}{b\gamma_n} = (1 + o(1)) \frac{n^{-p}}{b\gamma_n} \end{aligned}$$

and

$$b_n \leq \sum_{i=1}^{n-n^{2\alpha}} e^{-b(s_n-s_i)} \leq n e^{-b(s_n-s_{n^{2\alpha}})} \leq n e^{-bn^{1-\alpha}/2} \leq o(n^{-p}\gamma_n^{-1})$$

which proves the result.

The case $\gamma_n \longrightarrow 1$ is easily treated by noticing that

$$s_n - s_i = (1+\epsilon(n,i))\log(\frac{n}{i})$$

where

$$\lim_{n\longrightarrow\infty} \sup_{0<i<n} |\epsilon(n,i)| = 0.$$

References

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

[2] M.Métivier, P.Priouret, "Application of Kushner and Clark Lemma to General Classes of Stochastic Algorithms", IEEE-IT, vol 30, No 2, 1984.

[3] M.Duflo, *Méthodes récursives aléatoires*, Masson, Paris, 1990.

[4] O.Brandière, M.Duflo, "Les algorithmes stochastiques contournent-ils les pièges?", Internal Report, Université de Marne-la Vallée. To appear in "Annales de l'IHP".

[5] I.Fijalkow, Ph.Loubaton, P.A.Regalia, "A new algorithm for rational subspace estimation based on cascadable adaptive lossless filters", submitted to IEEE-SP.

[6] F.R.Gantmacher, *Theory of Matrices*, Paris, Dunod, 1966.

[7] P.W.Glynn, S.P.Meyn "A Lyapunov Bound for Solutions of Poisson Equation", to appear in Annals of Probability.

[8] P.Hall, C.C.Heyde, *Martingale Limit Theory and Its Applications*, Academic Press, 1980.

[9] T.Kato, *Perturbation theory for linear operators*, Springer-Verlag, 1966.

[10] H.J.Kushner, D.S.Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.

[11] Krasnovskii, *Stability of Motion*, Stanford University Press.

[12] M.Lavielle, E.Moulines, J-F. Cardoso, "A Maximum Likelihood Solution to DOA Estiamtion for Discrete Sources", IEEE Worshop on Statistical Signal and Array Processing, 1994, pp 349-352.

[13] V.A.Lazarev, "Convergence of Stochastic-Approximation Procedures in the Case of a Regression Equation with Several Roots", Problems of Information Transmission, 1993.

[14] S.P.Meyn, R.L.Tweedie *Markov Chains and Stochastic stability*, Springer-Verlag 1993.

[15] B.M.Nevel'son, R.Z.Khas'minskii, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society Translation of Math. Monographs, vol 47, 1976.

[16] E.Oja, "On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix", *Journal of Mathematical Analysis and Applications*, 106, 69-84 (1985).

[17] R.Pemantle, "Nonconvergence to Unstable Points in Urn Models and Stochastic Approximation", Annal of Probability, vol 18, no 2, 1990, 698-712.

[18] B.T.Polyak, "New stochastic Approximation Type Procedures", *Autom. & Remote Contr., 51(7).*

[19] V.U.Reddy, B.Egardt, T.Kailath, "Least Squares Type Algorithm for Adaptive Implementation of Pisarenko's Harmonic Retrieval Method", IEEE-ASSP, vol 30, No 3, June 1982.

[20] C.J.C.H. Watkins, P.Dayan, "$\mathcal{Q}$–Learning", Machine Learning, 8, 279-292 (1992).