A NEW DYNAMIC STOCHASTIC APPROXIMATION PROCEDURE

Dissertation for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY
DAVID RUPPERT

1977

This is to certify that the

thesis entitled

A New Dynamic Stochastic Approximation Procedure

presented by

David Ruppert

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in *Statistics*

Vaclav Fabian
Major professor

Date May 18, 1977

ABSTRACT

A NEW DYNAMIC STOCHASTIC APPROXIMATION PROCEDURE

By

David Ruppert

This paper considers Robbins-Monro stochastic approximation when the regression function changes with time. At time n, one can select X_n and observe an unbiased estimator of the regression function evaluated at X_n. Let θ_n be the root of the regression function at time n. Our goal is to select the sequence X_n so that $X_n - \theta_n$ converges to 0. It is assumed that $\theta_n = f(s_n)$ for s_n known at time n and f an unknown element of a class of functions. Under certain conditions on this class and on the sequence of regression functions, we obtain a random sequence X_n such that $|X_n - \theta_n|$ converges to 0 in Cesaro mean with probability 1. Under more stringent conditions, $X_n - \theta_n$ converges to 0 with probability 1.

A NEW DYNAMIC STOCHASTIC APPROXIMATION PROCEDURE

By

David Ruppert

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1977

To

Patricia

ACKNOWLEDGEMENTS

I wish to thank Professor Václav Fabian for having suggested this area of dynamic stochastic approximation to me, for giving guidance to the research as it progressed, and finally for his valuable criticism of earlier drafts of this thesis.

Also, I want to thank Professor Jan Mařík for a suggestion which led to theorem 3.11.

I believe that each member of the faculty of the Department of Statistics and Probability has helped me in some way during my four years of graduate study at Michigan State University and to them I am grateful.

Mrs. Noralee Burkhardt deserves thanks for typing my thesis excellently and for her cheerfulness and the patience while correcting my errors.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

This study has been motivated by practical situations in which a process is controlled by a variable X and it is desirable to choose X in such a manner that the response, $R_n(X)$, at time n is close to 0. If θ_n satisfies $R_n(\theta_n) = 0$, it would be enough to choose X_n, the value of X at time n, equal or close to θ_n. The basic information is provided by the process itself; for any choice of X_n we can obtain an unbiased estimate of $R_n(X_n)$.

If R_n, or at least θ_n, is independent of n and some regularity conditions are satisfied, then the stochastic approximation procedure of Robbins and Monro (1951) provides a method of selecting a sequence $\{X_n\}$ such that $X_n \to \theta_1$ almost surely.

We are concerned here with situations where θ_n does change with n. Dupač (1965, 1966) and Uosaki (1974) studied such situations, but their model is substantially different from ours; both models shall be compared later.

In our model we assume that $\theta_n = f(s_n)$ for an f in a family $\mathcal{S}$ of functions on a set S and for a sequence $\{s_n\}$ in S. Initially, only $\mathcal{S}$ is known, not f and not $\{s_n\}$. At time n, the value s_n becomes known, and, after X_n is selected, an unbiased estimate of $R_n(X_n)$ is observed.

The interpretation is that s_n summarizes the knowledge about the process at time n. For example, in the case of a process involving a chemical reactor, s_n can describe the age of the filter, the quality of the catalyzer, and the impurities of the input. In another example we may have $s_n = n$ and then the assumption concerning θ_n means simply that the function $n \rightsquigarrow \theta_n$ is in S.

We propose an approximation method, for which $X_n - \theta_n$ approaches 0 in a certain sense, for some families S. For example, S can be the family of all functions f on $[0,1]$ such that, for some K and $\alpha > \frac{1}{2}$, depending on f,

$$|f(x) - f(y)| \leq K|x - y|^{\alpha}$$

for all x,y in $[0,1]$ (cf. Theorem 3.9).

Another example, admittedly simpler, yet of considerable practical importance, is the case when S is the family of all linear combinations of k functions $f_1,f_2,\ldots,f_k$.

Both these examples are special cases of the more general condition (see assumption 2.3) that there exists an inner product space H and a function U on the set S into H such that $S \subset \{f_\beta;\ \beta \in H\}$ where f_β denotes the function defined on S and assigning to each s in S the value $\langle\beta, U(s)\rangle$.

Under certain additional regularity conditions we shall show that the proposed approximation procedure yields $\{X_n\}$ for which $X_n - \theta_n \to 0$ in Cesaro mean with probability one; under more stringent conditions $X_n - \theta_n \to 0$ with probability one.

Dupač (1965, 1966) considered Robbins-Monro type stochastic approximation methods when the root changes during the approximation process and Uosaki (1974) generalized his work. In these papers, the basic assumption is that θ_{n+1} is equal to $g_n(\theta_n)$, with g_n known, plus an unknown but small v_n. The procedure then is similar to the original Robbins-Monro procedure except that where the latter obtains the estimate X_{n+1} by adjusting X_n, the former adjusts $g_{n+1}(X_n)$ (and neglects v_n).

In our model, the procedure estimates the function f_β by estimating β. If H is infinite dimensional the procedure allows us to keep the estimates finite-dimensional in order that the procedure can be practically realizable.

In addition to the above problems we also consider, in theorem 4.5, the case where $U(s_n)$ is a random variable with values in R^k.

In summary we will show that under conditions similar to those used to prove the convergence of the Robbins-Monro method, $X_n - \theta_n \to 0$ in Cesaro mean with probability one, where θ_n is the unique root of $R_n(X) = 0$ and X_n is our estimate of θ_n. Of practical importance are similar generalizations of the Kiefer-Wolfowitz (1952) method of maximization (or minimization) of functions on R and Blum's (1954) multi-dimensional version of the Kiefer-Wolfowitz method. One can expect that the methods obtained by such generalizations would have a property analogous to the almost sure Cesaro mean convergence of $X_n - \theta_n$ to 0.

CHAPTER 2

NOTATION AND ASSUMPTIONS

2.1 Notation. The conventions introduced here hold throughout. Let R^k be k-dimensional Euclidean space. The space R^1 will be denoted simply as R. Denote the transpose of the matrix A by A^T. Then the inner product on R^k is defined by

$$<x,y> = x^T y \quad \text{for} \quad x,y \in R^k.$$

If A and B are sets, then A^B is the set of all functions from B to A.

Let (Ω,F,P) be a probability space. If $F \in F$, then I_F is the indicator of F.

If V is a normed vector space, then let $\underline{V}$ be the smallest σ-algebra containing all open balls, that is all sets of the form

$$\{X \in V: \|X + a\| < \varepsilon\} \quad \text{for} \quad \varepsilon > 0 \quad \text{and} \quad a \in V .$$

A map T in V^Ω is called a measurable transformation into V if $T^{-1}(\underline{V}) \subset F$. If $T_1,\ldots,T_n$ are measurable transformations into V, then $\sigma\{T_1,\ldots,T_n\}$ is the smallest σ-algebra on Ω containing

$$\bigcup_{i=1}^{n} T_i^{-1}(\underline{V}) .$$

All relations between measurable transformations are meant to hold with probability one.

If h_n is a sequence of numbers, then $O(h_n)$ denotes a sequence g_n of numbers such that for some K

$$|h_n^{-1} g_n| \leq K \quad \text{for all} \quad n .$$

<u>2.2 Assumption</u>. (i) Let S be a set and suppose $\mathcal{S} \subset R^S$. Suppose $f \in \mathcal{S}$.

(ii) Let $R_n \in R^R$, $\theta_n \in R$, and $A > 0$. Suppose

(1) $$(X - \theta_n) R_n(X) \geq 0$$

and

(2) $$|R_n(X)| \leq A(|X - \theta_n| + 1)$$

for all $X \in R$. Let $s_n \in S$ and suppose

(3) $$\theta_n = f(s_n).$$

<u>2.3 Assumption</u>. (i) Assumption 2.2(i) holds. Let H be a real vector space and suppose $\langle\cdot,\cdot\rangle$ is an inner product on H, i.e. $\langle\cdot,\cdot\rangle$ is a map from $H \times H$ to R such that if $x,y,z \in H$ and $a \in R$ then

$$\langle ax + y,z\rangle = a\langle x,z\rangle + \langle y,z\rangle,$$

$$\langle x,y\rangle = \langle y,x\rangle,$$

$$\langle x,x\rangle \geq 0,$$

and

$$\langle x,x\rangle = 0 \text{ implies } x = 0 .$$

For $x \in H$ define $\|x\| = \langle x,x\rangle^{\frac{1}{2}}$. Suppose there is a function U in H^S such that for each f in S there exists a β in H satisfying

$$f(s) = \langle\beta, U(s)\rangle \text{ for all } s \in S.$$

(ii) Assumption 2.2(ii) holds. Let $U_n = U(s_n)$.

2.4 Remark. We shall now consider the problem of estimating the sequence $\{\theta_n\}$. The experimenter knows H, $\langle\cdot,\cdot\rangle$, and U and he knows that assumption 2.3 holds. At time n he estimates β by an estimate β_n. Also at this time he learns the value of s_n and therefore of U_n; he uses U_n to estimate θ_n by $X_n = \langle\beta_n, U_n\rangle$. He can also observe a random variable Y_n, an unbiased (conditionally, given the past) estimator of $R_n(X_n)$. He then forms his next estimate

$$\beta_{n+1} = \beta_n - a_n Y_n U_n^*$$

with a_n a suitably chosen non-negative number and U_n^* either equal to U_n or a suitable approximation to U_n. For example, if $H = \ell_2$, then the experimenter may wish to use a finite dimensional approximation, U_n^*, to a U_n in ℓ_2.

We shall reformulate the construction of the β_n in the following assumption, where F_n is the σ-algebra associated with the "past" at time n.

2.5 Assumption. (i) Assumption 2.3 holds.

(ii) Let F_n be an increasing sequence of σ-algebras contained in F. Suppose $\{\beta_n\}$, $\{Y_n\}$, and $\{U_n^*\}$ are sequences of measurable transformations into H, random variables, and elements of H, respectively, such that with $X_n = \langle \beta_n, U_n \rangle$

(1) $\quad \beta_{n+1} = \beta_n - a_n Y_n U_n^*$ for some $a_n \geq 0$,

$\sigma(\beta_1, \ldots, \beta_n) \subset F_n$, and

(2) $E^{F_n} Y_n = R_n(X_n)$ and $E^{F_n}(Y_n - R_n(X_n))^2 \leq \sigma^2$ for some σ^2.

(iii) Assume that

(3) $$\langle \beta_n, U_n - U_n^* \rangle = 0 .$$

<u>2.6 Remark</u>. Suppose we wish to choose U_n^* not equal to U_n. Then (2.5.3) will still hold if for an increasing sequence of subspaces, $\{H_n\}$, $\beta_1 \in H_1$ and U_n^* is the projection of U_n onto H_{n+1}, for then by (2.5.1) $\beta_n \in H_n$ for all n.

CHAPTER 3

GENERAL RESULTS

We will be interested in the convergence to 0 of the sequences $\{\|\beta_n - \beta\|\}$ and $\{X_n - \theta_n\}$ defined in assumptions 2.2, 2.3, and 2.5. These two sequences are closely connected for under these assumptions $X_n - \theta_n = \langle\beta_n - \beta, U_n\rangle$. For practical purposes $\{X_n - \theta_n\}$ is of primary importance since X_n would be the value of the control variable at time n while θ_n would be a value of the control variable which for our intentions would be optimal at time n.

3.1 Lemma. Suppose assumption 2.5 holds and

(1) $$\Sigma\, a_n(1 + \|U_n\|)\,|\langle\beta, U_n - U_n^*\rangle| < \infty$$

and

(2) $$\Sigma\, a_n^2\|U_n^*\|^2(1 + \|U_n\|^2) < \infty .$$

Then,

(3) $$\|\beta_n - \beta\|^2 \text{ has a finite limit}$$

and

(4) $$\Sigma\, a_n R_n(X_n)(X_n - \theta_n) < \infty .$$

Proof: By (2.5.1) and (2.5.2)

(5) $$E^{F_n}\|\beta_{n+1} - \beta\|^2 \leq \|\beta_n - \beta\|^2 - 2a_nR_n(X_n)\langle\beta_n - \beta, U_n^*\rangle + a_n^2\|U_n^*\|^2(R_n^2(X_n) + \sigma^2)\ .$$

Now by (2.2.2),

(6) $$|R_n(X_n)| \leq A(|\langle\beta_n - \beta, U_n\rangle| + 1) \leq A(\|\beta_n - \beta\|\ \|U_n\| + 1).$$

Also by (2.5.3)

$$\langle\beta_n - \beta, U_n^*\rangle = (X_n - \theta_n) - \langle\beta, U_n^* - U_n\rangle.$$

Therefore

(7) $$R_n(X_n)\langle\beta_n - \beta\ , U_n^*\rangle \geq R_n(X_n)(X_n - \theta_n) - |\langle\beta, U_n^* - U_n\rangle| A((\|\beta_n - \beta\|^2 + 1)\|U_n\| + 1).$$

Substituting (6) and (7) into (5), one obtains

(8) $$E^{F_n}\|\beta_{n+1} - \beta\|^2 \leq \|\beta_n - \beta\|^2(1 + f_n) - 2a_nR_n(X_n)(X_n - \theta_n) + g_n$$

where $f_n = O(a_n^2\|U_n\|^2\|U_n^*\|^2 + a_n|\langle\beta, U_n - U_n^*\rangle|\,\|U_n\|)$ and $g_n = O(a_n^2\|U_n^*\| + a_n|\langle\beta, U_n - U_n^*\rangle|(1 + \|U_n\|))$ and $f_n, g_n \geq 0$.

By (1) and (2), $\Sigma\ f_n + g_n < \infty$. Thus (3) and (4) hold by Theorem 1 of Robbins and Siegmund (1971).

3.2 Remark. Condition (3.1.1) involves β which, of course, is unknown. However β depends on f and f is known to be in the class S. Thus it may be possible to verify condition (3.1.1) by using properties of S.

3.3 Theorem. Let assumption 2.5 hold. Let α, γ, and ε be numbers satisfying

$$\gamma \geq 0,$$

$$\tfrac{1}{2} + 2\gamma < \alpha \leq 1,$$

and

$$\alpha + \varepsilon > 1.$$

Suppose for $a > 0$,

$$a_n = an^{-\alpha},$$

$$\|U_n\| + \|U_n^*\| = O(n^{\gamma}),$$

and

$$(1 + \|U_n\|)|\langle\beta, U_n - U_n^*\rangle| = O(n^{-\varepsilon}).$$

If for all $\eta > 0$

(9)
$$\inf_n \inf_{\eta \leq |X-\theta_n|} |R_n(X)| > 0$$

or if $\gamma = 0$ and for all $\eta > 0$

(10)
$$\inf_n \inf_{\eta \leq |X-\theta_n| \leq \eta^{-1}} |R_n(X)| > 0$$

then

$$n^{-1} \sum_{k=1}^{n} |X_k - \theta_k| \to 0.$$

Proof: First, (3.1.1) holds since

$$a_n(1 + \|U_n\|)|\langle\beta, U_n - U_n^*\rangle| = O(n^{-(\alpha+\varepsilon)})$$

and $\alpha + \varepsilon > 1$. Next (3.1.2) holds for

$$a_n^2 \|U_n^*\|^2 (1 + \|U_n\|^2) = O(n^{-(2\alpha-4\gamma)})$$

and $2\alpha - 4\gamma > 1$. Thus by lemma 3.1, $\Sigma\, a_n R_n(X_n)(X_n - \theta_n) < \infty$ and $\lim \|\beta_n - \beta\|$ exists and is finite. From now until the end of the proof we look at an ω for which the two properties hold and write ξ instead of $\xi(\omega)$ for any random variable ξ. For every $\eta > 0$ there is a $\delta(\eta) > 0$ such that

(11) $\quad |X_n - \theta_n| \geq \eta$ implies $|R_n(X_n)| > \delta(\eta)$.

This follows directly from (9); if (10) holds and $\gamma = 0$ then $|X_n - \theta_n| \leq \|U_n\|\, \|\beta_n - \beta\|$ is a bounded sequence and (11) holds again. Let $\eta > 0$, set $I_n = 1$ if $|X_n - \theta_n| \geq \eta$ and 0 otherwise. Then the finiteness of $\Sigma\, a_n R_n(X_n)(X_n - \theta_n)$ and (11) imply

$$\Sigma\, n^{-\alpha} I_n |X_n - \theta_n| < \infty.$$

By Kronecker's lemma (see Loève (1963), p. 238)

$$n^{-\alpha} \sum_{k=1}^{n} |X_k - \theta_k| I_k \to 0.$$

Since $\alpha \leq 1$,

$$\limsup_{n\to\infty} n^{-1} \sum_{k=1}^{n} |X_n - \theta_n| \leq \eta + \limsup_{n\to\infty} n^{-1} \sum_{k=1}^{n} |X_n - \theta_n| I_n = \eta$$

for all $\eta > 0$.

<u>3.4 Remarks</u>. The conclusion $n^{-1} \sum_{1}^{n} |X_k - \theta_k| \to 0$ is of practical importance since if $n^{-1} \sum_{1}^{n} |X_k - \theta_k|$ is small then the process would have run at near optimal conditions for most of the first n runs.

Without additional assumptions, the conclusion of theorem 3.3 cannot be strengthened to $(X_n - \theta_n) \to 0$, as can be seen in example 4.7 below. Moreover, example 4.8 below shows that under the hypotheses of theorem 3.3 $(\beta_n - \beta) \to 0$ may fail even if $(X_n - \theta_n) \to 0$.

Since U_n^* is intended to be an approximation to U_n we can expect that $\|U_n^*\| = O(\|U_n\|)$ and in that case the condition

(1) $$\|U_n\| + \|U_n^*\| = O(n^\gamma)$$

would be known to hold with $\gamma = 0$ if U is bounded. If U is unbounded then it might be difficult to verify that (1) holds; however, the theorem has been formulated to allow $\gamma > 0$.

<u>3.5 Assumption</u>. Let D be a countable set. Define the real vector space ℓ_D^2 and the inner product $\langle\cdot,\cdot\rangle_D$ on ℓ_D^2 by

$$\ell_D^2 = \{g \in R^D: \sum_{d\in D} g^2(d) < \infty\}$$

and

$$\langle g,h\rangle_D = \sum_{d\in D} g(d)h(d) \quad \text{for} \quad g,h \in \ell_D^2 .$$

For $f \in \ell_D^2$ define $\|f\|_D = \langle f,f\rangle_D^{1/2}$.

Let $\{D_n\}$ be a sequence of finite subsets of D with $D_n \subset D_{n+1}$.

Suppose assumption 2.2(i) holds. Let U be a map from S to ℓ_D^2, let $U_n = U(s_n)$, and define U_n^* by

$$\begin{aligned} U_n^*(d) &= U_n(d) \quad \text{if} \quad d \in D_{n+1} \\ &= 0 \quad\quad\;\; \text{if} \quad d \notin D_{n+1}. \end{aligned}$$

Suppose assumption 2.5(ii) holds with $a_n = an^{-\alpha}$ for some $a > 0$ and $\alpha > \frac{1}{2}$. Suppose

$$\beta_1(d) = 0 \quad \text{if} \quad d \notin D_1.$$

3.6 Remark. If assumption 3.5 holds, then it can be easily shown by induction (see Remark 2.6) that assumption 2.5(iii) holds.

3.7 Assumption. Assumption 3.5 holds with the following choices of S, $\mathcal{S}$, D, D_n, and U.

$$S = [0,1].$$

$$\mathcal{S} = \{f \in R^{[0,1]}: \text{ for some } K > 0 \text{ and } \gamma > \tfrac{1}{2}, \; |f(x) - f(y)| \le K|x - y|^{\gamma} \text{ whenever } x,y \in [0,1]\}.$$

$$D = \{(k,m): \; m = k = 0 \text{ or } m \text{ and } k \text{ are integers satisfying } m \ge 0 \text{ and } 1 \le k \le 2^m\}.$$

$$D_n = \{(k,m): \; (k,m) \in D \text{ and } 2^m \le n\}.$$

$$U(x)(k,m) = 1 \quad \text{if} \quad (k,m) = (0,0).$$

For $(k,m) \neq (0,0)$,

$$\begin{aligned} U(x)(k,m) &= (m+1)^{-1} \quad \text{if} \quad x \in \left(\frac{k-1}{2^m}, \frac{k-\frac{1}{2}}{2^m}\right) \\ &= -(m+1)^{-1} \quad \text{if} \quad x \in \left(\frac{k-\frac{1}{2}}{2^m}, \frac{k}{2^m}\right) \\ &= 0 \quad \text{if} \quad x \in \left(\frac{\ell-1}{2^m}, \frac{\ell}{2^m}\right) \text{ with } \ell \neq k \text{ and } 1 \le \ell \le 2^m. \end{aligned}$$

As a function of x, $U(x)(k,m)$ is continuous at 0 and 1 and at points of discontinuity it equals the arithmetic mean of its left and right limits.

3.8 Remark. Note that $U(\cdot)(k,m)$ is a multiple of the Haar function with indices k and m as defined by Alexits (1961), page 46.

3.9 Theorem. Let assumption 3.7 hold. Then,

$$n^{-1} \sum_{k=1}^{n} |X_k - \theta_k| \to 0 .$$

Proof: We need only show that the hypotheses of theorem 3.3 hold. First we will show that assumption 2.3 holds with $H = \ell_D^2$. Let $\beta \in R^D$ be defined by

$$\beta(k,m) = 2^m(m+1)^2 \int_0^1 f(x)U(x)(k,m)dx$$

for $(k,m) \in D$. By the definition of S we can and shall choose a $\xi > \frac{1}{2}$ such that $|f(x) - f(y)| \leq K|x - y|^{\xi}$ for some K and all $x,y \in [0,1]$. Then,

$$|\beta(k,m)| = (m + 1)2^m|\int_0^{2^{-(m+1)}} f(2^{-m}(k - \tfrac{1}{2}) - x)$$

$$- f(2^{-m}(k - \tfrac{1}{2}) + x)dx|$$

$$\leq k(m+1)2^{m+\xi}\int_0^{2^{-(m+1)}} X^{\xi}dx = O((m+1)2^{-\xi m}) .$$

Therefore $\beta \in \ell_D^2$ since

$$\sum_{m=0}^{\infty} \sum_{k=1}^{2^m} (\beta(k,m))^2 = \sum_{m=0}^{\infty} O((m+1)^2)2^{m(1-2\xi)})$$

and $1 - 2\xi < 0$.

Now $f(x) = \langle\beta,U(x)\rangle_D$ for $x \in [0,1]$ by Alexits(1961), theorem 1.6.2. Thus assumption 2.3 holds; therefore assumption 2.5 holds.

For $x \in [0,1]$, $\sum_{k=1}^{2^m} (U(x)(k,m))^2 \leq (m+1)^{-2}$. Thus $\sup_x \|U(x)\|_D^2 \leq 1 + \sum_{m=1}^{\infty} m^{-2} < \infty$ and therefore $\|U_n\|_D + \|U_n^*\|_D = O(1)$.

Finally $\langle \beta, U_n - U_n^* \rangle_D = O(n^{-\xi})$ by Alexits (1961), 4.6.1 Therefore the hypotheses of theorem 3.3 are satisfied with $\gamma = 0$ and $\varepsilon = \xi$.

<u>3.10 Assumption</u>. Assumption 3.5 holds with the following choices of S, $\mathcal{S}$, D, D_n, and U.

$$S = [0, 2\pi],$$

$$K = \{g \in R^{[0,2\pi]}: \ g \text{ is Lebesgue measurable,}$$

$$\int_0^{2\pi} g(\mu)d\mu = 0 \text{ and } \int_0^{2\pi} (g(\mu))^2 d\mu < \infty\},$$

and

$$\mathcal{S} = \{h \in R^{[0,2\pi]}: \ h(x) = \int_0^x h'(\mu)d\mu + c \text{ where } c \in R \text{ and}$$
$$h' \in K\}.$$

$$D = \{(1,0)\} \cup \{(i,k): \ i = 1,2 \text{ and } k \geq 1\},$$

$$D_n = \{(i,k) \in D: \ k \leq n\} \text{ for } n \geq 0,$$

and

$$\begin{aligned} U(x)(i,k) &= 1 && k = 0 \\ &= k^{-1} \cos kx && i = 1 \text{ and } k \geq 1 \\ &= k^{-1} \sin kx && i = 2 \text{ and } k \geq 1. \end{aligned}$$

<u>3.11 Theorem</u>. Let assumption 3.10 hold. Then

$$n^{-1} \sum_{k=1}^{n} |X_k - \theta_k| \to 0 .$$

<u>Proof</u>: Define $\beta \in R^D$ by

$$\beta(1,0) = \frac{1}{2\pi} \int_0^{2\pi} f(x)dx$$

and

$$\beta(1,k) = \frac{k}{\pi} \int_0^{2\pi} \cos (kx)f(x)dx$$

$$\beta(2,k) = \frac{k}{\pi} \int_0^{2\pi} \sin (kx)f(x)dx \ .$$

By the definition of S, f is the indefinite integral of f' and $f' \in K$. Since $\int_0^{2\pi} f'(x)dx = 0$, $f(0) = f(2\pi)$. Thus integration by parts shows that for $k \geq 1$,

$$\begin{aligned} \beta(i,k) &= -\frac{1}{\pi} \int_0^{2\pi} \sin (kx)f'(x)dx \quad \text{if} \quad i = 1 \\ &= \frac{1}{\pi} \int_0^{2\pi} \cos (kx)f'(x)dx \quad \text{if} \quad i = 2 \ . \end{aligned}$$

Since $\int_0^{2\pi}(f'(x))^2dx < \infty$, $\beta \in \ell_D^2$ by the Bessel inequality.

Since f is an indefinite integral, it is continuous and of bounded variation. Therefore

$$f(x) = \langle \beta, U(x) \rangle_D$$

by, e.g., Akhieser (1956), section III, 53. Then assumption 2.3 holds and therefore assumption 2.5 holds.

Note that $\sup\limits_{x \in [0,2\pi]} \|U(x)\|_D < \infty$.

Finally

$$\begin{aligned} \|U_n - U_n^*\|_D^2 &= \sum_{k=m+1}^{\infty} (U_k(1,k))^2 + (U_k(2,k))^2 \\ &\leq 2 \sum_{k=n+1}^{\infty} k^{-2} = O(\int_n^{\infty} x^{-2}dx) = O(n^{-1}), \end{aligned}$$

so by the Cauchy-Schwarz inequality $|\langle \beta, U_n - U_n^* \rangle_D| = O(n^{-\frac{1}{2}})$.
Therefore the hypotheses of theorem 3.3 are satisfied with $\gamma = 0$ and $\varepsilon = \frac{1}{2}$.

CHAPTER 4

RESTRICTION TO H A FINITE DIMENSIONAL VECTOR SPACE

4.1. Foreword. If assumption 2.2 holds with S equal to the vector space spanned by k functions $f_1,\ldots,f_k$, then assumption 2.3 holds with $H = R^k$ and U the map

$$s \to \begin{pmatrix} f_1(s) \\ \vdots \\ f_k(s) \end{pmatrix} .$$

In this case, since inner products in R^k are easily computed, it is reasonable to suppose that in assumption 2.4 U_n^* and U_n have been chosen so $U_n = U_n^*$ (see Remark 2.4).

Suppose assumption 2.5 or assumption 4.4, the analogue of assumption 2.5 when U_n is random, hold. As will be seen, $\|\beta_{n+1} - \beta_n\|$ converges to zero. Therefore it is possible to find conditions on U_n so that, roughly speaking, $\beta_n - \beta$ cannot be almost perpendicular to U_n too often and under these conditions, $\|\beta_n - \beta\| \to 0$.

4.2. Theorem. Let assumption 2.5 hold with $H = R^k$ and $U_n = U_n^*$. Fix $p \geq k$ and let W_n be the $k \times p$ matrix whose ith column is U_{n+i-1} for $i = 1,\ldots,p$. Let $\delta_{n,min}$ and $\delta_{n,max}$ be the minimum and maximum eigenvalues, respectively, of $W_n W_n^T$.

Suppose the sequence $\{a_n\}$ satisfies

(1) $\qquad \Sigma\, a_n^2\, \|U_n\|^2(1 + \|U_n\|^2) < \infty \qquad$ and

(2) $\qquad \sum_{k=1}^{\infty} \left(\min_{n_k \leq j \leq n_{k+p-1}} a_j \right) = \infty$

for a sequence of integers $\{n_k\}$ such that

(3) $\qquad n_{k+1} \geq n_k + p \quad$ for all k and

for some $\delta, \Delta > 0$,

(4) $\qquad \delta \leq \delta_{n_k,\min} \leq \delta_{n_k,\max} \leq \Delta \quad$ for all k.

Suppose that for all $\varepsilon > 0$,

(5) $\qquad \inf_n \left(\inf_{\varepsilon^{-1} > |x-\theta_n| > \varepsilon} |R_n(x)| \right) > 0$.

Then $\|\beta_n - \beta\| \to 0$. If in addition $\sup_n \|U_n\| < \infty$ then $X_n - \theta_n \to 0$.

Proof: All the assumptions of lemma 3.1 hold so (3.1.3) and (3.1.4) hold.

By (1)

$$E(\Sigma\, (a_n\|U_n\|(Y_n - R_n(X_n))^2)) \leq \sigma^2\, \Sigma\, a_n^2\|U_n\|^2 < \infty .$$

Thus

(6) $\qquad a_n\|U_n\|(Y_n - R_n(X_n)) \to 0$.

Also by (1)

(7) $\qquad a_n\|U_n\|(1 + \|U_n\|) \to 0.$

Then since (2.5.1) holds with $U_n = U_n^*$

(8) $\|\beta_{n+1} - \beta_n\| \leq a_n\|U_n\|(A(1 + \|\beta_n - \beta\|\|U_n\|) + |Y_n - R_n(X_n)|)$.

By (3.1.3), (6), (7), and (8)

$$\|\beta_{n+1} - \beta_n\| \to 0.$$

For any $k \geq 1$ define

$$A_k = \{k^{-1} < \lim\|\beta_n - \beta\| < k\} \cap \{\|\beta_{n+1} - \beta_n\| \to 0\}.$$

Since except for a set of probability 0

$$\{\lim \|\beta_n - \beta\| > 0\} = \bigcup_{k=1}^{\infty} A_k,$$

to prove $\|\beta_n - \beta\| \to 0$ we need only show that for any k, $P(A_k) = 0$.

We now fix k and fix $\omega \in A_k$. Until the end of the proof we write ξ instead of $\xi(\omega)$ for any random variable ξ. Now choose L_1 such that

$$\|\beta_{n_\ell} - \beta\| > k^{-1} \quad \text{whenever} \quad \ell \geq L_1.$$

Then for all $\ell \geq L_1$,

$$\sum_{i=0}^{p-1} (<U_{n_\ell+i}, \beta_{n_\ell} - \beta>)^2 = \|W_{n_\ell}^T(\beta_{n_\ell} - \beta)\|^2$$

$$= (\beta_{n_\ell} - \beta)^T W_{n_\ell} W_{n_\ell}^T (\beta_{n_\ell} - \beta)$$

$$\geq \delta_{n_\ell,\min}\|\beta_{n_\ell} - \beta\|^2 \geq \delta\ k^{-2}.$$

Here we used (4) and the result that if A is a positive definite $k \times k$ matrix with minimum eigenvalue, λ, then $x^T Ax \geq \lambda\ \|x\|^2$ for all $x \in R^k$ (see Rao (1973) (1f.2.1)).

Thus there exists a sequence $\{m_\ell\}$ such that m_ℓ is in the set $\{n_\ell, n_\ell+1,\ldots,n_\ell+p-1\}$ and

$$(9) \qquad (<U_{m_\ell}, \beta_{n_\ell} - \beta>)^2 \geq \frac{\delta k^{-2}}{p} \quad \text{whenever} \quad \ell \geq L_1.$$

By (3), $m_{\ell+1} \geq m_\ell$ for all ℓ. Also by (4), $(<U_{m_\ell}, X>)^2 \leq \|W_{n_\ell}^T x\|^2 \leq \Delta\|X\|^2$ for $x \in R^k$. Since $\|\beta_{m_\ell} - \beta_{n_\ell}\| \leq \sum_{i=n_\ell+1}^{m_\ell} \|\beta_i - \beta_{i-1}\|$,

$$(10) \quad (<U_{m_\ell}, \beta_{m_\ell} - \beta_{n_\ell}>)^2 \leq \Delta\Big(\sum_{i=n_\ell+1}^{n_\ell+p-1} \|\beta_i - \beta_{i-1}\|^2\Big) .$$

Since $\|\beta_{n+1} - \beta_n\| \to 0$ we have by (10) that for a number L_2

$$(11) \qquad (<U_{m_\ell}, \beta_{m_\ell} - \beta_{n_\ell}>)^2 \leq \frac{\delta k^{-2}}{4p} \quad \text{whenever} \quad \ell \geq L_2.$$

By (9) and (11), if we let $L = \max\{L_1,L_2\}$ then

$$\begin{aligned} |X_{m_\ell} - \theta_{m_\ell}| &= |<U_{m_\ell}, \beta_{m_\ell} - \beta>| \\ &\geq |<U_{m_\ell}, \beta_{n_\ell} - \beta>| - |<U_{m_\ell}, \beta_{m_\ell} - \beta_{n_\ell}>| \\ &\geq \frac{k^{-1}}{2}\sqrt{\frac{\delta}{p}} \quad \text{whenever} \quad \ell \geq L. \end{aligned}$$

Also $|X_{m_\ell} - \theta_{m_\ell}| = |<\hat{\beta}_{m_\ell} - \beta_{m_\ell}, U_{m_\ell}>| \leq \Delta \sup_n\{\|\beta_n - \beta\|\} < \infty$.

Thus by (5) there exists $\Gamma > 0$ such that

$$R_{m_\ell}(X_{m_\ell})(X_{m_\ell} - \theta_{m_\ell}) \geq \Gamma \quad \text{whenever} \quad \ell \geq L.$$

Then since

$$\sum_{n=1}^{\infty} a_n R_n(X_n)(X_n - \theta_n) \geq$$

$$\sum_{\ell=L}^{\infty} a_{m_\ell} R_{m_\ell}(X_{m_\ell})(X_{m_\ell} - \theta_{m_\ell})$$

$$\geq \sum_{\ell=L}^{\infty} (\min_{n_\ell \leq j \leq n_\ell + p-1} a_j)\Gamma = \infty$$

it follows from (3.1.4) that $P(A_k) = 0$.

Finally since $X_n - \theta_n = \langle\beta_n - \beta_1, U_n\rangle$, $X_n - \theta_n \to 0$ if $\sup\|U_n\| < \infty$.

4.3. Remark. Until now we have assumed that s_n is a fixed element of S. Assumption 4.4 is an analogue of assumption 2.5 when $s_n \in S^\Omega$ and θ_n is a random variable.

At time n, the expected output of the process, given the past, depends on both X_n and θ_n. When θ_n was non-random we wrote the expected output as $R_n(X_n)$; the dependence of the output on θ_n is implicit in this expression. When θ_n is random it is more convenient to denote the expected output as $R_n(X_n, \theta_n)$ where R_n is a mapping of R^2 into R.

4.4. Assumption. Assumption 2.2(i) holds. Let R_n be a Borel map from R^2 to R such that

$$(x - y)R_n(x,y) \geq 0 \quad \text{for all } x,y \in R.$$

$$|R_n(x,y)| \leq A(1 + |x - y|) \quad \text{for } A > 0.$$

Let $s_n \in S^\Omega$ and define

$$\theta_n = f(s_n).$$

Suppose assumption (2.3)(i) holds with $H = R^k$ and with $U_n = U(s_n)$, U_n is a measurable transformation into R^k.

Let $\{\beta_n\}$ and $\{Y_n\}$ be random sequences in R^k and R, respectively, such that with

$$F_n = \sigma\{\beta_1,\ldots,\beta_n, U_1,\ldots,U_n\}$$

we have

$$\beta_{n+1} = \beta_n - a_n Y_n U_n \quad \text{for some} \quad a_n \geq 0,$$

$$E^{F_n} Y_n = R_n(X_n,\theta_n),$$

and

$$E^{F_n}(Y_n - R_n(X_n,\theta_n)) \leq \sigma^2 < \infty.$$

<u>4.5. Theorem</u>. Let assumption 4.4 hold. Define

$$F_n^* = \sigma\{\beta_1,\ldots,\beta_n, U_1,\ldots,U_{n-1}\}.$$

Suppose Γ, $K > 0$. Assume

$$\inf_{x \in R^k} P^{F_n^*}(\Gamma\|X\| \leq |\langle U_n,X\rangle| \leq \Gamma^{-1}\|X\|) \geq \Gamma$$

and

$$E^{F_n^*}(\|U_n\|^2(1 + \|U_n\|^2)) < K.$$

If

(1) $\Sigma\, a_n = \infty$ and $\Sigma\, a_n^2 < \infty$

and for all $\varepsilon > 0$

(2) $$\inf_n \inf_{\varepsilon^{-1} > |x-y| > \varepsilon} |R_n(x,y)| > 0,$$

then

$$\|\beta_n - \beta\| \to 0.$$

If

$$a_n = an^{-\alpha} \text{ with } a > 0 \text{ and } \tfrac{1}{2} < \alpha < 1,$$

$$E\|\beta_1\|^2 < \infty,$$

and for some $c > 0$

(3) $$|R_n(x,y)| \geq c|x - y| \text{ for all } x,y \in R,$$

then

$$\sup_n n^{\alpha}\|\beta_n - \beta\|^2 < \infty \text{ and } \sup_n n^{\alpha}\, E|X_n - \theta_n| < \infty.$$

Proof: First,

(4) $$E^{F_n^*}\|\beta_{n+1} - \beta\|^2 = \|\beta_n - \beta\|^2 - 2a_n\, E^{F_n^*} Y_n(X_n - \theta_n) + a_n^2\, E^{F_n^*}(Y_n\|U_n\|)^2.$$

If we define $\eta(x)$ for $x \geq 0$ by

$$\eta(x) = \inf_n \inf_{X\Gamma \leq |y-z| \leq X\Gamma^{-1}} |R_n(y,z)|$$

then,

$$E^{F_n^*}(Y_n(X_n - \theta_n)) = E^{F_n^*}((X - \theta_n)E^{F_n}Y_n)$$

$$= E^{F_n^*}(X_n - \theta_n)R_n(X_n,\theta_n)$$

$$\geq \Gamma\|\beta_n - \beta\|\eta(\|\beta_n - \beta\|)P^{F_n^*}(\Gamma^{-1}\|\beta_n-\beta\|\geq|<U_n,\beta_n-\beta>|\geq\Gamma\|\beta_n-\beta\|)$$

Thus,

$$(5) \qquad E^{F_n^*}(Y_n(X_n - \theta_n)) \geq \Gamma^2\|\beta_n - \beta\|\eta(\|\beta_n - \beta\|).$$

Next,

$$E^{F_n^*}(Y_n\|U_n\|)^2 = E^{F_n^*}(\|U_n\|^2 E^{F_n}Y_n^2)$$

$$\leq E^{F_n^*}\|U_n\|^2(R_n^2(X_n,\theta_n) + \sigma^2)$$

and since

$$R_n^2(X_n,\theta_n) \leq 2A^2(\|\beta_n - \beta\|^2\|U_n\|^2 + 1)$$

$$(6) \qquad E^{F_n^*}(Y_n\|U_n\|)^2 = O(\|\beta_n - \beta\|^2 + 1) .$$

By using (4) - (6) we obtain

$$(7) \quad E^{F_n^*}\|\beta_{n+1} - \beta\|^2 \leq \|\beta_n - \beta\|^2(1 + f_n)$$

$$- 2a_n\Gamma\|\beta_n - \beta\|\eta(\|\beta_n - \beta\|) + g_n$$

with $f_n,g_n \geq 0$ and $f_n,g_n = O(a_n^2)$. Then by theorem 1 of Robbins and Siegmund (1971), $\lim \|\beta_n - \beta\|$ exists and is finite and

$$\Sigma\, a_n\|\beta_n - \beta\|\eta(\|\beta_n - \beta\|) < \infty .$$

Since by (1) and (2), $\Sigma\, a_n x_n \eta(x_n) = \infty$ if $\{x_n\}$ is any sequence of numbers satisfying $x_n \to x$ with $x \neq 0$, we have $\|\beta_n - \beta\| \to 0$.

Moreover, (3) implies

$$\eta(x) \geq c\ \Gamma\ |x|$$

and this with (7), $E\|\beta_1\|^2 < \infty$, and $a_n = an^{-\alpha}$ implies

$$E\|\beta_{n+1} - \beta\|^2 \leq E\|\beta_n - \beta\|^2(1 + f_n) - M\ E\ \|\beta_n - \beta\|^2\ n^{-\alpha} + g_n$$

for some $M > 0$ and with $f_n, g_n \geq 0$ and $f_n, g_n = O(n^{-2\alpha})$.

Then by a lemma of Chung (see Fabian (1971), lemma 3.1)

$$\sup_n \{n^\alpha E\|\beta_n - \beta\|^2\} < \infty.$$

Since $E|X_n - \theta_n| \leq E(\|\beta_n - \beta\|\|U_n\|) \leq (E\|\beta_n - \beta\|^2\ E\|U_n\|^2)^{1/2}$ and $E\|U_n\|^2 \leq K$

$$\sup_n \{n^\alpha\ E|X - \theta_n|\} < \infty.$$

<u>4.7. Example</u>. With this example we show that the assumptions of theorem 3.3 imply neither $X_n - \theta_n \to 0$ nor $\|\beta_n - \beta\| \to 0$.

Let $H = R^2$. Suppose e_1 and e_2 are the standard unit vectors in R^2, i.e., $e_1^T = (1,0)$ and $e_2^T = (0,1)$. Suppose β is the zero vector, $R_n(x) = x$ for all n, $\beta_1 = e_1$ and $a_n = an^{-1}$ for $0 < a < 1$.

Let G be a subsequence of the integers such that $\sum_{n \in G} a_n < \infty$. Assume that U_n is e_1 or e_2 according as $n \in G$ or $n \notin G$. Assume the process is deterministic, i.e. $Y_n = R_n(X_n)$.

For $\xi \in R^2$ let $\xi^{(i)}$ be the ith coordinate of ξ, $i = 1,2$.

If $n \notin G$, then $U_n^{(1)} = 0$ and therefore

$$(1) \qquad \beta_{n+1}^{(1)} = \beta_n^{(1)} \quad \text{if} \quad n \notin G.$$

If $n \in G$, then $Y_n = X_n = \langle\beta_n, U_n\rangle = \beta_n^{(1)}$ and $U_n^{(1)} = 1$, so

$$(2) \qquad \beta_{n+1}^{(1)} = \beta_n^{(1)}(1 - a_n) \quad \text{if} \quad n \in G .$$

Since $\beta_1^{(1)} = 1$, we have by (1) and (2) that

$$\beta_n^{(1)} = \prod_{\substack{k<n \\ k\in G}} (1 - a_k) \quad \text{for} \quad n > 1 .$$

Since $a < 1$, $(1 - a_n) \neq 0$ for all n. Then since $\sum_{n\in G} a_n < \infty$, there exists $d > 0$ such that

$$\lim_{n\to\infty} \Big(\prod_{\substack{k\in G \\ k<n}} (1 - a_k)\Big) = d .$$

Therefore $\beta_n^{(1)} \not\to 0 = \beta^{(1)}$. Moreover $X_n = \beta_n^{(1)}$ whenever $n \in G$ and therefore $X_n - \theta_n \not\to 0$.

4.8. Example. Here we have another example satisfying the conditions of theorem 3.3 but for which $\beta_n \not\to \beta$. However, in this case $X_n - \theta_n \to 0$.

Let $H = R^2$. Elements of H will be represented as complex numbers. Suppose $\beta = 0$ and

$$\begin{aligned} R_n(x) &= 1 \quad \text{for} \quad x > 0 \\ &= -1 \quad \text{for} \quad x \leq 0. \end{aligned}$$

Suppose $c_1 = 1$ and

$$c_n = e^{i \sum_{j=1}^{n-1} j^{-1}} \quad \text{for } n \geq 2.$$

Let $a_n = |e^{in^{-1}} - 1|$ and

$$U_n = (e^{in^{-1}} - 1)a_n^{-1}c_n .$$

Also assume $\beta_1 = 1$ and $Y_n = R_n(X_n)$. Then for all n

(1) $$\beta_n = c_n \quad \text{and}$$

(2) $$X_n - \theta_n = X_n = -\left(\frac{1 - \cos n^{-1}}{2}\right)^{\frac{1}{2}} ,$$

whence $X_n - \theta_n \to 0$ but $\|\beta_n - \beta\| = 1$ for all n.

To prove the last statement, first note that

$$a_n^2 = (e^{in^{-1}} - 1)(e^{-in^{-1}} - 1) = 2(1 - \cos n^{-1}).$$

Next, with $\operatorname{Re} \phi$ denoting the real part of the complex number ϕ,

$$\begin{aligned} \langle c_n, U_n \rangle &= \operatorname{Re}(\bar{c}_n U_n) \\ &= \operatorname{Re}((e^{in^{-1}} - 1)a_n^{-1}) \\ &= -\left(\frac{1 - \cos n^{-1}}{2}\right)^{\frac{1}{2}}. \end{aligned}$$

Thus if (1) holds for $n = k$, so does (2). Moreover (1) and (2) with $n = k$ imply (1) for $n = k+1$ by the following calculation:

$$\beta_{k+1} = c_k - a_k(R_k(-(\frac{1 - \cos k^{-1}}{2})^{\frac{1}{2}}))U_k$$

$$= c_k + (e^{ik^{-1}} - 1)c_k$$

$$= c_k e^{ik^{-1}} = c_{k+1} .$$

By observing that (1) holds for $n = 1$ the proof is completed.

Note that by Taylor's theorem

$$a_n^2 = 2(1 - \cos n^{-1}) = n^{-2} + O(n^{-4})$$

with $0 < d_n < n^{-1}$. It is then easy to see that the assumptions of theorem 3.3 hold if the theorem is trivially generalized by replacing the assumption $a_n = an^{-\alpha}$ by $a_n = c_n n^{-\alpha}$ with $0 < m \leq c_n \leq M < \infty$ for some m,M.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akhiezer, N.I. (1956). _Theory of Approximation_. Ungar, New York.

Alexits, G. (1961). _Convergence Problems of Orthogonal Series_. Pergamon Press, New York.

Blum, J.R. (1954). Multidimensional stochastic approximation methods. _Ann. Math. Statist._ _25_ 737-744.

Dupăc, V. (1965). A dynamic stochastic approximation method. _Ann. Math. Statist._ _36_ 1695-1702.

Dupăc, V. (1966). Stochastic approximations in the presence of trend. _Czech. Math. J._ _16_ (_91_) 454-461.

Fabian, V. (1971). Stochastic approximation. _Optimizing Methods in Statistics_ (J.S. Rustagi, ed.) 439-470. Academic Press, New York.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. _Ann. Math. Statist._ _23_ 462-466.

Loève, M. (1963). _Probability Theory_ (3rd edition). Van Nostrand, New York.

Rao, C.R. (1973). _Linear Statistical Inference and Its Applications_. John Wiley, New York.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. _Ann. Math. Statist._ _22_ 400-407.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. _Optimizing Methods in Statistics_ (J.S. Rustagi, ed.) 233-257. Academic Press, New York.

Uosaki, K. (1974). Some generalizations of dynamic stochastic approximation processes. _Ann. Statist._ _2_ 1042-1048.