# On the Convergence of Certain Recursive Algorithms

Ljung, Lennart

1976

*Document Version:*
Publisher's PDF, also known as Version of record

# ON THE CONVERGENCE OF CERTAIN RECURSIVE ALGORITHMS

L. LJUNG

Report 7625 (C) May 1976
Department of Automatic Control
Lund Institute of Technology

# ON THE CONVERGENCE OF CERTAIN RECURSIVE ALGORITHMS

Lennart Ljung
Department of Automatic Control
Lund Institute of Technology
P.O.Box 725, 220 07 Lund, Sweden

ABSTRACT

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

---

AMS Classification No. 62L20
Keywords: Recursive stochastic algorithms, stochastic approximation.

TABLE OF CONTENTS

## 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 and $\{\beta(n)\}$ is a null sequence. Algorithm (1) coincides with the one recently analysed by Kushner [1].

This algorithm has obvious relations to the Robbins-Monro procedure [2] which, perhaps is the best known one of the stochastic approximation algorithms. 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 $P(Y(x) \leq y) = H(y|x)$ such that

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

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 $x(1), \ldots, x(n-1), y(1), \ldots, y(n-1)$ is $H(y|x(n-1))$. The asymptotic properties of (2) have been studied by many authors, e.g. [2], [3], [4] etc.

If $\beta(n) = 0$ all $n$, and

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

the algorithm (1) is a special case of (2). However, in many 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 [5] for minimization of a function has a similar relationship to (1), and as is further described in [1] 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, like control theory, parameter estimation methods, etc. More general variants of (1) have been analysed by the present author [6], [7], [8] with particular emphasis on control theory applications. The approach in these references is to asssociate (1) with a deterministic differential equation, in terms of which strong convergence of (1) can be studied.

In the study by Kushner [1] a similar idea is persued, 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) as just a simple special case of the algorithms studied in [8]. 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

$$D_S = \{x | f(x) = 0\} \tag{4}$$

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 $\frac{d}{dx} V(x) = -f(x)$ (column vector)

A2: $\{x|V(x) \leq C\}$ is compact for all $C < \sup V(x)$

A3: The set $D_S$ consists 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 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 behaviour of (1) and about the properties of the sequences $\{e(n)\}$ and $\{\gamma(n)\}$ are introduced.

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

B2: With probability one, $|x(n)|$ does not tend to infinity as $n \to \infty$.

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

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

where $\{v(k)\}$ are independent random vectors with zero mean values and unit

covariance matrices, and such that $E|v(k)|^{2p} < C$ for some integer p. Furthermore $|h(n,k)| < \alpha_n \lambda^{n-k}$ where $\{\alpha_n\}$ is non-decreasing and $\lambda < 1$.

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

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

Moreover

$$\sum_{1}^{\infty} (\gamma(n)\alpha_n^2)^p < \infty$$

where $\alpha_n$ and p are defined in C1.

C3: $\sup \left| \frac{d}{dx} f(x) \right| < \infty$

C4: $D_{\bar{\delta}} = \{x \mid |f(x)| \leq \bar{\delta}\}$ is compact 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

$$D_M = \{x \mid f(x) = 0 \text{ and the matrix } \frac{d}{dx} f(x) \text{ is negative semi-definite}\} \qquad (5)$$

This requires the following additional assumptions.

D1: The sequence $\{\beta(n)\}$ is a deterministic sequence. Furthermore, the sequence $\{e(n)\}$ can be decomposed as $e(n) = \varepsilon(n) + v(n)$ where $\varepsilon(n)$ is independent of $v(n)$ and $e(k)$ $k \neq n$ such that $E\,\varepsilon(n) = 0$ and $E\,\varepsilon(n)\varepsilon(n)^T \geq cI$ for some $c > 0$ and all n.

D2: The set $D_S$ consists of isolated points.

Assumption D1 implies that e(n) is not exactly predictable from the other variables $e(k)$ $k \neq n$. It should not be regarded as a very restrictive condition.

## 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

$x(n) \to D_S = \{x | f(x) = 0\}$ w.p.1 as $n \to \infty$

Proof: Let $\Omega^*$ be a subspace of the sample space such that

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

Clearly $P(\Omega^*) = 1$. Consider from now on a fixed realisation $\omega^* \in \Omega^*$ and let us study the sequence $\{x(k)\} = \{x(k,\omega^*)\}$. We shall throughout suppress the argument $\omega^*$, 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 $|f(x^*)| = \delta^* > 0$. From (1) we obtain directly

$$x(j) = x(n_k) + \sum_{n_k+1}^{j} \gamma(k)e(k) + \sum_{n_k+1}^{j} \gamma(k)\beta(k) + \sum_{n_k+1}^{j} \gamma(k)f\big(x(k-1)\big) =$$

$$= x(n_k) + S_1(n_k,j) + S_2(n_k,j) + f(x^*) \sum_{n_k+1}^{j} \gamma(k) + R(n_k,j,x^*) \qquad (6)$$

where

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

$$S_2(n_k,j) = \sum_{n_k+1}^{j} \gamma(k)\beta(k)$$

$$R(n_k,j,x^*) = \sum_{n_k+1}^{j} \gamma(k)\Big(f\big(x(k-1)\big) - f(x^*)\Big)$$

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

$$\sum_{n_k+1}^{m(n_k,\tau)} \gamma(k) \nearrow \tau < \infty \text{ as } n_k \to \infty \tag{7}$$

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

$$S_i(n_k,j) \to 0 \text{ uniformly in } n_k \leq j \leq m(n_k,\tau) \text{ for fixed } \tau \text{ as } n_k \to \infty \tag{8}$$

Proof of claim:

i=2: $|S_2(n_k,j)| \leq \max_{n_k \leq i} |\beta(i)| \cdot \tau \to 0$ as $n_k \to \infty$ according to A5

i=1: From the definition of z(n) in B1 we have

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

or

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

according to B1.

Let $B(x^*,\rho) = \{x \mid |x-x^*| < \rho\}$ and $C^* = \max\left\{\sup_{x \in B(x^*,1)} \left|\frac{d}{dx} f(x)\right|, 1\right\}$

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 $\rho^*$ will be clear below. Clearly,

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

Choose from now on

$$\tau = \tau^* = \rho^*/\delta^* \quad (\leq 1/8C^*)$$

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

$$|x(j) - x^*| \leq |x(n_k) - x^*| + |S_1(n_k,j)| + |S_2(n_k,j)| + |f(x^*)| \sum_{n_k+1}^{j} \gamma(k) +$$

$$+ \tau^* \max_{n_k+1 \leq i \leq j} |f(x(i-1)) - f(x^*)| \qquad (9)$$

Choose $k > K_1$ so large that

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

and

$$|x(n_k) - x^*| < \rho^*/2 \quad \text{(Then in particular } x(n_k) \in B^*\text{)}$$

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

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

and

$$|x(j) - x^*| \leq \rho^*/2 + \rho^*/8 + \rho^*/8 + C^*\tau^* + C^*\rho^*\tau^* < \rho^*/2 + \rho^*/4 + \rho^*/8 + \rho^*/8 \leq \rho^*$$

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

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

In particular we have

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

where

$$|R_2(n_k)| \leq |x(n_k) - x^*| + |S_1(n_k,m_k^*)| + |S_2(n_k,m_k^*)| +$$

$$+ C^*\rho^*\tau^* + \delta^* \left| \tau^* - \sum_{n_k+1}^{m_k^*} \gamma(j) \right| \qquad (10b)$$

Now

$$V(x(m_k^*)) = V(x^*) + (x(m_k^*) - x^*)V_x(x^*) + (x(m_k^*) - x^*)V_{xx}(\xi)(x(m_k^*) - x^*) =$$

$$= V(x^*) - \tau^* f(x^*)V_x(x^*) + (x(m_k^*) - x^*)V_{xx}(\xi)(x(m_k^*) - x^*) +$$

$$+ R_2(n_k)V_x(x^*) \qquad (11)$$

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

$$|R_2(n_k)f(x^*) + (x(m_k^*) - x^*)f_x(\xi)(x(m_k^*) - x^*)| \leq$$

$$\leq \delta^* \{ |x(n_k) - x^*| + |S_1(n_k, m_k^*)| + |S_2(n_k, m_k^*)| +$$

$$+ C^*\rho^*\tau^* + \delta^* |\tau^* - \sum_{n_k}^{m_k^*} \gamma(j)| \} + C^*\rho^{*2} \qquad (12)$$

Choose $k > K^*$ so that

$$|x(n_k) - x^*| < \delta^*\tau^*/16 = \rho^*/8$$

$$|S_1(n_k, m_k^*)| + |S_2(n_k, m_k^*)| < \delta^*\tau^*/16 = \rho^*/8$$

$$\left| \tau^* - \sum_{n_k+1}^{m_k^*} \gamma(j) \right| < \tau^*/16 = \rho^*/8\delta^*$$

Then the RHS of (12) is less than

(recall that $\rho^* = \delta^*\tau^*$)

$$\frac{(\delta^*)^2}{16} \left( \tau^* + \tau^* + \frac{16C^*\rho^*}{\delta^*} \tau^* + \tau^* \right) + C^*\rho^*\tau^*\delta^* \leq$$

$$\leq \frac{(\delta^*)^2}{16} (3\tau^* + 2\tau^*) + C^*/8C^*\tau^*(\delta^*)^2 \leq (\delta^*)^2\tau^*/2$$

where the first inequality follows from $\tau^* < 1/8C^*$ and the second one from the definition of $\rho^*$.

Hence,

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

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 $j = m_k^*$) to

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

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 infinum 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 $x^o$ would obviously yield $V(x^o) = V^o > \bar{V}$. Then $V(x(k)) > \bar{V} + d$ infinitely often for some sufficiently small d. Since V is continuous we can according to A3 choose this d so small that the compact area

$$\bar{D} = \left\{x|\bar{V} + \frac{d}{2} \leq V(x) \leq \bar{V} + d\right\} \qquad (14)$$

has no point in common with $D_S$.

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

Let $\{x(n_k')\}$ be defined as follows:

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

$$V(x(n_k' + s_k)) > \bar{V} + d$$

where $s_k$ is the first s for which $x(n_k' + s) \notin \bar{D}$. Let $x(n_k') \to \tilde{x}$ as $k \to \infty$. Clearly $V(\tilde{x}) = \bar{V} + \frac{d}{2}$ and $|\tilde{f}(x)| = \tilde{\delta} > 0$. Now define $\tilde{\rho}$, $\tilde{\tau}$ as above and let $\tilde{\rho}$ be so small that $B(\tilde{x},\tilde{\rho})$ has no point in common with $\{x | V(x) \geq \bar{V} + d\}$. Then, from (13)

$$V\left(x(m(n_k',\tilde{\tau}))\right) < \bar{V} + \frac{d}{2} - \tilde{\delta}^2\tilde{\tau}^2/2$$

and $x(j) \in B(\tilde{x},\tilde{\rho})$; $n_k' \leq j \leq m(n_k',\tilde{\tau})$. This contradicts 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 neighbourhood of $D_S$. This concludes the proof of Theorem 1. □

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

<u>Remark</u>. 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 $S = \{z | V(x) = z, x \in D_S \text{ and } z \leq V^0\}$ 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 areas 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 which gives

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

where

$$\Gamma(n,k) = \prod_{i=k+1}^{n} \left(1 - \gamma(i)\right) \qquad k < n; \qquad \Gamma(n,n) = 1$$

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 [6]). The fairly common choice $\gamma(n) = 1/n$ gives $\Gamma(n,k) = \frac{1}{n}$ and then various "laws of large numbers" can be applied to (15). A, for the present context suitable variant is given by Cramér and Leadbetter [9], p. 94-96(where it is given for continuous time stochastic processes, but the proof is also valid for discrete time processes):

Let $\gamma(n) = 1/n$ and assume $E\, e(n) = 0$,

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

Then $z(n) \to 0$ w.p.1 as $n \to \infty$.

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 $L \leq 1$ then

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

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

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

and then linking such estimates together using Minkowski's inequality. The formal proof is given in [6].

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

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

and

$$\sum_{n=1}^{\infty} P\{|z(n)| > \varepsilon\} \leq \frac{C_p}{\varepsilon^{2p}} \sum_{n=1}^{\infty} \gamma^p(n) \alpha_n^{2p} < \infty$$

The Borel Cantelli lemma now assures

$z(n) \to 0$ as $n \to \infty$ w.p.1. □

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 and Lemma 3 $\left(V(z) = \frac{1}{2L} z^2\right)$ converges w.p.1 to zero if

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

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

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 $f(x) = -\frac{d}{dx} V(x)$ 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 $\bar{C} = \sup|f_x|$.
Let $\bar{C}$ and $\bar{\delta}$ replace $C^*$ and $\delta^*$ in the proof of Lemma 1. Then $\bar{\rho}$ and $\bar{\tau}$ can be chosen globally outside $D_{\bar{\delta}}$.
Further take $x^* = x(k)$ which gives with $\bar{m}_k = m(k,\bar{\tau})$ for (13)

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

for all $k > \bar{K}$ and such that $x(k) \notin D_{\bar{\delta}}$.

Therefore, if $|x(k)| \to \infty$, $x(k)$ would remain outside the compact area $D_{\bar{\delta}}$ from a certain $K_1$ on. With $K_0 = \max(\bar{K}_1, \bar{K})$ and $n_i = n_{i-1} + m(n_{i-1}, \bar{\tau})$; $n_0 = K_0$ we then would have

$$V(x(n_j)) < V(x(\bar{K}_0)) - j\bar{\tau}\,\bar{\delta}^2/2$$

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

The set $D_3$ 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, C1, C2 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

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

have non positive real part.

Remark: Since $f(x) = -\frac{d}{dx} V(x)$, the matrix $-\frac{d}{dx} f(x)\Big|_{x=x^*}$ is the second derivative matrix (the Hessian) of V in $x^*$. The condition is that this should be positive semidefinite.

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

$$f(x) = f(x^*) + A(x-x^*) + g(x-x^*) \qquad (17)$$

where

$$g(x)/|x| \to 0 \text{ as } x \to 0 \qquad (18)$$

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

Let L be a left eigenvector for this eigenvalue:

$$LA = \mu L; \qquad \mu > 0$$

Introduce the following notation

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

$$\bar{\varepsilon}(k) = L\varepsilon(k) \qquad \text{(cf D1)}$$

$$\bar{f}(k,n) = L \sum_{j=n}^{k} h(k,j)v(j) - \bar{\varepsilon}(k) \qquad \text{(cf C1)}$$

$$\tilde{f}(k,n) = L \sum_{j=0}^{n-1} h(k,j)v(j)$$

Then $Le(k) = \bar{\varepsilon}(k) + \bar{f}(k,n) + \tilde{f}(k,n)$ and these terms are mutually independent since $h(k,k)v(k) - \varepsilon(k) = v(k) - E[e(k)|e(k-1), \ldots, e(0)]$ is independent of $\varepsilon(k)$.

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

$$y_k = y_{k-1} + \gamma(k)[\mu y_{k-1} + \bar{\varepsilon}(k) + \bar{f}(k,n) + \tilde{f}(k,n) + \bar{\beta}_k + \bar{g}_k] \qquad (19)$$

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

$$y_m = \Gamma(m,n)\sqrt{\gamma(n)}[Y_n + \bar{F}_\varepsilon(m,n) + \bar{F}(m,n) + \tilde{F}(m,n) + \bar{B}(m,n) + \bar{G}(m,n)] \qquad (20)$$

where

$$\Gamma(m,n) = \prod_{n+1}^{m} (1 + \mu\gamma(k)) \sim e^{\mu \sum_{n+1}^{m} \gamma(k)}$$

$$Y_n = \frac{1}{\sqrt{\gamma(n)}} y_n$$

$$\bar{F}_\varepsilon(m,n) = \frac{1}{\sqrt{\gamma(n)}} \sum_{n+1}^{m} \gamma(k)\Gamma(k,n)^{-1}\bar{\varepsilon}(k)$$

and $\bar{F}$, $\tilde{F}$, $\bar{B}$ and $\bar{G}$ defined similarly from $\bar{f}(k,n)$, $\tilde{f}(k,n)$, $\bar{\beta}_k$ and $\bar{g}_k$ respectively.

Since $\mu > 0$ it is possible to for each n take $m = \bar{m}(n)$ such that

$$r(\bar{m}(n),n) \geq 1$$

Now we have

$$E\left[\bar{F}_\varepsilon(m(n),n)\right]^2 = \frac{1}{\gamma(n)} \sum_{k=n}^{\bar{m}(n)} \gamma(k)^2 r(k,n)^{-2} E\bar{\varepsilon}(k)^2 \geq \frac{c}{\gamma(n)} \sum_{n}^{\bar{m}(n)} \gamma(k)^2 r(k,n)^{-2} \geq c_1 > 0$$

where $c_1$ is independent of n. Here the last inequality follows readily from the definition of $\bar{m}(n)$ and the properties of $\{\gamma(n)\}$, cf [8] p. 86. Since $\bar{F}_\varepsilon(m(n_1),n_1)$ is independent of $\bar{F}_\varepsilon(m(n_2),n_2)$ for $n_1 > m(n_2)$ it follows from the second Borel-Cantelli lemma that

$|\bar{F}_\varepsilon(m(n),n)| \geq c_2 > 0$ for infinitely many n w.p.1, i.e. in particular a.e. on $\Omega^*$.

Moreover, since $\bar{F}_\varepsilon$ and $\bar{F}$ are independent, also

$$|\bar{F}_\varepsilon(m(n),n) + \bar{F}(m(n),n)| \geq c_2 \quad \text{i.o. a.e. on } \Omega^* \tag{21a}$$

Furthermore

$$\tilde{F}(m(n),n) = \frac{1}{\sqrt{\gamma(n)}} \sum_{k=0}^{n} r(n,k)\nu(k)$$

where

$$r(n,k) = \sum_{j=n}^{\bar{m}(n)} \gamma(j) r(j,n)^{-1} Lh(j,k)$$

and

$$|r(n,k)| \leq \gamma(n)\alpha_{\bar{m}(n)}^{-} \sum_{j=n}^{\bar{m}(n)} \lambda^{j-k} \leq c_3\gamma(n)\alpha_{\bar{m}(n)}^{-}\lambda^{n-k} \leq c_4\gamma(\bar{m}(n))\alpha_{\bar{m}(n)}^{-}\lambda^{n-k}$$

where the first inequality follows from C1 and the last one from C2 (cf [8] p. 54). Now

$$E|F(m(n),n)|^{2p} \leq c_5\gamma(m(n))^p\alpha_{m(n)}^{2p} \sum_{k_1,\ldots,k_{2p}=0}^{n} \prod_{i=1}^{2p} \lambda^{n-k_i} E \prod_{i=1}^{2p} |\nu(k_i)| \leq$$

$$\leq c_6\gamma(m(n))^p\alpha_{m(n)}^{2p}$$

From Chebyshev's inequality and the Borel-Cantelli lemma together with C2 it now follows that

$$\bar{F}(m(n),n) \to 0 \text{ w.p.1 as } n \to \infty \tag{21b}$$

It is easy to verify that $\sum_{k=n}^{\infty} \gamma(k)\Gamma(k,n)^{-1} = 1$. Therefore $Y_n$ will dominate over $\bar{G}(\bar{m}(n),n)$ for large n and for $\omega \in \Omega^*$ according to (18) since $x(k) \to x^*$ on $\Omega^*$.

[Remark: It might happen that $x(n) - x^*$ would be "very close" to the null space of L so that $y_n$ would be an order of magnitude smaller than $|x(n) - x^*|$. However, this cannot happen for all n larger than some $N_o$, since 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 neighbourhood of $x^*$ the distribution of x(n) will be non-degenerate.]

From (20) follows that since $y_{\bar{m}(n)} \to 0$ on $\Omega^*$ we must have

$$H(\bar{m}(n),n) \triangleq \bar{F}_\varepsilon(\bar{m}(n),n) + \bar{F}(\bar{m}(n),n) + \bar{\bar{F}}(\bar{m}(n),n) + \bar{B}(\bar{m}(n),n) +$$

$$+ Y_n + \bar{G}(\bar{m}(n),n) \to 0 \quad \text{on } \Omega^* \text{ as } n \to \infty \tag{22}$$

But $\bar{F}_\varepsilon + \bar{F}$ is independent of all terms in H except $\bar{G}$. Therefore (22) would imply that $\bar{F} + \bar{F}_\varepsilon$ + "part of" $\bar{G}$ tends to zero on $\Omega$, which in particular means that $\bar{G}$ does not tend to zero due to (21a). Since $\bar{G}$ is dominated by Y on $\Omega^*$, the term $Y_n$ would then tend to infinity. But $\bar{\bar{F}}$ tends to zero according to (21b) and $\bar{B}$ is deterministic. Therefore $Y_n$ would dominate H, which violates (22), and we have arrived at a contradiction to the assumption that $f_x(x^*)$ has a positive eigenvalue. □

## 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 into $D_S$ 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 $e(n) = 0 \quad n \geq 2$, $\gamma(n) = 1/n$, $\beta(n) = 0$ all n. Then if $x(0) = 0$, $x(n) = x(n-1) + \frac{1}{n}\left(- x(n-1)^3\right)$; $n \geq 2$; $x(1) = e(1)$. 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 area $D_2$. Then (1) takes the modified form

$$x(n) = \left[x(n-1) + \gamma(n)\left(f\left(x(n-1)\right) + e(n)\right)\right]_{D_1}^{D_2} \qquad (23)$$

where

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

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

For the modified algorithm (23) obviously B2 holds. However, Lemma 1 cannot be directly applied because of the modification of (1). The following result holds though.

Theorem 3. Consider the modified algorithm (23). Assume A1 to A5, B2, 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

$$\tilde{D} = \left\{ x \,|\, V(x) \leq V_2 - \frac{V_2 - V_1}{4} \; ; \qquad V(x) \geq V_1 + \frac{V_2 - V_1}{4} \right\}$$

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

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

is greater than zero due to (ii).

As in the proof of Lemma 3 $\tilde{\rho}$ and $\tilde{\tau}$ can be chosen from $\tilde{C}$ and $\tilde{\delta}$ globally in $\tilde{D}$, and for a fixed realization in $\Omega^*$

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

for all $k > \tilde{K}$ and $x(k) \in \tilde{D}$. Therefore $V\left(x(k)\right)$ 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 (23) 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. [2], 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

$$\sum_{1}^{\infty} \gamma(n)^2 < \infty \tag{24}$$

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

$$\sum_{1}^{\infty} \gamma(n)^p < \infty \tag{25}$$

This condition together with C2 are satisfied e.g. for $\gamma(n) = Cn^{-\alpha}$ $1/p < \alpha \leq 1$. There consequently is a trade-off between conditions on $\{\gamma(n)\}$ and on the moments of $\{e(n)\}$. The following example shows that (24) can be relaxed only if higher moments of $\{e(n)\}$ exist.

Example 1. Let $\{e(n)\}$ be a sequence of independent random variables where $e(n)$ has the distribution

$$e(n) = \begin{cases} 1/\gamma(n) & \text{with probability } \gamma(n)^r \\ 0 & \text{with probability } 1 - \gamma(n)^r \end{cases}$$

Then $P\left(|\gamma(n)e(n)| \geq 1\right) = \gamma(n)^r$.

The moments $E|e(n)|^s$ are uniformly bounded only for $s \leq r$.

Assume that

$$\sum_{n=1}^{\infty} \gamma(n)^r = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma(n)^{r+\varepsilon} < \infty \quad \text{for some } \varepsilon > 0$$

Then

$$\sum_{n=1}^{\infty} P\{|\gamma(n)e(n)| \geq 1\} = \sum_{n=1}^{\infty} \gamma(n)^r = \infty$$

and since the variables $\{e(n)\}$ are independent

$$|\gamma(n)e(n)| \geq 1 \quad \text{i.o. w.p.1}$$

from the Borel-Cantelli lemma. With $z(n)$ defined by the algorithm in B1

$$z(n) = (1 - \gamma(n))z(n-1) + \gamma(n)e(n)$$

$z(n)$ will consequently w.p.1 not converge to any limit. To be able to apply Lemma 2

$$E|e(n)|^{2(r+\varepsilon)}$$

would have to be uniformly bounded. Thus the moment conditions on $e(n)$ cannot be dispensed with. □

It can also be remarked that in many applications,

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

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

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

The term $(\lambda(n) - \bar{\lambda})f(x(n-1))$ then can be incorporated in $\beta(n)$. 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 encountered problem in applications is to find the minimum of a function $V(x)$ from noise corrupted measurements

$$y_i(x) = V(x) + w(i) \tag{26}$$

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

In the Kiefer-Wolfowitz procedure [5] 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

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

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 $\alpha_k$ 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

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

which is less restrictive a condition on $\{\gamma(n)\}$ and $\{c_n\}$ than the one given by Blum [3]:

$$\sum_1^\infty c_n\gamma(n) < \infty \quad \text{and} \quad \sum_1^\infty \left(\gamma(n)/c_n\right)^2 < \infty$$

## 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 $\left(\frac{d}{dx}V(x)\right)^T f(x) < 0$ 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 $\frac{d}{dx}V(x) = -f(x)$ and instead postulate the existence of such a function V. In the theory of differential equations, see e.g. [10] or [11], such a function is known as a Lyapunov function, and it guarantees that the solution of the differential equation

$$\frac{d}{d\tau}X(\tau) = f(X(\tau)) \qquad (27)$$

for any initial condition $X^o \in R^n$ at $\tau = 0$ tends to the set $D_S$ as $\tau$ tends to infinity. Conversely, the existence of an invariant set $D_S$ to the differential equation (d.e.) (27) 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 $\tau_o$ also belongs to $D_S$ for all other $\tau$, $-\infty < \tau < \infty$. 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. (27) 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. (27) 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. (27) 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 (27). In fact, under additional conditions the trajectories of (27) can be associated with the asymptotic behaviour of (1) in a more strict sense, cf [6] - [8].

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*. Lemma 4 then (essentially) states that x(n) may converge only to stable, stationary points of the d.e. (27).

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

$$x(n) = x(n-1) + \gamma(n)Q\big(n;x(n-1),e(n)\big) \tag{28}$$

A function f is defined as

$$f(x) = \lim_{n\to\infty} E\, Q\big(n,x,e(n)\big) \tag{29}$$

where the expectation is over the distribution of e(n), with x regarded as a fixed parameter. It is assumed that the limit exists. With f thus defined we may study the d.e. (27) and relate convergence of (28) to stability properties of (27) as above. Some further technicalities in the proof of the theorem are required in this case, but the basic paths of the proofs remain the same. The structure (28) is studied in detail in [6].

3) As a final increase of complexity, it may be assumed that the disturbance term e(n) in (28) depends on previous estimates x(k), k < n. In particular a structure like

$$\begin{aligned}\varphi(n) &= g\big(n,\varphi(n-1),x(n-1),v(n)\big)\\ e(n) &= h\big(n,\varphi(n),x(n-1)\big)\end{aligned} \tag{30}$$

or a linear variant

$$\begin{aligned}\varphi(n) &= A\big(x(n-1)\big)\varphi(n-1) + B\big(x(n-1)\big)v(n)\\ e(n) &= C\big(x(n-1)\big)\varphi(n)\end{aligned} \tag{31}$$

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 parameters estimation applications, cf e.g. Hannan [12]. They are treated at length in [8]; the theorems are also quoted in [7]. The analysis again follows that of the simpler variant (1). A variable $\bar{e}(n,x)$ is defined for each x by

$$\bar{\varphi}(n,x) = g\left(n,\bar{\varphi}(n-1,x),x,\nu(n)\right); \quad \bar{\varphi}(0,x) = 0$$

$$\bar{e}(n,x) = h\left(n,\bar{\varphi}(n,x),x\right)$$

and it is assumed that the limit

$$f(x) = \lim_{n\to\infty} E\, Q\left(n,x,\bar{e}(n,x)\right)$$

exists with expectation over $\{\nu(n)\}$. The corresponding d.e. (27) is then analysed for stability properties, and these are related to strong convergence of (28), (30) as above.

The proofs for the case (28) and (30) or (28) and (31) are considerably more technical than those given in Section 3, but differ from them essentially only by an increased amount of book-keeping over small terms.

## 7. CONCLUSIONS

Strong convergence of a certain recursive algorithm (1), has been the main topic of this paper. The approach of the convergence results has been to study the behaviour of the algorithm realization-wise outside a given nullset of realizations. The convergence results (Theorems 1 and 3) imply certain extensions compared to the classical 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, (28) and (30).

REFERENCES

[1] H.J. Kushner: General convergence results for stochastic approximations via weak convergence theory. Submitted to Annals of Statistics (1976)

[2] H. Robbins and S. Monro: A stochastic approximation method. Ann. Math. Stat. Vol. 22 (1961) pp. 400-407.

[3] J. Blum: Multidimensional stochastic approximation methods. Ann. Math. Stat. Vol. 25 (1954) pp. 737-744.

[4] A.E. Albert and L.A. Gardner: Stochastic Approximation and Nonlinear Regression, Research Monograph 42, The MIT Press, Cambridge, Mass., 1967.

[5] J. Kiefer and J. Wolfowitz: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. Vol. 23 (1952) pp. 462-466.

[6] L. Ljung: Convergence of recursive stochastic algorithms. Report 7403, Dept. of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1974.

[7] L. Ljung: Analysis of recursive stochastic algorithms. Submitted to IEEE Trans. on Automatic Control (1976).

[8] L. Ljung: Theorems for the analysis of recursive stochastic algorithms. Report 7522, Dept. of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1975.

[9] H. Cramér and M.R. Leadbetter: Stationary and Related Stochastic Processes. Wiley, New York, 1967.

[10] W. Hahn: Stability of Motion. Springer Verlag, Berlin, 1967.

[11] V.I. Zubov: Methods of A.M. Lyapunov and Their Applications. Groningen, P. Noordhoff, 1964.

[12] E.J. Hannan: The convergence of some recursions. Submitted for publication.