A KIEFER - WOLFOWITZ TYPE STOCHASTIC APPROXIMATION PROCEDURE

Dissertation for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

THOMAS EDWARD OBREMSKI

1976

This is to certify that the

thesis entitled

A KIEFER-WOLFOWITZ TYPE
STOCHASTIC APPROXIMATION PROCEDURE

presented by

Thomas Edward Obremski

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics and Probability

Václav Fabian
Major professor

Date April 30, 1976

O-7639

ABSTRACT

A KIEFER-WOLFOWITZ TYPE
STOCHASTIC APPROXIMATION PROCEDURE

By

Thomas Edward Obremski

In considering both the Robbins-Monro (RM) and Kiefer-Wolfowitz (KW) stochastic approximation procedures, Abdelhamid (1973) has shown that if the density g of the errors in estimating function values (RM case), and differences of function values (KW case) is known, then a transformation of observations leads to methods which under mild conditions have desirable asymptotic properties. Fabian (1973) obtained the same asymptotic results in the (RM) case without assuming knowledge of the density g. We study the analogous problem in the (KW) case, and obtain the same asymptotic results as Abdelhamid.

A KIEFER-WOLFOWITZ TYPE STOCHASTIC APPROXIMATION PROCEDURE

By

Thomas Edward Obremski

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

1976

To Kacey, who
(to the best of my
knowledge) never
doubted I could do it.

ACKNOWLEDGEMENTS

I am sincerely grateful to Professor Václav Fabian for the guidance he gave me during my thesis work. His patience and valuable suggestions were always a source of encouragement for me. I feel that I have learned from him not only about the content area of the thesis, but also a good deal about writing style.

I would also like to thank David Ruppert for reading the material and discussing it with me, and Mrs. Noralee Burkhardt for her excellent typing of the thesis.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

Consider the stochastic approximation procedure given by

$$(1.1) \qquad X_{n+1} = X_n - a_n c_n^{-1} Y_n \ , \quad n = 1,2,\ldots,$$

where X_n, Y_n are random variables and a_n, c_n are positive numbers. Included in (1.1) are both the Robbins-Monro (1951) procedure (RM), and the Kiefer-Wolfowitz (1952) procedure (KW).

Abdelhamid (1973) and, independently, Anbar (1973) investigated the possible effect that transforming the observed random variables Y_n might have on the almost sure convergence and the asymptotic normality. Abdelhamid's investigation included both the RM and KW cases, Anbar's only the RM case.

Specifically they studied the asymptotic behavior of the procedure given by

$$(1.2) \qquad X_{n+1} = X_n - a_n c_n^{-1} h(Y_n) \ , \quad n = 1,2,\ldots,$$

where h was assumed to belong to a class C of Borel measurable functions which preserve both the almost sure convergence of X_n to θ and the asymptotic normality of

$n^{\beta}(X_n - \theta)$, where in the RM case, θ is the unknown root of the function f and $\beta = \frac{1}{2}$, and in the KW case, θ is the point of minimum of f and β lies in the interval [1/4, 1/3], depending on the assumptions on f. Denote by $\mathcal{F}_n$ the sigma-algebra generated by $X_1, X_2, \ldots, X_n$. The following conditions were assumed for the random variables $V_n = Y_n - E^{\mathcal{F}_n}(Y_n)$: V_n are conditionally (given $\mathcal{F}_n$) distributed according to a symmetric distribution function G admitting density g; g has derivative almost everywhere with respect to G; $0 < I(g) = \int (g'(v)/g(v))^2 dG(v) < +\infty$.

Within the class $\mathcal{C}$, they sought that function h^* which would minimize the second moment of the asymptotic distribution. It was known that in the case where g is normal, such an h was given by the identity function, that is, when g is known to be normal, (1.1) cannot be improved upon by transformation of observations. They found in general that within the class $\mathcal{C}$, $h^*(v) = (-g'/g)(v)$, unique up to multiplicative constant. So for example if g is double exponential, then $(-g'/g)(v) = C\,\mathrm{sign}(v)$ with a constant $C > 0$, and the optimal procedure is

$$X_{n+1} = X_n - a_n c_n^{-1}\,\mathrm{sign}(Y_n), \quad n = 1,2,\ldots, \tag{1.3}$$

first suggested by Fabian (1960 and 1964).

Abdelhamid also suggested improvements in some cases where G is known but fails to satisfy all of the assumptions above.

Without assuming knowledge of the distribution G, Fabian (1973a) constructed a RM-type procedure which performs asymptotically as well as the transformed RM procedure (1.2) does when G is known. With $a_n = an^{-1}$, $c_n = 1$ in the RM case and $cn^{-\gamma}$ in the KW case, a and c positive numbers, and γ in the interval [1/6, 1/4], Abdelhamid had derived values of a and c optimal in the sense of minimizing the second moment of the asymptotic distribution. In the RM case the optimal choice of a is $(f'(\theta)I(g))^{-1}$. Fabian suggested methods of estimating $I(g)$, $-g'/g$, and $f'(\theta)$ and pointed out some of the problems inherent in such estimation.

The main purpose of this paper is to achieve asymptotic results in the KW case with G unknown which are as strong as those obtained by Abdelhamid. Much of the direction for this present paper is provided by Fabian's 1973 paper which we shall refer to henceforth as I. In some places we were able to apply results obtained in I directly to the KW case. These places are indicated in the text. Much of the actual estimation of unknown parameters that is outlined in our main result, Theorem (4.1) is carried out as in I.

But, as in previous cases, when properties were obtained first for the RM case, the proofs for the analogous properties in the KW case had to be changed at several points to cover the more difficult situation. The paper gives detailed treatment only to the new parts of the proof and refers to I for the other parts, to an extent that leaves the paper readable.

The same speeds of convergence (Theorem (2.11)) and asymptotic normality result (Theorem 3.1)) as those obtained by Abdelhamid (Theorems (4.4) and (4.5)) are achieved. The main result, Theorem (4.1), is a realization of the procedure suggested, indicating how to estimate the optimal values of a and c, as well as $I(g)$, $-g'/g$, $f''(\theta)$, and $f'''(\theta)$. As in the RM case there is more freedom in estimating $-g'/g$ than in estimating $I(g)$. This is reflected in the conditions $\varepsilon_n \geq n^{-\beta_1}$ in (4.2.vi) and $\varepsilon_n \geq (\log n)^{-\beta_0}$ in (4.2.iii).

In Abdelhamid's treatment of the KW case, and in many of the earlier treatments, the following two assumptions have appeared: First, there exist constants A and B such that

(1.4) $$|f(x+1) - f(x)| < A|x - \theta| + B, \text{ for every } x \text{ in } R,$$

and, secondly,

(1.5) $$E^{F_n}(V_n^2) \leq \sigma^2, \text{ for every natural number } n,$$

for a number σ and $V_n = Y_n - E^{F_n}(Y_n)$. The latter assumption may be omitted here if in the truncation of the Y_n given in (2.6.4), y_n are chosen to be $(\log (n \vee 2))^{1-2\varepsilon_1}$. The use of truncation in (2.6.4) also enables us to weaken the assumption (1.4) to f being bounded on bounded intervals. Without the truncated term in the recursion relation (2.6.3) we would need to assume not only (1.4) but also a similar type of condition for $E^{F_n}(-g'/g)(Y_n)$.

CHAPTER TWO

ALMOST SURE CONVERGENCE OF THE PROPOSED STOCHASTIC APPROXIMATION PROCEDURE

2.1 Basic Notation

All random variables are assumed to be defined on a probability space $(\Omega, \mathcal{F}, P)$. Relations between random variables, including convergence, are meant to hold almost surely, unless specified otherwise. The set of real numbers is denoted by R, positive reals by R^+, and the class of all Borel subsets of R by $\mathcal{B}$. The indicator function of a set S is denoted by χ_S. Let E denote expectation, and $E^{\mathcal{F}}$ conditional expectation, given the σ-algebra $\mathcal{F}$. If $Z_1, \ldots, Z_n$ are random variables, then $\mathcal{F}(Z_1, \ldots, Z_n)$ denotes the σ-algebra induced by $Z_1, \ldots, Z_n$.

If $\{b_n\}$ is a sequence of numbers and $\{Z_n\}$ a sequence of random variables, then we write $Z_n = \mathcal{O}(b_n)$ if $\limsup |b_n^{-1} Z_n(\omega)| < +\infty$ for almost all ω. Similarly we write $Z_n = \mathcal{O}_u(b_n)$ if there exists a K in R and an integer n_0 with $|b_n^{-1} Z_n| \leq K$, for all $n \geq n_0$.

If φ is a function on R and a is in R^+, then for each x in R, $\varphi^a(x)$ denotes the difference $\varphi(x+a) - \varphi(x-a)$; if k is a natural number, $D^k\varphi(x)$ denotes the k^{th} derivative of φ at x.

2.2 Remark

The following assumptions are listed for reference later. Assumptions (2.6) and (2.7) appear in the convergence results in this chapter, Assumption (2.8) in the asymptotic normality result, Theorem (3.1). Only Assumptions (2.3) and (2.4) appear in the main result, Theorem (4.1).

2.3 Assumption

Both θ and γ belong to R. We assume that f is a function on R such that either

$D^2 f$ exists, is continuous in a neighborhood of θ, and $\gamma = 1/4$,

or

$D^3 f$ exists, is continuous in a neighborhood of θ, and $\gamma = 1/6$.

We further assume that f is bounded on bounded intervals, that $D^2 f(\theta) = M > 0$, and that for every natural number k,

(2.3.1) $$\sup_{-k<x-\theta<-\frac{1}{k}} \overline{D}\, f(x) < 0; \quad \inf_{\frac{1}{k}<x-\theta<k} \underline{D}\, f(x) > 0,$$

where $\overline{D}\, f(x)$ and $\underline{D}\, f(x)$ denote respectively the upper and lower derivatives of f at x.

2.4 Assumption

Assumption (2.3) holds. We assume that $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are random variables, that F_n is a non-decreasing sequence of σ-algebras such that for each n, $\mathcal{F}_n$ contains the σ-algebra $\mathcal{F}(X_1,\ldots,X_n,Y_1,\ldots,Y_{n-1})$. For each n, C_n is a positive $\mathcal{F}_n$-measurable random variable with $c_n = C_n n^{-\gamma}$, and $f^{c_n}(X_n)$ is the $\mathcal{F}_n$-measurable random variable whose value at ω is

$$f(X_n(\omega) + c_n(\omega)) - f(X_n(\omega) - c_n(\omega)) .$$

For each n, $Y_n - f^{c_n}(X_n)$ is conditionally, given $\mathcal{F}_n$, distributed according to a distribution function G which is symmetric, has zero expectation, has a density g which has a continuous derivative Dg everywhere on R. The density g is non-increasing on $[0,\infty)$ and $0 < I(g) = \int(g^{-1}D(g))^2 dG < +\infty$.

2.5 Remark

The assumption of symmetry of G is a natural one in a Kiefer-Wolfowitz type of procedure, where Y_n is an unbiased estimator of $f^{c_n}(X_n)$. This requirement is satisfied, for example, if the errors in estimating $f(X_n + c_n)$ and $f(X_n - c_n)$, respectively, are independent and identically distributed, given n.

2.6 Assumption

Assumption (2.4) holds and h_n are measurable functions on $(\Omega \times R, \mathcal{F}_n \times \mathcal{B})$ such that for each ω, $h_n(\omega,\cdot)$ are odd, and are non-negative on $[0,+\infty)$. For each n, D_n is a non-negative $\mathcal{F}_n$-measurable random variable and

$$(2.6.1)\quad |h_n(\omega,t)| \le n^{\varepsilon_1}\chi_{(-n,n)}(t),$$

$$(\log n)^{-\varepsilon_0} \le C_n \le (\log n)^{\varepsilon_0},\ D_n \le n^{\varepsilon_1}$$

with numbers $\varepsilon_0, \varepsilon_1$ satisfying $0 < \varepsilon_1 < \gamma/2$ and $0 < \varepsilon_0 < \varepsilon_1/2$. Note that for both possible values of γ, we can (and will) select a μ_γ such that

$$(2.6.2)\qquad 1/2 - \gamma - 2\varepsilon_1 > \mu_\gamma > 0.$$

We shall write $h_n(t)$ for $h_n(\cdot,t)$, and $h_n(Y_n)$ for $h_n(\cdot,Y_n(\cdot))$.

The random variables $X_1, X_2, \ldots$ satisfy

$$(2.6.3)\quad X_{n+1} = X_n - (nc_n)^{-1}[D_n h_n(Y_n) + \log(n \vee 2)^{-1+\varepsilon_1}\tilde{Y}_n]$$

where

$$(2.6.4)\qquad \tilde{Y}_n = (Y_n \vee (-y_n)) \wedge y_n$$

with $y_n = n^{\varepsilon_1}$ if G has finite second moment and $y_n = (\log(n \vee 2))^{1-2\varepsilon_1}$ otherwise.

2.7 <u>Assumption</u>

Assumption (2.6) holds. For almost all ω, $h_n(\omega,\cdot) \to -g^{-1}(Dg)$ on the set $\{t;\ g(t) > 0\}$ and

$$(2.7.1)\qquad D_n \to (2M\ I(g))^{-1}.$$

2.8 <u>Assumption</u>

Assumption (2.7) holds and

$$(2.8.1)\quad \int[h_n(t + \eta_n(t)) + g^{-1}(Dg)]^2 dG \to 0$$

for every sequence $\{\eta_n\}$ of functions on $\Omega \times R$ with $|\eta_n| \leq |f^{c_n}(X_n)|$ and such that, for almost all ω, $h_n(\omega, t + \eta_n(\omega,t))$ are Borel measurable with respect to t. The random variables $C_1, C_2, \ldots$ satisfy

$$(2.8.2) \quad C_n \to \begin{cases} [(\frac{3}{2})(D^3 f(\theta))^{-2} I^{-1}(g)]^{1/6} & \text{if } \gamma = 1/6 \\ & \text{and } D^3 f(\theta) \neq 0 \\ C & \text{otherwise,} \end{cases}$$

where C is in R^+.

2.9 Remark

Suppose Assumption (2.6) holds. In the proof of Theorem (2.10) we shall require expressions for $E^{F_n} h_n(Y_n)$ and $E^{F_n} \tilde{Y}_n$. If k is a Borel measurable function, then the conditional expectation $E^{F_n} k(Y_n)$, provided it exists, is equal to $K(E^{F_n} Y_n)$ where $K(\Delta) = \int k(t + \Delta) g(t) dt$. For $k = h_n$ and $k(t) = (t \vee (-y_n)) \vee y_n$, several properties of K were established in I under the same conditions on g and on h_n as we assumed in (2.6). So using the results (I3.1.1), (I3.1.2), and (I3.1.9) we have

$$(2.9.1) \quad E^{F_n} h_n(Y_n) = \Psi_n(f^{c_n}(X_n)), \quad E^{F_n} \tilde{Y}_n = f^{c_n}(X_n) \kappa_n$$

where Ψ_n are functions satisfying

$$(2.9.2) \qquad \Delta \Psi_n(\Delta) \geq 0 \quad \text{for all } \Delta \text{ in } R,$$

(2.9.3) $\Delta^{-1}\Psi_n(\Delta) \leq kn^{\varepsilon_1}$ with a k in R^+, for all $\Delta \neq 0$,

and κ_n (equal to $\Phi_n'(\Delta_n)$ in I) are non-negative F_n-measurable random variables with $\kappa_n \to 1$ on the set of all ω for which $\{f^{c_n}(X_n(\omega))\}$ is a bounded sequence.

2.10 Theorem

If Assumption (2.6) holds, then $(\log n)^\beta(X_n - \theta) \to 0$, for every $\beta > 0$.

Proof:

Assume without loss of generality that $\theta = 0$. Let $\varepsilon > 0$. It is easy to see that there is a function φ on R such that $\varphi(x) = \varphi(-x)$ for all x, $\varphi = 0$ on $[0,\varepsilon]$ and $\varphi > 0$ on $(\varepsilon,+\infty)$, φ has a bounded second derivative and first derivative $\underset{\sim}{D}$ satisfying $|\underset{\sim}{D}(x)| \leq |x|$ for all x, and $\underset{\sim}{D}(x) = x$ for $x > 2\varepsilon$.

We have $c_n = C_n n^{-\gamma} \leq n^{\varepsilon_1 - \gamma}$ by (2.6.1) and it suffices to consider n so large that $c_n < \varepsilon/2$. Then by (2.9.1)

$$(2.10.1) \qquad \underset{\sim}{D}(X_n)E^{F_n}\tilde{Y}_n \geq 0,$$

since $\kappa_n \geq 0$ and sign $f^{c_n}(x) =$ sign x for $|x| > c_n$ by (2.3.1). Define $B_n = (c_n^{-1}\underset{\sim}{D}(X_n)E^{F_n}\tilde{Y}_n)^{\frac{1}{2}}$. Write (2.6.3) as $X_{n+1} = X_n - U_n$, and $N_n = E^{F_n}U_n$. Then with $\alpha_n = n^{-1}(\log n)^{-1+\varepsilon_1}$, we have $\underset{\sim}{D}(X_n)N_n = A_n + \alpha_n B_n^2$, where by (2.9.2), $A_n \geq 0$. So

(2.10.2) $$\underset{\sim}{D}(X_n)N_n \geq \alpha_n B_n^2 .$$

Also, by (2.6.1), (2.6.2), and (2.6.4)

(2.10.3) $$E^{F_n} U_n^2 = O_u(n^{-1-2\mu_\gamma}) \text{ with } \mu_\gamma > 0.$$

Relations (2.10.2) and (2.10.3) show that conditions (2), (3), and (4) of Lemma (3.3), Fabian (1971), are satisfied with $\gamma_n = \varepsilon_n = 0$, and $\beta_n = k (\log n)^{2\varepsilon_0} (n^{-1-2\mu_\gamma})$ with a k in R^+. Hence a subsequence $\{B_{n_i}\}$ of $\{B_n\}$ converges to 0 and the sequence $\{\varphi(X_n)\}$ converges to a random variable.

Let ω be a point at which both properties hold. Since $\varphi(X_{n_i}(\omega))$ converges, $X_{n_i}(\omega)$ is bounded, and then so is $f^{c_{n_i}}(X_{n_i}(\omega))$. Therefore $\kappa_{n_i}(\omega) \to 1$ and so $c_{n_i}^{-1}(\omega)\underset{\sim}{D}(X_{n_i}(\omega))f^{c_{n_i}}(X_{n_i}(\omega)) \to 0$. The latter convergence, the properties of $\underset{\sim}{D}$, and (2.3.1) imply that $\limsup |X_{n_i}(\omega)| \leq \varepsilon$. But, since $\varphi(X_n(\omega))$ converges, $\varphi(X_n(\omega)) \to 0$ and $\limsup |X_n(\omega)| \leq \varepsilon$. The final relation holds for all ω in a set of probability one. Since ε was chosen arbitrary and positive, $X_n \to 0$. (Note that as a consequence, $\kappa_n \to 1$.)

Now suppose that N is a neighborhood of $x = 0$ in which $D^2 f$ exists and is continuous if $\gamma = 1/4$, and in which $D^3 f$ exists and is continuous if $\gamma = 1/6$.

Expanding $f^{c_n}(X_n)$ in powers of c_n in N we obtain, with a proper choice of ξ_n, that

$$(2.10.4)\quad c_n^{-1} f^{c_n}(X_n) = \alpha_n X_n + \xi_n c_n^{-1+\frac{1}{2\gamma}},$$

where α_n and ξ_n are F_n-measurable random variables, with

$$(2.10.5)\quad \alpha_n \to 2M,\ \xi_n \to \xi_0 = \begin{cases} 0 & \text{if } \gamma = 1/4 \\ \frac{1}{3} f'''(0) & \text{if } \gamma = 1/6 . \end{cases}$$

Using expression (2.10.4) we obtain

$$(2.10.6)\quad N_n = n^{-1}(\alpha_n X_n + \xi_n c_n^{-1+\frac{1}{2\gamma}}).$$

$$[D_n(f^{c_n}(X_n))^{-1}\Psi_n(f^{c_n}(X_n)) + (\log n)^{-1+\varepsilon_1} \kappa_n].$$

For ε_0 as in (2.6.1), we obtain from (2.10.6) and (2.9.3) that

$$(2.10.7)\quad N_n = n^{-1}(\log n)^{-1+\varepsilon_0}(\alpha_n X_n + \xi_n c_n^{-1+\frac{1}{2\gamma}})\delta_n,$$

with

$$(2.10.8)\qquad 0 \le \delta_n = O_u((\log n)^{1-\varepsilon_0} n^{2\varepsilon_1}),\ \delta_n \to +\infty.$$

Then $X_n - N_n = X_n(1 - n^{-1}(\log n)^{-1+\varepsilon_0}\alpha_n\delta_n) - R_n$, where $R_n = n^{-1}(\log n)^{-1+\varepsilon_0}\xi_n c_n^{-1+\frac{1}{2\gamma}}\delta_n$. Note that from

(2.10.5), (2.10.8), (2.6.1) and (2.6.2) we have that

(2.10.9) $$R_n = O(n^{-1-\mu_\gamma}), \quad \mu_\gamma > 0.$$

Now, eventually, depending on ω, $0 \le 1 - n^{-1}(\log n)^{-1+\varepsilon_0}\alpha_n\delta_n \le 1$, $(1-n^{-1}(\log n)^{-1+\varepsilon_0}\alpha_n\delta_n)^2 \le 1 - n^{-1}(\log n)^{-1+\varepsilon_0}\alpha_n\delta_n \le 1 - n^{-1}(\log n)^{-1+\varepsilon_0}$, and

(2.10.10) $$(X_n - N_n)^2 \le X_n^2(1 - n^{-1}(\log n)^{-1+\varepsilon_0}) + 2|X_nR_n| + R_n^2 .$$

Writing $X_{n+1} = (X_n - N_n) - (U_n - N_n)$ we obtain from (2.10.10)

(2.10.11) $$X_{n+1}^2 \leqq (1 - A_n)X_n^2 - 2V_n + W_n + T_n$$

with

(2.10.12) $$A_n \geqq n^{-1}(\log n)^{-1+\varepsilon_0},$$

(2.10.13) $$V_n = (X_n - N_n)(U_n - N_n), \quad W_n = (U_n - N_n)^2$$

$$T_n = 2|X_nR_n| + R_n^2 .$$

Suppose now β_n are positive numbers satisfying (eventually)

(2.10.14) $$\beta_n^{-1}\beta_{n+1}(1-A_n) \leqq 1, \quad \beta_{n+1}X_n = O(n^{\mu_\gamma-\eta}),$$

$$\beta_n \leqq n^{2\mu_\gamma-\eta} \quad \text{for an } \eta > 0 .$$

We shall show that under these conditions

$$(2.10.15) \qquad \sum_{n=1}^{\infty} \beta_{n+1} W_n < +\infty, \quad \sum_{n=1}^{\infty} \beta_{n+1} T_n < +\infty$$

$$\text{and} \quad \sum_{n=1}^{\infty} \beta_{n+1} V_n < \infty .$$

This, (2.10.14), and (2.10.11) then easily imply

$$(2.10.16) \qquad \beta_n X_n^2 = O(1) .$$

The first relation in (2.10.15) follows from (2.10.3) since $EW_n \leqq EU_n^2$. The second relation follows since by (2.10.9), $R_n^2 = O(n^{-2-2\mu_\gamma})$ and $\beta_{n+1}|X_n||R_n| = O(n^{\mu_\gamma - \eta}) O(n^{-1-\mu_\gamma})$.

From (2.10. 3), $E^{F_n} V_n^2 = (X_n - N_n)^2 O_u(n^{-1-2\mu_\gamma})$.

But $(X_n - N_n)^2 \leqq X_n^2 + T_n$ by (2.10.10) and

$\sum_{n=1}^{\infty} \beta_{n+1}^2 T_n |O_u(n^{-1-2\mu_\gamma})| \leqq \sum_{n=1}^{\infty} \beta_{n+1} T_n < +\infty$ as we have already shown. Concerning the other term, we have

$$\beta_{n+1}^2 X_n^2 O_u(n^{-1-2\mu_\gamma}) = O(n^{-1-\eta}).$$

This shows that $\sum_{n=1}^{\infty} \beta_{n+1}^2 E^{F_n} V_n^2 < +\infty$ and $\sum_{n=1}^{\infty} \beta_{n+1} V_n$ converges by the generalized Borel-Cantelli Lemma (Lemma 10, Dubins and Freedman (1965)). Thus (2.10.15) holds.

Choose now $\beta_n = (\log n)^b$ so that (2.10.14) is satisfied and (2.10.16) holds. This proves the theorem.

2.11 Theorem

If Assumption (2.7) holds, then $n^{\beta}(X_n - \theta) \to 0$ for every $0 < \beta < \frac{1}{2} - \gamma - 2\varepsilon_1$.

Proof:

Relation (I3.1.16) can be rewritten as

$$\text{(2.11.1)} \qquad \liminf \mu_n \geq I(g)$$

with $\mu_n = f^{-1}(X_n)\Psi_n(f(X_n))$. This relation holds also in our case with $\mu_n = [f^{c_n}(X_n)]^{-1}\Psi_n[f^{c_n}(X_n)]$. So we obtain from (2.10.6) and (2.7.1) a strengthening of (2.10.7) to

$$\text{(2.11.2)} \quad N_n = n^{-1}k_n(\alpha_n X_n + \xi_n c_n^{-1+\frac{1}{2\gamma}})$$

with $\liminf k_n \geq (2M)^{-1}$, $k_n = \mathcal{O}_u(n^{2\varepsilon_1})$. Then, (2.10.11) holds with (2.10.12) strengthened to

$$\text{(2.11.3)} \qquad A_n \geqq 2n^{-1}\alpha_n k_n'$$

with $k_n - k_n' \to 0$.

Now suppose $n^{2\beta_0}X_n^2 \to 0$ for a β_0 in $[0,\mu_\gamma]$. We know this is true at least for $\beta_0 = 0$. Choose a β in (β_0, μ_γ) and set $\beta_n = n^{\beta+\beta_0}$. These β_n satisfy (2.10.14) and thus, also, (2.10.16). Thus $n^{\beta}X_n \to 0$ for every $\beta < \mu_\gamma$; but since μ_γ can be chosen as any number less than $\frac{1}{2} - \gamma - 2\varepsilon_1$ (see (2.6.2)), the assertion of the theorem holds.

CHAPTER THREE

ASYMPTOTIC NORMALITY OF THE PROPOSED PROCEDURE

To obtain the following asymptotic normality result for the procedure proposed in (2.6.3) we use a one-dimensional version of Theorem (2.2), Fabian (1968).

3.1 Asymptotic Normality Theorem

If Assumption (2.8) holds, then $n^{\frac{1}{2}-\gamma}(X_n - \theta)$ is asymptotically normal with

$$\text{(i)} \begin{cases} \text{mean} = 0, & \\ \text{variance} = [6\ I(g)M^2C^2]^{-1} & \text{if } \gamma = 1/4, \end{cases}$$

$$\text{(ii)} \begin{cases} \text{mean} = 0, & \text{if } \gamma = 1/6 \text{ and} \\ \text{variance} = [(16/3)I(g)M^2C^2]^{-1} & D^3f(\theta) = 0, \\ \text{mean} = -(Q/128)^{1/3}, & \text{if } \gamma = 1/6 \text{ and} \\ \text{variance} = (Q/2^{11/2})^{2/3}, & D^3f(\theta) \neq 0, \end{cases}$$

where $Q = 3D^3f(\theta)M^{-3}I^{-1}(g)$.

Proof:

Assume without loss of generality that $\theta = 0$. Suppose Assumption (2.8) holds. As in I, proof of Theorem (3.1,iii), use (2.8.1) and the Schwarz inequality to obtain $\limsup [f^{c_n}(X_n)]^{-1}\Psi_n[f^{c_n}(X_n)] \leq I(g)$. This, along with (2.11.1) gives

$$(3.1.1) \qquad [f^{c_n}(X_n)]^{-1}\Psi_n[f^{c_n}(X_n)] \to I(g),$$

and from (2.10.6)

$$(3.1.2) \qquad N_n = n^{-1}(\alpha_n X_n + \xi_n c_n^{-1+\frac{1}{2\gamma}})\lambda_n,$$

with $\lambda_n = D_n[f^{c_n}(X_n)]^{-1}\Psi_n[f^{c_n}(X_n)] + (\log n)^{-1+\varepsilon_1}\kappa_n$, where in the proof of Theorem (2.10), it was shown that $\kappa_n \to 1$. So by (3.1.1) and (2.7.1) we have

$$(3.1.3) \qquad \lambda_n \to (2M)^{-1}.$$

Denoting conditional variance, given F_n, by Var^{F_n}, we have

$$\mathrm{Var}^{F_n}[h_n(Y_n)] = \int h_n^2(t + f^{c_n}(X_n))dG(t) - \Psi_n^2[f^{c_n}(X_n)] \to I(g)$$

by (2.8.1) and since $\Psi_n[f^{c_n}(X_n)] \to 0$. Therefore

$$(3.1.4) \qquad D_n^2 \,\mathrm{Var}^{F_n}[h_n(Y_n)] \to (2M)^{-2}I^{-1}(g).$$

Now consider $\mathrm{Var}^{F_n}[(\log n)^{-1+\varepsilon_1}\tilde{Y}_n]$. If y_n in (2.6.4) are $(\log n)^{1-2\varepsilon_1}$, this variance is bounded by $(\log n)^{-2\varepsilon_1}$. On the other hand, if $y_n = n^{\varepsilon_1}$ then G has finite

second moment, say σ^2, and $\operatorname{Var}^{F_n}\tilde{Y}_n \le E^{F_n}\tilde{Y}_n^2 \le E^{F_n}Y_n^2 \le \sigma^2 + [f^{c_n}(X_n)]^2$. So on the set $\{X_n \to 0\}$, $\operatorname{Var}^{F_n}\tilde{Y}_n \to 0$.
In either case then we have

$$\text{(3.1.5)} \qquad \operatorname{Var}^{F_n}[(\log n)^{-1+\varepsilon_1}\tilde{Y}_n] \to 0.$$

The random variables $h_n(Y_n)$ and $\tilde{Y}_n$ are not independent, but by the Schwarz inequality it follows from (3.1.4) and (3.1.5) that

$$\text{(3.1.6)} \qquad (nc_n)^2 E^{F_n}(U_n - N_n)^2 \to I^{-1}(g)(2M)^{-2} .$$

Now we set $Z_n = nc_n(U_n - N_n)$ and suppose r is in R^+. By (2.6.1), $Z_n = O_u((\log n)^{\varepsilon_0} n^{2\varepsilon_1})$. So $\{Z_n^2 > rn\}$ is eventually empty and

$$\text{(3.1.7)} \qquad E\, Z_n^2 \chi_{\{Z_n^2 \ge rn\}} \to 0 .$$

Writing X_{n+1} as $(X_n - N_n) - (U_n - N_n)$ we obtain $X_{n+1} = (1 - n^{-1}\alpha_n\lambda_n)X_n - n^{-1}c_n^{-1}Z_n - n^{-1}c_n^{-1+1/(2\gamma)}\xi_n\lambda_n$. Using this, (3.1.3), (3.1.6), (3.1.7), and the measurability properties of α_n, λ_n, and ξ_n we obtain the desired result by applying Theorem (2.2), Fabian (1968) with U_n in Theorem (2.2) replaced by X_n here, Γ_n by $\alpha_n\lambda_n$, V_n by Z_n, Φ_n by $-C_n^{-1}$, and T_n by $-C_n^{-1+1/(2\gamma)}\xi_n\lambda_n$.

For the case $\gamma = 1/4$, by (2.8.2), (2.10.5), and (3.1.3) we have, $\Phi = -C^{-1}$ and $T = 0$. Similarly for the case when $\gamma = 1/6$ and $D^3 f(0) = 0$. Finally, if $\gamma = 1/6$ and $D^3 f(0) \neq 0$, Φ is $-[(2/3)[D^3 f(0)]^2 I(g)]^{1/6}$ and T is $-(6M)^{-1}[(3/2)[D^3 f(0)]I^{-1}(g)]^{1/3}$.

In all cases, $\Gamma = 1$, $\alpha = 1$, $\beta = \beta_+ = 1 - 2\gamma$, and by (3.1.6), $\Sigma = I^{-1}(g)(2M)^{-2}$.

CHAPTER FOUR

THE MAIN RESULT

In this chapter we state and prove the main result, a realization of the procedure given in (2.6.3). Only (2.3) and (2.4) are assumed to hold. This result is given in the following theorem.

4.1 Theorem

Suppose Assumptions (2.3) and (2.4) hold with $\mathcal{F}_n$ as defined below. Let $\{k_\ell\}$ be an increasing sequence of positive integers such that $\ell/k_\ell \to 0$. Suppose $\{U_\ell\}$ and $\{V_\ell\}$ are sequences of random variables such that with

$$\mathcal{F}_n = \mathcal{F}(\{X_1, Y_1, \ldots, Y_{n-1}\} \cup \{U_\ell;\ k_\ell < n\} \cup \{V_\ell;\ k_\ell < n\}),$$

we have

$$(4.1.1)\quad \begin{cases} E^{\mathcal{F}_{k_\ell}} U_\ell = (2d_\ell)^{-2}(f^{d_\ell})^{d_\ell}(X_{k_\ell}) , \\ E^{\mathcal{F}_{k_\ell}}(U_\ell - E^{\mathcal{F}_{k_\ell}} U_\ell)^2 = O_u(d_\ell^{-4}) , \\ E^{\mathcal{F}_{k_\ell}} V_\ell = (2d_\ell)^{-3}((f^{d_\ell})^{d_\ell})^{d_\ell}(X_{k_\ell}), \\ E^{\mathcal{F}_{k_\ell}}(V_\ell - E^{\mathcal{F}_{k_\ell}} V_\ell)^2 = O_u(d_\ell^{-6}), \end{cases}$$

with d_ℓ of the form

(4.1.2) $\quad d_\ell = d\ell^{-\delta}$, d in R^+, $0 < \delta < 1/6$.

Then the sequence $\{X_n\}$ as defined in (4.2) below converges to θ and $t_n^{\frac{1}{2}-\gamma}(X_n - \theta)$ is asymptotically normal with mean and variance as given in Theorem (3.1), (i) and (ii), where $2t_n = 2n + 7$ card $\{\ell;\ k_\ell < n\}$ is the number of observations needed to construct X_n.

4.2 The Procedure

(i) Estimation of $D^2f(\theta)$:

Set $\bar{U}_n$ equal to the arithmetic mean of all U_ℓ with $k_\ell < n$. Then set

(4.2.1) $$u_n = (0 \vee \bar{U}_n).$$

(ii) Estimation of $D^3f(\theta)$:

Set v_n equal to the arithmetic mean of all V_ℓ with $k_\ell < n$.

(iii) Estimation of $I(g)$:

Let $\beta_0 > 0$ and choose $\varepsilon_n \geq (\log n)^{-\beta_0}$. Then this estimation is carried out precisely as it is in (I4.2.b), that is by a sequence $\{w_n\}$ with

(4.2.2) $$w_n = \int (h_n^o)^2 dG_{n-1}$$

where G_n is the empirical distribution function of $Y_1, Y_2, \ldots, Y_n$, and h_n^o is defined in (I4.2.3).

(iv) The sequence C_n:

Set

$$(4.2.3)\quad C_n = \begin{cases} [(3/2)v_n^{-2}w_n^{-1})^{1/6} \vee (\log n)^{-\varepsilon_0}] \wedge (\log n)^{\varepsilon_0}, & \text{if } \gamma = 1/6 \text{ and } D^3 f(\theta) = 0 \\ C & \text{, otherwise} \end{cases}$$

with ε_0 as in (2.6.1), C as in (2.8.2), v_n as in (4.2.ii) and w_n as in (4.2.2).

(v) The sequence D_n:

Set

$$(4.2.4)\qquad D_n = (2u_n w_n)^{-1} \wedge n^{\varepsilon_1}.$$

(vi) The functions h_n:

Choose $\varepsilon_n \geq n^{-\beta_1}$ for $n = 2, \ldots$. Then the estimators are constructed precisely as in (I4.2.c), but for $0 < \beta_1 < \frac{1}{2} - \gamma - 2\varepsilon_1$, with γ as in Assumption (2.3) and ε_1 as (2.6.1).

(vii) The sequence X_n:

The recursion relation for X_n is given in (2.6.3).

Proof of Theorem (4.1):

We shall prove the theorem by verifying Assumptions (2.3), (2.4), (2.6), (2.7), and (2.8).

First, Assumptions (2.3) and (2.4) are assumed to hold in the theorem. The measurability conditions on C_n, D_n, and h_n and condition (2.6.1) are obvious from their definitions. Relation (2.6.3) holds by assumption. Thus Assumption (2.6) holds, and by Theorem (2.10), $(\log n)^{\beta}(X_n - \theta) \to 0$, for every $\beta > 0$.

To show that u_n converges to $D^2 f(\theta)$, it suffices to show that $\bar{U}_n$ does. Let $W_\ell = U_\ell - E^{F_{k_\ell}} U_\ell$. Then W_ℓ is an orthogonal sequence,

$$\sum_{\ell=1}^{\infty} (\log \ell)^2 \ell^{-2} E\, W_\ell^2 \leq C_1 d^{-4} \sum_{\ell=1}^{\infty} (\log \ell)^2 \ell^{-2+4\delta} < +\infty$$

for a C_1 in R^+. So by Theorem (33.1.B.ii), Loève, we have $\ell^{-1} \sum_{j=1}^{\ell} W_j \to 0$. Also $E^{F_{k_\ell}} U_\ell = D^2 f(X_{k_\ell} + \nu_\ell)$, where $|\nu_\ell| \leq 2d_\ell$. So eventually, depending on ω, we obtain using Assumption (2.3) that $E^{F_{k_\ell}} U_\ell \to D^2 f(\theta)$ and $\bar{U}_n \to D^2 f(\theta)$. The convergence of w_n to $I(g)$ follows from Theorem (2.2), Fabian (1973b). Verification of the assumptions of this theorem are given in (I4.3.ii). Therefore Assumption (2.7) holds, and by our Theorem (2.11), $n^{\beta}(X_n - \theta) \to 0$, for every $0 < \beta < \frac{1}{2} - \gamma - 2\varepsilon_1$.

Finally, (2.8.1) follows from Extension (2.3), Fabian (1973b). Details and verification of the assumptions of this extension are given in (I4.3.iii). The convergence of v_n to $D^3f(\theta)$ follows by an argument similar to that used to show $\bar{U}_n \to D^2f(\theta)$. Therefore Assumption (2.8) holds, and by our Theorem (3.1), X_n has the properties asserted in Theorem (4.1) since $t_n/n \to 1$ because $\ell/k_\ell \to 0$.

REFERENCES

REFERENCES

ABDELHAMID, S.N. (1973). Transformation of observations in stochastic approximation. Ann. Statist. 1 1158-1174.

ANBAR, D. (1973). On optimal estimation methods using stochastic approximation procedures. Ann. Statist. 1 1175-1184.

BLUM, J.R. (1954). Approximation methods which converge with probability one. Ann. Math. Statist. 25 737-744.

DUBINS, L.E. and FREEDMAN, D.A. (1965). A sharper form of the Borel-Cantelli lemma and the strong law. Ann. Math. Statist. 36 800-807.

FABIAN, V. (1960). Stochastic approximation methods. Czechoslovak Math. J. 10(85) 123-159.

FABIAN, V. (1964). A new one-dimensional stochastic approximation method for finding a local minimum of a function. Trans. Third Prague Conf. Information Theory, Statistical Decision Functions, Random Processes, Czechoslovak Academy of Sciences, Prague, 85-105.

FABIAN, V. (1968). On asymptotic normality in stochastic approximation. Ann. Math. Statist. 39 1327-1332.

FABIAN, V. (1971). Stochastic Approximation. Optimizing Methods in Statistics (J.S. Rustagi, ed.). 439-470. Academic Press, New York.

FABIAN, V. (1973a) (Referred to in the text as I). Asymptotically efficient stochastic approximation; the RM case. Ann. Statist. 1 486-495.

FABIAN, V. (1973b). Estimation of the derivative of the logarithm of a density. Ann. Statist. 1 557-561.

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. Ann. Math. Statist. 23 462-466.

LOÈVE, MICHEL (1963). Probability Theory, Third Edition. D. van Nostrand Company, Inc.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. Ann. Math. Statist. 22 400-407.

VENTER, J.H. (1967). An extension of the Robbins-Monro procedure. Ann. Math. Statist. 38 181-190.