On the Convergence of Certain Recursive Algorithms

Ljung, Lennart

ON THE CONVERGENCE OF CERTAIN RECURSIVE ALGORITHMS

L. LJUNG

Report 7505 (C) February 1975
Lund Institute of Technology
Department of Automatic Control

ON THE CONVERGENCE OF CERTAIN RECURSIVE ALGORITHMS

Lennart Ljung *

Abstract

This paper gives convergence (with probability one) theorems for a class of recursive (sequential) algorithms. The class contains the Robbins - Monro and the Kiefer - Wolfowitz schemes and also more general algorithms of practical interest. A basic idea is that a deterministic ordinary differential equation is associated with the algorithm, and it is shown that convergence analysis can be performed in terms of this differential equation. When applied to the stochastic approximation algorithms mentioned above, the theorems give convergence results that are more general than those usually given.

* Division of Automatic Control, Lund Institute of Technology, Fack S-220 07 LUND, Sweden

At present with the Information Systems Laboratory , Stanford University, Stanford, Calif. 94305.

This research was supported partly by the Swedish Board for Technical Development under contract No 733546.

AMS Classification No: 62L20

Keywords: Recursive stochastic algorithms, stochastic approximation.

ON THE CONVERGENCE OF CERTAIN RECURSIVE ALGORITHMS

Lennart Ljung
Division of Automatic Control, Lund Institute of Technology, Fack , S-220 07 LUND, Sweden

1. Introduction

Recursive algorithms for estimation of certain parameters, finding roots of regression functions etc is a topic of great interest in applied mathematics. Stochastic recursive algorithms, like stochastic approximation algorithms, have been studied by e.g. Robbins and Monro (1951), Kiefer and Wolfowitz (1952), Blum (1954), Dvoretzky (1956), Wasan (1969) etc., in all of which convergence is a main concern. To prove convergence with probability one usually the martingale convergence theorem is used, which gives elegant proofs, but introduces some additional assumptions. When the algorithms are used in practical applications it frequently happens that some of these assumptions are not satisfied. In this paper convergence of recursive algorithms will be studied using a different approach. The proofs will be more technical, but in return a larger class of algorithms can be treated.

2. A Class of Algorithms

Consider the recursive algorithm

$$x_{n+1} = x_n + \gamma_{n+1} Q_n(x_n, \varphi_{n+1}) \tag{1a}$$

where $\{\gamma_n\}$ is a sequence of positive scalars, such that

$$\gamma_n \to 0 \quad \text{as} \quad n \to \infty \qquad \text{and} \qquad \sum_1^\infty \gamma_n = \infty \tag{1b}$$

$x_n \in R^m$ is the estimate and φ_{n+1} is a vector valued observation obtained at time n+1 , and Q_n is a correction formed from this observation and the current estimate x_n. The observation φ_{n+1} can be thought of as formed from a sequence of random variables $\{e_k\}_1^{n+1}$ and a sequence of parameters $\{b_k\}_1^n$, such that b_n is at the observer's disposal at time n.

Example. In many cases the observations are generated recursively as

$$\varphi_{n+1} = A(b_n)\varphi_n + e_{n+1} \tag{2}$$

where $A(b_n)$ is a matrix, which depends on the parameter b_n, chosen by the observer. ∎

To denote this dependence explicitly, let

$$\varphi_n = \varphi_n(b_{n-1}, b_{n-2}, \ldots, b_1; e_n, e_{n-1}, \ldots, e_1) = \varphi_n(b(\cdot)).$$

Introduce also

$$\bar{\varphi}_n(b) \triangleq \varphi_n(b, b, \ldots, b; e_n, e_{n-1}, \ldots, e_1).$$

In algorithm (1), the following choice of b_k is made:

$$b_k = x_k.$$

Algorithm (1) is similar to the stochastic approximation algorithms treated e.g. in Blum (1954), Burkholder (1956) etc. However, the results of these papers require that

$$E\left\{Q_n(x_n, \varphi_{n+1}) / \mathcal{X}_n\right\} = f(x_n) \tag{3}$$

where $\mathcal{X}_n$ is the σ-algebra generated by $\{x_1, \ldots, x_n\}$. Condition (3) is not satisfied for (1) if φ_{n+1} depends on b_k, $k \neq n$. Moreover, even if $\varphi_{n+1} = e_{n+1}$ and $\{e_k\}$ are dependent, (3) is in general violated. This case is common in applications of the Robbins-Monro scheme and the Kiefer-Wolfowitz procedure.

It is reasonable to assume that φ_n depends on old b_k, $k \ll n$, only to a small extent. The following condition on Q_n therefore is imposed:

If $b_k \in D \subset R^m$, $k = n-N, \ldots, n$, then

$$\left| Q_n(b_n, \varphi_{n+1}(b(\cdot)) - Q_n(b, \bar{\varphi}_{n+1}(\bar{b}))\right| \leq w_n \left\{ \max_{n \geq k \geq n-N} \left| b_k - \bar{b}\right| + q(N)\right\} \tag{4}$$

where w_n is a random variable, measurable with respect to $\{e_k\}_1^{n+1}$ and $q(N) \to 0$ as $N \to \infty$.

Theorem 2. Suppose that assumtion (i) of theorem 1 is satisfied and that $\mathrm{Var}\, Q_n(b, \varphi_{n+1}) > \delta I$ (where Var denoted the covariance matrix of Q_n with respect to $\{e_k\}_1^{n+1}$). Assume that $x_n \to x^*$ with non zero probability. Then $f(x^*)=0$ and the matrix $\frac{d}{dx} f(x^*)$ has all eigenvalues in the left half plane ($\mathrm{Re} \leq 0$), provided $f(x)$ is twice differentiable

The proofs are given in the appendix.

A main result of the theorems is that a deterministic ODE (5) can be associated with the recursive stochastic algorithm (1). According to theorem 2 the algorithm can only converge to stable stationary points of (5). If a stationary point is globally asymptotically stable then x_n will converge w.p.1 to this point according to theorem 1. It can also be shown that the trajectories of (5) are in a certain sense the expected paths of (1) and the probability of deviation from the trajectories can be estimated, see Ljung (1974a).

Remark. Theorem 1 basically does not deal with convergence in a stochastic setting. A fixed realization for which (i) and (ii) hold is considered throughout the proof. Convergence of (1) is shown under these conditions. The theorem thus also can be applied for each realization and the convergence concept "with probability one" can be omitted. In particular, this means that the limit function $f(x)$, as well as the convergence point might be random variables: $f(x) = f(x,\omega)$. Then in condition (iii) the ODE $\frac{d}{dt} x = f(x,\omega)$ should be asymptotically stable with stationary point $x^*(\omega)$ for almost every ω, i.e. (iii) must hold with probability one.

The three conditions (i).(ii) and (iii) in theorem 1 for convergence can be called the noise condition, the boundedness condition and the stability condition, respectively, cf Ljung (1974b). In the standard convergence theorems on the Robbins-Monro and the Kiefer-Wolfowitz procedures it is possible to find similar interpretations of the conditions, see Ljung (1974a). Here, however, we have sought an explicit separation between the conditions, so that each one can be studied separately, and, if necessary, applied to specific problems using specially designed tools .

The stability condition concerns only the deterministic ODE (5). It can be checked using standard Lyapunov stability theory, see e.g. Krasovskij (1963). In practical situations, valuable insight into the stability properties of (5) may be obtained by numerical solution.

The boundedness condition (ii) is discussed in more detail in Ljung (1974a). It can be violated even if the ODE (5) is globally stable, either if $|Q_n(b,\varphi)|$ increases too rapidly with $|b|$ or if the variance of the noise increases so fast, that the random walk effect in (1) becomes predominating. Therefore, some conditions must be imposed that assure (ii) and hence rule out such cases. In e.g. Blum (1956) and Aizerman et al (1970) such criteria are based on functions that can be understood as Lyapunov functions for (5). Another approach, which is studied in e.g. Albert and Gardner (1967) and Ljung (1974a), is to consider algorithms where the estimates are projected into a bounded area.

From a practical point of view, the question of boundedness can be considered as uninteresting if the area D^0 can be taken as any bounded area. No implementation of (1) will allow that x_n tends to infinity. A simpleminded approach to solve the problem is to restart the algorithm in the origin, say, if $|x_n|$ becomes unrealistically large, say $|x_n| > C$. If the trajectory of the ODE (5) from the origin to x^* does not pass the area $|x_n| \geq C$, then this solution assures that the boundedness condition is satisfied.

The noise condition (i) is a purely stochastic criterion, which can, assuming the expectation exists, be rewritten

$$h_n \to 0 \quad \text{as } n \to \infty \text{ with probability one} \tag{6}$$

and

$$\lim_{n\to\infty} E\, Q_n(b^0, \bar{\varphi}_{n+1}(b^0)) = f(b^0) \tag{7}$$

where

$$h_{n+1} = h_n + \gamma_{n+1} (v_n - h_n) \qquad ; \quad h_0 = 0 \tag{8}$$

with

$$v_n = Q_n(b^0, \bar{\varphi}_{n+1}(b^0)) - E\, Q_n(b^0, \bar{\varphi}_{n+1}(b^0)).$$

The variable h_n thus defined is a weighted sum of random variables $\{v_k\}_1^{n-1}$ with zero mean values Convergence of such sums is relatively easy to treat, and a variaty of results in the present case are given in [11] and [12]. Here two results will be quoted

(a) If $\gamma_n = 1/n$ (which easily extends to the common case $\gamma_n = a_n/n$, $\lim_{n\to\infty} a_n = A$) a

$$|E\, v_n^T v_m| \leq C(n^\beta + m^\beta)/(1 + |n - m|^\alpha) \;, \quad 0 \leq 2\beta < \alpha < 1$$

then (6) holds.

This is an ergodicity result due to Cramer and Leadbetter (1967).

(b) If v_n can be represented on the form

$v_n = \sum_{k=0}^{\infty} g_{k,n}\, e_{n-k}$ where $|g_{k,n}| < \alpha_n \lambda^k$, $|\lambda| < 1$ and $\{e_k\}$ is a sequence of independent random variables such that $E\,|e_k|^p < C$, and $\{\alpha_n\}$ is a non-drecreasing sequence of scalars, (9)

and if $\{\gamma_n\}$ is decreasing with $\lim_{n\to\infty} \sup\,[1/\gamma_{n+1} - 1/\gamma_n] < \infty$ and

$$\sum_{1}^{\infty} \gamma_n^{p/2} \alpha_n^p < \infty \tag{10}$$

then (6) is satisfied.

This result is shown by estimating the moments of h_n and applying Chebysjev's inequality and the Borel-Cantelli lemma. It is consequently possible to trade off conditions on the sequence $\{\gamma_n\}$ against conditions on the moments of Q_n.

4. Applications

Theorem 1 and the results discussed above can be combined into a variety of convergence results. Suitable versions of the different conditions in theorem 1 can be chosen for specific applications. Various applications in the field of parameter estimation for dynamical systems and automatic control have been reported in e.g. [3], [12] and [13].

In this section the results will be applied to the Robbins-Monro scheme and the Kiefer-Wolfowitz procedure. These algorithms are of the type

$$x_{n+1} = x_n + \gamma_{n+1} Q(x_n, e_{n+1}) \tag{11}$$

which means that condition (4) takes the simple form

$$|Q_n(x_n, e_{n+1}) - Q_n(\bar{x}, e_{n+1})| \le K(e_{n+1})\, |x_n - \bar{x}| \tag{12}$$

i.e. it means that Q_n should be Lipschitz-continuous in $x \in D$.

The Robbins-Monro scheme is designed to solve the equation

$$E_e\, Q(x,e) = f(x) = 0$$

when observations $Q(x,e_k)$ are available for each x. The conditions on the noise

$$Q(x,e_k) - f(x)$$

and on the gain sequence $\{\gamma_n\}$ for convergence with probability one are then weaker according to theorem 1 and result (b) than reported in e.g. Blum (1954). The noise sequence may very well be dependent as long as it can be represented as in (9). If the p:th moment of the noise exists it is required that

$$\sum_1^\infty \gamma_n^{p/2} < \infty$$

which is satisfied e.g. for $\gamma_n = C\, n^{-\alpha}$ $2/p<\alpha\leq 1$. The condition usually given allows just $1/2<\alpha\leq 1$. Notice that slowly decreasing sequences $\{\gamma_n\}$ are of interest in practice to achieve fast convergence of the estimate sequence.
On the other hand, more regularity of the function Q with respect to x is required in order to apply theorem 1 than in Blum (1954).

The same result holds for the Kiefer-Wolfowitz procedure: Consider the problem to minimize a function P(x) (assumed to be three times differentiable and such that $P(x) \to \sup P$ as $|x|\to\infty$) when only noise corrupted measurements are available :

$$J(x,v_n) = P(x) + v_n$$

where $\{v_n\}$ is a sequence of random variables with zero mean values and uniformly bounded 2p moments and such that it can be represented as in (6). The Kiefer-Wolfowitz procedure , [7], (See also Kushner (1974)) gives

$$x_{n+1} = x_n + \gamma_{n+1}\, \Delta J(x_n, v_{mn}, \ldots, v_{mn+m}, a_n)$$

where ΔJ is an approximation of $\frac{d}{dx} P(x_n)$ based on measurements $J(x_n + a_n u_i, v_{mn+i})$, $\{u_i\}$ being the unit vectors in R^m. Applying the corollary of theorem 1, theorem 2 and result (b) for the noise, gives that with probability one x_n tends either to a local minimum point of P(x) ($\frac{d}{dx} P(x) = 0$ and $\frac{d^2}{dx^2} P(x)$ is positive (semi) definite) or to infinity as $n \to \infty$, provided

$$\sum_1^\infty \left[\gamma_n/a_n^2\right]^p < \infty \quad \text{and} \quad a_n \to 0 \text{ as } n \to \infty \tag{13}$$

The conditions (13) are weaker than the ones usually given, Blum (1954):

$$\sum_{1}^{\infty} a_n \gamma_n < \infty \qquad \sum_{1}^{\infty} (\gamma_n / a_n)^2 < \infty \text{ and } \quad a_n \to 0 \text{ as } n \to \infty .$$

and notice also that the results hold for dependent sequences $\{v_n\}$.

6. Conclusions

Recursive stochastic algorithms of a certain type have been considered in this paper. The class of algorithms contains stochastic approximation algorithms of the Robbins-Monro and the Kiefer-Wolfowitz type, but it includes also other types of estimation algorithms. It should, for example, be possible to treat algorithms for constrained optimization problems, recently introduced by Kushner [9],[10] in the same framework. Since the classical results for stochastic approximation algorithms are not applicable in many practical cases, it is believed that there is a need for general results valid for more complex approximation schemes.

A main idea in this paper has been to associate an Ordinary differential equation to the convergence problem. This technique gives results on convergence with probability one that in some respects extend previously given results for stochastic approximation algorithms and also have a wider applicability. It also gives a tool for analysing the properties of the algorithms, since the ODE can be investigated heuristically and numerically. The possibility of using theorem 2 to prove non-convergence for specific schemes is of particular interest, cf [13].

Appendix

Proof of theorem 1

Introduce a denumerable subset of the set in which (i) holds,

$$D_d = \{x^{(1)}, x^{(2)}, \ldots\} \subset D$$

which is dense in D.

Let Ω denote the sample space and denote the elements of Ω by ω.

Assumption (i) implies that $z_n(x^{(i)})$ converges w.p.1, i.e. for all $\omega \in \Omega^{(i)}$, where $\Omega^{(i)}$ has measure 1. Let

$$\Omega^* = \bigcap_1^\infty \Omega^{(i)} \cap \Omega_A$$

where Ω_A is the set of all realizations for which r_n converges, and condition (ii) holds Then also Ω^* has measure 1. In the rest of the proof only such realizations ω are considered that belong to Ω^*.

The basic idea of the proof is that the sequence of estimates $\{x_n\}$ obtained from algorithm (1) behaves like solutions of the ODE (5) asymptotically and locally. This result is shown in the following lemma:

Lemma A.1 : Let $\bar{x} \in D_d$ and $\omega \in \Omega^*$. Let $t \leq t_0$, where t_0 does not depend on $\bar{x}$ and ω. Define the sequence m(n,t) so that

$$\sum_{k=n}^{m(n,t)} \gamma_k \to t \quad \text{as} \quad n \to \infty .$$

Then, if $x_n(\omega)$ belongs to D,

$$x_{m(n,t)}(\omega) = x_n(\omega) + tf(\bar{x}) + R_1(t,n,\omega,\bar{x}) + R_2(t,n,\omega,\bar{x}) \tag{A.1}$$

where

$$|R_1(t,n,\omega,\bar{x})| < t\,K\,|x_n(\omega) - \bar{x}| + At^2$$

and

$$|R_2(t,n,\omega,x)| \to 0 \text{ as } n \to \infty .$$

Proof of lemma A 1 Consider the sequence $\{z_n(\bar{x},\omega)\}$ defined in the theorem ($b^0=\bar{x}$). Let $n \le j(n,\omega) \le m(n,t)$. Then

$$z_{j(n,\omega)}(\bar{x},\omega) = z_n(\bar{x},\omega) + \sum_{i=n+1}^{j(n,\omega)} \gamma_i \left\{ Q_{i-1}(\bar{x}, \bar{\varphi}_i(\bar{x})) - z_{i-1}(\bar{x},\omega)\right\}$$

Let now n tend to infinity. Since

$$\lim_{n\to\infty} z_{j(n,\omega)}(\bar{x},\omega) = \lim_{n\to\infty} z_n(\bar{x},\omega) = f(\bar{x})$$

it follows that

$$\sum_{i=n+1}^{j(n,\omega)} \gamma_i \; Q_{i-1}(\bar{x}, \bar{\varphi}_i(\bar{x})) - f(\bar{x}) \sum_{i=n+1}^{j(n,\omega)} \gamma_i = R_3(j(n),n,\omega,\bar{x}) \tag{A.2}$$

where $R_3(j(n),n,\omega\bar{x}) \to 0$ as $n \to \infty$.

Analogously

$$\lim_{n\to\infty} \sum_{n+1}^{m(n,t)} \gamma_i w_{i-1}(\omega) = r_c \; t \quad , \text{ where } r_c = \lim_{n\to\infty} r_n \tag{A.3}$$

Consider now

$$x_{j(n,\omega)}(\omega) = x_n(\omega) + \sum_{n+1}^{j(n,\omega)} \gamma_i \; Q_{i-1}[x_{i-1}(\omega), \varphi_i(x(\cdot))] =$$

$$= x_n(\omega) + \sum_{n+1}^{j(n,\omega)} \gamma_i Q_i(\bar{x}, \bar{\varphi}_i(\bar{x})) + \sum_{n+1}^{j(n,\omega)} \gamma_i \left\{ Q_{i-1}[x_{i-1}(\omega), \varphi_i(x(\cdot))] - Q_{i-1}(\bar{x}, \bar{\varphi}_i(\bar{x}))\right\} \tag{A.4}$$

The first sum of the RHS of (A.4) can be approximated using (A.2). To approximate the second sum, use condition (4) with $N= i - n$:

$$\left| Q_{i-1}[x_{i-1}(\omega), \varphi_i(x(\cdot))] - Q_{i-1}(\bar{x}, \bar{\varphi}_i(\bar{x}))\right| \le w_{i-1}(\omega) \left\{ \max_{n\le j\le i} |x_j(\omega) - \bar{x}| + q(i-n)\right\}$$

(Assume for the moment that x_i , $n \le i \le j(n,\omega)$ belong to D^0. This assumption is removed below.)

Hence

$$\left| \sum_{n+1}^{j(n,\omega)} \gamma_i \left\{ Q_{i-1}[x_{i-1}(\omega), \varphi_i(x(\cdot))] - Q_{i-1}(\bar{x}, \bar{\varphi}_i(\bar{x})) \right\} \right| \leq$$

$$\leq \max_{n \leq i \leq j} \left| x_i(\omega) - \bar{x} \right| \sum_{n+1}^{m(n,t)} \gamma_i w_{i-1}(\omega) + \sum_{n+1}^{m(n,t)} \gamma_i w_{i-1} q(i-n) \leq$$

$$\leq \max_{n \leq i \leq m(n,t)} \left| x_i(\omega) - \bar{x} \right| \left\{ r_c t + R_4(t,n,\omega) \right\} + R_8(t,n,\omega) \leq$$

$$\leq \left[\max_{n \leq i \leq m} |x_i(\omega) - x_n(\omega)| + |x_n(\omega) - \bar{x}| \right] \left[r_c t + R_4(t,n,\omega) \right] + R_8(t,n,\omega) \qquad \text{(A.5)}$$

where $R_4(t,n,\omega) \to 0$ as $n \to \infty$ according to (A.3) and $R_8(t,n,\omega) \to 0$ as $n \to \infty$ as $N \to \infty$, since $q(N) \to 0$ as $N \to \infty$.

Assume that

$$\max_{n \leq i \leq m(n,t)} \left| x_i(\omega) - x_n(\omega) \right| = S(n,t,\omega)$$

is attained for $i = j^*(n,\omega)$. Then taking $j = j^*$ and inserting (A.5) and (A.2) in (A.4) for this j gives

$$S(n,t,\omega) \left[1 - r_c t - R_4(t,n,\omega) \right] \leq |f(\bar{x})| t + R_3(j(n),n,\omega,\bar{x}) + |\bar{x} - x_n(\omega)| [r_c t + R_4] + R_8$$

For sufficiently small t, $t < t_0$, and sufficiently large n, $r_c t + R_4(t,n,\omega) < 1/2$ and since $x_n(\omega)$ and $\bar{x} \in D$ we have

$$|x_n(\omega) - \bar{x}| < C_1$$

Hence

$$S(n,t,\omega) \leq 2[\, |f(\bar{x})| + r_c C_1]\, t + R_3(j(n),n,\omega,\bar{x}) + C_1 R_4 + R_8 = C_2 t + R_5(j(n),n,\omega,\bar{x})$$

where $R_5(j(n),n,\omega,\bar{x}) \to 0$ as $n \to \infty$.

Now choose $j(n,\omega) = m(n,t)$ in (A.4) which gives, using (A.2) and (A.5),

$$|x_{m(n,t)}(\omega) - x_n(\omega) - tf(\bar{x})| \leq R_3(t,n,\omega,\bar{x}) + |f(\bar{x})| \left[\sum_n^m \gamma_i - t \right] +$$

$$+ \left\{ C_2 t + R_5(t,n,\omega,\bar{x}) + |\bar{x} - x_n(\omega)| \right\} \left\{ r_c t + R_4(t,n,\omega) \right\} = C_2 r_c t^2 + r_c t |\bar{x} - x_n(\omega)| +$$

$$+ R_2(t,n,\omega,\bar{x})$$

where $R_2(t,n,\omega,\bar{x}) \to 0$ as $n \to \infty$.

It remains now only to remove the assumption $x_i \in D^o$, $n \leq i \leq m(n,t)$. If this assumption does not hold, let $i = \bar{j}(n,\omega)$ be the first time $x_i \notin D^o$. Then apply the results above to $j(n,\omega) = \bar{j}(n,\omega)$, which gives

$$|x_{\bar{j}(n,\omega)} - x_n| \leq C_4 t + R_6 \qquad \text{where } R_6 \to 0 \text{ as } n \to \infty.$$

For sufficiently small t, this contradicts the definition of $\bar{j}$.

∎

It follows from the converse stability theorem (see Krasovskij(1963)) that assumption (iii) implies the existence of a function V(x) with properties

$V(x)$ is infinitely differentiable

$0 \leq V(x) \leq 1 \iff x \in D_1$ and $V(x) = 0 \iff x = x^*$

$\frac{d}{dt} V(x) = V'(x)f(x)$ is negative definite in D_1

Consider from now on a fixed realization $\omega \in \Omega^*$. All variables below depend on ω, but this argument will be suppressed. From condition (ii) there exists at least one cluster point $\tilde{x}$ to the sequence $\{x_k\}$ in D. Hence there is a subsequence $\{x_{n_k}\}$ that tends to $\tilde{x}$ as $k \to \infty$. Since D_d is dense in D, there is for arbitrary $\epsilon > 0$ an element $\bar{x} = \bar{x}(\tilde{x},\epsilon) \in D_d$ such that $|\bar{x} - \tilde{x}| < \epsilon/2$.

Consider now

$$V[x_{m(n_k,t)}] - V[x_{n_k}]$$

where m is defined as in Lemma A 1. Denote $n_k = k'$ and $m(n_k,t) = k''$, and use the mean value theorem. This gives

$$V(x_{k''}) - V(x_{k'}) = V'(\xi_k)(x_{k''} - x_{k'}) = V'(\bar{x})(x_{k''} - x_{k'}) + (\xi_k - \bar{x})^T V''(\xi_k')(x_{k''} - x_{k'}) \qquad (A.6)$$

where

$$\xi_k = x_{k'} + \theta_1(x_{k''} - x_{k'}) \quad ; \quad \xi_k' = x_{k'} + \theta_2(\xi_k - x_{k'}) \quad , \quad 0 \leq \theta_i \leq 1 .$$

Now apply Lemma A.1 to $x_{k''} - x_{k'}$ which gives

$$x_{k''} - x_{k'} = t\, f(\bar{x}) + R_1(t,n_k,\bar{x}) + R_2(t,n_k,\bar{x})$$

Insert this in (A.7):

$$V(x_{k''}) - V(x_{k'}) = tV'(\bar{x})f(\bar{x}) + R_7(t,n_k,\bar{x})$$

where

$$R_7(t,n_k,\bar{x}) = (\xi_k - x)^T V''(\xi_k')(x_{k''} - x_{k'}) + V'(\bar{x})[R_1(t,n_k,\bar{x}) + R_2(t,n_k,\bar{x})]$$

Now suppose that the cluster point $\tilde{x}$ is different from the desired convergence point x^*. Then $V'(\tilde{x})f(\tilde{x}) = -\delta$, $\delta > 0$. By choosing ϵ, t and K properly, it follows that

$$V[x_{m(n_k,t)}] < V(\tilde{x}) - t\,\delta/2 \qquad k > K \tag{A.7}$$

<u>Lemma A 2</u> Suppose (A.7) holds for any subsequence $\{x_{n_k}\}$ that converges to a point different from x^*. Then

$$\liminf_{n\to\infty} V(x_n) = 0 \quad : \text{i.e.} \quad x^* \text{ is a cluster point} \tag{A.8}$$

<u>Proof</u> : Consider inf $V(x)$ taken over all cluster points in D Let this value be U. Since the set of cluster points in D is compact, there exists a cluster point $\hat{x}$, such that $V(\hat{x}) = U$. If now $U > 0$, $V'(\hat{x})f(\hat{x})$ will be strictly negative $(=-\delta)$ and from (A.7) $V(x_k)$ takes a value less than $U - \delta t/2$ infinitely often, which contradicts $U > 0$ being the infimum. Hence $U = 0$, which means that x^* is a cluster point. ∎

To conclude the proof of the theorem it now remains only to prove the following lemma:

<u>Lemma A.3</u>. From (A.7) and (A.8) it follows that

$$\limsup_{n\to\infty} V(x_n) = 0 \quad , \text{ i.e. } \quad x^* \text{ is the only cluster point and } \quad x_k \to x^* \text{ as } n \to \infty.$$

<u>Proof:</u> If $x_n \in D$ the difference

$$|x_{n+1} - x_n| = |\gamma_{n+1} Q_n'[x_n, \varphi_{n+1}(x(\cdot))]| \leq \gamma_{n+1}|Q_n(x^*, \bar{\varphi}_{n+1}(x^*))| + \gamma_{n+1} w_n\{|x_n - x^*| + q(1)\} \leq$$

$$\leq |z_{n+1}(x^*) - z_n(x^*)| + \gamma_{n+1}|z_n(x^*)| + \{|x_n - x^*| + q(1)\}\{|r_{n+1} - r_n| + \gamma_{n+1} r_n\}$$

tends to zero as n increases, since $z_n(x^*)$ and r_n converge. Suppose that

$$\limsup_{n\to\infty} V(x_n) = A > 0$$

Consider the interval $I = [A/3, 2A/3]$. (If $A > 1$ take $I = [1/3, 2/3]$.) This interval is then crossed "upwards" and "downwards" infinitely many times. Since the step size $x_{n+1} - x_n$ tends to zero when $x_n \in D$, there will be a subsequence of $V(x_n)$ that belongs to I. Consider now such a special convergent sequence of "upcrossings": Let $\{x_{n'_k}\}$ be defined as follows:

$$V(x_{n'_k-1}) < A/3 \qquad V(x_{n'_k}) \geq A/3 \qquad V(x_{n'_k+s_k}) > 2A/3$$

where s_k is the first s for which $V(x_{n'_k}+s) \notin I$. Let $x_{n'_k} \to \tilde{x}$ as $k \to \infty$. Clearly $V(\tilde{x}) = A/3$ Now, from (A.7),

$$V(x_{m(n'_k,t)}) < A/3 - \delta t/2$$

This means that $V(x_{n'_k+s_k}) \notin I$ where $s_k = m(n'_k,t) - n'_k$. But, if t is sufficiently small, no s, smaller than s_k can have made $V(x_{n'_k+s}) > 2A/3$, according to Lemma A.1 and the continuity of V. This contradicts the definition of the subsequence n'_k. Hence no interval I can exist. A must be zero and the lemma follows.

∎

Proof of Theorem 2

The assertion $f(x^*)=0$ follows directly from Lemma A.1. To show the second assertion, consider first the special case

$$Q_n(x, \varphi_{n+1}) = Ax + e_{n+1}$$

where A is a m|m matrix and $\{e_n\}$ is a sequence of random variables with zero mean values. Suppose that A has an eigenvalue λ with $\mathrm{Re}\,\lambda > 0$, and let L be a corresponding left eigenvector. Let $\tau_n = Lx_n$, and $\epsilon_n = Le_n$. The condition on Var Q_n implies that τ_n is not identically zero. Then algorithm (1) can be written

$$\tau_{n+1} = \tau_n + \gamma_{n+1}(\lambda\tau_n + \epsilon_{n+1})$$

and

$$\tau_{m+n} = \Gamma_{n,m}\left\{\tau_n + \sum_{n+1}^{n+m} \beta_k^{n+m} \epsilon_k\right\}$$

where

$$\Gamma_{n,m} = \prod_{n+1}^{n+m}(1+\lambda\gamma_k) \sim \exp\left(\lambda\sum_{n+1}^{n+m}\gamma_k\right) \quad \text{and} \quad \beta_k^{n+m} = \gamma_k \prod_{n+1}^{k}(1+\lambda\gamma_j)^{-1}$$

Since τ_n and the sum of random variables is not completely correlated (according to (i)) and $\Gamma_{n,m}$ tends to infinity as m increases, it follows that τ_k will, with probability one, not tend to zero as k tends to infinity. Hence x_k will not converge to 0 ($=x^*$) with non zero probability.

The general case is proven by linearization around x^*, and the additional terms are taken care of using (4). Like in the proof of theorem 1, this leads to several technicalities, and the calculations are therefore omitted. ∎

References

[1] Aizerman,M A , Braverman,E.M. and Rozonoer,L.I. (1970). Metod potentsialnych funktsij v teorii obuchenija mashin (The Method of Potential Functions in the Theory of Machine Learning). Izd. Nauka, Moscow (In Russian).

[2] Albert,A.E and Gardner,L.A. (1967). Stochastic Approximation and Nonlinear Regression. Research Monograph 42, The MIT press, Cambridge, Mass.

[3] Åström,K.J ,Borisson,U.,Ljung,L. and Wittenmark,B. (1975). Theory and applications of adaptive regulators based on recursive parameter estimation. Preprints of the 5th IFAC World Congress, Boston.

[4] Blum,J. (1954). Multidimensional stochastic approximation methods. Ann. Math. Statist. 25, 737-744.

[5] Burkholder, D.L. (1956). On a class of stochastic approximation procedures. Ann. Math Statist. 27 1044-1059

[6] Dvoretzky, A (1956). On stochastic approximation. Proc. Third Berkeley Symp Math Statist. Prob. 1 39-55.

[7] Kiefer,J. and Wolfowitz,J (1952). Stochastic estimation of the maximum of a regression function. Ann.Math. Statist. 23 462-466.

[8] Krasovskij, N N. (1963). Stability of motion. Stanford University Press, Stanford,Calif.

[9] Kushner,H.J. (1974). Stochastic approximation algorithms for constrained optimization problems. To appear Ann. Statist.

[10] Kushner,H.J. and Gavin, T. (1974). Stochastic approximation type methods for constrained systems: Algorithms and numerical results. IEEE Trans. Aut. Control AC-19 No 4 349-358.

[11] Ljung,L (1974a). Convergence of recursive stochastic algorithms. Report 7403, Division of Automatic Control, Lund Institute of Technology,Lund, Sweden.

[12] Ljung,L. (1974b). Convergence of recursive stochastic algorithms. Preprints of the IFAC Symp. on Stochastic Control in Budapest.

[13] Ljung,L., Söderström,T. and Gustavsson,I. (1974). Counterexamples to general convergence of a commonly used recursive identification method. Submitted to IEEE Trans. Aut. Control.

[14] Robbins,H and Monro,S. (1951). A stochastic approximation method. Ann.Math. Statist.22 400-407.

[15] Wasan,M.T. (1969). Stochastic approximation. Cambridge, University Press.