

Managing interprocessor delays in distributed recursive algorithms

V S BORKAR and V V PHANSALKAR

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India

Abstract. For a class of distributed recursive algorithms, it is shown that a stochastic approximation-like tapering stepsize routine suppresses the effects of interprocessor delays.

Keywords. Distributed algorithms; tapering stepsize; stochastic approximation; strict Liapunov systems.

1. Introduction

In the distributed implementation of recursive algorithms, one often has to contend with the problem of interprocessor delays. There have been several attempts to model and analyse such phenomena mathematically, the earliest perhaps being the work of Chazan & Miranker (1969) on linear iterations. A comprehensive survey of the work up to late eighties, along with an extensive bibliography, can be found in (Bertsekas & Tsitsiklis 1989), chapters 6 and 7. In these references, one observes that the distributed algorithms are shown to work as desired only under rather strong conditions, such as boundedness of delays and other algorithm-specific restrictions. On a different note, the recent work of Gharavi & Anantharam (1992) analyses the complex behaviour that can result under stationary delays even in linear iterations.

The aforementioned works use a constant stepsize for deterministic algorithms. A tapering stepsize as in stochastic approximation theory by Benveniste *et al* (1990) is used only when there is a random 'noise' component and the aim there is to suppress the effect of noise on the asymptotic behavior of the algorithm. We argue here that the same scheme also suppresses the effect of delays. The main strength of our analysis is that we allow for very general delays, possibly unbounded, correlated and nonstationary, as long as a mild 'moment' condition is satisfied. Another important gain is that our analysis is not very algorithm-specific. It applies simultaneously to a large class of algorithms, viz., those whose continuous limit is a so called 'strict Liapunov system'. The latter subsumes a large number of gradient-like algorithms currently in vogue among the neural networks community (Hirsch 1987).

The paper is organized as follows. The next section describes the mathematical formulation of the problem and the assumptions. It also defines a strict Liapunov system and gives several examples of such systems. The third section proves our

main result. The last section concludes with some relevant remarks.

Problem description

Consider a recursive algorithm in R^d , $d \geq 1$, described by

$$X(n+1) = X(n) + a(n)F(X(n)) \quad (1)$$

here $X(n) = [X_1(n), \dots, X_d(n)]' \in R^d$ with $X(0)$ prescribed, and $F(\cdot) = [F_1(\cdot), \dots, F_d(\cdot)]' : R^d \rightarrow R^d$ is a bounded map satisfying the Lipschitz condition:

$$\|F(x) - F(y)\| \leq K \|x - y\|; x, y \in R^d$$

for some $K > 0$ (Prime denotes transposition of a vector or a matrix). $\{a(n)\}$ is a prescribed sequence of strictly positive numbers. The i -th row of this vector iteration reads

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n), \dots, X_d(n)). \quad (2)$$

In our model of distributed computation with delays, we replace (2) by

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n - \tau_{i1}(n)), \dots, X_d(n - \tau_{id}(n))) \quad (3)$$

where $\{\tau_{ij}(n), 1 \leq i, j \leq d\}$ are random delays encountered at the $(n+1)$ -st iteration. Our aim is to analyse the system (3) under the following assumptions:

(A1) $a(n) = 1/n, n = 1, 2, \dots, a(0) = 1$.

(A2) The delays $\{\tau_{ij}(n)\}$ satisfy: There exist $b, C > 0$ such that

$$E[(\tau_{ij}(n))^b / X(m), m \leq n, \tau_{pq}(k), k < n, 1 \leq p, q \leq d] \leq C \quad (4)$$

with probability one, for all i, j, n .

(A3) The ordinary differential equation (ODE) described by

$$\dot{x}(t) = F(x(t)) \quad (5)$$

is a 'strict Liapunov system' described later in this section.

Some comments are called for regarding these assumptions. (A1) implies $\sum_n a(n)^2 < \infty, \sum_n a(n) = \infty$, which is the condition used in stochastic approximation theory to suppress the effect of noise. Our choice of $\{a(n)\}$ coupled with (4) will simplify our analysis considerably. This choice, however, is not very rigid. One could use other stochastic approximation-like tapering stepsize routines in conjunction with an appropriate modification of (4), as will become evident later. Note that (4), in particular, implies that the conditional distributions of τ_{ij} given the past are 'tight' in probability theory parlance. Finally, the Lipschitz condition on F implies that (5) has a unique solution for all $t \geq 0$. Later on we shall comment on the possibility of relaxing this condition.

A continuously differentiable function $V: R^d \rightarrow R^+$ is said to be a Liapunov function for (5) if $\nabla V \cdot F \leq 0$ everywhere and a strict Liapunov function if in addition, $\nabla V \cdot F < 0$ outside the set $E_F = \{z \mid F(z) = 0\}$. Call (5) a strict Liapunov system if it has bounded trajectories, E_F consists of isolated points and a strict Liapunov

function V exists. (The latter would imply bounded trajectories if $V(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$.) Some examples of such systems are:

(i) Gradient system: Consider $F = -a\nabla f$ where $a > 0$ and $f: R^d \rightarrow R$ is a continuously differentiable, proper function (i.e., $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$) with bounded gradient and isolated critical points. $V(\cdot) = f(\cdot) + \min f(x)$ will do the job.

(ii) Newtonian systems (Vidyasagar 1978): For f as above, consider

$$\ddot{x}(t) + h(x(t))\dot{x}(t) + a\nabla f(x(t)) = 0 \quad (6)$$

where $h: R^d \rightarrow R^+$ is bounded Lipschitz. This is Newton's law in a potential field with position-dependent kinetic friction. This can be rewritten as

$$\dot{x}(t) = y(t), \dot{y}(t) = -h(x(t))y(t) - a\nabla f(x(t)),$$

which conforms with (5). Assuming $f \geq 0$, the 'total energy' $V(x, y) = 1/2y^2 + af(x)$ serves as a strict Liapunov function as long as $h > 0$ outside $E_f = \{x \mid \nabla f(x) = 0\}$. For constant h , (6) will be recognized as the continuous analog of 'gradient descent with a momentum term' for minimizing f . For constant h , if we approximate f by a positive definite quadratic form near a local minimum of f , (6) looks like the familiar second order linear system of control systems (Ogata 1970) and may be expected to behave like one locally. That is, if h is chosen to make the latter 'underdamped', it will quickly hit a local minimum and then oscillate around it before settling down. The oscillations may be quenched by making h position dependent, with low values away from E_f and high values on E_f . Preliminary numerical experiments by the second author suggest that this is a promising strategy for speed-up.

(iii) Neural networks: The area of analog neural networks has several examples of strict Liapunov systems (Hirsch 1987; Schurmann 1989). Hirsch (1987) describes a scheme for building more complicated systems from 'cascades' of simpler ones, while Cohen (1992) considers the 'inverse problem' of constructing a strict Liapunov system with a prescribed equilibrium behaviour.

(iv) Fixed point algorithms: Recall the norms $\|\cdot\|_p, p \in [1, \infty]$, on R^d defined by: for $x = [x_1, \dots, x_d]' \in R^d$,

$$\begin{aligned} \|x\|_p &= \left(\frac{1}{d} \sum_{i=1}^d |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty, \\ \|x\|_\infty &= \max_i |x_i|. \end{aligned}$$

Suppose $F(x) = G(x) - x$ where G satisfies:

$$\|G(x) - G(y)\|_p \leq \alpha \|x - y\|_p, \quad x, y \in R^d,$$

for some $\alpha \in (0, 1), p \in [1, \infty]$. Then by the 'contraction mapping' theorem (Bertsekas & Tsitsiklis 1989), E_F is a singleton $\{x^*\}$ (say). Then $V(x) = \|x - x^*\|_p$ serves as a strict Liapunov function (Borkar & Soumyanath 1994). (One should note that V is not differentiable everywhere for $p = 1, \infty$, but this does not pose any problems, Borkar & Soumyanath 1994).

Our results will imply that discrete time, distributed implementations of (i) - (iv) above will asymptotically track the continuous dynamics under (A1) - (A2).

We conclude this section by recalling from Hirsch (1987) a result that will be central to our analysis. For given $T, \delta > 0$, a (T, δ) -perturbed trajectory of (5) is a (possibly discontinuous) function $y : [0, \infty) \rightarrow R^d$ such that the following conditions hold: There is an increasing sequence $T_j \rightarrow \infty$ and solutions $x^j(t), t \in [T_j, T_{j+1}]$ of (5) such that $T_{j+1} - T_j \geq T$ for all $j = 1, 2, \dots$, and

$$\|y(t) - x^j(t)\| < \delta, \quad T_j \leq t \leq T_{j+1}, j \geq 1.$$

Theorem 1. *Hirsch (1987) for a strict Liapunov system (5), for any $T, \epsilon > 0$, there exists a $\delta_0 = \delta_0(T, \epsilon) > 0$ such that if $0 < \delta < \delta_0$, then any limit point of a (T, δ) -perturbed trajectory $y(\cdot)$ is within ϵ of a point $p \in E_F$. Moreover, if $y(0)$ is in the domain of attraction of a stable equilibrium point $q \in E_F$, then for sufficiently small δ , we can take $p = q$.*

The way we shall use this result is by showing that a suitably interpolated version of (3) is in fact a (T, δ) -perturbed trajectory of (5). The details are carried out in the next section.

3. Main result

We shall proceed through a sequence of lemmas. Let $T > 0$ and define time sequences $\{t_n\}$ and $\{T_n\}$ as follows:

$$\begin{aligned} t_0 &= T_0 = 0. \\ t_n &= \sum_{i=1}^n 1/i, \quad n \geq 1. \\ T_n &= \min\{t_m \mid t_m \geq T_{n-1} + T\}, n \geq 1. \end{aligned}$$

Then $T_n = t_{m(n)}$ for a strictly increasing sequence $\{m(n)\}$ and $T_n \in [T_{n-1} + T, T_{n-1} + T + 1]$ for all n . Let $I_n = [T_n, T_{n+1}]$, $n \geq 1$. Let $x(t), t \in I_n$, be a solution of (5) with $x(T_n) = X(m(n))$ and define $y(t), t \in I_n$, by:

$$y(T_n) = x(T_n),$$

$$y(t_{m(n)+k+1}) = y(t_{m(n)+k}) + F(y(t_{m(n)+k}))(t_{m(n)+k+1} - t_{m(n)+k}),$$

with $y(\cdot)$ linearly interpolated on the intervals $[t_{m(n)+k}, t_{m(n)+k+1}]$. Let $\epsilon > 0$.

Lemma 1. For sufficiently large n ,

$$\sup_{I_n} \|x(t) - y(t)\| < \epsilon.$$

Proof. For $t \in I_n$, let $g(t) = \max\{t_{m(n)+k} \mid t_{m(n)+k} \leq t\}$. Then for $t \in I_n$,

$$y(t) = y(T_n) + \int_{T_n}^t F(y(g(s)))ds,$$

$$x(t) = x(T_n) + \int_{T_n}^t F(x(s)) ds.$$

Letting K, K' be resp. the Lipschitz constant of and a bound on F , we have

$$\begin{aligned} \|x(t) - y(t)\| &\leq K \int_{T_n}^t \|x(s) - y(s)\| ds + K \int_{T_n}^t \|y(s) - y(g(s))\| ds \\ &\leq K \int_{T_n}^t \|x(s) - y(s)\| ds + KK' \sum_{k=m(n)}^{m(n+1)} (1/k^2), \end{aligned}$$

where we use the fact that

$$\|y(t) - y(t_i)\| \leq K' |t_{i+1} - t_i| \text{ for } t \in [t_i, t_{i+1}].$$

The claim now follows from the Gronwall lemma (Vidyasagar 1978, pp. 292) and the summability of $\sum (1/k^2)$. $\square$

Next, write (3) as a vector iteration

$$X(n+1) = X(n) + \frac{1}{n} Z(n)$$

with appropriately defined $Z(n) = [Z_1(n), \dots, Z_d(n)]'$ and set

$$\hat{Z}(n) = E[Z(n)/X(m), m \leq n, \tau_{ij}(k), k < n, 1 \leq i, j \leq d], n \geq 0,$$

where the conditional expectation is componentwise. Let

$$M_n = \sum_{i=1}^n \frac{1}{i} (Z(i) - \hat{Z}(i)), n \geq 0.$$

Lemma 2. $\{M_n\}$ converges as $n \rightarrow \infty$, with probability one.

Proof. $\{M_n\}$ is seen to be a d -dimensional martingale (Neveu 1975) satisfying

$$\|M_{n+1} - M_n\| \leq 2K'd/n.$$

Thus,

$$E[\|M_{n+1} - M_n\|^2 / M_k, k \leq n] \leq K''/n^2$$

for a suitable $K'' > 0$. Since the right hand side is summable, the claim follows from Proposition VII-2-3(c), pp.149-150 (Neveu 1975). $\square$

Lemma 3. There exist $\bar{C}, r > 0, \bar{N} \geq 1$ such that for all $n \geq \bar{N}$,

$$\|F(X(n)) - \hat{Z}(n)\| < \bar{C}/n^r.$$

Proof. Writing $\hat{Z}(n) = [\hat{Z}_1(n), \dots, \hat{Z}_d(n)]'$, we shall compare $F_i(X(n)), \hat{Z}_i(n)$, for a fixed $i, 1 \leq i \leq d$. Let $1 > c > 0$. We have

$$\begin{aligned} &|F_i(X(n)) - \hat{Z}_i(n)| \\ &\leq E[|F_i(X(n)) - Z_i(n)| I\{\tau_{ij}(n) \leq n^c \text{ for all } j\} / X(m), m \leq n, \tau_{pq}(k), \\ &k < n, 1 \leq p, q \leq d] + E[|F_i(X(n)) - Z_i(n)| I\{\tau_{ij}(n) > n^c \text{ for some } j\} / X(m), \\ &m \leq n, \tau_{pq}(k), k < n, 1 \leq p, q \leq d]. \end{aligned}$$

The second term can be bounded by $2K'Cd/n^{bc}$ in view of (4) and the Chebyshev inequality. Let $\bar{n}$ = the integer part of n^c . Let n be large enough so that $n > \bar{n}$. Then for $m \leq \bar{n}$,

$$\|X(n) - X(n-m)\| \leq K' \sum_{k=n-\bar{n}}^n \frac{1}{k} \leq \bar{K}/n^{1-c}$$

for a suitable constant $\bar{K} > 0$. Thus the first term above may be bounded by $K\bar{K}d/n^{1-c}$. The claim follows. $\square$

Now define $\bar{X}(t), t \geq 0$, by: $\bar{X}(t_n) = X(n)$ for $n \geq 0$, with $\bar{X}(\cdot)$ defined by linear interpolation on each interval $[t_n, t_{n+1}]$, $n \geq 0$. For the next lemma, fix a sample point for which the conclusions of lemma 2 hold.

Lemma 4. For sufficiently large n ,

$$\sup_{I_n} \|\bar{X}(t) - y(t)\| < \epsilon.$$

Proof. Let $n \geq 1$. For $i \geq m(n)$, we have

$$\bar{X}(t_{i+1}) = \bar{X}(t_i) + \frac{1}{i}F(\bar{X}(t_i)) + \frac{1}{i}(\hat{Z}(i) - F(\bar{X}(t_i))) + \frac{1}{i}(Z(i) - \hat{Z}(i)).$$

Define $\xi_i = M_i - M_{m(n)}$, $\hat{X}_{i+1} = \bar{X}(t_{i+1}) - \xi_i$ for $i \geq m(n)$ with $\hat{X}_{m(n)} = X(m(n))$. (Thus $\xi_{m(n)-1} = 0$). Then

$$\hat{X}_{i+1} = \hat{X}_i + \frac{1}{i}F(\hat{X}_i) + \frac{1}{i}(F(\hat{X}_i + \xi_{i-1}) - F(\hat{X}_i)) + \frac{1}{i}(\hat{Z}(i) - F(\bar{X}(t_i))).$$

Also,

$$y(t_{i+1}) = y(t_i) + \frac{1}{i}F(y(t_i)).$$

Thus for n sufficiently large, lemma 3 leads to

$$\|\hat{X}_{i+1} - y(t_{i+1})\| \leq (1 + \frac{K}{i}) \|\hat{X}_i - y(t_i)\| + \frac{K}{i} |\xi_{i-1}| + \frac{\bar{C}}{i^{1+r}}.$$

Let $\delta > 0$. By lemma 2, for sufficiently large n ,

$$\sup_{i \geq m(n)} |\xi_i| < \delta/2,$$

$$\sum_{i \geq n} (1/i^{1+r}) < \frac{\delta}{2}.$$

Thus for n sufficiently large, using the inequality $1 + \frac{K}{n} \leq e^{K/n}$ and iterating, one has

$$\sup_{m(n) \leq i \leq m(n+1)} \|\hat{X}_i - y(t_i)\| \leq e^{K(T+1)} (\bar{C} + K(T+1))\delta.$$

Since

$$\sup_{m(n) \leq i \leq m(n+1)} \|\hat{X}_i - \bar{X}(t_i)\| < \delta/2,$$

the claim follows on choosing δ sufficiently small and observing that $\bar{X}(\cdot), y(\cdot)$ are linearly interpolated from their values at $\{t_i\}$. $\square$

We are now ready to prove our main result.

Theorem 2. *The iteration (3) converges with probability one to a single point in E_F . Furthermore, the iterates visit the basin of attraction of a stable equilibrium point in E_F infinitely often if and only if they converge to the same.*

Proof. From lemmas 1, 4 and theorem 1, it follows that the limit points of the iteration (3) are in E_F . The fact that the stepsize $a(n)$ is monotonically decreasing to zero ensures that the set of limit points is connected. Since E_F consists of isolated points, the set of limit points must therefore be a singleton. The first claim follows. The second claim follows from the second half of theorem 1. $\square$

Note that the property of a point being in the basin of attraction of a stable equilibrium point is generic for strict Liapunov systems. Thus the behaviour described in the second half of the theorem is generic.

4. Concluding remarks

We conclude by pointing out additional advantages of our model and some possible generalizations.

(i) Note that our model does not require that each value of i -th component computed at some stage be transmitted to the processor computing j -th component, or that the latter use it even when he receives it. The only requirements are that (a) if computation of i -th component requires the previous value of j -th component, then the latter be transmitted to i infinitely often, (b) the received value of the same, if used by i , be used without too much delay, i.e., such that (A2) holds. Of course, this concerns only the convergence of (3), not the speed thereof, which is a far more delicate issue that needs further study.

(ii) Several of our assumptions could be relaxed at the expense of additional technicalities. We have already mentioned that other choices of $\{a(n)\}$ with $\sum_n a(n) = \infty$, $\sum_n a(n)^2 < \infty$, could be explored. Boundedness and Lipschitz conditions on F could be dropped to some extent. For example, F being locally Lipschitz with at most linear growth will suffice for the global existence of a solution to (5). Alternatively, the condition that V is proper will do. The use of Lipschitz condition in our proofs can be replaced by locally Lipschitz property if we show a priori the fact that the iterates of (3) remain in a bounded set with probability one. Stochastic Liapunov arguments may help here. Finally, the proof of theorem 1 in Hirsch (1987) seems to extend easily to the case when E_F is not necessarily a collection of isolated points, but contains connected sets of local minima any two of which are separated by an amount exceeding a fixed $\Delta > 0$. Hirsch (1987) does not deem this situation interesting because E_F being a set of isolated points is generic from structural stability point of view. Nevertheless, one does encounter the more general situation mentioned above in 'learning' and 'identification' applications because of overparametrized model sets and therefore merits attention.

(iii) In stochastic approximation (Benveniste *et al* 1990), one studies iterations of

the type

$$X(n+1) = X(n) + a(n)G(X(n), \xi(n)) \quad (7)$$

with a prescribed $X(0)$, $\{a(n)\}$ as above, and $\{\xi(n)\}$ (in the simplest case) is a sequence of independent and identically distributed random variables, interpreted as 'noise'. Under mild conditions, one shows (Benveniste *et al* 1990) that this iteration tracks the continuous dynamics (5) with F defined by

$$F(x) = \int G(x, y)\mu(dy),$$

μ being the law of ξ_1 . If this F satisfies the conditions of this paper, our analysis extends to the distributed version of (7) with delays as in (3), satisfying (A2). This requires very minor modifications in the proof. An important special case is when $G(x, y) = -\nabla f(x - y)$ with f as in example (i) of section 2. If μ has a density φ , F becomes

$$F(x) = - \int \varphi(x - y) \nabla f(y) dy. \quad (8)$$

If φ is a 'nice' i.e. sufficiently smooth function, F may be viewed as a low pass filtered version of ∇f , or equivalently, gradient of a low pass filtered version of F . If F is jagged with several small, sharp valleys, it is not amenable for gradient-based methods and the right dose of low pass filtering may be expected to replace it by a better behaved function without significantly perturbing the location of $\text{Argmin}(f)$. In this viewpoint, $\{\xi_n\}$ is not 'noise', but a benign randomization that may be deliberately introduced to achieve the same effect as the computationally expensive convolution in (8). One may hope to further improve on the performance of this algorithm by progressively decreasing the extent of low pass filtering, e.g., by replacing the above algorithm by

$$X(n+1) = X(n) + a(n)(-\nabla f(X(n) - b(n)\xi(n)))$$

where $\{\xi(n)\}$ are i.i.d., zero mean and $\{b(n)\}$ decrease to zero at a rate slower than that of $\{a(n)\}$. This is closely related to a special case of a broad class of algorithms studied by Gelfand & Mitter (1992). These algorithms asymptotically track not an ODE as in (5), but a time-inhomogeneous stochastic differential equation. (In fact, the latter is precisely the Langevin algorithm of continuous time simulated annealing.) It would be interesting to extend our analysis to algorithms of this type.

The research of the first author was supported by the grant ISTC/EE/VS/049 from the ISRO-IISc Space Technology Cell. The research of the second author was supported by U.S. Office of Naval Research Grant No.00014-92-J-1324 under an Indo-U.S. project.

References

- Benveniste A, Metivier M, Priouret P 1990 *Adaptive algorithms and stochastic approximations* (Berlin-Heidelberg: Springer-Verlag)
- Bertsekas D P, Tsitsiklis J N 1989 *Parallel and distributed computation* (Englewood Cliffs, NJ: Prentice Hall)
- Borkar V S, Soumyanath K 1994 A new analog parallel scheme for fixed point computation, part I: theory. Preprint
- Chazan D, Miranker W 1969 *Chaotic oscillations, linear algebra and its applications*. 2: 199-222
- Cohen M A 1992 The construction of arbitrary stable dynamics in neural networks. *Neural Networks* 5: 83-103
- Gelfand S B, Mitter S K 1992 Recursive stochastic algorithms for global optimization in R^d . *SIAM J. Contr. Optim.* 29: 999-1018
- Gharavi R, Anantharam V 1992 A structure theorem for partially asynchronous relaxations with random delays. Preprint
- Hirsch M W 1987 Convergent activation dynamics in continuous time networks. *Neural Networks* 2: 331-349
- Neveu J 1975 *Discrete parameter martingales* (Amsterdam: North-Holland)
- Ogata K 1970 *Modern control engineering* (Englewood Cliffs, NJ: Prentice Hall)
- Schurmann B 1989 Stability and adaptation in artificial neural systems. *Phys. Rev.* A40: 2681-2688
- Vidyasagar M 1978 *Nonlinear systems analysis* (Englewood Cliffs, NJ: Prentice Hall)