Stochastic approximation with decreasing gain: Convergence and asymptotic theory

Bernard Delyon [1]

June 28, 2000

[1]IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France; email delyon@irisa.fr

Contents

1 Introduction **3**

2 Pointwise convergence of general algorithms **7**
2.1 Introduction 7
2.2 Pointwise convergence in the bounded case 7
2.2.1 Theory 8
2.2.2 Example : Estimation of eigenvectors 10
2.3 Reduction to the bounded case under recurrence condition 11
2.3.1 Theory 12
2.3.2 Example : Linear regression 15
2.4 Projection 15
2.4.1 Theory : Chen's method 16
2.4.2 Example : Stochastic gradient 18
2.4.3 Adaptive noise cancelling 18
2.4.4 Example : Source separation 19
2.5 Exercises 20

3 Stochastic models and pointwise convergence **25**
3.1 Robbins-Monro algorithms 25
3.1.1 Theory 25
3.1.2 Example : Dosage 27
3.1.3 Example : The two-armed bandit 27
3.2 Mixing noises 28
3.3 Algorithms with Markovian dynamics 30
3.3.1 Theory in the bounded case 30
3.3.2 Example : Blind equalizers 32
3.3.3 Example : AR estimation 33
3.3.4 Example : Q-learning 35
3.3.5 General theory 36
3.3.6 Example : MA estimation 40
3.3.7 Example : ARMA estimation 42
3.4 Exercises 43

4 Limit–theorems **45**
4.1 Introduction 45
4.2 Tools for linear approximation 47
4.2.1 A useful result on series 47

4.2.2 Approximation with a linear algorithm . . . 49
4.3 Asymptotic normality . . . 51
4.3.1 A general asymptotic bound for the estimation error . . . 53
4.3.2 Asymptotic normality . . . 54
4.3.3 Application to the algorithms with Markovian dynamics . . . 57
4.4 Averaging . . . 59
4.5 Exercises . . . 62

A Martingales **63**

B Markov chains and the Poisson equation **67**
B.1 Introduction to the Poisson equation . . . 67
B.2 Geometric ergodicity . . . 68
B.2.1 The finite case . . . 68
B.2.2 The general case . . . 69
B.2.3 Example : Linear models. . . . 70
B.3 Continuity with respect to a parameter : the compact case . . . 71
B.3.1 Theory . . . 71
B.3.2 Example : Linear models. . . . 73

C Some basic facts on stability **75**
C.1 Lyapunov functions . . . 75
C.2 Stable matrices . . . 76

Chapter 1

Introduction

The adaptive algorithms are now utilized across varied applications such as system identification, adaptive control, transmission systems, adaptive filtering for signal processing, and several aspects of pattern recognition and learning [28, 2, 9, 17, 30]. The aim of these algorithms is to estimate recursively an unknown time invariant (or slowly varying) parameter vector traditionally denoted by θ from measurements whose statistics depends on it.

The first known stochastic algorithm has been traced back to 1890 by B. Bru in the European artillery regulations. The problem is to adjust the slant θ (an angle) of a cannon in order to obtain a specified range r. This was done by trial and error, firing one shell after another. The nth experiment was done with the adjustment corresponding to the $(n-1)$th estimated value of θ, and led to an improved value θ_n; the result of this experiment is an observed range r_n, with a misadjustment $r - r_n$. It had been observed that r_n was actually a random variable, what made the problem much more difficult. After some heuristics, the specialists of the army converged on the following algorithm

$$\theta_n = \theta_{n-1} - \frac{\lambda}{n}(r - r_n)$$

where λ is some fixed normalization constant. In the case where one observes only if the shot were too short or too long, the regulations were to use

$$\theta_n = \theta_{n-1} - \frac{\lambda}{n}\,\mathrm{sign}(r - r_n).$$

The important discovery here is that the right choice of gain is of order $1/n$. This is a common feature of stochastic algorithms to have a rather large gain due to the necessity to average out the noise coming from the randomness of the observations. We shall see that under suitable assumptions, θ_n converges to the value which cancels the expectation of the additive term of the right hand side : the first algorithm will lead a value such that the mean range is r; the second one to a value which balances the probability of being short and the one of being long.

The general situation for stochastic approximation is the following. We assume that the observations Y_n are related to the true parameter θ^* via the equation

$$E[H(\theta, Y_n)] = 0, \quad \text{if} \quad \theta = \theta^* \tag{1.1}$$

where H is known, but the distribution of Y is unknown; $E[\,]$ stands for the expectation. A simple example is the noisy gradient problem :

$$H(\theta, Y_n) = -\nabla Q(\theta) + Y_n,$$

Y_n is an i.i.d Gaussian sequence and Q is a function to be minimized. Equation (1.1) means essentially that we do not have direct access to the cost function $E[H(\theta, Y_n)]$ but only to random observations of it (i.e. $H(\theta, Y_n)$). In some cases, the distribution of Y may also depend on θ and the above equation becomes

$$E_\theta[H(\theta, Y)] = 0, \quad \text{if} \quad \theta = \theta^*. \tag{1.2}$$

This was the case in the previous example, where Y_n was r_n and $H(\theta, Y_n)$ was $r - Y_n$ or $\operatorname{sign}(r - Y_n)$.

The general structure considered for stochastic algorithms is the following :

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n) \tag{1.3}$$

γ_n is a non-negative non-increasing sequence (gain sequence), typically $1/n$ or a constant, Y_n is the "somehow stationary" observed sequence and θ_n is at each n an improved estimate of θ^*. The function

$$h(\theta) = E_\theta[H(\theta, Y)]$$

is the *mean field* of the algorithm, and (1.3) may be seen as a perturbation of the equation :

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

which is itself close to the so-called O.D.E. (ordinary differential equation)

$$\dot{\theta} = h(\theta)$$

since γ_n is small. Notice that in the case where $H(\theta, Y) = Y - \theta$, and $\gamma_n = 1/n$, θ_n is simply the arithmetic mean of the Y_n's :

$$\theta_n = \frac{1}{n}(Y_1 + Y_2 + ... Y_n).$$

Observation models. The easiest situation is when Y_n is a given sequence of i.i.d random variable. For instance in the simplest case of stochastic gradient, we want to find the solution θ^* to

$$\theta^* = \arg\min_\theta E[Q(\theta, Y)]$$

and we observe independent samples Y_n with the distribution of Y; in this case, the choice $H(\theta, y) = -\nabla Q(\theta, y)$, leads to an algorithm resembling a gradient algorithm where the expectation sign has been omitted; the arithmetic mean corresponds to $Q(\theta, y) = |y - \theta|^2/2$. Let us mention, as an example, the recursive maximum likelihood : let $p_\theta(y)$ be the density of the distribution of Y when the true parameter is θ; the choice $Q(\theta, y) = -\log(p_\theta(y))$, leads to the algorithm

$$\theta_n = \theta_{n-1} + \gamma_n \frac{\nabla_\theta p_{\theta_{n-1}}(Y_n)}{p_{\theta_{n-1}}(Y_n)}.$$

However, the Y_ns are not always independent : in the classical *Robbins-Monro algorithms* the distribution of Y_n depends on θ_{n-1} (but not on the previous values of θ_k or Y_k). This was the case in our first example; this is also the situation when one wants to tune a parameter θ (e.g. a dosage of a chemical products,...) so that some criterion $L(\theta, Y)$ has a prescribed mean value α, and Y_n is the outcome of an experiment made with θ_{n-1}; e.g. $L(\theta, Y)$ may be the effect of dosage

θ, measured by Y (heat produced,...) combined with the price of the constituents (a function of θ). In this case $H(\theta, y) = \alpha - L(\theta, y)$ and the equation is

$$E_\theta[L(\theta, Y)] = \alpha, \quad \theta = \theta^*.$$

Notice that the dependence of the sequence θ_n implies that the Y_ns cannot be independent.

Unfortunately, the Robbins-Monro situation does not cover all the practical examples and the most general setting is the *Markovian dynamics*, introduced in [2, 19], where Y_n is a random variable depending only on (θ_{n-1}, Y_{n-1}); more precisely, the distribution of Y_n conditional to the whole past $(Y_{n-1}, \theta_{n-1}, ..)$ is given by a transition kernel $\Pi_{\theta_n}(Y_{n-1}, dy)$:

$$P(Y_{n+1} \in B|\theta_0, ..\theta_n, Y_0, ..Y_n) = \Pi_{\theta_n}(Y_n, B);$$

Y_n may be seen as an inhomogeneous Markov chain governed by θ_n (this point of view is somehow untrue since θ_n depends itself on the past values of the Y_n process); typically, the evolution of Y_n will take the form

$$Y_{n+1} = f(Y_n, \theta_n) + W_{n+1}$$

where (W_n) is an i.i.d. sequence. In this case, the expectation above (equation (1.2)) has to be taken with respect to the stationary distribution of the Markov chain Y_n when θ is fixed.

The *Robbins-Monro* algorithms correspond to the situation where $\Pi_{\theta_n}(Y_n, B)$ does not depend on Y_n, i.e. Y_{n+1} is drawn "independently of the rest" according to a distribution which depends only on θ_n.

Decreasing gain. In the case of decreasing gain, which is the subject of this book, typical results which are proved are the almost sure convergence of θ_n to a solution to the equation $h(\theta) = 0$ (where $h(\theta) = E_\theta[H(\theta, Y)]$), the asymptotic normality of $\gamma_n^{-1/2}(\theta_n - \theta^*)$ and the law of iterated logarithm [10, 21, 2, 14] [1].

The main difficulty in the proof of the almost sure convergence is the stability, i.e. to ensure that θ_n does not tend to infinity. The traditional way of dealing with this problem is either to put this into the assumption set (this is what we do firstly in §2.2 and §2.3), or to use strong assumptions on Lyapunov functions; unfortunately this last method is rarely instrumental and at best leads to extremely laborious computations on most realistic situations. Instead of this, we have chosen to use Chen's projection, which is an alternate method always usable for stabilizing the algorithm without any additional knowledge on the mean field (cf §2.4). This leads, in Chapter 3, to a handy treatment of the convergence with probability one in the Robbins-Monro case as well as in the Markovian situation.

Chapter 4 is devoted to general results on the asymptotic distribution of $\gamma_n^{-1/2}(\theta_n - \theta^*)$. We tried to give the most precise account of what can be said under the assumption of a.s. convergence, and to clarify these results which are already known in specific contexts [21, 2]. Properties of the averaging method have also been included. In the case $\gamma_n = 1/n$, the asymptotic covariance matrix V of is the solution to the *Lyapunov equation* (cf §4.1)

$$-(K + I/2)V - V(K + I/2)^T = R(\theta^*)$$

[1] Let us just mention in this note that the algorithms with constant gain $\gamma_n = \gamma$ are used in the case where θ^* is slowly varying with time, in which case the objective is to find the value of γ which gives the algorithm the best tracking performances.

where

$$\begin{aligned} K &= \nabla h(\theta^*) \\ R(\theta^*) &= \sum_{n=-\infty}^{+\infty} \mathrm{Cov}(H(\theta^*, Y_n), H(\theta^*, Y_0)). \end{aligned}$$

The expectation is taken w.r.t the distribution of the sequence (Y_n) (in the general Markovian situation, one has to consider here the stationary distribution of the chain of transition kernel $\Pi_{\theta^*}(Y_{n-1}, dy)$). In particular, the asymptotic normality occurs only if $I/2 + K$ has only negative eigenvalues (if not, the Lyapunov equation above has no solution). Improving this covariance (minimizing V) may be done by insertion of a matrix gain Γ:

$$\theta_n = \theta_{n-1} + \frac{1}{n}\Gamma H(\theta_{n-1}, Y_n)$$

and a simple computation shows that the optimal gain is

$$\Gamma^* = -K^{-1}.$$

In some cases, this matrix may be estimated during the algorithm, but convergence becomes quite difficult to prove. With this choice of the gain, the Cramér-Rao bound is attained. Another way is to use the Polyak-Ruppert averaging method [27, 25]:

$$\begin{aligned} \theta_n &= \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n) \\ \bar\theta_n &= \bar\theta_{n-1} + \frac{1}{n}(\theta_{n-1} - \bar\theta_{n-1}). \end{aligned}$$

with a “larger” γ_n (typically $\gamma_n = n^{-2/3}$). One can prove the asymptotic optimality of this algorithm (cf §4.4).

Chapter 2

Pointwise convergence of general algorithms

2.1 Introduction

The study of stochastic algorithms will be made by rewriting (1.3) as

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n \tag{2.1}$$

where the mean field $h(\theta) = E_\theta[H(\theta, Y)]$ is the function of θ we want to cancel and $\eta_n = H(\theta_{n-1}, X_n) - h(\theta_{n-1})$ plays the role of a random perturbation (ideally a sequence of i.i.d. random variables). The proof of convergence is 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 (noise sequence $(\eta_1(\omega), \eta_2(\omega), ...)$) of the stochastic system.

The first point of this program is thoroughly studied in this chapter. Under stability conditions on θ_n and suitable properties of η_n, we will show that the convergence is strongly related to the convergence of the mean differential equation driven by the mean field h :

$$\frac{d\theta_t}{dt} = h(\theta_t), \qquad h(\theta) = E_\theta[H(\theta, Y)].$$

The second point will be treated on examples, and more completely studied in §3.

2.2 Pointwise convergence in the bounded case

We will assume that the noise term in (2.1) may be written in the form $\eta_n = e_n + r_n$ where

- the series $\sum_{i=1}^n \gamma_i e_i$ converges
- r_n tends to zero.[1]

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

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

Thus this assumption on η_n is also necessary, at least from a theoretical point of view.

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 §2.2.2). Other equivalent conditions under which the decomposition holds are used in the literature (see the exercises 2.5.1 and 2.5.2).

2.2.1 Theory

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 continuous vector field defined on $\mathcal{O}$; there exist a non-negative C^1 function V such that*

- $\forall x \in \mathcal{O},\ \langle \nabla V(x), h(x)\rangle \leq 0$
- *the set $\mathcal{S} = \{x; \langle \nabla V(x), h(x)\rangle = 0\}$ is such that $V(\mathcal{S})$ has empty interior.*

Remarks : Assumption (A) is based on the existence of a Lyapunov function V for the field h (cf Appendix C.1); this assumption is theoretically non-restrictive because it is strongly related to the convergence of the trajectories of h (cf Theorem 38 p.76 or see [13] theorem 5.3). If h is a gradient field, $h = -\nabla Q$, the choice $V = Q$ will work if Q is C^1.

Assumption (A) will ensure convergence to the set $\mathcal{S}$; typically $\mathcal{S}$ coincides with $\{x : h(x) = 0\}$ and is a finite set but this is not required.

For checking the second point of (A), Sard's theorem of differential geometry may be useful in difficult situations : it asserts that if V is C^d, then $V(\{\nabla V = 0\})$ has empty interior.

We shall consider here deterministic algorithms of the form

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n, \qquad \gamma_n \geq 0, \qquad \sum_{p=1}^{\infty} \gamma_p = \infty, \qquad \lim_{n\to\infty} \gamma_n = 0, \tag{2.2}$$

where $n \geq 1$, θ_0 is given, h is a vector field on $\mathcal{O} \subset \mathbf{R}^d$, e_n and r_n are perturbations, and γ_n is a non-negative scalar gain sequence.

The following definition contains the basic assumption for the convergence of θ_n :

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

$$\theta_n \quad \text{remains in a compact subset of} \quad \mathcal{O} \tag{2.3}$$

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n \quad \text{exists} \quad \text{and} \quad \lim_{n\to\infty} |r_n| = 0 \tag{2.4}$$

Remarks : We shall see later that a typical way of proving the convergence of series required in this definition is to use Theorem 29 p.63.

Theorem 2 *We assume that the algorithm (2.2) is A-stable. Under assumption (A), the distance of θ_n from the set $\mathcal{S}$, $d(\theta_n, \mathcal{S})$, converges to 0; in particular, if $\mathcal{S}$ is finite, θ_n converges to some point of $\mathcal{S}$.*

Proof: Set

$$F(\theta) = \langle \nabla V(\theta), h(\theta) \rangle.$$

We consider

$$\theta'_n = \theta_n + \sum_{i=n+1}^{\infty} \gamma_i e_i \tag{2.5}$$

$$\delta'_n = -\sum_{i=n}^{\infty} \gamma_i e_i \tag{2.6}$$

leading to

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

Thus

$$\theta'_n = \theta'_{n-1} + \gamma_n h(\theta'_{n-1}) + \gamma_n r'_n \tag{2.7}$$

where

$$r'_n = r_n - h(\theta'_{n-1}) + h(\theta'_{n-1} + \delta_n)$$

tends to zero (since δ_n tends to zero, θ'_n remains in some compact set $\mathcal{K}_0$ and h is uniformly continuous on compact subsets). Let us recall that V is C^1; hence we have, using Taylor's formula (θ'' is some point between θ'_{n-1} and θ'_n):

$$\begin{aligned} V(\theta'_n) &= V(\theta'_{n-1}) + \gamma_n \langle \nabla V(\theta''), h(\theta'_{n-1}) + r'_n \rangle \\ &= V(\theta'_{n-1}) + \gamma_n F(\theta'_{n-1}) + \gamma_n r''_n, \end{aligned}$$

where r''_n tends to zero as n tends to infinity. Consider any open neighborhood $\mathcal{N}$ of $\mathcal{S} \cap \mathcal{K}_0$. The continuity of F (any continuous function attains its maximum on any compact set) implies that if θ'_{n-1} is in the compact set $\mathcal{K}_0 \backslash \mathcal{N}$, then

$$F(\theta'_{n-1}) \leq -2\varepsilon$$

for some $\varepsilon > 0$. Finally, we have the existence of $\varepsilon, C > 0$ such that

$$V(\theta'_n) \leq V(\theta'_{n-1}) - \gamma_n \varepsilon + \gamma_n C 1_{\mathcal{N}}(\theta'_{n-1}) \tag{2.8}$$

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

Consider the neighborhood of $V(\mathcal{S} \cap \mathcal{K}_0)$:

$$A_\alpha = \{x \in \mathbf{R} : \ d(x, V(\mathcal{S} \cap \mathcal{K}_0)) < \alpha\}.$$

A_α is a union of intervals of length 2α, it is thus a finite union of intervals of length at least 2α; we fix a small $\alpha > 0$, and choose $\mathcal{N} = V^{-1}(A_\alpha)$. In that case, from (2.8), the sequence $u_n = V(\theta'_n)$ satisfies

$$u_n \leq u_{n-1} - \gamma_n \varepsilon + \gamma_n C 1_{A_\alpha}(u_{n-1}).$$

Each time that u_n is out of A_α, it diminishes of $-\gamma_n \varepsilon$; since $\sum \gamma_i$ is infinite, this implies that u_n will reach another interval of A_α corresponding to smaller values of u_n; since u_n cannot jump of more than $\gamma_n C$, this implies that $d(u_n, A_\alpha)$ tends to zero. Since $V(\mathcal{S} \cap \mathcal{K}_0)$ is closed (compact), this implies that the set of limit points of u_n is contained in $V(\mathcal{S}\cap\mathcal{K}_0)$; this set of limit points is an interval (because $u_{n+1} - u_n$ tends to zero), and on the other hand, $V(\mathcal{S})$ has empty interior; this implies that u_n converges to some limit : $V(\theta'_n)$ converges to some point of $V(\mathcal{S})$.

Consider again an open neighborhood $\mathcal{N}$ of $\mathcal{S}$. The convergence of $V(\theta'_n)$ implies that for any $\tau > 0$, there exists $n(\tau)$ such that

$$|V(\theta'_n) - V(\theta'_p)| < \tau, \qquad n(\tau) < n < p.$$

This implies (cf (2.8))

$$\varepsilon \sum_{k=n+1}^{p} \gamma_k - C \sum_{k=n+1}^{p} \gamma_k 1_{\mathcal{N}}(\theta'_{k-1}) \leq \tau, \qquad n(\tau) < n < p. \tag{2.9}$$

Consider $n > n(\tau)$ and let p be the first integer larger than n such that $\varepsilon(\gamma_{n+1} + ... + \gamma_p) > \tau$; (2.9) implies that there exists $k \in [n, p]$ such that $\theta'_k \in \mathcal{N}$; since $\theta'_k - \theta'_n$ is arbitrary small (of order $O(\tau/\varepsilon)$, cf (2.7)), θ'_n is itself at distance $O(\tau/\varepsilon)$ 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. ∎

2.2.2 Example : Estimation of eigenvectors

We are given a sequence of random matrices A_k with common expectation A, a symmetric matrix; the following algorithm for estimating the first eigenvector of A is considered in [23] :

$$\theta_n = \frac{\theta_{n-1} + \gamma_n A_n \theta_{n-1}}{|\theta_{n-1} + \gamma_n A_n \theta_{n-1}|}. \tag{2.10}$$

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 [9, 23]).

Theorem 3 *We consider the algorithm (2.10) where A_k is a sequence of independent matrices and*

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

The sequence θ_n converges almost surely to some eigenspace of this matrix. In particular, if A has no multiple eigenvalue, θ_n converges to some eigenvector.

Proof :

Step 1 : Rewriting (2.10) into the form (2.2). Using the relation true for $|\theta| = 1$ and $|\varepsilon| < 1$

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

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 (2.2) is

$$\begin{aligned}
\mathcal{O} &= \mathbf{R}^d\backslash\{0\} \\
h(x) &= Ax-(x^TAx)x+\lambda x-\lambda|x|^2x=(I-xx^T)(A+\lambda I)x \\
e_n &= e_n'+e_n'' \\
e_n' &= (A_n-A)\theta_{n-1}-(\theta_{n-1}^T(A_n-A)\theta_{n-1})\theta_{n-1} \\
e_n'' &= O(\gamma_n\|A_n\|^2).
\end{aligned}$$

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 (since $x^Th(x)=(1-|x|^2)(x^T(A+\lambda I)x)$).

Step 2: Checking the A-stability. Since $E[\sum_n\gamma_n^2|A_n|^2]<\infty$, the series $\sum_n\gamma_n e_n''$ is a.s. absolutely convergent. On the other hand, e_n' being an L_2-bounded martingale-increment the series $\sum_n\gamma_n e_n''$ is also a.s. convergent (cf Theorem 29 p.63) and the A-stability is satisfied.

Step 3: Checking (A). Set

$$B=A+\lambda I$$

and take

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

Then

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

∎

It is possible to prove that θ_n converges to some point even if there is multiple eigenvalues [8]. It is also possible to show that θ_n converges to the eigenvector corresponding to the maximum eigenvalue [9]; amazingly, this last point requires rather involved considerations in probability theory (the more general problem of the convergence toward stable points, i.e. $\nabla h(\theta)<0$, is treated in [9]).

2.3 Reduction to the bounded case under recurrence condition

A-stability is the basic assumption for convergence of θ_n required in Theorem 2 p.8; the boundedness of the sequence θ_n (equation (2.3)) is however generally difficult to check. The aim of this section is to give conditions under which this boundedness assumption may be replaced with the assumption that θ_n does not tend to infinity (recurrence, equation (2.14)). Proposition 9 p.14 provides a simple but useful tool for checking the recurrence (see also Exercise 2.5.8 p.22).

2.3.1 Theory

The following assumption will be essential; the Lyapunov function V below guarantees that the field h is attractive enough when x is far away (i.e. close to $\partial\mathcal{O}$) :

(B) *h is a continuous vector field defined on $\mathcal{O}$; there exist a non-negative C^1 function V and a compact set $\mathcal{K} \subset \mathcal{O}$ such that*

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

If h is a gradient field, $h = -\nabla Q$, the choice $V = Q$ will work if Q is C^1 and tends to ∞ in the vicinity of $\partial\mathcal{O}$.

The proofs of the theorems of this section will be based on the following

Lemma 4 *Under assumption (B), consider a real number M such that*

$$\mathcal{K} \subset \{\theta; V(\theta) \le M\}.$$

Then there exists ε, depending on M, V and h, such that for any fixed $N \le \infty$, the relations

$$V(\theta_0) \le M$$

$$\forall\ 1 \le n \le N, \qquad \left|\sum_{i=1}^{n} \gamma_i e_i\right| < \varepsilon, \quad |r_n| < \varepsilon, \quad |\gamma_n| < \varepsilon$$

(if $N = +\infty$ the convergence of the series is assumed) imply for θ_n of (2.2)

$$V(\theta_n) \le M + 1, \quad n \le N.$$

Proof : We define

$$\begin{aligned} \theta'_n &= \theta_n + \sum_{i=n+1}^{N} \gamma_i e_i \\ \delta_n &= -\sum_{i=n}^{N} \gamma_i e_i. \end{aligned}$$

Notice that

$$|\theta'_n - \theta_n| \le 2\varepsilon, \quad |\delta_n| \le 2\varepsilon$$

and the sequence θ'_n satisfies

$$\theta'_n = \theta'_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n r_n. \tag{2.11}$$

We shall show by recursion that $V(\theta'_n) < M + 1/2$. This will prove the result since if ε is small enough, $V(\theta') \le M + 1/2$ and $|\theta - \theta'| \le 2\varepsilon$ implies $V(\theta) \le M + 1$.

Using Taylor formula we obtain from (2.11):

$$V(\theta'_n) = V(\theta'_{n-1}) + \gamma_n \langle \nabla V(\theta''_{n-1}), h(\theta_{n-1}) + r_n \rangle$$

for some θ''_{n-1} lying in between θ'_{n-1} and θ'_n. Since

$$|\theta''_{n-1} - \theta'_{n-1}| \quad \le \quad |\theta'_n - \theta'_{n-1}| \le \varepsilon(\sup_{V(\theta)<M} |h(\theta)| + \varepsilon)$$

one has

$$V(\theta'_n) \leq V(\theta'_{n-1}) + \gamma_n \langle \nabla V(\theta'_{n-1}), h(\theta'_{n-1}) \rangle + \gamma_n \varepsilon' \tag{2.12}$$

where ε' can be made arbitrarily small by an adequate choice of ε. On the other hand, since the function $\langle \nabla V, h \rangle$ is continuous and negative on the compact set $\{\theta; M \leq V(\theta) \leq M + 1/2\}$, there exists $\alpha > 0$ such that

$$\langle \nabla V(\theta'), h(\theta') \rangle < -\alpha \quad \text{if} \quad V(\theta') \in [M, M + 1/2].$$

We can choose ε such that $\varepsilon' < \alpha$, implying

$$V(\theta'_n) \leq V(\theta'_{n-1}) \quad \text{if} \quad V(\theta'_{n-1}) \in [M, M + 1/2] \tag{2.13}$$

From equations (2.12,2.13), we can conclude that, for ε small enough

- if $V(\theta'_{n-1}) \in [0, M]$, then $V(\theta'_n) < M + 1/2$
- if $V(\theta'_{n-1}) \in [M, M + 1/2]$, then $V(\theta'_n) < V(\theta'_{n-1})$.

Thus $V(\theta'_n)$ remains smaller than $M + 1/2$. ∎

Theorem 5 *We assume that assumption (B) is satisfied and the existence of a compact subset $\mathcal{K}_0 \subset \mathcal{O}$ such that*

$$\theta_n \in \mathcal{K}_0 \quad \textit{infinitely often.} \tag{2.14}$$

If equation (2.4) is satisfied the algorithm (2.2) is A-stable.

Proof: Consider a time n_0 such that $\theta_{n_0} \in \mathcal{K}_0$ and

$$\forall n \geq n_0 + 1, \qquad \left| \sum_{i=n_0+1}^{n} \gamma_i e_i \right| < \varepsilon, \quad |r_n| < \varepsilon, \quad |\gamma_n| < \varepsilon$$

where ε has the value given in Lemma 4. The theorem is a straightforward application of Lemma 4 to the algorithm starting at time n_0, with $N = +\infty$ and M such that

$$\mathcal{K} \cup \mathcal{K}_0 \subset \{\theta; V(\theta) \leq M\}.$$

∎

In many cases of interest, e_n is some sort of martingale-increment whose variance depends on the past (for example in the case of the Robbins-Monro procedure or in §3.1 and §3.3). For these situations, it would be fruitful to have a theorem where the series in (2.4) would be restricted to those term for which θ_i is well controlled. Theorem 6 below gives an answer to this question (as well as Lemma 8 in a more sophisticated way).

Theorem 6 *We assume that assumption (B) is satisfied and the existence of a compact subset $\mathcal{K}_0 \subset \mathcal{O}$ such that*

$$\theta_n \in \mathcal{K}_0 \quad \textit{infinitely often.} \tag{2.15}$$

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

$$\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i 1_{V(\theta_{i-1}) \leq M} \quad \textit{exists and} \quad \lim_{n\to\infty} |r_n| 1_{V(\theta_{n-1}) \leq M} = 0. \tag{2.16}$$

Then the algorithm (2.2) is A-stable.

In a stochastic context, condition (2.15) is almost surely satisfied as soon as the sequence θ_n has a uniformly bounded moment, as shown in Proposition 9 below. The problem of recurrence is also considered in §2.4.

Proof: This is also an application of Lemma 4. Choose M such that

$$\mathcal{K} \cap \mathcal{K}_0 = \{\theta;\ V(\theta) \leq M\}$$

and m such that $\theta_m \in \mathcal{K}_0$ and for any $n > m$

$$\left| \sum_{i=m+1}^{n} e_i 1_{V(\theta_{i-1}) \leq M+2} \right| < \varepsilon, \quad |r_n| 1_{V(\theta_{n-1}) \leq M+2} < \varepsilon, \quad |\gamma_n| < \varepsilon.$$

where ε has the value given in Lemma 4. Then this lemma implies, by recursion on k, that $V(\theta_k) \leq M+1$, $k = m+1, m+2, \ldots$ ∎

We turn now to Lemma 8 which is a refined version of Theorem 6, useful in the case of algorithms with Markovian dynamics (Section 3.3). This lemma explains how, as soon as recurrence is satisfied, stability may be inferred from information about the excursions $(\theta_n, \theta_{n+1}, , ..\theta_p)$ for which $\theta_n \in \mathcal{K}_0$ and none of the θ_i is out of some compact set; this lemma prepares Theorem 10 p.16. We need first a definition

Definition 7 *We say that an increasing sequence $(\mathcal{K}_q)_{q \geq 0}$ of compact sets is an increasing compact covering of $\mathcal{O}$ if*

$$\mathcal{K}_{q-1} \subset int(\mathcal{K}_q), \qquad \cup_q \mathcal{K}_q = \mathcal{O}.$$

We say that $(\rho(k))_{k>0}$ is a recurrence time sequence if $\theta_{\rho(k)} \in \mathcal{K}_0$ and $\rho(k-1) < \rho(k)$.

Lemma 8 *We assume that assumption (B) is satisfied and that we are given an increasing compact covering of $\mathcal{O}$ and a recurrence time sequence $\rho(k)$. If for any q large enough*

$$\lim_{k \to \infty} \sup_{n \geq \rho(k)} I(\theta_{\rho(k)} \ldots, \theta_n \in \mathcal{K}_q) \left\{ |r_{n+1}| + \left| \sum_{i=\rho(k)+1}^{n+1} \gamma_i e_i \right| \right\} = 0 \tag{2.17}$$

then the sequence θ_n is contained in some compact set $\mathcal{K}_q$.

Proof: We choose M and q such that (cf assumption (B))

$$\mathcal{K} \cup \mathcal{K}_0 \subset \{\theta; V(\theta) \leq M\}$$
$$\{\theta; V(\theta) \leq M+1\} \subset \mathcal{K}_q$$

and consider a time $\rho(k)$ such that for all $n \geq \rho(k)$ satisfying $\theta_{\rho(k)} \ldots, \theta_n \in \mathcal{K}_q$ one has

$$\left| \sum_{i=\rho(k)+1}^{n+1} \gamma_i e_i \right| < \varepsilon, \quad |r_{n+1}| < \varepsilon, \quad |\gamma_{n+1}| < \varepsilon$$

where ε has the value given in Lemma 4 p.12. Applying Lemma 4, we obtain immediately by induction that $\theta_p \in \mathcal{K}_q$, $p = \rho(k)+1, \rho(k)+2 \ldots.$ ∎

For applying Theorem 6, the following result may be useful (cf §2.3.2)

Proposition 9 *Let (X_n) be a sequence of random variables such that*

$$\sup_n E[|X_n|] < \infty$$

then

$$P(|X_n| \longrightarrow \infty) = 0.$$

Proof: Indeed, using Fatou's lemma, we get

$$E[\liminf_n |X_n|] \leq \liminf_n E[|X_n|] < \infty$$

and consequently $\liminf_n |X_n|$ is a.s. finite. ∎

2.3.2 Example: Linear regression

We give in this section a typical application of Theorem 6.

We observe a sequence (y_n, φ_n) satisfying

$$y_n = \varphi_n^T \theta^* + \xi_n$$

where (φ_n) and (ξ_n) are two mutually independent sequence of i.i.d. variables. We shall consider the following estimation algorithm for θ^*:

$$\theta_n = \theta_{n-1} + \gamma_n f(y_n - \varphi_n^T \theta_{n-1})\varphi_n. \tag{2.18}$$

We will assume that for all $x \neq 0$

$$\begin{aligned}
\sum \gamma_n^2 &< \infty, \quad \sum \gamma_n = \infty & (2.19)\\
xE[f(x+\xi_n)] &> 0, \quad x \neq 0 & (2.20)\\
|f(x)| &\leq C(1+|x|) & (2.21)\\
E[|\varphi_n|^4] &< \infty, \quad E[|\xi_n|^2] < \infty & (2.22)\\
E[\varphi_n \varphi_n^T] &> 0 & (2.23)\\
& & (2.24)
\end{aligned}$$

Typical examples of functions f are

$$\begin{aligned}
f(x) &= x\\
f(x) &= \operatorname{sign}(x)\\
f(x) &= -\frac{p'_\xi(x)}{p_\xi(x)} = -\partial_x \log p_\xi(x)
\end{aligned}$$

where p_ξ is the density of ξ_t. This last choice corresponds to the recursive maximum likelihood. The proof of convergence is left in Exercise 2.5.3.

2.4 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 (this is not always the case as we have seen in §2.3.2). 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. On the other hand, practical implementations are generally performed with some projection which brings back θ into a safe region where we know the true parameter lies.

The most naive way is, if we know in advance that the true parameter is in some set $\mathcal{K}_0$, to bring θ back into this set each time it wanders out. Unfortunately, this algorithm may not converge if some trajectories of h go out of $\mathcal{K}_0$; that is, even the algorithm without any perturbation term will fail to converge. This difficulty may be avoided by imposing additional constraints on the set

$\mathcal{K}_0$ (namely, that the trajectories starting in this set remain in it); this solution however may be impossible to implement (since h is not known), and is theoretically difficult to deal with.

There exists a more subtle and elegant way due to H. F. Chen [7] of doing projections with which we get recurrence without spoiling convergence : we choose any increasing sequence of compact sets $\mathcal{K}_n$, $n = 0, 1, ..,$ whose union is $\mathcal{O}$, and start with $n = 1$; each time θ_k is out of $\mathcal{K}_n$, its value is reset arbitrarily in $\mathcal{K}_0$ and n is incremented by one. More precisely we now consider the modified algorithm defined as :

$$\begin{aligned} \hat{\theta}_n &= \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n, \\ \text{if } \hat{\theta}_n \in \mathcal{K}_{s(n-1)} &: \quad \theta_n = \hat{\theta}_n, \qquad s(n) = s(n-1) \\ \text{if } \hat{\theta}_n \notin \mathcal{K}_{s(n-1)} &: \quad \theta_n = \theta'_n \in \mathcal{K}_0, \qquad s(n) = s(n-1) + 1. \end{aligned} \tag{2.25}$$

θ'_n is arbitrarily chosen in $\mathcal{K}_0$ (for instance it may be a function of the past values, a randomly chosen point....). The initial conditions are any admissible values for θ_0 and $s(0)$. We introduce a special notation for this algorithm :

$$\theta_n \stackrel{pr}{=} \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n \tag{2.26}$$

2.4.1 Theory : Chen's method

Theorems of this section guarantee the validity of the method under reasonable assumptions. Theorem 10 makes the fundamental point concerning stability; it is general and will be used later in §3.3 for proving the convergence of the algorithms with Markovian dynamics. Theorems 11 and 12 are simpler statements.

Theorem 10 *Consider the algorithm (2.26) where $\mathcal{K}_q$ is an increasing compact covering (cf Definition 7 p.14) and assumption (B) is satisfied; let $\tau(n)$ stand for the last projection time before n ($\tau(n) \le n$). If for any q large enough*

$$\lim_{n\to\infty} I(\theta_{\tau(n)}..., \theta_n \in \mathcal{K}_q) \left\{ |r_{n+1}| + \left| \sum_{i=\tau(n)+1}^{n+1} \gamma_i e_i \right| \right\} = 0, \tag{2.27}$$

then the sequence θ_n is contained is some compact set $\mathcal{K}_q$, and the projections are done only a finite number of times. This means that (2.26) rewrites

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n + \gamma_n r'_n$$

where r'_n is finitely often non-zero.

Proof : We can assume that there is an infinite number of projection (if not the result is already true). We shall apply Lemma 8. This algorithm can be put in a form similar to (2.2):

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r''_n$$

where

$$\begin{aligned} \text{if } \hat{\theta}_n \in \mathcal{K}_{s(n-1)} &: \quad r''_n = r_n \\ \text{if } \hat{\theta}_n \notin \mathcal{K}_{s(n-1)} &: \quad r''_n = \gamma_n^{-1}(\theta'_n - \theta_{n-1}) - h(\theta_{n-1}) - e_n. \end{aligned}$$

Consider the equation (2.17) where the $\rho(k)$s are the projection times; since for k large enough $s(\rho(k)) > q$ and $\theta_{\rho(k)-1} \notin \mathcal{K}_q$, the indicator in (2.17) vanishes if $\rho(k) \neq \tau(n)$. Thus (2.17) reduces to

$$\lim_{n\to\infty} I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i e_i = 0, \qquad I_n = I_n^q = I(\theta_{\tau(n)} \ldots, \theta_n \in \mathcal{K}_q)$$
$$\lim_{n\to\infty} I_n \, |r''_{n+1}| = 0.$$

The first equation is already satisfied. We show now that

$$I_n \, |r''_{n+1}| \neq I_n \, |r_{n+1}| \tag{2.28}$$

can happen only a finite number of times and the proof will be complete. (2.28) implies that $r''_{n+1} \neq r_{n+1}$ and $I_n = 1$; under these circumstances, $\hat{\theta}_{n+1} \notin \mathcal{K}_{s(n)}$ and $\theta_n \in \mathcal{K}_q$. The equation

$$\gamma_{n+1} e_{n+1} = \sum_{i=\tau(n)+1}^{n+1} \gamma_i e_i - \sum_{i=\tau(n)+1}^{n} \gamma_i e_i.$$

implies that $\gamma_{n+1} e_{n+1}$ is arbitrarily small if n is large enough by virtue of (2.27). This, and $\theta_n \in \mathcal{K}_q$, imply that if k is large enough $\hat{\theta}_{n+1} \in \mathcal{K}_{q+1}$. Thus (2.28) would imply $\hat{\theta}_{n+1} \in \mathcal{K}_{q+1}$ and $\theta_n \notin \mathcal{K}_{s(n)}$; this cannot happen infinitely often (since that would mean that $s(n)$ tends to infinity and $s(n) < q+1$) and the proof is complete. ∎

Theorem 11 *Consider the algorithm (2.26) where $\mathcal{K}_q$ is an increasing compact covering (cf Definition 7 p.14). If assumption (B) and equation (2.4) are satisfied then the algorithm is A-stable.*

In particular the projections are done a finite number of times.

Proof: This is an immediate consequence of Theorem 10 since (2.4) implies (2.27). ∎

Theorem 12 *Consider the algorithm (2.26) where $\mathcal{K}_q$ is an increasing compact covering and assumption (B) is satisfied; if for any $q \in \mathbf{N}$,*

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n 1_{V(\theta_{n-1})\leq q} \quad \textit{exists and} \quad \lim_{n\to\infty} |r_n| 1_{V(\theta_{n-1})\leq q} = 0, \tag{2.29}$$

then the algorithm is A-stable. In particular the projections are done a finite number of times.

Proof: This result is an immediate consequence of Theorem 10 with

$$\mathcal{K}_q = \{\theta;\, V(\theta) \leq q\}.$$

Equation (2.29) implies that

$$\lim_{n,p\to\infty} \sum_{i=n}^{p} \gamma_i e_i 1_{V(\theta_{i-1})\leq q} = 0. \tag{2.30}$$

Equation (2.27) is thus satisfied. ∎

2.4.2 Example: Stochastic gradient

Consider the simplest stochastic gradient algorithm

$$\theta_n = \theta_{n-1} - \gamma_n \nabla Q(\theta_{n-1}) + \gamma_n e_n$$

where e_n is, say a sequence of i.i.d. variables; checking the recurrence (i.e. equation (2.14)), if it happens to be true, is generally tedious and requires additional assumptions on Q.

It is however very simple to prove that the algorithm

$$\theta_n \stackrel{pr}{=} \theta_{n-1} - \gamma_n \nabla Q(\theta_{n-1}) + \gamma_n e_n$$

converges (Exercise 2.5.4).

2.4.3 Adaptive noise cancelling

The typical problem here is the following [30]: you record someone speaking in a noisy environment ('signal corrupted with additive noise') and you would like to reconstruct the original signal free of any noise. The principle is quite simple; it consists in using a second microphone for recording the noise only (this one should be located behind the speaker!) and in combining the to signals in order to get more speach and less noise. This rewrites in its simplest form

$$\hat{s}_n = x_n - \theta y_n.$$

where $\hat{s}_n$ is the reconstructed speach, x_n and y_n are the two measured signals and θ is parameter whose value has to be optimized. Let us assume the oversimplified model

$$\begin{aligned} x_n &= s_n + \theta^* w_n + e_n \\ y_n &= w_n. \end{aligned}$$

where s_n is the speach and w_n and e_n are independent noises. The optimal value for θ is θ^* and satisfies

$$\theta^* = \arg\min_\theta E[(x_n - \theta^* y_n)^2].$$

The gradient algorithm for the estimation of θ^* is thus

$$\theta_n \stackrel{pr}{=} \theta_{n-1} + \gamma_n (x_n - \theta_{n-1} y_n) y_n. \tag{2.31}$$

One checks easily that under the (unrealistic!) assumption that (x_n, y_n) is an i.i.d sequence, this algorithm will converge to $\theta^* = E[x_n y_n]/E[y_n^2]$.

In practice, because of propagation delays, θ is a vector and the algorithm writes:

$$\begin{aligned} \theta_n &\stackrel{pr}{=} \theta_{n-1} + \gamma_n (x_n - \theta_{n-1}^T Y_n) Y_n, \quad Y_n = (y_n, ... y_{n-p+1})^T \\ \hat{s}_n &= x_n - \theta_n^T Y_n. \end{aligned}$$

2.4.4 Example: Source separation

This section is more informal and we will not prove the convergence of the presented algorithms. The reader may try to find specific assumptions under which convergence holds if the projection is done, and study the sets of limit points; see also the Exercise 2.5.7.

We observe a process $x_n = As_n$ where A is square matrix, and s_n is a process with independent components and unit variance. The objective is to reconstruct s_n; an estimate of $B = A^{-1}$ will be used. The idea is to find a matrix B such that the process Bx_n has independent components with non-zero variance. It is easily seen that if s_n were Gaussian, any matrix $B = OA^{-1}$ where O is orthogonal would be a solution to this problem. However, if the components of s_n are non-Gaussian, the only solution are for O the permutation matrix (cf exercice 2.5.10); this means that one can reconstruct s_n up to a permutation.

Let us assume first that the distribution of s is known. In this case a natural way to go is the recursive maximum likelihood. The log-likelihood of a sample x under the hypothesis $A^{-1} = B$ is

$$\log p_s(Bx) + \log|\det(B)|$$

and its gradient w.r.t B is[2]

$$\nabla_B \left(\log p_s(Bx) + \log|\det B|\right) = \frac{\nabla p_s}{p_s}(Bx)x^T + B^{-T} = \left(\frac{\nabla p_s}{p_s}(y)y^T + I\right) B^{-T}.$$

A proposed algorithm is

$$B_n \quad = \quad B_{n-1} + \gamma_n \left(\frac{\nabla p_s}{p_s}(y_n)y_n^T + I\right) B_{n-1}, \qquad y_n = B_{n-1}x_n. \tag{2.32}$$

This algorithm is not exactly the gradient above but has the same stationary points; on these points, we have the uncorrelation property

$$E[y(i)(\log p_j)'(y(j))] = 0, \qquad i \neq j$$

where p_j is the density of the jth component of s. This does not ensure in general that the components of y are independent. Algorithm (2.32) has the advantage of being correctly normalized, and to act directly on the matrix $C_n = B_nA$

$$C_n \quad = \quad C_{n-1} + \gamma_n \left(\frac{\nabla p_s}{p_s}(C_{n-1}s_n)s_n^TC_{n-1}^T + I\right) C_{n-1}.$$

This equivariance property [6] implies that its behavior is essentially independent of A.

If the distribution of s is unknown or unsmooth (e.g. if s takes a finite number of possible values), other choices are possible. Among them let us mention one of the first one (the Jutten & Hérault algorithm)

$$B_n(i,j) = B_{n-1}(i,j) - \gamma_n y_n(i)^3 y_n(j) \qquad i \neq j, y_n = B_{n-1}x_n$$

[2]We use here the convention that the gradient of a function $f(B)$ is a matrix G such that the first order approximation is

$$f(B + \varepsilon H) \simeq f(B) + \varepsilon \ \mathrm{trace}(G^TH).$$

This convention implies that the gradient may be identified with an element of the original space endowed with the inner product $(M, N) \to \mathrm{trace}(M^TN)$.

which solves the equation $E[y_t(i)^3 y_t(j)] = 0$ (its assumed that $E[s] = 0$). Clearly, other expressions of the form $g_i(y_n(i))h_j(y_n(j))$ may be used.

The following has been also proposed [6]:

$$B_n = B_{n-1} - \gamma_n(y_n y_n^T - I + g(y_n)y_n^T - y_n g(y_n)^T)B_{n-1}.$$

In this last algorithm, the matrix $yy^T - I$ may be seen as a whitening term whilst $g(y)y^T - yg(y)^T$ is a skew-symmetric matrix contributing in multiplying B by an approximate orthogonal transformation[3].

Notice that if no projection is done, the algorithms of this section may suffer strong instability problems since the mean field is not sublinear.

2.5 Exercises

Exercise 2.5.1 Show that, if $\gamma_i \geq 0, \sum \gamma_i = \infty, \gamma_i \to 0$, the decomposability of η_n as $\eta_n = e_n + r_n$ with equation (2.4) is equivalent to the existence of some increasing sequence $\sigma(i)$ such that

$$\lim_{i\to\infty} \sup_{\sigma(i)\leq p\leq\sigma(i+1)} \left| \sum_{n=\sigma(i)}^{p} \gamma_n \eta_n \right| = 0$$

and for some a

$$0 < a \leq \sum_{n=\sigma(i)}^{\sigma(i+1)} \gamma_n, \quad i = 1, 2, ...$$

(Hint: For proving (2.4) from the above equations, set

$$r_n = \left(\sum_{j=\sigma(i)}^{\sigma(i+1)-1} \gamma_j \right)^{-1} \sum_{j=\sigma(i)}^{\sigma(i+1)-1} \gamma_j \eta_j, \qquad \sigma(i) \leq n < \sigma(i+1)$$

and $e_n = \eta_n - r_n$).

Exercise 2.5.2 Show that if $\gamma_i \geq 0, \sum \gamma_i = \infty, \gamma_i \to 0$, the decomposability of η_n as $\eta_n = e_n + r_n$ with equation (2.4) is equivalent to the convergence to 0 of the sequence τ_n:

$$\tau_n = p_n \sum_{i=1}^{n} \frac{\gamma_i}{p_i} \eta_i, \qquad p_n = (1-\gamma_1)...(1-\gamma_n).$$

(Hint: If $\tau_n \to 0$, try $r_n = \tau_{n-1}, e_n = \eta_n - \tau_{n-1}$. If the decomposition holds, prove that $\tau_n \to 0$ from the equation

$$\tau_n = \tau_{n-1} - \gamma_n \tau_{n-1} + \gamma_n e_n + \gamma_n r_n$$

seen as a stochastic algorithm: prove first that $\tau_n + \sum_{i>n} \gamma_i e_i$ is bounded, and then use Theorem 2).

Exercise 2.5.3 (Convergence of the linear regression) We consider the algorithm (2.18) with the assumptions given in §2.3.2.

[3] If A is a skew-symmetric matrix, e^A is orthogonal and $I + \gamma A \simeq e^A$.

- Set $\tilde{\theta}_n = \theta_n - \theta^*$ and write a recurrence relation (inequality) for $E[|\tilde{\theta}_n|^2]$ (use (2.20), (2.21) and (2.22))
- Prove that $u_n = E[|\tilde{\theta}_n|^2] + 1$ satisfies for some C:

$$u_n \leq e^{C \sum_{i=1}^{n} \gamma_i^2} u_0$$

and that, with probability 1, θ_n does not tend to infinity.
- Compute the mean field and prove, using (2.20), that the function

$$V(\theta) = |\theta - \theta^*|^2/2$$

is a valid Lyapunov function.
- Show, by using (2.23) that $\langle \nabla V(\theta), h(\theta) \rangle$ is zero if and only if $\theta = \theta^*$.
- Prove that (2.16) holds with $r_n = 0$ (use the fact that (φ_n, ξ_n) is an i.i.d sequence and Theorem 29 p.63).
- Prove that Theorem 6 and Theorem 2 apply and that θ_n converges to θ^*.
- Prove that if the function $\log p_\xi(x)$ is concave, condition (2.20) is satisfied with $f(x) = -\partial_x \log p_\xi(x)$ (p_ξ is the density of ξ_t).

Exercise 2.5.4 (Noisy gradient) We consider the algorithm of §2.4.2. Under the following assumptions

- e_n is an i.i.d. sequence, and $E[|e_1|^2] < \infty$
- $\sum \gamma_n = \infty$, $\sum \gamma_n^2 < \infty$
- Q is a C^1 function and $\nabla Q(x) = 0$ has finitely many solutions

show that θ_n tends a.s. to a solution to $\nabla Q(\theta) = 0$ (use Theorems 10, 29, and 2).

Exercise 2.5.5 (Quantile estimation) Let X_n be an i.i.d. sequence. We assume that the density $p(x)$ satisfies the regularity condition

$$\int |x|(|p'(x)| + |p''(x)|)\, dx < \infty.$$

For some $0 < \alpha < 1$, we want to minimize w.r.t θ the function

$$V(\theta) = \frac{1}{2} E[|\theta - X|] + \left(\frac{1}{2} - \alpha\right) \theta$$

where the variable X has the distribution of X_n.

- Write the corresponding stochastic gradient algorithm.
- Compute the mean field. What are the stationary points?
- Prove that $V(\theta)$ tends to $+\infty$ when $|\theta|$ tends to $+\infty$.

- Show that Exercise 2.5.8 applies and prove the convergence.

Exercise 2.5.6 (Noise cancellation) We assume the model

$$\begin{aligned} x_n &= s_n + \theta^* w_n + e_n \\ y_n &= \alpha s_n + w_n + f_n. \end{aligned}$$

with s_n, w_n, e_n and f_n are independent with variances σ_s, σ_w, σ_e, σ_f, and α is a real number. What is the mean field of (2.31)? What are the limit points?

Exercise 2.5.7 We consider the algorithms of Section 2.4.4 and look for Lyapunov functions V satisfying $\langle \nabla V, h\rangle \le 0$.

- Find a Lyapunov function for the algorithm (2.32).
- Let $\psi(x)$ be a real-valued function of the vector x. For x fixed, give the expression of the gradient of the function of the matrix B: $B \to \psi(Bx)$ (cf the note Section 2.4.4).
- Prove that for any square matrix M

  $$\mathrm{trace}(MM^T + M^2) \ge 0.$$

  (Hint: Use the Cauchy-Shwarz inequality or use that $\mathrm{trace}(AB) = \mathrm{trace}(BA)$, $\mathrm{trace}(A^T) = \mathrm{trace}(A)$ and $\mathrm{trace}(S^2) \ge 0$ for S symmetric.)
- Let g be the gradient of some function G. Find a Lyapunov function for the algorithm

  $$B_n = B_{n-1} - \gamma_n (g(y_n) y_n^T - y_n g(y_n)^T) B_{n-1}.$$

Exercise 2.5.8 (Recurrence and Lyapunov functions) We consider the general algorithm (2.2) with $r_n = 0$. Assume the existence of a Lyapunov function $V \ge 0$ with, for some constant C

$$\begin{aligned} &\sum_n \gamma_n^2 < \infty, \\ &|\nabla V(x+y) - \nabla V(x)| \le C|y|, \\ &|h|^2 \le C(1+V), \\ &E[e_n|\mathcal{F}_{n-1}] = 0, \\ &E[|e_n|^2|\mathcal{F}_{n-1}] < C(1+V(\theta_{n-1})). \end{aligned}$$

- Using the Taylor formula, write $V_n = V(\theta_n)$ as a function of V_{n-1}, $h(\theta_{n-1})$, e_n, and a second order term (use that ∇V is Lipschitz continuous).
- We assume that for any x, $\langle \nabla V(x), h(x)\rangle \le 0$. Calculate $E[V_n|\mathcal{F}_{n-1}]$ and show that for some C, the process

  $$S_n = -p_n(1 + V(\theta_n)), \quad p_n = (1 + C\gamma_{n+1}^2)(1 + C\gamma_{n+2}^2)...$$

  is a submartingale. Prove that almost surely the sequence $V(\theta_n)$ converges (use Theorem 28 p.63).

  Hence if $\lim_{x\to\partial\mathcal{O}} V(x) = +\infty$, θ_n remains in a compact subset of $\mathcal{O}$.

- We assume now that
$$\lim \sup_{x \to \partial \mathcal{O}} \langle \nabla V(x), h(x) \rangle < 0.$$
Show that the process
$$T_n = -p_n(1 + V(\theta_n)) + \sum_{i=0}^{n-1} p_{i+1}\gamma_{i+1}\langle \nabla V(\theta_i), h(\theta_i)\rangle, \quad n \geq 0$$
is a submartingale.
Prove that $T_n - \sup_{k<n} T_k$ is bounded by some constant depending only on V, h and θ_0.
(Hint: Consider the time $k(n) = \sup\{0 \leq j < n, \langle \nabla V(\theta_j), h(\theta_j)\rangle \geq 0\}$ with $k(n) = 0$ if this set is empty; note that $\theta_{k(n)}$ necessarily belongs to a fixed compact set).
Prove that almost surely the sequence θ_n does not tend to $\partial\mathcal{O}$ (use Theorem 31 p.64).

Exercise 2.5.9 (Costa loop) Consider the modulated phase signal a_n with the form
$$a_n = e^{i\psi_n}, \quad \psi_n \in \left\{\frac{\pi}{4} + k\frac{\pi}{2}, k = 0, 1, 2, 3\right\}.$$
We assume that (a_n) is an i.i.d. sequence and that the distribution of a_n is uniform over the set with 4 elements (each a_n represents two bits of information). We observe
$$y_n = (a_n + \xi_n)e^{i\theta^*}.$$
We assume that ξ_n is a complex Gaussian white noise:
$$\xi_n = \alpha_n + i\beta_n$$
where α_n and β_n are two i.i.d. independent sequences with variance $\sigma^2/2$, so that $E[\xi_n\bar{\xi}_n] = \sigma^2$. We consider the algorithm
$$\theta_n = \theta_{n-1} - \gamma_n \mathrm{Im}(y_n^4 e^{-4i\theta_{n-1}}).$$

- Compute for any non-negative integers p, q the value of $E[a_n^p \bar{a}_n^q]$.
- For $\lambda, \mu \in \mathbf{R}$, compute $E[e^{\lambda\xi_n + \mu\bar{\xi}_n}]$ and deduce, for any p and q the value of $E[\xi_n^p\bar{\xi}_n^q]$.
- Compute the mean field. What are the stationary points? What are the stable ones ($\nabla h(\theta) < 0$)?
- a_n is reconstructed from y_n through the formula
$$\hat{a}_n = \mathrm{sign}(y_n e^{-i\theta_n}), \quad \text{where} \quad \mathrm{sign}(a + ib) = \mathrm{sign}(a) + i\,\mathrm{sign}(b).$$
What is the influence of the limit point on the sequence $\hat{a}_n$?

Exercise 2.5.10 Let X be a square integrable two dimensional r.v. with independent components. Let A be a full rank matrix such that AX has two independent components.

- Show that there exist functions μ_1 and μ_2 such that the characteristic function of X writes $E[\exp(i(\alpha, \beta)X)] = \mu_1(\alpha)\mu_2(\beta)$
- Express the characteristic function $\nu(\alpha, \beta)$ of AX using μ_1 and μ_2
- Prove that either X is Gaussian (or constant) or A is a permutation matrix (hint: differentiate $\log(\nu)$ w.r.t. α and β).

Chapter 3

Stochastic models and pointwise convergence

In the preceding section we focused on the deterministic aspects of the algorithms, although the randomness was considered in the examples.

We will give models for the stochastic structure of the noise entering the algorithm, thus restraining the class in which equation (2.2) lies. This restriction will help to relate our results to practical situations and will allow us to study limit–theorems in the next chapter. We stick here to pointwise (almost sure) convergence.

Section 3.1 considers the class of the so-called Robbins-Monro algorithms. This class is not as general as the one considered in §3.3 (algorithms with Markovian dynamics), but has the enormous advantage to be much easier to study and to contain interesting examples.

3.1 Robbins-Monro algorithms

3.1.1 Theory

In this section we will study an important class of algorithms, the Robbins-Monro algorithm. These algorithms have the form

$$\theta_n \overset{pr}{=} \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n) \tag{3.1}$$

where Y_n is drawn from a distribution $P_{\theta_{n-1}}$ which depends on θ_{n-1} but Y_n does not depend otherwise on the past :

$$P(Y_n \in A | Y_{n-1}, Y_{n-2}, ..., Y_1, \theta_0) = P_{\theta_{n-1}}(A). \tag{3.2}$$

The algorithm tries to solve

$$E_\theta[H(\theta, Y)] = 0$$

where E_θ denotes the expectation under P_θ. If we consider the mean field

$$h(\theta) = E_\theta[H(\theta, Y)],$$

we can rewrite

$$\theta_n \overset{pr}{=} \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n$$

with

$$e_n = H(\theta_{n-1}, Y_n) - h(\theta_{n-1}).$$

We will make the following assumptions

(R0) $\sum \gamma_n = \infty, \quad \sum \gamma_n^2 < \infty.$

(R1) *h is continuous, the set $\mathcal{S} = \{\theta, h(\theta) = 0\}$ is finite and there exists a continuously differentiable function V such that*

$$\begin{aligned}
&\lim_{x \to \partial\mathcal{O}} V(x) = \infty, \\
&\langle \nabla V(\theta), h(\theta) \rangle \leq 0, \\
&\{\langle \nabla V(\theta), h(\theta) \rangle = 0\} = \mathcal{S}.
\end{aligned} \tag{3.3}$$

(R2) *For any compact set $\mathcal{K}$,*

$$\sup_{\theta \in \mathcal{K}} E[|H(\theta, Y)|^2] < \infty.$$

Theorem 13 *Under the assumptions (R0, R1, R2), the algorithm (3.1,3.2) converges to some point θ^*, $h(\theta^*) = 0$ with probability one.*

If no projection is done and we know that θ_n remains in some compact subset of $\mathcal{O}$ the assumption that V tends to infinity is unnecessary.

Remarks: We do not mean to give here the most general assumption set. It is easily shown that assumption (R1) may be weakened in order to look like assumption (A) and also that one could use two Lyapunov functions instead of one (one satisfying assumption (A) and the other one satisfying assumption (B)).

Exercise 2.5.8 may be used to prove that θ_n does not tend to $\partial\mathcal{O}$, and in this case the projections are superfluous.

Proof: For the projected algorithm, we will first apply Theorem 12 p.17. Assumption (B) is clearly satisfied. In order to check (2.29), we can apply Theorem 29 p.63 with $p = 2$ and

$$X_n = \gamma_n e_n 1_{V(\theta_{n-1}) \leq M} \tag{3.4}$$

$$S_n = \sum_{k=1}^{n} X_k. \tag{3.5}$$

We have indeed

$$E[|X_n|^2 | \mathcal{F}_{n-1}] = \gamma_n^2 1_{V(\theta_{n-1}) \leq M} E_{\theta_{n-1}}[|e_n|^2] \leq \gamma_n^2 \sup_{V(\theta) \leq M} E[|H(\theta, Y) - h(\theta)|^2] \leq C_M \gamma_n^2$$

for some constant C_M (since, by the assumption (3.3), $\{V \leq M\}$ is compact). Hence S_n converges a.s., Theorem 12 applies and the algorithm is A-stable. The conclusion is now a consequence of Theorem 2 p.8.

If no projection is done but we know that θ_n remains in some compact set, the same calculation proves directly that the algorithm is A-stable and we conclude in the same way. ∎

3.1.2 Example: Dosage

θ is here a real number which represents the dosage of several products. For each dosage we can make an experiment whose outcome is a real number Y_n (quantity of heat produced...) and we want to find the dosage which solves the equation

$$E_\theta[Y] = \alpha$$

where α is a given number. We assume that $E_\theta[Y]$ is a continuous strictly increasing function of θ. The algorithm is

$$\theta_n \stackrel{pr}{=} \theta_{n-1} + \gamma_n(\alpha - Y_n).$$

The mean field is here $h(\theta) = \alpha - E_\theta[Y]$. We can set

$$V(\theta) = -\int_0^\theta h(x)\ dx$$

and one shows easily the convergence of the algorithm to the solution θ^* if

$$\forall\ M > 0, \quad \sup_{|\theta|<M} E[Y^2] < \infty$$
$$\min_x h(x) < \alpha < \max_x h(x).$$

3.1.3 Example: The two-armed bandit

A slot machine has two arms. The winning probability with arm A (resp. B) is the unknown quantity p_A (resp. p_B). We want to estimate which machine has the highest winning probability. The strategy is as follows: at time n, we choose machine A (resp. B) with probability θ_{n-1} (resp. $1 - \theta_{n-1}$) and

$$\theta_n = \begin{cases} \theta_{n-1} & \text{if we lost} \\ \theta_{n-1} + \gamma_n(1 - \theta_{n-1}) & \text{if we won on machine } A \\ \theta_{n-1} - \gamma_n\theta_{n-1} & \text{if we won on machine } B \end{cases}$$

The point is that we hope that θ_n will converge to 1 (resp. 0) if A (resp. B) is the best machine.

Clearly, if $\theta_0 \in]0,1[$ then θ_n will remain in this interval; we may choose for instance $\mathcal{O} =]-1,2[$. The mean field is

$$\begin{aligned} h(\theta) &= P_\theta(\text{choosing } A \text{ and winning on it})(1-\theta) \\ &\quad +P_\theta(\text{choosing } B \text{ and winning on it})(-\theta) \\ &= \theta p_A(1-\theta) - (1-\theta)p_B\theta \\ &= \theta(1-\theta)(p_A - p_B). \end{aligned}$$

One can choose $V(\theta) = -(p_A - p_B)(\theta^2/2 - \theta^3/3)$ and (R1) is satisfied. By definition of h, $e_n = H(\theta_{n-1}, Y_n) - h(\theta_{n-1})$ is a martingale increment ($Y_n \in \{1,2,3,4\}$ combines here the choice of the machine and the outcome of the game); moreover, it is bounded by 2. Thus the algorithm is A-stable and the required assumptions are satisfied with $\mathcal{S} = \{0,1\}$. It remains to prove that the algorithm converges to the good solution θ^*. Unfortunately, we do not know if this is true! However, one checks easily that if θ^* if zero (resp. 1) then $E[\theta_n]$ is decreasing (increasing).

Open problem : Does θ_n converges to θ^* w.p.1 ?

3.2 Mixing noises

In the preceding sections, we explained why it is natural to try to write an algorithm in the form

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n + \gamma_n r_n$$

where

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n \quad \text{exists} \quad \text{and} \quad \lim_{n\to\infty} |r_n| = 0. \tag{3.6}$$

In §2.2.2 or §3.1, e_n appeared as a martingale increment. In many situations however e_n is not a martingale-increment but has more or less a mixing property. The best example will be given in §3.3. Next theorem deals with this situation. We will not use this result; it is here only to relate the assumption (3.6) with probability theory and for making use the model (3.7) in the future.

Theorem 14 *Let e_n be a process measurable w.r.t. an increasing sequence of σ-fields $\mathcal{F}_n$, and such that*

$$\sup_{n>0} \sum_{k=0}^{\infty} \|E[e_{n+k}|\mathcal{F}_n]\|_2 = \sigma < \infty$$

then, e_n rewrites

$$e_n = \xi_n + \nu_n - \nu_{n-1} \tag{3.7}$$

where ξ_n is a martingale-increment sequence, and $\|\nu_n\|_2 \le \sigma$, $\|\xi_n\|_2 \le 2\sigma$. If

$$\sum_n \gamma_n^2 < \infty, \qquad \sum_n |\gamma_n - \gamma_{n+1}| < \infty$$

then

$$\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n \quad \textit{exists a.s. and in } L_2.$$

Proof : Define the $\mathcal{F}_n$-adapted processes ξ_n and ν_n as

$$\begin{aligned}
\nu_n &= -\sum_{k=1}^{\infty} E[e_{n+k}|\mathcal{F}_n] \\
\xi_n &= e_n - \nu_n + \nu_{n-1}
\end{aligned}$$

then clearly

$$\|\nu_n\|_2 \le \sigma, \quad \|\xi_n\|_2 \le \|e_n - \nu_n\|_2 + \|\nu_{n-1}\|_2 \le 2\sigma$$

and ξ_n is a martingale-increment sequence since :

$$\begin{aligned}
\xi_n &= e_n + \sum_{k=1}^{\infty} E[e_{n+k}|\mathcal{F}_n] - \sum_{k=1}^{\infty} E[e_{n+k-1}|\mathcal{F}_{n-1}] \\
&= \sum_{k=0}^{\infty} E[e_{n+k}|\mathcal{F}_n] - E[e_{n+k}|\mathcal{F}_{n-1}].
\end{aligned}$$

And if $\sum \gamma_n^2 < \infty$,

$$\begin{aligned}\sum_{n=1}^{p} \gamma_n e_n &= \sum_{n=1}^{p} \gamma_n \xi_n + \sum_{n=1}^{p} \gamma_n(\nu_n - \nu_{n-1}) \\ &= \sum_{n=1}^{p} \gamma_n \xi_n + \sum_{n=1}^{p-1} (\gamma_n - \gamma_{n+1})\nu_n - \gamma_1\nu_0 + \gamma_n\nu_n.\end{aligned}$$

The first term is convergent because of Theorem 29 p.63. For the second, it is sufficient to prove that

$$\sum_{n=1}^{\infty} |\gamma_n - \gamma_{n+1}|\ |\nu_n| < \infty \quad a.s.$$

and this is true because the expectation is finite. It remains to demonstrate the convergence of $\gamma_n\nu_n$ to zero. It suffices to prove that

$$\sum_{n=1}^{\infty} \gamma_n^2 |\nu_n|^2 < \infty \quad a.s.$$

which is true since the expectation is finite. ∎

3.3 Algorithms with Markovian dynamics

3.3.1 Theory in the bounded case

This section is devoted to the case of algorithms with Markovian dynamics in the case of bounded field; the general case is technically difficult and will be treated in Section 3.3.

In this model introduced in [19, 2], the stochastic algorithm is defined by the following equation

$$\theta_n \overset{pr}{=} \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) \tag{3.8}$$

where the state X belongs to some space $\mathcal{X}$ (generally discrete or Euclidean), $H(\theta, X)$ is a function $\mathbf{R}^d \times \mathcal{X} \longrightarrow \mathbf{R}^d$, and X_{n+1} is randomly chosen from X_n through the θ-dependent transition probability $\Pi_{\theta_n}(x, dy)$:

$$P(X_{n+1} \in B | \theta_0, ..\theta_n, X_0, ..X_n) = \Pi_{\theta_n}(X_n, B). \tag{3.9}$$

θ_0 and X_0 are arbitrary; the projection done as explained in §2.4 with, in addition a projection of X_n in a compact subset $\mathcal{X}$, each time a projection of θ_n is made (the modified value of X_n may depend arbitrarily on the past)[1]. This form encompasses most algorithms. 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, we are in the Robins-Monro case.

We will denote also by Π_θ the transition operator of the Markov chain :

$$(\Pi_\theta f)(x) = \int f(y) \Pi_\theta(x, dy) = E_\theta[f(X_1 | X_0 = x)].$$

With this notation, one has

$$(\Pi_\theta^n f)(x) = E_\theta[f(X_n | X_0 = x)].$$

The proof requires some assumptions on the Markov chains parameterized with θ which are not very restrictive but not always easy to check; they are essentially borrowed to [19] :

(MS) *For every θ, there exists $h(\theta)$ such that the Poisson equation*

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

has a solution; for any compact set $\mathcal{C} \subset \mathcal{O}$,

$$\sup_{\theta \in \mathcal{C}, x \in \mathcal{X}} |H(\theta, x)| + |v_\theta(x)| < \infty$$

and there exist a continuous function $\phi_\mathcal{C}$, $\phi_\mathcal{C}(0) = 0$, such that for any $\theta, \theta' \in \mathcal{C}$,

$$\sup_{x \in \mathcal{X}} |\Pi_\theta v_\theta(x) - \Pi_{\theta'} v_{\theta'}(x)| \leq \phi_\mathcal{C}(|\theta - \theta'|).$$

[1] In some cases, convergence may also be proved without doing any projection [2], but the price to pay is much more stringent assumptions and very intricate proofs.

Remarks : Typically, for every θ the Markov chain driven by Π_θ has a unique invariant probability $\mu_\theta(dx)$ and $h(\theta)$ is the expectation of $H(\theta, x)$ under this distribution : $h(\theta) = \int H(\theta, x)\mu_\theta(dx)$. One method for verifying (MS) is to write v_θ as a rapidly converging series :

$$v_\theta(.) = \tilde{H}(\theta, .) + \Pi_\theta \tilde{H}(\theta, .) + \Pi_\theta^2 \tilde{H}(\theta, .) + ... \qquad \tilde{H} = H - h. \tag{3.10}$$

Conditions under which this solution exists and satisfies the regularity condition are given in Theorem 35 p.71. The boundedness condition on H is rather strong, but simplifies enormously the proof.

We have the following result

Theorem 15 *Assume that θ_n is given by algorithm (3.24) and assume that*

(i) *assumptions (MS), (A) and (B) are satisfied (h is given by (MS)),*

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

then θ_n converges almost surely to $\mathcal{S}$.

Remarks : It appears clearly from the second step of the proof that the projections are unneccesary if one can prove that even without projection θ_n remain a.s. in a compact set (which may depend on the sample path).

Proof : We shall use Theorems 10 p.16 and 2 p.8. The proof reuses the ideas of [19], but the projections here greatly simplify the task. The basic idea is 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 :

$$\begin{aligned} H(\theta_n, X_{n+1}) &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= h(\theta_n) + v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_n) \\ &\quad + \Pi_{\theta_n} v_{\theta_n}(X_n) - \Pi_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) \\ &\quad + \Pi_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= h(\theta_n) + \xi_{n+1} + (\nu_n - \nu_{n+1}) + r_{n+1}. \end{aligned} \tag{3.11}$$

We are going to prove that for any q, almost surely (cf (2.27))

$$\lim_n I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i(\xi_i + \nu_{i-1} - \nu_i) = 0,$$
$$\lim_n I_n r_{n+1} = 0, \quad I_n = I_n^q = I(\theta_{\tau(n)}..., \theta_n \in \mathcal{K}_q).$$

$\tau(n)$ is the last projection time before n ($\tau(n) \leq n$).

We study separately each term on the sample paths such that $\tau(n)$ tends to infinity. Assumption (MS) implies that for $\theta \in \mathcal{K}_q$

$$|\Pi_\theta v_\theta(x)| \leq |v_\theta(x)| + |H(\theta, x)| + |h(\theta)| \leq C.$$

Thus $I_n \xi_{n+1}$ is an L_2-bounded martingale-increment sequence and the martingale $\sum \gamma_i I_{i-1} \xi_i$ converges; this implies that

$$\lim_n I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i \xi_i = \lim_n I_n \left(\sum_{i=1}^{n+1} \gamma_i I_{i-1} \xi_i - \sum_{i=1}^{\tau(n)} \gamma_i I_{i-1} \xi_i \right) = 0.$$

On the other hand, we have

$$I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i(\nu_{i-1} - \nu_i) \quad = \quad I_n \sum_{i=\tau(n)+1}^{n} (\gamma_{i+1} - \gamma_i)\nu_i + \gamma_{\tau(n)+1} I_n \nu_{\tau(n)} - \gamma_{n+1} I_n \nu_{n+1}.$$

The boundedness of ν_k for $\theta_k \in \mathcal{K}_q$ implies that the first sum is absolutely convergent and that both last terms tends to zero. The convergence of the last remainder to zero is almost immediate: for n large enough, the projection set contains $\mathcal{K}_{q+1}$, and on the other hand $\theta_n \in \mathcal{K}_q$ implies that $\hat\theta_{n+1} \in \mathcal{K}_{q+1}$; hence no projection occurs if $I_n = 1$ and we have

$$\begin{aligned} I_n |r_{n+1}| &= I_n |\Pi_{\theta_{n+1}} v_{\theta_{n+1}}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1})| \\ &\leq I_n \phi_{\mathcal{K}_{q+1}}(|\theta_{n+1} - \theta_n|) \\ &= I_n \phi_{\mathcal{K}_{q+1}}(\gamma_n |H(\theta_n, X_{n+1})|). \end{aligned}$$

Equation (2.27) is thus satisfied. Hence Theorem 10 applies and the sequence θ_n is a.s. bounded. The same arguments on series show that A-stability occurs and Theorem 2 applies. ∎

3.3.2 Example: Blind equalizers

A zero-mean i.i.d. sequence (a_i) (the symbol sequence) with finite number of possible real values s_1,...s_n is emitted and the received signal is

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

where (c_i) is the channel filter and ξ_n an i.i.d. bounded white noise sequence with smooth density, independent of the a_n's. The receiver has to reconstruct the (a_i) sequence from the y_is. A popular simple method is to use a vector θ which approximately inverts the channel

$$\sum_{i=0}^{q} \theta_i c_{n-i} \simeq 1_{n=0}.$$

and to reconstruct a_n as

$$\hat a_n = \arg\min_{s_i} |Y_n^T \theta - s_i|, \quad Y_n = (y_n, ..y_{n-p})^T$$

(this works if the signal to noise ratio is really small). If the channel is unknown, θ has to be estimated directly from the observations. A proposed algorithm [1] is a gradient one based on some contrast function ψ:

$$\theta_n \stackrel{pr}{=} \theta_{n-1} - \gamma_n Y_n \psi'(Y_n^T \theta_{n-1}) \tag{3.12}$$

This algorithm is a stochastic gradient with cost function

$$V(\theta) = E[\psi(Y_n^T \theta)].$$

The consistency of the design requires in particular that

$$E[Y_n \psi'(a_n)] = 0$$

that is

$$E[a_n \psi'(a_n)] = 0.$$

A possible choice for ψ may be $\psi(x) = \psi_1(x) - \alpha\psi_2(x)$ where ψ_1 and ψ_2 are some function and $\alpha = E[a_n\psi_1'(a_n)]/E[a_n\psi_2'(a_n)]$; the case $\psi_1(x) = x^2/2$ and $\psi_2(x) = |x|$ is considered in [2]. One can show that a minimizer of $V(\theta)$ will be a good candidate for the inverse filter (i.e. $Y_n^T\theta^* \simeq a_{n-p}$ for some p, cf [2] pp. 60-70); however this property depends strongly on the choice of ψ.

We are in the situation of §3.3 with

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

X_n is a Markov chain with a simple structure and $\Pi_\theta = \Pi$ is independent of θ, satisfying

$$\Pi^{p+q+1} = \pi.$$

Hence

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

Assumption (MS) is straightforwardly checked if ψ is smooth enough. Since $h(\theta) = -\nabla V(\theta)$, V is a good candidate for the Lyapunov function of assumptions (A) and (B).

3.3.3 Example: AR estimation

We consider here the case of the estimation of an AR(2) process. The extension to AR(n) is immediate.

Construction of the algorithm

We observe a stationary process y_n of the form

$$y_n = e_n - \theta^* y_{n-1} \tag{3.13}$$

where e is an i.i.d. bounded sequence with zero mean and variance σ (if boundedness is not satisfied, one has to use the results of Section 3.3.5). θ^* is the true unknown parameter, lying in $]-1,1[$. It is easy to verify that for any value $\theta \in]-1,1[$, if we define the stationary process

$$\varepsilon_n = y_n + \theta y_{n-1} \tag{3.14}$$

then $E[\varepsilon_n^2] > E[e_n^2]$ if $\theta \neq \theta^*$ and $e_n = \varepsilon_n$ if $\theta = \theta^*$; thus

$$\theta^* = \arg\min_\theta E[\varepsilon_n^2]. \tag{3.15}$$

The corresponding gradient algorithm is

$$\theta_n \overset{pr}{=} \theta_{n-1} - \gamma_n\varepsilon_n y_{n-1} \tag{3.16}$$
$$\varepsilon_n = y_n + \theta_{n-1}y_{n-1}. \tag{3.17}$$

The projections may be made with for instance $\mathcal{K}_n =]-n, n[$. We are in the situation of §3.3 with

$$\begin{aligned} X_n &= (y_n, y_{n-1}) \\ H(\theta, x) &= -(x_1 + \theta x_2)x_2 \end{aligned}$$

and

$$X_n = \begin{pmatrix} -\theta^* & 0 \\ 1 & 0 \end{pmatrix} X_{n-1} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} e_n = FX_{n-1} + Ge_n.$$

Notice that $\Pi_\theta = \Pi$ does not depend on θ.

Checking the assumptions

Let us start with assumptions (A,B) and calculate the mean field. From equation (3.17), we observe that y_n is orthogonal to $e_{n-1}, e_{n-2}, ...$, and one obtains after elementary calculations (here $\theta_n = \theta$ is fixed)

$$\begin{aligned} E[y_n^2] &= \sigma^2/(1-\theta^{*2}) \\ E[y_n y_{n-1}] &= -\theta^* \sigma^2/(1-\theta^{*2}). \end{aligned}$$

Thus

$$h(\theta) = -\sigma^2 \frac{\theta - \theta^*}{1-\theta^{*2}}$$

which is the negated gradient of

$$V(\theta) = \frac{\sigma^2}{2} \frac{(\theta-\theta^*)^2}{1-\theta^{*2}} = \frac{E[\varepsilon_n^2] - \sigma^2}{2}$$

(verification of the last equality is left to the reader). This function is an adequate Lyapunov function for assumptions (A) and (B). Let us proceed with assumptions (MS) and calculate the solution to the Poisson equation. Notice that

$$\begin{aligned} \Pi^n H(\theta, x) &= -E[(\xi_1 + \theta \xi_2)\xi_2], \qquad \xi = F^n x + F^{n-1} G e_1 + F^{n-2} G e_2 + ... G e_n \\ &= -((F^n x)_1 + \theta (F^n x)_2)(F^n x)_2 + c_n(\theta) \end{aligned}$$

where $c_n(\theta)$ is some function of θ only. This implies that a solution to the Poisson equation is a linear function of θ:

$$v_\theta(x) \;=\; -\sum_{n=0} ((F^n x)_1 + \theta (F^n x)_2)(F^n x)_2 = x^T Q x + \theta\, x^T R x$$

for some matrices Q and R. Assumption (MS) is now immediate.

3.3.4 Example: Q-learning

This example is slightly more complicated than the Robbins-Monro case but simpler that the general Markovian dynamics since Π_θ does not depend on θ.

We describe here the problem as stated in [29] (see also [4] for extensions). Consider a computational agent (controller) moving around some finite world, choosing one from a finite collection of actions (controls) at every time step. The world constitutes a Markov process controlled by the agent. At step n the controller is equipped to register the state $X_n \in \mathcal{X}$ of the world and can choose its control $U_n \in \mathcal{U}$ accordingly. This action has a random cost C_{n+1} whose distribution knowing the past depends only on the state and the action; the state changes according to a probabilistic law which depends on U_n:

$$\begin{aligned}
&\mathcal{F}_n = \sigma(X_i, U_i, C_i,\ i = 1, ...n)\\
&P(C_{n+1} \in A|\mathcal{F}_n) = P(C_{n+1} \in A|X_n, U_n),\\
&E[C_{n+1}|\mathcal{F}_n] = c(x, u), \quad x = X_n \quad u = U_n,\\
&P(X_{n+1} = y|\mathcal{F}_n) = p^u_{xy}, \quad x = X_n \quad u = U_n,
\end{aligned}$$

In the classical control problem (which is not ours), the three-dimensional array p^u_{xy} is known to the controller, as well as the function $c(x, u)$. The task facing him is that of determining an optimal policy π (a function which determines U_n from the past values of the state), i.e. one that minimizes the cost-to-go ($\rho < 1$)

$$J_\pi(x_0) = \sum_{n=0}^{\infty} \rho^n E_\pi[c(X_n, U_n)], \qquad X_0 = x_0. \tag{3.18}$$

E_π denotes the distribution of the sequence $X_1, U_1, ...$ when the policy π is chosen, and here $X_0 = x_0$. The parameter ρ (discount factor) guarantees the convergence and is not critical; one can avoid it by considering costs on a finite window but the theory in this case gives similar results with more complicated proofs. We have the following classical result

Theorem 16 *An optimal policy π^* exists which is independent of the initial state x_0 and is a function only of the last observed state:*

$$U_n = \pi^*(X_n).$$

$J^* = J_{\pi^*}$ is the solution to the Bellman equation

$$J^*(x) = \min_u c(x, u) + \rho \sum_y p_{xy}(u) J^*(y)$$

and $\pi^*(x)$ is the argument of the min *above* [2].

Unlike in Section 3.1.3, we will not try to simultaneously estimate π^* and go to it: a policy π_r will be given and the estimation of π^* will go on while choosing the controls with π_r. A choice of the controls with the current estimate of π^* would lead to the convergence problems encountered in Section 3.1.3.

[2] Notice that the solution to the Bellman equation is unique due to the fact that the mapping

$$J(.) \longrightarrow \min_u c(x, u) + \rho \sum_y p_{.y}(u) J(y)$$

is a contraction for the L_∞ norm.

For any given policy π, we define

$$Q_\pi(x,u) = c(x,u) + \rho \sum_y p_{xy}(u) J_\pi(y)$$

which is the cost-to-go if the control u (non-necessarily $\pi(x)$) is chosen at the first step; then clearly

$$J^*(x) = \min_u Q^*(x,u).$$

Our objective reduces to the estimation Q^*. The traditional Q-learning algorithm is the following scheme

$$\begin{aligned} Q_n(x,u) &= \begin{cases} Q_{n-1}(x,u) + \gamma_n(C_n + \rho J_{n-1}(X_n) - Q_{n-1}(x,u)), & \text{if} \quad (x,u) = (X_{n-1}, U_{n-1}) \\ Q_{n-1}(x,u) \quad \text{otherwise.} \end{cases} \\ J_{n-1}(y) &= \min_u Q_{n-1}(y,u) \end{aligned} \tag{3.19}$$

It remains now to specify how the controls U_n are chosen; in order to update all the entries of Q_n, it is necessary to introduce some randomization in the policy (dual control). We assume that U_n is chosen from a randomized policy

$$U_n = \pi_r(X_n, W_n) \tag{3.20}$$

where W_n is some i.i.d sequence. We will assume the following:

$$\|C_n\|_\infty < \infty \tag{3.21}$$

$$\forall x, u, \quad P(\pi_r(x, W_n) = u) > 0 \tag{3.22}$$

$$\min_{xy} \max_u p_{xy}(u))_{xy} > 0. \tag{3.23}$$

One proves that under these assumptions, the sequence Q_n of the Q-learning algorithm (3.19), (3.20), converges to Q^* (Exercise 3.4.7).

3.3.5 General theory

We consider here the case where $H(\theta, X)$ is not bounded (cf assumption (MS)). The notations we use here refer to those of Section 3.3.1. The stochastic algorithm is defined by the following equation

$$\theta_n \overset{pr}{=} \theta_{n-1} + \gamma_n \psi_n(H(\theta_{n-1}, X_n)) \tag{3.24}$$

where we are in the situation of Section 3.3.1:

$$P(X_{n+1} \in B | \theta_0, ..\theta_n, X_0, ..X_n) = \Pi_{\theta_n}(X_n, B). \tag{3.25}$$

the projections are done as explained in §2.4, with, in addition a projection of X_n in a compact subset $\mathcal{Q}$ of $\mathcal{X}$, each time a projection of θ_n is made (the modified value of X_n may depend arbitrarily on the past). The function ψ_n is a projection on a ball of increasing radius which is necessary only if Π_θ depends on θ and X (cf the assumptions in the theorem below); it ensures that $\theta_n - \theta_{n-1}$ tends to zero, a key condition for the stability of the process X_n.

We shall show in the proof that the projection ψ_n is made only a finite number of times.

The proof requires some assumptions on the Markov chains parameterized with θ which are not very restrictive but not always easy to check (see examples below and Appendix B); they are essentially borrowed to [19]:

(M1) *For every compact set $\mathcal{K}$ there exist a continuous function $V(x) \geq 1$, $\alpha < 1$, $n_0 > 0$ and C such that for all $\theta, \theta' \in \mathcal{K}$*

$$\begin{aligned} \Pi_\theta^{n_0} V &\leq \alpha V + C \\ \Pi_\theta V &\leq CV \\ |\Pi_\theta^n V - \Pi_{\theta'}^n V| &\leq C|\theta - \theta'|V, \quad 1 \leq n \leq n_0 - 1. \end{aligned}$$

(M2) *For every θ, there exists $h(\theta)$ such that the Poisson equation*

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

has a solution; for every compact set $\mathcal{K}$ there exists $C > 0$ such that $\theta, \theta' \in \mathcal{K}$ implies

$$\begin{aligned} &|H(\theta, .)| + |v_\theta| \leq CV^{1/2}, \\ &|\Pi_\theta v_\theta - \Pi_{\theta'} v_{\theta'}| \leq C|\theta - \theta'|V^{1/2}. \end{aligned}$$

Remarks : The Lyapunov function V is the same in both assumptions and may depend on $\mathcal{K}$.

If Π_θ has a unique invariant mesure $\pi_\theta(dx)$, $h(\theta)$ will be necessarily $\pi_\theta(H(\theta, x))$ (apply π_θ to both sides of the Poisson equation).

We have the following result :

Theorem 17 *Assume that θ_n is given by algorithm (3.24) and assume that*

(i) *assumptions (M1,M2) and (A,B) are satisfied (h is given by (M2)),*

(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 $n_0 = 1$ then $\psi_n(x) = x$. Otherwise*

- *for some $C > 0$ and $\eta < 1$: $\sum \gamma_n^{2\eta} < \infty$ and $\psi_n(x) = x$ if $|x| < C\gamma_n^{-\eta}$*
- $\lim_{n\to\infty} \gamma_n \|\psi_n\|_\infty = 0$,

then θ_n converges almost surely to $\mathcal{S}$. Furthermore, the sequence $V(X_n)$ is bounded in L_1 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.

The function ψ_n is typically the projection on the ball of radius $\gamma_n^{-\eta}$.

Proof: We shall use Theorem 10 p.16. The basic idea is to show that the projection ψ_n is made only a finite number of times and then to decompose $e_{n+1} = H(\theta_n, X_{n+1}) - h(\theta_n)$ by using Poisson equation and making appear two remainders :

$$\begin{aligned} H(\theta_n, X_{n+1}) - h(\theta_n) &= v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_n) \\ &\quad + \Pi_{\theta_n} v_{\theta_n}(X_n) - \Pi_{\theta_n} v_{\theta_n}(X_{n+1}) \\ &= \xi_{n+1} + \xi'_{n+1}. \end{aligned}$$

We shall finally prove that, for any increasing compact covering $\mathcal{K}_q$, (2.27) is satisfied with $r_n = 0$:

$$\lim_n I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i(\xi_i + \xi'_i) = 0, \quad I_n = I_n^q = I(\theta_{\tau(n)} ..., \theta_n \in \mathcal{K}_q). \tag{3.26}$$

Notice that since q is fixed, the subscript q will be omitted when unnecessary. This will give us the boundedness of θ_n. We will then prove that A-stability occurs.

Step 1. The aim of this step is to give the following bound on $V(X_n)$:

$$\sup_n E[I_{n-1} V(X_n)] < \infty. \tag{3.27}$$

We shall use an auxiliary Lyapunov function W defined as

$$W(\theta, x) = \beta^{n_0 - 1} V(x) + \beta^{n_0 - 2}(\Pi_\theta V)(x) + ... + (\Pi_\theta^{n_0 - 1} V)(x) \tag{3.28}$$

(n_0 comes from (M1) with $\mathcal{K} = \mathcal{K}_q$) where β is such that

$$\beta^{n_0} = \alpha.$$

Notice that, for $\theta, \theta' \in \mathcal{K}_q$,

$$\begin{aligned} W \le CV &\le C^2 W \\ \Pi_\theta W(\theta, x) &\le \beta W(\theta, x) + C \\ |W(\theta, x) - W(\theta', x)| &\le C|\theta - \theta'| W(\theta, x) \end{aligned}$$

for some constant C depending on q. The first and the third property are obvious from (M1); the second one follows from (3.28), (M1) and the choice of β.

We consider first the case where the projections ψ_n are used (cf assumption (iii)); set

$$u_n = \gamma_n \|\psi_n\|_\infty$$

and notice that if θ_n is not projected:

$$|\theta_n - \theta_{n-1}| \le u_n.$$

This happens if $\tau(n) < n$. Set $\mathcal{F}_n = \sigma(\theta_0, \theta_1, ...\theta_n, X_0, X_1, ...X_n)$, then on the $\mathcal{F}_n$-measurable set where $\tau(n) < n$:

$$\begin{aligned} E[I_n W(\theta_n, X_{n+1}) | \mathcal{F}_n] &= I_n (\Pi_{\theta_n} W)(\theta_n, X_n) \\ &\le I_n \beta W(\theta_n, X_n) + C \\ &\le I_n \beta (W(\theta_{n-1}, X_n) + C u_n W(\theta_{n-1}, X_n)) + C \\ &\le I_{n-1} \beta (1 + C u_n) W(\theta_{n-1}, X_n) + C \end{aligned}$$

because $\tau(n) < n$ implies $I_n = I_{n-1} = 1$. If $\tau(n) = n$:

$$\begin{aligned} E[W(\theta_n, X_{n+1})|\mathcal{F}_n] &= (\Pi_{\theta_n} W)(\theta_n, X_n) \\ &\leq \sup_{\theta\in\mathcal{K}_0} \sup_{x\in\mathcal{Q}} E[W(\theta, X_1)|\theta_0 = \theta, X_0 = x] \end{aligned}$$

Finally, in any case

$$E[I_n W(\theta_n, X_{n+1})|\mathcal{F}_n] \quad \leq \quad I_{n-1}(\beta + C u_n) W(\theta_{n-1}, X_n) + C.$$

Since $I_{n-1}u_n$ can be made arbitrarily small by choosing n is large enough, i.e. $n > n_1$, and n_1 is a non-random integer, we have

$$E[I_n W(\theta_n, X_{n+1})|\mathcal{F}_n] \quad \leq \quad \beta' I_{n-1} W(\theta_{n-1}, X_n) + C.$$

For some $\beta' < 1$ and C (C has increased for taking into account the terms for $n < n_1$). Taking expectation on both sides, we have, for any n

$$E[I_n W(\theta_n, X_{n+1})] \quad \leq \quad \beta' E[I_{n-1} W(\theta_{n-1}, X_n)] + C.$$

The expectations are bounded and $V \leq CW$ implies that equation (3.27) holds.

If $\psi_n(x) = x$ (no projection), $n_0 = 1$ or Π_θ does not depend on θ; in this situation we take $W = V$, u_n is not used (since W does not depend on θ), and the same conclusion is easily obtained.

Step 2. We prove here that Theorem 10 applies.

Assumption (M2) and (iii) imply that $\sum \gamma_n^{2\eta} |H(\theta_{n-1}, X_n)|^2 I_{n-1}$ is a.s. finite (since the expectation is finite), hence $\gamma_n^\eta I_{n-1} H(\theta_{n-1}, X_n)$ tends a.s. to zero, and the projection ψ_n is almost surely made only a finite number of times, for the ns such that $I_{n-1} \neq 0$, that is

$$\exists\ n(\omega), \quad \forall n > n(\omega), \quad I_{n-1}(\theta_n - \theta_{n-1}) = I_{n-1}\gamma_n H(\theta_{n-1}, X_n).$$

The main thing now is to prove that there is only finitely many Chen-projections; reasonning, by contradiction, we will consider the sample paths such that $\tau(n)$ tends to infinity and show that Theorem 10 applies. The remark above concerning the ψ_n-projections implies that the almost sure convergence of the series (2.27) to zero is equivalent to the convergence of (3.26). We study separately each term of this series.

Assumptions (M1-3) imply that for some C'

$$|\Pi_\theta v_\theta(x)| \leq |v_\theta(x)| + |H(\theta, x)| + |h(\theta)| \leq C' V(x)^{1/2}.$$

Thus $I_n \xi_{n+1}$ is an L_2-bounded martingale-increment sequence and the martingale $\sum \gamma_i I_{i-1} \xi_i$ converges; since $I_n = I_n I_i$ for $\tau(n) \leq i \leq n$, this implies that

$$\lim_n I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i \xi_i = \lim_n I_n \left(\sum_{i=1}^{n+1} \gamma_i I_{i-1} \xi_i - \sum_{i=1}^{\tau(n)} \gamma_i I_{i-1} \xi_i \right) = 0.$$

We consider now ξ'_n.

$$I_n \sum_{i=\tau(n)+1}^{n+1} \gamma_i \xi'_i \quad = \quad I_n \sum_{i=\tau(n)}^{n} \gamma_{i+1} (\Pi_{\theta_i} v_{\theta_i}(X_i) - \Pi_{\theta_i} v_{\theta_i}(X_{i+1}))$$

$$\begin{aligned} &= I_n \sum_{i=\tau(n)+1}^{n} I_{i-1}\gamma_{i+1}(\Pi_{\theta_i}v_{\theta_i}(X_i) - \Pi_{\theta_{i-1}}v_{\theta_{i-1}}(X_i)) \\ &\quad + I_n \sum_{i=\tau(n)+1}^{n} I_{i-1}(\gamma_{i+1}-\gamma_i)\Pi_{\theta_{i-1}}v_{\theta_{i-1}}(X_i) \\ &\quad + I_n\gamma_{\tau(n)+1}\Pi_{\theta_{\tau(n)}}v_{\theta_{\tau(n)}}(X_{\tau(n)}) - I_n\gamma_n\Pi_{\theta_n}v_{\theta_n}(X_{n+1}) \end{aligned}$$

For the first term we have absolute convergence:

$$\begin{aligned} \sum_{i=1}^{\infty} I_{i-1}\gamma_{i+1}|\Pi_{\theta_i}v_{\theta_i}(X_i) - \Pi_{\theta_{i-1}}v_{\theta_{i-1}}(X_i)| &\leq C'\sum_{i=1}^{\infty} I_{i-1}\gamma_{i+1}|\theta_i - \theta_{i-1}|V(X_i)^{1/2} \\ &\leq C''\sum_{i=1}^{\infty} I_{i-1}\gamma_i\gamma_{i+1}|H(\theta_{i-1},X_i)|V(X_i)^{1/2} \\ &\leq C'''\sum_{i=1}^{\infty} I_{i-1}(\gamma_i^2+\gamma_{i+1}^2)V(X_i) \\ &< \infty \quad a.s. \end{aligned}$$

since the expectation is finite. For the second term also:

$$\sum_{i=1}^{\infty} I_{i-1}|\gamma_{i+1}-\gamma_i|\ |\Pi_{\theta_{i-1}}v_{\theta_{i-1}}(X_i)| \leq C\sum_{i=1}^{\infty} I_{i-1}V(X_i)^{1/2}|\gamma_i-\gamma_{i+1}| < \infty \quad a.s.$$

The second term term tends to zero since both $\theta_{\tau(n)}$ and $X_{\tau(n)}$ are in compact sets. The last term tends to zero since it is square summable:

$$\sum_{n=1}^{\infty} I_n\gamma_n^2|\Pi_{\theta_n}v_{\theta_n}(X_{n+1})|^2 \leq \sum_{n=1}^{\infty}\gamma_n^2 V(X_{n+1}) < \infty \quad a.s.$$

Finally, since (3.26) converges, we have proved that (2.27) holds and Theorem 10 applies. Thus $\tau(n)$ is a.s. bounded, the sequence θ_n remains a.s. in a compact set and the A-stability ensues by the same decomposition of series as above. Theorem 2 implies the convergence of θ_n to some point of $\mathcal{S}$. ∎

3.3.6 Example: MA estimation

We proceed now with the case of the estimation of an MA(2). Such a process y_n has the form

$$y_n = e_n + \theta^* e_{n-1} \tag{3.29}$$

where e is an i.i.d. sequence with zero mean and variance σ. θ^* is the true unknown parameter, lying in $]-1,1[$. It is easy to verify that, if for any value $\theta \in]-1,1[$ we define the stationary process

$$\varepsilon_n = y_n - \theta\varepsilon_{n-1} = y_n - \theta y_{n-1} + \theta^2 y_{n-2} + \ldots \tag{3.30}$$

then, in the stationary regime, $E[\varepsilon_n^2] > E[e_n^2]$ if $\theta \neq \theta^*$ and $e_n = \varepsilon_n$ if $\theta = \theta^*$; thus

$$\theta^* = \arg\min_\theta E_\theta[\varepsilon_n^2]. \tag{3.31}$$

The idea of the RML algorithm is to built a gradient algorithm based on (3.31). For this we rewrite the cost as

$$E_\theta[\varepsilon_n^2] = E[\varepsilon_n(\theta)^2] \tag{3.32}$$

where ε_n is now considered as a function of θ through equation (3.30). The derivative ε_n' of ε_n w.r.t. θ is well defined as the solution to

$$\varepsilon_n' = -\theta\varepsilon_{n-1}' - \varepsilon_{n-1}.$$

These considerations lead to the algorithm

$$\theta_n \stackrel{pr}{=} \theta_{n-1} - \gamma_n \varepsilon_n \varepsilon_n' \tag{3.33}$$
$$\varepsilon_n = y_n - \theta_{n-1}\varepsilon_{n-1} \tag{3.34}$$
$$\varepsilon_n' = -\theta_{n-1}\varepsilon_{n-1}' - \varepsilon_{n-1}. \tag{3.35}$$

The projections may be made with for instance $\mathcal{K}_n = [-1+1/n, 1-1/n]$. We are in the situation of §3.3 with

$$\begin{aligned} X_n &= (e_n, \varepsilon_n, \varepsilon_n')^T \\ H(\theta, x) &= -x_2 x_3 \end{aligned}$$

and

$$X_n = \begin{pmatrix} 0 & 0 & 0 \\ \theta^* & -\theta & 0 \\ 0 & -1 & -\theta \end{pmatrix} \begin{pmatrix} e_{n-1} \\ \varepsilon_{n-1} \\ \varepsilon_{n-1}' \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} e_n = F_\theta X_{n-1} + G e_n.$$

Checking the assumptions

We start with assumptions (A,B) and calculate the mean field. From equations (3.34, 3.35) we observe that ε_n' is orthogonal to $e_n, e_{n-1}, \ldots$, and one obtains after elementary calculations (here $\theta_n = \theta$ is fixed)

$$\begin{aligned} E[\varepsilon_n e_n] &= \sigma^2 \\ E[\varepsilon_n^2] &= (1 + |\theta^*|^2 - 2\theta^*\theta)\sigma^2/(1-\theta^2) \\ E[\varepsilon_n \varepsilon_n'] &= -\theta^2 E[\varepsilon_{n-1}\varepsilon_{n-1}'] + \sigma^2(\theta - \theta^*)\frac{1-\theta\theta^*}{1-\theta^2}. \end{aligned}$$

Thus

$$h(\theta) = -\sigma^2(\theta - \theta^*)\frac{1-\theta\theta^*}{(1-\theta^2)^2}.$$

This function is the negated gradient of

$$V(\theta) = \sigma^2 \frac{1 + \theta^{*2} - \theta\theta^*}{2(1-\theta^2)}$$

which corresponds to (3.32). Hence $V(\theta)$ is an adequate Lyapunov function for assumptions (A) and (B).

Let us proceed with assumptions (M1) and (M2), and calculate the solution to the Poisson equation. Notice that

$$\begin{aligned} \Pi_\theta^n H(\theta, x) &= E[-(F_\theta^n x + F_\theta^{n-1} G e_1 + F_\theta^{n-2} G e_2 + \ldots G e_n)_2 (F_\theta^n x + F_\theta^{n-1} G e_1 + \ldots)_3] \\ &= -(F_\theta^n x)_2 (F_\theta^n x)_3 + c_n(\theta) \end{aligned}$$

where c_n is some function of θ only. This implies that a solution to the Poisson equation is

$$\begin{aligned} v_\theta(x) &= \sum_{n=0} -(F_\theta^n x)_2 (F_\theta^n x)_3 \\ &= -\frac{1}{2i\pi} \int [(1-zF_\theta)^{-1}x]_2 [(1-z^{-1}F_\theta)^{-1}x]_3 \frac{dz}{z}. \end{aligned}$$

Assumptions (M1) and (M2) are now easily checked with $V(x) = |x|^4 + 1$ (Exercise 3.4.8). The above formula leads to

$$v_\theta(x) = (1-\theta^2)^{-2}(x_2 - \theta\theta^* x_1)(\theta^* x_1 - \theta x_2 - (1-\theta^2)x_3).$$

3.3.7 Example: ARMA estimation

The model for the true signal is now

$$y_n + a_1^* y_{n-1} \ldots + a_p^* y_{n-p} = e_n + b_1^* e_{n-1} \ldots + b_q^* e_{n-q}$$

which rewrites

$$\begin{aligned} y_n &= e_n + f_n^T \theta^* &\quad (3.36) \\ f_n &= (-y_{n-1}, \ldots, -y_{n-p}, e_{n-1}, \ldots e_{n-q})^T &\quad (3.37) \end{aligned}$$

with the notation

$$\theta = (a_1 \ldots, a_p, b_1 \ldots, b_q)^T.$$

The prediction error is $y_n - f_n^T\theta$. Considerations inspired from sections 3.3.3 and 3.3.6 lead to the algorithm

$$\begin{aligned} \theta_n &\overset{pr}{=} \theta_{n-1} - \gamma_n \varepsilon_n \Phi_n \\ \varepsilon_n &= y_n - \varphi_n^T \theta_{n-1} \\ \varphi_n &= (-y_{n-1} \ldots, -y_{n-p}, -\varepsilon_{n-1} \ldots, -\varepsilon_{n-q}) \\ \Phi_n &= -\varphi_n - \sum_{i=1}^{q} b_i(n-1)\Phi_{n-i}. \end{aligned}$$

ε_n is the estimated prediction error, φ_n is the estimation of f_n, and Φ_n is its gradient w.r.t. θ. Projection sets may be defined by constraints on the zeroes or on the reflection coefficients of the MA part $(b_1 \ldots, b_q)$. With the state

$$X_n = (e_n \ldots, e_{n-q}, \varphi_n, \Phi_n \ldots, \Phi_{n-q+1}),$$

the algorithm rewrites

$$\begin{aligned} \theta_n &\overset{pr}{=} \theta_{n-1} - \gamma_n H(\theta_{n-1}, X_n) \\ X_n &= F(\theta_{n-1})X_{n-1} + Ge_n. \end{aligned}$$

The study is more technical but essentially similar to what has been done in Sections 3.3.3 and 3.3.6.

3.4 Exercises

Exercise 3.4.1 Prove the convergence of the algorithm of Section 3.1.2.

Exercise 3.4.2 We consider the algorithm of Section 3.1.3. Prove that $E[\theta_n]$ is decreasing if $\theta^* = 0$.

Exercise 3.4.3 We consider the algorithm of Section 3.3.3. Prove that the projection sets $[-1+1/n, 1-1/n]$ are valid.

(Hint: Use the Lyapunov function $V(\theta) = (\theta - \theta^*)^2/(1-\theta^2)$).

Exercise 3.4.4 We consider the algorithm of Section 3.3.2. Check that Πv_θ is Lipschitz on any compact subset of $\mathcal{O}$.

Exercise 3.4.5 We consider the algorithm of Section 3.3.3. Prove that V is a valid Lyapunov function for assumptions (A) and (B).

Exercise 3.4.6 (Proof of Theorem 16 p.35) For any policy π, we denote by $\pi(x_0, x_1, ...x_n)$ the value of the control U_n chosen at time n from the past values of the state. Consider the function

$$J^*(x) = \min_\pi J_\pi(x)$$

Prove that

$$J_\pi(x_0) = c(x, \pi(x_0)) + \rho \sum_y p_{xy}(\pi(x_0)) J_\pi.$$

By taking the $\min_\pi$ in this equation prove that J^* satisfies the Bellman equation. Prove that $J^* = J_{\pi^*}$ where π^* is defined by the Bellman equation.

(Hint: The policy π may be split into two independent parts π_0 and π_1. π_0 decides on U_0 and is a function of X_0; π_1 decides on U_i from $X_0, ...X_i$ and is an infinite family of functions. Take the min w.r.t π_1 first.)

Exercise 3.4.7 (Convergence of the Q-learning) We consider the algorithm (3.19).

- Show that the sequence Q_n is bounded.
- Show that (X_n, U_n) is a Markov chain and give its transition probabilities.
- Show that its invariant measure $p(x, u)$ on $\mathcal{X} \times \mathcal{U}$ is unique and that geometric ergodicity holds (Theorem 32 p.68). Show that $p(x, u) > 0$.
- Give the expression for the mean field $h(Q)$. Show that it rewrites $h(Q) = h^0(Q).p$ ($M.N$ denotes the dot product: $(M.N)_{ij} = M_{ij}N_{ij}$), where $Q \to h^0(Q) + Q$ is a contraction.
- Show that for any non-negative sequence $(x_1, x_2...x_n)$ and $s \geq 1$,

$$\sum_i x_i^{s-1} \leq n^{1/s} \|x\|_\infty^{-1} \sum_i x_i^s$$

and that for q large enough, the following function is a Lyapunov function for assumption (A):

$$V(Q) = \sum_{x,u} \frac{1}{p(x,u)} (Q(x,u) - Q^*(x,u))^{2q}$$

This function may be seen as a suitably smoothed version of $\|Q - Q^*\|_\infty$.

- Prove that Theorems 15 and 35 apply.
- Show that if the function $c(x,u)$ is known and if $p_{xy}(u)$ does not depend on u, the control problem reduces to the estimation of J. Study the algorithm which does this estimation.

Exercise 3.4.8 Check that (M1) and (M2) are satisfied in the example of Section 3.3.6 (notice that, as n tends to $+\infty$, $\|F_\theta^n\| \longrightarrow 0$ uniformly w.r.t. θ, as shown by Proposition 40 p.77)

Chapter 4

Limit–theorems

4.1 Introduction

We will see that under rather mild assumptions the asymptotic variance of the estimation error of the algorithm

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n) = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n, \quad \gamma_n = n^{-\alpha}$$

is, up to a factor γ_n the solution V to

$$U + KV + VK^T = 0, \quad \alpha < 1 \tag{4.1}$$

$$U + KV + VK^T + V = 0, \quad \alpha = 1 \tag{4.2}$$

where

$$K = \nabla E[H(\theta^*, Y)] = \nabla h(\theta^*)$$

and U is essentially the asymptotic variance of (cf Theorem 24 p.54 and the following remark)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \simeq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H(\theta^*, Y_i). \tag{4.3}$$

In particular the condition

$$\begin{array}{ccc} K & \text{stable if} & \alpha < 1 \\ K + I/2 & \text{stable if} & \alpha = 1 \end{array}$$

is necessary for having an estimation error of order $\sqrt{\gamma_n}$. In the stationary case, U rewrites

$$U = \sum_{i=-\infty}^{\infty} R(i) = R(0) + \sum_{i=1}^{\infty} R(i) + R(i)^T$$

where

$$R(i) = \mathrm{Cov}(H(\theta^*, Y_k), H(\theta^*, Y_{i+k})).$$

In the Markov case, this quantity has to be calculated when the Y_n process is in its stationary distribution under Π_{θ^*}.

The factor $\sqrt{\gamma_n}$ let think that $\gamma_n = 1/n$ is the best choice of gain. One important question is then to know which matrix gain Γ permits to attain the best rate in the modified algorithm

$$\theta_n = \theta_{n-1} + \frac{1}{n}\Gamma H(\theta_{n-1}, Y_n).$$

Since equation (4.2) becomes

$$\Gamma U \Gamma^T + \Gamma K V + V K^T \Gamma^T + V = 0$$

a standard calculation leads to the solution

$$\Gamma^* = -K^{-1}$$

and the optimal asymptotic variance is [2]

$$V^* = K^{-1} U K^{-T}$$

in the sense that $x^T V x \geq x^T V^* x$ for any x and any choice of Γ. However, K is not known in advance. Notice that if Γ is chosen too small, the algorithm may converge with a rate smaller than $n^{-1/2}$, because $\Gamma K + I/2$ will fail to be stable. On the other hand, if Γ is too large, V increases. The choice of Γ is thus a difficult point. If the distribution of Y_n does not depend on θ_{n-1}, a first possibility is to estimate this matrix empirically :

$$\hat{K}_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta H(\theta_{i-1}, Y_i) \tag{4.4}$$

and to plug this estimate in the algorithm :

$$\theta_n = \theta_{n-1} - \frac{1}{n}\hat{K}_n^{-1} H(\theta_{n-1}, Y_n).$$

When applied to the problem of regression §2.3.2, this leads to the RLS (Recursive Least Squares) algorithm, by transformation of the algorithm

$$\theta_n = \theta_{n-1} + \frac{1}{n}(y_n - \varphi_n^T \theta_{n-1})\varphi_n$$

into

$$\theta_n = \theta_{n-1} + \frac{1}{n}R_n^{-1}(y_n - \varphi_n^T \theta_{n-1})\varphi_n$$

with

$$R_n = \frac{1}{n}\sum_{i=1}^{n} \varphi_i \varphi_i^T.$$

Notice that θ_n in this case coincides with

$$\theta_n = (nR_n)^{-1}\sum_{i=1}^{n} \varphi_i y_i$$

which is exactly the least squares estimate. The stability of algorithms built on this principle, equation (4.4), is generally a rather complicated matter; moreover, it is not possible to apply this idea in the Robbins-Monro situation. An alternative algorithm, the Ruppert-Polyak averaging method [27, 25], is proposed in §4.4 for solving this problem, without knowledge or estimation of K. It consists in the following

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) \tag{4.5}$$

$$\bar{\theta}_n = \bar{\theta}_{n-1} + \frac{1}{n}(\theta_{n-1} - \bar{\theta}_{n-1}) = \frac{1}{n}(\theta_0 + \theta_1 + ... + \theta_{n-1}). \tag{4.6}$$

The gain γ_n is larger than $1/n$, typically $n^{-2/3}$. We shall show that the convergence of $\bar{\theta}_n$ to θ^* hold with an optimal asymptotic rate. This algorithm may have a long transitory period, and one may feel in practice that the averaging does not forget fast enough the initial values of θ_n, which are not so good. A practical way to avoid this difficulty is to start the averaging (equation (4.6)) with a certain delay, so that it takes place it the 'stochastic' part of algorithm (4.5), i.e. when the variance of the estimate θ_n is larger than its bias.

4.2 Tools for linear approximation

This section is concerned with purely technical results which will be used later. Section 4.2.1 is devoted to bounds for expressions like (4.7). In §4.2, we prove that θ_n (algorithm (4.12)) is well approximated by μ_n (equation (4.14)), in the sense of equation (4.13). The point is that μ_n is much easier to study since it appears as a linear combination of the noise terms. Theorem 20 will be an important step for proving the central limit–theorem.

4.2.1 A useful result on series

This section is devoted to the statement and the proof of a result concerning summation of series, for extensive use in this chapter. Theorem 19 considers two non-negative sequences x_n ε_n are such that

$$x_n \leq (1-\gamma_n)x_{n-1} + \gamma_n \varepsilon_n. \tag{4.7}$$

It is clear that nothing like $x_n = O(\varepsilon_n)$ can be true in general (because ε_n may be accidentally zero); however we shall see that if $\varepsilon_n = O(\gamma_n^p)$ (or $o(\gamma_n^p)$) then x_n has the same property (equation (4.10)).

Lemma 18 *Let u_n, v_n and γ_n be non-negative sequences and such that*

$$\sum_n \gamma_n = \infty, \quad \gamma_n \longrightarrow 0$$

and

$$u_n \leq (1-\gamma_n)\ u_{n-1} + \gamma_n v_n \tag{4.8}$$

then

$$\limsup_n u_n \leq \limsup_n v_n.$$

The same holds if (4.8) is replaced with

$$u_n \leq e^{-\gamma_n}\ u_{n-1} + \gamma_n v_n.$$

Proof: We put $x_+ = x\ 1_{x>0}$. One has

$$(x+y)_+ \leq x_+ + y_+.$$

Choose $v > \limsup_n v_n$, then, for n large enough (so that $v_p < v$ and $\gamma_p \leq 1$ for $p \geq n$),

$$\begin{aligned} u_n - v &\leq (1-\gamma_n)\ (u_{n-1} - v) + \gamma_n(v_n - v) \\ (u_n - v)_+ &\leq (1-\gamma_n)\ (u_{n-1} - v)_+ + \gamma_n(v_n - v)_+ \\ &\leq (1-\gamma_n)\ (u_{n-1} - v)_+ \end{aligned}$$

This implies that $(u_n - v)_+$ tends to zero (because $\sum_n \gamma_n = \infty$), and since v is arbitrarily close to $\limsup v_n$, the result follows.

The second assertion is a consequence of the first one since in this case u_n rewrites

$$\begin{aligned} &u_n \leq (1-\gamma'_n)\ u_{n-1} + \gamma'_n v'_n \\ &\gamma'_n = 1 - e^{-\gamma_n} \\ &v'_n = \frac{\gamma_n}{\gamma'_n} v_n. \end{aligned}$$

∎

Theorem 19 *Let γ_i be a non-negative sequence converging to zero, and b and p be two real numbers, $b > 0$, $p \geq 0$ and consider two non-negative sequences x_n and ε_n such that*

$$x_n = e^{-b\gamma_n} x_{n-1} + \gamma_n^{p+1} \varepsilon_n \tag{4.9}$$

that is

$$x_n = \sum_{i=1}^{n} \gamma_i e^{-b(s_n - s_i)} \gamma_i^p \varepsilon_i + e^{-bs_n} x_0, \quad s_n = \gamma_1 + \gamma_2 ... + \gamma_n.$$

The following is true

- *if* $\frac{1}{\gamma_n} - \frac{1}{\gamma_{n-1}} \longrightarrow 0$ *then for any p*

$$\limsup_n \gamma_n^{-p} x_n \leq b^{-1} \limsup_n \varepsilon_n \tag{4.10}$$

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

$$\limsup_n \gamma_n^{-p} x_n \leq (b-p)^{-1} \limsup_n \varepsilon_n, \quad \textit{if} \quad p < b.$$

Remarks: The conditions on $\gamma_n$are satisfied if $\gamma_n = n^{-\alpha}$, $0 < \alpha \leq 1$.

In the case of inequality in (4.9):

$$x_n \leq e^{-b\gamma_n} x_{n-1} + \gamma_n^{p+1} \varepsilon_n$$

one has

$$x_n \leq \sum_{i=1}^{n} \gamma_i e^{-b(s_n - s_i)} \gamma_i^p \varepsilon_i + e^{-bs_n} x_0$$

and the bounds are still valid.

Proof: Set

$$u_n = \gamma_n^{-p} x_n.$$

Then,

$$\begin{aligned} u_n &= \left(\frac{\gamma_n}{\gamma_{n-1}}\right)^{-p} e^{-b\gamma_n} u_{n-1} + \gamma_n \varepsilon_n \\ &= \exp(p \log(\gamma_{n-1}/\gamma_n) - b\gamma_n)\ u_{n-1} + \gamma_n \varepsilon_n \\ &= (1 - b_n \gamma_n)\ u_{n-1} + b_n \gamma_n \frac{\varepsilon_n}{b_n} \end{aligned} \tag{4.11}$$

where

$$b_n = \gamma_n^{-1}(1 - \exp(p \log(\gamma_{n-1}/\gamma_n) - b\gamma_n)).$$

In the first case γ_n/γ_{n-1} tends to 1 and

$$\log(\gamma_{n-1}/\gamma_n) = (\gamma_{n-1}/\gamma_n - 1)(1 + o(1)) = \gamma_{n-1}(1/\gamma_n - 1/\gamma_{n-1})(1 + o(1)) = o(\gamma_n)$$

thus b_n converges to b and we conclude with Lemma 18. In the case $n\gamma_n \longrightarrow 1$ and $b > p$, b_n converges to $b - p$ and we conclude in the same way. ∎

4.2.2 Approximation with a linear algorithm

In this section, we will work with the following assumption

(C) *The algorithm (4.16) is A-stable and θ_n converges to some limit θ^*. h is C^1 in some neighborhood of θ^* with first derivatives Lipschitz, and the matrix $K = \nabla h(\theta^*)$, has all its eigenvalues with negative real part.*

And we shall give a comparison theorem between the solution of

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n \tag{4.12}$$

and the process μ_n defined by

$$\mu_n = \mu_{n-1} + \gamma_n K \mu_{n-1} + \gamma_n \eta_n.$$

For technical reasons, the process μ_n will be actually

$$\mu_n = e^{\gamma_n K} \mu_{n-1} + \gamma_n \eta_n.$$

Theorem 20 will be an important step for proving the central limit–theorem since the asymptotic normality of θ_n, will be a consequence of the two facts:

- $A_n^{-1}\mu_n$ tends in distribution to the normal variable, where $A_n = E[\mu_n \mu_n^T]^{1/2}$
- the distance between $\theta_n - \theta^*$ and μ_n is $o(\|A_n\|)$ in mean (the results of §4.2.1 will be useful).

In this section, we will consider the assumption

Theorem 20 *Let assumption (C) hold and the sequence $\gamma_n > 0$ be decreasing with $\sum \gamma_n = \infty$; then the algorithm (4.12) satisfies*

$$\theta_n - \theta^* = \mu_n + O\left(e^{-\kappa s_n} + \sum_1^n \gamma_i e^{-\kappa(s_n - s_i)}(|\mu_{i-1}|^2 + \gamma_i^2)\right) \tag{4.13}$$

where

$$\mu_n = \sum_1^n e^{(s_n - s_i)K}\gamma_i \eta_i, \tag{4.14}$$

κ *is any real number smaller than the smallest real part of the eigenvalues of* $-K$,

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

and μ_n tends to zero.

In this theorem, the constant hidden in the $O(\ )$ depends on everything (κ included) except on n.

Proof: We need first to prove that μ_n tends to zero. We have

$$\begin{aligned}
\mu_n &= e^{\gamma_n K}\mu_{n-1} + \gamma_n \eta_n, \qquad \mu_0 = 0\\
&= e^{\gamma_n K}\mu_{n-1} + \theta_n - \theta_{n-1} - \gamma_n h(\theta_{n-1}).
\end{aligned}$$

If we set $z_n = \tilde\theta_n - \mu_n$, $\tilde\theta_n = \theta_n - \theta^*$, this equation rewrites

$$\begin{aligned}
z_n &= \theta_{n-1} - \theta^* - e^{\gamma_n K}\mu_{n-1} + \gamma_n h(\theta_{n-1})\\
&= e^{\gamma_n K} z_{n-1} + (I - e^{\gamma_n K})\tilde\theta_{n-1} + \gamma_n h(\theta_{n-1})\\
&= e^{\gamma_n K} z_{n-1} + \gamma_n O(\gamma_n|\tilde\theta_{n-1}| + |\theta_{n-1} - \theta^*|^2)\\
&= e^{\gamma_n K} z_{n-1} + \gamma_n O(\gamma_n^2 + |\tilde\theta_{n-1}|^2)\\
&= \sum_1^n e^{(s_n - s_i)K}\gamma_i O(\gamma_i^2 + |\tilde\theta_{i-1}|^2) + e^{s_n K} z_0
\end{aligned}$$

which leads to the bound (cf Proposition 39 p.76)

$$|z_n| \le C\sum_1^n e^{-\nu(s_n - s_i)}\gamma_i(\gamma_i^2 + |\tilde\theta_{i-1}|^2) + Ce^{-\kappa s_n}|z_0| \tag{4.15}$$

for some $\nu > \kappa$. Theorem 19 p.48 (with $p = 0$) implies that the left hand side tends to zero, and so does μ_n. $|z_n|$ is thus smaller than

$$x_n = C\sum_1^n e^{-\nu(s_n - s_i)}\gamma_i(\tau_i + |z_{i-1}|^2) + Ce^{-\nu s_n}|z_0|, \quad \tau_n = \gamma_n^2 + |\mu_{n-1}|^2.$$

x_n tends to zero (Theorem 19) and

$$\begin{aligned}
x_n &= e^{-\gamma_n \nu}x_{n-1} + C\gamma_n|z_{n-1}|^2 + C\gamma_n\tau_n\\
&\le e^{-\gamma_n \nu}x_{n-1} + C\gamma_n x_{n-1}^2 + C\gamma_n\tau_n\\
&\le e^{-\gamma_n \kappa}x_{n-1} + C\gamma_n\tau_n
\end{aligned}$$

for n large enough ($\nu > \kappa$ and x_n tends to zero). Hence for some constant C'

$$x_n \le C'\sum_1^n e^{-\kappa(s_n - s_i)}\gamma_i\tau_i + C'e^{-\kappa s_n}$$

what implies (4.13) since $|z_n| \le x_n$. ∎

4.3 Asymptotic normality

In this section, we will consider algorithms

$$\theta_n \quad = \quad \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n \eta_n \tag{4.16}$$

with the assumptions (C) and

(D) *The gain sequence in decreasing and satisfies either*

Case 1: $$\frac{1}{\gamma_n} - \frac{1}{\gamma_{n-1}} \longrightarrow 0$$

or

Case 2: $$\gamma_n n \longrightarrow 1$$

In Case 2 the strict stability of $K + I/2$ is also required (cf assumption (C)).

(E) *There exist a random variable $X > 0$ such that*

$$\left\| X \sum_{k=n}^{p} \gamma_k \eta_k \right\|_2 < \sqrt{\gamma_n}, \qquad \text{if} \qquad \sum_{i=n}^{p} \gamma_i \leq 1. \tag{4.17}$$

Remarks: The sequences $\gamma_n = n^{-\alpha}$, $0 < \alpha \leq 1$, satisfy the condition of assumption (D).

Under (D), $\sum_n^p \gamma_i^2 \leq \gamma_n \sum_n^p \gamma_i$, and assumption (E) is satisfied if η_n is an L_2-bounded martingale increment sequence.

Assumption (D) implies that $\gamma_n > C/n$ for some $C > 0$.

Because of the introduction of the density X, assumption (E) is tailored for having the possibility of handling L_2 properties as well as a.s. relations, (cf point 3 of Theorem 21). In particular, the extra remainder due to the projection made in the Chen procedure (cf Theorem 10 p.16) has no influence since is non-zero only finitely often (notice that its L_2-norm would be impossible to estimate).

Theorem 21 *If γ_n is decreasing, any of the following conditions implies (E)*

1. $\eta_n = \eta_n' + \eta_n''$ *where* η_n' *and* η_n'' *satisfy (E).*
2. $\|X\eta_n\|_2 < \sqrt{\gamma_n}, \quad n > 0.$
3. $\sup_n \eta_n/\sqrt{\gamma_n} < \infty$ *a.s.*
4. $\|X \sum_{k=n}^{p} \gamma_k \eta_k\|_2 < 1/\sqrt{\gamma_n}$, *if n and p are such that* $\sum_{i=n}^{p} \gamma_i \leq 1$.

Proof:

1. Take $X = \min(X', X'')$.

2. Obvious.

3. Take the inverse of this supremum for X in point 2 of this theorem.

4. Setting $S_k = \sum_{i=n}^{k} \eta_i$:

$$\sum_{k=n}^{p} \gamma_k \eta_k = \sum_{k=n}^{p} \gamma_k (S_k - S_{k-1}) = \sum_{k=n}^{p-1} (\gamma_k - \gamma_{k+1}) S_k + \gamma_p S_p$$

hence

$$\|X \sum_{k=n}^{p} \gamma_k \eta_k\|_2 \le \sum_{k=n}^{p-1} (\gamma_k - \gamma_{k+1}) \frac{1}{\sqrt{\gamma_n}} + \gamma_p \frac{1}{\sqrt{\gamma_n}} = O(\sqrt{\gamma_n}).$$

∎

We shall need the following consequence of (E):

Lemma 22 *Assumptions (D) and (E) imply the existence of two processes e_n and r_n such that*

$$\eta_n = e_n + r_n$$
$$\left\| X \sum_{k=n}^{\infty} \gamma_k e_k \right\|_2 < \sqrt{\gamma_n} \tag{4.18}$$
$$\|X r_n\|_2 < \sqrt{\gamma_n}. \tag{4.19}$$

Notice that the converse is true: If the conditions of the lemma are satisfied, assumption (E) holds.

Proof: We consider a sequence $p_1, p_2, ...$ such that

$$p_1 = 1, \quad \sum_{k=p_i}^{p_{i+1}-1} \gamma_k \le 1 < \sum_{k=p_i}^{p_{i+1}} \gamma_k$$

This relation with $\gamma_k > C/k$ (this comes from assumption (D)) implies that $p_{i+1}/p_i \le C'$. Set

$$\begin{aligned} r_n &= \left(\sum_{k=p_i}^{p_{i+1}-1} \gamma_k \right)^{-1} \sum_{k=p_i}^{p_{i+1}-1} \gamma_k \eta_k, \quad p_i \le n < p_{i+1} \\ e_n &= \eta_n - r_n. \end{aligned}$$

Equation (4.19) is an immediate consequence of (E) and of $p_{i+1}/p_i \le C'$. Notice that for any k

$$\sum_{n=p_i}^{p_{i+1}-1} \gamma_n r_n = \sum_{k=p_i}^{p_{i+1}-1} \gamma_k \eta_k.$$

If we denote by n_+ the smallest p_i such that $p_i \geq n$, and p_- the largest p_i such that $p_i \leq p$ we have

$$\begin{aligned}
\sum_{k=n}^{p} \gamma_k e_k &= \sum_{k=n}^{p} \gamma_k \eta_k - \sum_{k=n}^{p} \gamma_k r_k \\
&= \sum_{k=n}^{n_+-1} \gamma_k \eta_k + \sum_{k=p_-}^{p} \gamma_k \eta_k - \sum_{k=n}^{n_+-1} \gamma_k r_k - \sum_{k=p_-}^{p} \gamma_k r_k \\
\left\| X \sum_{k=n}^{p} \gamma_k e_k \right\|_2 &\leq C\sqrt{\gamma_n} + C\sqrt{\gamma_n} + C \sum_{k=n}^{n_+-1} \gamma_k \sqrt{\gamma_n} + C \sum_{k=p_-}^{p} \gamma_k \sqrt{\gamma_n} \\
&\leq 4C\sqrt{\gamma_n}
\end{aligned}$$

∎

4.3.1 A general asymptotic bound for the estimation error

Theorem 23 *Let assumptions (C), (D) and (E) be satisfied. Then there exist a random variable Z and an L_1-bounded sequence Y_n (also L_2-bounded in Case 1 of assumption (D)) such that*

$$|\theta_n - \theta^*| \leq \sqrt{\gamma_n} Z\ Y_n. \tag{4.20}$$

Furthermore we have the limit in probability

$$\frac{1}{\sqrt{\gamma_n}} \left(\theta_n - \theta^* - \sum_{i=1}^{n} e^{(s_n - s_i)K} \gamma_i \eta_i \right) \xrightarrow{P} 0. \tag{4.21}$$

Proof: We will use equation (4.13) where

$$\mu_n = \sum_{i=1}^{n} e^{(s_n - s_i)K} \gamma_i \eta_i.$$

Choose any $\kappa > 0$ larger than the smallest real part of the eigenvalues of $-K$ (by assumption (D) one can choose $\kappa > 1/2$ if $n\gamma_n \to 1$). Rewriting η_n as in Lemma 22 we have

$$\|X \sum_{i=1}^{n} e^{(s_n - s_i)K} \gamma_i r_i\|_2 \leq C \sum_{i=1}^{n} e^{-(s_n - s_i)\kappa} \gamma_i^{3/2} = O(\sqrt{\gamma_n})$$

(Theorem 19 p.48, $b = \kappa, p = 1/2$) and setting $S_j = \sum_{i=j}^{\infty} \gamma_i e_i$

$$\begin{aligned}
\sum_{i=1}^{n} e^{(s_n - s_i)K} \gamma_i e_i &= \sum_{i=1}^{n} e^{(s_n - s_i)K} (S_i - S_{i+1}) \\
&= \sum_{i=2}^{n} e^{(s_n - s_i)K} (1 - e^{\gamma_i K}) S_i + e^{(s_n - s_1)K} S_1 - S_{n+1} \\
\|X \sum_{i=1}^{n} e^{(s_n - s_i)K} \gamma_i e_i\|_2 &\leq C \sum_{i=1}^{n} e^{-(s_n - s_i)\kappa} \gamma_i \sqrt{\gamma_i} + C\sqrt{\gamma_n} \\
&\leq C' \sqrt{\gamma_n}.
\end{aligned}$$

Finally $\gamma_n^{-1/2} X \mu_n$ is bounded in L_2, that is

$$|\mu_n| \leq \sqrt{\gamma_n} Z Y_n \tag{4.22}$$

where Y_n is bounded in L_2 and $Z = 1/X$ is fixed. We can now estimate the last term of equation (4.13). Applying Theorem 19 ($b = \kappa$, $p = 1/2$),

$$\begin{aligned}\|(Z^2+1)^{-1}\sum_1^n \gamma_i e^{-\kappa(s_n-s_i)}(\gamma_i^2+|\mu_i|^2)\|_1 &\le C\sum_1^n \gamma_i e^{-\kappa(s_n-s_i)}\gamma_i \\ &= C'\begin{cases}\gamma_n & \text{(Case 1)}\\ \gamma_n^{1/2+\varepsilon} & \text{(Case 2)}\end{cases}\end{aligned}$$

(in Case 2 of assumption (D), the condition $b > p$ in Theorem 19 forces us to bound γ_i by $\gamma_i^{1/2+\varepsilon}$ with $1/2+\varepsilon < \kappa$). Finally

$$\sum_1^n \gamma_i e^{-\kappa(s_n-s_i)}(\gamma_i^2+|\mu_i|^2) \le Z'Y_n'\begin{cases}\gamma_n & \text{(Case 1)}\\ \gamma_n^{1/2+\varepsilon} & \text{(Case 2)}\end{cases} \tag{4.23}$$

where Y_n' is bounded in L_1 and $Z' = C'(Z^2+1)$. This and equation (4.13) imply equation (4.21) (the term $e^{-\kappa s_n}$ of (4.13) is immediately treated with Theorem 19).

It remains now to prove that (4.20) holds true. Combining equations (4.13, 4.22, 4.23), we obtain in Case 2

$$|\theta_n-\theta^*| \le \sqrt{\gamma_n}ZY_n + \gamma_n^{1/2+\varepsilon}Z''Y_n' \le C''\sqrt{\gamma_n}(Z+Z'')(Y_n+Y_n')$$

where the variable Z'' comes from Z' and the $O(\ )$ of (4.13). This is equation (4.20). In Case 1 we have

$$|\theta_n-\theta^*| \le \sqrt{\gamma_n}ZY_n + \gamma_n Z''Y_n'$$

and using that for $a, b, x \ge 0$ the bound $x \le a+b$ implies that $x^2 \le a^2+2xb$ we get

$$|\theta_n-\theta^*|^2 \le \gamma_n Z^2Y_n^2 + 2\gamma_n(|\theta_n-\theta^*|Z'')Y_n'.$$

This is equation (4.20). ∎

4.3.2 Asymptotic normality

In order to obtain limits in distribution, we need some precise assumptions on η_n. Motivated by Section 3.3.1 and Section 3.2 (Theorem 14 p.28), we shall consider a noise of the form

$$\eta_n = \xi_n + \nu_n - \nu_{n-1} + r_n$$

where ξ_n is a martingale increment sequence and ν_n and r_n are additional variables.

Theorem 24 *Let assumptions (C) and (D) be satisfied, and if $n\gamma_n \to 1$, we require that $\gamma_n = 1/n$. Furthermore assume that for some matrix U and $\varepsilon > 0$ and some random variables $X, X', X'' > 0$*

$$\begin{aligned}&E[\xi_i|\xi_1\dots,\xi_{i-1}] = 0, \quad a.s\\ &\sup_i \|\xi_i\|_{2+\varepsilon} < \infty\\ &\gamma_n^{-1/2}\|Xr_n\|_1 \longrightarrow 0\\ &\gamma_n^{1/2}\|X'\nu_n\|_1 \longrightarrow 0\\ &\gamma_n\|X''\sum_{i=1}^n(\xi_i\xi_i^T-U)\|_1 \longrightarrow 0.\end{aligned}$$

Then

$$\frac{\theta_n - \theta^*}{\sqrt{\gamma_n}} \xrightarrow{d} \mathcal{N}(0, V)$$

where V is the solution to the following Lyapunov equation depending on the Case of assumption (D)

$$\begin{aligned} Case\ 1 &: \quad U + KV + VK^T = 0, \qquad K = \nabla h(\theta^*) \\ Case\ 2 &: \quad U + (K + I/2)V + V(K + I/2)^T = 0. \end{aligned}$$

Remarks: The assumption on r_n is satisfied if

$$\gamma_n^{-1/2} r_n \to 0, \quad a.s$$

(set $X^{-1} = \sup_n \gamma_n^{-1/2}|r_n|$ and use the Lebesgue theorem). An analogous statement holds for the two other L_1 bounds of the assumption.

Under these assumptions, Theorem 30 p.63 implies that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_i \xrightarrow{d} \mathcal{N}(0, U).$$

Proof: Note that assumption (E) is satisfied (cf the remark following the statement of assumption (E) and Theorem 21). By an application of Proposition 39 p.76 and Theorem 19, the sum

$$\gamma_n^{-1/2} \sum_0^n e^{(s_n - s_i)K} \gamma_i r_i$$

converges a.s. to zero and

$$\begin{aligned} \sum_{i=1}^n e^{(s_n - s_i)K} \gamma_i (\nu_i - \nu_{i-1}) &= \sum_{i=1}^{n-1} e^{(s_n - s_i)K} (\gamma_i - \gamma_{i+1} e^{\gamma_{i+1} K}) \nu_i + e^{s_n K} \gamma_1 \nu_0 + \gamma_n \nu_n \\ \|X' \sum_{i=1}^n e^{(s_n - s_i)K} \gamma_i (\nu_i - \nu_{i-1})\|_1 &\leq C \sum_{i=1}^n e^{-(s_n - s_i)\kappa} o(\gamma_i^{3/2}) \\ &= o(\sqrt{\gamma_n}). \end{aligned}$$

Thus Theorem 23 implies that the limit in distribution of $\gamma_n^{-1/2}(\theta_n - \theta^*)$ coincides with the limit of

$$\gamma_n^{-1/2} \sum_0^n e^{(s_n - s_i)K} \gamma_i \xi_i. \tag{4.24}$$

This kind of sum of martingale increments is a martingale array, cf [10], and we shall use Theorem 30 p.63 with

$$X_{ni} = \gamma_n^{-1/2} e^{(s_n - s_i)K} \gamma_i \xi_i.$$

Notice that for δ small enough

$$\sum_{i=1}^n E|[X_{ni}|^{2+\delta}] \leq C \gamma_n^{-(1+\delta/2)} \sum_i e^{-(2+\delta)(s_n - s_i)\kappa} \gamma_i \gamma_i^{\delta+1}$$

where $\kappa > 0$ is any real number larger than the smallest real part of the eigenvalues of $-K$ (by assumption one can choose $\kappa > 1/2$ if $\gamma_n = 1/n$). For δ small enough we can apply Theorem 19 (with $b = (2+\delta)\kappa$, $p = \delta + 1$) and

$$\sum_{i=1}^{n} E|[X_{ni}|^{2+\delta}] \;\leq\; C\gamma_n^{-(1+\delta/2)}\gamma_n^{\delta+1} = C\gamma_n^{\delta/2}$$

which tends to zero. Since this inequality is valid also for $\delta = 0$, equations (A.1) and (A.2) are satisfied. The rest of the proof consists now in the computation of the asymptotic variance V of equation (A.3). Set

$$V_n \;=\; \sum_{i=1}^{n} X_{ni}X_{ni}^T, \quad U_i = \xi_i\xi_i^T, \quad S_n = \sum_{i=1}^{n} U_i - U$$

then

$$V_n \;=\; \gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{(s_n-s_i)K}U_i e^{(s_n-s_i)K^T} = \bar{V}_n + \tilde{V}_n$$

with

$$\begin{aligned}\bar{V}_n &= \gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{(s_n-s_i)K}U e^{(s_n-s_i)K^T}\\ \tilde{V}_n &= \gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{(s_n-s_i)K}(U_i - U)e^{(s_n-s_i)K^T}.\end{aligned}$$

Let us show first that $\tilde{V}_n$ tends in probability to zero :

$$\begin{aligned}\tilde{V}_n &= \gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{(s_n-s_i)K}(S_i - S_{i-1})e^{(s_n-s_i)K^T}\\ &= \gamma_n^{-1}\sum_{i=1}^{n-1}\gamma_i^2 e^{(s_n-s_i)K}S_i e^{(s_n-s_i)K^T} - \gamma_{i+1}^2 e^{(s_n-s_{i+1})K}S_i e^{(s_n-s_{i+1})K^T} + \gamma_n^{-1}\gamma_n^2 S_n.\end{aligned}$$

We use now the matrix inequality

$$|ABA - A'BA'| = |(A-A')BA - A'B(A'-A)| \leq |A-A'|\;|B|\;(|A| + |A'|)$$

with $A = A_i = \gamma_i e^{(s_n-s_i)K}$ and $A' = A_{i+1}$; then

$$\begin{aligned}|A - A'| &= |(\gamma_i - \gamma_{i+1})e^{(s_n-s_i)K} + \gamma_{i+1}e^{(s_n-s_i)K}(1 - e^{-\gamma_{i+1}K})|\\ &\leq C\gamma_i^2 e^{-(s_n-s_i)\kappa}\end{aligned}$$

and

$$|\tilde{V}_n| \;\leq\; C'\gamma_n^{-1}\sum_{i=1}^{n-1}\gamma_i e^{-2(s_n-s_i)\kappa}\gamma_i^2|S_i| + \gamma_n|S_n|$$

where C depends only on the matrix K and on the sequence γ_n. Setting $\varepsilon_i = \gamma_i E[X|S_i|]$, we obtain

$$E[X''|\tilde{V}_n|] \;\leq\; C'\gamma_n^{-1}\sum_{i=1}^{n-1}\gamma_i e^{-2(s_n-s_i)\kappa}\gamma_i\varepsilon_i + \varepsilon_n.$$

Theorem 19 implies that the sum is $o(\gamma_n)$, and $\tilde{V}_n$ tends in probability to zero. We compute now $\bar{V}_n$. Notice first that equation (4.25) implies that V_n is bounded :

$$|\bar{V}_n| \leq C\gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{-2\kappa(s_n-s_i)} \leq C'$$

where we used again Theorem 19. On the other hand

$$\begin{aligned}
\bar{V}_n &= \gamma_n^{-1}\sum_{i=1}^{n}\gamma_i^2 e^{(s_n-s_i)K}Ue^{(s_n-s_i)K^T} \qquad (4.25)\\
&= \gamma_n U + \frac{\gamma_{n-1}}{\gamma_n}e^{\gamma_n K}\bar{V}_{n-1}e^{\gamma_n K^T}\\
&= \gamma_n U + \frac{\gamma_{n-1}}{\gamma_n}(1+\gamma_n K + O(\gamma_n^2))\bar{V}_{n-1}(1+\gamma_n K^T + O(\gamma_n^2))\\
&= \gamma_n U + \bar{V}_{n-1} + \frac{\gamma_{n-1}-\gamma_n}{\gamma_n}\bar{V}_{n-1} + \gamma_{n-1}(K\bar{V}_{n-1} + \bar{V}_{n-1}K^T) + O(\gamma_n\gamma_{n-1})\\
&= \bar{V}_{n-1} + \gamma_n(U + K\bar{V}_{n-1} + \bar{V}_{n-1}K^T) + \frac{\gamma_{n-1}-\gamma_n}{\gamma_n}\bar{V}_{n-1} + O(\gamma_n\gamma_{n-1} + |\gamma_{n-1}-\gamma_n|).
\end{aligned}$$

If $\gamma_n = 1/n$ we obtain

$$\bar{V}_n = \bar{V}_{n-1} + \frac{1}{n}(U + K\bar{V}_{n-1} + \bar{V}_{n-1}K^T + \bar{V}_{n-1}) + O(\frac{1}{n^2}) \qquad (4.26)$$

else (i.e. Case 1), (D) implies that

$$\bar{V}_n = \bar{V}_{n-1} + \gamma_n(U + K\bar{V}_{n-1} + \bar{V}_{n-1}K^T) + o(\gamma_n) \qquad (4.27)$$

Denoting by small letters the vectorized matrices, this algorithm rewrites

$$v_n = v_{n-1} + \gamma_n(u + Av_{n-1}) + o(\gamma_n).$$

Notice that the matrix A has all its eigenvalues with negative real part (these eigenvalues are $\lambda_i + \lambda_j$ where λ_i and λ_j are eigenvalues of K in Case 1, and $\lambda_i + \lambda_j + 1$ in Case 2). We can apply Theorem 2 p.8 for the convergence to the solution to $u + Av = 0$, i.e. $U + KV + VK^T = 0$ ($U + KV + VK^T + V = 0$ in Case 2).

This convergence of $\bar{V}_n$ implies the convergence of V_n, and by Theorem 30 p.63, the process of equation (4.24) converges to $\mathcal{N}(0, V)$ and so does $(\theta_n - \theta^*)/\sqrt{\gamma_n}$. ∎

4.3.3 Application to the algorithms with Markovian dynamics

We consider the algorithm (3.8) under assumptions (MS), (C) and (D), and we assume additionnally that $\phi_\mathcal{C}$ (cf (MS)) can be chosen as

$$\phi_\mathcal{C}(x) = K_\mathcal{C}\sqrt{x}$$

where $K_\mathcal{C}$ is a $K_\mathcal{C}$-dependent constant.

It is elementary to prove from equation (3.11) that assumption (E) is satisfied (this is left as an exercise; use Theorem 21 and the remark following the statement of assumption (E)). We can apply Theorem 23 and conclude to the existence of Z and Y_n such that

$$|\theta_n - \theta^*| \leq \sqrt{\gamma_n}ZY_n.$$

It is also easy to prove that the assumptions of Theorem 24, except the last one, are satisfied. The only problem now is thus the fifth one :

$$\lim \gamma_n \|X \sum_{i=1}^{n} (\xi_i \xi_i^T - U)\|_1 = 0$$

where the process ξ_n is (cf equation (3.11))

$$\xi_{n+1} = v_{\theta_n}(X_{n+1}) - \Pi_{\theta_n} v_{\theta_n}(X_n).$$

Set $v_*(x) = v_{\theta^*}(x)$, $\Pi_* = \Pi_{\theta^*}$ and

$$\xi^*_{n+1} = v_*(X_n) - \Pi_* v_*(X_{n-1})$$

Then, under assumptions (MS)

$$|\xi^*_{n+1} - \xi_{n+1}| \leq C|\theta_n - \theta^*| \leq \sqrt{\gamma_n} Z Y_n.$$

This implies that

$$\begin{aligned}
\gamma_n \|Z^{-1} \sum_{i=1}^{n} \xi_i^* \xi_i^{*T} - \xi_i \xi_i^T\|_1 &= \gamma_n \|Z^{-1} \sum_{i=1}^{n} \xi_i^* (\xi_i^* - \xi_i)^T - (\xi_i - \xi_i^*) \xi_i^T\|_1 \\
&\leq C\gamma_n \sum_{i=0}^{n} \|Z^{-1}(\theta_i - \theta^*)\|_1 \\
&\leq C\gamma_n \sum_{i=0}^{n} \sqrt{\gamma_i} \|Y_i\|_1 \\
&= O(n\gamma_n^{3/2}).
\end{aligned}$$

We shall use the Poisson equation again : Assume that we can find a bounded function $w(x)$ and a matrix U such that

$$w - \Pi_* w = v_* v_*^T - \Pi_* v_* \Pi_* v_*^T - U$$

and set

$$l(y,x) = w(y) + (\Pi_* v \Pi_* v_*^T)(x) - v_*(y) \Pi_* v_*^T(x) - \Pi_* v_*(x) v_*(y)^T.$$

Then we have, because of (MS)

$$E[l(X_{n+1}, X_n)|\mathcal{F}_n] = \int l(y, X_n) \Pi_{\theta_n}(X_n, dy) = \int l(y, X_n) \Pi_*(X_n, dy) + O(|\theta_n - \theta^*|)$$

and the first term rewrites

$$\begin{aligned}
\int l(y, X_n) \Pi_*(X_n, dy) &= \Pi_* w(X_n) - (\Pi_* v_* \Pi_* v_*^T)(X_n) \\
&= (w - v_* v_*^T)(X_n) + U \\
&= l(X_n, X_{n-1}) - (\Pi_* v_* \Pi_* v_*^T)(X_{n-1}) \\
&= l(X_n, X_{n-1}) - \xi_n \xi_n^T + U.
\end{aligned}$$

Hence

$$\begin{aligned}
\sum_{i=1}^n \xi_i\xi_i^T - U &= \sum_{i=1}^n l(X_i, X_{i-1}) - E[l(X_{i+1}, X_i)|\mathcal{F}_i] + \sum_{i=1}^n O(|\theta_i - \theta^*|) \\
\|\min(1, 1/Z)\sum_{i=1}^n \xi_i^*\xi_i^{*T} - U\|_1 &\leq \|\sum_{i=1}^n l(X_i, X_{i-1}) - E[l(X_{i+1}, X_i)|\mathcal{F}_i]\|_2 + C\sum_{i=1}^n \sqrt{\gamma_i}\|Y_i\|_1 \\
&= O(\sqrt{n} + n\sqrt{\gamma_n})
\end{aligned}$$

finally, we have

$$\gamma_n\|\min(1, 1/Z)\sum_{i=1}^n \xi_i\xi_i^T - U\|_1 = O(\sqrt{n}\gamma_n + n\gamma_n^{3/2})$$

which is $o(1)$ if $\gamma_n = n^{-\alpha}$, $\alpha > 2/3$. We have proved the

Theorem 25 *Consider the algorithm (3.8) under the assumptions (MS), (C) and (D)with* $\gamma_n = n^{-\alpha}$, $\alpha > 2/3$. *We assume in addition that there exist a bounded function* w *and a matrix* U *such that*

$$w - \Pi_{\theta^*} w = v_{\theta^*} v_{\theta^*}^T - \Pi_{\theta^*} v_{\theta^*} \Pi_{\theta^*} v_{\theta^*}^T - U.$$

Then

$$\frac{\theta_n - \theta^*}{\sqrt{\gamma_n}} \xrightarrow{d} \mathcal{N}(0, V)$$

where V *is the solution to*

$$\begin{aligned}
2/3 < \alpha < 1 \quad &: \quad U + KV + VK^T = 0, \qquad K = \nabla h(\theta^*), \\
\alpha = 1 \quad &: \quad U + (K + I/2)V + V(K + I/2)^T = 0.
\end{aligned}$$

The matrix U is of course

$$U = E_{\theta^*}[v_{\theta^*}(X)v_{\theta^*}(X)^T - \Pi_{\theta^*} v_{\theta^*}(X)\Pi_{\theta^*} v_{\theta^*}(X)^T]$$

where E_{θ^*} is the invariant measure of the chain under Π_{θ^*}.

4.4 Averaging

We will consider here the so-called Polyak-Ruppert algorithm [27, 25]:

$$\begin{aligned}
\theta_n &= \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n(\xi_n + \nu_n - \nu_{n-1} + r_n) \\
\bar{\theta}_n &= \bar{\theta}_{n-1} + \frac{1}{n}(\theta_{n-1} - \bar{\theta}_{n-1}).
\end{aligned}$$

Theorem 26 *Let assumptions (C) and (D) be satisfied, and* $\gamma_n = n^{-\alpha}$, $1/2 < \alpha < 1$. *We assume in addition that* ξ_n *is a martingale increment sequence with* $(|\xi_n|^2)_{n>0}$ *uniformly integrable, and the bound*

$$\sup_n \sqrt{\gamma_n}\|X\nu_n\|_2 + \frac{\|Xr_n\|_2}{\sqrt{\gamma_n}} < \infty$$

for some r.v. $X > 0$*. Furthermore, the following convergences in probability hold*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_n \quad \xrightarrow{P} \quad 0,$$

$$\frac{1}{n} \sum_{1}^{n} \xi_i \xi_i^T \quad \xrightarrow{P} \quad U$$

for some matrix U*. Then*

$$\sqrt{n}\ (\bar{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, V) \qquad \textit{in distribution} \tag{4.28}$$

where

$$V = K^{-1} U K^{-T}.$$

Remarks: Inspection of the proof reveals that the choice $\alpha = 2/3$ gives a good first order balance between different remainders which have an order of magnitude of γ_n or $(n\sqrt{\gamma_n})^{-1}$. In this theorem, the i-th row of K contains the gradient of the i-th component of h.

If the projection is used, the projected algorithm rewrites

$$\theta_n \quad = \quad \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n(\xi_n + \nu_n - \nu_{n-1} + r_n + r'_n)$$

where r'_n is zero but a finite number of times. Theorem 21 implies that the additional remainder r'_n has no influence on the validity of the assumptions.

The proof of Theorem 26 is based on Lemma 27 below.

Proof of theorem 26: Without loss of generality, we will assume in this proof that $\theta^* = 0$. Let us first prove that

$$\sqrt{n}\ \bar{\theta}_n + \frac{K^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \xrightarrow{P} 0. \tag{4.29}$$

The L_2 bounds of the assumptions, Theorem 21 p.51 and the remark following assumption (E), imply that assumption (E) is satisfied and Theorem 23 p.53 applies.

Using Lemma 27 with equation

$$\begin{aligned} \theta_n &= \theta_{n-1} + \gamma_n K \theta_{n-1} + \gamma_n O(|\theta_{n-1}|^2) + \gamma_n \eta_n, \quad \eta_n = \xi_n + r_n + \nu_n - \nu_{n-1} \\ \bar{\theta}_n &= \bar{\theta}_{n-1} + \frac{1}{n}(\theta_{n-1} - \bar{\theta}_{n-1}) \end{aligned} \tag{4.30}$$

we obtain, thanks to Theorem 23 p.53

$$\begin{aligned} \bar{\theta}_n + \frac{K^{-1}}{n} \sum_{i=1}^{n} \eta_i &= O\left(\frac{1}{n} \sum_{i=1}^{n} |\theta_{i-1}|^2\right) + O\left(\frac{|\theta_n|}{n\gamma_n} + \frac{|\theta_0|}{n\gamma_1} + \frac{1}{n} \sum_{i=2}^{n} |\theta_{i-1}|(\gamma_i^{-1} - \gamma_{i-1}^{-1})\right) \\ |\sqrt{n}\ \bar{\theta}_n + \frac{K^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \eta_i| &\le \frac{Z'}{\sqrt{n}} \left(\sum_{i=1}^{n} \gamma_i Y_i^2 + \frac{Y_n + |\theta_0|}{\sqrt{\gamma_n}} + \sum_{i=2}^{n} Y_i \sqrt{\gamma_i}(\gamma_i^{-1} - \gamma_{i-1}^{-1})\right) \end{aligned} \tag{4.31}$$

(Z' is a random variable taking into account Z from Theorem 23 p.53 and the random variable implicitly contained in the $O(\ )$ of (4.30)). Notice that, for C large enough

- $n^{-1/2} \sum_{i=1}^{n} \gamma_i E\left[Y_i^2\right] \le C\gamma_n \sqrt{n} = o(1)$
- $(n\gamma_n)^{-1/2} E[Y_n + |\theta_0|] \le C/\sqrt{n\gamma_n} = o(1)$

- and

$$
\begin{aligned}
C^{-1}E\sum_{i=2}^{n} Y_i\sqrt{\gamma_i}(\gamma_i^{-1}-\gamma_{i-1}^{-1}) &\le \sum_{i=2}^{n}\sqrt{\gamma_i}(\gamma_i^{-1}-\gamma_{i-1}^{-1}) \\
&= \sum_{i=2}^{n-1}\gamma_i^{-1}(\sqrt{\gamma_i}-\sqrt{\gamma_{i+1}})+\sqrt{\gamma_n}\gamma_n^{-1}-\sqrt{\gamma_2}\gamma_1^{-1} \\
&\le \sum_{i=2}^{n-1}\frac{\sqrt{\gamma_i}-\sqrt{\gamma_{i+1}}}{\sqrt{\gamma_i\gamma_{i+1}}}+1/\sqrt{\gamma_n} \\
&= \frac{2}{\sqrt{\gamma_n}}.
\end{aligned}
$$

These relations prove that the right hand side of (4.31) converges in probability to zero and the assumptions on r_n, ν_n achieve the proof equation (4.29). Equation (4.28) is now a consequence of Theorem 30 p.63 applied to the variables $X_{ni} = n^{-1/2}\xi_i$, $1 \le i \le n$. ∎

Lemma 27 *Let the sequence $\bar\theta_n \in R^d$ be defined by the equation*

$$
\left\{\begin{array}{lcl}
\theta_n &=& \theta_{n-1}+\gamma_n K\theta_{n-1}+\gamma_n\zeta_n \\
\bar\theta_n &=& \frac{1}{n}\sum_0^{n-1}\theta_i
\end{array}\right. \tag{4.32}
$$

with $\theta_0 \in R^d$, $\zeta_n \in R^d$, matrix $K \in R^{d\times d}$ is non-degenerate, and γ_n be a positive decreasing sequence. Then

$$
\bar\theta_n = -\frac{1}{n}\sum_1^n K^{-1}\zeta_i + w_n
$$

with

$$
|w_n| \le K_0|\theta_n|/(n\gamma_n) + K_0|\theta_0|/\gamma_1 + \frac{K_0}{n}\sum_2^n |\theta_{i-1}|(\gamma_i^{-1}-\gamma_{i-1}^{-1}), \qquad K_0 = \|K^{-1}\|.
$$

Proof: From the first equation of the system (4.32) we derive that

$$
\theta_{n-1} = -\gamma_n^{-1}K^{-1}(\theta_{n-1}-\theta_n) - K^{-1}\zeta_n
$$

Taking average we obtain

$$
\begin{aligned}
\bar\theta_n &= \frac{1}{n}\sum_0^{n-1}\theta_i = -\frac{1}{n}\sum_1^n K^{-1}\zeta_i - \frac{1}{n}\sum_1^n \gamma_i^{-1}K^{-1}(\theta_{i-1}-\theta_i) \\
&= -\frac{1}{n}\sum_1^n K^{-1}\zeta_i + w_n
\end{aligned}
$$

Let us estimate w_n. Summing by parts we have

$$
w_n = -K^{-1}\theta_0/(n\gamma_1) + K^{-1}\theta_n/(n\gamma_n) - \frac{1}{n}\sum_2^n K^{-1}\theta_{i-1}(\gamma_i^{-1}-\gamma_{i-1}^{-1}).
$$

∎

4.5 Exercises

Exercise 4.5.1 (Necessity of assumption (E)) Show that if the solution to

$$\theta_n \quad = \quad (1-\gamma_n)\theta_{n-1} + \gamma_n \eta_n$$

satisfies (4.20), then (E) is satisfied.

Exercise 4.5.2 Give a heuristical justification of (4.3).

Exercise 4.5.3 (Quantile estimation) We consider the algorithm of Exercise 2.5.5.

- Prove that Theorem 24 applies.
- Compute the asymptotic variance (matrix V).
- What are the optimal gain γ^* and the corresponding variance V^*?

Exercise 4.5.4 (AR estimation) Compute the asymptotic variance of the algorithm of §3.3.3

Exercise 4.5.5 (MA estimation) Compute the asymptotic variance of the algorithm of §3.3.6

Exercise 4.5.6 (Costa loop) We consider the algorithm of Exercise 2.5.9.

- Give the explicit expression of $E[H(\theta, X_n)^2]$ as a fourth order polynomial in the variable $\tau = \sigma^2$ (use the results of Exercise 2.5.9).
- Compute the asymptotic variance when $\gamma_n = \gamma/n$. Which is the optimal γ and the corresponding variance?

Appendix A

Martingales

If $\mathcal{F}_n$ is an increasing sequence of σ-field, a process $(S_n)_{n>0}$ is a $\mathcal{F}_n$-martingale if

$$E[S_{n+1}|\mathcal{F}_n] = S_n, \quad n > 0.$$

and a $\mathcal{F}_n$-submartingale if

$$E[S_{n+1}|\mathcal{F}_n] \geq S_n, \quad n > 0.$$

Generally $\mathcal{F}_n = \sigma(S_1, ...S_n)$. The fundamental theorem on martingales is (Theorem 2.5 of [10]) :

Theorem 28 *If S_n is a submartingale such that*[1]

$$\sup_n E[(S_n)_+] < +\infty$$

then S_n converges almost surely to a r.v. of L_1.

Most useful results on discrete-time martigales may be found in the classical book [10]; for example the following classical theorem (Theorem 2.17) :

Theorem 29 *Let $\mathcal{F}_n$ be an increasing sequence of σ-fields and S_n an $\mathcal{F}_n$-martingale. Set $X_n = S_n - S_{n-1}$. If for some $1 \leq p \leq 2$*

$$\sum_{n=1}^{\infty} E[|X_n|^p|\mathcal{F}_{n-1}] < \infty \qquad a.s.$$

then S_n converges a.s.

Next theorem is concerned with the central limit-theorem :

Theorem 30 *Let $\{S_{ni}, \mathcal{F}_{ni}, 1 \leq i \leq n, 1 \leq n\}$ be a martingale array :*

$$E[S_{n,i+1}|\mathcal{F}_{ni}] = S_{ni}$$

and assume that the variables $X_{ni} = S_{ni} - S_{n,i-1}$ satisfy the conditions, for some $\varepsilon > 0$

$$\lim_n \sum_{i=1}^{n} E[|X_{ni}|^{2+\varepsilon}] = 0, \tag{A.1}$$

$$\sup_n \sum_{i=1}^{n} E[|X_{ni}|^2] < \infty \tag{A.2}$$

[1] We set $x_+ = (|x| + x)/2$.

and

$$\sum_{i=1}^{n} X_{ni} X_{ni}^T \xrightarrow{P} V \tag{A.3}$$

then

$$S_{nn} \xrightarrow{d} \mathcal{N}(0, V).$$

Proof: This theorem is a consequence of Theorem 3.2 of [10]. In this reference, the convergence holds if X_{ni} is one dimensional and if

$$\sup_i |X_{ni}| \xrightarrow{P} 0$$

and

$$E[\sup_i |X_{ni}|] \quad \text{is bounded in } n.$$

But under the assumptions (A.1) and (A.2), for any $\alpha > 0$,

$$P(\sup_i |X_{ni}| > \alpha) \leq E\left[\sum_{i=1}^{n} 1_{|X_{ni}|>\alpha}\right] \leq E\left[\sum_{i=1}^{n} \frac{|X_{ni}|^{2+\epsilon}}{\alpha^{2+\epsilon}}\right]$$

tends to zero and

$$E[\sup_i |X_{ni}|^2] \leq E[\sum_i |X_{ni}|^2]$$

is bounded. For the extension to arbitrary dimensions we apply the one dimensional result to $\langle S_{ni}, u\rangle$ for any vector u, and obtain the convergence in distribution of $\langle S_{nn}, u\rangle$ to $\mathcal{N}(0, u^T V u)$; this implies, using characteristic functions, the convergence of S_{nn} to $\mathcal{N}(0, V)$. ∎

Next theorem is a slightly modified version of Theorem 5.28 of [5]; it is used in Exercise 2.5.8 and concerns real valued martingales:

Theorem 31 *Let S_n be a submartingale such that for any $a > 0$*

$$E[\sup_n (S_n)_+ 1_{S_1,\ldots S_{n-1}<a}] < \infty.$$

If the sets A_1, A_2 are defined by

$$A_1 = \{\omega;\ \lim_n S_n \text{ exists}\}$$
$$A_2 = \{\omega;\ \sup_n S_n = +\infty\}$$

then $P(A_1 \cup A_2) = 1$.

Remarks: The theorem is not true without the condition. For instance consider independent v.a. X_n such that $P(X_n = n^2) = 1 - P(X_n = -1) = 1/(n^2+1)$; the Borel-Cantelli lemma implies that $X_n = n^2$ finitely often and $S_n = X_1 + \ldots + X_n$ is a martingale which tends to $-\infty$.

The condition is satisfied if

$$E[\sup_n (S_n - \sup_{k<n} S_k)_+] < \infty \quad \text{or if} \quad E[\sup_n (S_n - S_{n-1})_+] < \infty.$$

Proof: For any $a > 0$, denote by S_n^a the submartingale S_n stopped at its first entrance in the set $[a, +\infty[$:

$$\begin{aligned} S_n^a &= S_{n\wedge\tau_a} \\ \tau_a &= \min\{n, S_n \geq a\}. \end{aligned}$$

This new submartingale satisfies

$$(S_n^a)_+ \leq 1_{\tau_a > n} a + 1_{\tau_a \leq n}(S_{\tau_a})_+ \leq a + \sup_p (S_p)_+ 1_{S_1,\ldots S_{p-1} < a}.$$

Hence

$$E[(S_n^a)_+] \leq a + E[\sup_p (S_p)_+ 1_{S_1,\ldots S_{p-1} < a}].$$

Since $(S_n^a)_+$ is bounded in L_1 the submartingale S_n^a converges a.s. On the set $\{\omega; \sup_n S_n < a\}$, S_n and S_n^a coincide, and thus S_n converges. Since this is true for any a, we obtain the convergence of S_n on $\{\omega; \sup_n S_n < +\infty\}$. ∎

Appendix B

Markov chains and the Poisson equation

B.1 Introduction to the Poisson equation

Poisson equation is by far the simplest tool for obtaining limit theorems for Markov chains; it reduces the study of Markov chain to classical martingale results.

Consider a Markov chain X_n with transition kernel $\Pi(x, dy)$ and invariant measure π

$$E[f(X_n)|X_{n-1} = x] = (\Pi f)(x) = \int f(y)\Pi(x, dy)$$
$$\pi(f) = \int f(y)\pi(dy).$$

Π denotes both the transition kernel and the transition operator. Suppose we are given a function $f(x)$. The Poisson equation is

$$(I - \Pi)v = f - \pi(f). \tag{B.1}$$

If it exists, its solution v is the sum (converging in some sense)

$$v = \sum_{k=0}^{\infty} \Pi^k(f - \pi(f)) = \sum_{k=0}^{\infty}(\Pi^k(f) - \pi(f)). \tag{B.2}$$

In many cases (geometric ergodicity, property (GE) below) this series is exponentially convergent, under weak conditions on f. If X_n is a realization of the Markov chain, one can write

$$\begin{aligned} f(X_1) + f(X_2) + ... + f(X_n) &= n\pi(f) + (I - \Pi)v(X_1) + (I - \Pi)v(X_2) + ... + (I - \Pi)v(X_n) \\ &= n\pi(f) + v(X_1) + (v(X_2) - \Pi v(X_1)) + ... + (v(X_n) - \Pi v(X_{n-1})) + \Pi v(X_n) \\ &= n\pi(f) + v(X_1) + M_n + \Pi v(X_n) \end{aligned}$$

where M_n is a martingale. From this formula, one can easily deduce large number laws and the central limit-theorem for this kind of sums, by simply using well-known limit theorems on martingles [10].

The existence of the solution to the Poisson equation(B.1) will clearly depend on the rate of convergence of Π^n to π (cf equation (B.2)). For this purpose, we present in Section B.2 some conditions under which the convergence of Π^n to π is geometric. Section B.3 is devoted to the dependence (regularity) w.r.t θ of the solutions $v_\theta(x)$ given a parametrised familly $f_\theta(x)$.

In this appendix we give conditions under which the solution to the Poisson equation exists for a quite large class of functions f. This implies some restrictions on the Markov chain X_n, especially in Section B.3. Another way to deal with this problem would be to impose more restrictions on f (regularity), and less on Π, as is done in the usual operator-theoretic treatment of Markov chains [22].

B.2 Geometric ergodicity

B.2.1 The finite case

Consider a Markov chain X_n on a finite state space $\mathcal{X}$. In this situation, it is well known that if the convergence of Π^n to π holds, it is geometric; let us give a condition under which this happens.

Π may be seen as a matrix whose p_{ij} entry denotes the transition probability from state i to state j, and the transition operator acts as left product by Π:

$$E[f(X_n)|X_{n-1}=x] = (\Pi f)(i) = \sum p_{ij} f(j),$$

the function f being identified with a vector. We denote by $p_{ij}^{(k)}$ the k-step transition probability :

$$E[f(X_n)|X_{n-k}=x] = (\Pi^k f)(x) = \sum p_{ij}^{(k)} f(j)$$

The convergence of the series (B.2) is governed by the rate of convergence of the $p_{ij}^{(k)}$ to the invariant measure π_j. It is standard that this convergence holds if and only if the matrix Π enjoys the following properties :

- irreducibility : the eigenspace associated to 1 is unique and has dimension 1.
- aperiodicity : non-existence an eigenvalue of modulus one different from 1.

We consider the assumption (IA) which is equivalent to these properties ;

(IA) *(Irreducibility and aperiodicity) There exists a state $e \in \mathcal{X}$ and $n > 0$, such that for any $i \in \mathcal{X}$, $p_{ie}^{(n)} > 0$ (i.e. there exists $i_0 = i, i_1, i_2 ... i_{n-1}, i_n = e$, such that $p_{i_{k-1}i_k} > 0$, $k = 1, ...n$).*

This means that some power of the transition matrix has a full positive column. If n were allowed to depend on i in this assumption, only irreducibility would hold and aperiodicity would not be guaranteed (consider for instance the chain : $\mathcal{X} = \{0,1\}$, $p_{00} = p_{11} = 0$, $p_{01} = p_{10} = 1$). We have the classical theorem :

Theorem 32 *Under assumption (IA), the invariant measure π is unique and the transition probabilities $p_{ij}^{(k)}$ converge at a geometric rate to π_j.*

In particular, if (IA) is satisfied, equation (B.1) has always a solution. This solution is unique up to an additive constant and one of them is given by equation (B.2).

B.2.2 The general case

We consider a topological space $\mathcal{X}$ (typically a part of $\mathbf{R}^d$), endowed with its Borel σ-field. For any positive function V on $\mathcal{X}$, we will denote

$$\mathcal{C}_V = \{f \text{ continuous}, \sup_x \frac{|f(x)|}{V(x)} < \infty\}$$
$$\|f\|_V = \sup_x \frac{|f(x)|}{V(x)} \quad f \in C_V$$

and the operator norm

$$\|\Pi\|_V = \sup_f \frac{\|\Pi f\|_V}{\|f\|_V}.$$

We consider also a Markov transition kernel Π, defined at least for bounded Borel functions and non-negative Borel functions:

$$(\Pi f)(x) = \int_{\mathcal{X}} f(y)\Pi(x, dy) = E[f(X_1)|X_0 = x]$$

Consider the property:

(GE) Π *is a Markov transition probability with a unique invariant probability* π *such that, for some function* $V \geq 1$, $\|\Pi\|_V < \infty$ *and the spectral radius of* $\Pi - \pi$ *in* $\mathcal{C}_V$ *is smaller than one:*

$$\|\Pi^n - \pi\|_V \leq C\rho^n, \qquad \rho < 1. \tag{B.3}$$

and the assumptions:

(ST) *(Stability) There exist a continuous function* $V \geq 1$ *which tends to infinity*[1], *and* $p, C, \varepsilon > 0$ *such that*

$$\Pi^p V \leq (1-\varepsilon)\ V + C, \tag{B.4}$$
$$\Pi V \leq CV. \tag{B.5}$$

(AP) *(Aperiodicity) There does not exist sets* $E_1, ...E_n$ *such that* $n > 1$ *and*

$$\mathcal{X} = \cup E_i \tag{B.6}$$
$$\Pi(x, E_{i+1}) = 1, \quad x \in E_i, \quad i = 1, ...n \tag{B.7}$$

with $E_{n+1} = E_1$.

(IR) *(ψ-irreducibility) There exist a measure* ψ, *whose support has a non-empty interior, such that for any* ψ*-positive set* A

$$P_x(X_n \in A \text{ for some } n \geq 1) > 0$$

for all x.

[1] If $\mathcal{X}$ is compact, the function 1 tends to infinity ad (ST) is automatically satisfied.

Then we have the following result consequence of Theorems 6.0.1, 6.2.5, and 15.0.1 of [20]:

Theorem 33 *Let $\Pi(x, dy)$ be a Feller Markov transition probability. If Π satisfies (ST,AP,IR) then (GE) holds.*

It is worth noticing that if X_n converges in distribution to some variable of law ψ, then (AP) and (IR) are satisfied. Indeed, on the one hand (B.7) implies that $\psi(E_i) = 1$, $i = 1, ...n$, hence $n = 1$, and on the other hand for any measurable set A we have:

$$\begin{aligned} P_x(X_n \in A \text{ for some } n \geq 1) &= E_x[\sup_{n\geq 1} 1_{X_n \in A}] \\ &= 1 - E_x[\inf_{n\geq 1} 1_{X_n \notin A}] \\ &\geq 1 - E_x[\liminf_n 1_{X_n \notin A}] \\ &\geq 1 - \liminf_n E_x[1_{X_n \notin A}] \quad \text{(Fatou)} \\ &= \psi(A). \end{aligned}$$

B.2.3 Example: Linear models.

We consider here models of the form

$$X_n = FX_{n-1} + GW_n$$

where $X_n \in \mathbf{R}^d$, W is an $\mathbf{R}^p$-valued i.i.d. sequence, and the pair (F, G) belongs to the set of stable controllable pairs

$$\mathcal{M}_c = \{F \in \mathbf{R}^{d\times d}, G \in \mathbf{R}^{d\times p}, \rho(F) < 1, \text{ the matrix } [F^{d-1}G|, \ldots |FG|G] \text{ has rank } d\}.$$

($\rho(.)$ denotes the spectral radius).

Theorem 34 *Let $(F, G) \in \mathcal{M}_c$. We assume that the absolutely continuous part of the distribution of W_n has a density bounded away from 0 on some open set and that for some $r > 0$*

$$E[|W_n|^r] < \infty.$$

Then assumption (GE) is satisfied with $V(x) = 1 + |x|^r$. If $\|W_n\|_\infty < \infty$, assumption (GE) is satisfied with $V = 1$.

Proof: We shall prove that (G1), (AP) and (IR) hold true. Equations (B.4) and (B.5) are immediate thanks to the equation

$$\begin{aligned} X_{n+k} &= F^k X_n + F^{k-1}GW_{n+1} + \ldots FGW_{n+k-1} + GW_{n+k} \qquad \text{(B.8)} \\ (\Pi^k V)(x) &= E[V(F^k x + F^{k-1}GW_1 + \ldots FGW_{k-1} + GW_k)] \end{aligned}$$

and the stability of F (F^k is arbitrarily small if k is large enough). Equation (B.8) proves also that X_k converges in distribution to

$$X'_k = \sum_{j=0}^{+\infty} F^j GW_{k-j}$$

because

$$E[V(X'_k - X_k)] = E[V(\sum_{j=k}^{+\infty} F^j GW_{j+k} - F^k X_n)] = O(\|F^k\|^r).$$

This inplies that (AP) and (IR) are satisfied (cf the remark following Theorem 33 p.70). ∎

B.3 Continuity with respect to a parameter : the compact case

B.3.1 Theory

The problem of the continuity with respect to a parameter is quite difficult. We treat here only the compact case, that is the situation where $\mathcal{X}$ is compact and $V = 1$ (of course this covers the case where $\mathcal{X}$ finite).

Theorem 35 *We are given a family of Markov transition operators Π_θ, $\theta \in \mathcal{K}$; $\mathcal{K}$ is a compact set of $\mathbf{R}^d$. We assume that assumption (GE) is satisfied for each Π_θ, where $V = 1$, and n and C may depend on θ. We assume also that for some $q \geq 1$*

$$\int |\Pi_\theta^q(x, dy) - \Pi_{\theta'}^q(x, dy)| \leq C|\theta - \theta'| \qquad x \in \mathcal{X}, \theta, \theta' \in \mathcal{K}. \tag{B.9}$$

Then, for each Π_θ, (GE) is satisfied, and ρ can be chosen independently of θ. The invariant measure of Π_θ will be denoted by π_θ.

Consider functions H_θ which satisfy

$$|H_\theta(x)| \leq C \tag{B.10}$$

$$|\Pi_\theta^k H_\theta(x) - \Pi_{\theta'}^k H_{\theta'}(x)| \leq C|\theta - \theta'|, \quad x \in \mathcal{X}, \theta, \theta' \in \mathcal{K}, 0 \leq k < q. \tag{B.11}$$

Then for all $\theta \in \mathcal{K}$, the solution v_θ to the Poisson equation exists :

$$(I - \Pi_\theta)v_\theta = H_\theta - \pi_\theta(H_\theta) \tag{B.12}$$

and satisfies

$$|v_\theta(x)| \leq C' \tag{B.13}$$

$$|v_\theta(x) - v_{\theta'}(x)| \leq C'|\theta - \theta'|, \qquad x \in \mathcal{X}, \theta, \theta' \in \mathcal{K}. \tag{B.14}$$

Remarks : In many situations, the choice $q = 1$ may work; we shall see however that it happens in common cases that q has to be chosen larger; the point is that the support of Π_θ may depend heavily on θ, in which case we can have

$$|\Pi_\theta(x, dy) - \Pi_{\theta'}(x, dy)| = \Pi_\theta(x, dy) + \Pi_{\theta'}(x, dy)$$

and (B.9) will not be satisfied, whereas Π_θ^q depends really smoothly on θ. This is the situation for the chain (X_n, Y_n) :

$$(X_n, Y_n) = (\theta X_{n-1} + Y_{n-1}, \xi_n)$$

where ξ_n is an i.i.d. sequence.

Proof : We start with the case $q = 1$; the general case will be an easy consequence. Since property (GE) holds for each Π_θ, then $\Pi_\theta = \pi_\theta + R_\theta$ where π_θ is the invariant measure and R_θ is an operator with spectral radius smaller than one. Since equation (B.9) means exactly that the function

$$\theta \longrightarrow \Pi_\theta$$

is continuous (w.r.t. the norm $\|.\|_V$ previously defined), theorem 3.16 p.212 of [12] (cf also IV.3.5 p.213) implies that the decomposition

$$\Pi_\theta = \pi_\theta + R_\theta \tag{B.15}$$

holds continuously w.r.t. θ with the explicit expression

$$\pi_\theta = -\frac{1}{2i\pi}\int_\Gamma (\Pi_\theta - zI)^{-1} z dz$$

where Γ is a small contour in the complex plane around 1; notice that the convergence of the integral and its continuity with respect to θ follows from the property (GE) and equation (B.9) : 1 is an isolated point of the spectrum of Π_θ (we refer to [12] for details). In particular, we have for θ' close enough to θ (so that Γ can be chosen independently of θ')

$$\begin{aligned}
\|\pi_\theta - \pi_{\theta'}\|_V &\leq \sup_{z\in\Gamma} \|(\Pi_\theta - zI)^{-1} - (\Pi_{\theta'} - zI)^{-1}\|_V \\
&= \sup_{z\in\Gamma} \|(\Pi_\theta - zI)^{-1}(\Pi_\theta - \Pi_{\theta'})(\Pi_{\theta'} - zI)^{-1}\|_V \\
&= \sup_{z\in\Gamma} \|(\Pi_\theta - zI)^{-1}\|_V \ \|\Pi_\theta - \Pi_{\theta'}\|_V \sup_{z\in\Gamma} \ \|(\Pi_{\theta'} - zI)^{-1}\|_V \\
&\leq C|\theta - \theta'|
\end{aligned}$$

The continuity of the decomposition (B.15), the upper semi-continuity of the spectral radius and the compactness of $\mathcal{K}$ implies that ρ may be chosen independently of θ.

Property (GE) implies that the series

$$v_\theta = \sum_{k=0}^{\infty} (\Pi_\theta - \pi_\theta)^k H_\theta \tag{B.16}$$

converges exponentially fast in C_V; the first equality implies that v_θ is solution to the Poisson equation (B.12) and (B.13) holds.

Equation (B.16) implies that $\pi_\theta v_\theta = 0$ and, setting

$$\tilde{H} = H_\theta - \pi_\theta(H_\theta),$$

the following formula holds

$$\begin{aligned}
(I - \Pi_\theta + \pi_\theta)(v_\theta - v_{\theta'}) &= (I - \Pi_\theta)v_\theta - (I - \Pi'_\theta)v_{\theta'} + (\Pi_\theta - \pi_\theta - \Pi'_\theta)v_{\theta'} \\
&= \tilde{H}_\theta - \tilde{H}_{\theta'} + (\Pi_\theta - \Pi_{\theta'})v_{\theta'} + (\pi_\theta - \pi_{\theta'})v_{\theta'}.
\end{aligned}$$

Since the operator $R_\theta = \Pi_\theta - \pi_\theta$ has a spectral radius smaller than 1, $(I - R_\theta)^{-1}$ has a finite norm and

$$\begin{aligned}
\|v_\theta - v_{\theta'}\|_V &\leq C'\|\tilde{H}_\theta - \tilde{H}_{\theta'} + (\Pi_\theta - \Pi_{\theta'})v_{\theta'} + (\pi_\theta - \pi_{\theta'})v_{\theta'}\|_V \\
&\leq C''|\theta - \theta'|.
\end{aligned}$$

This proves the result in the case $q = 1$. If $q > 1$, we can apply the result with

$$\begin{aligned}
\Pi'_\theta &= \Pi_\theta^q \\
H'_\theta &= (I + \Pi_\theta + \dots \Pi_\theta^{q-1})H_\theta.
\end{aligned}$$

We obtain the existence of some function v_θ such that

$$\begin{aligned}
(I - \Pi_\theta^q)v_\theta &= (I + \Pi_\theta + \dots \Pi_\theta^{q-1})H_\theta - \pi_\theta((I + \Pi_\theta + \dots \Pi_\theta^{q-1})H_\theta) \\
&= (I + \Pi_\theta + \dots \Pi_\theta^{q-1})(H_\theta - \pi_\theta(H_\theta)).
\end{aligned}$$

Multiplying both sides by $I - \Pi_\theta + \pi_\theta$ leads to

$$(I - \Pi_\theta^q)(I - \Pi_\theta + \pi_\theta)v_\theta = (I - \Pi_\theta^q)(H_\theta - \pi_\theta(H_\theta))$$

and since $\Pi_\theta \pi_\theta = \pi_\theta \Pi_\theta = \pi_\theta$ this formula rewrites

$$(I - \Pi_\theta^q + \pi_\theta)(I - \Pi_\theta)v_\theta \quad = \quad (I - \Pi_\theta^q + \pi_\theta)(H_\theta - \pi_\theta(H_\theta)).$$

Since the spectral radius of $\Pi_\theta^q - \pi_\theta$ is strictly smaller than one, $I - \Pi_\theta^q + \pi_\theta$ is invertible and we conclude that the Poisson equation is satisfied

$$(I - \Pi_\theta)v_\theta \quad = \quad H_\theta - \pi_\theta(H_\theta).$$

∎

B.3.2 Example : Linear models.

We consider here models of §B.2.3 with $\theta = (F, G)$.

Theorem 36 *Let $\mathcal{K}$ be a compact subset of $\mathcal{M}_c$ and let the assumptions of Theorem 34 hold and assume that W_n has a C^1 density with compact support, then (B.9) holds with $q = d$.*

Proof : We have to bound

$$\int \left| \Pi_\theta^d(x, dy) - \Pi_{\theta'}^d(x, dy) \right| .$$

We will use the smoothing effect of the density of W_n. We have for any bounded function f

$$\int f(y)\Pi_\theta^d(x, dy) \quad = \quad \int f(F^d x + Mw)p(w)dw \tag{B.17}$$

with

$$\begin{aligned} M &= [F^{d-1}G|\ldots|FG|G] \\ \bar{W} &= \begin{pmatrix} W_1 \\ \vdots \\ W_{d-1} \\ W_d \end{pmatrix}. \end{aligned}$$

We are going to take the derivative of equation (B.17) w.r.t. M. For this, we will use the convention that the matrix derivative of a scalar function $\partial_M u(M)$ is the transposed of the matrix of $\partial_{M_{ij}} u(M)$ and gradients are row vectors, so that $\partial_x f(Mx) = \nabla f(Mx)M$ and $\partial_M f(Mx) = x\nabla f(Mx)$. We will also denote by M^+ the pseudo-inverse of the full-rank matrix M (i.e. $MM^+ = I$). Let f be a C^1 function :

$$\begin{aligned} \partial_M \int f(F^d x + Mw)p(w)dw &= \partial_M \int g(Mw)p(w)dw, \quad g(u) = f(F^d x + u) \\ &= \int w\nabla g(Mw)p(w)dw \\ &= \int w\nabla g(Mw)Mp(w)dw \; M^+ \\ &= \int w(\partial_w g(Mw))p(w)dw \; M^+ \\ &= -\int \partial_w(wp(w)) \; g(Mw)dw \; M^+ \\ &= -\int \partial_w(wp(w))f(F^d x + Mw)dw \; M^+. \end{aligned} \tag{B.18}$$

For the regularity w.r.t. F, we simply write

$$\int f(F^d x + Mw)p(w)dw = \int f(Mz)p(z - M^+ F^d x)dz. \tag{B.19}$$

From equation (B.18) and (B.19) , there exists some constant C such that

$$|\partial_\theta \int f(y)\Pi_\theta^d(x, dy)| \leq C\|f\|_\infty$$

for any C^1 function f; by density, this extends to Borel functions. Since there exists a function $s(y)$, with value in $\{-1, 1\}$, such that

$$\int \left|\Pi_\theta^d(x, dy) - \Pi_{\theta'}^d(x, dy)\right| = \int s(y)\left(\Pi_\theta^d(x, dy) - \Pi_{\theta'}^d(x, dy)\right)$$

we have

$$\int \left|\Pi_\theta^d(x, dy) - \Pi_{\theta'}^d(x, dy)\right| = \int s(y)\Pi_\theta^d(x, dy) - \int s(y)\Pi_{\theta'}^d(x, dy) \leq C|\theta - \theta'|.$$

∎

Appendix C

Some basic facts on stability

C.1 Lyapunov functions

This section is a simple introduction to Lyapunov functions and their use for the study of the stability of dynamical systems. Let us mention first the classical Lyapunov theorem which is evidently closely related to the assumptions of Section 2:

Theorem 37 *(Lyapunov) Let h be a continuous vector field defined on an open set $\mathcal{O}$ and x_t be the solution to*

$$\frac{dx_t}{dt} = h(x_t), \quad x_0 \in \mathcal{O}. \tag{C.1}$$

We assume that there exists a function $V \geq 0$, continuously differentiable on $\mathcal{O}$ such that

$$\lim_{x \to \partial\mathcal{O}} V(x) = +\infty$$
$$\langle \nabla V(x), h(x) \rangle \leq 0.$$

Then x_t is bounded and

$$\lim_{t\to\infty} \langle \nabla V(x_t), h(x_t) \rangle = 0.$$

Proof: We do not give all the details. One has

$$\frac{dV(x_t)}{dt} = \langle \nabla V(x_t), h(x_t) \rangle.$$

This implies that $V(x_t)$ decreases to some limit. In particular x_t remains in a compact subset $\mathcal{K}$ of $\mathcal{O}$. This equation implies also that

$$\int_0^\infty \langle \nabla V(x_t), h(x_t) \rangle \, dt < \infty.$$

Since the function $t \longrightarrow f(t) = \langle \nabla V(x_t), h(x_t) \rangle$ is uniformly continuous (x_t remains in $\mathcal{K}$), it is routine to prove (by contradition) that its integrability implies that it tends to zero. ∎

This theorem is generally used in situations where we have also an inequality like

$$\langle \nabla V(x), h(x) \rangle \leq -C \, |h(x)|^2$$

in which case we can deduce that $h(x_t)$ tends to zero.

There exists also theorems ensuring the existence of a Lyapunov function if the trajectories of h converge (cf [13], Theorem 5.3). For instance, the Theorem 3.2 of [31]:

Theorem 38 *Let $\mathcal{S}$ be a compact set such that , the solution x_t of (C.1) satisfies the uniform stability condition*

$$\lim_{x_0 \to \mathcal{S}} \sup_{t > 0} \sup d(\mathcal{S}, x_t) = 0$$

then there exists a Lyapunov function $V \geq 0$, infinitely differentiable

$$\begin{aligned}\lim_{x \to \partial \mathcal{O}} V(x) &= +\infty \\ \langle \nabla V(x), h(x) \rangle &\leq 0 \\ \langle \nabla V(x), h(x) \rangle &< 0, \quad x \notin \mathcal{S}.\end{aligned}$$

The typical example of a Lyapunov function is the following : consider a matrix A strictly stable (all its eigenvalues have a negative real part), and the dynamical system

$$\frac{dx_t}{dt} = A\ x_t, \quad x_0 = x. \tag{C.2}$$

Because of the stability of A, the norm of e^{tA} decreases exponentially fast to zero (cf Proposition 39 below) and the following matrix is well defined :

$$P = \int_0^\infty e^{tA^T} e^{tA} dt.$$

Notice that

$$A^T P + PA = \int_0^\infty A^T e^{tA^T} e^{tA} + e^{tA^T} e^{tA} A\ dt = \left[e^{tA^T} e^{tA} \right]_0^\infty = -I.$$

Hence, the function

$$V(x) = x^T P x$$

satisfies

$$\langle \nabla V(x), h(x) \rangle = 2x^T PAx = x^T (A^T P + PA) x = -|x|^2. \tag{C.3}$$

Thus V is a Lyapunov function for the system.

In the same way, one shows the stability of a time varying system

$$\frac{dx_t}{dt} = A_t\ x_t, \quad x_0 = x$$

if the matrices A_t keep close enough to A so that

$$A_t^T P + PA_t < 0 \quad \forall\ t.$$

C.2 Stable matrices

Proposition 39 *Let A be a matrix; for any number κ which is strictly larger than the largest real part of the eigenvalues of A, there exists a constant C (depending on κ and A) such that*

$$\|e^{tA}\| \leq C e^{t\kappa}, \qquad t \geq 0.$$

Proof: One can write $A = P^{-1}JP$ where J is a Jordan matrix and

$$e^{tA} = P^{-1}e^{tJ}P.$$

On the other hand, e^{tJ} is just the block-diagonal matrix made of the exponential of the Jordan blocks. These two remarks reduce the proof to the case where A is a Jordan block. In this case, $A = \lambda I + N$ where λ is the eigenvalue and N a nilpotent matrix and

$$e^{tA} = e^{t\lambda}e^{tN} = e^{t\lambda}\sum_{i=1}^{d-1}\frac{t^i}{i!}N^i$$

where d is the dimension, $N^d = 0$; hence

$$\|e^{tA}\| \leq C\ e^{tRe(\lambda)}(1+t^d)$$

what proves the result. ∎

Proposition 40 *Let $A(\theta)$ be a continuous function defined on a compact set $\mathcal{K}$ with values in the set of matrices with spectral radius strictly smaller than 1. There exists m such that*

$$\sup_{\theta\in\mathcal{K}} |A(\theta)^m| < 1.$$

Proof: Set

$$\mathcal{O}_n = \{\theta \in \mathcal{K}, |A(\theta)^n| < 1/2\}.$$

Since for each θ the spectral radius of $A(\theta)$ is smaller than one, we have

$$\cup_n \mathcal{O}_n = \mathcal{K}.$$

Since $\mathcal{O}_n$ is a sequence of open sets and $\mathcal{K}$ is compact, this implies that there exists n such that

$$\cup_{k=1}^n \mathcal{O}_k = \mathcal{K}.$$

Now, consider a number p and $\theta \in \mathcal{K}$. For some $k \leq n$, $\theta \in \mathcal{O}_k$, and p rewrites $p = ak+r$, $0 \leq r < k$; hence

$$|A(\theta)^p| = |A(\theta)^{ak+r}| \leq |A(\theta)|^r 2^{-a} = |A(\theta)|^r 2^{r/k}(2^{1/k})^{-p}.$$

Since $r < k \leq n$ and $|A(\theta)|$ is bounded on $\mathcal{K}$ one has

$$|A(\theta)^p| \leq C(2^{1/n})^{-p}$$

where C is independent of p. ∎

Bibliography

[1] A. Benveniste, M. Goursat, Blind Equalizers, *IEEE Trans on Com.*, vol 32, No 8, Aug. 1984.

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

[3] A. Berman, R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic-Press, 1979.

[4] D.P. Bertsekas, *Dynamic programming and optimal control*, Athena scientific, 1995.

[5] L. Breiman, *Probability*, Addison-Wesley,1968.

[6] J.-F. Cardoso, B. Laheld, "Equivariant adaptive source estimation", IEEE-SP, 1995 (ou 1996).

[7] H. F. Chen, L. Guo, A.-J. Gao, "Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds", *Stoch. Proc. Appl.* 27, 1988, No 2, 217-231.

[8] B. Delyon, General Results on the Convergence of Stochastic Algorithms, *IEEE-AC*, vol 41, No 9, Sept 1996, pp 1245-1256.

[9] M. Duflo, *Algorithmes Stochastiques*, Springer, 1996.

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

[11] M. Iosifescu, R. Theodorescu, *Random Processes and Learning*, Springer-Verlag, 1969.

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

[13] N. N. Krasovskii, *Stability of Motion*, Stanford University Press, 1963.

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

[15] H. J. Kushner, *Approximation and weak convergence for random processes, with applications to stochastic systems theory*, MIT Press, 1984.

[16] L. Ljung, P. Priouret, "A result on the mean square error tracking obtained using general tracking algorithms", *International J. Adaptive Control and Signal Proc.*, Vol. 5, 231-250, 1991.

[17] L. Ljung, T. Soderström, *Theory and Practice of Recursive Identification*, MIT Press, 1983.

[18] O. Macchi, E. Eweda, Bases théoriques pour l'égalisation adaptative en mode autodidacte, *Annales de télécommunication*, 41, No 5-6, mai-juin 1986.

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

[20] S.P. Meyn, R.L. Tweedie, *Markov chains and stochastic stability*, Springer, 1993.

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

[22] N. F. Norman, *Markov Processes and Lerning Models*, Academic Press, 1972.

[23] 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).

[24] G. Pflug, "Applicational aspects of Stochastic Approximation", in L. Ljung, G. Pflug, H. Walk, *Stochastic approximation and optimization of random systems*, Birkaüser, 1992.

[25] B. T. Polyak, New Stochastic Approximation Type Procedures, *Automatica i Telemekh.*, July 1990, pp. 98-107 (in Russian); translated as Autom. & Remote Contr., 51(7).

[26] W. Rudin, *Real and complex analysis*,Mc Graw-Hill, 1974.

[27] D. Ruppert, *Efficient estimators from a Slowly Convergent Robbins-Monro process* Technical report N. 781, School of Oper. Res. and Ind. Eng., Cornell University Ithaca, N.Y., 1988.

[28] G. N. Saridis, "Stochastic Approximation Methods for Identification and Control – A Survey", *IEEE-AC*, vol 19, No 6, décembre 1974.

[29] C. J. Watkins, P. Dayan, "Q–Learning", Machine Learning, 8, 279-292 (1992).

[30] B. Widrow, J. R. Glover, J. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, R. C. Goodlin, "Adaptive Noise Cancelling : Principles and Applications", *Proc. IEEE*, Vol. 63, No 12, Dec 1975.

[31] F. W. Wilson, "Smoothing Derivatives of Functions and Applications", *Trans. Am. Math.* 139, 413-428, 1969.