

Stochastic Approximation with ‘Bad’ Noise

V. Anantharam

Department of Electrical Engg. and Computer Science
University of California
Berkeley, CA 94720, USA
Email: ananth@eecs.berkeley.edu

Vivek S. Borkar

School of Technology and Computer Science
Tata Institute of Fundamental Research
Homi Bhabha Road, Mumbai 00005, India
Email: borkar.vs@gmail.com

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.

I. INTRODUCTION

Stochastic approximation schemes arise in many communications applications, particularly in those related to resource allocation and estimation, and take the form of gradient seekers, saddle point seekers, fixed point seekers, etc. – see [1] for examples. It is well known that the noise processes in the Internet and several other situations arising in communications are long range dependent and heavy tailed, a fact which has also been theoretically justified through limit theorems such as [6]. The presence of long range dependent and heavy tailed noise puts these stochastic approximation schemes outside the ambit of conventional convergence analysis. Motivated by this, we extend this analysis to cover such noise. In (1) below, these aspects are captured by a (sampled) fractional Brownian motion $\{B_n\}$ and a stable process $\{S_n\}$, resp. That this does introduce significant additional complications for stochastic approximation schemes is reflected in the fact that the convergence claim in (4) below 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 (†) below. In particular, it is *not* ‘a.s.’ as is usually the case [3]. (We can, however, improve the claim to ‘a.s.’ if the heavy tailed component is missing.) We follow the ‘o.d.e.’ approach to the analysis of stochastic approximation, see, e.g., [3] for the classical version. The idea is to treat (1) as a noisy discretization of (3) and then argue that the errors due to both discretization and noise become asymptotically negligible in ξ' -th mean under the stated hypotheses.

II. THE MODEL

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

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

where

- $h : \mathcal{R}^d \rightarrow \mathcal{R}^d$ 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

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

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

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

which are standard conditions for stochastic approximation. Clearly $\sup_n a(n) < \infty$.

Consider the related o.d.e.:

$$\dot{x}(t) = h(x(t)). \quad (3)$$

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) $V : \mathcal{R}^d \rightarrow \mathcal{R}^+$ 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^* . 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 $1 < \xi' < \xi$,

$$E[\|x_n - x^*\|^{\xi'}] \rightarrow 0. \quad (4)$$

Here (\dagger) is a ‘stability of iterates’ condition. A sufficient condition for (\dagger) is discussed later.

III. SKETCH OF THE PROOF

The above result is proved through a sequence of lemmas, listed below. Details appear in [1]. The o.d.e. approach is based on comparing with trajectories of (3) 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 (3) on $[t(n), \infty)$ with $x^n(t(n)) = \bar{x}(t(n)) := x_n$. Fix $T > 0$ and for $n \geq 0$, let

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

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

$$\begin{aligned} & \sup_{t \in [t(n), t(n)+T]} \|\bar{x}(t) - x^n(t)\| \\ & \leq K(T) \left(\sum_{i=n}^{m(n)} a(i)^2 (1 + \|\bar{x}(t(n))\|) \right. \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) \zeta_{i+1} \right\| \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) M_{i+1} \right\| \\ & \quad + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) R(i) B_{i+1} \right\| \\ & \quad \left. + \sup_{n \leq j \leq m(n)} \left\| \sum_{i=n}^j a(i) D(i) S_{i+1} \right\| + a(n) \right). \quad (5) \end{aligned}$$

This is a straightforward consequence of the discrete Gronwall inequality. The rest of the exercise consists of establishing that each term on the right hand side is asymptotically negligible. Of these, the last is trivially so in view of (2), whereas the first can be handled easily in view of (\dagger) and (2). Likewise, the second term is an immediate consequence of our hypothesis regarding $\{\zeta_n\}$. The third uses (\dagger) and the martingale convergence theorem. The treatment of these terms is exactly as in the classical set-up (see, e.g., [3], Chapter 2). For the fourth term, we have:

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

This uses the properties of fractional Brownian motion and an inequality of Fernique ([2], section 10.1). For the fifth term, we use the results of Joulin [5] to establish:

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

Together, these lead to Theorem 1 by standard arguments (see, e.g., [3], Chapter 2).

The stability test of [4] can be adapted to the present scenario and implies (\dagger) 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 [4] that

$$h_\infty(x) := \lim_{c \uparrow \infty} h_c(x) \quad (6)$$

exists. Consider the o.d.e.

$$\dot{x}_c(t) = h_c(x_c(t)) \quad (7)$$

for $0 < c \leq \infty$. The key condition of [4] which we adapt here in a stronger form is the following:

(*) For $c = \infty$, (7) has the origin as the globally exponentially stable equilibrium.

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

If there is no heavy tailed noise, we have the stronger claim:

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

$$\sup_n \|x_n\| < \infty \text{ a.s.}, \quad (8)$$

then $x_n \rightarrow x^*$ a.s. Also, (8) holds if the origin is the globally asymptotically stable equilibrium for (7).

IV. VARIANTS

Some further variants of these results can be established:

1) *Constant stepsize*: For $a(n) \equiv a > 0$, we have the weaker claim

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

for a suitable $\chi > 0$.

2) *General limit sets*: A more general version of Theorem 1 that covers situations more general than unique asymptotically stable equilibrium is: $x_n \rightarrow \{x : \langle V(x), h(x) \rangle = 0\}$ in the ξ' th mean.

3) *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 space¹ S with $|S| = s$, and satisfying:

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

where $q_x(\cdot | \cdot)$ is a transition probability on S smoothly parametrized by x . Thus if $x_n \equiv x \forall n$, $\{Y_n\}$ 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

¹Extension to more general state spaces is possible – see [3], Chapter 6.

distribution $m_x(i), i \in S$. Then the asymptotic o.d.e. is

$$\dot{x}(t) = \sum_i m_{x(t)}(i)h(x(t), i). \quad (9)$$

With (9) replacing (3), the theory is similar to the above.

- 4) *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. See [3], Chapter 7, for a full description of this situation for the classical framework, including the additional technical conditions needed. As in [3], the conclusion is that the limiting o.d.e. (3) gets replaced by

$$\dot{x}(t) = \Lambda(t)h(x(t)) \quad (10)$$

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. Chapter 7 of [3] describes instances where (10) is adequate for convergence analysis, and also discusses ways of modifying step-sizes to work around this problem.

ACKNOWLEDGMENT

The work of V. Anantharam was 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. The work of V. S. Borkar was supported 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.

REFERENCES

- [1] V. Anantharam and V. S. Borkar, “Stochastic approximation with long range dependent and heavy tailed noise”, *submitted*, 2009.
- [2] S. M. Berman, *Sojourns and Extremes of Stochastic Processes*, Belmont, CA: Wadsworth and Brooks / Cole, 1992.
- [3] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, New Delhi: Hindustan Publishing Agency, and Cambridge, UK: Cambridge University Press, 2008.
- [4] V. S. Borkar and S. P. Meyn, “The ODE method for convergence of stochastic approximation and reinforcement learning”, *SIAM Journal of Control and Optimization* vol. 38 no. 2, pp. 447–469, 2000.
- [5] A. Joulin, ‘On maximal inequalities for stable stochastic integrals’, *Potential Analysis* vol. 26 no. 1, pp. 57–78, 2007.
- [6] T. Mikosch, S. Resnick, H. Rootzen and A. Stegeman, “Is network traffic approximated by stable Lévy motion or fractional Brownian motion?”, *The Annals of Applied Probability* vol. 12 no. 1, pp. 23–68, 2002.
- [7] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pt. I, pp. 379–423, 1948; pt. II, pp. 623–656, 1948.