The Annals of Statistics
1978, Vol. 6, No. 3, 680–696

STRONG CONVERGENCE OF A STOCHASTIC APPROXIMATION ALGORITHM

BY LENNART LJUNG

Linköping University

Convergence with probability one of a recursive stochastic approximation algorithm is considered. Some extensions of previous results for the Robbins–Monro and the Kiefer–Wolfowitz procedures are given. An inportant feature of the approach taken here is that the convergence analysis can be directly extended to more complex algorithms.

1. Introduction. Stochastic approximation algorithms of different variants have now long been studied in many contexts. In this paper the following particular recursive algorithm will be studied:

$$x(n) = x(n-1) + \gamma(n)[f(x(n-1)) + e(n) + \beta(n)] , \tag{1}$$

where $f(x)$ is the negative gradient of a real valued function $V(x)$, $\{e(n)\}$ is a sequence of random vectors, $\{\beta(n)\}$ is a (possibly random) sequence tending to zero, and $\{\gamma(n)\}$ is a (possibly random) sequence of positive scalars. Algorithm (1) coincides with the one recently analysed by Kushner [7].

This algorithm has obvious relations to the Robbins–Monro procedure [11] which perhaps is the best known stochastic approximation algorithm. The Robbins–Monro procedure is a way of stochastically solving the equation

$$f(x) = 0 ,$$

where to each value x there corresponds a random variable $Y = Y(x)$ with distribution $P(Y(x) \leqq y) = H(y \mid x)$ such that

$$f(x) = \int_{-\infty}^{\infty} y \, dH(y \mid x)$$

is the expectation of Y for given x. The Robbins–Monro procedure for finding the root of $f(x)$ then is

$$x(n) = x(n-1) + \gamma(n)y(n) , \tag{2}$$

where $y(n)$ is a random vector whose distribution function for given $x(1), \cdots, x(n-1), y(1), \cdots, y(n-1)$ is $H(y \mid x(n-1))$. The asymptotic properties of (2) have been studied by many authors, e.g., [1], [2], [11], etc.

If in (1) $\beta(n) = 0$ all n, and $\{e(n)\}$ is a sequence of identically distributed random vectors obeying

$$E[e(n) \mid e(n-1), \cdots, e(1)] = 0 , \tag{3}$$

then algorithm (1) can be put in the framework of (2). However, in many

Received January 1975; revised June 1977.

AMS 1970 *subject classifications.* Primary 62L20; Secondary 93E10.

Key words and phrases. Recursive stochastic algorithms, stochastic approximation.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

applications where (1) is used, the disturbances $\{e(n)\}$ are correlated, which violates (3) and then (1) no longer can be described in terms of (2).

The Kiefer–Wolfowitz procedure [6] for minimization of a function has a similar relationship to (1), and as is further described in [7] and in Section 5 below, the inclusion of the terms $\{\beta(n)\}$ then is essential. Algorithms of the form (1) are also widely used in many applied fields, such as control theory, parameter estimation methods, etc. More general variants of (1) have been analysed by the present author [8], [9], [10] with particular emphasis on control theory applications. The approach in these references is to associate (1) with a deterministic differential equation, in terms of which strong convergence of (1) can be studied. In the study by Kushner [7] a similar idea is pursued, though with an entirely different technique and for convergence in probability.

The conditions that have to be imposed on the algorithm (1) are described in Section 2, while Section 3 contains the basic lemmas of the analysis. The main results about strong convergence of (1) are given in Section 4. Applications of the results to the Robbins–Monro and to the Kiefer–Wolfowitz procedures are treated in Section 5. As remarked above, algorithm (1) is just a simple special case of the algorithms studied in [10]. In Section 6 extensions of the convergence results to these more general algorithms are described.

2. General assumptions. A main concern of this paper is to prove convergence with probability one (w.p. 1) of (1) into the set

$$D_S = \{x \mid f(x) = 0\}\,. \tag{4}$$

To do this, certain assumptions have to be imposed on (1) and these conditions will now be stated.

The following general assumptions will be used throughout the paper.

A1: $V(x)$ is a twice continuously differentiable function on R^n and $(d/dx)V(x) = -f(x)$ (column vector).

A2: $\{x \mid V(x) \leqq C\}$ is compact for all $C < \sup V(x)$.

A3: The set D_S consists of a finite number of isolated connected sets.

A4: $\{\gamma(n)\}$ is a (possibly random) sequence of positive scalars, such that $\gamma(n) \to 0$ and $\sum_1^\infty \gamma(n) = \infty$ (w.p. 1).

A5: $\{\beta(n)\}$ is a sequence of R^n valued random variables, such that $\beta(n) \to 0$ w.p. 1 as $n \to \infty$.

By A3 it is meant that D_S can be written as a union of connected sets, such that each of these sets has a strictly positive distance to the union of the other sets. The assumption A3 can be replaced by

A3′: The function $V(x)$ is n times continuously differentiable, where n is $\dim x$,

as is explained after Lemma 1.

In the main lemma the following two assumptions about the behavior of (1) and about the properties of the sequences $\{e(n)\}$ and $\{\gamma(n)\}$ are introduced.

B1: Let $z(n)$ be defined by $z(n) = z(n-1) + \gamma(n)(e(n) - z(n-1))$; $z(0) = 0$. Then $z(n) \to 0$ w.p. 1 as $n \to \infty$.

B2: $\liminf_{n\to\infty} |x(n)| < \infty$ w.p. 1.

Notice that assumption B2 as such does not preclude that a subsequence of $\{x(k)\}$ may tend to infinity.

These conditions are fairly implicit, and more easily checked ones are desirable. Several ways of verifying B1 and B2 are possible, and in two lemmas it will be shown that, e.g., the following conditions ensure B1, B2.

C1: $e(n)$ has an innovations representation

$$e(n) = \textstyle\sum_{k=0}^{n} h(n, k)\nu(k)$$

where $\{\nu(k)\}$ are independent random vectors with zero mean values and unit covariance matrices, and such that $E|\nu(k)|^{2p} < C$ for some integer p. Furthermore $|h(n, k)| < \alpha_n \lambda^{n-k}$ where $\{\alpha_n\}$ is nondecreasing and $\lambda < 1$.

C2: $\{\gamma(n)\}$ is a deterministic, nonincreasing sequence such that

$$\limsup_{n\to\infty} \left| \frac{1}{\gamma(n)} - \frac{1}{\gamma(n-1)} \right| < \infty .$$

Moreover

$$\textstyle\sum_1^\infty (\gamma(n)\alpha_n^2)^p < \infty$$

where α_n and p are defined in C1.

C3: $\sup |(d/dx)f(x)| < \infty$.

C4: $D_{\bar\delta} = \{x \mid |f(x)| \leqq \bar\delta\}$ is compact and nonempty for some $\bar\delta > 0$.

The reason for including a sequence $\{\alpha_n\}$ in C1 that may tend to infinity is to allow treatment of schemes like the Kiefer–Wolfowitz procedure, where the variance of the disturbances increases to infinity.

Finally, it will be shown that the set D_S defined by (4) into which the estimates converge may be replaced by the smaller set

$$D_M = \left\{ x \mid f(x) = 0 \text{ and the matrix } \frac{d}{dx} f(x) \text{ is negative semidefinite} \right\} . \tag{5}$$

This requires the following additional assumptions.

D1: The sequences $\{\beta(n)\}$ and $\{\gamma(n)\}$ are deterministic sequences. Furthermore, the sequence $\{e(n)\}$ consists of independent random vectors with zero mean values and covariance matrices obeying

$$c_2 \cdot \alpha_n \cdot I \leqq Ee(n)e(n)^T \leqq c_1 \cdot \alpha_n \cdot I$$

for some strictly positive scalars c_1, c_2 and where $\{\alpha_n\}$ is a nondecreasing sequence of strictly positive scalars. Moreover,

$$E|e(n)|^4/(E|e(n)|^2)^2 \leqq c_3 .$$

D2: The set D_S consists of isolated points.

3. Basic lemmas. Convergence of (1) w.p. 1 follows from the following main lemma.

LEMMA 1. *Assume* A1 *to* A5 *and* B1 *to* B2. *Then*

$$x(n) \to D_S = \{x \mid f(x) = 0\} \quad \text{w.p. } 1 \qquad \textit{as} \quad n \to \infty .$$

PROOF. Let Ω^* be a subspace of the sample space such that

$$\Omega^* = \{\text{B1 holds}\} \cap \{\text{B2 holds}\} \cap \{\text{A4 holds}\} \cap \{\text{A5 holds}\} .$$

Clearly $P(\Omega^*) = 1$. Consider from now on a fixed realization $\omega^* \in \Omega^*$ and let us study the sequence $\{x(k)\} = \{x(k, \omega^*)\}$. We shall throughout suppress the argument ω^*, on which most of the variables (including subsequences) below depend. In view of B2 there exists a cluster point x^* to $\{x(k)\}$. Let n_k be a subsequence such that $x(n_k) \to x^*$. Suppose that $|f(x^*)| = \delta^* > 0$. ("$|\cdot|$" denotes here, as everywhere else, the Euclidean norm.) From (1) we obtain directly

$$\begin{aligned} x(j) &= x(n_k) + \textstyle\sum_{n_k+1}^{j} \gamma(k)e(k) + \sum_{n_k+1}^{j} \gamma(k)\beta(k) + \sum_{n_k+1}^{j} \gamma(k)f(x(k-1)) \\ &= x(n_k) + S_1(n_k, j) + S_2(n_k, j) + f(x^*) \textstyle\sum_{n_k+1}^{j} \gamma(k) + R(n_k, j, x^*) , \end{aligned} \tag{6}$$

where

$$\begin{aligned} S_1(n_k, j) &= \textstyle\sum_{n_k+1}^{j} \gamma(k)e(k) \\ S_2(n_k, j) &= \textstyle\sum_{n_k+1}^{j} \gamma(k)\beta(k) \\ R(n_k, j, x^*) &= \textstyle\sum_{n_k+1}^{j} \gamma(k)(f(x(k-1)) - f(x^*)) . \end{aligned}$$

Now suppose that $n_k \leqq j \leqq m(n_k, \tau)$ where $m(n_k, \tau)$ is such that

$$\textstyle\sum_{j=n_k+1}^{m(n_k,\tau)} \gamma(j) \nearrow \tau < \infty \qquad \text{as} \quad n_k \to \infty . \tag{7}$$

The number $m(n_k, \tau)$ is finite for any k and any $\tau < \infty$ due to A4. Then we claim that

$$\begin{aligned} &S_i(n_k, j) \to 0 \quad \text{uniformly in} \quad n_k \leqq j \leqq m(n_k, \tau) \\ &\text{for fixed} \quad \tau \quad \text{as} \quad n_k \to \infty . \end{aligned} \tag{8}$$

PROOF OF CLAIM.

$i = 2$: $|S_2(n_k, j)| \leqq \max_{n_k \leqq i} |\beta(i)| \cdot \tau \to 0$ as $n_k \to \infty$ according to A5.
$i = 1$: From the definition of $z(n)$ in B1 we have

$$z(j) = z(n_k) + S_1(n_k, j) - \textstyle\sum_{n_k+1}^{j} \gamma(k)z(k-1)$$

or

$$|S_1(n_k, j)| \leqq |z(j)| + |z(n_k)| + \tau \max_{n_k \leqq i} |z(i)| \to 0 \qquad \text{as} \quad n_k \to \infty$$

according to B1.

Let $B(x^*, \rho) = \{x \mid |x - x^*| < \rho\}$ and $C^* = \max\{\sup_{x \in B(x^*,1)} |(d/dx)f(x)|, 1\}$. Choose from now on a fixed sphere $B^* = B(x^*, \rho^*)$ with

$$0 < \rho^* < \min(1, \delta^*/8C^*) .$$

We recall that $\delta^* = |f(x^*)|$. The reason for this particular choice of ρ^* will be

clear below. Clearly,

$$|R(n_k, j, x^*)| \leqq \max_{n_k+1 \leqq i \leqq j} |f(x(i-1)) - f(x^*)| \cdot \tau .$$

Choose from now on

$$\tau = \tau^* = \rho^*/2\delta^* \quad (\leqq 1/(16C^*))$$

and denote $m(n_k, \tau^*) = m_k^*$. From (6) we have that for $j \leqq m_k^*$,

$$(9) \qquad |x(j) - x^*| \leqq |x(n_k) - x^*| + |S_1(n_k, j)| + |S_2(n_k, j)| + |f(x^*)| \textstyle\sum_{n_k+1}^{j} \gamma(k) + \tau^* \max_{n_k+1 \leqq i \leqq j} |f(x(i-1)) - f(x^*)| .$$

According to (8) there exists an integer K_1, such that for $k > K_1$

$$|S_i(n_k, j)| < \rho^*/8 \qquad i = 1, 2 , \quad n_k \leqq j \leqq m_k^*$$

and

$$|x(n_k) - x^*| < \rho^*/8 . \quad \text{(Then in particular } x(n_k) \in B^*.)$$

If $x(i) \in B^*$ for $i = n_k, \cdots, j-1$, then

$$\max_{n_k+1 \leqq i \leqq j} |f(x(i-1)) - f(x^*)| < C^*\rho^*$$

and

$$|x(j) - x^*| \leqq \rho^*/8 + \rho^*/8 + \rho^*/8 + \delta^*\tau^* + C^*\rho^*\tau^* < \rho^*/8 + \rho^*/4 + \rho^*/2 + \rho^*/16 \leqq \rho^* .$$

Hence also $x(j) \in B^*$. By induction it follows that

$$x(j) \in B^* \qquad n_k \leqq j \leqq m_k^* ; \quad k > K_1 .$$

In particular we have

$$(10\text{a}) \qquad x(m_k^*) - x^* = \tau^* f(x^*) + R_2(n_k)$$

where

$$(10\text{b}) \qquad |R_2(n_k)| \leqq |x(n_k) - x^*| + |S_1(n_k, m_k^*)| + |S_2(n_k, m_k^*)| + C^*\rho^*\tau^* + \delta^*|\tau^* - \textstyle\sum_{n_k+1}^{m_k^*} \gamma(j)| .$$

Now (subscript x denoting derivative)

$$(11) \qquad \begin{aligned} V(x(m_k^*)) &= V(x^*) + (x(m_k^*) - x^*)^T V_x(x^*) \\ &\quad + \tfrac{1}{2}(x(m_k^*) - x^*)^T V_{xx}(\xi)(x(m_k^*) - x^*) \\ &= V(x^*) + \tau^* f(x^*)^T V_x(x^*) \\ &\quad + \tfrac{1}{2}(x(m_k^*) - x^*)^T V_{xx}(\xi)(x(m_k^*) - x^*) + R_2^T(n_k) V_x(x^*) . \end{aligned}$$

Now $V_x(x^*) = -f(x^*)$ and $V_{xx}(\xi) = -f_x(\xi)$; $\xi \in B^*$. Hence

$$(12) \qquad \begin{aligned} &|R_2^T(n_k) f(x^*) + \tfrac{1}{2}(x(m_k^*) - x^*)^T f_x(\xi)(x(m_k^*) - x^*)| \\ &\qquad \leqq \delta^*\{|x(n_k) - x^*| + |S_1(n_k, m_k^*)| + |S_2(n_k, m_k^*)| \\ &\qquad\quad + C^*\rho^*\tau^* + \delta^*|\tau^* - \textstyle\sum_{n_k}^{m_k^*} \gamma(j)|\} + C^*\rho^{*2} . \end{aligned}$$

Since $x(n_k)$ tends to x^* and according to (7) and (8), there exists an integer K^*, such that for $k > K^*$

$$|x(n_k) - x^*| < \delta^*\tau^*/32$$
$$|S_1(n_k, m_k^*)| + |S_2(n_k, m_k^*)| < \delta^*\tau^*/32$$
$$|\tau^* - \textstyle\sum_{n_k+1}^{m_k^*} \gamma(j)| < \tau^*/16 .$$

Then the RHS of (12) is less than (recall that $\rho^* = 2\delta^*\tau^*$)

$$\delta^* \left(\frac{\delta^*\tau^*}{32} + \frac{\delta^*\tau^*}{32} + \frac{\delta^*\tau^*}{8} + \frac{\delta^*\tau^*}{16}\right) + C^* \cdot 4(\delta^*\tau^*)^2$$
$$\leqq \tfrac{1}{4}(\delta^*)^2 \cdot \tau^* + 4C^*(\delta^*)^2\tau^* \cdot \frac{C^*}{16} = \tfrac{1}{2}\tau^*(\delta^*)^2$$

where the inequality follows from $\tau^* < 1/(16C^*)$.

Hence,

$$V(x(m_k^*)) < V(x^*) - \tau^*|f(x^*)|^2 + (\delta^*)^2\tau^*/2 < V(x^*) - \tau^*|f(x^*)|^2/2 \tag{13}$$

for $k > K^*$. This holds for all cluster points x^* such that

$$|f(x^*)| > 0 .$$

Therefore, if x^* is any cluster point with $V(x^*) = V^*$ and $|f(x^*)| = \delta^* > 0$ then (13) implies that $x(j)$ belongs infinitely often (namely for $j = m_k^*$, each $k > K^*$) to

$$D^* = \{x \mid V(x) \leqq V^* - \tau^*(\delta^*)^2/2\}$$

which is compact according to assumption A2. Hence there is at least one cluster point in D^*, and if this does not belong to D_S we may repeat the argument.

Let $\bar{V} = \inf V(x)$ where the infimum is taken over the cluster points of $\{x(k)\}$. Since the set of cluster points is closed, it follows that there is a cluster point $\bar{x}$ with $V(\bar{x}) = \bar{V}$. Obviously $\bar{x} \in D_S$; otherwise we could use (13) to infer the existence of a cluster point with still lower value of V. Similarly, all cluster points $\bar{x}$ with $V(\bar{x}) = \bar{V}$ must belong to D_S.

We shall now proceed to show that there can be no cluster point outside D_S. Such a point x^0 would obviously yield $V(x^0) = V^0 > \bar{V}$. Then for some sufficiently small $d > 0$, $V(x(k)) > \bar{V} + d$ infinitely often. Since V is continuous we can according to A3 choose this d so small that the compact set

$$\bar{D} = \left\{x \mid \bar{V} + \frac{d}{2} \leqq V(x) \leqq \bar{V} + d\right\} \tag{14}$$

has no point in common with D_S.

Since the "step size" $|x(n+1) - x(n)|$ tends to zero when $x(n) \in \{x \mid V(x) \leqq \bar{V} + d\}$, it follows that $x(k)$ would be inside $\bar{D}$ and cross it infinitely often "uphill" and "downhill." Consider now a subsequence of "upcrossings" of $\bar{D}$. Let $\{x(n_k')\}$ be defined as follows:

$$V(x(n_k' - 1)) < \bar{V} + \frac{d}{2}; \qquad V(x(n_k')) \geqq \bar{V} + \frac{d}{2};$$

and assume that the smallest positive s for which $x(n_k' + s) \notin \bar{D}$ yields $V(x(n_k' + s)) > V + d$. That is, n_k' is the kth time the iterate is $\geqq \bar{V} + d/2$ just after it is $< \bar{V} + d/2$ and also where the sequence $\{x_i\}$ does not enter $\{x \mid V(x) < \bar{V} + d/2\}$ until it first leaves $\bar{D}$.

Let $\tilde{x}$ be a cluster point to the sequence $x(n_k')$ and let $x(n_k'')$ be a subsequence of this sequence tending to $\tilde{x}$. Clearly, $V(\tilde{x}) = \bar{V} + d/2$; and let $|f(\tilde{x})|$ be denoted by $\tilde{\delta}$, which is strictly greater than zero. Now denote

$$\max\left\{\sup_{x \in B(\tilde{x},1)} \left|\frac{d}{dx} f(x)\right|, 1\right\} = \tilde{c}$$

and let $\tilde{\rho} > 0$ be less than $\min(1, \tilde{\delta}/8\tilde{c}, d/2)$ and so small that $B(\tilde{x}, \tilde{\rho})$ has no point in common with the set $\{x \mid V(x) \geqq \bar{V} + d\}$. Let $\tilde{\tau} = \tilde{\rho}/2\tilde{\delta}$. Now, apply result (13) to the cluster point $\tilde{x}$, which gives, for sufficiently large k,

$$V(x(m(n_k'', \tilde{\tau}))) < \bar{V} + \frac{d}{2} - \tilde{\rho}\tilde{\delta}/4 .$$

From the analysis preceding equation (10a) it also follows that $x(j) \in B(\tilde{x}, \tilde{\rho})$; $n_k'' \leqq j \leqq m(n_k'', \tilde{\tau})$. Consequently, the sequence $\{x_i, i \geqq n_k''\}$ will not enter the set $\{x \mid V(x) \geqq \bar{V} + d\}$ when it first leaves $\bar{D}$, which is a contradiction to the definition of n_k'' as a sequence of upcrossings. Therefore $\bar{D}$ will not be crossed upwards infinitely many times, and since there is a cluster point in D_S, the sequence $\{x(k)\}$ will remain in any arbitrarily small neighborhood of D_S for sufficiently large k. There are consequently no cluster points outside D_S, and the proof of Lemma 1 is complete. □

Note that in the proof a fixed realization is considered throughout. Therefore the conclusion of the theorem holds for *any* sequences $\{e(n)\}$, $\{\gamma(n)\}$, $\{\beta(n)\}$ (regarded as realizations of stochastic processes or not) such that B1, B2, A4 and A5 hold.

REMARK. Notice that assumption A3 is used only to infer the existence of the set $\bar{D}$ in (14) disjoint from D_S. For a general set D_S but under the additional assumption A3′ it follows from the Morse and Sard theorem that the set $S = \{z \mid V(x) = z, x \in D_S \text{ and } z \leqq V^0\}$ is a compact set of measure zero. This also implies that a set $\bar{D}$ can be chosen disjoint from D_S. Notice also that it follows from the proof that $\{x(n)\}$ cannot oscillate between the isolated sets in D_S.

In order to verify assumption B1 certain conditions on the sequences $\{\gamma(n)\}$ and $\{e(n)\}$ have to be introduced. The recursion in B1 can be solved by

$$z(n) = \textstyle\sum_{k=1}^{n} \psi(n, k) e(k) \tag{15}$$

where

$$\psi(n, k) = \gamma(k) \textstyle\prod_{i=k+1}^{n} (1 - \gamma(i)) \quad k < n ; \qquad \psi(n, n) = \gamma(n) .$$

If $\{e(n)\}$ is a sequence of independent random variables, many approaches to prove convergence of $z(n)$ are available, but we shall not pursue that here (cf.

[8]). The fairly common choice $\gamma(n) = 1/n$ gives $\psi(n, k) = 1/n$ and then various "laws of large numbers" can be applied to (15). In Cramér and Leadbetter [3], pages 94–96, the following variant is given:

Let $\gamma(n) = 1/n$ and assume $Ee(n) = 0$, and that

$$|Ee(n)e(m)| \leqq K \frac{n^p + m^p}{1 + |n - m|^q} \quad 0 \leqq 2p < q < 1 . \tag{16}$$

Then $z(n) \to 0$ w.p. 1 as $n \to \infty$. (In [3] this result is given for continuous time stochastic processes, but the proof is valid also for discrete time processes.)

Another result that appears to be useful in applications is the following.

LEMMA 2. *Assume* C1 *and* C2. *Then* B1 *holds.*

PROOF. Let

$$L = \limsup_{n\to\infty} \left| \frac{1}{\gamma(n)} - \frac{1}{\gamma(n-1)} \right| .$$

The moments of $z(n)$ are estimated in the following claim.

CLAIM. If $L \leqq 1$ then, for some positive constant C_r

$$E|z(n)|^r < C_r(\alpha_n)^r(\gamma(n))^{r/2} ; \quad 1 < r \leqq 2p .$$

The claim is shown by straightforward calculation of the moments of sums like

$$T_k = \textstyle\sum_{i=n_k}^{n_{k+1}} \gamma(i)\Gamma(n_{k+1}, i)e(i) \qquad \text{where} \quad \lim_{k\to\infty} \sum_{i=n_k}^{n_{k+1}} \gamma(i) = \tau > 0$$

and then linking such estimates together using Minkowski's inequality. The formal proof is given in [8].

With this claim, Chebyshev's inequality can be applied to yield

$$P(|z(n)| > \varepsilon) \leqq \frac{E|z(n)|^{2p}}{\varepsilon^{2p}} \leqq \frac{C_{2p}\gamma^p(n)\alpha_n{}^{2p}}{\varepsilon^{2p}}$$

and

$$\textstyle\sum_{n=1}^{\infty} P(|z(n)| > \varepsilon) \leqq \frac{C_{2p}}{\varepsilon^{2p}} \sum_{n=1}^{\infty} \gamma^p(n)\alpha_n{}^{2p} < \infty .$$

The Borel–Cantelli lemma now assures

$$z(n) \to 0 \qquad \text{as} \quad n \to \infty \quad \text{w.p. } 1 . \qquad \square$$

If $L > 1$ we take

$$z(n) = z(n-1) + L\gamma(n)\left(\frac{1}{L}e(n) - \frac{1}{L}z(n-1)\right)$$

which, according to Lemma 1 ($V(z) = (1/2L)z^2$) converges w.p. 1 to zero if

$$\tilde{z}(n) = \tilde{z}(n-1) + L\gamma(n)\left(\frac{1}{L}e(n) - \tilde{z}(n-1)\right)$$

does. But this latter algorithm converges w.p. 1 according to the first part of this proof. $\square$

The reason for assumption B2 is that it very well may happen that the sequence $\{x(n)\}$ tends to infinity even when assumptions A and B1 are satisfied. Further conditions on the functions $V(x)$ and $f(x) = -(d/dx)V(x)$ have to be introduced to ensure B2.

LEMMA 3. *Assume* A1 *to* A5, B1 *and* C3 *to* C4. *Then* B2 *holds.*

PROOF. Consider as in the proof of Lemma 1 a fixed realization $\omega^* \in \Omega^*$. Let $\bar{\delta} > 0$ be a constant, such that C4 holds and denote $\bar{c} = \sup |f_x|$, which is finite according to C3. Introduce, analogously to the proof of Lemma 1, $\bar{\rho} = \min(1, \bar{\delta}/8\bar{c})$ and $\bar{\tau} = \bar{\rho}/2\bar{\delta}$. Denote $m(k, \bar{\tau}) = \bar{m}_k$. According to (7) and (8) there exists an integer $\bar{K}$, such that for $k > \bar{K}$

$$|S_1(k, \bar{m}_k)| + |S_2(k, \bar{m}_k)| < \bar{\delta}\bar{\tau}/32$$
$$|\bar{\tau} - \textstyle\sum_{k+1}^{\bar{m}_k} \gamma(j)| < \bar{\tau}/16$$

(cf. expressions below equation (12)). This implies, following the arguments leading to (13) and taking $x^* = x(k)$ that if for some $k > \bar{K}$ $x(k) \notin D_{\bar{\delta}}$, then

$$V(x(\bar{m}_k)) < V(x(k)) - \bar{\tau}\bar{\delta}^2/2 .$$

Therefore, if $|x(k)| \to \infty$, $x(k)$ would remain outside the compact set $D_{\bar{\delta}}$ from a certain K_1 on. With $K_0 = \max(K_1, \bar{K})$ and $n_i = m(n_{i-1}, \tau)$; $n_0 = K_0$ we then would have

$$V(x(n_j)) < V(x(K_0)) - j\bar{\tau}\bar{\delta}^2/2$$

which would imply that $V(x) \to -\infty$. This is impossible since V is bounded from below, according to A2. □

The set D_S consists both of local minima, local maxima and saddle points of V. In fact, as might be expected only the local minima are possible convergence points as shown in the following lemma.

LEMMA 4. *Assume* A1 *to* A5, B1 *and* D1 *and that* $x(n) \to x^*$ *on a set of positive measure as* $n \to \infty$. *Then* $f(x^*) = 0$ *and all eigenvalues of the matrix*

$$\left.\frac{d}{dx} f(x)\right|_{x=x^*}$$

have nonpositive real parts.

REMARK. Since $f(x) = -(d/dx)V(x)$, the matrix $-(d/dx)f(x)|_{x=x^*}$ is the second derivative matrix (the Hessian) of V at x^*. The condition is that this should be positive semidefinite.

PROOF. Let $x(n) \to x^*$ on Ω^*, with $P(\Omega^*) > 0$. Denote $f_x(x^*) = A$. Then

$$f(x) = f(x^*) + A(x - x^*) + g(x - x^*) \tag{17}$$

where

$$g(x)/|x| \to 0 \quad \text{as} \quad x \to 0 . \tag{18}$$

It follows directly from the proof of Lemma 1 that $f(x^*) = 0$. Suppose that the assertion of the lemma is not true, i.e., that at least one eigenvalue of A is positive.

Let L be a left eigenvector for this eigenvalue:

$$LA = \mu L\,; \quad \mu > 0\,.$$

Introduce the following notation:

$$L(x(n) - x^*) = y_n\,, \qquad L\beta(n) = \bar{\beta}_n\,, \qquad Lg(x(n) - x^*) = \bar{g}_n\,, \qquad \bar{e}_k = Le(k)\,.$$

Then $E(\bar{e}_k)^2 = L[Ee(k)e(k)^T]L^T \equiv \bar{\alpha}_k$.

Multiplying (1) by L from the left gives, using (17),

$$y_k = y_{k-1} + \gamma(k)[\mu y_{k-1} + \bar{e}_k + \bar{\beta}_k + \bar{g}_{k-1}]\,. \tag{19}$$

Solving (19) from $k = n + 1$ to $k = m$ gives

$$y_m = \Gamma(m, n)[y_n + f(m, n) + B(m, n) + G(m, n)] \tag{20}$$

where

$$\begin{aligned} \Gamma(m, n) &= \textstyle\prod_{n+1}^{m}(1 + \mu\gamma(k)) \\ f(m, n) &= \textstyle\sum_{n+1}^{m}\gamma(k)\Gamma(k, n)^{-1}\bar{e}_k \\ B(m, n) &= \textstyle\sum_{n+1}^{m}\gamma(k)\Gamma(k, n)^{-1}\bar{\beta}_k \\ G(m, n) &= \textstyle\sum_{n+1}^{m}\gamma(k)\Gamma(k, n)^{-1}\bar{g}_{k-1}\,. \end{aligned} \tag{21}$$

Let

$$A(m, n)^2 \equiv Ef(m, n)^2 = \textstyle\sum_{n+1}^{m}\gamma(k)^2\Gamma(k, n)^{-2}\bar{\alpha}_k\,.$$

Now, choose $m = \bar{m}(n)$ so large that

$$A(\bar{m}(n), n)\Gamma(\bar{m}(n), n) \geqq 1\,. \tag{22}$$

This is possible, since $\Gamma(m, n)$ tends to infinity as $m \to \infty$ for any fixed n, according to (21), (A4) and to the fact that $\mu > 0$.

Now, the random variable

$$\xi_n \equiv A(\bar{m}(n), n)^{-1}f(\bar{m}(n), n)$$

has zero mean, unit variance and fourth moment

$$\begin{aligned} E\xi_n^4 &= [3\textstyle\sum_{k_1,k_2=n+1}^{\bar{m}(n)}\gamma(k_1)^2\gamma(k_2)^2\Gamma(k_1, n)^{-2}\Gamma(k_2, n)^{-2}E\bar{e}_{k_1}^2\bar{e}_{k_2}^2]A(\bar{m}(n), n)^{-4} \\ &= 3 + [3\textstyle\sum_{k=n+1}^{\bar{m}(n)}\gamma(k)^4(E\bar{e}_k^4 - \bar{\alpha}_k^2)]A(\bar{m}(n), n)^{-4} \leqq 3(1 + C_3)\,. \end{aligned}$$

The last inequality follows from Schwarz' inequality since

$$E\bar{e}_k^4 - \bar{\alpha}_k^2 \leqq C_3\bar{\alpha}_k^2$$

according to assumption D1.

Since ξ_n has unit variance and bounded fourth moments there exist constants $C_1 > 0$, $C_2 > 0$, independent of n, such that

$$P(|\xi_n| \geqq C_1) \geqq C_2\,.$$

The random variable ξ_n is constructed from $\{e_k, n \leqq k \leqq \bar{m}(n)\}$. It is therefore

independent of $\xi_{n'}$ for $n' > \bar{m}(n)$. The second Borel–Cantelli lemma can therefore be applied to an infinite sequence of independent ξ_{n_j}, giving that

$$|\xi_n| \geqq C_1 > 0 \quad \text{for infinitely many } n \text{ w.p. 1, i.e., in particular on } \Omega^* . \tag{23}$$

Now from (20), with $\bar{m} = \bar{m}(n)$

$$y_{\bar{m}} = A(\bar{m}, n)\Gamma(\bar{m}, n)H(\bar{m}, n) \tag{24}$$

where

$$H(\bar{m}, n) = A(\bar{m}, n)^{-1}[y_n + B(\bar{m}, n) + G(\bar{m}, n)] + \xi_n .$$

Since we assume that $y_{\bar{m}(n)}$ tends to zero on Ω^*, (24) implies according to (22) that

$$H(\bar{m}(n), n) \to 0 \quad \text{on} \quad \Omega^* \quad \text{as} \quad n \to \infty . \tag{25}$$

Since ξ_n is independent of y_n and of $B(\bar{m}, n)$, it follows from (23) that also

$$|\xi_n + A(\bar{m}, n)^{-1}(y_n + B(\bar{m}, n))|$$

is greater than a strictly positive constant i.o. w.p. 1. Therefore (25) would imply that $A(\bar{m}, n)^{-1}G(\bar{m}, n)$ does not tend to zero on Ω^*. Consequently, for some subsequence n_k (with $\bar{m}_k = \bar{m}(n_k)$)

$$\begin{aligned} 0 < C &< A(\bar{m}_k, n_k)^{-1}G(\bar{m}_k, n_k) \\ &\leqq A(\bar{m}_k, n_k)^{-1} \textstyle\sum_{n_k+1}^{\bar{m}_k} \gamma(j)\Gamma(j, n_k)^{-1}|\bar{g}_{j-1}| \\ &\leqq \frac{1}{\mu} A(\bar{m}_k, n_k)^{-1} \max_{n_k \leq j} |\bar{g}_j| , \end{aligned}$$

where we in the last inequality used the fact that

$$\textstyle\sum_n^\infty \gamma(j)\Gamma(j, n)^{-1} = \frac{1}{\mu} \qquad \text{all} \quad n .$$

Moreover, from (18)

$$|\bar{g}_j| \leqq r(x(j) - x^*) \cdot |x(j) - x^*|$$

where $r(z) \to 0$ as $z \to 0$. Hence, on Ω^* as $x_j \to x^*$, it follows that

$$A(\bar{m}_k, n_k)^{-1} \max_{n_k \leq j} |x(j) - x^*| \to \infty \qquad \text{as} \quad k \to \infty .$$

Let the maximum be attained for $j = r_k$. Since $A(\bar{m}, n)$ is decreasing as n increases it follows that ($s_k = \bar{m}(r_k)$)

$$A(s_k, r_k)^{-1}|x(r_k) - x^*| \to \infty .$$

According to assumption D1 a full rank random vector, independent of previous data is added to $x(n)$ for each n. Therefore, in a sufficiently small neighborhood of x^* the distribution of $x(n)$ will be nondegenerate. It consequently follows that

$$A(s_k', r_k')^{-1}y_{r_k'} \to \infty \qquad \text{as} \quad k \to \infty$$

for some subsequence $\{r_k'\}$ of $\{r_k\}$. But if this is the case, this term would

dominate $H(\bar{m}, n)$, which violates (25), and we have arrived at a contradiction to the assumption that $f_x(x^*)$ has a positive eigenvalue. □

REMARK. It may be of interest to comment somewhat on the role of the independence assumption about $\{e(n)\}$ in the proof. It is used on two occasions: firstly, in the calculation of the variance and fourth moment of ξ_n. It is clear that under assumption C1 with some restrictions on $h(\cdot, \cdot)$ similar results would be obtained. Secondly, the independence assumption is used in the application of the second Borel–Cantelli lemma and to infer that ξ_n and y_n are independent. Again, under assumptions C1 and C2 it can be shown that "the part" of ξ_n that has its origin in innovations $\nu(k)$ for $k < n$ tends to zero w.p. 1. Therefore this part can be treated separately, and the independence property of the rest can be used in the same way as in the proof above.

4. Main results. The lemmas of the previous section can be combined into several results. It should be noticed that, in addition, Lemmas 1 and 2 are results of independent interest. Two theorems will now be given concerning convergence of (1).

THEOREM 1. *Assume* A1 *to* A5 *and* C1 *to* C4. *Then* $x(n) \to D_S$ w.p. 1 *as* $n \to \infty$.

PROOF. Follows from Lemmas 1 to 3.

THEOREM 2. *Assume* A1 *to* A5, C1 *to* C4 *and* D1 *to* D2. *Then* $x(n)$ *tends to a point in* D_M w.p. 1 *as* $n \to \infty$ (D_M *defined by* (5)).

PROOF. It follows from Theorem 1 that $x(n)$ converges to D_S w.p. 1 and, as remarked after Lemma 1, $\{x(n)\}$ cannot oscillate between isolated points in D_S. Therefore, except on a set of measure zero, $x(n)$ will converge to a point in D_S. Obviously D_S consists of at most a denumerable number of points. Any such point to which $x(n)$ converges on a set of positive measure must satisfy the conditions of Lemma 4. This concludes the proof of Theorem 2.

Now, if $V(x)$ is such that C3 or C4 do not hold it might happen that $x(n)$ tends to infinity. This can be seen from the following simple example.

EXAMPLE 1. Let $V(x) = \frac{1}{4}x^4$ and $e(n) = 0$ $n \geqq 2$, $\gamma(n) = 1/n$, $\beta(n) = 0$ all n. Then if $x(0) = 0$, $x(n) = x(n-1) + (1/n)(-x(n-1)^3)$; $n \geqq 2$; $x(1) = e(1)$; clearly, $x(n)$ will tend to infinity if $|e(1)| > 2$.

However, in any application of the algorithm (1) this will certainly be prevented somehow. A very straightforward idea is to project the estimate $x(n)$ into a compact set D_1 if it is outside another set D_2. Then (1) takes the modified form

$$x(n) = [x(n-1) + \gamma(n)(f(x(n-1)) + e(n) + \beta(n)]_{D^2}^{D_2} \tag{26}$$

where

$$\begin{aligned}[Z]_{D_1}^{D_2} &= Z \quad \text{if} \quad Z \in D_2 \\ &= \text{some value in} \quad D_1, \quad \text{if} \quad Z \notin D_2\end{aligned}$$

and where D_1, D_2 are compact sets such that $D_1 \subset D_2$.

For the modified algorithm (26) obviously B2 holds. However, Lemma 1 cannot be directly applied because of the modification of (1). The following holds though.

THEOREM 3. *Consider the modified algorithm* (26). *Assume* A1 *to* A5, B1, *and that*

(i) $\sup_{x \in D_1} V(x) < \inf_{x \notin D_2} V(x)$,
(ii) $D_S \subset D_1$ (D_S *defined by* (4)).

Then $x(n) \to D_S$ w.p. 1 *as* $n \to \infty$.

PROOF. Let $\sup_{x \in D_1} V(x) = V_1$ and $\inf_{x \notin D_2} V(x) = V_2$ and introduce

$$\tilde{D} = \left\{ x \mid V(x) \leqq V_2 - \frac{V_2 - V_1}{4} \,;\; V(x) \geqq V_1 + \frac{V_2 - V_1}{4} \right\} .$$

Then $\sup_{x \in \tilde{D}} |f_x| = \tilde{C}$ is less than infinity since $\tilde{D}$ is bounded and

$$\inf_{x \in \tilde{D}} |f(x)| = \tilde{\delta}$$

is greater than zero due to (ii).

Define $\tilde{\rho}$ and $\tilde{\tau}$ such that $0 < \tilde{\rho} < \min(1, \tilde{\delta}/8\tilde{C})$ and $\tilde{\tau} = \tilde{\rho}/2\tilde{\delta}$. Then, as in the proof of Lemma 3 we conclude that, for a fixed realization in Ω^* (defined in the beginning of the proof of Lemma 1) there exists an integer $\tilde{K}$, such that if $x(k) \in \tilde{D}$ and $k > \tilde{K}$ then

$$V(x(m(k, \tilde{\tau}))) < V(x(k)) - \tilde{\tau}\tilde{\delta}^2/2 .$$

Therefore $V(x(k))$ is strictly decreasing in $\tilde{D}$ from a certain k on. Since, as before, the step size $x(n) - x(n-1)$ tends to zero in D_2 it follows that $x(k)$ cannot pass from D_1 to a value outside D_2 after a certain value of k. Hence from this value on the algorithm (26) coincides with (1) and Theorem 1 now completes the proof of Theorem 3. □

Clearly, this theorem can be combined with Lemmas 2 and 4 to yield obvious variants.

5. The Robbins–Monro and Kiefer–Wolfowitz procedures. The analysis gives some extensions of the "classical" convergence results on the Robbins–Monro procedure, e.g., [11], even though the results given so far deal with a fairly special structure. In the first place it is possible to treat the case with dependent disturbances $\{e(n)\}$ in (1). Moreover, the frequently cited condition

$$\sum_1^\infty \gamma(n)^2 < \infty \tag{27}$$

has been shown to be unnecessary. When the disturbances $\{e(n)\}$ satisfy C1 (with $\alpha_k =$ constant), and when $\{\gamma(n)\}$ satisfies C2 it is sufficient that

$$\sum_1^\infty \gamma(n)^p < \infty . \tag{28}$$

This condition together with C2 are satisfied, e.g., for $\gamma(n) = Cn^{-\alpha}$ where $1/p < \alpha \leqq 1$. There consequently is a trade-off between conditions on $\{\gamma(n)\}$

and on the moments of $\{e(n)\}$. It can also be shown that (27) can be relaxed only if higher moments of $\{e(n)\}$ exist ([8]).

It can also be remarked that often

$$\gamma(n) = \lambda(n)/n$$

appears to be a suitable choice of gain sequence, where $\lambda(n)$ is a possibly random sequence tending a.e. to a positive constant $\bar{\lambda}$. Then (1) can be written

$$x(n) = x(n-1) + \frac{1}{n}(\bar{\lambda}f(x(n-1)) + (\lambda(n) - \bar{\lambda})f(x(n-1)) + \lambda(n)e(n) + \lambda(n)\beta(n)) .$$

The term $(\lambda(n) - \bar{\lambda})f(x(n-1))$ then can be incorporated in $\beta(n)$ if f is bounded. The result quoted in Section 3 for the choice $\gamma(n) = 1/n$ then can be applied to infer B1 from mild conditions on $\{e(n)\}$ and $\{\lambda(n)\}$.

A frequently studied problem ([6], [7]) is to find the minimum of a function $V(x)$ from noise corrupted measurements

$$y_i(x) = V(x) + w(i) \tag{29}$$

where $\{w(i)\}$ is a sequence of random variables with zero mean values.

In the Kiefer–Wolfowitz procedure [6] it is suggested to form an estimate of the negative gradient at $x = x^*$

$$d(x^*, c)$$

based on (linear operations of) at least $n + 1$ measurements of $V(x)$ in the sphere around x^* with radius c. Then

$$d(x^*, c) = -\frac{d}{dx}V(x)\Big|_{x=x^*} + \beta + e$$

where

$$|\beta| \leqq c|V''(\xi)| \qquad \xi \text{ belongs to } B(x^*, c)$$

and e is formed from the variables $w(i)$ and has a variance

$$E|e|^2 \sim Ew(i)^2/c^2 .$$

The Kiefer–Wolfowitz procedure amounts to choosing a decreasing sequence $c_n \to 0$ and then take

$$x(n) = x(n-1) + \gamma(n)\{d(x(n-1), c_n)\} .$$

Suppose that the function $V(x)$ satisfies A1, A2, A3, C3, C4 and $\{\gamma(n)\}$ satisfies C2, A4 and $w(i)$ satisfies C1 with α_k constant (which implies that the corresponding sequence $\{e(n)\}$ satisfies C1 with $\alpha_k = 1/c_k$). Then Theorem 1 implies that $x(n)$ tends to D_S w.p. 1 as $n \to \infty$ if

$$\textstyle\sum_1^\infty (\gamma(n)/c_n^2)^p < \infty$$

which is less restrictive a condition on $\{\gamma(n)\}$ and $\{c_n\}$ then the one given by Blum [2]:

$$\textstyle\sum_1^\infty c_n\gamma(n) < \infty \qquad \text{and} \qquad \sum_1^\infty (\gamma(n)/c_n)^2 < \infty .$$

6. Extensions. In this section it will be discussed how the results of Sections 3 and 4 can be extended to more general algorithms than (1).

(1) First, it is not necessary that $f(x)$ is the negative gradient of V. It is clear from the proofs that what matters is only that V is a twice continuously differentiable function, subject to A2, such that the scalar product $((d/dx)V(x))^T f(x) < 0$ outside a compact set D_S. Then under the appropriate additional assumption convergence of $x(n)$ to D_S follows. Therefore we may dispense with the assumption $(d/dx)V(x) = -f(x)$ and instead postulate the existence of such a function V. In the theory of differential equations (see, e.g., [4] or [12]), such a function is known as a *Lyapunov function*, and it guarantees that the solution of the differential equation

$$\frac{d}{d\tau} X(\tau) = f(X(\tau)) \tag{30}$$

for any initial condition $X^0 \in R^n$ at $\tau = 0$ tends to the set D_S as τ tends to infinity. Conversely, the existence of an invariant set D_S to the differential equation (d.e.) (30) such that for all initial conditions, the solution tends to D_S implies the existence of a function $V(x)$ with the aforementioned properties. (An *invariant set* D_S of a d.e. is a set such that a solution that belongs to D_S for a certain τ_0 also belongs to D_S for all other τ, $-\infty < \tau < \infty$. The set of all values x^0 such that solutions starting at x^0 tend to D_S is known as the *domain of attraction* of D_S.)

Therefore A1 and A2 can be replaced by

A1′: The d.e. (30) has an invariant set D_S with global domain of attraction.

Actually, if an invariant set does not have a global domain of attraction A1, A2 and B1 may be replaced by

A1″: The d.e. (30) has an invariant set D_S with domain of attraction D_A.
B1′: $x(n) \in \bar{D}$ i.o. w.p. 1 where $\bar{D}$ is a compact subset of D_A.

To make the d.e. (30) meaningful, we here assume that f is an everywhere defined locally Lipschitz-continuous function.

Actually, in the proof of Lemma 1, it was shown that the sequence $\{x(n)\}$ locally and asymptotically follows the trajectories of (30). In fact, under additional conditions the trajectories of (30) can be associated with the asymptotic behavior of (1) in a more strict sense, cf. [8]—[10].

It may also be remarked that the derivative of f in Lemma 4 can be interpreted as the system matrix for the linearized d.e. around a stationary point x^*. The interpretation of Lemma 4 then is that $x(n)$ may converge only to *stable, stationary points of the* d.e. (30).

(2) The analysis can be applied not only to the structure (1) with additive disturbances but also to the case

$$x(n) = x(n-1) + \gamma(n)Q(n; x(n-1), e(n)) \tag{31}$$

where $Q(\cdot, \cdot, \cdot)$ is a function from $R \times R^n \times R^m$ to R^n satisfying certain regularity conditions. A function f is defined as

$$f(x) = \lim_{n\to\infty} EQ(n, x, e(n)) \tag{32}$$

where the expectation is over the distribution of $e(n)$, with x regarded as a fixed parameter. It is assumed that the limit in (32) exists. With f thus defined we may study the d.e. (30) and relate convergence of (31) to stability properties of (30) as above. Some further technicalities in the proof of the the theorem are required in this case, but the basic paths of the proofs remain the same. The structure (31) is studied in detail in [8].

(3) As a final increase of complexity, it may be assumed that the disturbance term $e(n)$ in (31) depends on previous estimates $x(k)$, $k < n$. In particular a structure like

$$\begin{aligned} \varphi(n) &= g(n, \varphi(n-1), x(n-1), \nu(n)) \\ e(n) &= h(n, \varphi(n), x(n-1)) \end{aligned} \tag{33}$$

or a linear variant

$$\begin{aligned} \varphi(n) &= A(x(n-1))\varphi(n-1) + B(x(n-1))\nu(n) \\ e(n) &= C(x(n-1))\varphi(n) \end{aligned} \tag{34}$$

can be postulated, where $\{\nu(n)\}$ is assumed to be a sequence of independent random vectors. These structures are of particular interest in control theory and in certain sequential parameter estimation applications, cf., e.g., Hannan [5]. They are treated at length in [9] and [10]. The analysis again follows that of the simpler variant (1). A variable $\bar{e}(n, x)$ is defined for each x by

$$\begin{aligned} \bar{\varphi}(n, x) &= g(n, \bar{\varphi}(n-1, x), x, \nu(n)); \qquad \bar{\varphi}(0, x) = 0 \\ \bar{e}(n, x) &= h(n, \bar{\varphi}(n, x), x) \end{aligned}$$

and it is assumed that the limit

$$f(x) = \lim_{n\to\infty} EQ(n, x, \bar{e}(n, x))$$

exists with expectation over $\{\nu(n)\}$. The corresponding d.e. (30) is then analysed for stability properties, and these are related to strong convergence of (31), (33) as above.

The proofs for the case (31) and (33) or (31) and (34) are considerably more technical than those given in Section 3, but differ from them essentially only by an increased amount of bookkeeping over small terms.

7. Conclusions. Strong convergence of a certain recursive algorithm (1), has been the main topic of this paper. The convergence results have been obtained by studying the behavior of the algorithm on each realization outside a given null set of realizations. The convergence results (Theorems 1 and 3) imply certain extensions compared to previous results on strong convergence of stochastic approximation algorithms. Also the classification of possible convergence points, Theorem 2, seems to be new.

It is believed, though, that the important merit of the present approach is that the method of proof extends directly to more complex algorithms as described in Section 6, while it does not seem to be clear how the conventional technique would be applied to, say, (31) and (33).

Acknowledgment. The author is indebted to the anonymous referees and to his former colleagues at the Department of Automatic Control, Lund Institute of Technology, for many useful comments and suggestions on previous versions of this paper.

REFERENCES

[1] ALBERT, A. E. and GARDNER, L. A. (1967). *Stochastic Approximation and Nonlinear Regression.* Research Monograph 42, MIT Press.

[2] BLUM, J. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Statist.* **25** 737–744.

[3] CRAMÉR, H. and LEADBETTER, M. R. (1967). *Stationary and Related Stochastic Processes.* Wiley, New York.

[4] HAHN, W. (1967). *Stability of Motion.* Springer-Verlag, Berlin.

[5] HANNAN, E. J. (1976). The convergence of some recursions. *Ann. Statist.* **4** 1258–1270.

[6] KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

[7] KUSHNER, H. J. (1976). General convergence results for stochastic approximations via weak convergence theory. To appear in *J. Math. Anal. Appl.* Also available as LCDS TR 76-1, Brown Univ.

[8] LJUNG, L. (1974). Convergence of recursive stochastic algorithms. Report 7403, Dept. of Automatic Control, Lund Institute of Technology, Sweden.

[9] LJUNG, L. (1975). Theorems for the analysis of recusive stochastic algorithms. Report 7522, Dept. of Automatic Control, Lund Institute of Technology, Sweden.

[10] LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control* **AC-22** 551–575.

[11] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

[12] ZUBOV, V. I. (1964). *Methods of A. M. Lyapunov and Their Application.* Noordhoff, Groningen.

DEPARTMENT OF ELECTRICAL ENGINEERING
LINKÖPING UNIVERSITY
INSTITUTIONEN FÖR SYSTEMTEKNIK
S-581 83 LINKÖPING, SWEDEN