

STOCHASTIC APPROXIMATION WITH LONG RANGE DEPENDENT AND HEAVY TAILED NOISE

V. ANANTHARAM¹ AND V. S. BORKAR²

ABSTRACT: Stability and convergence properties of stochastic approximation algorithms are analyzed when the noise includes a long range dependent component (modeled by a fractional Brownian motion) and a heavy tailed component (modeled by a symmetric stable process), in addition to the usual ‘martingale noise’. This is motivated by the emergent applications in communications. The proofs are based on comparing suitably interpolated iterates with a limiting ordinary differential equation. Related issues such as asynchronous implementations, Markov noise, etc. are briefly discussed.

Key words: stochastic approximation, long range dependence, heavy tailed noise, o.d.e. limit, convergence in ξ th mean

¹Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA. Research supported by the ARO MURI grant W911NF-08-1-0233 “Tools for the Analysis and Design of Complex Multi-Scale Networks” and by the NSF grants CCF-0500234, CCF-0635372 and CNS-0627161.

²School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India. Research supported in part by the ARO MURI grant W911NF-08-1-0233 “Tools for the Analysis and Design of Complex Multi-Scale Networks” and the J. C. Bose Fellowship.

1 Introduction

We consider a stochastic approximation scheme in \mathcal{R}^d of the type

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1} + R(n)B_{n+1} + D(n)S_{n+1} + \zeta_{n+1}], \quad (1)$$

where

- $h = [h_1, \dots, h_d]^T : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is Lipschitz,
- $B_{n+1} := \tilde{B}(n+1) - \tilde{B}(n)$, where $\tilde{B}(t), t \geq 0$, is a d -dimensional fractional Brownian motion with Hurst parameter $\nu \in (0, 1)$,
- $S_{n+1} := \tilde{S}(n+1) - \tilde{S}(n)$, where $\tilde{S}(t), t \geq 0$, is a symmetric α -stable process with $1 < \alpha < 2$,
- $\{\zeta_n\}$ is an ‘error’ process satisfying $\sup_n \|\zeta_n\| \leq K_0 < \infty$ a.s. and $\zeta_n \rightarrow 0$ a.s.,
- $\{R(n)\}$ is a bounded deterministic sequence of $d \times d$ random matrices,
- $\{D(n)\}$ is a bounded sequence of $d \times d$ random matrices adapted to $\{\mathcal{F}_n\}$, for $\mathcal{F}_n := \sigma(x_i, B_i, M_i, S_i, \zeta_i, i \leq n)$.
- $\{M_n\}$ is a martingale difference sequence w.r.t. $\{\mathcal{F}_n\}$ satisfying

$$E[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K_1(1 + \|x_n\|^2). \quad (2)$$

- $\{a(n)\}$ are positive non-increasing stepsizes which are $\Theta(n^{-\kappa})$ for some $\kappa \in (\frac{1}{2}, 1]$. In particular, they satisfy:

$$\sum_n a(n) = \infty, \quad \sum_n a(n)^2 < \infty, \quad (3)$$

which are standard conditions for stochastic approximation. Clearly $\sup_n a(n) < \infty$. We assume without loss of generality that $\sup_n a(n) \leq 1$, this restriction does not affect our arguments in any essential way.

Consider the related o.d.e.:

$$\dot{x}(t) = h(x(t)). \quad (4)$$

We assume that this o.d.e. has a unique asymptotically stable equilibrium x^* , with an associated continuously differentiable Liapunov function (whose existence is guaranteed by the ‘smooth’ versions of converse Liapunov theorems – see, e.g., Theorem 3.2, p. 425, of [15]) $V : \mathcal{R}^d \rightarrow \mathcal{R}^+$ satisfying $\lim_{\|x\| \uparrow \infty} V(x) = \infty$ and $\langle \nabla V, h \rangle(x) < 0$ for $x \neq x^*$. In turn, existence of such a V implies global asymptotic stability of x^* (*ibid.*). Our main result will be:

Theorem 1 Suppose

(†) $K_2 := \sup_n E[\|x_n\|^\xi] < \infty$ for some $\xi \in [1, \alpha)$.

Then for $1 < \xi' < \xi$,

$$E[\|x_n - x^*\|^{\xi'}] \rightarrow 0. \quad (5)$$

Here (†) is a ‘stability of iterates’ condition. A sufficient condition for (†) is given in section 4.

The result is motivated by the several applications of stochastic approximation in communication networks. Some common scenarios are:

1. *Gradient schemes*: Here $h = -\nabla F$ for some $F : \mathcal{R}^d \rightarrow \mathcal{R}$ which we seek to minimize. Suppose F has a unique global minimizer x^* . Then $V \equiv F$ will serve as the Liapunov function required by our theorem. See [6] for an example.
2. *Saddle point seekers*: Consider $F : \mathcal{R}^m \times \mathcal{R}^m \rightarrow \mathcal{R}$ which is strictly convex in its first argument for each value of the second and strictly concave in its second argument for each value of the first, with a unique saddle point $x^* = (y^*, z^*) \in \mathcal{R}^d$ for $d = 2m$. Let $h(y, z) := [-\nabla^y F(y, z) : \nabla^z F(y, z)]^T$, where ∇^y, ∇^z denote the gradients in y and z variables, resp. Then $V(x) = \|x - x^*\|^2$ serves as a Liapunov function. See [16] for details and a specific scenario.
3. *Fixed point seekers*: Let $h(x) = F(x) - x$, i.e., $h(x^*) = 0 \iff x^*$ is a fixed point of F . If $-F$ is monotone, i.e., $\langle F(x) - F(y), x - y \rangle \leq 0$, then $V(x) = \|x - x^*\|^2$ serves as a Liapunov function.

$y) < 0$ whenever $x \neq y$, and has x^* as a fixed point, then this fixed point is unique and $V(x) := \|x - x^*\|^2$ again serves as a Liapunov function. See [7] for an instance of this. Also, if F is a contraction w.r.t. the norm $\|\cdot\|_p, 1 < p < \infty$, then also F has a unique fixed point x^* by the contraction mapping principle and $\|x - x^*\|_p$ works as a Liapunov function. In fact, this extends to $p = \infty$ if the continuous differentiability condition on V is replaced by continuity alone and the requirement ' $\langle \nabla V, h \rangle(x) < 0$ for $x \neq x^*$ ' is replaced by the requirement: ' $V(x(t))$ decreases along any nonconstant trajectory of (4)'. This is a case of great interest in approximate dynamic programming [1].

The key contribution of this work is to consider such stochastic approximation schemes with *long range dependent* and *heavy tailed* noise. In (1), these aspects are captured by the processes $\{B_n\}$ and $\{S_n\}$ resp. It is well known that the noise processes in the Internet and several other situations arising in communications exhibit such behavior, a fact which has also been theoretically justified through limit theorems such as [11]. That this does introduce significant additional complications for stochastic approximation schemes is reflected in the fact that the convergence claim in (5) is 'in ξ' th mean' for $1 < \xi' < \xi < \alpha$ where α is the index of stability of the heavy tailed part of the noise and ξ is as in (†), and not 'a.s.' as is usually the case [5]. (We can, however, improve the claim to 'a.s.' if the heavy tailed component is missing, as we observe later in section 5.)

We follow the 'o.d.e.' approach to the analysis of stochastic approximation, see, e.g., [5] for the classical version. The idea is to treat (1) as a noisy discretization of (4) and then argue that the errors due to both discretization and noise become asymptotically negligible in ξ' th mean under the stated hypotheses. It then follows that (1) has the same asymptotic limit in ξ' th mean as that of (4). In section 2, we use Gronwall inequality to get a bound on the maximum deviation in norm between a certain piecewise linear interpolation of the iterates on one hand and the solution of the differential equation on the other, over a time window of fixed width, so that both agree at the beginning of the window. This estimate is quite standard in the o.d.e. approach to stochastic approximation (see, e.g., [5], Lemma 1, p. 12, also [2], [3], [12]) and the only difference is the additional error terms on the right hand side. Nevertheless we include it in some detail because it is key to the development that follows. In section 3, we obtain moment estimates for the

error terms. Whereas the heavy tailed component of the noise is responsible for the weakening of the claim from ‘almost sure’ to ‘in ξ 'th mean’ (as will become apparent later), the error estimate for this component is in fact easy thanks to already available estimates. It is the long range dependent component of the noise that takes the bulk of the effort. Section 4 proves Theorem 1 and gives a sufficient condition for (†). Section 5 strengthens the conclusions to ‘almost sure’ convergence for a special case. Section 6 sketches the corresponding developments for the constant stepsize algorithms, i.e., when $a(n) \equiv a > 0$. Section 7 concludes with assorted comments about generalizations of use in applications.

Throughout this paper, $C > 0$ will denote a generic constant which may differ from place to place, even within the same string of equations / inequalities. $\|\cdot\|$ will denote the standard Euclidean norm unless otherwise specified.

2 Preliminaries

The o.d.e. approach is based on comparing with trajectories of (4) the continuous interpolation $\{\bar{x}(t), t \geq 0\}$ of the iterates $\{x_n\}$ defined as follows: Let $t(0) = 0, t(n) = \sum_{i=0}^{n-1} a(i), n \geq 1$. Then $t(n) \uparrow \infty$. Set $\bar{x}(t(n)) = x_n \forall n$ and interpolate linearly on $[t(n), t(n+1)]$ for all $n \geq 0$. For $n \geq 0$, let $x^n(t), t \geq t(n)$, denote the trajectory of (4) on $[t(n), \infty)$ with $x^n(t(n)) = \bar{x}(t(n)) := x_n$. Fix $T > 0$ and for $n \geq 0$, let

$$m(n) := \min\{j \geq n : t(j) \geq t(n) + T\}.$$

Since $\sup_n a(n) \leq 1, t(m(n)) \in [t(n) + T, t(n) + T + 1]$. We then have:

Lemma 1 For a constant $K(T) > 0$ depending on T and the Lipschitz constant of h ,

$$\begin{aligned} & \sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^n(t)\| \\ & \leq K(T) \left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|\bar{x}(t(n))\|) + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) \zeta_{i+1} \right\| \right) \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) M_{i+1} \right\| + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \end{aligned}$$

$$+ \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)D(i)S_{i+1} \right\| + a(n). \quad (6)$$

Proof We have

$$\begin{aligned} \bar{x}(t(n+k)) &= \\ & \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(\bar{x}(t(n+i))) \\ & + \sum_{i=0}^{k-1} a(n+i)M_{n+i+1} + \sum_{i=0}^{k-1} a(n+i)R(n+i)B_{n+i+1} \\ & + \sum_{i=0}^{k-1} a(n+i)D(n+i)S_{n+i+1} + \sum_{i=0}^{k-1} a(n+i)\zeta_{n+i+1}. \end{aligned} \quad (7)$$

Compare this with

$$\begin{aligned} x^n(t(n+k)) &= \\ & \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(x^n(t(n+i))) \\ & + \sum_{i=0}^{k-1} \int_{t(n+i)}^{t(n+i+1)} (h(x^n(y)) - h(x^n(t(n+i))))dy. \end{aligned} \quad (8)$$

Note that for $t(\ell) \leq t \leq t(\ell+1)$,

$$x^n(t) - x^n(t(\ell)) = \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell))))ds + \int_{t(\ell)}^t h(x^n(t(\ell)))ds.$$

Since h is Lipschitz and therefore of linear growth,

$$\|x^n(t) - x^n(t(\ell))\| \leq C \int_{t(\ell)}^t \|x^n(s) - x^n(t(\ell))\|ds + C(1 + \|x^n(t(\ell))\|)a(\ell).$$

By the Gronwall inequality,

$$\sup_{t \in [t(\ell), t(\ell+1)]} \|x^n(t) - x^n(t(\ell))\| \leq Ca(\ell)(1 + \|x^n(t(\ell))\|).$$

By the Lipschitz property of h ,

$$\sup_{t \in [t(\ell), t(\ell+1)]} \left\| \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell))))ds \right\| \leq Ca(\ell)^2(1 + \|x^n(t(\ell))\|).$$

On the other hand, a standard argument based on the Gronwall inequality shows that

$$\sup_{t \in [t(n), t(m(n))]} \|x^n(t)\| \leq C \|\bar{x}(t(n))\|.$$

Thus

$$\sup_{t \in [t(\ell), t(\ell+1)]} \left\| \int_{t(\ell)}^t (h(x^n(s)) - h(x^n(t(\ell)))) ds \right\| \leq Ca(\ell)^2 (1 + \|\bar{x}(t(n))\|). \quad (9)$$

Subtracting (8) from (7), using (9) and the discrete Gronwall inequality, we have

$$\begin{aligned} & \sup_{n \leq i \leq m(n)} \|\bar{x}(t(i)) - x^n(t(i))\| \\ & \leq C \left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|\bar{x}(t(n))\|) \right) + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) \zeta_{i+1} \right\| \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) M_{i+1} \right\| + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\|. \end{aligned}$$

The claim now follows as in [5], p. 14. □

3 Moment estimates

We begin by analyzing the error term in (6) due to the fractional Brownian motion. We have

$$E[\|\tilde{B}(t) - \tilde{B}(s)\|^2] = C|t - s|^{2\nu}, \quad t \geq s,$$

and,

$$\begin{aligned} & E[(\tilde{B}(t) - \tilde{B}(s))(\tilde{B}(u) - \tilde{B}(v))^T] \\ & = \frac{C}{2} [|t - v|^{2\nu} + |s - u|^{2\nu} - |t - u|^{2\nu} - |s - v|^{2\nu}] I, \quad v \leq u \leq s \leq t, \end{aligned}$$

where ‘ I ’ denotes the identity matrix. Hence, using the fact that the $\{R(n)\}$ are bounded,

$$\begin{aligned} E[\|\sum_{i=n}^N a(i)R(i)(\tilde{B}(i+1) - \tilde{B}(i))\|^2] &\leq C\left(\sum_{i=n}^N a(i)^2 + \right. \\ &\quad \left. 2 \sum_{n \leq i < k \leq N} a(i)a(k) |k-i+1|^{2\nu} + |k-i-1|^{2\nu} - 2|k-i|^{2\nu} \right) \quad (10) \\ &:= \hat{\sigma}^2(n, N). \end{aligned}$$

Lemma 2 $\hat{\sigma}^2(n, m(n)) \leq C \left(\frac{1}{n^\gamma}\right)$ for $\gamma := 2\kappa(1-\nu) > 0$ for $\nu > \frac{1}{2}$ and $\gamma := \kappa$ for $\nu \leq \frac{1}{2}$.

Proof For any f , we have

$$|f(x+1) + f(x-1) - 2f(x)| \leq 2 \max_{y \in [x-1, x+1]} |f''(y)|.$$

Using $f(x) = |x|^{2\nu}$ for $\nu > \frac{1}{2}$ and $f(x)$ a modification of $|x|^{2\nu}$ suitably smoothed near $x = 0$ when $\nu \leq \frac{1}{2}$, it can be verified that

$$| |k-m+1|^{2\nu} + |k-m-1|^{2\nu} - 2|k-m|^{2\nu} | \leq C(|k-m|^{-\eta} \wedge 1) \quad (11)$$

for $\eta := 2 - 2\nu \in (0, 2)$ and $k \neq m+1$, where we interpret $0^{-\eta}$ as $+\infty$ for $\eta > 0$. For $k = m+1$, we have the left hand side above equal to $2^{2\nu} - 2$. Combining, we have (11) for all k, m . Thus

$$\begin{aligned} &2 \sum_{n \leq i < k \leq N} a(i)a(k) |k-m+1|^{2\nu} + |k-m-1|^{2\nu} - 2|k-m|^{2\nu} | \\ &\leq C \sum_{n \leq i \leq k \leq N} a(i)a(k) \psi(|k-m|) \end{aligned} \quad (12)$$

where $\psi(x) := \frac{1}{x^\eta} \wedge 1, x \geq 0$. By the Perron–Frobenius theorem, the maximum eigenvalue of the matrix $[[\psi(|k-m|)]]_{k,m}$ is bounded by its maximum row sum, which is

$$\begin{aligned} 2 \sum_{i=1}^{\lceil (N-n)/2 \rceil} \frac{1}{i^\eta} &\leq C(|N-n|^{1-\eta}) \quad \text{for } \eta < 1, \\ &\leq C \quad \text{for } \eta \geq 1. \end{aligned}$$

From the definition of $m(n)$, we have $m(n) - n = \Theta(n^\kappa)^3$. Specifically, if $a(n) \geq cn^{-\kappa}$ for some $c > 0$, then $a(n), \dots, a(2n) \geq c2^{-\kappa}n^{-\kappa}$. So for large n , $t(n + T2^\kappa c^{-1}n^\kappa) \geq t(n) + T$, implying $m(n) - n \leq T2^\kappa c^{-1}n^\kappa$, which is the desired estimate. Hence $\sum_{i=n}^{m(n)} a(i)^2 = \Theta\left(\frac{1}{n^\kappa}\right)$. Combining these observations, we see that for a constant $C(T) > 0$ depending on T ,

$$\begin{aligned}\hat{\sigma}^2(n, m(n)) &\leq C(T) \left(\frac{n^{\kappa(1-\eta)}}{n^\kappa} \right) \quad \text{for } \eta < 1, \\ \hat{\sigma}^2(n, m(n)) &\leq \frac{C(T)}{n^\kappa} \quad \text{for } \eta \geq 1.\end{aligned}\tag{13}$$

This completes the proof. \square

In fact, the same argument shows a more general fact, which will be used later:

Lemma 3 Let $0 \leq s < t \leq T + 1$, $m_t(n) := \min\{n' \geq n : \sum_{i=n}^{n'} a(i) \geq t\}$, and $m_s(n) := \min\{n' \geq n : \sum_{i=n}^{n'} a(i) \geq s\}$. Then

$$E\left[\left\|\sum_{i=m_s(n)}^{m_t(n)} a(i)R(i)(B(i+1) - B(i))\right\|^2\right] \leq \frac{C(T)}{n^\gamma}$$

for $C(T), \gamma$ as above. Furthermore, $C(T)$ can be chosen such that

$$\lim_{T \downarrow 0} C(T) = 0.$$

Lemma 4 $E[\sup_{n \leq j \leq m(n)} \|\sum_{i=n}^j a(i)R(i)B_{i+1}\|^2] \rightarrow 0$.

Proof We first obtain a bound for

$$E\left[\sup_{n \leq N \leq m(n)} \left\|\sum_{i=n}^N a(i)R(i)B_{i+1}\right\|^2\right]\tag{14}$$

in terms of the foregoing. Consider a continuous time process $X(t), t \in [n, m(n)]$, defined by: $X(n) =$ the zero vector in \mathcal{R}^d and

$$X(t) := \int_n^t \tilde{R}(s)d\tilde{B}(s), \quad n \leq t \leq m(n),$$

³Here and elsewhere, $f_n = \Theta(g_n)$ will hold for the statement: *both* $f_n = O(g_n)$ and $g_n = O(f_n)$ hold simultaneously.

where $\tilde{R}(t) := a(i)R(i)$ for $t \in [i, i + 1)$, $n \leq i < m(n)$. We shall be using a variant of Fernique's inequality from [4], section 10.1 (see appendix for a full statement). For this, define

$$\begin{aligned}\phi(x) &:= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \\ \Psi(x) &:= \int_x^\infty \phi(y) dy, \\ \varphi^n(u) &:= \max_{n \leq s < t \leq m(n): t-s \leq u(m(n)-n)} E[\|X(t) - X(s)\|^2]^{\frac{1}{2}}, \\ K &:= \frac{5}{2} e^2 \sqrt{2\pi}, \\ \Gamma &:= \sqrt{5}, \\ Q^n(u) &:= \varphi^n(u) + (2 + \sqrt{2}) \int_1^\infty \varphi^n(ue^{-y^2}) dy, \quad u > 0.\end{aligned}$$

Both $\varphi^n(u)$ and $Q^n(u)$ increase with u . By the preceding lemma, $\varphi^n(u)$ converges to 0 as $u \rightarrow 0$. Clearly, $Q^n(u)$ does so as well. We shall prove that $Q^n(1)$ is $o(1)$. By Lemma 3 and the definition of $\varphi^n(\cdot)$, we have

$$\varphi^n(u) \leq \frac{C}{n^{\gamma/2}} \quad \text{for } u \leq 1.$$

Thus

$$Q^n(1) \leq C \left(\frac{1}{n^{\gamma/2}} + \int_1^\infty \varphi^n(e^{-y^2}) dy \right).$$

Now, $\varphi^n(u) \leq Cn^{\kappa(\nu-1)}u^\nu$ for $u < a(m(n))/(T+1)$, whereby

$$\varphi^n(e^{-y^2}) < Cn^{\kappa(\nu-1)}e^{-\nu y^2} \tag{15}$$

for

$$y > \sqrt{\log \left(\frac{T+1}{a(m(n))} \right)} := g(n).$$

Using Lemma 3 we have,

$$\begin{aligned}& \int_1^\infty \varphi^n(e^{-y^2}) dy \\ &= \int_1^{g(n)} \varphi^n(e^{-y^2}) dy + \int_{g(n)}^\infty \varphi^n(e^{-y^2}) dy\end{aligned}$$

$$\begin{aligned}
&\leq \frac{Cg(n)}{n^{\gamma/2}} + C \int_{g(n)}^{\infty} n^{\kappa(\nu-1)} e^{-\nu y^2} dy \\
&\leq \frac{Cg(n)}{n^{\gamma/2}} + n^{\kappa(\nu-1)} \frac{C e^{-\nu g(n)^2}}{g(n)} \\
&= \frac{Cg(n)}{n^{\gamma/2}} + n^{\kappa(\nu-1)} \frac{C m(n)^{-\nu\kappa}}{g(n)(T+1)^\nu} \\
&\leq \frac{Cg(n)}{n^{\gamma/2}} + \frac{C}{g(n)n^\kappa}.
\end{aligned}$$

The first inequality follows from Lemma 2 and (15), the third inequality follows from $m(n) = \Theta(n)$. Thus

$$Q^n(1) \leq G(n) := \frac{C}{n^{\gamma/2}} + \frac{Cg(n)}{n^{\gamma/2}} + \frac{C}{g(n)n^\kappa}. \quad (16)$$

Note that $g(n) = O(\sqrt{\log n})$. Thus $G(n) = o(1)$. Let $Z(u) := X(n + u(m(n) - n))$, $u \in [0, 1]$. By (10.1.9) of p. 198, [4], we have

$$P(\max_{u \in [0,1]} \|Z(u)\| > x) \leq dK\Psi\left(\frac{x}{Q^n(1)}\right)$$

for $x \geq \Gamma Q^n(1)$. Hence for $x \geq 0$,

$$\begin{aligned}
P(\max_{t \in [n, m(n)]} \|X(t)\| > x) &= P(\max_{u \in [0,1]} \|Z(u)\| > x) \\
&\leq dK\Psi\left(\frac{x}{Q^n(1)}\right).
\end{aligned}$$

$\forall x > \Gamma Q^n(1)$. Then, for $\delta > 0$,

$$\begin{aligned}
&E[\sup_{t \in [n, m(n)]} \|X(t)\|^2] \\
&= 2 \int_0^\infty x P(\sup_{t \in [n, m(n)]} \|X(t)\| \geq x) dx \\
&\leq 2\delta + 2 \int_\delta^\infty x P(\sup_{t \in [n, m(n)]} \|X(t)\| > x - \delta) dx \\
&\leq 2\delta C + 2 \int_0^\infty x P(\sup_{t \in [n, m(n)]} \|X(t)\| > x) dx \\
&= 2\delta C + 2 \int_0^{\Gamma Q^n(1)} x P(\sup_{t \in [n, m(n)]} \|X(t)\| > x) dx
\end{aligned}$$

$$\begin{aligned}
& + 2 \int_{\Gamma Q^n(1)}^{\infty} x P(\sup_{t \in [n, m(n)]} \|X(t)\| > x) dx \\
& \leq 2\delta C + (\Gamma G(n))^2 + 2Kd \int_{\Gamma Q^n(1)}^{\infty} x \Psi\left(\frac{x}{Q^n(1)}\right) dx \\
& \leq 2\delta C + (\Gamma G(n))^2 + \\
& \quad 2Kd \int_{\Gamma Q^n(1)}^{\infty} x \left(\frac{Q^n(1)}{x}\right) \phi\left(\frac{x}{Q^n(1)}\right) dx \\
& \leq 2\delta C + (\Gamma G(n))^2 + 2KdCG(n) \xrightarrow{n \uparrow \infty} 2\delta C.
\end{aligned}$$

Since $\delta > 0$ was arbitrary,

$$E[\sup_{t \in [n, m(n)]} \|X(t)\|^2] \rightarrow 0,$$

from which the claim follows. \square

Next consider the error term $\sup_{n \leq j \leq m(n)} \|\sum_{i=n}^j a(i)D(i)S_{i+1}\|$.

Lemma 5 $E[\sup_{n \leq j \leq m(n)} \|\sum_{i=n}^j a(i)D(i)S_{i+1}\|^\xi] \rightarrow 0$.

Proof Recall that $\{D(n)\}$ are bounded. By the scaling property of stable processes, $\sum_{i=n}^{m(n)} a(i)D(i)S_{i+1}$ has the same law as

$$\sum_{i=n}^{m(n)} a(i)D(i)a(i)^{-\frac{1}{\alpha}} (\tilde{S}_{\sum_{k=n}^{i+1} a(k)} - \tilde{S}_{\sum_{k=n}^i a(k)}).$$

By Theorem 3.2, p. 65, of [10], we then have

$$P\left(\sup_{n \leq j \leq m(n)} \left\|\sum_{i=n}^j a(i)D(i)S_{i+1}\right\| \geq x\right) \leq \frac{C(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1})^{\frac{\alpha}{\alpha+1}}}{x^\alpha} \quad (17)$$

for $x > C(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1})^{\frac{1}{\alpha+1}}$. Note that

$$\epsilon(n, \alpha) := C\left(\sum_{i=n}^{m(n)} a(i)^{\frac{\alpha^2-1}{\alpha}+1}\right)^{\frac{1}{\alpha+1}} \leq C(T+1)^{\frac{1}{\alpha+1}} a(n)^{\frac{\alpha-1}{\alpha}} \xrightarrow{n \uparrow \infty} 0. \quad (18)$$

Thus for $1 < \xi < \alpha$,

$$E\left[\sup_{n \leq j \leq m(n)} \left\|\sum_{i=n}^j a(i)D(i)S_{i+1}\right\|^\xi\right]$$

$$\begin{aligned}
&\leq C \int_0^\infty x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
&= C \int_0^{\epsilon(n, \alpha)} x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
&\quad + C \int_{\epsilon(n, \alpha)}^\infty x^{\xi-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| \geq x \right) dx \\
&\leq C \epsilon(n, \alpha)^\xi + C \int_{\epsilon(n, \alpha)}^\infty x^{\xi-1} \left(\frac{\epsilon(n, \alpha)^\alpha}{x^\alpha} \wedge 1 \right) dx \\
&\leq C \epsilon(n, \alpha)^\xi + C \int_{\epsilon(n, \alpha)}^\infty x^{\xi-1} \left(\frac{\epsilon(n, \alpha)^\alpha}{x^\alpha} \right) dx \\
&= C \epsilon(n, \alpha)^\xi \\
&\rightarrow 0
\end{aligned}$$

as $n \uparrow \infty$. The claim follows. \square

An alternative proof can be given by using the classical Burkholder-Davis-Gundy inequalities, but we use Joulin's 'concentration' inequality as it paves way for the analysis of finite time behavior of the scheme as in [5], sections 4.1, 4.2. We do not, however, pursue this theme here.

4 Main results

Proof of Theorem 1:

By (†), we have

$$\begin{aligned}
E \left[\left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|x_n\|) \right)^\xi \right] &\leq C \left(\sum_{i=n}^{m(n)} a(i)^2 \right)^\xi \\
&\rightarrow 0 \text{ as } n \uparrow \infty.
\end{aligned} \tag{19}$$

By (2) and the inequality in the 'Remark' on p. 151, [13], we have

$$E \left[\sup_{n \leq k \leq m(n)} \left\| \sum_{i=n}^k a(i) M_{i+1} \right\|^\xi \right]$$

$$\begin{aligned}
&\leq CE\left[\left(\sum_{i=n}^{m(n)} a(i)^2 E[\|M_{i+1}\|^2 | \mathcal{F}_i]\right)^{\frac{\xi}{2}}\right] \\
&\leq CE\left[\left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|x_i\|)^2\right)^{\frac{\xi}{2}}\right] \\
&\leq CE\left[\sum_{i=n}^{m(n)} a(i)^\xi (1 + \|x_i\|)^\xi\right] \\
&\leq Ca(n)^{\xi-1} (T+1)(1+K_2) \\
&\rightarrow 0
\end{aligned} \tag{20}$$

because $a(n)^{\xi-1} \rightarrow 0$ as $n \uparrow \infty$. (The third inequality above follows from the subadditivity of $x^a : \mathcal{R}^+ \rightarrow \mathcal{R}^+$ for $a \in (0, 1)$.) Our conditions on $\{\zeta_n\}$ imply

$$E\left[\sup_{n \leq i \leq m(n)} \|\zeta_i\|^\xi\right] \rightarrow 0. \tag{21}$$

Lemma 4 in particular implies that

$$E\left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^\xi\right] \rightarrow 0. \tag{22}$$

Now take the norm $\|\cdot\|_\xi := E[\|\cdot\|^\xi]^{\frac{1}{\xi}}$ on both sides of (6) and use (19), (20), (21), (22) and Lemma 4 to conclude that

$$\lim_{n \uparrow \infty} E\left[\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^n(t)\|^\xi\right] = 0.$$

With a small additional calculation – see, e.g., [5], Chapter 2, p. 14 – we can improve this to

$$\lim_{n \uparrow \infty} E\left[\sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\|^\xi\right] = 0, \tag{23}$$

where $x^s(\cdot)$ is the solution to (4) on $t \geq s$ with $x^s(s) = \bar{x}(s)$. Let $\epsilon > 0$, $M \gg 0$ (we choose M depending on ϵ later), and pick $T > 0$ such that for any $x(\cdot)$ satisfying (4) with $\|x(0)\| \leq M$, we have $\|x(t) - x^*\| < \frac{\epsilon}{2} \forall t \geq T$. Then for $1 < \xi' < \xi$,

$$\begin{aligned}
&E[\|\bar{x}(s+T) - x^*\|^{\xi'}] \\
&\leq E[\|x^s(s+T) - x^*\|^{\xi'} I\{\|\bar{x}(s)\| \leq M\}] \\
&\quad + E\left[\sup_{t \in [s, s+T]} \|\bar{x}(t) - x^s(t)\|^\xi I\{\|\bar{x}(s)\| \leq M\}\right] \\
&\quad + E[\|\bar{x}(s+T) - x^*\|^{\xi'} I\{\|\bar{x}(s)\| > M\}].
\end{aligned} \tag{24}$$

The first term on the right is $< \frac{\epsilon}{2}$ by our choice of M, T . The second is $< \frac{\epsilon}{4}$ for s large enough, by (23). The third is $< \frac{\epsilon}{4}$ for M large enough because (†) and $\xi' < \xi \implies \|\bar{x}(s+T) - x^*\|^{\xi'}$ is uniformly integrable. Thus the right hand side can be made smaller than any $\epsilon > 0$ for $s+T$ sufficiently large. The claim follows. \square

We show next that the stability test of [8] can be adapted to the present scenario and implies (†) for any $\xi \in (1, \alpha)$, when $E[\|x_0\|^\xi] < \infty$.

Let $h_c(x) := \frac{h(cx)}{c}$ for $c > 0$. We assume as in [8] that

$$h_\infty(x) := \lim_{c \uparrow \infty} h_c(x) \quad (25)$$

exists. Consider the o.d.e.

$$\dot{x}_c(t) = h_c(x_c(t)) \quad (26)$$

for $0 < c \leq \infty$. The key condition of [8] which we adapt here is the following:

(*) For $c = \infty$, (26) has the origin as the globally exponentially stable equilibrium.

This is a stronger condition than the one used in [8], where only global asymptotic stability was needed. See [9] for an interesting perspective on the two notions of stability.

Note that $h_c, c > 0$, are Lipschitz with the same Lipschitz constant as h , therefore equicontinuous. Thus the convergence in (25) is uniform on compacts. Using this, a simple argument based on the Gronwall inequality as in Lemma 2, pp. 23, of [5] shows that for a fixed initial condition, $x_c(\cdot) \rightarrow x_\infty(\cdot)$ uniformly on compacts as $c \uparrow \infty$. Fix $1 < \xi < \alpha$. For $n \geq 0$, define $\bar{x}^n(t), t \geq t(n)$, by:

$$\bar{x}^n(t(m)) = \frac{x_m}{E[\|x_n\|^\xi]^{\frac{1}{\xi}} \vee 1}, \quad m \geq n,$$

with linear interpolation on $[t(m), t(m+1)]$ for $m \geq n$. Also define $\tilde{x}^n(t), t \geq t(n)$, to be the solution to (26) with $c = c(n) := E[\|x_n\|^\xi]^{\frac{1}{\xi}} \vee 1$ and $\tilde{x}^n(t(n)) = \bar{x}^n(t(n))$.

Lemma 6 $\sup_{t \in [t(n), t(n)+T]} E[\|\bar{x}^n(t) - \tilde{x}^n(t)\|^\xi] \rightarrow 0$ as $n \uparrow \infty$.

Proof Follows from (21), Lemmas 1–5, and an application of the Gronwall inequality, exactly as in the proof of Theorem 1. Note that

$$\sup_n \sup_{n \leq m \leq m(n)} E[\|\bar{x}^n(t(m))\|^\xi] < \infty \quad (27)$$

by construction, which replaces (\dagger) in the proof of Theorem 1. \square

Theorem 2 Under above hypotheses, (\dagger) holds.

Proof Let $T > 0$, which we specify later. Suppose there exists a subsequence $\{n(k)\}$ such that

$$E[\|x_{n(k)}\|^\xi] \uparrow \infty.$$

Define $\{T_\ell\}$ by: $T_0 := 0, T_{\ell+1} := \min\{t(m) \geq T_\ell : t(m) - T_\ell \geq T\}, \ell \geq 0$. A standard application of the discrete Gronwall inequality shows that

$$E[\sup_{n \leq i < m(n)} \|x_i\|^\xi] < \tilde{C} E[\|x_n\|^\xi], \quad (28)$$

where the constant \tilde{C} depends on T , but not on n . In particular, if $\{\ell(k)\}$ are such that $T_{\ell(k)} \leq t(n(k)) < T_{\ell(k)+1}$, then we must have

$$\infty \leftarrow \frac{1}{\tilde{C}} E[\|x_{n(k)}\|^\xi] \leq E[\|\bar{x}(T_{\ell(k)})\|^\xi].$$

Thus we have

$$E[\|\bar{x}(T_{\ell(k)})\|^\xi] \rightarrow \infty. \quad (29)$$

Pick $T > 0$ such that $\|x_\infty(t)\| < \frac{1}{8}\|x_\infty(0)\|$ for $t \geq T$. This is possible by $(*)$. Then there exists $c_0 > 1$ such that:

$$\|x_c(t)\| < \frac{1}{4}\|x_c(0)\| \quad \forall t \in [T, T+1] \quad \text{when } c \geq c_0. \quad (30)$$

By (29), we may assume without any loss of generality that $E[\|\bar{x}(T_{\ell(k)})\|^\xi]^{\frac{1}{\xi}} > c_0 \forall k$. Let $n^*(\ell)$ be defined by: $T_\ell = t(n^*(\ell))$. Note that $n^*(\ell+1) = m(n^*(\ell))$. By Lemma 6, we have for any $\frac{1}{4} > \epsilon > 0$ and k sufficiently large,

$$E[\|\bar{x}^{n^*(\ell(k))}(T_{\ell(k)+1})\|^\xi]^{\frac{1}{\xi}}$$

$$\begin{aligned}
&\leq E[\|\tilde{x}^{n^*(\ell(k))}(T_{\ell(k)+1})\|^\xi]^{\frac{1}{\xi}} + \epsilon \\
&\leq \frac{1}{4}E[\|\tilde{x}^{n^*(\ell(k))}(T_{\ell(k)})\|^\xi]^{\frac{1}{\xi}} + \frac{1}{4} \\
&= \frac{1}{4}E[\|\bar{x}^{n^*(\ell(k))}(T_{\ell(k)})\|^\xi]^{\frac{1}{\xi}} + \frac{1}{4} \\
&= \frac{1}{2}.
\end{aligned}$$

Here the first inequality follows from Lemma 6 and the second from (30) and our choice of ϵ . The first equality follows from the equality of $\bar{x}^n(t(n))$ and $\tilde{x}^n(t(n))$, and the second from the fact that once $E[\|\bar{x}(T_\ell)\|^\xi]^{\frac{1}{\xi}} \geq 1$, $E[\|\bar{x}^{n^*(\ell)}(T_\ell)\|^\xi]^{\frac{1}{\xi}} = 1$. Thus

$$E[\|\bar{x}(T_{\ell(k)+1})\|^\xi] \leq \frac{1}{2}E[\|\bar{x}(T_{\ell(k)})\|^\xi],$$

i.e.,

$$E[\|x_{n^*(\ell(k)+1)}\|^\xi] \leq \frac{1}{2}E[\|x_{n^*(\ell(k))}\|^\xi].$$

Hence for n sufficiently large, if $E[\|\bar{x}(T_n)\|^\xi] > c_0$, then $E[\|\bar{x}(T_k)\|^\xi], k \geq n$, falls back to c_0 at an exponential rate. Therefore for such n , $E[\|\bar{x}(T_{n-1})\|^\xi]$ is either even larger than $E[\|\bar{x}(T_n)\|^\xi]$, or is $\leq c_0$. Hence there is a subsequence along which $E[\|\bar{x}(T_n)\|^\xi]$ jumps from a value $\leq c_0$ to one that is increasing to ∞ . This contradicts (28), implying that $\sup_n E[\|x_n\|^\xi] < \infty$. \square

Remark: As observed in [5], Chapter 3, there is no universal scheme for establishing boundedness of iterates even in the classical set-up (where the desired boundedness is ‘a.s.’ as opposed to ‘in ξ th mean’ here). What one has is a family of tests, each with its own domain of utility. One expects the same here, i.e., the above is but one way to ensure (\dagger) , not necessarily the only or the ‘best’ one. Alternatively, one can project iterates to a large but compact convex set A to keep them bounded by design. The set A should be large enough to contain the desired equilibrium x^* , which implies an a priori judgement about $\|x^*\|$. The limiting o.d.e. is the projected dynamics corresponding to (4), with a correction term at the boundary ∂A of A that forces it to remain inside A . If h is transversal to ∂A at all points and points inwards, this correction term is zero and the foregoing goes through. If not, one has to allow for spurious equilibria or other attractors in ∂A created by the projection operation. One can also ‘grow’ A very slowly to the whole

space. See [5], section 5.4, for a discussion of these issues.

5 Almost sure convergence in the absence of the heavy tailed noise

If $D(n) \equiv 0 \forall n$ in the above (i.e., the noise is ‘light tailed’ albeit long range dependent) and (*) holds, we can improve the conclusions of Theorem 1 to ‘ $x_n \rightarrow x^*$ a.s.’ To see this, proceed as follows: Let $E[\|x_0\|^2] < \infty$.

- We have $\sup_n E[\|x_n\|^2] < \infty$ by arguments analogous to the ones leading to Theorem 2 above. Then in particular, by (2),

$$\begin{aligned} \sum_n a(n)^2 E[\|x_n\|^2] < \infty &\implies \sum_n a(n)^2 \|x_n\|^2 < \infty \text{ a.s.} \\ &\implies \sum_n a(n)^2 E[\|M_{n+1}\|^2 | \mathcal{F}_n] < \infty \text{ a.s.} \\ &\implies \sum_n a(n) M_{n+1} \text{ converges, a.s.} \end{aligned}$$

The last implication follows from Proposition VII-2-3(c) of [13], p. 149. Thus $\sup_{n \leq k \leq m(n)} \|\sum_{i=n}^k a(i) M_{i+1}\|^2 \rightarrow 0$ a.s.

- $\zeta_n \rightarrow 0$ a.s. implies that $\sup_{n \leq k \leq m(n)} \sum_{i=n}^k a(i) \|\zeta_i\| \rightarrow 0$ a.s.
- Since for $m \geq 1$,

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] = m \int_0^\infty x^{m-1} P \left(\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \geq x \right) dx,$$

we may argue as in the proof of Lemma 4 to obtain, for $0 < \delta < 1$,

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] \leq C(\delta + G(n)^m).$$

For n large, choose $\delta = G(n)^m$ to get

$$E \left[\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\|^m \right] \leq CG(n)^m. \quad (31)$$

Since $G(n) = O(n^{-a'})$ for some $a > 0$, $G(n)^m = O(n^{-ma})$. Pick m such that $ma > 1$. A standard argument using the Borel-Cantelli lemma then yields

$$\sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i)R(i)B_{i+1} \right\| \rightarrow 0 \text{ a.s.}$$

- In view of the foregoing, the arguments of [8] go through to conclude that $\sup_n \|x_n\| < \infty$ a.s. (in fact, as in [8], it suffices to have ‘asymptotic stability’ in place of ‘exponential stability’ in (*), since initial conditions of the o.d.e. trajectories $x^n(\cdot)$ above can be taken to lie in a possibly sample path dependent compact set.), whence $(\sum_{i=n}^{m(n)} a(i)^2)(1 + \|x_n\|) \rightarrow 0$ a.s.

The claim then follows from Lemma 1.

We summarize the above as:

Theorem 3 If $D(n) \equiv 0 \forall n$ and

$$\sup_n \|x_n\| < \infty \text{ a.s.}, \tag{32}$$

then $x_n \rightarrow x^*$ a.s. Also, (32) holds if the origin is the globally asymptotically stable equilibrium for (26).

6 Constant stepsize schemes

Very often, e.g., in tracking algorithms with a slowly varying environment, it is more convenient to use a constant small stepsize $a(n) \equiv a > 0$. Then, as pointed out in section 9.1 of [5], one cannot expect a.s. convergence to x^* , but only an asymptotic concentration of probability in a neighborhood of x^* . Mimicking the steps on section 9.2, [5], we set $t(n) = na, n \geq 0$, and take $T > 0$ of the form $T = Na$ for some $N \geq 1$. Assume (†). Then setting $a(i) \equiv a$ in sections 2 – 4, we have the following:

- 1.

$$E\left[\left(\sum_{i=nN}^{(n+1)N} a^2(1 + \|x_{nN}\|)\right)^\xi\right]^{\frac{1}{\xi}}$$

$$\leq C \left(\sum_{i=nN}^{(n+1)N} a^2 \right) = Ca. \quad (33)$$

2. As in (20),

$$\begin{aligned} & E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aM_{i+1} \right\|^\xi \right]^{\frac{1}{\xi}} \\ & \leq CE \left[\left(\sum_{i=nN}^{(n+1)N} a^2 (1 + \|x_i\|^2) \right)^{\frac{\xi}{2}} \right]^{\frac{1}{\xi}} \\ & \leq Ca^{\frac{\xi-1}{\xi}} \end{aligned} \quad (34)$$

3. We closely follow the steps for Lemmas 2 – 4 in section 3 with $a(n) \equiv a$. We then have the following counterpart of (13):

$$\hat{\sigma}^2(nN, (n+1)N) \leq Ca^{\eta \wedge 1}.$$

where we use $N = \frac{T}{a}$. A similar analog of Lemma 3 holds. Mimicking the proof of Lemma 4 then leads to

$$\begin{aligned} & E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aR(i)B_{i+1} \right\|^\xi \right]^{\frac{1}{\xi}} \\ & \leq E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aR(i)B_{i+1} \right\|^2 \right]^{\frac{1}{2}} \\ & \leq Ca^{\eta \wedge 1/2} + Ca^{\frac{\eta \wedge 1}{2}} \sqrt{\log((T+1)/a)} + \frac{Ca}{\sqrt{\log((T+1)/a)}}. \end{aligned} \quad (35)$$

4. Using (17) as before,

$$\begin{aligned} & E \left[\sup_{nN \leq k \leq (n+1)N} \left\| \sum_{i=nN}^k aD(i)S_{i+1} \right\|^\xi \right]^{\frac{1}{\xi}} \\ & \leq Ca^{\frac{\alpha-1}{\alpha}}. \end{aligned} \quad (36)$$

Let $\chi := \min(\frac{\xi-1}{\xi}, \frac{\eta \wedge 1}{2} - \epsilon)$ (since $\frac{\alpha-1}{\alpha} > \frac{\xi-1}{\xi}$), where $\epsilon > 0$ may be chosen to be arbitrarily small. Then (33) – (36) combined with Lemma 1 yields

$$E \left[\sup_{t \in [nN, (n+1)N]} \|\bar{x}(t) - x^{nN}(t)\|^\xi \right]^{\frac{1}{\xi}} \leq Ca^\chi. \quad (37)$$

Now fix $1 < \xi' < \xi$. Pick $M \gg 1$ such that for any $t > s > 0$,

$$\sup_{t \in [nN, (n+1)N]} E[\|\bar{x}(t) - x^*\|^{\xi'} I\{\|\bar{x}(s)\| > M\}] < a^\chi.$$

This is possible by (\dagger) and the ensuing uniform integrability of $\{\bar{x}(t), \|\bar{x}(t) - x^*\|^{\xi'}, t \geq 0\}$. Now pick $T = Na > 0$ such that for $x(\cdot)$ satisfying (4), $\|x(0)\| \leq M \implies \|x(t) - x^*\| < a^\chi \forall t \geq T$. Then by (24) and the foregoing,

$$\limsup_{n \uparrow \infty} E[\|x_n - x^*\|^{\xi'}]^{\frac{1}{\xi'}} \leq Ca^\chi,$$

which gives a quantitative measure of asymptotic concentration of the iterates around x^* .

Recall our hypothesis $\sup_n \|\zeta_n\| \leq K_0$. If $K_0 = O(a^\chi)$, the foregoing continues to hold even without the requirement that $\zeta_n \rightarrow 0$ a.s.

That $(*)$ implies (\dagger) follows as before for the constant stepsize case (see, e.g., pp. 110-111 of [5], also [8]).

7 Miscellaneous remarks

Many of the variations on the basic convergence theory of stochastic approximations in the classical set-up of [5] have their counterparts in the present framework. We sketch some of them in outline, pointing to [5] for greater detail while making them reasonably self-contained.

1. *General limit sets:* One important observation is that in the more general case when there exists a C^1 Liapunov function V satisfying the conditions $\lim_{\|x\| \uparrow \infty} V(x) = \infty$ and $\langle \nabla V(x), h(x) \rangle \leq 0$, a similar argument shows that $x_n \rightarrow \{x : \langle \nabla V(x), h(x) \rangle = 0\}$ in the ξ' th mean, $\xi' < \xi$.
2. *Markov noise:* Suppose we replace the term $h(x_n)$ on the r.h.s. of (1) by $h(x_n, Y_n)$ where $\{Y_n\}$ is a process taking values in a finite state

space⁴ S with $|S| = s$, and satisfying:

$$P(Y_{n+1} = i | \mathcal{F}_n, Y_m, m \leq n) = q_{x_n}(i | Y_n) \quad \forall n \geq 0,$$

where $q_x(\cdot | \cdot)$ is a transition probability on S smoothly parametrized by x . W.l.o.g., let $S = \{1, \dots, s\}$. Thus if $x_n \equiv x \quad \forall n$, $\{Y_n\}$ would be a Markov chain, hence the appellation ‘Markov noise’. We assume that for each x , the corresponding Markov chain is irreducible and thus has a unique stationary distribution $m_x(i), i \in S$. Then the asymptotic o.d.e. is

$$\dot{x}(t) = \sum_i m_{x(t)}(i) h(x(t), i). \quad (38)$$

With this replacing (4), the theory is similar to the above. To see this, define the process $\mu(t) = [\mu_1(t), \dots, \mu_s(t)], t \geq 0$, taking values in $\mathcal{P}(S) :=$ the probability simplex on S , as follows: $\mu_i(t) := \delta_{Y_n i}, 1 \leq i \leq s, t \in [t(n), t(n+1))$, where δ_{jk} is the Kronecker delta. Then

$$\sum_i \mu_i(t) h(\bar{x}(t), i) = h(x_n, Y_n), \quad t \in [t(n), t(n+1)).$$

Mimicking the arguments above, this suggests that we consider $\tilde{x}^n(t), t \geq t(n)$, the solution to the o.d.e.

$$\dot{\tilde{x}}^n(t) = \sum_i \mu_i(t) h(\tilde{x}^n(t), i), \quad (39)$$

with $\tilde{x}^n(t(n)) = \bar{x}(t(n))$. Let $x^n(t), t \geq t(n)$, be the solution to the expected ‘limiting o.d.e.’ (38) with $x^n(t(n)) = \bar{x}(t(n))$. Let $1 < \xi < \alpha$. Assume that $\sup_n E[\|x_n\|^\xi] < \infty$, which by familiar Gronwall-based arguments yields

$$\sup_n E\left[\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t)\|^\xi \right] < \infty. \quad (40)$$

Argue as in the preceding sections to claim that for $\xi' < \xi$,

$$E\left[\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - \tilde{x}^n(t)\|^{\xi'} \right] \rightarrow 0. \quad (41)$$

We now mimic the arguments of section 6.3 of [5], pp. 73-74. Consider $\mu(\cdot) = [\mu_1(\cdot), \dots, \mu_s(\cdot)]$ restricted to $[0, T], T > 0$, as an element

⁴Extension to more general state spaces is possible – see [5].

of $\mathcal{U}_T := (L_2[0, T])^s$ with the weak* topology and $\mu(\cdot)$ itself as an element of the space $\mathcal{U} := \{u(\cdot) : u(\cdot)|_{[0, T]} \in \mathcal{U}_T \ \forall T > 0\}$ with the inductive topology. It is easy to see that this is a compact metrizable space. A simple application of the Arzela-Ascoli theorem shows that $\tilde{x}^n(t(n) + \cdot), n \geq 1$, is a relatively compact sequence in $C([0, \infty); \mathcal{R}^d)$. Let $(x'(\cdot), \mu'(\cdot))$ denote a limit point of $(\tilde{x}^n(t(n) + \cdot), \mu(t(n) + \cdot))$ in $C([0, \infty); \mathcal{R}^d) \times \mathcal{U}$ as $n \uparrow \infty$. Henceforth we consider this subsequence, denoted by $\{n\}$ again by abuse of notation. Define for $1 \leq j \leq s$,

$$Z_n^j := \sum_{m=0}^n a(m)(I\{Y_{m+1} = j\} - p(j|x_m, Y_m)), \quad n \geq 0.$$

This is a square-integrable martingale with $\sup_n E[\|Z_n^j\|^2] \leq C \sum_n a(n)^2 < \infty$. Hence it converges a.s., implying in particular that for $t > s \geq 0$,

$$\sum_{k=m_s(n)}^{m_t(n)} a(k)(I\{Y_{k+1} = j\} - p(j|x_k, Y_k)) \rightarrow 0$$

a.s.. Dividing by $\sum_{k=m_s(n)}^{m_t(n)} a(k) \approx t - s$ and letting $n \uparrow \infty$, we get

$$\int_s^t \sum_k (I\{k = j\} - p(j|x'(r), k)) \mu'_k(r) dr = 0.$$

By Lebesgue's theorem,

$$\sum_k (I\{k = j\} - p(j|x'(r), k)) \mu'_k(r) = 0$$

for a.e. r , where the qualification 'a.e.' may be dropped by choosing a suitable version. It follows that $\mu'(t) = \pi_{x'(t)} \forall t$. Passing to the limit as $n \uparrow \infty$ in (39), we get (38) with $x'(\cdot)$ replacing $x(\cdot)$. It follows that

$$\sup_{t \in [t(n), t(n)+T]} \|x^n(t) - \tilde{x}^n(t)\|^{\xi'} \rightarrow 0 \quad \text{a.s.}$$

By (40) and the dominated convergence theorem, we have

$$E\left[\sup_{t \in [t(n), t(n)+T]} \|x^n(t) - \tilde{x}^n(t)\|^{\xi'} \right] \rightarrow 0. \quad (42)$$

Combining (41) and (42), we have

$$E\left[\sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - \tilde{x}^n(t)\|^{\xi'} \right] \rightarrow 0.$$

The rest follows as before.

See [5], sections 6.2-6.3 for a detailed treatment of the classical case, which in particular serves as a pointer to some extensions (among them, a more general state space and an additional ‘control’ process).

3. *Asynchronous schemes:* One often has to consider situations where different components of (1) are computed by different processors, possibly not all at the same time, and with different local clocks, with the results being transmitted to each other with random transmission delays. Thus let $Y_n := \{j \in \{1, \dots, d\} : j\text{th component is updated at time } n\}$, possibly random. Also, let $\tau_{ij}(n)$ denote the bounded random delay with which the value of j th component has been received at processor i at time n . In other words, at time n the i th processor knows $x_{n-\tau_{ij}(n)}(j)$ but not $x_m(j)$ for $m > n - \tau_{ij}(n)$. Let $\nu(i, n) = \sum_{m=0}^n I\{i \in Y_m\}$ denote the number of times the i th component got updated till time n , i.e., the ‘local clock’ at processor i . One then replaces the stepsize $a(n)$ in the i th component of (1) by $a(\nu(i, n))I\{i \in Y_n\}$ and $h_i(x_n)$ by

$$h_i(x_{n-\tau_{i1}(n)}(1), \dots, x_{n-\tau_{id}(n)}(d)).$$

Assume that the iterates are bounded a.s. As in [5], Chapter 7, the conclusion is that the limiting o.d.e. (4) gets replaced by

$$\dot{x}(t) = \Lambda(t)h(x(t)) \tag{43}$$

where $\Lambda(t)$ for each t is a diagonal matrix with nonnegative diagonal entries. Intuitively, this reflects the differing rates at which the different components are getting updated. The manner in which this factor arises is as follows. For simplicity, assume a common clock for all processors. Recall that (1) is an iteration in \mathcal{R}^d . Let $\mu'(t) = [\mu'_1(t), \dots, \mu'_d(t)]$ denote a process taking values in $\{0, 1\}^d$ and defined by: $\mu'_i(t) := I\{i \in Y_n\}$ for $t \in [t(n), t(n+1))$. Then $\Lambda(t), t \in [0, T]$, arises as a weak* limit point of $\text{diag}(\mu(t(n) + t)), t \in [0, T]$, as $n \uparrow \infty$. The effect of different clocks can also be absorbed in this analysis. As for the delays, as long as they are bounded (this can be relaxed to some extent), their effect on the asymptotics of the algorithm can be ignored. This is because their net effect is to contribute in (6) yet another error

term of the order

$$a(\nu(i, n)) \sum_j |x_n(j) - x_{n-\tau_{ji}(n)}(j)|.$$

The j th summand can be bounded by

$$\begin{aligned} & \left| \sum_{m=n-\tau_{ji}(n)}^n a(\nu(i, n)) I\{i \in Y_n\} h_i(x_{n-\tau_{i1}(n)}(1), \dots, x_{n-\tau_{id}(n)}(d)) \right| \\ & + \left| \sum_{m=n-\tau_{ji}(n)}^n a(\nu(i, m)) M_{m+1}(i) \right| \\ & + \left| \sum_{m=n-\tau_{ji}(n)}^n a(\nu(i, n)) (R(m) B_{m+1})(i) \right| \\ & + \left| \sum_{m=n-\tau_{ji}(n)}^n a(\nu(i, n)) (D(m) S_{m+1})(i) \right| \\ & + \left| \sum_{m=n-\tau_{ji}(n)}^n a(\nu(i, n)) \zeta_{m+1}(i) \right|, \end{aligned}$$

where the notation is self-explanatory. The first term is bounded by $a(\nu(i, n - M))KM$ where $M > 0$ is any bound on $\tau_{k\ell}(j)$, $1 \leq k, \ell \leq d, j \geq 0$, and $K > 0$ is any (possibly random) bound on $|h_i(x_{k-\tau_{j1}(k)}(1), \dots, x_{k-\tau_{jd}(k)}(d))|$, $1 \leq j \leq d, k \geq 0$. Note that such a bound exists a.s. by our hypothesis of bounded iterates. It follows that this term goes to zero as $n \uparrow \infty$ a.s.. The remaining terms except the penultimate go to zero a.s. as well by familiar arguments, the penultimate one does so in ξ 'th mean, again by familiar arguments.

Intuitively, the time scaling $n \rightarrow t(n)$ asymptotically ‘squeezes out’ time intervals of any given width and therefore the error contributed by delays is asymptotically negligible. The implications of (43) to convergence of the algorithm are discussed in *ibid*. In particular, it does not affect the convergence behavior of (1) for a few important special cases such as gradient schemes and fixed point seekers for max-norm contractions, as long as the diagonal terms in $\Lambda(t)$ remain bounded away from zero. However, it may affect the rate of convergence. See [5], Chapter 7 for more details and possible generalizations (in particular, a possible relaxation on the boundedness hypothesis on delays).

Appendix: Fernique's inequality

Let $I = [0, 1]$ and $(X_t, t \in I)$ a zero mean scalar Gaussian process. Define for $h > 0$,

$$\varphi(h) = \max_{\|t-s\| \leq h, s, t \in I} E[|X_t - X_s|^2]^{\frac{1}{2}}.$$

Assume $\lim_{h \downarrow 0} \varphi(h) = 0$, so that X is stochastically continuous. Let $k \geq 2$ and define

$$\begin{aligned} K &:= \frac{5}{2} k^2 \sqrt{2\pi}, \\ \gamma &:= \sqrt{1 + 4 \log(k)}. \end{aligned}$$

Then Fernique's inequality says that for any interval $J \subset I$ of width at most $h > 0$,

$$P\left(\max_{t \in J} |X_t| \geq x \left[\max_{t \in J} E[X_t^2]^{\frac{1}{2}} + (2 + \sqrt{2}) \int_1^\infty \varphi(hk^{-y^2}) dy \right]\right) \leq K \Psi(x) \quad (44)$$

$\forall x \geq \gamma$, where $\Psi(x) := (2\pi)^{-\frac{1}{2}} \int_x^\infty e^{-\frac{y^2}{2}} dy$ as usual. The consequence of important to us is the following ((10.1.9) of [4]): For

$$Q(t) := \varphi(t) + (2 + \sqrt{2}) \int_1^\infty \varphi(tk^{-y^2}) dy,$$

J as above, and $t_0 \in J$,

$$P\left(\max_{t \in J} |X_t - X_{t_0}| > x\right) \leq K \Psi\left(\frac{x}{Q(h)}\right).$$

See pp. 197–198 of [4].

References

- [1] ABOUNADI, J., BERTSEKAS, D. P., AND BORKAR, V. S. (2003) ‘Stochastic approximation for nonexpansive maps: application to Q-learning algorithms’, *SIAM Journal of Control and Optimization* 41(1), pp. 1-22.

- [2] BENAÏM, M. (1996) ‘Dynamics of stochastic approximation’, in *Le Séminaire de Probabilités*, J. Azema and M. Emery, M. Ledoux and M. Yor (eds.), Springer Lecture Notes in Mathematics No. 1709, Springer Verlag, Berlin - Heidelberg, 1–68.
- [3] BENVENISTE, A., METIVIER, M., AND PRIOURET, P. (1990) *Adaptive Algorithms and Stochastic Approximation*, Springer Verlag, Berlin - New York.
- [4] BERMAN, S. M. (1992) *Sojourns and Extremes of Stochastic Processes*, Wadsworth and Brooks / Cole, Belmont, CA.
- [5] BORKAR, V. S. (2008) *Stochastic Approximation: A Dynamical Systems Viewpoint*, Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK.
- [6] BORKAR, V. S. (2007) ‘Some examples of stochastic approximation in communications’, in ‘*Network Control and Optimization*’, T. Chahed and B. Tuffin (eds.), Lecture Notes in Computer Science 4465, pp. 150–157.
- [7] BORKAR, V. S., AND KUMAR, P. R. (2003) ‘Dynamic Cesaro-Wardrop equilibration in networks’, *IEEE Transactions on Automatic Control* 48(3), pp. 382–396.
- [8] BORKAR, V. S., AND MEYN, S. P. (2000) ‘The ODE method for convergence of stochastic approximation and reinforcement learning’, *SIAM Journal of Control and Optimization* 38(2), pp. 447-469.
- [9] GRÜNE, L., SONTAG, E. D., AND WIRTH, F. R. (1999) ‘Asymptotic stability equals exponential stability, and ISS equals finite energy gain – if you twist your eyes’, *Systems and Control Letters* 38, pp. 127-134.
- [10] JOULIN, A. (2007) ‘On maximal inequalities for stable stochastic integrals’, *Potential Analysis* 26, pp. 57-78.
- [11] MIKOSCH, T., RESNICK, S., ROOTZEN, H., AND STEGEMAN, A. (2002) ‘Is network traffic approximated by stable Lévy motion or fractional Brownian motion?’, *The Annals of Applied Probability* 12(1), pp. 23–68.

- [12] KUSHNER, H. J., AND YIN, G. (2003) *Stochastic Approximation and Recursive Algorithms and Applications* (2nd ed.), Springer Verlag, New York.
- [13] NEVEU, J. (1975) *Discrete-Parameter Martingales*, North-Holland, Amsterdam.
- [14] SARKAR, S., AND TASSIULAS, L. (2002) ‘A framework for routing and congestion control for multicast information flows’, *IEEE Transactions on Information Theory* 48(10), pp. 2690-2708.
- [15] WILSON, F. W., Jr. (1969) ‘Smoothing derivatives of functions and applications’, *Transactions of the American Mathematical Society* 139, pp. 413-428.
- [16] ZHANG, J., ZHENG, D., AND CHIANG, M. (2008) ‘The impact of stochastic noisy feedback on distributed network utility maximization’, *IEEE Transactions on Information Theory* 54(2), pp. 645–665.