Convergence of Recursive Stochastic Algorithms

Ljung, Lennart

1974

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

REPORT 7403
FEBRUARY 1974

Convergence of Recursive Stochastic Algorithms

LENNART LJUNG

LTH

Division of Automatic Control · Lund Institute of Technology

Convergence of Recursive Stochastic Algorithms

Lennart Ljung

ABSTRACT

Convergence with probability one for a class of recursive, stochastic algorithms is considered. The class contains stochastic approximation algorithms like the Robbins - Monro scheme and the Kiefer - Wolfowitz procedure. It also contains other estimation and control algorithms that are common in stochastic control theory.

It is shown that the problem of convergence can be separated into a deterministic problem and a stochastic one. The analysis of the deterministic problem reduces to stability analysis of an ordinary differential equation (ODE). For the stochastic problem it is sufficient to show consistency for a simple algorithm that estimates the mean value of a random variable.

Using this technique, the usual conditions for convergence can be extended. Correlated observations can be treated and the conditions on the gain sequence can be traded off against conditions on the moments of the noise.

The behaviour of the algorithm can also be described using the ODE that is associated with the convergence problem. Based on the trajectories of the ODE future values generated by the algorithm can be predicted. Numerical solution of the ODE therefore is a valuable tool to analyse the asymptotic properties of the recursive stochastic algorithm.

TABLE OF CONTENTS.

1. ALGORITHMS.

Recursive algorithms where stochastic observations enter, occur in many fields of applications, such as estimation, filtering and control theory. In this report convergence of a certain class of recursive algorithms will be considered. The class includes stochastic approximation algorithms and also other algorithms that are common in control applications. The obtained results are more general than earlier reported ones.

In this chapter some examples of recursive, stochastic algorithms that fit in the formulation are given. Chapter 2 contains a short survey and classification of previously reported convergence criteria. A theorem that separates the considered problem of convergence into a deterministic problem and a purely stochastic one is shown in Chapter 3. The stochastic part is further discussed in Chapter 4. In Chapter 5 conditions are given that assure bounded estimates. In Chapter 6 the question of convergence rate is discussed and a theorem is shown, which connects the behaviour of the recursive algorithm to that of a deterministic ordinary differential equation. Finally, in Chapter 7, the results are summarized and discussed.

1.1. A General Recursive Algorithm.

A general version of a recursive algorithm can be written

$$x_n = x_{n-1} + H_n(x_{n-1},\varphi_n) \qquad x_0 = x^0 \tag{1.1}$$

where $\{x_n\}$ is a sequence of vectors. These vectors will be called estimates and it is supposed that they are estimates of some desired or optimal value x^*, which is independent of n. It is then desirable that the sequence $\{x_n\}$ tends to x^* as n tends to infinity.

The correction $H_n(x_{n-1},\varphi_n)$ is a function of the previous estimate x_{n-1} and of an observation φ_n obtained at time n.

The observations are in general functions

$$\varphi_n = \varphi_n(e_n,e_{n-1},\dots,e_0;x_{n-1},\dots,x_0) \tag{1.2}$$

of the previous estimates $\{x_i\}$ and of a sequence of vector valued random variables $\{e_i\}$ that is independent of $\{x_i\}$. In case the experimenter has some test signal at his disposal, this can be included in the sequence $\{e_i\}$. In many cases the observations do not depend on previous estimates. Then the sequence $\{e_i\}$ can be taken as the observations themselves:

$$\varphi_n = e_n \tag{1.3}$$

Another common special case is that the observation depends only on x_{n-1}, i.e.

$$\varphi_n = \varphi_n(x_{n-1},e_n) \tag{1.4}$$

Since it is desired that $\{x_i\}$ converges, the correction $H_n(x_{n-1},\varphi_n)$ caused by a single stochastic observation φ_n must tend to zero as n tends to infinity, i.e.

$$H_n(x_{n-1},\varphi_n) = \gamma_n\tilde{H}_n(x_{n-1},\varphi_n)$$

where $\{\gamma_n\}$ is a sequence of positive scalars tending to zero. The variables may either be predetermined scalars, which gives the algorithm

$$x_n = x_{n-1} + \gamma_n \tilde{H}_n(x_{n-1},\varphi_n) \tag{1.5}$$

or functions of the observations, giving

$$x_n = x_{n-1} + \gamma_n(\varphi_n,\ldots,\varphi_0)\tilde{H}_n(x_{n-1},\varphi_n) \tag{1.6}$$

In some cases it is convenient to specially treat the case when $\tilde{H}_n$ is multiplied by a matrix $S_n(\varphi_n,\ldots,\varphi_0)$:

$$x_n = x_{n-1} + \gamma_n S_n(\varphi_n,\ldots,\varphi_0)\tilde{H}_n(x_{n-1},\varphi_n) \tag{1.7}$$

In (1.6) and (1.7) γ_n and S_n respectively have to be updated recursively. Therefore, it is always possible to rewrite (1.6) and (1.7) in the form (1.5) by extending the estimate vector x_n. However, in some applications it is more favourable to directly consider (1.6) and (1.7).

The results of this report are mainly concerned with the cases (1.3) and (1.4). Then $\tilde{H}_n(x_{n-1},\varphi_n)$ can be written explicitely as a function of x_{n-1} and e_n

$$\tilde{H}_n\big(x_{n-1},\varphi_n(x_{n-1},e_n)\big) = Q_n(x_{n-1},e_n)$$

Hence the basic algorithm to be considered here is

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1},e_n) \tag{1.8}$$

The techniques that are used also apply for the general case (1.2) if the influence of old estimates $\{x_i\}$ on φ_n decreases sufficiently fast. Certain algorithms with φ_n as in (1.2) are treated in Ljung-Wittenmark (1974).

In Sections 1.2 and 1.3 several examples of algorithm (1.1) are given. These examples will, hopefully, clarify the classification of H_n, and show that a number of control theory applications fit in the structures (1.1) - - (1.8).

1.2. Stochastic Approximation Algorithms.

> "Stochastic approximation is concerned with schemes converging to some sought value when, due to the stochastic nature of the problem, the observations involve errors." (Dvoretzky (1956))

The original scheme was devised by Robbins and Monro (1951).

Example 1.1 - The Robbins-Monro (RM) scheme.

Consider the problem to solve

$$E_e Q(x,e) = 0 \tag{1.9}$$ [1]

for x. Let the solution be x^*. Observations $\varphi_n = Q(x,e_n)$, n=1,... are available for any x. Robbins-Monro proved that under certain conditions the scheme

$$x_n = x_{n-1} + \gamma_n Q(x_{n-1},e_n) \tag{1.10}$$

converges to x^*. These results were extended by Blum (1954b) to the multidimensional case. The conditions for convergence are discussed in Chapter 2. Let us just remark that among the necessary conditions we have

[1] "E_e" denotes expectation with respect to e, while the vector x is considered as a fixed parameter

$$\sum_1^\infty \gamma_n = \infty \tag{1.11a}$$

$$\sum_1^\infty \gamma_n^2 < \infty \tag{1.11b}$$

In this report it is shown that (1.11b) can be replaced by a weaker condition.

Comparing with the formalism of the previous section the observation φ_n has the form (1.4), and (1.10) is thus a simple special case of (1.8). The sequence $\{\gamma_n\}$ is deterministic as in (1.5). In some applications, see Section 1.3, $Q(x_{n-1},e_n)$ is itself not a primary observation, but formed from observations and from x_{n-1}.

As a simple example of the RM scheme, consider the problem to find the mean value of a stochastic variable e. Let $Ee = x^*$. This value x^* can be found as the solution of

$$E_e(e-x) = 0$$

Now take in (1.9) $Q(x,e) = e - x$ and apply the RM scheme:

$$x_n = x_{n-1} + \gamma_n(e_n-x_{n-1}) \tag{1.12}$$

With the choice $\gamma_n = 1/n$, which clearly satisfies (1.11) we obtain

$$x_n = \frac{1}{n}\sum_{k=1}^{n} e_k$$

which is an efficient estimate of x^*.

□

In many applications it is interesting to minimize a function

$$E_v J(x,v) = P(x) \tag{1.13}$$

with respect to x. If the derivative $\frac{\partial}{\partial x} J(x,v)$ can be calculated the stationary points of (1.13) can be found as solutions of

$$E_v \left(- \frac{\partial}{\partial x} J(x,v) \right) = 0$$

This is a problem that can be solved using the RM scheme. If the derivative cannot be calculated, it seems natural to replace it with some difference approximation. This was suggested by Kiefer and Wolfowitz (1952):

Example 1.2 - The Kiefer-Wolfowitz (KW) procedure.

Consider the problem to minimize (1.13) with respect to x. Observations $J(x,v_n)$, $n = 1,\ldots,$ of the criterion are available for each chosen x. The distribution of v_n is independent of x. Kiefer-Wolfowitz (1952) and Blum (1954b) suggested that the minimizing point x^* should be estimated recursively.

$$x_n = x_{n-1} + \gamma_n \bar{J}(x_{n-1}, a_n, \bar{v}_n)/a_n \tag{1.14}$$

where

$$\bar{J}(x,a,\bar{v}) = \big(J(x-au_i, v_{j_1}) - J(x,v_i), \ldots, \\ J(x-au_m, v_{j_m}) - J(x,v_i) \big)$$

m is here the dimension of the vector x and $\{u_i\}$ are the unit vectors in R^m. Consequently, to advance one step with (1.14), m+1 measurements are made and m+1 outcomes of the noise v enter.

Blum (1954b) has shown convergence with probability one for (1.14) under certain conditions. The conditions on the sequences $\{\gamma_n\}$ and $\{a_n\}$ are

$$\lim_{n\to\infty} a_n = 0; \quad \sum_1^\infty \gamma_n = \infty, \quad \sum_1^\infty a_n\gamma_n < \infty, \quad \sum_1^\infty(\gamma_n/a_n)^2 < \infty$$

In this procedure the observation φ_n consists of $J(x_{n-1}+$ $+au_i, v_{ji})$ and is of the type (1.4). The noise vector e_n in (1.4) must be chosen to include m+1 values of v. Clearly (1.14) is a special case of the basic algorithm (1.8). Notice that $Q_n(x_{n-1},e_n) = J(x_{n-1},a_n,v_n)/a_n$ actually is time varying in this case.

□

In the KW procedure the minimization is performed with a steepest descent method. Recently, Kushner (1972), (1973), and Kushner-Gavin (1973) have considered more general minimization routines. This approach seems to promise better numerical behaviour of the algorithms.

Dvoretzky (1956), considers a version of stochastic approximation algorithms, which includes both the RM and KW schemes. He writes the recursion

$$x_{n+1} = T_n(x_n,\dots,x_1) + e_{n+1} \tag{1.15}$$

where

$$E(e_{n+1} \mid x_1,\dots,x_n) = 0$$

Under certain conditions on the transformations T_n, (1.15) converges with probability one. Dvoretzky considers primarily the case with scalar valued x. Extensions to the multidimensional case has been given by Derman-Sacks (1959). Also (1.15) fits in the formulation (1.1) with general observations (1.2).

The term "stochastic approximation algorithms" will be used frequently in this report. By this term will be meant the algorithms discussed in this section, in particular the RM and KW procedures.

1.3. Applications to Control Theory.

Learning systems.

Tsypkin (e.g. 1968, 1971, 1973) has applied stochastic approximation techniques to a broad variety of problems in control theory. The approach is known as "learning systems". In short, the "goal of learning" is defined as to minimize some criterion P(x) with respect to the vector x. However, only noisy observations J(x,v) of the criterion are available, where

$$E_v J(x,v) = P(x) \tag{1.16}$$

If the derivative $\frac{\partial}{\partial x} J(x,v)$ can be formed the RM scheme can be applied to

$$E_v\left[-\frac{\partial}{\partial x} J(x,v)\right] = 0$$

If not, the KW procedure can be applied to (1.16).

In this framework estimation and identification problems ("learning models"), adaptive systems, supervised and unsupervised pattern recognition ("learning pattern recognition systems" and "Self-learning systems of classification") etc. can be treated. Similar approaches are considered by several other authors, see e.g. Fu (1969), Saridis et al (1969), Sakrison (1967).

Basically, the learning algorithms rely upon the RM scheme (or KW scheme) and convergence criteria for these can be applied. However, it may happen that the usual criteria are not applicable in a given case.

An approach that is related to stochastic approximation is suggested by Aizerman, Braverman and Rozonoer (e.g. 1964ab, 1970). Their "Potential Function Method" can be applied to various problems in machine learning. The connection with the RM scheme is discussed in Aizerman et al (1965).

Estimation.

A common problem is to estimate the coefficients in a difference equation

$$y(t+1) + a_1y(t) + \dots + a_my(t-m+1) =$$

$$= b_0u(t) + \dots + b_mu(t-m) + v(t+1) \qquad (1.17)$$

where $\{v(t)\}$ is a sequence of independent, random variables with zero mean values. The variables y, u and e are scalars. Measurements of the input u and of the output y are available.

Let x^* denote the vector of true values:

$$x^* = (a_1,\dots,a_m,b_0,\dots,b_m)^T$$

and let

$$\xi_t = \left(-y(t),\dots,-y(t-m+1),u(t),\dots,u(t-m)\right)^T$$

denote the vector of observations. Then (1.17) can be written

$$y(t+1) - \xi_t^T x^* = v(t+1) \tag{1.18}$$

A reasonable "goal of learning" for this estimation problem is to find the vector x that minimizes the criterion

$$P(x) = E\left(y(t+1) - \xi_t^T x\right)^2 \tag{1.19}$$

This function can be minimized by taking the derivative and applying the RM scheme:

$$-P'(x)=E_e Q(x,e) = 0 \tag{1.20}$$

where

$$Q(x,e_{t+1}) = \xi_t y(t+1) - \xi_t \xi_t^T x \;; \quad e_{t+1} = \left(\xi_t, y(t+1)\right)$$

Then

$$x_{n+1} = x_n + \gamma_{n+1}\left\{\xi_n y(n+1) - \xi_n \xi_n^T x_n\right\} \tag{1.21}$$

The variables γ_n can be chosen in several ways. Let it first be a sequence of scalars. This stochastic approximation version of least squares estimation is treated e.g. by Wieslander (1969), Tsypkin (1973) and Mendel (1973), where also other variants are discussed.

For normalization reasons γ_n is often chosen as

$$\gamma_{n+1} = \frac{1}{n}\left(\xi_n^T\xi_n\right)^{-1}$$

or

$$\gamma_{n+1} = \left(\sum_{k=0}^{n} \xi_k^T\xi_k\right)^{-1}$$

Comparing with Section 1.1, the observation $\varphi_{n+1} = (\xi_n, y(n+1))$ does not depend on $\{x_i\}$ and is consequently of the type (1.3). The algorithm (1.21) with the discussed choices of $\{\gamma_n\}$ is a special case of (1.6).

Generally speaking, when we are faced with a problem (1.9), to solve

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

for x, it seems desirable to solve

$$\frac{1}{n}\sum_{k=1}^{n} Q(x,e_k) = 0 \tag{1.22}$$

for x at time n. This is likely to give a good estimate x_n. In the case (1.4) the basic observations $\varphi(x_{k-1},e_k)$, from which $\tilde{H}(x_{k-1},\varphi_k) = Q(x_{k-1},e_k)$ is formed, depend on x_{k-1}. It is then not clear how the function $Q(x,e_k)$ could be formed from $\varphi(x_{k-1},e_k)$. Hence, it is not possible to solve (1.22).

In the present estimation problem, however, (1.22) can be solved, since the function $Q(x,e_{k+1}) = (\xi_k y(k+1) - \xi_k \xi_k^T x)$ can be formed for any x, as soon as y(k+1) and ξ_k are known. Furthermore, in this special case Q(x,e) is linear in x. Then it is actually possible to solve (1.22) recursively with a special (matrix) choice of γ_n in (1.21):

Example 1.3 - Real time least squares.

For the estimation problem (1.20) Eq. (1.22) is

$$\frac{1}{n}\sum_{k=1}^{n} Q(x,e_k) = \frac{1}{n}\sum_{k=1}^{n} \xi_k[y(k+1) - \xi_k^T x] = 0 \qquad (1.23)$$

which can be solved for x if only

$$\frac{1}{n+1}\sum_{k=0}^{n} \xi_k \xi_k^T = R_n \qquad \text{(2m×2m matrix)}$$

is known. The solution of (1.23) can be written recursively (Åström (1968)):

$$x_{n+1} = x_n + \frac{1}{n+1} R_n^{-1}\left[\xi_n y(n+1) - \xi_n \xi_n^T x_n\right] \qquad (1.24)$$ [1)]

In Eq. (1.24) the correction can be written

$$H_n = \gamma_n S_n(e_n,\dots,e_0) \cdot Q(x_{n-1},e_n)$$

where $S_n = R_n^{-1}$ and $\gamma_n = 1/n$, which is of type (1.7).

[1)] This expression holds only asymptotically as n tends to infinity.

parameters are assumed to be x. This value can be obtained as the solution of

$$\hat{y}(s+1|s,x) + \hat{c}_1\hat{y}(s|s-1,x) + \dots + \hat{c}_m\hat{y}(s-m+1|s-m,x) =$$
$$= (\hat{c}_1-\hat{a}_1)y(s) + \dots + (\hat{c}_m-\hat{a}_m)y(s-m+1) +$$
$$+ \hat{b}_0u(s) + \dots + \hat{b}_mu(s-m) \qquad (1.27)$$

where s = 0,...,t. Suitable initial values must be chosen. Eq. (1.26) corresponds to minimization of the prediction error, and clearly (1.19) is a special case ($c_i = 0$) of (1.26).

One way to determine the estimate x at time n is to minimize

$$\frac{1}{n}\sum_{1}^{n}[y(t) - y(t|t-1,x)]^2 \qquad (1.28)$$

with respect to x. This is possible to do, using (1.27) if the observations {u(t),...,u(0),y(t),...,y(0)} are available. However, in the general case it is not possible to write the sequence of estimates that minimize (1.28) recursively as in Example 1.3. This is due to the fact that the criterion is a more complex function of x in this case.

The RM scheme can be applied to the derivative of (1.26). This gives

$$x_{t+1} = x_t - \gamma_{t+1}\left\{\frac{\partial}{\partial x}\,\hat{y}(t+1|t,x_t)[\hat{y}(t+1|t,x_t) - y(t+1)]\right\} \qquad (1.29)$$

It is important to notice that it is possible to exactly calculate $\hat{y}(t+1|t,x_t)$ according to (1.27) only if infinitely many old y(s) and u(s) are known. This means that ever, for adaptive systems, calculation of this derivative requires that the characteristics of the system are known. Therefore this approach cannot be applied straightforwardly. The KW procedure can be used instead.

A different approach is to estimate the system dynamics and use the estimates for design of the regulator. The regulator is updated recursively. In Ljung (1972) is discussed how such adaptive structures fit in the formulation (1.1).

Example 1.5 - A class of self-tuning regulators.

Suppose algorithm (1.21) or (1.24) is used to estimate the system dynamics. The estimate at time t can be used to determine the next input from old input output data:

$$u(t+1) = h(x_t, \xi_t) \tag{1.30}$$

Åström-Wittenmark (1973) consider a minimum variance control law, which gives several nice features to the resulting adaptive (or self-tuning) regulator.

It is important to notice that since u(t) depends on x_t, also the observations ξ_t and y(t+1) will depend on x_t, $x_{t-1}, \ldots, x_0$. This means that φ has the general form (1.2) and Q defined in (1.20)

$$Q(x_t, e_{t+1}) = \tilde{H}(x_t, \varphi_{t+1}) \xi_t(x_t, x_{t-1}, \ldots, x_0) \cdot$$

$$\cdot \left[y(t+1; x_t, \ldots, x_0) - \xi_t^T(x_t, \ldots, x_0) x_t\right]$$

is no longer a linear function of x_t. In particular (1.24) no longer gives the estimates that are solutions to (1.22).

The convergence properties of the class of self-tuning regulators under consideration are treated in detail in Ljung-Wittenmark (1974).

□

It is of interest to compare Examples 1.3, 1.4 and 1.5. In all these cases the objective is to solve

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

for x. In Examples 1.3 and 1.4 the basic observations {y(t)} and {u(t)} do not depend on x. From these observations the function

$$Q_t(x,e_t) \quad \text{or} \quad Q_t(x,e_t,\ldots,e_0)$$

can be formed for any x. It is then possible to solve

$$\frac{1}{n}\sum_{1}^{n} Q(x,e_t) = 0 \tag{1.31}$$

for x, which gives the estimate x_n.

In Example 1.3, because of the linear dependence of x in Q(x,e), it is possible to calculate the sequence of solutions to (1.31) recursively as in (1.24). In Example 1.4 Q depends on x in a more complex way, and the solution to (1.31) cannot be written as recursions containing a fixed and finite number of observations. In Example 1.5 also the basic observations y(t) and u(t) depend on x_t. The estimate at time t is used for a decision that affects future observations, which is a typical feature of an adaptive system. It is then not possible to calculate $Q(x,e_t)$ for arbitrary x and (1.31) cannot be solved.

An automatic (self learning)classifier is an example of an adaptive way to estimate mean values of two stochastic variables.

Example 1.6 - Self learning classification (Unsupervised pattern recognition).

A classifier receives scalar valued signals e_n, which may belong to either of two a priori unknown classes A and B. The classifier must find a classification rule, i.e. a number c_n such that e. is classified as A if $e_n \leqslant c_n$ and B otherwise. The number c_n can e.g. be determined as follows

$$c_n = (x_n^A + x_n^B)/2$$

where

$$x_n^A = \begin{cases} x_{n-1}^A + \gamma_n(e_n - x_{n-1}^A) & \text{if } e_n \text{ is classified as A} \\ x_{n-1}^A & \text{otherwise} \end{cases} \qquad (1.32)$$

x_n^B is defined analogously. Clearly, x_n^A is the mean value of the outcomes classified as A. This scheme is proposed e.g. by Tsypkin (1968).

Algorithm (1.32) can be considered as an RM scheme to solve

$$x^A = E\{e|e \in A(x)\} = E\{e|e \leqslant \tfrac{1}{2}(x^A+x^B)\} = E_e Q^A(x^A,x^B,e) = 0$$

$$x^B = E\{e|e \in B(x)\} = E\{e|e > \tfrac{1}{2}(x^A+x^B)\} = E_e Q^B(x^A,x^B,e) = 0$$

where

$$Q^A(x^A,x^B,e) = \begin{cases} e & \text{if } e \leq \frac{1}{2}(x^A+x^B) \\ x^A & \text{if } e > \frac{1}{2}(x^A+x^B) \end{cases} \quad \text{and } Q^B \text{ analogously}$$

Since the classification e ∈ A or e ∈ B depends on x, the right hand side depends on x. For the simple algorithm (1.12), where the mean value of a stochastic variable is estimated, this is not the case. As in Example 1.5, the adaptive nature of the algorithm makes Q a more complex function of x.

□

To summarize, we have in this chapter seen examples of a variety of algorithms in control theory, that have a structure given by (1.1). In this report the convergence properties of such algorithms are treated. The results will also imply new convergence criteria for stochastic approximation algorithms. The results are summarized in Chapter 7.

2. CLASSIFICATION OF CONVERGENCE CRITERIA.

The convergence properties of some of the algorithms discussed in Chapter 1 have been treated by many authors. However, there does not seem to exist a unified approach.

Convergence of stochastic approximation algorithms is treated in a number of papers. A selection of these follows below. However, most of the given results cannot be applied to cases with correlated observations.

The potential function method is extensively treated in Aizerman, Braverman and Rozonoer (1970).

Convergence of the real time least squares method (Example 1.3) follows from the consistency of least squares estimation, see e.g. Åström and Eykhoff (1971). Convergence for the more complex cases of general recursive identification schemes (Example 1.4) and self-tuning regulators (Example 1.5) does not seem to be treated in the literature.

The objective of this chapter is to illustrate what types of conditions that are usually imposed to assure convergence with probability one (w.p.1). The discussion here is basically confined to the Robbins-Monro scheme. This procedure and variants thereof are extensively treated in the literature. The convergence criteria can be classified into three classes. This conclusion is supported by several examples. The chapter can therefore also be read as a short and incomplete survey of previous results. As such it is not essential for the rest of this report, but it serves as a background to the new results presented here.

2.1. Problem Formulation.

In Example 1.1 we have introduced the RM scheme as a recursive method to solve

$$E_eQ(x,e) = f(x) = 0 \tag{2.1}$$

by

$$x_n = x_{n-1} + \gamma_n Q(x_{n-1},e_n) \tag{2.2}$$

Many papers and books deal with convergence w.p.1 of (2.2) to the desired point x^*, see e.g. Blum (1954ab), Dvoretzky (1956), Burkholder (1956), Derman and Sacks (1959), Gladysjev (1965), Albert and Gardner (1967), Wazan (1969) and Aizerman et al (1970). Various criteria have been suggested.

It is possible to classify the conditions into three main classes.

Eq. (2.2) can be written

$$x_n = x_{n-1} + \gamma_n[f(x_{n-1}) + (Q(x_{n-1},e_n) - f(x_{n-1}))]$$

The term $Q(x_{n-1},e_n) - f(x_{n-1})$ can heuristically be regarded as noise added to the deterministic algorithm

$$x_n = x_{n-1} + \gamma_n f(x_{n-1}) \tag{2.3}$$

Since the noise term is not likely to improve the convergence of (2.2), it is reasonable to require first of all that the algorithm (2.3) should converge to x^*. Convergence of (2.3) which is an ordinary difference equation, depends on the "step size" γ_n, as well as on the function f(x). A necessary property of $f(x)$ is that (2.3) shall converge for sufficiently small γ_n. The corresponding maximal magnitude of γ_n may depend on the initial value x_0. Conditions that assure this property will be referred to as stability conditions.

The variable γ_n tends to zero as n tends to infinity. Now, in the beginning of the recursion (2.3) γ_n can still be too large and $|x_n|$ may increase rapidly. If then also $|f(x_n)|$ increases, the sequence $\{\gamma_n\}$ may not decrease sufficiently fast to yield small corrections $\gamma_n f(x_n)$ and convergence of (2.3). Therefore conditions that assure boundedness of x_n must be introduced. These will be called boundedness conditions. Boundedness and stability conditions together imply convergence of (2.3) to the desired point x^*, if

$$\gamma_n \to 0 \quad \text{as} \quad n \to \infty \quad \text{and} \quad \sum_1^\infty \gamma_n = \infty \tag{2.4}$$

The second condition is necessary, since otherwise x_n could move only a given distance from the initial value.

Furthermore, certain conditions on the noise $Q(x,e_n) - f(x)$, and on the sequence γ_n must be introduced to assure that the influence of the randomness in algorithm (2.2) is sufficiently small. Such conditions will be referred to as noise conditions.

2.2. Stability Conditions.

In the original paper by Robbins and Monro (1951), only mean square convergence of (2.2) was considered. Convergence w.p.1 was first shown by Blum (1954a). He considered the scalar case and introduced the following conditions on f(x):

$$\begin{aligned} f(x) &> 0 \quad \text{for} \quad x < x^* \\ f(x) &< 0 \quad \text{for} \quad x > x^* \end{aligned} \tag{2.5}$$

where x^* is the solution of $f(x) = 0$. Blum also assumed that

$$\inf_{\delta_1 < |x-x^*| < \delta_2} |f(x)| > 0 \quad \text{for all } \delta_2 > \delta_1 > 0$$

In this simple case the stability of the iterative methods to find the solution of $f(x) = 0$ is, of course, determined by the way the curve $f(x)$ intersects the x-axis.

Albert and Gardner (1967) use as criterion that $\frac{d}{dx} f(x)$ be negative for all x.

Dvoretzky's (1956) criterion is valid not only for the RM case, but also for the more general algorithm (1.15). The criterion is based on a contraction mapping property and implies (2.5) for algorithm (2.2) in the case of scalar valued x. For vector valued x and Q Dvoretzky suggests the following criterion:

$$|x_n - x^* - \gamma_n f(x_n)| < F_n |x_n - x^*|$$

where

$$\prod_{n=1}^{\infty} F_n = 0$$

Blum (1954b) has suggested to use functions with Lyapunov properties. The existence of a function V(x) with properties

$$V(x) \geqslant 0$$

$$\inf_{\varepsilon<|x-x^*|} |V(x) - V(x^*)| > 0 \qquad \forall \ \varepsilon > 0$$

$$\sup_{\varepsilon<|x-x^*|} V'(x) \cdot f(x) < 0 \qquad \forall \ \varepsilon > 0$$ [1)]

is assumed. These properties obviously guarantee that the ordinary differential equation

$$\frac{d}{dt} x = f(x)$$

is stable.

Braverman and Rozonoer (1969) and Aizerman et al (1970) have given similar and slightly more general stability criteria.

In e.g. Braverman and Rosenoer (1969) and Krasulina (1972) special attention is paid to the case when the equation

$$f(x) = 0$$

has several roots.

Gladysjev (1965) has introduced a less general criterion,

[1)] The derivative V' is regarded as a row vector.

which is more easily checked than the one above. It has the form

$$\inf_{\varepsilon<|x-x^*|<1/\varepsilon} (x-x^*)^T f(x) > 0 \qquad \forall\ \varepsilon > 0 \tag{2.6}$$

Obviously it is a special case of Blum's approach, with $V(x) = \frac{1}{2}|x - x^*|^2$. The criterion (2.6) is also used by Tsypkin (1971).

When the RM scheme is applied to minimize a function as in (1.13) the function $P(x) = E_v J(x,v)$ can be used as a Lyapunov function for the problem. This approach has been persued by Litvakov (1968) and Devyaterikov et al (1969).

2.3. Boundedness Conditions.

Consider first an example that shows how divergence in the algorithm (2.3) can occur, even if the stability conditions are satisfied.

Example 2.1. Let $f(x) = -x^3$. Then (2.3) gives

$$x_n = x_{n-1} - \gamma_n x_{n-1}^3 \tag{2.7}$$

Clearly f(x) satisfies any of the cited stability conditions. However, with $\gamma_n = 1/n$ and $x_0 = 2$ the sequence of estimates is

$x_1 = -6 \quad x_2 = 102 \quad x_3 = -353634 \quad \ldots$

and $|x_n|$ tends to infinity as n tends to infinity. For this initial value, convergence would be obtained if

$\gamma_n \leqslant 1/2$, $n = 1,\ldots$ In general γ_n must be less than $2/x_0^2$ to assure convergence. There consequently exists no sequence $\{\gamma_n\}$ that yields convergence for (2.7) irrespectively of the initial value.

□

Some restrictions to rule out cases like Example 2.1 must be introduced. For (2.2) there is a non zero probability that x_n may belong to a given area arbitrarily far away from x^*. Therefore such restrictions cannot be obtained by conditions on the sequence $\{\gamma_n\}$ related to the initial value x_0, but conditions on $f(x)$ must be introduced.

Blum (1954a) uses

$$|f(x)| \leqslant a + b|x|$$

for the scalar case. Albert and Gardner (1967) consider truncated algorithms of the type

$$\begin{aligned}
x_n' &= x_{n-1} + \gamma_n Q(x_{n-1}, e_n) \\
x_n &= x_n' && \text{if } B < x_n' < A \\
x_n &= B && \text{if } x_n' \leqslant B \\
x_n &= A && \text{if } x_n' \geqslant A
\end{aligned} \tag{2.8}$$

For the Lyapunov function approach, additional assumptions on the Lyapunov function $V(x)$ must be introduced. Blum (1954b) assumes that

$$V_a(x) = E_e W_a(x,\theta,e) \leqslant K \quad \text{for all } a,x$$

where

$$W_a(x,\theta,e) = Q(x,e)^T V''\left(x+\theta a Q(x,e)\right) Q(x,e)$$

The variable θ is a number between 0 and 1, that may depend on e. $V''(x)$ is the matrix of second order derivatives (the Jacobian) of $V(x)$, the function introduced in Section 2.2.

In Aizerman et al (1970) a more general condition is assumed:

$$E_e\left[\max_{0\leqslant\theta\leqslant 1} W_{\gamma_n}(x,\theta,e)\right] \leqslant C_1V(x) - C_2V'(x)f(x) + C_3 \qquad (2.9)$$

Devyaterikov et al (1969) use a similar condition.

Gladysjev (1965), Litvakov (1968) and Tsypkin (1971) use a criterion that is similar to the one due to Blum in the scalar valued case:

$$E_e[Q(x,e)^TQ(x,e)] \leqslant C_1(1+x^Tx)$$

This is a special case of (2.9), with $V(x) = |x - x^*|^2$.

All these boundedness criteria state, in various forms, that Q(x,e) must not increase faster than $|x|$ as $|x|$ tends to infinity.

2.4. Noise Conditions.

In statistical literature, like in Blum (1954ab), Burkholder (1956), the problem (2.1) is often formulated as follows. Consider a family of stochastic variables Y(x) having distribution $H(\cdot|x)$ and conditional expectations $f(x) = EY(x)$. The function f(x) is called the regression function with respect to the family Y(x). The equation $f(x) = 0$ is then solved recursively

$$x_{n+1} = x_n - \gamma_n y_n$$

where y_n is a random variable whose conditional distribution has the property

$$H(\cdot|x_n, x_{n-1}, \ldots, x_1; y_{n-1}, \ldots, y_1) = H(\cdot|x_n)$$

For the formulation (2.2) this means that the distribution of $Q(x_n, e_{n+1})$ given $Q(x_k, e_{k+1}), k = n-1, \ldots, 0$ and $x_k, k = n, \ldots, 0$ may depend only on x_n. In particular, the distribution of e_{n+1} must not depend on $e_n, \ldots, e_0$. Consequently, the sequence $\{e_i\}$ must consist of independent random variables. When the Lyapunov function approach is used, as in e.g. Blum (1954b), Aizerman et al (1970), Litvakov (1968), this independence assumption is critical to evaluate

$$E_e\left\{V'(x_n)Q(x_n, e_{n+1})|V(x_1), \ldots, V(x_n)\right\}$$

Comer (1964) points out that the assumption on independence is not very realistic in process control applications. He considers the case

$$Q(x_n, e_{n+1}) = f(x_n) + e_{n+1}$$

where the variables $\{e_n\}$ fulfil a condition weaker than independence. However, he does not show convergence with probability one for such a process. Wasan (1969) uses Comer's result for a convergence theorem for dependent noise. This theorem, however, does not seem to be correct.

A more general stochastic approximation algorithm is considered by Albert and Gardner (1967). They allow time varying regression functions $f_n(x)$ and

$$\gamma_n = \gamma_n(x_1,\ldots,x_n)$$

They give a sufficient condition for convergence of such an algorithm also for dependent noise. This result is not applicable in the present case, since it requires that

$$\sum_{n=1}^{\infty} \inf_x f_n'(x) \quad \inf_x \gamma_n = \infty$$

$$\sum_1^{\infty} \sup_x \gamma_n < \infty$$

which obviously cannot be fulfilled for time invariant regression functions.

The conclusion is that the assumption on independent observations has been critical to prove convergence w.p.1 for (2.2). In this report results valid also for dependent observations are presented.

Together with this assumption, it is also usually assumed that

$$E_e[Q(x,e) - f(x)]^T[Q(x,e) - f(x)] \leqslant \sigma^2(x) \quad \text{for all } x \qquad (2.10)$$

and

$$\sum_1^\infty \gamma_n^2 < \infty$$

In some papers a condition (2.10) is included in conditions of type (2.9).

Krasulina (1969) has shown convergence for the Kiefer-Wolfowitz procedure in case

$$E_v \, | J(x,v) - P(x) |^p \leqslant C \qquad 1 < p < 2$$

and

$$\sum_{n=1}^\infty (\gamma_n/a_n)^p < \infty$$

where a_n is the search length as in Example 1.2. It is thus not assumed that the variance of the noise exists. Krasulina (1972) has also shown convergence in case

$$E_e \, | Q(x,e) - f(x) |^2 \geqslant \sigma^2 > 0 \quad \text{for all } x$$

$$E_e \, | Q(x,e) - f(x) |^{4+\delta} \leqslant C < \infty \quad \delta > 0$$

$$\gamma_n = n^{-1/2}$$

where the condition

$$\Sigma \gamma_n^2 < \infty$$

is not satisfied.

In this report convergence with probability one for a set of algorithms that includes (2.2) is considered. Some new conditions of the discussed types are derived. The noise conditions in the convergence theorem of this report are more general than those discussed above. Also, less restrictions will in general be imposed on the sequence $\{\gamma_n\}$.

3. A SEPARATION THEOREM.

Consider the basic algorithm (1.8)

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1}, e_n) \qquad (3.1)$$

where $\{\gamma_n\}$ is a sequence of scalar valued variables that may be random:

$$\gamma_n = \gamma_n(e_n, \ldots, e_0)$$

The RM scheme (2.2) is a special case of (3.1). In the previous chapter different criteria to assure convergence of the RM scheme were discussed. It was found to be convenient to classify the criteria into three classes:
A) Noise Conditions, B) Boundedness Conditions, and C) Stability Conditions.

In this chapter conditions for the convergence w.p.1 of (3.1) are given. They are of general nature and can usually be applied in practice only after further investigations. Such analysis is given in the following chapters. The main idea of the theorem is that, the question of convergence is separated into three conditions that can be studied as problems of their own.

In Section 3.1 a heuristic interpretation of the theorem is given. The separation theorem is stated in Section 3.2 and in Section 3.3 some examples are given.

3.1. A Heuristic Interpretation.

Suppose that we shall solve the equation (1.9)

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

where measurements $Q(x,e_n)$ are available for any chosen x. An intuitive and simple-minded approach to this problem is as follows:

1) Fix an x^i.

2) Obtain a large number of samples $Q(x^i,e_k)$, $k=1,\ldots,n$.

3) Form an estimate of $f(x^i)$ as a weighted sum of these samples:

$$\hat{f}(x^i) = \sum_{k=1}^{n} \beta_k^n Q(x^i,e_k) = z_n(x^i)$$

If β_k^n can be expressed as

$$\beta_k^n = \gamma_k \prod_{i=k+1}^{n} (1-\gamma_i)$$

the sum can also be defined recursively as

$$z_k(x^i) = z_{k-1}(x^i) + \gamma_k[Q(x^i,e_k) - z_{k-1}(x^i)] \;;\; z_0=0 \qquad (3.2)$$

4) Based on this estimate, determine a new x-value

$$x^{i+1} = x^i + \tilde{\gamma}_{i+1} z_i(x^i) \qquad (3.3)$$

5) Take this x^{i+1} as the new x and repeat from 2).

This scheme has two phases: an estimation phase (3.2) and a decision phase (3.3). Now, let the number of samples in each estimation phase tend to infinity. The resulting, hypothetic, algorithm would then converge if a) the estimation phases give consistent estimates: $z_n(x^i) \to f(x^i)$ w.p.1 as $n \to \infty$, and b) the decision phase, which is a deterministic difference equation with $z_n(x^i)$ replaced by $f(x^i)$, converges to the solution of $f(x) = 0$.

Now, the Robbins-Monro scheme (3.1) can in fact be seen as an ingenious mixing of the two phases. A decision is taken in each step, but as n tends to infinity, more effort is paid to the estimation, since γ_n tends to zero.

The separation theorem states that, in spite of the mixing of the phases, convergence of (3.1) still follows from consistency in the estimation phase and convergence in the decision phase. More precisely, the conditions

a) $z_n(x^0) \to f(x^0)$ w.p.1 as $n \to \infty$ for all x^0 where

$$z_k = z_{k-1}(x^0) + \gamma_k\{Q(x^0,e_k) - z_{k-1}(x^0)\}$$

b) $x_n \to x^*$ as $n \to \infty$ where $f(x^*) = 0$ and

$$x_k = x_{k-1} + \gamma_k f(x_{k-1})$$

are the main conditions for convergence of (3.1) to the desired value x^*.

In Theorem 3.1 the condition b) above is split up into a boundedness condition and a stability condition.

3.2. Separation.

Theorem 3.1. Consider the algorithm (3.1)

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1}, e_n) \qquad x \in R^m$$

and let D be a compact subset of R^m. Let $Q_n(x,e)$ be Lipschitz continuous in an open region $D^0 \supset D$ for fixed e, with Lipschitz constant $K_n(e)$. Assume that the sequence of positive scalars (random variables) $\{\gamma_n\}$ satisfies

$$\gamma_n \to 0 \text{ as } n \to \infty \quad \text{and} \quad \sum \gamma_n = \infty \qquad \text{(w.p.1)}$$

Let $z_n(x^0)$ and r_n be recursively defined by

$$z_n(x^0) = z_{n-1}(x^0) + \gamma_n[Q_n(x^0, e_n) - z_{n-1}(x^0)] \qquad z_0 = 0 \qquad (3.4)$$

$$r_n = r_{n-1} + \gamma_n[K_n(e_n) - r_{n-1}] \qquad r_0 = 0 \qquad (3.5)$$

where x^0 is a fixed element in D.

Assume that

a) $z_n(x^0)$ converges w.p.1 for any $x^0 \in D$ and define the function

$$f(x) = \lim_{n\to\infty} z_n(x)$$

r_n converges w.p.1.

b) $x_n \in D$ infinitely often (i.o.) w.p.1. (This means that w.p.1 there exists a subsequence $\{x_{n_k}\}$ that belongs to the region.)

c) The ordinary differential equation

$$\frac{d}{dt} x = f(x) \qquad (3.6)$$ [1]

has a stationary point x^* which is an asymptotically stable solution with domain of attraction $D_1 \supset D$. (That is, all solutions with initial values in D_1 tend to x^* as t tends to infinity.)

Then $x_n \to x^*$ w.p.1 as $n \to \infty$.

□

The proof is given in Appendix A.

Notice that the only condition in Theorem 3.1 for the separation to hold is that $Q_n(x,e)$ is Lipschitz continuous for fixed e. This is quite a weak condition. In particular, the separation is obtained without any conditions on the noise e.

Condition a) in Theorem 3.1 will be called <u>the noise condition</u>. It concerns the convergence w.p.1 of the two algorithms (3.4) and (3.5). These have the same structure as the simple example (1.12) of the RM scheme where the mean value of a stochastic variable is estimated. The convergence of these algorithms is investigated in Chapter 4. There $Q_n(x^0,e_n)$ is considered as a random variable for which $E_eQ_n(x^0,e_n)$ exists. If $z_n(x^0) \to f(x^0)$ w.p.1, it then follows under weak assumptions, that

$$f(x^0) = \lim_{n\to\infty} E_eQ_n(x^0,e)$$

This connects the two different definitions of f(x), the

[1] Existence and uniqueness of solutions to (3.6) follows from the Lipschitz continuity of Q and from condition a).

one in (2.1) and the one in Theorem 3.1.

Condition b) will be called the <u>boundedness condition</u>. In its given form it is clearly necessary for convergence, but may be difficult to apply directly. Conditions that imply the boundedness condition are given in Chapter 5. Notice that if it does not hold, and D can be taken as any compact region, this means that x_n tends to infinity with non zero probability.

Condition c) clearly is the <u>stability condition</u>. It can be checked using Lyapunov stability theory, see e.g. Krasovskij (1963).

The techniques to find Lyapunov functions are not discussed here. In practical situations, sufficient insight into the stability properties of (3.6) might be obtained by numerical solution. The importance of the ODE (3.6) is, however, not restricted to the question of convergence of (3.1). In Chapter 6 it is shown that the trajectories of (3.6) are related to the asymptotic behaviour of (3.1). Therefore, when investigating the properties of (3.1), numerical solution of (3.6) can be a valuable complement to simulation of (3.1).

<u>Remark</u>. Basically, the theorem does not deal with convergence in a stochastic setting. A fixed realization for which a) holds on a dense subset of D and for which b) and the conditions on $\{\gamma_n\}$ hold is considered throughout the proof. Convergence of (3.2) is shown under these conditions. The theorem thus also can be applied for each realization, and the stochastic convergence concept "w.p.1" can be omitted. In particular, this means that the limit function f(x), as well as the convergence point x^* might be random variables: $f(x) = f(x,\omega)$. Then in condition c) the ODE $\dot{x} = f(x,\omega)$ should be asymptotically stable with

stationary point $x^*(\omega)$ for almost every ω, i.e. c) must hold w.p.1.

Several extensions of the theorem are possible. It holds also for the algorithm

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1},\ldots,x_0;e_n)$$

if the dependence of old x_i on Q_n decreases sufficiently fast. In Ljung-Wittenmark (1974) such an extension is proved for a special class of such algorithms.

In some applications, like real time least squares (Example 1.3), Q_n is multiplied by a matrix that depends on old values of the noise:

$$x_n = x_{n-1} + \gamma_n S_n(e_1,\ldots,e_n)Q_n(x_{n-1},e_n) \tag{3.7}$$

This case is considered in the following corollary.

Corollary 1. Consider algorithm (3.7). Suppose the conditions of Theorem 3.1 hold with Q_n replaced by S_nQ_n. (The limit function f(x) may be a random variable.) Then $x_n \to$ $\to x^*$ w.p.1 as $n \to \infty$.

In case there are several stationary points of (3.6) it may be easier to use the following variant of the stability condition.

Corollary 2. Consider algorithm (3.1). Suppose that conditions a) and b) of Theorem 3.1 hold. Assume that there exists a twice differentiable function V(x), $x \in D_1$, where D_1 is an open set that contains D, such that

$V'(x)f(x) \leq 0$, $\forall\, x \in D_1$, and $V'(x)f(x) = 0 \Leftrightarrow x \in D_c$. Assume further that $V(\bar{x})$ = const and $V'(\bar{x})f(\bar{x}) < 0$ for $\bar{x}$ belonging to the boundary of D. (This assumption can be omitted if D_1 can be taken as R^m.) Then $x_n \to D_c$ w.p.1 as $n \to \infty$ [1]. Furthermore, if D_c consists of isolated points, then $x_n \to x^*(\omega)$ w.p.1 as $n \to \infty$, where $x^*(\omega) \in D_c$.

The proof of this corollary is indicated in Appendix A.

3.3. Examples of Application of the Theorem.

Theorem 3.1 is applicable to a variety of algorithms. Some examples of cases that can be treated are given below. In general no convergence results can be obtained at this stage, since the noise and the boundedness conditions remain to be analysed. Hence it is demonstrated only how the separation can be achieved and how the conditions can be reformulated. Some of the examples are continued in Chapter 7.

Example 3.1 - Minimization of a function when derivatives are available.

This common case can be formulated: Minimize $P(x) = E_v J(x,v)$, $x \in R^m$, with respect to x. The derivative $\frac{\partial}{\partial x} J(x,v)$ is available, and the Robbins-Monro scheme can be applied to solve

$$f(x) = -E_v \frac{\partial}{\partial x} J(x,v) = 0$$

[1] By this is meant that $\inf_{\tilde{x} \in D_c} |x_n - \tilde{x}| \to 0$ as $n \to \infty$ w.p.1.

for x. Examples include various estimation problems as shown in Chapter 1. In this case the boundedness and stability conditions can be interpreted in a special way.

Suppose that P(x) is continuously differentiable. Under weak conditions it follows that $f^T(x) = -P'(x)$. It is now possible to choose P(x) as the function V(x) in Corollary 2. Then $V'(x)f(x) = P'(x)f(x) = -f^T(x)f(x) \leqslant 0$ $\forall\, x \in R^m$. Consequently, if x_n belongs to some bounded region i.o. it will tend to a solution of $f(x) = 0$. If it is not bounded i.o. it tends to infinity. To summarize: If the noise condition is satisfied, x_n either tends to infinity or to a stationary point of P(x) w.p.1 as $n \to \infty$.

Example 3.2 - The Kiefer-Wolfowitz procedure.

If the minimum of P(x) in the previous example is to be found and no derivatives are available, the Kiefer-Wolfowitz procedure can be applied, as described in Example 1.2. The sequence $\{a_n\}$ tends to zero as n tends to infinity. Consider (1.14). To apply the theorem, take

$$Q_n(x_{n-1},e_n) = \Big[\big(J(x_{n-1}+a_n u_1,v_{j_1}) - J(x_{n-1},v_i)\big)/a_n,\ldots\Big]^T$$

Suppose that P(x) is twice continuously differentiable with respect to x. Then

$$Q_n(x_{n-1},e_n) = -P'(x_{n-1}) + a_n P''(\xi_n) + q_n(x_{n-1},e_n)/a_n$$

where

$$q_n(x_{n-1},e_n) = \Big[J(x_{n-1}+a_n u_1,v_{j_1}) - E_{v_{j_1}} J(x_{n-1}+a_n u_1,v_{j_1}) -$$

$$- J(x_{n-1},v_i) - E_{v_i} J(x_{n-1},v_i),\ldots\Big]^T$$

The condition on Lipschitz continuity of Q_n gives a condition

$$|q_n(x_1,e) - q_n(x_2,e)| \leqslant K_n(e)|x_1 - x_2| \tag{3.8}$$

The noise condition takes the form

$$z_n(x^0) = z_{n-1}(x^0) + \gamma_n\big(Q_n(x^0,e_n) - z_{n-1}(x^0)\big) \Rightarrow \tag{3.9}$$

$$\Rightarrow z_n(x^0) \to -P'(x^0) \text{ w.p.1 as } n \to \infty.$$

$$r_n = r_{n-1} + \gamma_n\big(K_n(e_n) - r_{n-1}\big) \Rightarrow r_n \to r_\infty \text{ w.p.1 as } n \to \infty \tag{3.10}$$

Now, $Q_n(x^0,e)$ contains a deterministic component $- P'(x^0) +$ $+ a_n P''(\xi_n)$ where $\xi_n \to x^0$ as $n \to \infty$. This implies that $P''(\xi_n)$ is bounded and so $a_n P''(\xi_n) \to 0$ as $n \to \infty$. It is easy to show that (3.9) is equivalent to

$$\tilde{z}_n(x^0) = \tilde{z}_{n-1}(x^0) + \gamma_n\big(q_n(x^0,e_n)/a_n - \tilde{z}_{n-1}(x^0)\big)$$

$$\Rightarrow \tilde{z}_n(x^0) \to 0 \text{ w.p.1 as } n \to \infty \tag{3.11}$$

Consequently, if (3.8), (3.10) and (3.11) are satisfied, then x_n tends w.p.1 either to infinity or to a stationary point of $P(x)$.

Example 3.3 - Minimization of a function using noise corrupted measurements.

A common special case of Example 3.3 is that

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

where the distribution of e_n does not depend on x. If the KW procedure is applied to this case, the noise conditions are simpler than in Example 3.2. Obviously, since q_n does not depend on x,(3.8) and (3.10) are trivially satisfied. Hence x_n tends w.p.1 either to infinity or to a stationary point of P(x) if $z_n \to 0$ w.p.1 as $n \to \infty$, where

$$z_n = z_{n-1} + \gamma_n(e_n/a_n - z_{n-1});\ z_0 = 0$$

Example 3.4 - Real time least squares.

In Example 1.3 it was found that the solution of (1.23) can be written as

$$x_{n+1} = x_n + \frac{1}{n+1} R_n^{-1}\left[\xi_n y(n+1) - \xi_n\xi_n^T x_n\right] \tag{3.12}$$

where

$$R_n = \frac{1}{n+1}\sum_{k=0}^{n} \xi_k\xi_k^T \tag{3.13}$$

Assume that $\{\xi_k\}$ is a stationary stochastic process. Then

$$R_n \to R = E\xi_k\xi_k^T \quad \text{as } n \to \infty \text{ w.p.1}$$

It is assumed that R is nonsingular. In this case

$$Q_n(x,e) = R_n^{-1}\xi_n y(n+1) - \xi_n\xi_n^T x_n$$

This function clearly is Lipschitz continuous, since

$$|Q_n(x_1,e_n) - Q_n(x_2,e_n)| \leq |R_n^{-1}\xi_n\xi_n^T||x_1 - x_2| \leq$$

$$\leq \Big[|R^{-1}\xi_n\xi_n^T| + |R_k^{-1} - R^{-1}||\xi_n\xi_n^T|\Big]|x_1 - x_2| =$$

$$= K_n(e_n)|x_1 - x_2|$$

Therefore Corollary 1 of the separation theorem can be applied to (3.12). The noise condition is that

$$z_n(x^0) = z_{n-1}(x^0) + \frac{1}{n}\Big[R_n^{-1}\Big(\xi_n\xi_n^T x^0 + \xi_n y(n+1)\Big) - z_{n-1}(x^0)\Big] \quad (3.14)$$

converges w.p.1 to some limit $f(x^0)$.

The solution of (3.17) is

$$z_n(x^0) = \frac{1}{n}\sum_{k=1}^{n} R_k^{-1}\Big[\xi_k\xi_k^T x^* - \xi_k\xi_k^T x^0 + \xi_k e(k+1)\Big] =$$

$$= R^{-1}\Big[\Big\{\frac{1}{n}\sum_1^n \xi_k\xi_k^T\Big\}\Big\{x^* - x^0\Big\} + \frac{1}{n}\sum_1^n \xi_k e(k+1)\Big] +$$

$$+ \frac{1}{n}\sum_1^n (R_k^{-1} - R^{-1})\Big[\xi_k\xi_k^T(x^* - x^0) + \xi_k e(k+1)\Big]$$

It is not difficult to show that

$$z_n(x^0) \to R^{-1}E\xi_k\xi_k^T(x^* - x^0) = x^* - x^0 = f(x^0) \quad \text{w.p.1 as } n \to \infty$$

The second noise condition is that r_n converges w.p.1 where r_n is defined by (3.4). $K_n(e_n)$ is equal to $|R^{-1}\xi_n\xi_n^T| + |R_n^{-1} - R^{-1}||\xi_n\xi_n^T|$ and

$$r_n = \frac{1}{n}\sum_1^n |R^{-1}\xi_k\xi_k^T| + \frac{1}{n}\sum_1^n |R_k^{-1} - R^{-1}||\xi_k\xi_k^T|$$

which converges w.p.1 since $|R^{-1}\xi_k\xi_k^T|$ is a stationary process and since $|R^{-1} - R_k^{-1}||\xi_k\xi_k^T| \to 0$ w.p.1 as $n \to \infty$.

The ODE in the stability condition becomes

$$\frac{d}{dt} x = x^* - x$$

which clearly is globally asymptotically stable.

If the boundedness condition does not hold, x_n would tend to infinity with non zero probability. This is easily contradicted. Hence all conditions in Theorem 3.1 are satisfied and

$$x_n \to x^* \text{ w.p.1 as } n \to \infty$$

follows.

4. THE NOISE CONDITION.

The noise condition in Theorem 3.1 is that the two algorithms

$$z_n(x^0) = z_{n-1}(x^0) + \gamma_n[Q_n(x^0,e_n) - z_{n-1}(x^0)] \qquad z_0 = 0 \quad (4.1)$$

$$r_n = r_{n-1} + \gamma_n[K_n(e_n) - r_{n-1}] \qquad r_0 = 0 \quad (4.2)$$

converge w.p.1 (for fixed x^0).

In Theorem 3.1 no assumption about the statistics of the variables $Q_n(x^0,e_n)$ was made. In this chapter $Q_n(x^0,e_n)$ is considered as a vector valued random variable for which $E_{e_n}Q_n(x^0,e_n)$ exists for every n. Furthermore, the sequence $\{\gamma_n\}$ is here supposed to consist of deterministic scalars as in (1.5). If the original sequence $\{\gamma_n\}$ in (3.1) is stochastic, γ_n and Q_n in (3.1) can be redefined as

$$\gamma_n Q_n(x,e_n) = \bar{\gamma}_n\left(1 + \frac{\tilde{\gamma}_n}{\bar{\gamma}_n}\right)Q_n(x,e_n) = \bar{\gamma}_n Q_n^*(x,e_n^*)$$

where $\bar{\gamma}_n = E\gamma_n$ and $\tilde{\gamma}_n = \gamma_n - \bar{\gamma}_n$ which gives a deterministic sequence $\{\bar{\gamma}_n\}$.

For a specific application, it is possible to study the convergence of (4.1) using the special structure of the problem as in Example 3.4. In this chapter general conditions to assure convergence of (4.1) w.p.1 are discussed. The objective is to give results that cover practical situations, rather than to elaborate on the sharpness of the theorems.

In Section 4.1 it is shown that the convergence problem for (4.1) can equivalently be formulated for a simpler algorithm. In Section 4.2 the common case when the sequence $\{\gamma_n\}$ asymptotically decreases as $1/n$ is considered.

Expressions for the absolute moments of z_n are derived in Section 4.3 and these are used in Section 4.4 to obtain more general convergence results.

4.1. An Equivalent Problem Formulation.

In this section it is shown that it is sufficient to analyze convergence of the algorithm

$$y_n = y_{n-1} + \gamma_n[f_n - y_{n-1}] \qquad y_0 = 0 \tag{4.3}$$

where $\{y_n\}$ are scalars, $\{\gamma_n\}$ a sequence of deterministic positive scalars and $\{f_n\}$ a sequence of scalar valued random variables with $Ef_n = 0$ all n.

Algorithm (4.1) is more complex than (4.3). It involves vector valued random variables with time varying mean values.

Lemma 4.1. Let $z_n(x^0)$ and y_n be defined by (4.1) and (4.3) respectively. Suppose that

$$E_eQ_n(x^0,e) \to f(x^0) \text{ as } n \to \infty \tag{4.4}$$

and that

$$0 \leqslant \gamma_n \leqslant 1 \qquad \sum_1^\infty \gamma_n = \infty$$

Then

$$z_n(x^0) \to f(x^0) \quad \text{w.p.1 as } n \to \infty$$

if and only if

$$y_n \to 0 \quad \text{w.p.1 as } n \to \infty \qquad \text{for } f_n = Q_n^{(i)}(x^0,e_n) - E_e Q_n^{(i)}(x^0,e_n)$$

$i = 1,\ldots,m$, where $Q^{(i)}$ denotes the i:th row of the column vector Q.

Proof. Since (4.1) is linear in z_n

$$z_n^{(i)}(x_0) = y_n + v_n$$

where v_n is defined by

$$v_n = v_{n-1} + \gamma_n[d_n - v_{n-1}] \qquad v_0 = 0$$

where

$$d_n = E_e Q_n^{(i)}(x^0,e_n)$$

It clearly is sufficient to show that (4.4) implies that $v_n \to f^{(i)}(x^0)$ as n tends to infinity. This is done as follows.

Eq. (4.4) means that $|d_n - f^{(i)}(x^0)| < \varepsilon$ for $n > N_0(\varepsilon)$. Then

$$v_{N_0+m} = \prod_{j=N_0+1}^{N_0+m} (1-\gamma_j)v_{N_0} + \sum_{j=N_0+1}^{N_0+m} \beta_j^{N_0+m} d_j$$

where

$$\beta_j^N = \gamma_j \prod_{i=j+1}^{N} (1-\gamma_i) \quad \text{if } j < N \quad \text{and} \quad \beta_N^N = \gamma_N \tag{4.5}$$

Now $v_{N_0} = 1$ and $d_j = 1$ gives $v_j = 1$; $j \geqslant N_0$, which means that

$$\prod_{j=N_0+1}^{N_0+m} (1-\gamma_j) + \sum_{j=N_0+1}^{N_0+m} \beta_j^{N_0+m} = 1$$

and consequently

$$\left|v_{N_0+m} - f^{(i)}(x^0)\right| \leqslant \left|\prod_{j=N_0+1}^{N_0+m} (1-\gamma_j)\right| \left|d_{N_0} - f^{(i)}(x^0)\right| +$$

$$+ \left|\sum_{j=N_0+1}^{N_0+m} \beta_j^{N_0+m}\right| \varepsilon$$

Now

$$\prod_{j=N_0+1}^{N_0+m} (1-\gamma_j) \to 0 \quad \text{as } m \to \infty \quad \text{since} \quad \sum_1^\infty \gamma_j = \infty$$

Hence

$$|v_n - f^{(i)}(x^0)| < 2\varepsilon$$

for sufficiently large n and so

$$\lim_{n\to\infty} v_n = f^{(i)}(x^0)$$

□

Remark. Assumption (4.4) can be replaced by the weaker condition $v_n \to f^i(x^0)$ as $n \to \infty$, where v_n is defined as in the proof of the lemma.

As a consequence of the lemma it is sufficient to study convergence of the simple algorithm (4.3).

4.2. The Case $\gamma_n \sim A/n$ For Large n.

It has already been remarked that Eqs. (4.1) and (4.3) correspond to estimation of mean values. As shown in Example 1.1 a suitable choice of $\{\gamma_n\}$ then is $\gamma_n = 1/n$. Also in the case (3.1) $\gamma_n = B/n$ for large n (but not for small n, cf. Chapter 6) seems to be a good choice. The convergence properties for the case $\gamma_n = b_n/n$ where $b_n \to B$ as $n \to \infty$, therefore deserve special interest.

In algorithm (4.3) $\gamma_n Q_n(x_n, e_{n+1}) = b_n/n \; Q_n(x_n, e_{n+1})$ can be redefined as

$$1/n \; Q_n^*(x_n, e_{n+1}) \text{ where } Q_n^* = b_n Q_n$$

Apply the separation theorem and (4.1) becomes

$$z_{n+1}(x^0) = z_n(x^0) + \frac{1}{n}\left(Q_n^*(x^0, e_{n+1}) - z_n(x^0)\right)$$

Consequently the convergence analysis for (4.3) with $\gamma_n =$ $= 1/n$ covers all sequences that asymptotically behave like B/n.

With $\gamma_n = 1/n$ in (4.3)

$$y_n = \frac{1}{n} \sum_{k=1}^{n} f_k \tag{4.6}$$

and ergodic theory can be applied to obtain convergence of y_n.

According to Cramer-Leadbetter (1967) (4.6) converges to zero w.p.1 if

$$Ef_k f_s \leqslant \frac{k^p + s^p}{1 + |k - s|^q} \qquad 0 \leqslant 2p < q < 1 \tag{4.7}$$

The condition (4.7) imposes a restriction on the dependence of the sequence $\{f_k\}$, that is quite weak.

4.3. Asymptotic Moments of y_n.

We will now consider a general sequence $\{\gamma_n\}$ and general distributions of f_n. As a convenient regularity condition on f_n will be chosen that the absolute moments up to a certain order p exist. The corresponding moments of y_n then also exist. In this section upper bounds for these moments are calculated.

To facilitate the calculation, certain conditions on the sequence $\{\gamma_n\}$ are introduced. They are chosen as

a) $0 \leqslant \gamma_n \leqslant 1$

b) $\sum_{n=1}^{\infty} \gamma_n = \infty$

(4.8)

c) $\gamma_{n+1} \geqslant \gamma_n(1-\gamma_{n+1})$

d) $\{\gamma_n\}$ is decreasing as a function of n

Notice that it is always possible to redefine γ_n and Q_n so that the sequence γ_n can be scaled arbitrarily. Condition (4.8a) therefore is not restricting. Condition (4.8c) states that in the sum y_n each observation f_k has a weight no less than the previous one. As will be discussed in Chapter 6, this is the interesting case. The conditions (4.8) are satisfied for the common choice

$$\gamma_n = A/n^{\alpha} \quad 0 < \alpha < 1 \quad \text{all } A; \quad \alpha = 1 \quad A \geq 1 \tag{4.9}$$

Some conditions on the dependence between the variables f_k also must be imposed. Conditions involving only second moments as (4.7) are not sufficient in the general case. Since most stochastic processes occuring in control theory have been generated as white noise through some linear (time varying) filter, we adopt the following condition:

Let f_n be obtained from white noise as

$$f_n = \sum_{k=0}^{\infty} h_{k,n} e_{n-k} \quad \text{where} \quad |h_{k,n}| < \alpha_n \lambda^k \quad \lambda < 1 \tag{4.10}$$

and $(e_k,\ k = 0, \pm 1, \ldots)$ is a sequence of independent random variables with zero mean values.

Remark. If $\{f_n\}$ is a stationary, regular stochastic process, it can always be represented as filtered white noise as in (4.10), Doob (1953). The conditions on $h_{k,n}$, however, do not follow automatically from stationarity only.

It is now possible to prove the following lemma.

Lemma 4.2. Consider the algorithm (4.3)

$$y_n = y_{n-1} + \gamma_n(f_n - y_{n-1}) \qquad y_0 = 0$$

Assume that the sequence $\{\gamma_n\}$ satisfies (4.8). Assume further that f_n satisfies (4.10) and that $\{\alpha_n\}$ is a non decreasing sequence of numbers and

$$E|e_k|^p < C$$

which implies

$$E|f_n|^p < C' \cdot \alpha_n^p$$

where p is an even integer. Then

$$E|y_n|^r \leqslant K_r(\alpha_n)^r(\gamma_n)^{r/2} \qquad 1 < r \leqslant p \qquad (4.11)$$

The proof is given in Appendix B.

□

The lemma extends the results given by Chung (1954). There, (4.11) is obtained in the special case

$$\alpha_n = 1 \quad \gamma_n = n^{-\alpha} \quad 1/2 < \alpha \leqslant 1 \quad \text{and } f_n \text{ indep. variables.}$$

However, Chung considers a more general regression function.

The estimates on the moments can be used to obtain convergence criteria. This is treated in the next section.

4.4. Convergence With Probability One.

Using the estimates of the absolute moments of y_n it is easy to establish convergence of y_n to zero w.p.1:

Theorem 4.1. Consider the algorithm (4. 3) with the same assumptions as in Lemma 4.2. Suppose

$$\sum_{n=1}^{\infty} \gamma_n^{p/2}\alpha_n^p < \infty \quad \text{where p is defined in Lemma 4.2.}$$

Then $y_n \to 0$ as $n \to \infty$ w.p.1.

Proof. From Chebysjev's inequality and Lemma 4.2

$$P(|y_n|>\varepsilon) \leqslant \frac{E|y_n|^p}{\varepsilon^p} \leqslant \frac{k_p\gamma_n^{p/2}\alpha_n^p}{\varepsilon^p}$$

and

$$\sum_{n=1}^{\infty} P(|y_n|>\varepsilon) \leqslant \frac{k_p}{\varepsilon^p}\sum_{n=1}^{\infty} \gamma_n^{p/2}\alpha_n^p < \infty \quad .$$

The Borel Cantelli lemma now assures

$y_n \to 0$ as $n \to \infty$ w.p.1. □

Theorem 4.1 shows that it is possible to trade off conditions on the sequence (γ_n, n = 1,...) against conditions on the moments of f_n.

Thus the usually given criterion

$$\sum_{n=1}^{\infty} \gamma_n^2 < \infty \quad \text{for the Robbins-Monro case} \tag{4.12}$$

is in fact not necessary to achieve convergence w.p.1. It can be violated if more regularity of f_n is required. For example, if all moments of f_n exist and are uniformly bounded, it is sufficient that

$$\sum_{n=1}^{\infty} \gamma_n^p < \infty \quad \text{for some sufficiently large } p$$

which is satisfied e.g. for $\gamma_n = n^{-\alpha}$, $0 < \alpha \leqslant 1$.

However, (4.12) can be violated only if higher moments of f_n exist. This is seen from the following example.

Example 4.1. Let f_n, $n = 1,\dots,$ be a sequence of independent random variables where f_n has the distribution

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

Then $P(|\gamma_n f_n| \geqslant 1) = (\gamma_n)^r$.

The moments $E|f_n|^s$ are uniformly bounded only for $s \leqslant r$.

Assume that

$$\sum_{n=1}^{\infty} \gamma_n^r = \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \gamma_n^{r+\varepsilon} < \infty \qquad \text{for some } \varepsilon > 0$$

Then

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

and since the variables (f_n) are independent

$$|\gamma_n f_n| \geq 1 \qquad \text{i.o. w.p.1}$$

from the Borel-Cantelli lemma. With algorithm (4.3)

$$y_n = (1-\gamma_n)y_{n-1} + \gamma_n f_n$$

y_n will consequently w.p.1 not converge to any limit. To be able to apply the theorem $E|f_n|^{2(r+\varepsilon)}$ would have to be uniformly bounded. Thus the moment conditions on f_n cannot be dispensed with.

□

A common case is when the variables f_n are normally distributed. Then also y_n has normal distribution. The probability that $|\gamma_n|>\varepsilon$ can now be determined directly which gives the following theorem:

Theorem 4.2. Consider algorithm (4.3)

$$y_n = y_{n-1} + \gamma_n(f_n-y_{n-1})$$

Let the variables f_n have normal distribution with zero mean value. Assume that

$$|Ef_n f_{n+r}| < \lambda^r \alpha_n \alpha_{n+r} \quad \text{all } n \text{ and some } \lambda < 1,$$

where $\{\alpha_n\}$ is a non decreasing sequence.

Let the sequence $\{\gamma_n\}$ satisfy (4.8) and suppose that

$$\sum_{n=1}^{\infty} e^{-\varepsilon/\gamma_n\alpha_n^2} < \infty \quad \text{for all } \varepsilon > 0$$

Then $y_n \to 0$ as $n \to \infty$ w.p.1.

Proof. As in Lemma 4.2 it follows that

$$Ey_n^2 < K_2\gamma_n\alpha_n^2$$

Now

$$P(|y_n|>\varepsilon) = \frac{1}{\sqrt{2\pi Ey_n^2}} \int_{|x|>\varepsilon} e^{-x^2/2Ey_n^2}dx \leqslant \frac{1}{\varepsilon} Ce^{-\varepsilon^2/K_2\gamma_n\alpha_n^2}$$

Application of the Borel-Cantelli lemma completes the proof.

□

So far in this section the cases with dependent and independent random variables have been dealt with simultaneously. If we confine ourselves to the case of independent variables a refinement of Theorem 4.1 can be obtained:

Theorem 4.3. Consider the algorithm (4.3)

$$y_n = y_{n-1} + \gamma_n(f_n-y_{n-1})$$

where $\{f_n\}$ is a sequence of independent random variables. Suppose $\{\gamma_n\}$ satisfies (4.8). Let $E|f_n|^p \leqslant \alpha_n^p$ for some real $p > 1$, where $\{\alpha_n\}$ is a non decreasing sequence, and suppose that

$$\sum_{n=1}^{\infty} \gamma_n^{p'} \alpha_n^p < \infty \quad \text{where } p' = \min(p, 1+p/2)$$

Then

$y_n \to 0$ as $n \to \infty$ w.p.1.

The proof is given in Appendix C.

□

Remark. The condition (4.8cd) is not used in case $1 < p \leqslant 2$.

Krasulina's (1969) result corresponds to the case $1 < p < 2$. The usual condition

$$Ef_n^2 < C \quad \text{and} \quad \Sigma\gamma_n^2 < \infty$$

is obtained as a special case of the theorem ($p=2$).

Summing up, Theorems 4.1 - 4.3 give weaker and more general conditions for convergence of (4.1) than usually reported. Dependent random variables can also be treated. These results can now be applied to the general algorithm (3.1) via Theorem 3.1. In Chapter 7 some applications of this kind are given.

5. THE BOUNDEDNESS CONDITION.

In the separation theorem in Chapter 3 it is assumed known that the estimates$\{x_n\}$,obtained from

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1}, e_n) \qquad (5.1)$$

with probability one, infinitely often are inside a certain bounded region D. Criteria which assure such behaviour will be discussed in this chapter. Here $\{\gamma_n\}$ is assumed to be a sequence of deterministic scalars.

In Section 2.3 some suggested criteria were reviewed. Condition (2.9) restricts the choice of Lyapunov functions for the problem. A similar condition is discussed in Section 5.1. It is shown that under weak conditions on the noise and on the sequence $\{\gamma_n\}$ convergence in probability to the desired value x^* can be established. Now, convergence in probability implies convergence w.p.1 along a subsequence. Consequently, x_n will w.p.1 belong to any open region containing x^* i.o. This gives the desired boundedness property for any region D.

From a practical point of view the question of boundedness of the estimates may seem uninteresting. In many cases the desired convergence point x^* of (5.1) is a priori known to belong to some bounded region. It is therefore natural to construct the estimates x_n such that they belong to this area. A straightforward way is to project the right hand side of (5.1) into the area in question. Such algorithms are discussed in Section 5.2.

5.1. Lyapunov Function Approach.

In this section Q is supposed to be time invariant. With slight modifications the results hold also for general Q_n. Denote

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

A Lyapunov function for the ODE

$$\frac{d}{dt} x = f(x); \quad f(x) = 0 \Rightarrow x = x^* \tag{5.2}$$

will now be introduced:

Let V(x) be a twice continuously differentiable function satisfying

$$\text{a)} \quad V(x) \geq 0; \quad V(x) = 0 \Leftrightarrow x = x^*$$

$$\text{b)} \quad V'(x)f(x) = W(x) \leq - CV(x), \; C > 0 \tag{5.3}$$

$$\text{c)} \quad E_e[Q(x,e)^T V''(\xi_n(e))Q(x,e)] \leq - AW(x) + B; \quad x \in R^m, \; A,B > 0$$

$$\text{where } \xi_n(e) = x + \theta\gamma_n Q_n(x,e) \quad 0 \leq \theta \leq 1 \quad n \geq N_0$$

Condition (5.3c) restricts the choice of functions V(x). All Lyapunov functions to (5.2) do not satisfy (5.3c). Also, as will be shown below, there exists functions f(x) for which (5.3abc) cannot be satisfied for any V(x), but for which (5.3ab) can be satisfied.

Notice that in (5.3c) $\xi_n(e)$ depends on e. This condition thus also requires certain regularity of the noise.

In (5.3) like everywhere before, the expectation of a function $h(x,e)$

$$E_e h(x,e)$$

is taken with respect to e, while x is regarded as a fixed parameter. Let now x_n be generated by the algorithm (5.1). Then x_n depends on the noise terms $e_1,\ldots,e_n$. If $\{e_i\}$ is a sequence of independent variable e_n is independent of x_{n-1}. Hence

$$E_{e_n} h(x_{n-1},e_n) = E[h(x_{n-1},e_n)|x_{n-1}]$$

where the RHS denotes conditional expectation given x_{n-1} and

$$Eh(x_{n-1},e_n) = E_{x_{n-1}} E_{e_n} h(x_{n-1},e_n) \tag{5.4}$$

If $\{e_i\}$ are dependent, also x_{n-1} and e_n are dependent and (5.4) is no longer true. However, x_{n-1} contains information about recent e_k only to a decreasing extent. Therefore, under quite mild conditions on the noise $\{e_i\}$, x_{n-1} and e_n become less dependent as n increases. Then the following relation holds:

$$Eh(x_{n-1},e_n) - E_{x_{n-1}} E_{e_n} h(x_{n-1},e_n) \to 0 \quad \text{as} \quad n \to \infty \tag{5.5}$$

for any function h for which Eh exists.

Based on the function V in (5.3) a theorem that guarantees convergence in probability can be shown.

Theorem 5.1. Consider the algorithm

$$x_n = x_{n-1} + \gamma_n Q(x_{n-1},e_n) \tag{5.6}$$

where

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

Let the function $V(x)$ satisfy (5.3). Assume that the stochastic process $\{e_i\}$ is such that condition (5.5) is satisfied and that $EV(x_n)$ is finite for all n (but not necessarily uniformly bounded).

Then $x_n \to x^*$ in probability as $n \to \infty$ and consequently $x_n \in D$ i.o. w.p.1, where D is any open region containing x^*.

Proof. By expansion into Taylor series:

$$V(x_n) = V[x_{n-1}+\gamma_n Q(x_{n-1},e_n)] = V(x_{n-1}) + \gamma_n h(x_{n-1},e_n)$$

where

$$h(x_{n-1},e_n) = V'(x_{n-1})Q(x_{n-1},e_n) + \gamma_n Q(x_{n-1},e_n)^T V''(\xi_n) \cdot$$
$$\cdot\ Q(x_{n-1},e_n)$$

$$E_{e_n} h(x_{n-1},e_n) = V'(x_{n-1})f(x_{n-1}) + \gamma_n E_{e_n} Q(x_{n-1},e_n)^T V''(\xi_n) \cdot$$
$$\cdot\ Q(x_{n-1},e_n) \leq W(x_{n-1}) - \gamma_n AW(x_{n-1}) + B\gamma_n \leq$$
$$\leq -(1-\gamma_n A)CV(x_{n-1}) + B\gamma_n$$

Now

$$EV(x_n) = EV(x_{n-1}) + \gamma_n Eh(x_{n-1},e_n) =$$

$$= E[V(x_{n-1}) + \gamma_n E_{e_n} h(x_{n-1},e_n)] + \gamma_n g_n \leq$$

$$\leq E[(1-\gamma_n C+CA\gamma_n^2)V(x_{n-1})+B\gamma_n^2] + \gamma_n g_n$$

where $g_n \to 0$ as $n \to \infty$ according to (5.5)

For sufficiently large n, $\gamma_n < 1/2A$ and then we have

$$EV(x_n) \leq EV(x_{n-1}) + \gamma_n \frac{C}{2}\left[(g_n+B\gamma_n)\frac{2}{C} - EV(x_{n-1})\right]$$

It now follows from Lemma 4.1 that since $g_n+B\gamma_n \to 0$, we have $EV(x_n) \to 0$ as $n \to \infty$. This implies, according to (5.3a) that $x_n \to x^*$ in probability as $n \to \infty$. □

Remark. Notice that the function V(x) assures the boundedness condition as well as the stability condition in Theorem 3.1. If also the noise condition is satisfied, the conclusion of the theorem can be strenghtened to yield convergence w.p.1.

Example 5.1. Consider the simple case

$$Q(x,e) = e - x, \quad Ee = 0; \quad Ee^2 = 1, \quad \{e_n\} \text{ indep. variables}$$

which gives

$$x_n = x_{n-1} + \gamma_n(e_n-x_{n-1}); \quad \gamma_n \to 0 \quad \sum_{n=1}^{\infty} \gamma_n = \infty$$

Choose as function $V(x) = x^2$. Then (5.3a) is trivially satisfied. Since

$$W(x) = V'(x)E_e(e-x) = -2x^2$$

also condition (5.3b) is satisfied with $C = 2$. Now

$$E_e[(e-x)2(e-x)] = 2 + 2x^2$$

and so condition (5.3c) holds with $B = 2$ and $A = 1$. Theorem 5.1 now states that $x_n \rightarrow 0$ in probability.

If, on the other hand,

$$Q(x,e) = e - x^3$$

we can still try $V(x) = x^2$. Again (5.3ab) are satisfied, but

$$E_e[(e-x^3)2(e-x^3)] = 2 + 2x^6$$

and (5.3c) cannot be satisfied for any B and A. Indeed, x_n may very well tend to infinity as shown in Example 2.1.

□

5.2. Projection Algorithms.

In most applications algorithm (5.1) will in fact be

$$x_n = [x_{n-1} + \gamma_n Q_n(x_{n-1}, e_n)]_D \qquad (5.7)$$

where

$$[f]_D = \begin{cases} f \text{ if } f \in D \\ \text{some interior or boundary point of } D \text{ if } f \notin D \end{cases}$$

where D is some closed bounded region.

The sequence $\{x_n\}$ will thus by definition belong to a bounded area. This means that condition b) of Theorem 3.1 is automatically satisfied. However, the theorem cannot be straightforwardly applied, since the behaviour of $\{x_n\}$ close to the boundary of D is not governed by (5.1). To use the separation theorem, it must be shown that

$$x_n \in D^0 \quad \text{i.o. w.p.1}$$

where D^0 is a subset of D such that $\partial D^0 \cap \partial D = \phi$ (∂D = boundary of D). Within the set D^0, the projection algorithm (5.7) coincides with (5.1) for large n.

It is assumed that D is described by

$$D = \{x | U(x) \leqslant A\}$$

where U is a twice continuously differentiable non negative function. The region D cannot be chosen arbitrarily. Loosely, the algorithm (5.1) shall not have a tendency to move out of the region. It is thus assumed that the trajectories of (5.2) do not intersect ∂D "outwards", i.e.

$$\sup_{x \in \partial D} U'(x)f(x) < 0 \tag{5.8}$$

where

$$f(x) = \lim_{n \to \infty} E_e Q_n(x,e)$$

(Assume that the convergence is uniform in $x \in D$).

Assume further that

$$E_e\left[Q_n(x,e)^T U''\big(\xi_n(e)\big)Q_n(x,e)\right] \leq B \qquad x \in D \tag{5.9}$$

where

$$\xi_n(e) = x + \theta\gamma_n Q_n(x,e) \quad \text{some } \theta \quad 0 \leq \theta \leq 1$$

This condition much resembles (5.3c). However, (5.3c) is basically a, rather restrictive, condition on $Q(x,e)$ as a function of x as shown in Example 5.1. Condition (5.9) is a quite weak condition on the noise e.

Theorem 5.2. Consider algorithm (5.7) where D is defined as above. Assume that (5.8) and (5.9) hold, and that $f(x)$ is continuous in a neighbourhood of ∂D. Let $\{\gamma_n\}$ and $\{e_n\}$ satisfy the conditions of Theorem 5.1. Then $x_n \in D^0$ i.o. w.p.1 where D^0 is a subset of D, such that $\partial D^0 \cap \partial D = \phi$.

Proof. Since $W(x) = U'(x)f(x)$ is continuous, (5.8) implies that

$$\sup_{x \in \Delta D_\varepsilon} W(x) = \hat{\delta} < 0$$

where ΔD_ε is some neighbourhood of ∂D. Define D^0 as the difference between D and ΔD_ε:

$$D^0 = D \cap [D \cap \Delta D_\varepsilon]^* \qquad {}^* = \text{complement}$$

Suppose that

$$x_n(\omega) \notin D^0 \quad \text{all } n > N(\omega) \text{ for } \omega \in \Omega' \text{ where } P(\Omega') = \delta > 0.$$

Define the random variable

$$N(\omega) \begin{cases} \text{as above for } \omega \in \Omega' \\ 0 \text{ for } \omega \notin \Omega' \end{cases}$$

and consider the following modification of algorithm (5.7), yielding the estimates $\{\bar{x}_n\}$.

Let $\tilde{x}_n$ be defined as

$$\tilde{x}_n = \bar{x}_{n-1} + \gamma_n Q_n(\bar{x}_{n-1}, e_n)$$

and let

$$\bar{x}_n = \bar{x}_{n-1} \qquad \text{if } n \geq N(\omega) \text{ and } \bar{x}_{n-1} \in D^0$$

$$\bar{x}_n = [\tilde{x}_n]_D \qquad \text{otherwise}$$

Then $\bar{x}_n(\omega) = x_n(\omega)$ for $\omega \in \Omega'$ and

$$P\left(x_n(\omega) \in \Delta D_\varepsilon\right) > \delta/2 \quad \text{for } n > N_0$$

Introduce

$$\tilde{U}(x) = \begin{cases} U(x) & x \in \Delta D_\varepsilon \\ 0 & x \in D^0 \end{cases}$$

Then

$$E\tilde{U}(\bar{x}_{n+1}) - E\tilde{U}(\bar{x}_n) \leqslant E\tilde{U}(\tilde{x}_{n+1}) - E\tilde{U}(\bar{x}_n) \leqslant$$

$$\leqslant \gamma_n \Big\{ - \hat{\delta}\ \delta/2\ P(N<n) + AP(N\geqslant n) + EU'(x_n) \cdot$$

$$\cdot [E_e Q_n(x_n,e) - f(x_n)] + \gamma_n B \Big\}$$

But $P(N\geqslant n) \to 0$ and $E_e Q_n(x_n,e) - f(x_n) \to 0$ as $n \to \infty$. Since $\Sigma\gamma_n = \infty$ this implies that $E\tilde{U}(\bar{x}_n) \to -\infty$ which is impossible.

Hence $P(\Omega') = 0$, i.e. $x_n \in D^0$ i.o. w.p.1. □

Remark. By modifying the proof, condition (5.5) can be replaced by the (essentially stronger) condition a) of Theorem 3.1.

Example 5.2. Consider again the case $Q(x,e) = e - x^3$. Choose $D = [-\sqrt{A},\sqrt{A}]$ and $U(x) = x^2$. It is easy to see that $U(x)$ satisfies (5.6) and (5.7). Theorem 5.2 now guarantees that x_n is strictly interior to D i.o. if algorithm (5.6) is used. From Theorem 3.1 then follows that $x_n \to 0$ w.p.1 as $n \to \infty$ under weak conditions on the noise e_n. The projection into a bounded area is thus not only a formal trick to achieve theorems on convergence. It also makes divergent schemes converge. Intuitively, the estimate "rests" at the boundary until γ_n is so small that the adjustments $\gamma_n Q(x_{n-1},e_n)$ force x_n into the interior of the area D.

6. CONVERGENCE RATE.

In this chapter the importance of the ordinary differential equation (3.6) associated with the algorithm (1.8) is discussed. It is shown that the ODE is intimately connected with the selection of the sequence $\{\gamma_n\}$ to obtain fast convergence of $\{x_n\}$. The usefulness of slow convergence of $\{\gamma_n\}$ is illustrated and explained. The importance of small γ_n initially is also discussed.

In Section 6.1 some numerical examples are given, where slow convergence of $\{\gamma_n\}$ is favourable. A heuristic analysis of the connection between convergence rate and choice of $\{\gamma_n\}$ is given in Section 6.2. In Section 6.3 it is shown that the sequence $\{\gamma_n\}$ in many cases must be bounded from above to obtain acceptable stability properties. A theorem that connects the trajectories of (3.6) with the sequence $\{x_n\}$ defined by (1.8) is proved in Section 6.4. There also the implications of this result on the choice of $\{\gamma_n\}$ are illustrated.

6.1. Choice of $\{\gamma_n\}$.

Consider as in the Robbins-Monro case the problem to solve

$$E_e Q(x,e) = f(x) = 0 \qquad (6.1)$$

for x. As remarked in Chapter 1, a suitable estimate of x at time n is obtained as the solution of

$$\frac{1}{n}\sum_{k=1}^{n} Q(x,e_k) = 0 \qquad (6.2)$$

In the simple case (1.12) when $Q(x,e) = e - x$ the equation (6.2) is linear in x and the solution can be obtained recursively as

$$x_n = x_{n-1} + \frac{1}{n}(e_n - x_{n-1}) \qquad (6.3)$$

which is the RM scheme with $\gamma_n = 1/n$. The same analysis can also be done in the more general case when x is a vector, see (1.24). The resulting algorithm is of the form (1.7) with $\gamma_n = 1/n$ and S_n tending to a constant matrix.

However, in general when $Q(x,e)$ does not depend linearly on x, a more complex situation arises. This is the case for the adaptive algorithms in Examples 1.5 and 1.6, for the general recursive estimation algorithms of Example 1.4, and a variety of other cases. It is recognized by most people who have applied such algorithms that a considerable increase in convergence rate is obtained if γ_n is chosen to decrease more slowly than 1/n. Some specific examples are given below.

Example 6.1 - Self-tuning regulator.

An example with a self-tuning regulator (see Example 1.5) is shown in Fig. 6.1 (from Wittenmark (1973)). It is readily seen that the parameters tend to the desired values more rapidly for a constant $\gamma_n = \gamma_0$ than for $\gamma_n = 1/n$.

Example 6.2 - Recursive maximum likelihood.

In Fig. 6.2 (from Söderström (1973)) the result of recursive approximate maximum likelihood estimation is shown. Cf. Example 1.4. Again the curve that corresponds to a slower decrease in γ_n shows faster convergence.

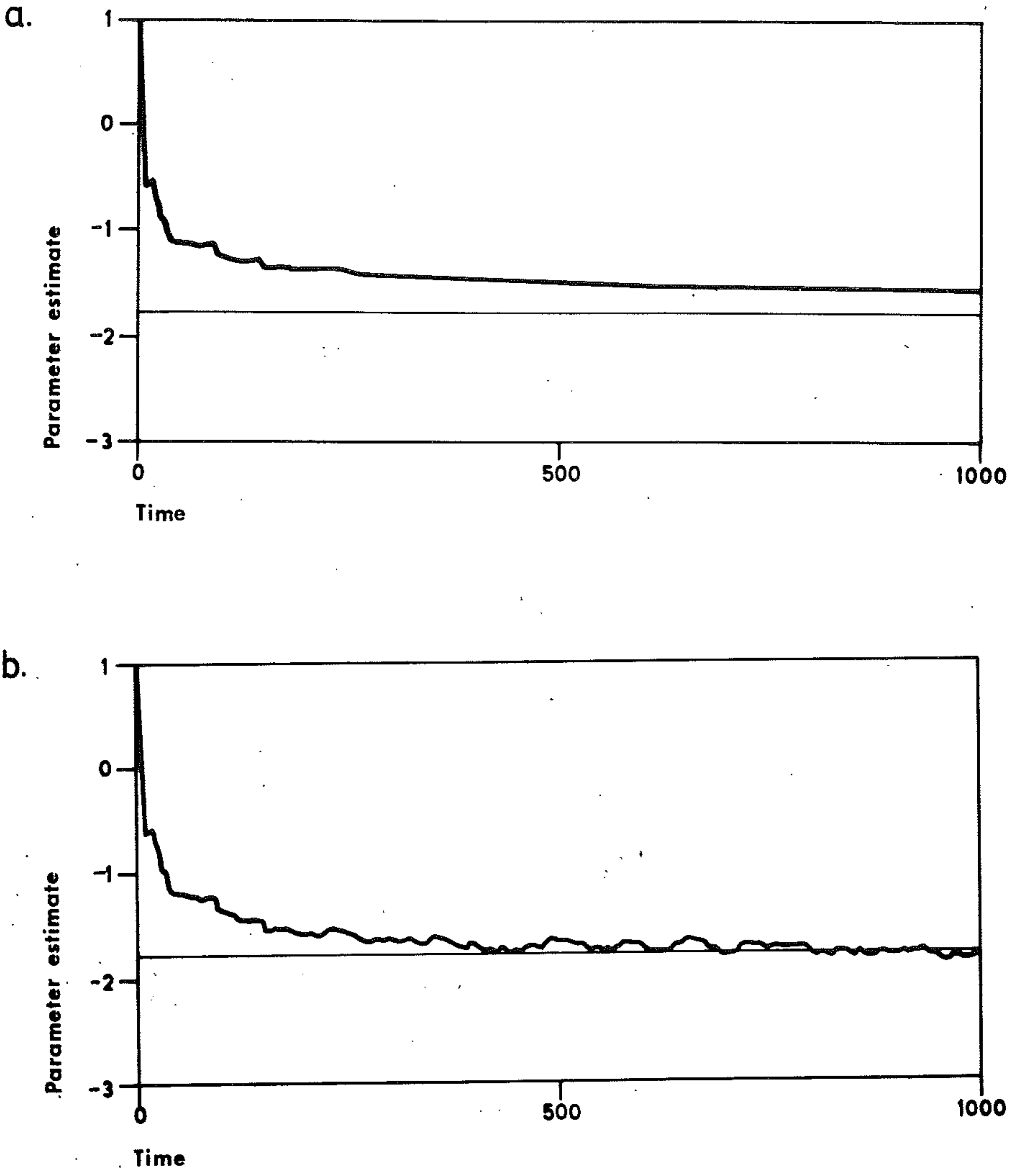
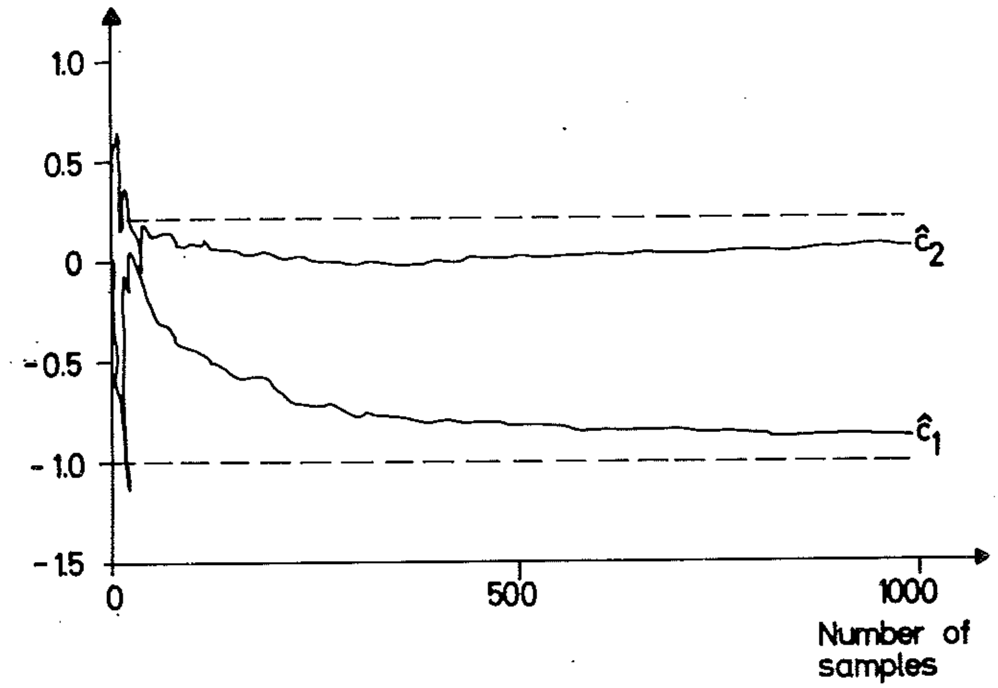


Fig. 6.1 (Wittenmark (1973)) - Convergence for a self-tuning regulator, with one parameter. The algorithm is of type (1.24), (1.30) with a) $\gamma_n = 1/n$ and b) γ_n = constant = 0.01.



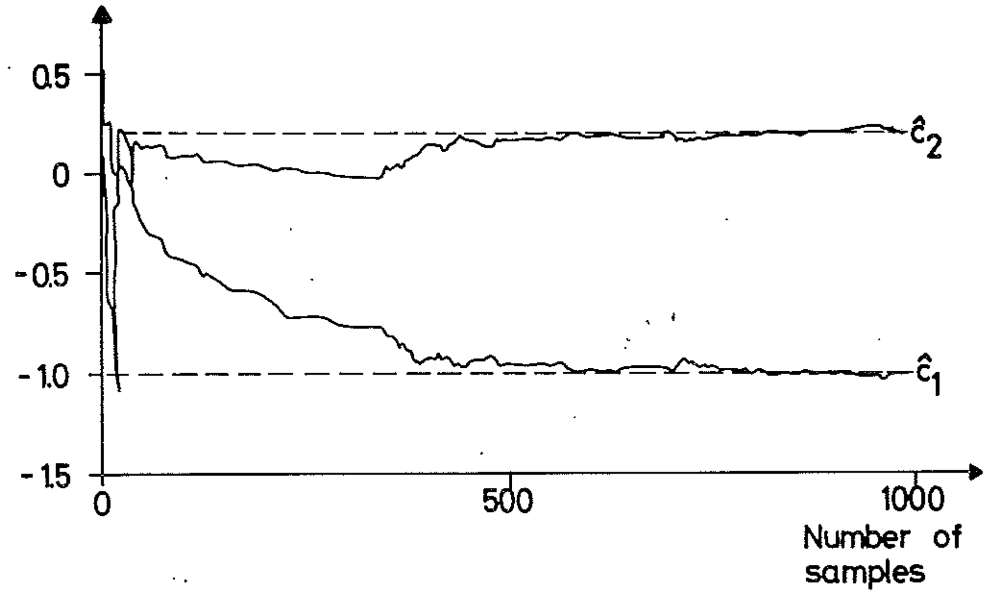


Fig. 6.2 (Söderström (1973)) - Convergence for C parameter estimates of a linear system (1.25). An approximate maximum likelihood method (1.29) has been used. The upper curves correspond to $\gamma_n = 1/n$ in (1.29). The lower curve is obtained if restarts are used. These make γ_n decrease more slowly than 1/n.

Also in the RM scheme with improved convergence rate, suggested by Kesten (1958) $\{\gamma_n\}$ is reduced slowly.

6.2. Heuristic Analysis.

Consider again the problem to solve (6.1) from measurements $Q(x,e_k)$, $k = 1,...$ It is convenient to introduce

$$\tilde{Q}(x,e) = Q(x,e) + x$$

Then (6.1) can be written

$$E_e\tilde{Q}(x,e) = x \tag{6.4}$$

The Robbins-Monro scheme (1.10) is one possibility to obtain estimates x_n recursively:

$$x_n = x_{n-1} + \gamma_n Q(x_{n-1},e_n); \quad x_0 = 0 \tag{6.5}$$

It is straightforward to show that x_n defined by (6.5) can be written

$$x_n = \sum_{k=1}^{n} \beta_k^n \tilde{Q}(x_{k-1},e_k) \tag{6.6}$$

where

$$\beta_k^n = \gamma_k \prod_{j=k+1}^{n} (1-\gamma_j) \qquad k < n; \quad \beta_n^n = \gamma_n$$

In this section (6.6) is heuristically interpreted as an approximate solution of (6.2), which can be written as

$$x_n = \frac{1}{n} \sum_{k=1}^{n} \tilde{Q}(x_n, e_k) \triangleq \tilde{f}_n(x_n) \tag{6.7}$$

Eq. (6.7) is to be solved for x_n. This cannot be done straightforwardly. One difficulty is that $\tilde{f}_n(x)$ is not available as a function of x. Even if it were, (6.7) is a non linear equation that must be solved. This can e.g. be done in the iterative manner

$$x_n^{(i)} = \tilde{f}_n\left(x_n^{(i-1)}\right) \tag{6.8}$$

If just one iteration is made and the previous estimate x_{n-1} is used as a starting value we have

$$x_n = \tilde{f}_n(x_{n-1}) = \frac{1}{n} \sum_{k=1}^{n} \tilde{Q}(x_{n-1}, e_k) \tag{6.9}$$

Now, $\tilde{Q}(x_{n-1}, e_k)$ are not known, and $\tilde{f}_n(x_{n-1})$ cannot be calculated. One possibility is to approximate $\tilde{Q}(x_{n-1}, e_k)$ with $\tilde{Q}(x_{k-1}, e_k)$ and so the sum in (6.9) is replaced by

$$\frac{1}{n} \sum_{k=1}^{n} \tilde{Q}(x_{k-1}, e_k) \tag{6.10}$$

Now, the last terms in (6.10) are likely to be better approximations than the first ones. Therefore it is reasonable to assume that a weighted sum

$$\sum_{k=1}^{n} \beta_k^n \tilde{Q}(x_{k-1}, e_k) \tag{6.11}$$

is a better approximation of the RHS of (6.9) if β_k^n increases with k, than if $\beta_k^n = 1/n$ as in (6.10). Combining (6.11) with (6.9) the following approximate solution to

(6.7) is obtained

$$x_n = \sum_{k=1}^{n} \beta_k^n \tilde{Q}(x_{k-1}, e_k)$$

where β_k^n are suitable weighting coefficients:

$$\sum_{k=1}^{n} \beta_k^n = 1 \qquad \beta_k^n < \beta_{k+1}^n \qquad \text{all } k < n \tag{6.12}$$

The estimate given by (6.12) is exactly of the form that the RM scheme gives, i.e. (6.6).

It is interesting to see what the property (6.12) means in terms of γ_n. Some calculation shows that

$$\beta_k^n < \beta_{k+1}^n \leftrightarrow \frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k} < 1 \tag{6.13}$$

This means that $1/\gamma_k$ increases more slowly than k, or that γ_k decreases more slowly than 1/k. The fact that such choices give faster convergence rates in practice can therefore be explained with the more suitable weighting of old observations in (6.6).

Remark. The analysis is just a heuristic one. The function $\tilde{Q}(x,e)$ was chosen as $Q(x,e) + x$. A similar analysis could be performed for

$$\tilde{Q}_\lambda(x,e) = Q(x,e) + \lambda x$$

The estimate x_n defined by (6.5) then is

$$x_n = \sum_{k=1}^{n} \beta_k^n(\lambda) Q(x_{k-1}, e_k) \quad \text{where} \quad \beta_k^n(\lambda) = \gamma_k \prod_{k+1}^{n-1} (1-\lambda\gamma_i)$$

Also the iterative scheme (6.8) has another convergence dynamics for this $\tilde{Q}$. The goodness of the correction obtained in one iteration, as in (6.9) depends on λ. There is consequently some interaction between the convergence dynamics of (6.8) and the weighting coefficients in (6.6). This is not stringently accounted for in the above analysis.

6.3. Bounds on the Sequence $\{\gamma_n\}$.

The choice of $\{\gamma_n\}$ affects not only the convergence rate. In many cases too large values of γ_n may cause instability effects in the algorithm (6.5).

Example 6.3. Consider the following scheme:

$$x_n = x_{n-1} + \gamma_n(e_n - Ax_{n-1}) \tag{6.14}$$

with

$$A = \begin{pmatrix} \delta & \omega \\ -\omega & \delta \end{pmatrix} ; \qquad \delta > 0$$

Introduce $P_n = Ex_n x_n^T$. Suppose that $\{e_n\}$ is a sequence of independent variables with normal distribution and zero mean value. Then

$$P_{n+1} = P_n - \gamma_n(AP_n + P_n A^T) + \gamma_n^2[AP_n A^T + \Sigma]$$

where $\Sigma = Ee_n e_n^T$. Suppose $\Sigma = I$. It is then straightforward to show that $\text{tr}\, P_{n+1} > \text{tr}\, P_n$ unless

$$\gamma_n < \frac{2\delta\, \text{tr}\, P_n}{2 + (\delta^2+\omega^2)\text{tr}\, P_n}$$

A necessary condition for this relation to hold obviously is

$$\gamma_n < \frac{2\delta}{\delta^2 + \omega^2} \tag{6.15}$$

However, applying Theorems 5.1, 3.1 and 4.3 to this algorithm, it can be shown that $x_n \to 0$ w.p.1 for any sequence $\{\gamma_n\}$ such that

$$\Sigma\gamma_n = \infty \qquad \text{and} \qquad \Sigma\gamma_n^p < \infty \quad \text{some real } p > 0 \tag{6.16}$$

Consequently, bounds on $\{\gamma_n\}$ like (6.15) are not necessary to achieve convergence. Such bounds, however, are of great importance to obtain convergence in practice. Algorithm (6.14) with $\omega = 5$ and $\delta = 0.5$ has been simulated for some choices of $\{\gamma_n\}$ that all satisfy (6.16) and theoretically give convergence. The results are shown in Table 6.1. □

Another example of bounded sequences $\{\gamma_n\}$ is the estimation algorithm (1.21):

$$x_{n+1} = x_n + \gamma_{n+1}\left\{\xi_n y(n+1) - \xi_n\xi_n^T x_n\right\}$$

To avoid that x_n assumes too large values $\{\gamma_n\}$ must be normalized. The choices

$$\gamma_n = \frac{1}{n}\left(\xi_n^T\xi_n\right)^{-1}$$

or

$$\gamma_n = \left(\sum_{k=1}^{n} \xi_k^T\xi_k\right)^{-1}.$$

have the effect that the sequence $\{\gamma_n\}$ is appropriately bounded. Notice that the real time least squares algorithm corresponds to

$$\gamma_n S_n = \left(\sum_1^n \xi_k\xi_k^T\right)^{-1}$$

Table 6.1 - Simulation of (6.14) with $\delta = 0.5$, $\omega = 5$ and $\Sigma = I$ for some sequences $\{\gamma_n\}$. The numbers shown are $|x_n|$.

n \ γ_n	$1/n$	$1/n^{0.1}$	$0.04/n$	$5/(500+n)$
0	1.42	1.42	1.42	1.42
1	6.99	6.99	1.39	1.41
2	17.55	31.56	1.41	1.41
5	66.99	2691.2	1.43	1.42
10	129.21	$3\cdot 10^6$	1.42	1.42
100	121.28	$\sim 10^{55}$	1.36	1.31
25000	8.70	$\sim 10^{75}$	1.22	0.02

6.4. Trajectories.

So far, it has been observed that $\{\gamma_n\}$ must be chosen to be sufficiently small to avoid unstable behaviour of (6.5). We have also indicated that (6.13) should be satisfied in order to improve the convergence rate. However, we have not been able to give any rules or quantitative estimates how to choose $\{\gamma_n\}$. It may be argued that the more $Q(x,e)$ changes with x, the less weight should the first terms in the sum (6.11) have. Clearly, if Q is independent of x, all terms should have the same weight. We will try and formalize such an argument.

The separation theorem 3.1 states that the ODE

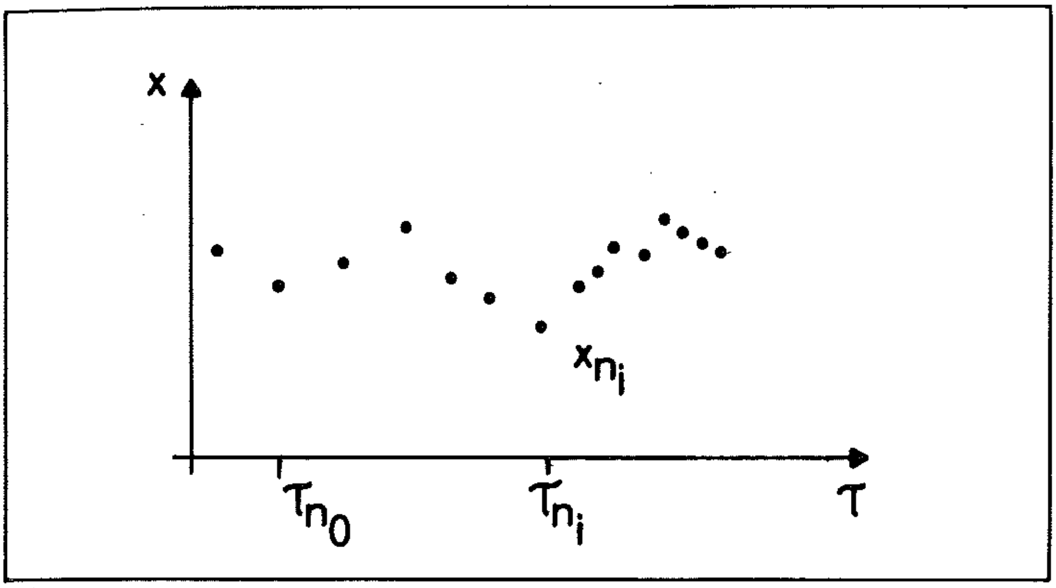
$$\frac{d}{d\tau} x = f(x) = E_e Q(x,e) \tag{6.17}$$

is important to decide convergence of (6.5). It can, in fact, be shown that the trajectories of (6.17) also govern the behaviour of the estimates x_n, obtained from (6.6). Loosely, the trajectories are the "expected paths" of $\{x_n\}$.

The result is formulated as follows. Let x_i, $i = n_0, \ldots,$ be generated by (6.5). The values can be plotted with the sample numbers i as the abscissa. It is also possible to introduce a fictitious time τ by

$$\tau_n = \sum_{i=1}^{n} \gamma_i$$

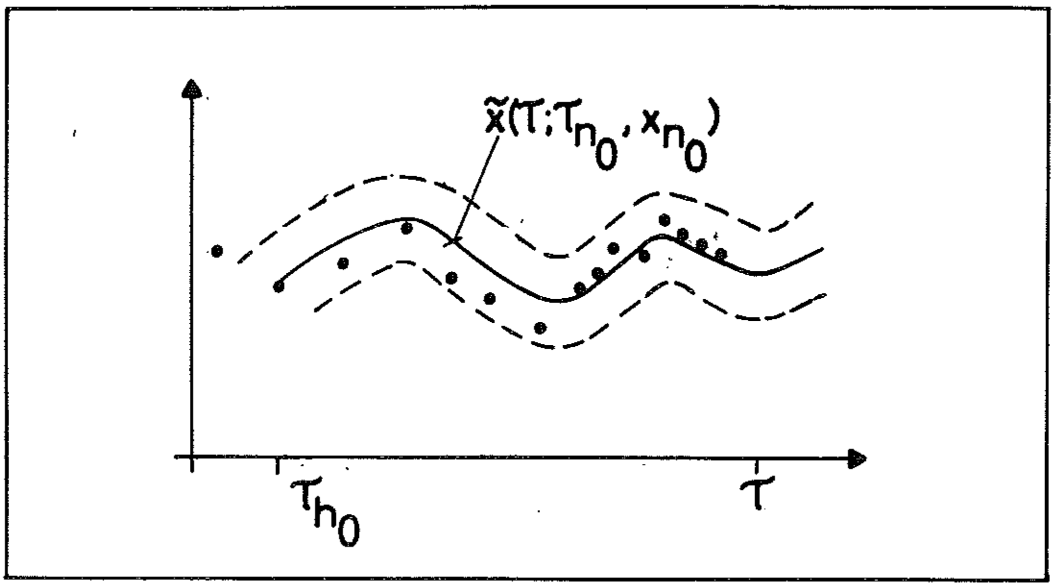
Suppose that the estimates x_i are plotted against this time τ:



Let $\tilde{x}(\tau;\tau_{n_0},x_{n_0})$ be the solution of (6.17) with initial value x_{n_0} at time τ_{n_0}:

$$\tilde{x}(\tau_{n_0};\tau_{n_0},x_{n_0}) = x_{n_0}$$

Plot also this solution in the same diagram:



Let I be a set of integers. The probability that all points x_i, $i \in I$, simultaneously are within a certain distance ε from the trajectory is estimated in the following theorem.

<u>Theorem 6.1</u>. Consider algorithm (6.5). Let $Q(x,e)$ be Lipschitz continuous for fixed e with Lipschitz constant $K(e)$. Assume that $E|Q(x^0,e_n)|^{2p} < C$ and $E|K(e_n)|^{2p} < C$ and suppose that $Q(x^0,e_n)$, $K(e_n)$ and $\{\gamma_n\}$ satisfy the conditions of Lemma 4.2. Assume that $f(x)$ is continuously differentiable. Denote

$$\sum_{i=1}^{n} \gamma_i = \tau_n$$

and denote the solution of (6.17) with initial condition $x(\tau_{n_0}) = x^0$ by $\tilde{x}(\tau;\tau_{n_0},x^0)$. Consider the ODE (6.17) linearized around this solution:

$$\frac{d}{d\tau} \Delta x = f'\left(x(\tau;\tau_{n_0},x_{n_0})\right)\Delta x$$

Assume that there exists a quadratic Lyapunov function for this linear, time varying ODE (see e.g. Brockett (1970)). Let I be a set of integers such that $\inf|\tau_i-\tau_j| = D > 0$ where $i \neq j$ and $i,j \in I$. Then there exists a K and an ε_0 such that for $\varepsilon < \varepsilon_0$

$$P\left\{\sup_{n\in I}|x_n - \tilde{x}(\tau_n;\tau_{n_0},x_{n_0})| > \varepsilon\right\} \leqslant \frac{K^{r}}{\varepsilon^{4r}} \sum_{j=n_0}^{N} (\gamma_j)^r \qquad r\leqslant p \quad (6.18)$$

where $N = \sup_{i\in I} i$, which may be ∞.

The proof is given in Appendix D.

□

If the sum $\Sigma\gamma_n^r$ is convergent, the right hand side of (6.18) can, for fixed ε, be chosen arbitrarily small by taking n_0 sufficiently large. Thus the theorem states that the trajectories of the ODE (6.17) arbitrarily well describe the behaviour of the algorithm (6.5) for sufficiently large time points.

We have in Section 6.3 observed that in some applications $\{\gamma_n\}$ has to be bounded by quite a small constant. For example, in the estimation algorithm (1.21) with $\gamma_n =$ $= \left(\Sigma \xi_k^T \xi_k\right)^{-1}$ a large observation $|\xi_k|$ causes all γ_n to be small. In these cases

$$\sum_{i=k}^{\infty} \gamma_i^r$$

is small and so the probability that the points x_n are outside a certain region around the trajectory also is small.

Although the proof of the theorem provides an estimate of K from given constants, we do not intend to use (6.18) to obtain numerical bounds for the probability. The point of the theorem is that a connection between the ODE (6.17) and the algorithm (6.5) is established.

Example 6.4. Consider again Example 6.3, with $\delta = 0.5$, $\omega = 5$ and $\Sigma = I$. Then, according to (6.5), γ_n has to be smaller than 0.04 to assure a stable behaviour. The sequence $\{\gamma_n\}$ has been chosen as

a) $\gamma_n = 0.04/n$

b) $\gamma_n = 1/(100+n)$

c) $\gamma_n = 0.01$

In Fig. 6.3 the results from simulations are shown. The sequence $\{x_n\}$ is there plotted against n and against the fictitious time

$$\tau_n = \sum_1^n \gamma_k$$

Corresponding phase planes are also shown.

The different choices of $\{\gamma_n\}$ can be seen as different scaling of the time. The slower $\{\gamma_n\}$ decreases, the faster runs the corresponding time.

Due to Theorem 6.1 the estimates have to follow the solution of the associated ODE. This is shown in Fig. 6.4. The way to make the estimates approach the desired point $x^* = 0$ fast, is to speed up the time, i.e. to make

$$\sum_1^n \gamma_k$$

as large as possible, while keeping the bound $\gamma_k < 0.04$. This clearly implies that $\{\gamma_n\}$ shall decrease slowly. However, the effect of the noise must be taken into consideration when the estimates are close to the origin. A comparison between Fig. 6.3c and Fig. 6.4 shows that the effect is not negligable.

The agreement between the simulations and the trajectories depends critically on the bound on γ_n that has to be chosen to avoid instability effects. If the trajectories are "straight", γ_n can be chosen larger and the noise has greater relative influence. In Fig. 6.5 simulations of (6.14) with $\delta = 5$, $\omega = 5$ and $\Sigma = I$ are shown. In this case large $\{\gamma_n\}$ can be chosen initially, and x_n quickly gets close to the origin. Then γ_n must decrease faster than in

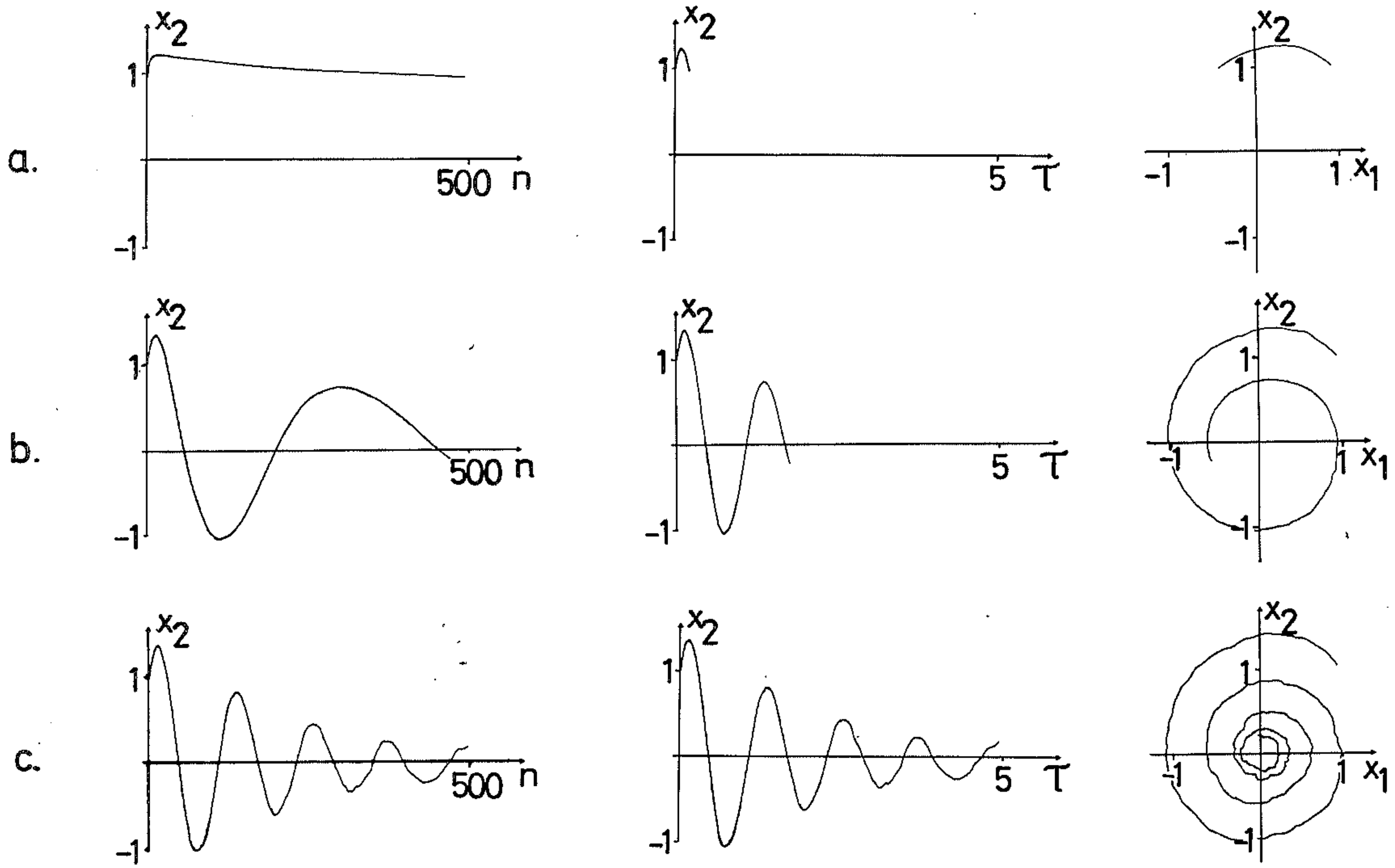
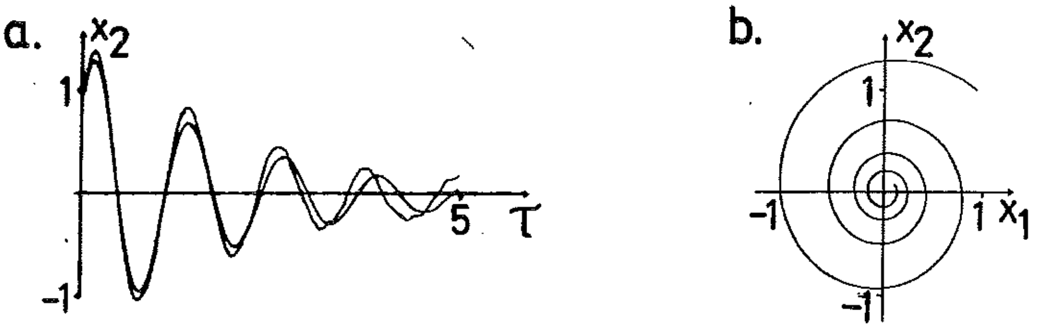
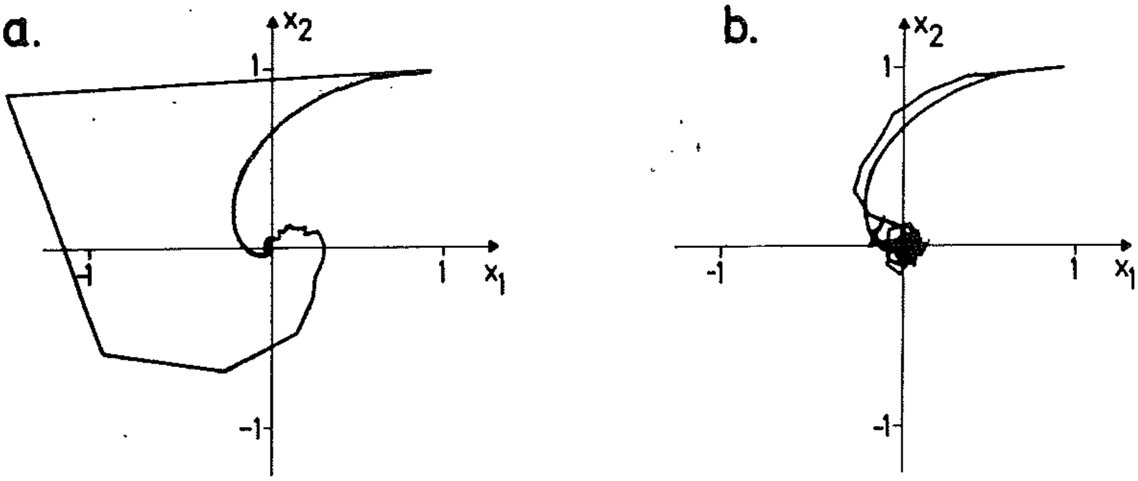


Fig. 6.3 - Simulation of (6.14) with $\delta = 0.5$, $\omega = 5$ and $\Sigma = I$. The second component of x_n is plotted against n and τ_n and against the first component of x_n for different choices of $\{\gamma_n\}$ a) $\gamma_n = 0.04/n$ b) $\gamma_n = 1/(100+n)$ c) $\gamma_n = 0.01$.



<u>Fig. 6.4</u> - Solution of the ODE associated to the algorithm (6.14) with $\delta = 0.5$, $\omega = 5$. In Fig. a, also the curve of Fig. 6.3c is shown.



<u>Fig. 6.5</u> - Simulation of (6.14) with $\delta = 5$, $\omega = 5$ and $\Sigma = I$ for a) $\gamma_n = 0.25/n$ and b) $\gamma_n = 0.05$, $n = 1,\ldots,100$. The solution of the associated ODE is also shown.

the previous case, in order to reduce the predominating influence of the noise.

In Fig. 6.6 simulations of the self-tuning regulator of Example 1.5 are shown. The regulator parameter estimates α and β for the system

$$y(t+1) - 0.99y(t) = u(t) + 0.5u(t-1) + e(t) - 0.7e(t-1)$$

where $\{e(t)\}$ is white noise, are plotted in a phase plane. These curves are compared with the trajectories of the corresponding ODE. It is seen that the trajectories well describe the behaviour of the algorithm.

To summarize, Theorem 6.1 and these examples show that numerical solution of the ODE (6.17) can be a valuable complement to simulation of the algorithm (6.5). The effect of various choices of the sequence $\{\gamma_n\}$ can also be understood in terms of this ODE. The advantage with sequences that decrease slowly (in the beginning of the procedure) can be clearly seen as in Example 6.4. This conclusion is the same as that of the heuristic analysis in Section 6.2.

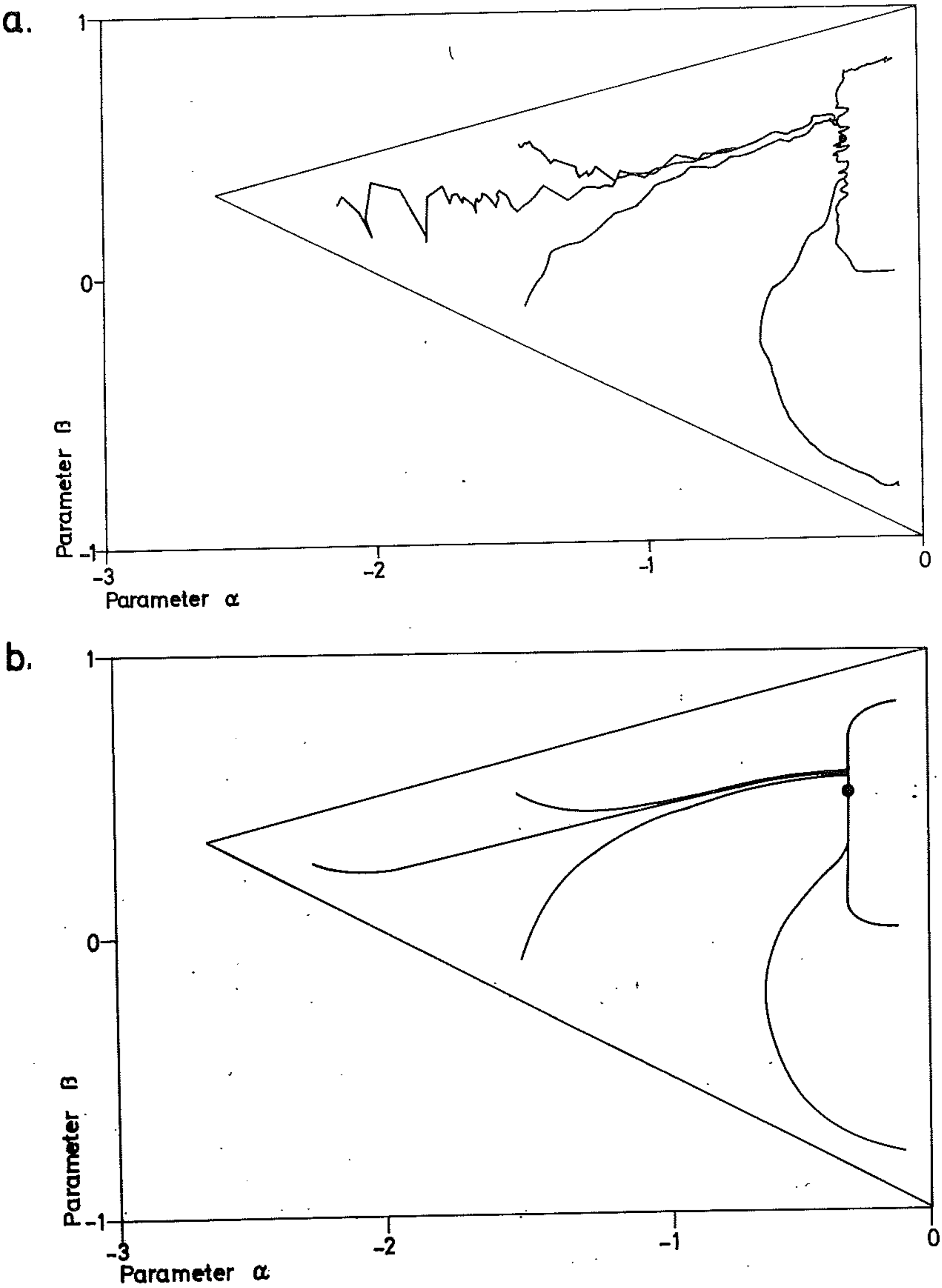


Fig. 6.6 - a) Simulation of the self-tuning regulator (1.21) with (1.30). The sequence γ_n is chosen as $0.002n^{-0.1}$.

b) Trajectories of the corresponding ODE.

7. SUMMARY AND DISCUSSION.

The separation theorem together with the results of Chapters 4 and 5 can be combined into several theorems on convergence. For a specific application it is,of course, possible also to tailor noise and boundedness conditions. In Section 7.1 two convergence theorems are given that seem to cover a large variety of applications. Some examples are given in Section 7.2. In Section 7.3 the significance of the results of the report are discussed.

7.1. Convergence Theorems.

Suppose in this section that $\{\gamma_n\}$ is deterministic. The sequence $\{\gamma_n\}$ in the algorithm

$$x_n = x_{n-1} + \gamma_n Q_n(x_{n-1}, e_n) \tag{7.1}$$

is often chosen such that

$$\gamma_n n \to A \text{ as } n \to \infty \tag{7.2}$$

For this choice Theorem 3.1 can be combined with (4.7) and Theorem 5.2 to yield the following result.

Theorem 7.1. Consider algorithm (7.1) with $\{\gamma_n\}$ satisfying (7.2). Assume that the estimates are projected into an area $D = \{x | U(x) \leqslant A\}$ where U satisfies conditions (5.8) and (5.9). Let $Q_n(x,e_n)$ be Lipschitz continuous in D for fixed e_n with Lipschitz constant $K_n(e_n)$. Introduce

$$f(x) = \lim_{n\to\infty} E_e Q_n(x,e)$$

and assume that the convergence is uniform in $x \in D$.

Suppose that the ODE

$$\dot{x} = f(x)$$

has an asymptotically stable solution $x(t) = x^*$, with domain of attraction $\supset D$. Suppose that

$$\mathrm{Cov}\big(K_n(e_n),K_m(e_m)\big) \leqslant C \cdot \frac{n^p + m^p}{1 + |n - m|^q} \qquad 0 \leqslant 2p < q < 1$$

and

$$\mathrm{Cov}[Q_n(x^0,e_n),Q_m(x^0,e_m)] \leqslant C \cdot \frac{n^p + m^p}{1 + |n - m|^q} \qquad 0 \leqslant 2p < q < 1$$

Then $x_n \to x^*$ w.p.1 as $n \to \infty$.

□

For more general sequences $\{\gamma_n\}$ the conditions on the noise terms must be somewhat strengthened:

Theorem 7.2. Consider algorithm (7.1) with the same conditions on D and f(x) as in Theorem 7.1. Suppose that $\{\gamma_n\}$ satisfies

$$\sum_{n=1}^{\infty} \gamma_n = \infty; \qquad \sum_{n=1}^{\infty} \alpha_n^{2p}\gamma_n^p < \infty \text{ (p integer)};$$

$$\{\gamma_n\} \text{ decreasing}; \qquad \limsup_{n\to\infty}\left[\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n}\right] < \infty$$

Assume that

$$E|Q_n(x^0,e_n)|^{2p} \leqslant \alpha_n^{2p}L_1(x^0) \quad \text{and} \quad E|K_n(e_n)|^{2p} \leqslant \alpha_n^{2p}L_2$$

where $\{\alpha_n\}$ is nondecreasing, and that the stochastic processes

$$Q_n(x^0,e_n) - E_{e_n}Q(x^0,e_n) \quad \text{and} \quad K_n(e_n) - EK_n(e_n)$$

can be considered as filtered white noise as in (4.10). Then $x_n \to x^*$ w.p.1 as $n \to \infty$.

□

These theorems are only examples of convergence results that can be synthesized from the results of Chapters 3-5. Some specific applications are given in the next section.

7.2. Applications.

1. The Robbins-Monro scheme (cf. Example 1.1)

Theorems 7.1 and 7.2 are directly applicable to the RM scheme. In terms of the classification of Chapter 2 the stability condition has another formulation, but is essentially the same as the one due to Blum (1954b). The boundedness condition (projection algorithm) is maybe more natural. The most important feature is that the noise condition is substantially weaker. Correlated observations can be handled and the conditions on the sequence $\{\gamma_n\}$ are weaker. If all moments of the noise exist, it is only required that

$$\Sigma\gamma_n^p < \infty \text{ some } p$$

This is e.g. true for $\gamma_n = n^{-\alpha}$, $0 < \alpha \leq 1$. The usual results allow just $1/2 < \alpha \leq 1$. The discussion of Chapter 6 shows that slowly decreasing sequences $\{\gamma_n\}$ are of interest as a way to obtain faster convergence.

Example 7.1. Consider the algorithm of Example 6.3. Suppose that $\{e_n\}$ is a sequence of normally distributed, zero mean valued random variables, with a covariance function $Ee_n^T e_{n+s}$ that tends to zero exponentially as s tends to infinity. In this case

$$Q(x_{n-1},e_n) = - Ax_{n-1} + e_n$$

which clearly is Lipschitz continuous, with a Lipschitz constant that does not depend on e_n. The ODE $\dot{x} = - Ax$ is globally asymptotically stable with stationary point $x^* = 0$. There exists a quadratic Lyapunov function that satisfies (5.3). Consequently, if

$$\sum_1^\infty \gamma_n = \infty; \sum_1^\infty \gamma_n^p < \infty \text{ some } p > 0 \left(\text{or even } \sum_1^\infty \exp(-\varepsilon/\gamma_n) < \infty \text{ all } \varepsilon > 0\right)$$

and

$$\limsup_{n\to\infty} \left[\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n}\right] < \infty$$

then $x_n \to 0$ w.p.1 as $n \to \infty$.

2. Minimization of a function using noise corrupted measurements (the Kiefer-Wolfowitz procedure). (Cf. Examples 1.2 and 3.3)

Consider as in Example 3.3 minimization of a function h(x) when only noise corrupted measurements are available:

$$J(x,e_n) = h(x) + e_n$$

where e_n is a sequence of random variables with zero mean and uniformly bounded 2p moments. Let $\{e_n\}$ be obtained from independent variables by linear, exponentially stable filtering. The minimization is performed with the Kiefer-Wolfowitz procedure with step size γ_n and increments a_n for numerical differentiation. Combining the results of Example 3.3 and Theorem 4.1 we have:

x_n tends w.p.1 either to infinity or to a stationary point of h(x) if for some p

$$\sum_1^\infty \gamma_n = \infty; \quad \{\gamma_n\} \text{ decreasing}; \quad \limsup_{n\to\infty} \left[\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n}\right] < \infty;$$

$$\sum_1^\infty (\gamma_n/a_n^2)^p < \infty; \quad a_n \to 0 \tag{7.3}$$

These conditions are considerably more general than the usually reported ones:

$$\Sigma\gamma_n = \infty; \quad \Sigma(\gamma_n/a_n)^2 < \infty; \quad a_n \to 0; \quad \Sigma\gamma_n a_n < \infty \tag{7.4}$$

Taking $\gamma_n = c_1 n^{-\alpha}$ and $a_n = c_2 n^{-\beta}$ conditions (7.4) imply that (α,β) must lie in the shaded region in Fig. 7.1. If we assume that all moments of the noise exist, our conditions, (7.3), require (α,β) to lie in the triangle A, which contains the shaded region.

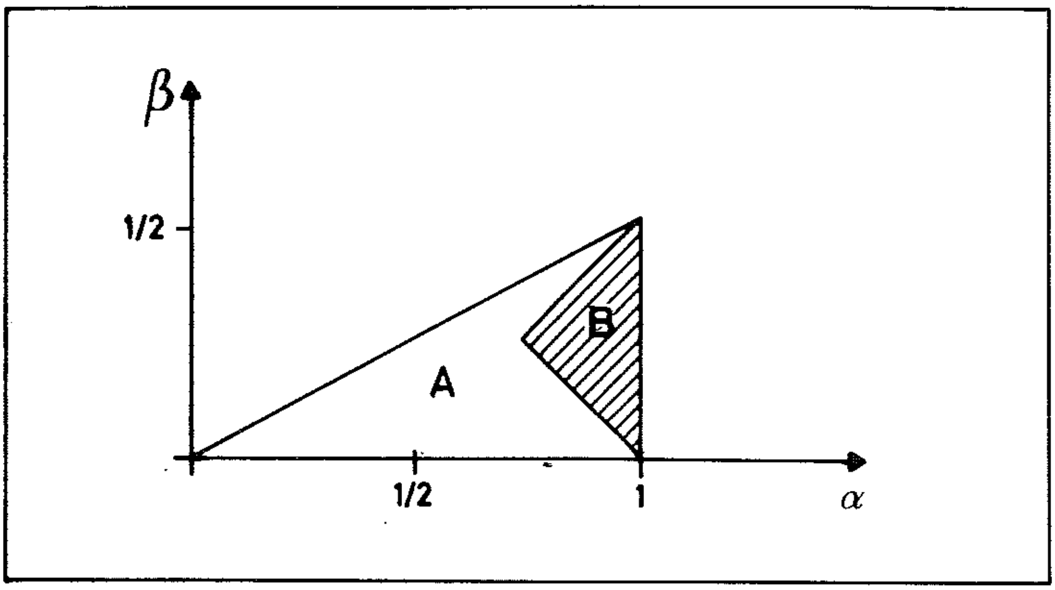


Fig. 7.1 - Values of (α,β) that give convergence in the Kiefer-Wolfowitz procedure with $\gamma_n = n^{-\alpha}$; $a_n = n^{-\beta}$.
Region A: according to (7.3). Region B: according to (7.4).

3. Adaptive systems.

Applications to real time least squares estimation has been considered in Example 3.4. The more complex problem with self-tuning regulators based on least squares estimation (Example 1.5) is treated in Ljung-Wittenmark (1974).

The self-learning classifier, suggested by Tsypkin (1968) (see also Braverman (1966)), can be analysed using the same tools:

Example 7.2. Consider the self-learning classifier defined in Example 1.6. Let e_n have the distribution shown in Fig. 7.2, consisting of two triangular distributions:

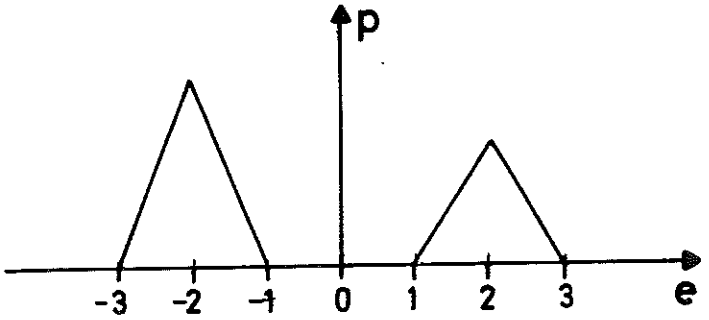


Fig. 7.2 - Probability density function of the random variable to be classified.

The probability of outcomes in the left triangle is λ. Clearly, it is desirable that the classification rule, the number c_n, should converge to some value between -1 and +1. Introduce as in Example 1.6

$$x_n = \begin{pmatrix} x_n^A \\ \\ x_n^B \end{pmatrix} .$$

Then (1.32) can be written

$$x_n = x_{n-1} + \gamma_n Q(x_{n-1}, e_n)$$

$E_e Q(x,e) = f(x)$ is readily computed as follows. For a given x the corresponding classification point is $c(x) = (x_1+x_2)/2$. $f_1(x)$ is then the mean value of the distribution left of the point c(x), minus x_1. $f_2(x)$ is found correspondingly. The algebraic expression for f(x) as a function of x and λ is lengthy and is omitted.

The trajectories of the ODE $x = f(x)$ are shown in Fig. 7.3 for two choices of λ. In case $\lambda = 0.5$ the variable x converges to $x^* = (-2,2)$ which gives a correct classification rule $c^* = 0$. The case $\lambda = 0.99$ corresponds to a common situation, when errors that occur rather seldom (1%), "outliers", shall be detected. In this case there are two possible convergence points, $x^* = (-2,2)$ and $x^{**} = (-2.3, -1.4)$. The latter gives a classification rule $c^{**} = -1.8$ which classifies 39% of "correct values" as outliers. For any starting value x_0 there is a positive probability that the classification rule converges to c^{**}.

In this example it is straightforward to numerically solve the corresponding ODE. It is quite cumbersome to find suitable Lyapunov functions for the problem.

□

7.3. Conclusion.

In this report it has been shown that an ordinary differential equation can be defined for certain classes of recursive, stochastic algorithms. These algorithms cover a variety of control applications. The ODE has been shown to contain information about convergence of the algorithms as well as about convergence rates and behaviour of the algorithm. In the analysis it has been possible to separate the stochastic part from the rest of the problem.

Martingale theory, which is the traditional tool to show convergence, has not been used. This causes some proofs to be more technical, because instead of just referring to the martingale theorem it is in fact necessary to go through some arguments that are used in the proof of it. However, the application of martingale theory requires some extra conditions, that can be removed with our approach.

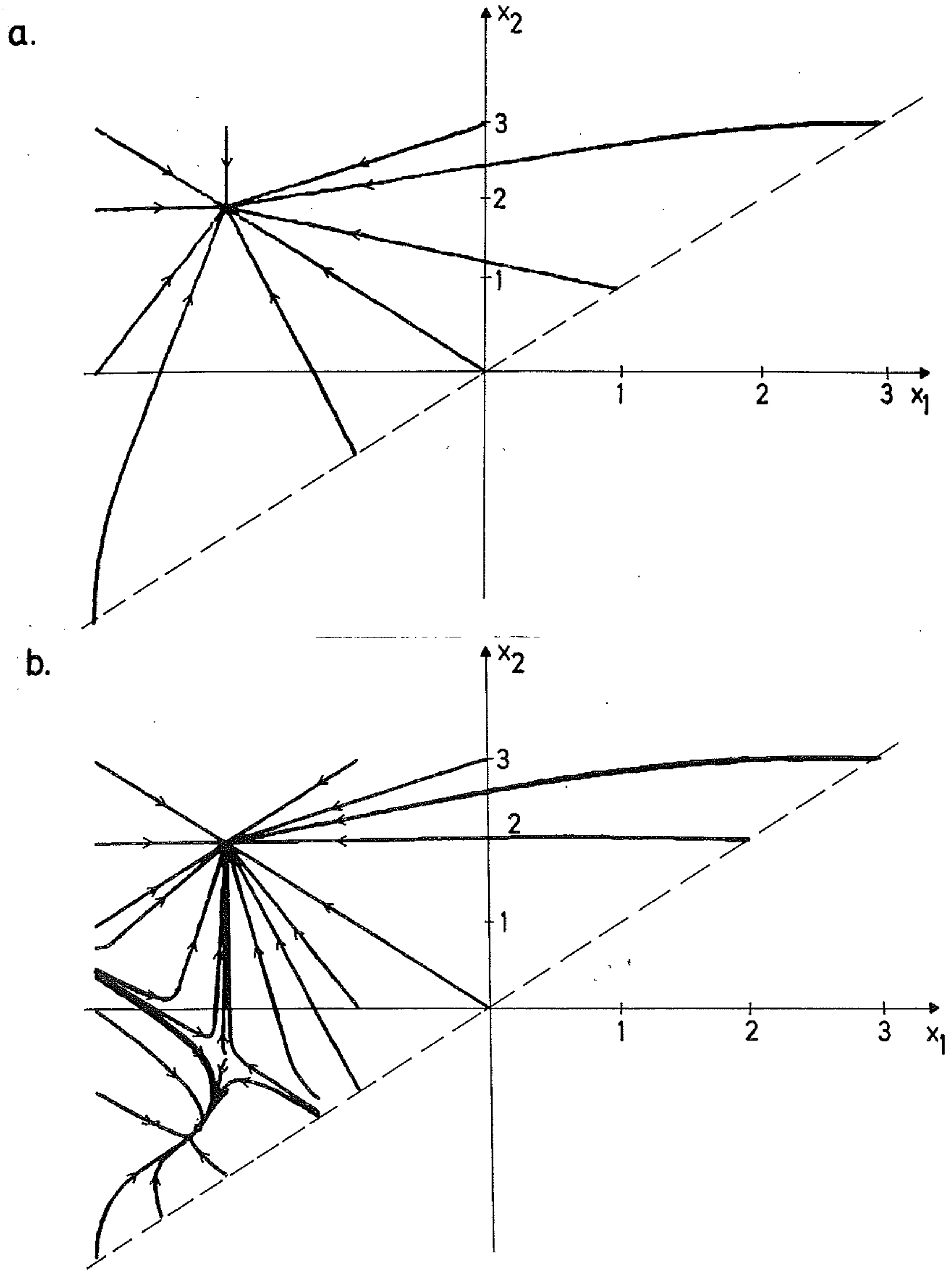


Fig. 7.3 - Trajectories for the ODE that corresponds to self-learning classification for the distributions defined in Example 7.2.

a) λ = 0.5   b) λ = 0.99

8. ACKNOWLEDGEMENTS.

I am glad to express my sincere thanks to Professor K.J. Åström who has given me most valuable guidance throughout this work.

The main part of the research has been carried out during my visit to Professor Ya.Z. Tsypkin's laboratory at the Institute for Control Problems in Moscow. I would like to express my deep gratitude to the whole laboratory for their hospitality and support during my stay. In particular I thank Professor Ya.Z. Tsypkin and Dr. B.T. Polyak for many stimulating discussions on the subject of this report.

I would also like to thank Per Hagander, Jacques de Maré and Torkel Glad who have contributed with important comments on the manuscript. Thanks also to Gudrun Christensen who has typed the manuscript and to Britt-Marie Carlsson who has prepared the figures.

Part of this work has been supported by the Swedish Board for Technical Development under contract No. 733546.

9. REFERENCES.

Aizerman, M.A., Braverman, E.M., and Rozonoer, L.I. (1964a) Theoretical Foundations of Potential Function Method in Pattern Recognition. Avtomat. i Telemek, Vol. 25, No. 6, pp.917-937.

Aizerman, M.A., Braverman, E.M., and Rozonoer, L.I. (1964b) Method of Potential Functions in the Problem of Restoration of Functional Converter Characteristics by Means of Points Observed Randomly. Avtomat. i Telemek, Vol. 25, No. 12, pp. 1705-1714.

Aizerman, M.A., Braverman, E.M., and Rozonoer, L.I. (1965) The Robbins-Monro Process and the Potential Functions Method. Avtomat. i Telemek, Vol. 26, No. 11, pp. 1951-1955.

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).

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

Åström, K.J. (1968) Lectures on the Identification Problem - The Least Squares Method. Report 6806, Div. of Aut. Control, Lund Inst. of Technology.

Åström, K.J., and Eykhoff, P. (1971) System Identification - A Survey. Automatica, 7, pp. 123-162.

Åström, K.J., and Wittenmark, B. (1973) On Self-Tuning Regulators, Automatica, Vol. 9, pp. 185-194.

von Bahr, B., and Esseen, C.-G. (1965)
Inequalities for the rth Absolute Moment of a Sum of Random Variables, 1 ≤ r ≤ 2. Ann. Math. Stat. 36, pp. 299-303.

Blum, J. (1954a)
Approximation Methods Which Converge With Probability One. Ann. Math. Stat. 25, pp. 382-386.

Blum, J. (1954b)
Multidimensional Stochastic Approximation Methods. Ann. Math. Stat. 25, pp. 737-744.

Braverman, E.M. (1966)
The Method of Potential Functions in the Problem of Training Machines to Recognize Patterns Without a Trainer. Avtom. i Telemek, Vol. 27, No. 10, pp. 1748-1770.

Braverman, E.M., and Rozonoer, L.I. (1969)
Convergence of Random Process in the Theory of Learning Machines I & II. Avtom. i Telemek, Vol. 30, No. 1, pp. 44-64, No. 3, pp. 386-402.

Brockett, R.W. (1970)
Finite Dimensional Linear Systems. Wiley, New York.

Burkholder, D.L. (1956)
On a Class of Stochastic Approximation Procedures. Ann. Math. Stat. 27, pp. 1044-1059.

Chung, K.L. (1954)
On Stochastic Approximation Methods. Ann. Math. Stat. 25, pp. 463-483.

Comer, J.P. (1964)
Some Stochastic Approximation Procedures For Use in Process Control. Ann. Math. Stat. 35, pp. 1136-1146.

Cramér, H., and Leadbetter, M.R. (1967)
Stationary and Related Stochastic Processes. Wiley, New York.

Derman, C., and Sacks, J. (1959)
On Dvoretzky's Stochastic Approximation Theorem. Ann. Math. Stat. 30, pp. 601-606.

Devyaterikov, I.P., Kaplinskij, A.I., and Tsypkin, Ya.Z. (1969)
Convergence of Learning Algorithms. Avtom. i Telemek, Vol. 30, No. 10, pp. 1619-1625.

Dharmadhikari, S.W., and Jogdeo, K. (1969)
Bounds on Moments of Certain Random Variables. Ann. Math. Stat. 40, pp. 1506-1508.

Doob, J.L. (1953)
Stochastic Processes. Wiley, New York.

Dvoretzky, A. (1956)
On Stochastic Approximation. Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability, I, pp. 39-55.

Gladyshev, E.G. (1965)
On Stochastic Approximation. Teorija verojatnostej i jeje primenenija. 10, No. 2, pp. 297-300 (in Russian).

Fu, K.S. (1969)
Learning System Theory. In L.A. Zadeh and E. Polak (Ed.): System Theory, pp. 425-466, McGraw-Hill, New York.

Kesten, H. (1958)
Accelerated Stochastic Approximation. Ann. Math. Stat. 29, pp. 41-59.

Kiefer, J., and Wolfowitz, J. (1952)
Stochastic Estimation of the Maximum of a Regression Function. Ann. Math. Stat. 23, 462-466.

Krasovskij, N.N. (1963)
Stability of Motion. Stanford University Press, Standford, Calif.

Krasulina, T.P. (1969)
Application of Stochastic Approximation Algorithms to Automatic Control Problems in the Presence of Strong Interference. Avtom. i Telemek, Vol. 30, No. 5, pp. 745-749.

Krasulina, T.P. (1972)
Robbins-Monro Process in the Case of Several Roots. Avtom. i Telemek, Vol. 33, No. 4, pp. 580-586.

Kushner, H.J. (1972)
Stochastic Approximation Algorithms for the Local Optimization of Functions with Nonunique Stationary Points. IEEE Trans. Aut. Control AC-17, pp. 646-655.

Kushner, H.J. (1973)
Stochastic Approximation Algorithms for Constrained Optimization Problems. Research Report, Div. of Applied Math. and Eng., Brown University, Providence, R.I.

Kushner, H.J., and Gavin, T. (1973)
Stochastic Approximation Type Methods for Constrained Systems: Algorithms and Numerical Results. Research Report, Div. of Applied Math. and Eng., Brown University, Providence, R.I.

Litvakov, B.M. (1968)
Convergence of Recurrent Algorithms for Pattern Recognition Learning. Avtom. i Telemek, Vol. 29, No. 1, 121-129.

Ljung, L. (1972)
Convergence Concepts for Adaptive Structures. Report 7218, Div. of Automatic Control, Lund Inst. of Technology.

Ljung, L., and Wittenmark, B. (1974)
Asymptotic Properties of Self-Tuning Regulators. Report 7404. Division of Automatic Control, Lund Institute of Technology.

Mendel, J.M. (1973)
Discrete Techniques of Parameter Estimation. Marcel Dekker Inc., New York.

Petrov, V.V. (1972)
Summy nezavisimykh sluchajnykh velichin (Sums of Independent Random Variables). Nauka, Moscow (in Russian).

Robbins, M., and Monro, S. (1951)
A Stochastic Approximation Method. Ann. Math. Stat. 22, pp. 400-407.

Sakrison, D.J. (1967)
The Use of Stochastic Approximation to Solve the System Identification Problem. IEEE Trans. AC-12, pp. 563-567.

Saridis, G.N., Nikolic, Z.J., and Fu, K.S. (1969)
Stochastic Approximation Algorithms for System Identification, Estimation, and Decomposition of Mixtures. IEEE Trans. Systems Science and Cybernetics, SSC-5, pp. 8-15.

Söderström, T. (1973)
An On-Line Algorithm for Approximate Maximum Likelihood Identification of Linear Dynamic Systems. Report 7308, Div. of Aut. Control, Lund Institute of Technology.

Tsypkin, Ya,Z. (1968)
Self-Learning - What Is It? IEEE Trans. Aut. Control AC-13, pp. 608-612.

Tsypkin, Ya.Z. (1971)
Adaption and Learning in Automatic Systems. Academic Press, New York.

Tsypkin, Ya.Z. (1973)
Foundations of the Theory of Learning Systems. Academic Press, New York.

Wasan, M.T. (1969)
Stochastic Approximation. Cambridge, University Press.

Wieslander, J. (1969)
Real Time Identification - Part I. Report 6908, Div. of Automatic Control, Lund Institute of Technology.

Wittenmark, B. (1973)
A Self-Tuning Regulator. Report 7311, Division of Automatic Control, Lund Institute of Technology.

Zolotarev, V.M. (1965)
On Some Inequalities of Chebyshev Type. Litovskij matem. sb., Vol. 5, No. 2, pp. 233-250.

APPENDIX A.

Proof of Theorem 3.1.

Before prodeeding to the proof, let us first remark that although $z_n(x^0)$ converges w.p.1 for each fixed x^0, this does not imply convergence if x^0 is a random variable. To treat such problems in a strict way, introduce a denumerable subset of D

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

which is dense in D.

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

Assumption a) 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

$$\Omega^* = \bigcap_{i=1}^{\infty} \Omega^{(i)} \cap \Omega_A$$

where Ω_A is the set of all realizations for which r_n converges, condition b) is satisfied and $\gamma_n \to 0$; $\Sigma\gamma_n = \infty$. 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 (3.1) behaves like solutions of the ODE (3.6) 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 $t \leqslant t_0$, where t_0 does not depend on $\bar{x}$ and ω. Define the sequence $m(n,t,\omega)$ so that

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

Then, if $x_n(\omega)$ belongs to a closed subset of D^0,

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

where

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

and

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

Proof of Lemma A.1. Consider the sequence $\{z_n(\bar{x},\omega)\}$ defined by (3.4). Let $n \leqslant j(n,\omega) \leqslant m(n,t,\omega)$. Then

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

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

$$\lim_{n\to\infty} \sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega)\left\{Q_i\left(\bar{x},e_i(\omega)\right) - z_{i-1}(\bar{x},\omega)\right\} = 0$$

or

$$\sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega)Q_i\left(\bar{x},e_i(\omega)\right) - f(\bar{x}) \sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega) =$$

$$= R_3\left(j(n),n,\omega,\bar{x}\right) \tag{A.2}$$

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

Analogously

$$\lim_{n\to\infty} \sum_{n+1}^{m(n,\omega)} \gamma_i(\omega)K_i\left(e_i(\omega)\right) = r_\infty \lim_{n\to\infty} \sum_{n+1}^{m(n,\omega)} \gamma_i(\omega) = r_\infty t \tag{A.3}$$

where $r_\infty = \lim\limits_{n\to\infty} r_n$

Consider now

$$x_{j(n,\omega)}(\omega) = x_n(\omega) + \sum_{n+1}^{j(n,\omega)} \gamma_i(\omega)Q_i\left(x_{i-1}(\omega),e_i(\omega)\right) =$$

$$= x_n(\omega) + \sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega)Q_i\left(\bar{x},e_i(\omega)\right) +$$

$$+ \sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega)\left\{Q_i\left(x_{i-1}(\omega),e_i(\omega)\right) - Q_i\left(\bar{x},e_i(\omega)\right)\right\} \tag{A.4}$$

The first sum of the RHS of (A.4) can be approximated using (A.2). To approximate the second sum, use the Lipschitz continuity of Q_i:

$$\left|Q_i\left(x_{i-1}(\omega),e_i(\omega)\right) - Q_i\left(\bar{x},e_i(\omega)\right)\right| \le \left|x_{i-1}(\omega) - \bar{x}\right| K_i\left(e_i(\omega)\right)$$

(Assume for a moment that $x_i; n \leqslant i \leqslant j(n,\omega)$ belong to D^0, so that the Lipschitz continuity holds. This assumption is removed below.)

Hence

$$\left|\sum_{i=n+1}^{j(n,\omega)} \gamma_i(\omega)\left\{Q_i\left(x_{i-1}(\omega),e_i(\omega)\right) - Q_i\left(\bar{x},e_i(\omega)\right)\right\}\right| \leqslant$$

$$\leqslant \max_{n\leqslant i\leqslant j} |x_i(\omega) - \bar{x}| \sum_{i=n+1}^{m(n,t,\omega)} \gamma_i(\omega)K_i\left(e_i(\omega)\right) \leqslant$$

$$\leqslant \max_{n\leqslant i\leqslant m} |x_i(\omega) - \bar{x}|\{r_\infty t + R_4(t,n,\omega)\} \leqslant$$

$$\leqslant [\max_{n\leqslant i\leqslant m} |x_i(\omega) - x_n(\omega)|][r_\infty t + R_4(t,n,\omega)] \qquad (A.5)$$

where $R_4(t,n,\omega) \to 0$ as $n \to \infty$ according to (A.3).

Assume that

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

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

$$S(n,t,\omega) = |x_j(\omega) - x_n(\omega)| \leqslant \left|f(x) \sum_{n+1}^{j} \gamma_i(\omega) + R_3\left(j^{(n)},n,\omega,\bar{x}\right)\right| +$$

$$+ [S(n,t,\omega) + |x_n(\omega) - \bar{x}|][r_\infty t + R_4(t,n,\omega)]$$

or

$$S(n,t,\omega)[1 - r_\infty t - R_4(t,n,\omega)] \leqslant |f(x)|t + R_3(j^{(n)},n,\omega,\bar{x}) +$$

$$+ |\bar{x} - x_n(\omega)|[r_\infty t + R_4(t,n,\omega)]$$

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

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

Hence

$$S(n,t,\omega) \leqslant 2[|f(x)| + r_\infty C_1]t + R_3(j(n),n,\omega,\bar{x}) + C_1 R_4(t,n,\omega) =$$

$$= C_2 t + R_5(j(n),n,\omega,\bar{x})$$

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

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

$$|x_{m(n,t,\omega)}(\omega) - x_n(\omega) - tf(\bar{x})| \leqslant$$

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

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

$$= C_2 r_\infty t^2 + r_\infty t|\bar{x} - x_n(\omega)| + R_2(t,n,\bar{x},\omega)$$

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

It now remains only to remove the assumption $x_i \in D^0$, $n \leqslant i \leqslant m(n,t,\omega)$. If this assumption does not hold, let

$i = \bar{j}(n,\omega)$ be the first time $x_i \notin D^0$. Then apply the results above to $j(n,\omega) = \bar{j}(n,\omega)$, which gives

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

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

□

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

o $V(x)$ is infinitely differentiable

o $0 \leq V(x) < 1 \Rightarrow x \in D_1$ and $V(x) = 0 \Leftrightarrow x = x^*$

o $\frac{d}{dt} V(x) = V'(x)f(x)$ 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.

An outline of the rest of the proof is as follows:

<u>Step 1</u>: A convergent subsequence $\{x_{n_k}\}$ tending to $\tilde{x}$ is considered. Then $\{x_{n_k}\}$ is close to $\tilde{x}$ infinitely often, and according to Lemma A.1, $x_{m(n_k,t)}$ will approximately be $x_{n_k} + tf(\tilde{x})$. This means that $V(x_{m(n_k,t)})$ is strictly less than $V(x_{n_k})$ if $\tilde{x} \neq x^*$. A complication in this step is that $\tilde{x}$ may not belong to D_d. The formal proof is somewhat lengthy and involves several elaborate choices of constants. The result is, however, intuitively clear. The proof of step 1 follows over the next few pages and extends to eq. (A.10).

Step 2: From the above result it is quite clear that x^* must be a cluster point to $\{x_n\}$, since $V(x_n)$ has a tendency to decrease everywhere in D except for $x = x^*$. That this actually is the case is shown in Lemma A.2.

Step 3: If there is another cluster point to $\{x_n\}$ than x^*, say $\hat{x}$, the sequence must move from x^* to $\hat{x}$ infinitely many times. But then $V(x_n)$ is increasing, which contradicts the result of step 1. Hence only one cluster point exists and convergence follows. The formal proof of this claim is given in Lemma A.3.

From condition b) 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 $\varepsilon > 0$ an element $\bar{x} = \bar{x}(\tilde{x},\varepsilon) \in D_d$ such that $|\bar{x} - \tilde{x}| < \varepsilon/2$. Consequently

$$|x_{n_k} - \bar{x}| < \varepsilon \qquad k > K_0(\varepsilon) \tag{A.6}$$

Consider now

$$V\left(x_{m(n_k,t)}\right) - V(x_{n_k})$$

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

$$V(x_{k''}) - V(x_{k'}) = V'(\xi_k)\ (x_{k''}-x_{k'}) =$$

$$= V'(\bar{x})\ (x_{k''}-x_{k'}) + (\xi_k-\bar{x})^T V''(\xi_k')(x_{k''}-x_{k'}) \tag{A.7}$$

where

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

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

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

where

$$|R_1(t,n_k,x)| < tK|x_{n_k} - \bar{x}| + At^2 \qquad (A.8)$$

and

$$R_2(t,n_k,x) \to 0 \text{ as } k \to \infty \qquad (A.9)$$

Insert this in (A.7):

$$V(x_{k''}) - V(x_{k'}) = tV'(\bar{x})\, f(\bar{x}) + R_6(t,n_k,\bar{x})$$

where

$$R_6(t,n_k,\bar{x}) = (\xi_k-\bar{x})^T V''(\xi_k')(x_{k''}-x_{k'}) + V'(\bar{x})\,\{R_1 + R_2\}$$

Now suppose that the cluster point $\tilde{x}$ is different from the desired convergence point x^*. Then $V'(\tilde{x})f(\tilde{x}) = -\delta$, $\delta > 0$. This implies that $\exists\ \varepsilon_0$ such that

$$V'(\bar{x})f(\bar{x}) < -\delta/2 \quad \text{if} \quad |\bar{x} - \tilde{x}| < \varepsilon_0$$

Denote

$$\sup_{|\xi-\bar{x}|<\varepsilon_0} |V''(\xi)| = C_1, \quad \sup_{x\in D} |V'(x)| = C_3,$$

$$|f(\bar{x})| + At_0 + \varepsilon = C_2(\varepsilon)$$

Then

$$|(\xi_k-\bar{x})^T V''(\xi_k')(x_{k''}-x_{k'})| \leq C_1[tC_2(\varepsilon) + R_2]^2$$

First choose $\varepsilon = \min\left(\varepsilon_0,\ \delta/(4C_3K)\right)$ and $k > K_0(\varepsilon)$. Then

$$|V'(\bar{x})R_1| < t[\delta/4 + AC_3t]$$

$$|R_6(t,n_k,\bar{x})| \leq C_1[tC_2(\varepsilon) + R_2]^2 + t(\delta/4+At) + R_2 =$$

$$= t\delta/4 + t^2\left(C_1C_2^2(\varepsilon) + C_3A\right) + R_2C_3 +$$

$$+ C_1R_2^2 + 2R_2C_1C_2(\varepsilon)$$

Now choose

$$t \leq \frac{\delta}{8C_2^2(\varepsilon) + C_3A}$$

which gives

$$|R_6(t,n_k,\bar{x})| \leq 3t\delta/8 + R_2C_3 + C_1R_2^2 + 2C_1C_2(\varepsilon)R_2$$

Finally choose $K > K_0(\varepsilon)$ so that

$$R_2C_3 + C_1R_2^2 + 2C_1C_2(\varepsilon)R_2 < t\delta/16 \quad \text{for } k > K$$

which is possible since $R_2(t,n_k,x) \to 0$ as $k \to \infty$.

Hence

$$V(x_{k''}) - V(x_{k'}) < -t\delta/2 + R_6(t,n_k,\bar{x}) < -t\delta/32$$

or

$$V\left(x_{m(n_k,t)}\right) < V\left(x_{n_k}\right) - t\delta/32 \quad k > K$$

Since $x_{n_k} \to \tilde{x}$ as $k \to \infty$ and V is continuous this implies

$$V\left(x_{m(n_k,t)}\right) < V(\tilde{x}) - t\delta/64 \qquad k > K_1 \tag{A.10}$$

This means that if $\tilde{x}$ is a cluster point different from x^* the sequence x_n will i.o. be strictly interior to $\{x|V(x) \leqslant V(\tilde{x}) - t\delta/64\}$. This region is compact. Consequently another cluster point must exist that yields a smaller value of V. In Lemma A.2 it is shown that this implies that also x^* must be a cluster point, i.e. that

$$\liminf_{n\to\infty} V(x_n) = 0 \tag{A.11}$$

To conclude the proof it must also be shown that

$$\limsup_{n\to\infty} V(x_n) = 0 \tag{A.12}$$

This is done in Lemma A.3.

Lemma A.2. Suppose (A.10) holds for any subsequence $\{x_{n_k}\}$ that converges to a point different from x^*. Then (A.11) holds.

Proof. 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.10) $V(x_k)$ takes a value less than $U - \delta t/64$ infinitely often, which contradicts U being the infimum. Hence $U = 0$, which means that x^* is a cluster point.

□

Lemma A.3. From (A.10) and (A.11) it follows that

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

Proof. If $x_n \in D$ the difference

$$|x_{n+1} - x_n| = |\gamma_n Q_n(x_n, e_{n+1})| \leqslant \gamma_n |Q_n(x^*, e_{n+1})| + $$
$$+ |x_n - x^*| \gamma_n |K_n(e_{n+1})| \leqslant$$
$$\leqslant |z_{n+1}(x^*) - z_n(x^*)| + \gamma_n |z_n(x^*)| +$$
$$+ |x_n - x^*| \left\{ |r_{n+1} - r_n| + \gamma_n r_n \right\}$$

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

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

Consider the interval $I = [A/3, 2A/3]$ [1]. 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:

$$V\left(x_{n_k'-1}\right) < A/3 \qquad V\left(x_{n_k'}\right) \geqslant A/3 \qquad V\left(x_{n_k'+s_k}\right) > 2A/3$$

where s_k is the first s for which $V\left(x_{n_k'+s}\right) \notin I$ and $x_{n_k'} \to \tilde{x}$ as $k \to \infty$. Clearly $V(\tilde{x}) = A/3$.

[1] If A > 1 take $I = [1/3, 2/3]$.

Now, from (A.10)

$$V\left(x_{m(n'_k,t)}\right) < A/3 - \delta t/64$$

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.

□

Lemma A.3 implies that $x_n \to x^*$ for the chosen realization. The set of all ω for which this holds, Ω^*, has measure 1. This concludes the proof of the theorem.

□

Proof of Corollary 2. The proof of Corollary 2 is very much like the proof of the theorem. Eq. (A.10) is obtained straightforwardly if $\tilde{x} \notin D_c$. The only differences are caused by the weaker properties of V(x) in this case. In particular need the area $\{x|V(x) \leq V(\tilde{x})\}$ not be bounded. Therefore the closure of the set

$$\{x|V(x) \leq V(\tilde{x})\} \cap D$$

is considered instead. The conclusion of the corresponding Lemma A.2 therefore is that inf V(x) taken over the cluster points in D is assumed at a cluster point $\hat{x}$, such that either $\hat{x} \in \partial D$ or $\hat{x} \in D_c$.

If $\hat{x} \in \partial D$, eq. (A.10) implies that $\{x_n\}$ is outside D infinitely often. Since it also is inside D infinitely often, there must be another cluster point $\hat{\hat{x}}$ on the boundary. At this point the trajectories are pointing into the region D and V(x) is decreasing. This contradicts $\hat{x}$ giving the infimum of $V(\hat{x})$ since $V(\hat{x}) = V(\hat{\hat{x}})$.

Hence $\hat{x} \in D_c$ is a cluster point. Now suppose that there is another cluster point $\tilde{x} \notin D_c$. Then clearly, $V(\tilde{x}) > V(\hat{x})$ and $\{x_n\}$ moves from $\hat{x}$ to $\tilde{x}$ infinitely many times, corresponding to increasing V(x). As in Lemma A.3 this is contradicted, and the corollary follows.

□

APPENDIX B

Proof of Lemma 4.2.

Lemma 4.2. Consider the algorithm (4.3)

$$y_n = y_{n-1} + \gamma_n(f_n - y_{n-1}) \qquad y_0 = 0 \qquad (B.1)$$

Assume that the sequence $\{\gamma_n\}$ satisfies (4.8). Assume further that f_n satisfies (4.10) and that $\{\alpha_n\}$ is a non decreasing sequence of numbers and

$$E|e_k|^p < C$$

which implies

$$E|f_n|^p < C'\alpha_n^p$$

where p is an even integer. Then

$$E|y_n|^r \leqslant K_r(\alpha_n)^r(\gamma_n)^{r/2} \qquad 1 < r \leqslant p \qquad (B.2)$$

Proof It is evidently sufficient to show (B.2) for $r = p$, since Lyapunov's inequality

$$\left(E|x|^r\right)^{1/r} \leqslant \left(E|x|^{r'}\right)^{1/r'} \qquad 1 < r \leqslant r' < \infty$$

gives

$$\left[E\left|\frac{y_n}{\alpha_n\sqrt{\gamma_n}}\right|^r\right]^{1/r} \leqslant \left[E\left|\frac{y_n}{\alpha_n\sqrt{\gamma_n}}\right|^p\right]^{1/p} \leqslant (K_p)^{1/p} \qquad 1 < r \leqslant p$$

and so

$$E(|y_n|^r) \leqslant (K_p)^{r/p}\alpha_n^r\gamma_n^{r/2}$$

The solution of (B.1) can be written

$$y_n = \left[\prod_{k=m+1}^{n} (1-\gamma_k)\right]y_m + \sum_{k=m+1}^{n} \beta_k^n f_k \qquad (B.3)$$

where β_k^n is defined by (4.5).

Then, according to (4.8c)

$$\beta_k^n \leqslant \beta_n^n = \gamma_n \qquad k \leqslant n$$

Choose a subsequence n_k, such that

$$\sum_{i=n_k+1}^{n_{k+1}} \gamma_i \to t \quad \text{as } k \to \infty$$

This is possible since

$$\sum_1^\infty \gamma_i$$

diverges and $\gamma_n \to 0$. From (4.8d) we have

$$\sum_{n_{k-1}}^{n_k} \gamma_i \geqslant (n_k-n_{k-1})\gamma_{n_k}$$

and thus

$$(n_k - n_{k-1}) \leq K_1 / \gamma_{n_k} \tag{B.4}$$

Introduce

$$T_k = \sum_{i=n_{k-1}+1}^{n_k} \beta_i^{n_k} f_i$$

The lemma will now be proved by first (Lemma B.1) estimating $E|T_k|^p$ and then extending (Lemma B.2) this estimate to y_n.

Lemma B.1. With the assumptions of the theorem

$$E|T_k|^p \leq K_p \alpha_{n_k}^p \gamma_{n_k}^{p/2}$$

Proof of Lemma B.1.

$$E\ T_k^p \leq \sum_{j_1=n_{k-1}+1}^{n_k} \cdots \sum_{j_p=n_{k-1}+1}^{n_k} \beta_{j_1}^{n_k} \cdot \ldots \cdot \beta_{j_p}^{n_k} |Ef_{j_1} \cdot \ldots \cdot f_{j_p}| \leq$$

$$\leq \gamma_{n_k}^p \sum_{j_1} \cdots \sum_{j_p} |Ef_{j_1} \cdot \ldots \cdot f_{j_p}| \tag{B.5}$$

Now consider

$$\sum_{j_1} \cdots \sum_{j_p} |Ef_{j_1} \cdot \ldots \cdot f_{j_p}| \leq$$

$$\leq \sum_{j_1} \cdots \sum_{j_p} \left[\sum_{k_1=0}^{\infty} \cdots \sum_{k_p=0}^{\infty} |h_{k_1,j_1} \ldots h_{k_p,j_p} Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}| \right] \leq$$

$$\leq \sum_{j_1} \cdots \sum_{j_p} \left[\sum_{k_1=0}^{\infty} \cdots \sum_{k_p=0}^{\infty} \alpha_{n_k}^{p} \lambda^{k_1+\ldots+k_p} |Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}| \right]$$

Consider

$$Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}$$

If some index j_i-k_i occurs only once in the product above, the expectation value is zero, since e_n and e_m are independent with mean value zero for $n \neq m$.

Suppose that $k_1,\ldots,k_p$ are fixed. The term $Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}$ is independent of the order of the indices. Each term can be obtained in at most p! ways. In every non zero term each index occurs at least twice. Therefore

$$\sum_{j_1} \cdots \sum_{j_p} |Ef_{j_1} \ \ldots \ f_{j_p}| \leq p!\alpha_{n_k}^{p} \sum_{j_1}{}' \cdots \sum_{j_p}{}' \sum_{k_1} \cdots \sum_{k_p} \lambda^{k_1+\ldots+k_p} \cdot$$

$$\cdot \ |Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}|$$

where the prime denotes that the summation is restricted to terms for which each index j_i-k_i, $1 \leq i \leq p/2$, equals at least one of the indices j_n-k_n, $p/2+1 \leq n \leq p$, and vice versa.

Suppose now that $k_1,\ldots,k_p,j_1,\ldots,j_{p/2}$ are fixed in the sum above. Then the number of terms that are obtained as $j_{p/2+1},\ldots,j_p$ vary over $[n_{k-1},n_k]$ is less than or equal to $(p/2)^{p/2}$, and independent of n_k-n_{k-1}.

Since $|Ee_{j_1-k_1} \cdot \ldots \cdot e_{j_p-k_p}| \leq E|e|^p \leq C$ this gives

$$\sum_{j_1}\cdots\sum_{j_p}|Ef_{j_1}\cdot\ldots\cdot f_{j_p}| \leqslant \alpha^p_{n_k}\, p!\, (p/2)^{p/2} C \sum_{j_1}\cdots\sum_{j^{p/2}}\left(\sum_{k=0}^{\infty} \lambda^k\right)^p \leqslant$$

$$\leqslant \alpha^p_{n_k} C'(p,\lambda)(n_k-n_{k-1})^{p/2}$$

and from (B.5)

$$ET_k^p \leqslant \gamma^p_{n_k}\alpha^p_{n_k}(n_k-n_{k-1})^{p/2}C(\lambda) \leqslant C\gamma^{p/2}_{n_k}\alpha^p_{n_k}$$

□

<u>Lemma B.2</u>. Under the assumptions of the theorem

$$E|T_k|^p < B_1\alpha^p_{n_k}\gamma^{p/2}_{n_k}$$

implies

$$E|y_{n_k}|^p < B_2\alpha^p_{n_k}\gamma^{p/2}_{n_k} \quad \text{where } B_2 = B_1/(1-e^{t/4})^p \qquad (B.6)$$

<u>Remark</u>. This result holds for general $p > 0$.

<u>Proof of Lemma B.2</u>. Introduce

$$A_k = \prod_{i=n_{k-1}+1}^{n_k}(1-\gamma_i)$$

It follows that $e^{-2t} < A_k < e^{-t/2}$ for sufficiently large k. Then from (B.3)

$$y_{n_k} = T_k + A_k y_{n_{k-1}} \qquad (B.7)$$

Suppose that (B.6) is valid for k-1:

$$\left[E|y_{n_{k-1}}|^p\right]^{1/p} \leqslant B_2^{1/p}\gamma_{n_{k-1}}^{1/2}\alpha_{n_{k-1}}$$

Minkowski's inequality, applied to (B.7) then yields

$$\left[E|y_{n_k}|^p\right]^{1/p} \leqslant \left[E|T_k|^p\right]^{1/p} + A_k\left[E|y_{n_{k-1}}|^p\right]^{1/p}$$

Now

$$A_k\gamma_{n_{k-1}} = \beta_{n_{k-1}}^{n_k} \leqslant \beta_{n_k}^{n_k} = \gamma_{n_k}$$

since γ_k satisfies (4.8). Hence

$$\left[E|y_{n_k}|^p\right]^{1/p} \leqslant \left[B_1^{1/p}\gamma_{n_k}^{1/2} + A_k^{1/2}B_2^{1/p}\gamma_{n_k}^{1/2}\right]\alpha_{n_k} \leqslant$$

$$\leqslant \left[B_1^{1/p} + e^{t/4}B_2^{1/p}\right]\gamma_{n_k}^{1/2}\alpha_{n_k} = B_2^{1/p}\gamma_{n_k}^{1/2}\alpha_{n_k}$$

Consequently

$$E|y_{n_k}|^p \leqslant B_2\gamma_{n_k}^{p/2}\alpha_{n_k}^p \qquad k = 1,2,\ldots$$

since

$$E|y_{n_1}|^p = E|T_1|^p < B_1\gamma_{n_1}^{p/2}\alpha_{n_1}^p < B_2\gamma_{n_1}^{p/2}\alpha_{n_1}^p$$

□

To complete the proof of the theorem it remains only to be shown that (B.2) is valid not only for the subsequence n_k as in (B.6) but for all values of n. But for $n_{k-1} < j \leqslant \leqslant n_k$

$$y_j = \prod_{i=n_{k-1}+1}^{j} (1-\gamma_i) y_{n_{k-1}} + \prod_{i=j+1}^{n_k} (1-\gamma_i)^{-1} \sum_{i=n_{k-1}+1}^{j} \beta_i^{n_k} f_i =$$

$$= \prod_{i=j+1}^{n_k} (1-\gamma_i)^{-1} \left[A_k y_{n_{k-1}} + \sum_{n_{k-1}+1}^{j} \beta_i^{n_k} f_i \right]$$

As in (B.5)

$$E\left|\sum_{n_{k-1}+1}^{j} \beta_i^{n_k} f_i\right|^p \leqslant \gamma_{n_k}^p \sum_{j_1=n_{k-1}+1}^{j} \cdots \sum_{j_p=n_{k-1}+1}^{j} |Ef_{j_1} \cdots f_{j_p}| \leqslant$$

$$\leqslant C' \gamma_{n_k}^{p/2} \alpha_j^p$$

Application of Minkowski's inequality as in Lemma B.2 now gives

$$(E|y_j|^p)^{1/p} \leqslant \left[\prod_{n_{k-1}+1}^{n_k} (1-\gamma_i)^{-1}\right] B_2 \gamma_{n_k}^{1/2} \alpha_j$$

and hence

$$E|y_j|^p \leqslant e^{2pt} B_2 \gamma_j^{p/2} \alpha_j^p = C' \gamma_j^{p/2} \alpha_j^p$$

□

APPENDIX C.

Proof of Theorem 4.3.

Theorem 4.3. Consider the algorithm (4.3)

$$y_n = y_{n-1} + \gamma_n(f_n - y_{n-1})$$

where $\{f_n\}$ is a sequence of independent random variables. Suppose $\{\gamma_n\}$ satisfies (4.8). Let $E|f_n|^p \leq \alpha_n^p$ for some real $p > 1$, where $\{\alpha_n\}$ is a nondecreasing sequence and suppose that

$$\sum_{n=1}^{\infty} \gamma_n^{p'} \alpha_n^p < \infty \quad \text{where } p' = \min(p, 1+p/2)$$

Then

$$y_n \to 0 \quad \text{as} \quad n \to \infty \text{ w.p.1}$$

Proof. As in the proof of Theorem 4.1 we will introduce

$$T_k = \sum_{n_{k-1}+1}^{n_k} \beta_i^{n_k} f_i \; ; \qquad \sum_{n_{k-1}}^{n_k} \gamma_j \to t \quad \text{as } k \to \infty$$

and first obtain estimates of $E|T_k|^p$ where p is a real number > 1. Thus Lemma B.1 cannot be applied, since p is there assumed to be an even integer. Consider first $1 < p \leq 2$. In this case a theorem due to von Bahr-Esseen (1965) is applicable:

Suppose S_n is a sum of independent variables

$$S_n = \sum_1^n e_i$$

Then

$$E|S_n|^p \leqslant (2-1/n) \sum_1^n E|e_i|^p \qquad 1 < p \leqslant 2$$

In our case:

$$E|T_k|^p \leqslant 2 \sum_{j=n_{k-1}+1}^{n_k} \left(\beta_j^{n_k}\right)^p E|f_j|^p \leqslant K \sum_{j=n_{k-1}+1}^{n_k} \gamma_j^p \alpha_j^p$$

Clearly, using Chebysjev's inequality and the Borel Cantelli lemma

$\sum_{j=1}^{\infty} (\gamma_j \alpha_j)^p < \infty$ implies that $T_k \to 0$ as $k \to \infty$ w.p.1.

As follows from the lemma below, this implies that $y_n \to 0$ as $n \to \infty$ w.p.1.

Lemma C.1 (Petrov (1972)). Let $B_n \to \infty$ as $n \to \infty$, and suppose there exists a subsequence B_{n_k} and constants $c_1 > 1$ and c_2, such that

$$c_1 \leqslant \frac{B_{n_k}}{B_{n_{k+1}}} \leqslant c_2$$

for all sufficiently large k. Let S_n be a sum of independent variables with zero mean value. Let $T_k = (S_{n_k} - S_{n_{k-1}})B_{n_k}$.

Then

$B_n S_n \to 0$ w.p.1 as $n \to \infty$

if and only if

$T_k \to 0$ w.p.1 as $n \to \infty$ □

Introduce

$$\delta_n = \gamma_n \prod_{i=1}^{n} (1-\gamma_i)^{-1}$$

$$S_n = \sum_{1}^{n} \delta_k f_k$$

$$B_n = \prod_{1}^{n} (1-\gamma_i)$$

Then $y_n = B_n S_n$, $\beta_k^n = \delta_k B_n$ and

$$(S_{n_k} - S_{n_{k-1}}) B_{n_k} = \sum_{n_{k-1}+1}^{n_k} \delta_j B_{n_k} f_j = \sum_{n_{k-1}+1}^{n_k} \beta_j^{n_k} f_j = T_k$$

The proof of the assertion now follows from Lemma C.1 in case $1 < p \leqslant 2$.

Now, we turn to the case $p > 2$. The following result, due to Dharmadhikari and Jogdeo (1969) will be used:

Let $\{e_i\}$ be a sequence of independent zero mean valued random variables

$$S_n = \sum_1^n e_i$$

Then

$$E|S_n|^p \leqslant C_p n^{p/2-1} \sum_1^n E|e_i|^p \qquad p \geqslant 2$$

Applying this result to T_k the following estimates are obtained:

$$E|T_k|^p \leqslant C(n_k-n_{k-1})^{(p/2)-1} \sum_{n_{k-1}+1}^{n_k} (\beta_j^{n_k})^p \alpha_j^p E|f_j|^p \leqslant$$

$$\leqslant C_p(n_k-n_{k-1})^{p/2} \gamma_{n_k}^p \alpha_{n_k}^p F_p \leqslant K_p' \gamma_{n_k}^{p/2} \alpha_{n_k}^p$$

The second inequality follows from property (4.8). The third inequality follows from (B.4). From Lemma B.2 now follows that also

$$E|y_n|^p \leqslant K_p \gamma_n^{p/2} \alpha_n^p \qquad \text{(C.3)}$$

The estimate (C.3) will now be applied to the following lemma by Zolotarev (1965).

Lemma C.2 (Zolotarev (1965)). Let $e_1, \ldots, e_n$ be independent random variables with zero mean values. Let

$$0 \leqslant B_n \leqslant B_{n-1} \leqslant \ldots \leqslant B_1$$

Let $\varphi(x)$ be a function defined for $x \geqslant 0$, nondecreasing, convex and

$$\varphi(0) = 0 \qquad \varphi(xy) \leqslant \varphi(x)\varphi(y) \qquad x,y \geqslant 0$$

Let

$$S_n = \sum_1^n e_i$$

Then

$$P(\max_{m\leqslant k\leqslant n} B_k S_k \geqslant x) \leqslant \frac{1}{\varphi(x)}\Big\{\varphi(B_n)E^+\varphi(S_n) + \\ + \sum_{k=m}^{n-1}\big(\varphi(B_k) - \varphi(B_{k+1})\big)E^+\varphi(S_k)\Big\}$$

for any x.

Here

$$E^+\varphi(x) = \int_0^\infty \varphi(x)dF(x)$$

where F(x) is the distribution function for x.

□

Take now in Lemma C.2.

$$B_n = \prod_1^n (1-\gamma_j); \qquad \varphi(x) = |x|^p \qquad S_n = \sum_1^n \delta_k f_k$$

where

$$\delta_k = \gamma_k \prod_{i=1}^{k} (1-\gamma_i)^{-1}$$

Then $y_n = c_n S_s$ and

$$P(\max_{m\leqslant k\leqslant n} |y_k|>\varepsilon) \leqslant \frac{1}{|\varepsilon|^p}\left\{E|y_n|^p + \sum_{k=m}^{n-1} (B_k^p - B_{k+1}^p) E|S_k|^p\right\}$$

Now

$$B_k^p - B_{k+1}^p = B_k^p[1 -(1-\gamma_{k+1})^p] \leqslant p\gamma_{k+1}C_k^p \text{ since } 0 \leqslant \gamma_{k+1} \leqslant 1$$

Choose a subsequence n_k such that

$$\sum_1^\infty \gamma_{n_k}^{p/2}\alpha_{n_k}^p < \infty$$

This is possible since

$$\sum_1^\infty \gamma_n^{p'}\alpha_n^p < \infty \Rightarrow \gamma_n\alpha_n \to 0$$

Then

$$\sum_{k=1}^\infty P(\max_{n_k\leqslant j\leqslant n_{k+1}} |y_j|>\varepsilon) \leqslant \frac{1}{|\varepsilon|^p}\left\{\sum_{k=1}^\infty E|y_{n_k}|^p + \sum_1^\infty p\gamma_k E|y_k|^p\right\} \leqslant$$

$$\leqslant C\left\{\sum_1^\infty \gamma_{n_k}^{p/2}\alpha_{n_k}^p + \sum_1^\infty \gamma_k^{p/2+1}\alpha_k^p\right\} < \infty$$

and consequently $y_n \to 0$ w.p.1 as $n \to \infty$.

This completes the proof of the theorem.

□

APPENDIX D

Proof of Theorem 6.1.

Lemma A.1 states that the sequence $\{x_n\}$ locally and asymptotically follows the trajectories of (6.17). This result will be extended to global estimates by linking a chain of local estimates (A.1).

Order the set of indices $I = \{n_i\}$ such that $n_1 < n_2 < \ldots < < n_k < n_{k+1} < \ldots$ Denote $\Delta\tau_k = \tau_{n_{k+1}} - \tau_{n_k}$. Then, taking $\bar{x} = x_n$ in (A.1) yields

$$x_{n_{k+1}} = x_{n_k} + \Delta\tau_k f(x_{n_k}) + A\Delta\tau_k^2 + R_2(\Delta\tau_k, n_k, x_{n_k}) \qquad \text{(D.1)}$$

By going through the details of the proof of Lemma A.1 it is found that

$$R_2(\Delta\tau_k, n_k, x_{n_k}) = \sum_{i=n_k+1}^{n_{k+1}} \gamma_i \Big\{ [2Q(x_{n_k}, e_i) - K(e_i)] - $$

$$- E[2Q(x_{n_k}, e_i) - K(e_i)]\Big\}$$

Applying Lemma 4.2 gives

$$E|R_2|^{2p} \leq L_1 \gamma_{n_k}^p$$

and so from Chebysjev's inequality

$$P(|R_2(\Delta\tau_k, n_k, x_{n_k})| > \varepsilon_1) \leq L_1 \gamma_{n_k}^p / \varepsilon_1^{2p} \qquad \text{(D.2)}$$

Also,

$$\tilde{x}(\tau_{n_k+1};\tau_{n_k},x_{n_k}) = x_{n_k} + \Delta\tau_k f(x_{n_k}) + L_2\Delta\tau_k^2 \qquad (D.3)$$

Combine (D.1) and (D.3):

$$|x_{n_{k+1}} - \tilde{x}(\tau_{n_{k+1}};\tau_{n_k},x_{n_k})| \leqslant (A+L_2)\Delta\tau_k^2 + R_2(\Delta\tau_k,n_k,x_k) \quad (D.4)$$

Define $L_3 = A + L_2$.

According to the assumptions of the theorem there exists a function $V(\Delta x,\tau)$ which is quadratic in Δx and such that

$$C_1|\Delta x|^2 \leqslant V(\Delta x,\tau) \leqslant C_2|\Delta x|^2 \qquad (D.5)$$

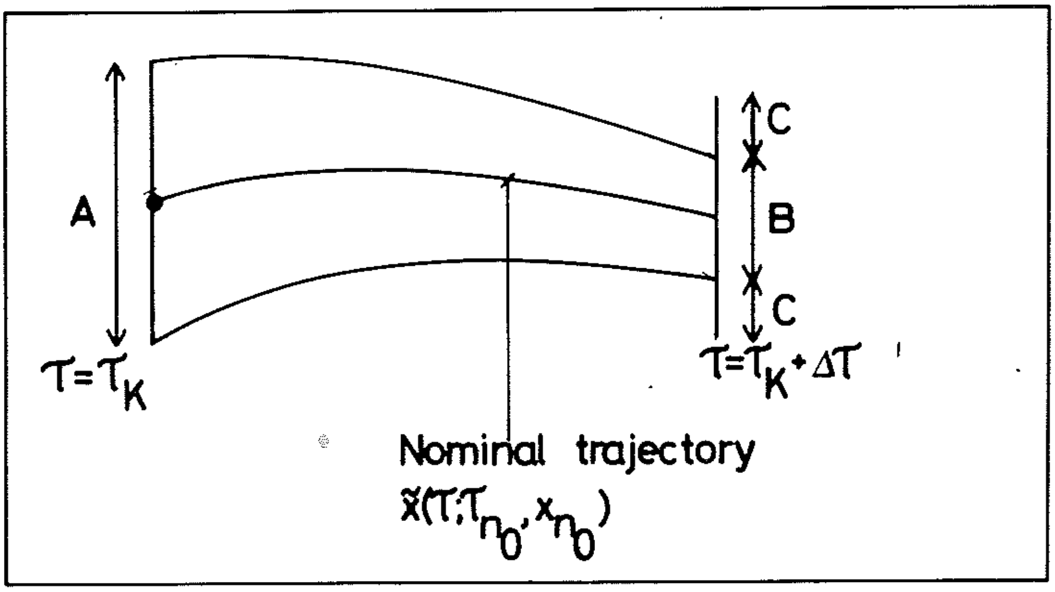
and

$$\frac{d}{d\tau} V(\Delta x,\tau) < - \lambda|\Delta x|^2 \; ; \quad \lambda > 0 \qquad (D.6)$$

along solutions of the variational equation.

We will now give an outline and a heuristic interpretation of the rest of the proof before we proceed to the formal treatment.

The idea of the proof can geometrically be expressed as follows:



Assume that the estimate at time τ_k is in the interval A. The trajectories that start in A belong at time $\tau_k + \Delta\tau_k$ to the interval B. The length of B is given by (D.6). If $V(\Delta x,\tau) = |\Delta x|^2$, then $B \leqslant (1-\lambda\Delta\tau_k)A$. Now, the estimates obtained by the algorithm differ from the trajectories with less than $L_3\Delta\tau_k^2 + R_2(\Delta\tau_k,n_k,x_{n_k})$ according to (D.4). Denote this distance by C. During the time interval $\Delta\tau_k$, the estimates have not diverged from the nominal trajectory if $A \leqslant B + 2C$, i.e. if

$$A \leqslant (1-\lambda\Delta\tau_k)A + L_3\Delta\tau_k^2 + R_2(\Delta\tau_k,n_k,x_{n_k})$$

or

$$\lambda\Delta\tau_k A \leqslant L_3\Delta\tau_k^2 + R_2(\Delta\tau_k,n_k,x_{n_k})$$

To achieve this, A and $\Delta\tau_k$ must be chosen with care. The interval $\Delta\tau_k$ must be large enough to let the trajectories converge sufficiently, and small enough to limit second order effects and the noise influence.

We now turn to the formal proof.

Select first ε (corresponding to A in the discussion above) such that

$$\varepsilon < \frac{8DL_3C_1C_2}{\lambda} = \varepsilon_0$$

Since $\Delta\tau_k > D$, it follows that

$$\frac{\lambda\varepsilon}{C_1C_2L_3 \cdot 8} < \Delta\tau_k \qquad \text{all } k.$$

Possibly by extending the set I, it is thus possible to obtain

$$\frac{\lambda\varepsilon}{C_1C_2L_3 \cdot 8} < \Delta\tau_k < \frac{3\lambda\varepsilon}{C_1C_2L_3 \cdot 8} \tag{D.7}$$

Now suppose that

$$|R_2(\Delta\tau_k, n_k, x_{n_k})| < \frac{\lambda^2\varepsilon^2 3}{L_3C_2^2C_1^2 \cdot 64}$$

and

$$V^{1/2}[(\tilde{x}(\tau_{n_k};\tau_{n_0},x_{n_0}) - x_{n_k}),\tau_{n_k}] \leqslant \varepsilon$$

Then

$$V^{1/2}[(\tilde{x}(\tau_{n_{k+1}};\tau_{n_0},x_{n_0}) - x_{n_{k+1}}),\tau_{n_{k+1}}] \leqslant$$

$$\leqslant V^{1/2}[\{\tilde{x}(\tau_{n_{k+1}};\tau_{n_0},x_{n_0}) - \tilde{x}(\tau_{n_{k+1}};\tau_{n_k},x_{n_k})\},\tau_{n_{k+1}}] +$$

$$+ V^{1/2}\Big[[\tilde{x}(\tau_{n_{k+1}};\tau_{n_k},x_{n_k}) - x_{n_{k+1}}],\tau_{n_{k+1}}\Big] \leqslant$$

$$\leqslant \left(1 - \frac{\lambda}{2C_1}\Delta\tau_k\right)V^{1/2}[(\tilde{x}(\tau_{n_k};\tau_{n_0},x_{n_0}) - x_{n_k}),\tau_{n_k}] +$$

$$+ C_2|\tilde{x}(\tau_{n_{k+1}};\tau_{n_k},x_{n_k}) - x_{n_{k+1}}| \leqslant$$

$$\leqslant \left(1 - \frac{\lambda}{2C_1}\Delta\tau_k\right)\varepsilon + C_2L_3\Delta\tau_k^2 + C_2|R_2(\Delta\tau_k,n_k,x_{n_k})| \leqslant$$

$$\leqslant \varepsilon + C_2L_3\left\{\left(\Delta\tau_k - \frac{\lambda\varepsilon}{8C_1C_2L_3}\right)\left(\Delta\tau_k - \frac{3\lambda\varepsilon}{8C_1C_2L_3}\right)\right\} \leqslant \varepsilon$$

The first inequality follows since $V^{1/2}$ is a norm, the second follows from the properties of V. The third and fourth inequalities follow from the assumptions made just above. The last inequality follows from (D.7). In other words, if

$$V^{1/2}[(\tilde{x}(\tau_{n_k};\tau_{n_0},x_{n_0}) - x_{n_k}),\tau_{n_k}] \leqslant \varepsilon$$

then

$$V^{1/2}[(\tilde{x}(\tau_{n_{k+1}};\tau_{n_0},x_{n_0}) - x_{n_{k+1}}),\tau_{n_{k+1}}] \leqslant \varepsilon$$

with probability at least

$$P\left\{|R_2(\Delta\tau_k, n_k, x_{n_k})| < \frac{3\lambda^2\varepsilon^2}{64C_1^2C_2^2L_3}\right\} >$$

$$> 1 - L_1\left(\frac{64L_3C_2^2C_1^2}{3\lambda^2}\right)^{2p} \gamma_{n_k}^p/\varepsilon^{4p}$$

Now the event

$$\Omega = \{\sup_{n\in I}|x_n - \tilde{x}(\tau_n;\tau_{n_0},x_{n_0})| > \varepsilon\} \subset$$

$$\subset \{\sup_{n\in I} V^{1/2}[(x_n - \tilde{x}(\tau_n;\tau_{n_0},x_{n_0})),\tau_n] > \varepsilon C_1\} \subset \bigcap_1^N \Omega_k$$

where

$$\Omega_k = \Big\{V^{1/2}[(x_{n_j} - \tilde{x}(\tau_{n_j};\tau_{n_0},x_0)),\tau_{n_j}] \leq \varepsilon C_1 \quad j \leqslant k;$$

$$V^{1/2}[(x_{n_{k+1}} - \tilde{x}(\tau_{n_{k+1}},\tau_{n_0},x_0)),\tau_{n_{k+1}}] > \varepsilon C_1\Big\}$$

and thus

$$P(\Omega) \leqslant \sum_1^N P(\Omega_k) \leqslant \left[\frac{64C_2^2C_1L_3}{3\lambda^2}\right]^{2p} \frac{L_1}{\varepsilon^{4p}} \sum_{j=n_0}^N \gamma_j^p$$

and the theorem is proved.

□