SIAM J. OPTIMIZATION
Vol. 3, No. 4, pp. 868–881, November 1993

012

ACCELERATED STOCHASTIC APPROXIMATION*

BERNARD DELYON† AND ANATOLI JUDITSKY†

Abstract. A technique to accelerate convergence of stochastic approximation algorithms is studied. It is based on Kesten's idea of equalization of the gain coefficient for the Robbins–Monro algorithm. Convergence with probability 1 is proved for the multidimensional analog of the Kesten accelerated stochastic approximation algorithm. Asymptotic normality of the delivered estimates is also shown. Results of numerical simulations are presented that demonstrate the efficiency of the acceleration procedure.

Key words. stochastic approximation, accelerated algorithms, optimal algorithms

AMS subject classifications. 62L20, 93B30

1. Introduction. Let us consider the problem of searching for the stationary point x^* of the vector field $\varphi(x) : R^N \to R^N$. The observations y_t of the function $\varphi(\cdot)$ are available at any point $x_{t-1} \in R^N$ and contains random disturbance ξ_t:

$$y_t = \varphi(x_{t-1}) + \xi_t. \tag{1}$$

The problem is to find x^* under the assumption that a unique solution exists.

The method of stochastic approximation (SA) (which takes its origin from [10]) is well studied for this problem. To obtain a sequence of estimates of the solution x^*, the following recursive procedure is used:

$$x_t = x_{t-1} - \gamma_t y_t, \tag{2}$$

where γ_t is a gain coefficient and x_0 is an arbitrary fixed point in R^N. In the study of this algorithm the main focus of attention was the asymptotic analysis of the method when $\gamma_t = \gamma t^{-1}$. For this case conditions have been obtained under which almost sure convergence and asymptotic normality take place (see [6] and [15]). Asymptotically optimal versions (algorithms that ensure the highest asymptotic rate of convergence) of that method have also been developed in the works of Venter [14], Fabian [3], and Polyak and Tsypkin [9].

On the other hand, nonasymptotic properties of SA algorithms are the main focus of the interest in applications. Unfortunately, as is well known to engineers (see the discussion in [13]), asymptotically optimal methods behave badly in finite time: the choice of the gain γt^{-1} is too "cautious" if the disturbance ξ_t is small with respect to the initial error $x_0 - x^*$. Several heuristic procedures have been suggested in order to accelerate convergence in a finite time interval (see, for instance, [13, Chap. 5])[1].

In particular, the accelerated SA procedure has been studied in the work of Kesten [4] for the one-dimensional case. It is based on the idea that frequent changes of the sign of the difference $x_t - x_{t-1} = \gamma_t y_t$ indicate that the estimates are close to the real solution and are significantly disturbed by noise, whereas few fluctuations of the sign indicate that x_t is still far from x^*. In fact, the number s_t of changes of the sign of y_i for $i = 1, \ldots, t-1$ constitutes a new time scale. According to this scale, small values of s_t mean that large gains γ_t (in other words, large magnitudes of correction) should be used at the tth step and, in turn, large values of s_t mean that the procedure has "reached" its asymptotic region and $\gamma_t = \gamma t^{-1}$ should be used. Almost sure convergence of that procedure has been proved.

---

* Received by the editors on June 17, 1991; accepted for publication (in revised form) on July 25, 1992.

† Institut de Recherche en Informatique et Systèmes Aléatoires/Institut National de Recherche en Informatique et en Automatique, Campus de Beaulieu, 35042, Rennes, France.

[1] As noted in [4], the investigation of this problem had been suggested by Robbins in his first works on SA.

The principal issue of this paper is a result of the almost sure convergence of the multidimensional analog of Kesten's algorithm. Based on that in §3 we obtain conditions for asymptotic normality for the accelerated version of the usual SA procedure. In §4 we study Kesten-like modification of the Ruppert–Polyak (see [8] and [12]) SA algorithm. Section 5 contains results of numerical simulations.

2. Kesten's algorithm. In order to obtain the estimates x_t of x^* we use the following algorithm:

$$x_t = x_{t-1} - \gamma_t y_t, \qquad x_0 \in R^N, \tag{3}$$

where the scalar gain γ_t is defined by the equations

$$s_{t+1} = s_t + I(y_t^T y_{t-1} < 0), \tag{4}$$

$$\gamma_{t+1} = \gamma(s_{t+1}) \tag{5}$$

(here $\gamma(t)$ is a deterministic sequence).

We suppose that we have a probability space $(\Omega, \mathcal{F}, P)$ with an increasing family of σ-fields $\mathcal{F}_t = \sigma(x_0, \xi_1, \ldots, \xi_t)$. Let us consider the following assumptions on the problem.

ASSUMPTION 1. ξ_t is a sequence of random variables such that the conditional distribution $P_x(d\xi)$ of ξ_t, knowing the past, depends only on $x_{t-1} = x$. Furthermore, $E(\xi_t|x_{t-1}) = 0$ and for some S_M, $E(\xi_t \xi_t^T|x_{t-1}) \le S_M$. The measures P_x satisfy

$$\lim_{x \to x^*} \|P_x - P_{x^*}\| = 0, \tag{6}$$

where $\|\cdot\|$ denotes the total variation. Moreover, for any hyperplane H containing the origin, $P_{x^*}(H) = 0$. For any $R > 0$ and $\delta > 0$,

$$\min_{|x|<R} P_x(|\xi| \le \delta) > 0. \tag{7}$$

ASSUMPTION 2. $\varphi(x)$ is a continuous function of x. There exists $\rho > 0$ such that for any $\gamma^* \le \rho$ and any starting point x_0, the deterministic sequence

$$x_{t+1} = x_t - \gamma^* \varphi(x_t) \tag{8}$$

converges to x^*. There exists a function $V(x) : R^N \to R^+$, positive $\beta, \bar{R}$, and a matrix $M > 0$ such that

$$\begin{aligned} &V(x^*) = 0, \\ &\nabla^2 V(x) \le M \quad \text{for all } x, \\ &\varphi(x)^T \nabla V(x) \ge \frac{\rho}{2}(\varphi(x)^T M \varphi(x) + tr(S_M M)) + \beta \end{aligned}$$

for any x such that $|x - x^*| \ge \bar{R}$. Moreover,

$$\varphi(x)^T \nabla V(x) > 0 \quad \text{for any } x \ne x^*.$$

ASSUMPTION 3. The gain coefficient $\gamma(n) > 0$ satisfies

$$\sup_n \gamma(n) \le \rho, \quad \sum_{n=1}^{\infty} \gamma(n) = \infty, \quad \sum_{n=1}^{\infty} \gamma^2(n) < \infty.$$

Note that Assumption 1 implies that

$$\lim_{x \to x^*} E_x \xi \xi^T = E_{x^*} \xi \xi^T = S(x^*) \tag{9}$$

exists, and there is $v > 0$ such that $S(x^*) > vI$ (here E_x denotes the expectation with respect to P_x). Denote $P^* = P_{x^*} \otimes P_{x^*}$.

Comment. We present here an example of the procedure when the conditions stated above are satisfied. Let us consider the following *nonlinear* algorithm for estimating x^*:

$$x_t = x_{t-1} - \gamma_t f(y_t).$$

Here $f(x) : R^N \to R^N$ is a nonlinear function. We can rewrite this algorithm in a form similar to (3):

$$x_t = x_{t-1} - \gamma_t \psi(x_{t-1}) - \gamma_t \xi_t,$$

where $\psi(x) = E_x f(\varphi(x) + \xi)$ and $\zeta_t = f(\varphi(x) + \xi) - \psi(x)$. Suppose that Assumption 1 is satisfied. We require that $|f(x)| \le K_0(1 + |x|)$ and $f(x)$ is continuous. This implies that the functions ψ and $\chi(x) = E_x \zeta_t \zeta_t^T$ are correctly defined and there is K_1 such that $|\chi(x)| \le K_1$. Furthermore, if the distribution P_x of ξ is absolutely continuous, then Assumption 1 with respect to ζ_t holds true. Given some additional assumptions on f, Assumption 2 can also be verified.

Moreover, one can study the case of nonadditive disturbances (when $y_t = \varphi(x_{t-1}, \xi_t)$; see [1]) in the same way.

THEOREM 1. *Let Assumptions* 1–3 *hold. Then the process defined by* (3)–(5) *satisfies*

$$x_t \to x^* \quad a.s.,$$
$$\lim_{t \to \infty} \frac{s_t}{t} - P^*(\xi_1^T \xi_2 < 0) \to 0 \quad a.s.$$

Comment. A result similar to the first proposition of Theorem 1 has been stated for the one-dimensional case in [4]. The second proposition of the theorem states that the new time scale, defined by (4), is asymptotically equivalent (up to a coefficient) to the original scale.

Assumption 3 is typical when dealing with stochastic approximation algorithms. Assumption 2 is specific to the Kesten algorithm. It guarantees the stability of the Markov chain, defined by (3) when $\gamma_t \equiv \gamma$. As we shall see later, it ensures a certain regularity in the increase of s_t.

Note that we cannot directly utilize classical results on almost sure convergence of SA procedures (see, for instance, [5, Thm. 2.3.3] and [1, Thm. 2.5.1]). Indeed, the conditions of these results demand, at least, that $\gamma_t \to 0$ as $t \to \infty$, which is not obvious for the algorithm under consideration.

Proof of Theorem 1. In the proofs of the theorems let us adopt the following conventions: we denote by K, δ, and α generic positive constants. All relations between random variables are supposed to be true almost surely.

An outline of the proof is as follows. We show first the positive recurrency of the process x_t in some vicinity of x^*, where the disturbance ξ_t forces s_t to increase regularly, so that $\gamma_t \to 0$. Next we prove that (x_t) visits any neighborhood of x^* infinitely often. We conclude the proof by showing that x_t escapes an arbitrary neighborhood of x^* only a finite number of times.

For the sake of simplicity let us put $x^* = 0$. The following lemma will be used to prove that s_t tends to infinity; it is actually slightly stronger than we need.

LEMMA 1. *For any starting point $z_0 \in R^N$, and any $\gamma^* \leq \rho$, the Markov chain z_t resulting from the equation*

$$z_{t+1} = z_t - \gamma^*(\varphi(z_t) + \xi_{t+1}) \tag{10}$$

satisfies

(i) *$P(z_t \in B(\bar{R})$ infinitely often$) = 1$, where $B(\bar{R})$ is a ball $\{|x| < \bar{R}\}$ and $\bar{R}$ is defined as in Assumption 2.*

(ii) *There exist $\epsilon > 0$ and n_0 such that*

$$P(z_{n_0}^T z_{n_0+1} < 0 | z_0) > \epsilon$$

for any $z_0 \in B(\bar{R})$.

Proof. Put $V_t = V(z_t)$ and $\varphi_t = \varphi(z_t)$. As z_t satisfies (10), we have

$$V_{t+1} \leq V_t - \gamma^* \varphi_t^T \nabla V_t - \gamma^* \xi_t^T \nabla V_t + \gamma^{*2}(\varphi_t + \xi_t)^T M(\varphi_t + \xi_t)/2,$$

and from Assumptions 1 and 2

$$\begin{aligned} E(V_{t+1}|z_t) &\leq V_t - \gamma^* \varphi_t^T \nabla V_t + \gamma^{*2}(\varphi_t^T M \varphi_t + tr(S_M M))/2 \\ &\leq V_t - \gamma^* \varphi_t^T \nabla V_t + \gamma^{*2}(\varphi_t^T \nabla V_t - \beta)/\rho + KI(|z_t| < \bar{R}) \\ &\leq V_t - \gamma^{*2}\beta/\rho + KI(|z_t| < \bar{R}). \end{aligned} \tag{11}$$

Define a stopping time $\nu = \inf\{t \geq 1 : |z_t| \leq \bar{R}\}$. Then we derive from (11)

$$\begin{aligned} 0 &\leq EV_{t+1} I(t < \nu) \\ &\leq EV_t I(t < \nu) - \gamma^{*2}\beta/\rho EI(t < \nu) \\ &\leq EV_t I(t-1 < \nu) - \gamma^{*2}\beta/\rho EI(t < \nu) \\ &\leq V_0 + K - \gamma^{*2}\beta/\rho E\left(\sum_{i=0}^{t} I(i < \nu)\right) \\ &= V_0 + K - E\nu\gamma^{*2}\beta/\rho. \end{aligned}$$

Thanks to Assumption 2, $V_0 \leq K|z_0|^2$. Thus

$$E\nu \leq K(|z_0|^2 + 1)\rho/(\gamma^{*2}\beta) \quad \text{and} \quad \nu < \infty \quad \text{a.s.}$$

Hence

$$P(z_t \in B(\bar{R}) \text{ i.o.}) = 1. \tag{12}$$

Note that

$$z_0^T z_1 = |z_0|^2 - \gamma^* z_0^T \varphi_0 - \gamma^* z_0^T \xi_1$$

and the distribution $P_0(d\xi)$ is nondegenerate. Thus the continuity of the $\phi(\cdot)$ along with condition (6) implies the existence of $\delta_1 > 0$ and $\epsilon_1 > 0$ such that

$$P(z_0^T z_1 < 0 | z_0) > \epsilon_1 \tag{13}$$

for any $z_0 \in B(\delta_1)$. On the other hand, from the convergence of the deterministic counterpart (8) of the algorithm, condition (7), and, again, the continuity of $\phi(\cdot)$, we obtain that for any $\delta_1 > 0$ there exist n_0 and $\epsilon_2 > 0$ such that

$$P(|z_{n_0}| \leq \delta_1) \geq \epsilon_2 \tag{14}$$

for any $z_0 \in B(\bar{R})$. Hence we get from (14) and (13) that

$$P(z_{n_0}^T z_{n_0+1} < 0 | z_0) > \epsilon_2 \epsilon_1 = \epsilon$$

as soon as $z_0 \in B(\bar{R})$. □

LEMMA 2. $s_t \to \infty$ *almost surely.*

Proof. For any integer s^*

$$\begin{aligned} P(\lim_{t\to\infty} s_t = s^*) &\le \sum_t P(s_i = s^* \text{ for any } i > t) \\ &= \sum_t E(P(s_i = s^* \text{ for any } i > t | \mathcal{F}_t)). \end{aligned}$$

It follows from the strong Markov property that the conditional to the $\mathcal{F}_t$ law of the process (γ_{t+i}, x_{t+i}) if s_{t+i} remains equal to s^* coincides with the law Q_{x_t} of the Markov chain given by (10) with $\gamma^* = (s^*)^{-1}$ and starting point $z_0 = x_t$. Consequently,

$$P(\lim_{t\to\infty} s_t = s^*) \le \sum_t E(Q_{x_t}(z_i^T z_{i-1} \ge 0 \text{ for all } i \ge 1)). \tag{15}$$

However, by standard manipulations (see [2, Problem 9, Chap. 5.6]), we get from Lemma 1 (ii) that for any $z_0 \in R^N$

$$\{z_i^T z_{i-1} < 0 \text{ i.o.}\} \supset \{z_i \in B(\bar{R}) \text{ i.o.}\} \text{ a.s.}$$

Hence by Lemma 1 (i), we conclude that $Q_{z_0}(z_i^T z_{i-1} < 0 \text{ infinitely often}) = 1$ for any z_0. This implies

$$Q_x(x_i^T z_{i-1} \ge 0 \quad \text{for all } i \ge 1) = 0 \quad \text{for any } x$$

and consequently $P(\lim_{t\to\infty} s_t = s^*) = 0$ for any s^*. □

LEMMA 3. *For any* $\epsilon > 0, x_t \in B(\epsilon)$ *infinitely often* (*in other words,* x_t *visits any neighborhood of zero infinitely often*).

Proof. Define for any $\gamma^* > 0$ the stopping times

$$\begin{aligned} \sigma &= \inf(t : \gamma_t \le \gamma^*), \\ \tau &= \inf(t \ge \sigma : |x_t| < \epsilon). \end{aligned}$$

We have from Assumption 2 that for γ^* small enough for all $|x| > \epsilon$ and $\gamma \le \gamma^*$

$$\begin{aligned} &\varphi(x)^T \nabla V(x) - \frac{\gamma}{2}(\varphi(x)^T M \varphi(x) + trMS_M) \\ &\ge \Big(\varphi(x)^T \nabla V(x) - \frac{\gamma}{2}(\varphi(x)^T M \varphi(x) + trMS_M)\Big) I(|x| > \bar{R}) \\ &\quad + \varphi(x)^T \nabla V(x) I(\epsilon < |x| \le \bar{R}) - \gamma K \ge \delta(\epsilon) \end{aligned}$$

with $\delta(\epsilon) > 0$. Thus we obtain from (3) for all $t > \sigma$ (since $\{t > \sigma\}$ is $\mathcal{F}_{t-1}$ measurable)

$$\begin{aligned} E(V_t I(t \le \tau)|\mathcal{F}_{t-1}) &\le I(t-1 \le \tau) E(V_t|\mathcal{F}_{t-1}) \\ &\le V_{t-1} I(t-1 \le \tau) - \gamma_t \varphi(x_{t-1})^T \nabla V_{t-1} I(t-1 \le \tau) \\ &\quad + \frac{\gamma_t^2}{2}(\varphi(x_{t-1})^T M \varphi(x_{t-1}) + trMS_M) I(t-1 \le \tau) \\ &\le V_{t-1} I(t-1 \le \tau) - \gamma_t \delta(\epsilon) I(t-1 \le \tau), \end{aligned} \tag{16}$$

with $\delta(\epsilon) > 0$. Hence, taking expectation with respect to $\mathcal{F}_\sigma$ and summing up to τ, we obtain from (16)

$$\delta(\epsilon) E\left(\sum_{i=\sigma+1}^{\tau} \gamma_i | \mathcal{F}_\sigma\right) \le V_\sigma.$$

Due to Lemma 2, $\sigma < \infty$ and hence $V_\sigma < \infty$. It is clear that

$$E\left(\sum_{i=\sigma+1}^{\tau} \gamma(i) | \mathcal{F}_\sigma\right) \le E\left(\sum_{i=\sigma+1}^{\tau} \gamma_i | \mathcal{F}_\sigma\right) < \infty.$$

From the fact that $\sum_{i=1}^{\infty} \gamma(i) = \infty$, we conclude that $\tau < \infty$. $\square$

For $\epsilon > 0$ small enough, let us define the stopping times

$$\begin{aligned} \tau &= \inf(t : V(x_t) > \epsilon), \\ \sigma_k &= \inf(t : s_t = k). \end{aligned} \tag{17}$$

LEMMA 4. *There exists $\delta_v > 0$ such that if $V(x_0) \le \epsilon/2$ then $P(\tau < \infty) \le K(\epsilon) \sum_{i=1}^{\infty} \gamma(i)^2$ for any $\epsilon < \delta_v$.*

Proof. Let us choose δ_v such that $|\varphi(x)| < \delta_1$ if $V(x) < \delta_v (\delta_1$ has been defined in (14)). From (3) we obtain by Assumption 2,

$$\begin{aligned} P(\tau < \infty) \le P\Bigg(&-\sum_{i=0}^{\tau-1} \gamma_{i+1} \varphi(x_i)^T \nabla V(x_i) - \sum_{i=0}^{\tau-1} \gamma_{i+1} \nabla V(x_i)^T \xi_{i+1} \\ &+ K \sum_{i=0}^{\tau-1} \gamma_{i+1}^2 (|\varphi(x_i)|^2 + |\xi_{i+1}|^2) \ge \epsilon/2\Bigg) \end{aligned} \tag{18}$$

$$\begin{aligned} &\le P\Bigg(\Bigg| \sum_{i=0}^{\tau-1} \gamma_{i+1} \nabla V(x_i)^T \xi_{i+1} \Bigg| \\ &\qquad + K \sum_{i=0}^{\tau-1} \gamma_{i+1}^2 (|\varphi(x_i)|^2 + |\xi_{i+1}|^2) \ge \epsilon/2\Bigg) \\ &\le P\Bigg(\Bigg| \sum_{i=0}^{\tau-1} \gamma_{i+1} \nabla V(x_i)^T \xi_{i+1} \Bigg| \ge \epsilon/4\Bigg) \end{aligned}$$

$$+ P\left(\sum_{i=0}^{\tau-1} \gamma_{i+1}^2 (|\varphi(x_i)|^2 + |\xi_{i+1}|^2) \ge K\epsilon/4\right) = I_1 + I_2. \tag{19}$$

Define the martingale

$$M_t = \sum_{i=1}^{t \wedge \tau} \gamma_i \nabla V(x_{i-1})^T \xi_i$$

(where $t \wedge \tau = \min(t, \tau)$). Then by the Doob inequality, we have

$$I_1 \le P(\sup_t |M_t| \ge \epsilon/4) \le \frac{32}{\epsilon^2} E\left(\sum_{i=1}^{\infty} \gamma_i^2 |\nabla V(x_{i-1})|^2 |\xi_i|^2 I(i \le \tau)\right)$$

$$\leq K/\epsilon^2 E\left(\sum_{i=1}^{\infty}\gamma_i^2 \sum_{|x|\leq\epsilon} |\nabla V(x)|^2 E(|\xi_i|^2 I(i\leq\tau)|\mathcal{F}_{i-1})\right)$$
$$\leq K(\epsilon)\sum_{i=1}^{\infty} E\gamma_i^2 I(i\leq\tau).$$

In an analogous way we get

$$I_2 \leq K'(\epsilon)\sum_{i=1}^{\infty} E\gamma_i^2 I(i\leq\tau).$$

Hence

$$P(\tau<\infty)\leq K\sum_{i=1}^{\infty} E\gamma_i^2 I(i\leq\tau). \tag{20}$$

Now we will show that we can substitute $\gamma(i)$ for γ_i in (20). When x_t is close to zero the noise ξ_t forces the s_t to increase regularly. Indeed, due to Assumption 1 there is $\mu>0$ such that for all x small enough

$$\mu = \max_v P_x((\xi+u)^T v<0), \qquad |u|<\delta_1. \tag{21}$$

Since the function $\varphi(\cdot)$ is continuous, we conclude from (21) that

$$\mu \leq \max_v Px((\xi+\varphi(x))^T v<0)$$

as soon as $V(x)<\delta_v$. In other words,

$$P((\xi_t+\varphi(x_{t-1}))^T(\xi_{t-1}+\varphi(x_{t-2}))<0|\mathcal{F}_{t-1})\geq\mu$$

for $|x_{t-1}|$ such that $V(x_{t-1})<\delta_v$. Hence

$$P(\{s_{t+1}-s_t>0\}\cap\{t\leq\tau\}|\mathcal{F}_{t-1})\geq \mu I(t\leq\tau). \tag{22}$$

Define $\nu_k=\min(\sigma_k,\tau)$. Then we have

$$\sum_{i=1}^{\tau}\gamma_i^2=\sum_{i=1}^{\tau}\gamma^2(s_i)\leq\sum_{k=0}^{\infty}\gamma^2(k)(\nu_{k+1}-\nu_k).$$

Next, for any $n\geq 0$ we obtain from (22)

$$\begin{aligned}&P(\nu_{k+1}-\nu_k\geq n|\mathcal{F}_{\nu_k})\\ &\quad = P(\{\text{there is no change of the gain coefficient on } n \text{ steps}\}\cap\{\nu_k+n\leq\tau\}|\mathcal{F}_{\nu_k})\\ &\quad \leq (1-\mu)^n,\end{aligned}$$

which implies that $E(\nu_{k+1}-\nu_k)\leq\mu^{-1}$. Therefore,

$$E\sum_{i=1}^{\tau}\gamma_i^2\leq\sum_{k=0}^{\infty}\gamma^2(k)E(\nu_{k+1}-\nu_k)\leq\frac{1}{\mu}\sum_{k=0}^{\infty}\gamma^2(k);$$

hence

$$P(\tau < \infty) \leq \frac{K(\epsilon)}{\mu} \sum_{k=0}^{\infty} \gamma^2(k). \qquad \square$$

LEMMA 5. $x_t \to 0$ *almost surely.*

Proof. Denote

$$A = \{|x_t| > \epsilon \text{ i.o.}\}. \tag{23}$$

Define the stopping time $\tau_k = \inf(t \geq \sigma_k : x_t \in B(\epsilon/2))$ with σ_k defined in (17). From Lemmas 2 and 3, we have that the sequence τ_k is strictly increasing and finite. The Markov property then implies that for all k

$$P(A) = P(A \cap \{\tau_k < \infty\}) \leq E(I(\tau_k < \infty) P_{\tau_k}(A)) \leq K(\epsilon) \sum_{i=k}^{\infty} \gamma_i^2.$$

Thus $P(A^c) = 1$. Due to the arbitrary choice of ϵ in (23), we obtain the desired proposition. $\square$

The objective of the following proposition is to obtain an estimate of the speed of convergence of s_t/t to its limit.

PROPOSITION 1. $s_t/t \to P^*(\xi_1^T \xi_2 < 0)$ *almost surely.*

Proof. Put

$$\Psi_x(a) = \max_{|u|=1} P_x(|u^T \xi| \leq a).$$

Note that $\Psi_x(a)$ is the highest probability of a stripe of width $2a$ "centered in 0" under the conditional law of ξ. We use the decomposition

$$s_t = \sum_{i=1}^{t-1} s_{i+1} - s_i - I(\xi_i^T \xi_{i-1} < 0) + \sum_{i=1}^{t-1} I(\xi_i^T \xi_{i-1} < 0),$$

and setting $\varphi_i = \varphi(x_i)$, we obtain the following bound for the first term:

$$\begin{aligned}
&|s_{t+1} - s_t - I(\xi_t^T \xi_{t-1} < 0)| \\
&\quad = |I((\varphi_{t-1} + \xi_t, \varphi_{t-2} + \xi_{t-1}) < 0) - I(\xi_t^T \xi_{t-1} < 0)| \\
&\quad \leq |I((\varphi_{t-1} + \xi_t, \varphi_{t-2} + \xi_{t-1}) < 0) - I((\xi_t, \varphi_{t-2} + \xi_{t-1}) < 0)| \\
&\qquad + |I((\xi_t, \varphi_{t-2} + \xi_{t-1}) < 0) - I(\xi_t^T \xi_{t-1} < 0)| \\
&\quad \leq I(|(\xi_t, \varphi_{t-2} + \xi_{t-1})| < |(\varphi_{t-1}, \varphi_{t-2} + \xi_{t-1})|) + I(|\xi_t^T \xi_{t-1}| < |\xi_t^T \varphi_{t-2}|). \\
&\quad = u_t + v_t.
\end{aligned}$$

From the Neveu martingale theorem [7], we have

$$\begin{aligned}
\sum_{i=1}^{t-1} u_i &= \sum_{i=1}^{t-1} u_i - E(u_i|\mathcal{F}_{i-1}) + \sum_{i=1}^{t-1} E(u_i|\mathcal{F}_{i-1}) \\
&= o(t^{1/2+\alpha}) + \sum_{i=1}^{t-1} \Psi_{x_{i-1}}(|\varphi_{i-1}|).
\end{aligned}$$

Let us estimate

$$E(v_t|\mathcal{F}_{t-2}) = E(P_{x_{t-1}} \otimes P_{x_{t-2}}(|\xi_t^T \xi_{t-1}| < |\xi_t^T \varphi_{t-2}|)|\mathcal{F}_{t-2}).$$

Substituting the law P_0 for $P_{x_{t-1}}$, we get

$$\begin{aligned} E(v_t|\mathcal{F}_{t-2}) &\leq P_0 \otimes P_{x_{t-2}}(|\xi^T \xi_{t-1}| < |\xi^T \varphi_{t-2}|) + E(\|P_{x_{t-1}} - P_0\||\mathcal{F}_{t-2}) \\ &\leq \Psi_{x_{t-2}}(|\varphi_{t-2}|) + (E(\|P_{x_{t-1}} - P_0\||\mathcal{F}_{t-2}) - \|P_{x_{t-1}} - P_0\|) + \|P_{x_{t-1}} - P_0\|. \end{aligned}$$

Using again the Neveu theorem, we obtain

$$\begin{aligned} \sum_{i=1}^{t-1} v_i &= \sum_{i=1}^{t-1} v_i - E(v_i|\mathcal{F}_{i-1}) + \sum_{i=2}^{t-1} E(v_i|\mathcal{F}_{i-1}) - E(v_i|\mathcal{F}_{i-2}) + \sum_{i=2}^{t-1} E(v_i|\mathcal{F}_{i-2}) + Ev_1 \\ &= o(t^{1/2+\alpha}) + \sum_{i=1}^{t-2} \Psi_{x_{i-1}}(|\varphi_{i-1}|) + \|P_{x_{i-1}} - P_0\|. \end{aligned}$$

Summing up, we have

$$\begin{aligned} s_t &= o(t^{1/2+\alpha}) + 2\sum_{i=1}^{t-1} \Psi_{x_{i-1}}(|\varphi_{t-1}|) \\ &\quad + \sum_{i=1}^{t-2} \|P_{x_{i-1}} - P_0\| + \sum_{i=1}^{t-1} I(\xi_i^T \xi_{i-1} < 0) \\ &= o(t^{1/2+\alpha}) + 2\sum_{i=1}^{t-1} \Psi_0(|\varphi_{i-1}|) + 3\sum_{i=1}^{t-1} \|P_{x_{i-1}} - P_0\| + tP^*(\xi_1^T \xi_2 < 0). \end{aligned}$$

Note that since $x_t \to 0$ and the function $\varphi(\cdot)$ is continuous, we derive that

$$\frac{1}{t}\sum_{i=1}^{t-1} \Psi_0(|\varphi(x_{i-1})|) + \frac{1}{t}\sum_{i=1}^{t-1} \|P_{x_{i-1}} - P_0\| \to 0.$$

Hence $s_t/t \to P^*(\xi_1^T \xi_2 < 0)$. □

3. Asymptotic normality of the SA procedure. Consider algorithm (3)–(5) with the special choice of the gain sequence: $\gamma(t) = \gamma t^{-1}$. Denote $\zeta^{-1} = P^*(\xi_1^T \xi_2 < 0)$. We shall show that the accelerated algorithm is asymptotically equivalent to the usual SA procedure with the gain $\gamma_t = \gamma\zeta t^{-1}$. Let us consider the following assumptions.

ASSUMPTION 2′. Assumption 2 holds. Moreover,

$$|\varphi(x) - \varphi'(x^*)(x - x^*)| = o(|x - x^*|).$$

The matrix $I/2 - \gamma\zeta\nabla\varphi(x^*)$ is Hurwitz, i.e., has all strictly negative eigenvalues.

ASSUMPTION 3′. $\gamma(t) = \gamma t^{-1}$ with $\gamma < \rho$ for ρ defined in Assumption 2.

THEOREM 2. *Let assumptions* 1′–3′ *hold. Then*

$$\begin{aligned} &x_t \to x^* a.s., \\ &\sqrt{t}(x_t - x^*) \xrightarrow{D} \mathcal{N}(0, V), \end{aligned}$$

where matrix V is a unique positive definite solution of the Lyapunov equation

$$\left(\gamma\zeta\nabla\varphi(x^*) - \frac{I}{2}\right)V + V\left(\gamma\zeta\nabla\varphi(x^*) - \frac{I}{2}\right)^T = (\gamma\zeta)^2 S(x^*). \tag{24}$$

In other words, normalized errors of algorithm (3)–(5) are asymptotically normal with zero mean and covariance matrix V.

Proof. Put $x^* = 0$. Note that as soon as all conditions of Theorem 1 hold

$$x_t \to 0, \qquad s_t - \zeta^{-1}t \to 0,$$

which means that $t\gamma_t - \zeta\gamma \to 0$. The following simple lemma will be useful in further developments.

LEMMA 6. *Let P (v_t) be a random sequence of real numbers, such that $v_t \to 0$ almost surely as $t \to \infty$. Then there exists a deterministic sequence (a_t) such that*

$$a_t \to 0 \quad \textit{and} \quad v_t/a_t \to 0\, a.s.$$

Proof. Let us construct the sequence $w_t = \max\{|v_i|, i \geq t\}$. Obviously, (w_t) is decreasing and $w_t \xrightarrow{P} 0$. Thus there exists a sequence (ϵ_t) such that $\epsilon_t > 0, \epsilon_t \to 0$, and $P(w_t > \epsilon_t) < \epsilon_t$ as $t \to \infty$. So, $w_t/\sqrt{\epsilon_t} \xrightarrow{P} 0$. This means that there is a subsequence t_k of times such that $w_{t_k}/\sqrt{\epsilon_{t_k}} \to 0$ almost surely. Let us define a sequence (a_j) in the following way:

$$a_j = \sqrt{\epsilon_{t_k}} \quad \text{for } t_k \leq j < t_{k+1}.$$

Then we have for all $j \geq 1$

$$|v_j|/a_j \leq w_j/a_j \leq w_{t_k}/\sqrt{\epsilon_{t_k}} \to 0 \quad \text{as } j \to \infty. \qquad \square$$

Theorem 1, along with Lemma 6, yields that there exists a sequence (a_t) of nonrandom positive numbers such that

$$a_t \to 0 \quad \text{and} \quad (\gamma\zeta - t\gamma_t)/a_t \to 0, \qquad x_t/a_t \to 0 \text{ a.s.} \tag{25}$$

Let us define the stopping times

$$\tau_R = \inf\{t : |\gamma\zeta - t\gamma_t| \geq R|a_t|\}, \qquad \sigma_R = \inf\{t : |x_t| \geq R|a_t|\} \tag{26}$$

for $\alpha > 0$ and $\nu = \min(\tau_R, \sigma_R)$. From Lemma 6 and (25) we conclude that for any $\epsilon > 0$ one can choose $R < \infty$ such that

$$P(\nu = \infty) \geq 1 - \epsilon. \tag{27}$$

Consider along with the process (3)–(5) a new linearized process z_t, which is defined by the equation

$$z_t = z_{t-1} - \frac{\gamma\zeta}{t}(\varphi'(0)z_{t-1} + \xi_t), \qquad z_0 = x_0. \tag{28}$$

Asymptotic properties of this process have been completely studied. For example, all of the conditions of the Nevel'son–Khasminskij theorem [6] are satisfied; thus

$$\begin{aligned} &z_t t^{1/2-\alpha} \to 0 \quad \text{for all } \alpha > 0 \quad \text{and} \quad E|z_t|^2 \leq \tfrac{K}{t}, \\ &\sqrt{t}z_t \xrightarrow{D} \mathcal{N}(0, V), \end{aligned} \tag{29}$$

where the matrix V is defined in (24). Hence to prove the assertion of the theorem, it suffices to show asymptotic equivalence of the processes (x_t) and (z_t).

Denote $\Delta_t = x_t - z_t$.

PROPOSITION 2. $\sqrt{t}\Delta_t \xrightarrow{P} 0$.

Proof. For Δ_t, we have from (3) and (28)

$$\Delta_t = \Delta_{t-1} - \frac{\gamma\zeta}{t}\varphi'(0)\Delta_{t-1} + \left(\frac{\gamma\zeta}{t} - \gamma_t\right)\varphi'(0)x_{t-1} + \gamma_t(\varphi'(0)x_{t-1} - \varphi(x_{t-1})) + \left(\frac{\gamma\zeta}{t} - \gamma_t\right)\xi_t \tag{30}$$

$$= \Delta_{t-1} - \frac{\gamma\zeta}{t}\varphi'(0)\Delta_{t-1} + |x_{t-1}|\frac{u_{t-1}}{t} + \left(\frac{\gamma\zeta}{t} - \gamma_t\right)\xi_t, \tag{31}$$

where u_t is an $\mathcal{F}_t$ measurable random variable satisfying

$$|u_t| \le \max\left\{|\varphi'(0)|Ra_t, (\gamma\zeta + Ra_t)\sup_{|x|\le Ra_t}(|\phi(x) - \phi'(0)x|/|x|)\right\} \stackrel{\text{def}}{=} b_t.$$

Note that $\lim_{t\to\infty} b_t = 0$. From Assumption 2′ and the Lyapunov theorem, we conclude that there is a solution $A = A^T > 0$ of the Lyapunov equation

$$\left(\frac{I}{2} - \gamma\zeta\varphi'(0)\right)A + A\left(\frac{I}{2} - \gamma\zeta\varphi'(0)\right)^T = -I.$$

Thus we obtain

$$\gamma\zeta(A^T\varphi'(0) + \varphi'(0)^T A) \ge (1+\beta)A \tag{32}$$

for some $\beta > 0$. Let us put $V_t = \Delta_t^T A\Delta_t$. Using the inequality

$$(a+b+c+d)^2 \le a^2 + 3(b^2+c^2+d^2) + 2a(b+c+d),$$

we obtain from (31) for any $t \le \nu$

$$\begin{aligned}
V_t &\le V_{t-1} + 3|A|(\gamma^2\zeta^2 t^{-2}|\varphi'(0)|^2|\Delta_{t-1}|^2 + |x_{t-1}|^2|u_{t-1}|^2t^{-2} + R^2a_t^2t^{-2}|\xi_t|^2) \\
&\quad + t^{-1}(-(1+\beta)V_{t-1} + 2|\Delta_{t-1}||x_{t-1}||u_{t-1}| + 2(\gamma\zeta - t\gamma_t)\xi_t^T A\Delta_{t-1}) \\
&\le V_{t-1} + K(t^{-2}V_{t-1} + a_{t-1}t^{-2} + a_{t-1}t^{-2}|\xi_t|^2) \\
&\quad + t^{-1}(-(1+\beta)V_{t-1} + 4(|\Delta_{t-1}|^2 + |z_{t-1}|^2)b_{t-1} + 2(\gamma\zeta - t\gamma_t)\xi_t^T A\Delta_{t-1}) \\
&\le V_{t-1}\left(1 - \frac{1+\beta/2}{t}\right) + Ka_{t-1}t^{-2} + a_{t-1}t^{-2}|\xi_t|^2 \\
&\quad + 2|z_{t-1}|^2b_{t-1}/t + 2(\gamma\zeta/t - \gamma_t)\xi_t^T A\Delta_{t-1}
\end{aligned}$$

if t is large enough. And now, taking expectations on both sides, we obtain

$$EV_tI(t<\nu) \le EV_tI(t-1<\nu) \le \left(1 - \frac{1+\beta/2}{t}\right)EV_{t-1}I(t-1<\nu) + o(t^{-2}).$$

Therefore, we get for $W_t = tV_tI(t<\nu)$

$$E(W_t|\mathcal{F}_{t-1}) \le \left(I - \frac{\beta/2}{t-1}\right)W_{t-1} + o(t^{-1}).$$

Hence $EW_t \to 0$ and $\sqrt{t}\Delta_tI(t<\nu) \xrightarrow{P} 0$ for any value of R. Due to the arbitrary choice of ϵ in (27) we obtain the desired proposition. □

4. Algorithm with averaging of trajectories. Let us consider the Polyak–Ruppert algorithm [8], [12] for the stochastic approximation problem:

$$\begin{cases} x_t = x_{t-1} - \gamma_t y_t, \\ \bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i, \qquad x_0 \in R^N, \end{cases} \tag{33}$$

$$y_t = \varphi(x_{t-1}) + \xi_t$$

with the sequence of scalar gain coefficients γ_t defined by (4) and (5).

The first equation of (33) along with (4) and (5) constitutes an accelerated stochastic approximation algorithm that is analogous to that considered in §2. The averaging in (33) ensures the asymptotical optimality of the method (see [8] for details). We impose the following assumptions:

ASSUMPTION 5. There exists a function $U(x) : R^N \to R^+$ such that for some $\kappa > 0, \alpha > 0, \epsilon > 0, L > 0$ and any $x, y \in R^N$, the following conditions hold:

$$\begin{aligned} &U(x) \geq \alpha|x|^2, \\ &|\nabla U(x) - \nabla U(y)| \leq L|x-y|, \\ &U(x^*) = 0, \qquad \nabla U(x)^T\varphi(x) > 0 \quad \text{for } x \neq x^*, \\ &\nabla U(x)^T\varphi(x) \geq \kappa U \quad \text{for } |x - x^*| < \epsilon. \end{aligned}$$

ASSUMPTION 6. There exists a matrix $\varphi'(x^*) > 0$ and $K_\varphi < \infty, 0 < \lambda \leq 1$ such that

$$|\varphi(x) - \varphi'(x^*)(x - x^*)| \leq K_\varphi |x - x^*|^{1+\lambda}.$$

ASSUMPTION 7. $\gamma(t) = \gamma t^{-\mu}$ with $\gamma > 0$ and $(1+\lambda)^{-1} < \mu < 1$.

Comment. In fact, Assumptions 2 and 5 declare the existence of two Lyapunov functions for the system. The probe function U in condition 5 describes the local properties of the function $\varphi(\cdot)$ in the neighborhood of x^*, and V declared in Assumption 2 is, in turn, a "global" one that guarantees the global stability of the system.

THEOREM 3. *If conditions* 1–7 *are satisfied then*

$$\begin{aligned} &\bar{x}_t \to x^* a.s., \\ &\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} \mathcal{N}(0, V), \end{aligned}$$

where

$$V = \varphi'(x^*)^{-1} S(x^*)(\varphi'(x^*)^{-1})^T.$$

Proof. We will verify the assumptions of Theorem 2 in [8]. Assumptions 1, 5, and 6 ensure that conditions 3.1–3.4 of Theorem 2 in [8] hold. It suffices to show that

$$\sum_{i=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty. \tag{34}$$

Note that all of the conditions of Theorem 1 are satisfied; thus

$$\frac{s_t}{t} - \zeta^{-1} \to 0.$$

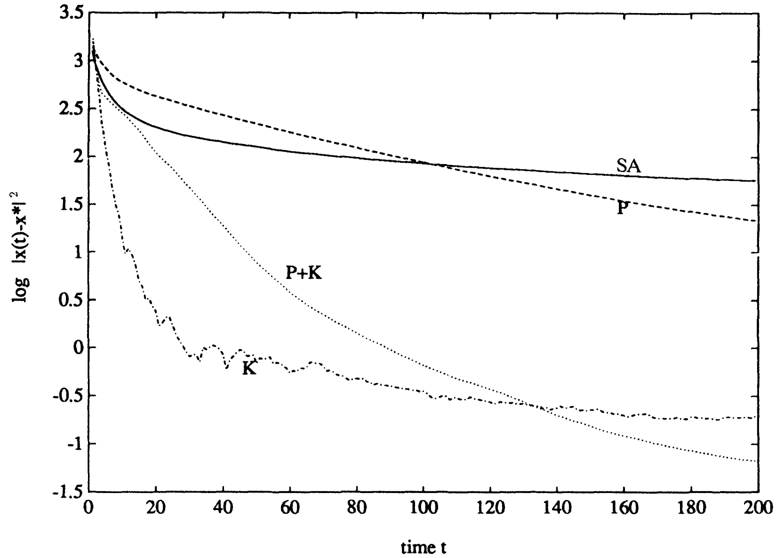

FIG. 1

This means that there are $\alpha > 0$ and $t_\alpha < \infty$ such that $s_t \geq \alpha t$ for $t \geq t_\alpha$. Thus we obtain by Assumption 7

$$\begin{aligned}
\sum_{i=1}^{\infty} \gamma_i^{(1+\lambda)/2} i^{-1/2} &= \sum_{i=1}^{t_\alpha} \gamma_i^{(1+\lambda)/2} i^{-1/2} + \sum_{i=t_\alpha+1}^{\infty} \gamma_i^{(1+\lambda)/2} i^{-1/2} \\
&\leq K + \sum_{i=t_\alpha+1}^{\infty} (i\alpha)^{-\mu(1+\lambda)/2} i^{-1/2} \\
&\leq K + K \sum_{i=t_{\alpha+1}}^{\infty} i^{-1+\alpha'}
\end{aligned}$$

for some $\alpha' > 0$. Hence the series (34) is summable. $\square$

5. Numerical examples. Consider a stochastic approximation problem for the vector field in R^2

$$\varphi(x) = \left(\frac{x_1 - x_1^*}{1 + \sqrt{|x - x^*|}}, \frac{8(x_2 - x_2^*)}{1 + \sqrt{|x - x^*|}} \right)^T$$

with disturbances $\xi_t \in R^2$ that are independent and identically distributed Gaussian random variables with zero mean and covariance

$$S(x^*) = \begin{bmatrix} 1.0 & 0 \\ 0 & 1.0 \end{bmatrix}.$$

The initial error is $x_0 - x^* = (20, 20)^T$.

The trajectories of the *logarithm* of the error variance averaged by 10 samples for the ordinary SA algorithm, the Polyak–Ruppert algorithm (P), and their accelerated versions (K and P + K, respectively) are presented in Fig. 1. First we compare algorithm (3)–(5) with $\gamma(t) = t^{-1}$ to the ordinary stochastic approximation algorithm

$$x_t = x_{t-1} - \frac{\gamma}{t}(\varphi(x_{t-1}) + \xi_t). \tag{35}$$

In this example the accelerated algorithm (K) significantly outperforms the ordinary one (SA). Next we can compare this behavior to that of the Ruppert–Polyak algorithm. ($\gamma(t) = t^{-0.6}$ was arbitrarily chosen for the first equation of the Ruppert–Polyak method (33).) We can see that the Ruppert–Polyak algorithm (P) and its Kesten-like modification (P + K) asymptotically outperform their ordinary counterparts (algorithms without averaging of the trajectories).

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[2] L. Breiman, *Probability*, Addison-Wesley, Reading, MA, 1968.

[3] V. Fabian, *On asymptotically efficient recursive estimation*, Ann. Statist., 6(1978), pp. 854–866.

[4] H. Kesten, *Accelerated stochastic approximation*, Ann. Math. Statist., 29(1958), pp. 41–59.

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

[6] M. B. Nevel'son and R. Z. Khas'minskij, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society, Providence, RI, 1973.

[7] J. Neveu, *Martingales á Temps Discret*, Masson, Paris, 1972.

[8] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM, J. Control Optim., 29(1991), pp. 838–855.

[9] B. T. Polyak and Ya. Z. Tsypkin, *Adaptive estimation algorithms* (*convergence, optimality, stability*), Automation and Remote Control, 40 (1980), pp. 378–389.

[10] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22(1951), pp. 400–407.

[11] H. Robbins and D. Siegmund, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Method in Statistics, J. S. Rustaji, ed., Academic Press, New York, pp. 233–257.

[12] D. Ruppert, *Efficient Estimators from a Slowly Convergent Robbins–Monro Process*, Tech. Rep. N. 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1985.

[13] Ya. Z. Tsypkin, *Foundations of Informational Theory of Identification*, Nauka, Moscow, 1984. (In Russian.)

[14] J. H. Venter, *An extension of the Robbins–Monro procedure*, Ann. Math. Statist., 38(1967), pp. 181–190.

[15] M. Wasan, *Stochastic Approximation*, Cambridge University Press, London, 1970.