STOCHASTIC APPROXIMATION OF MINIMA WITH IMPROVED ASYMPTOTIC SPEED

By Václav Fabian

Czechoslovak Academy of Sciences, Prague

1. Summary. It is shown that the Keifer-Wolfowitz procedure—for functions f sufficiently smooth at θ, the point of minimum—can be modified in such a way as to be almost as speedy as the Robbins-Monro method. The modification consists in making more observations at every step and in utilizing these so as to eliminate the effect of all derivatives $\partial^j f/[\partial x^{(i)}]^j$, $j = 3, 5 \cdots, s - 1$. Let δ_n be the distance from the approximating value to the approximated θ after n observations have been made. Under similar conditions on f as those used by Dupač (1957), the result is $E\delta_n^2 = O(n^{-s/(s+1)})$. Under weaker conditions it is proved that $\delta_n^2 n^{s/(s+1)-\epsilon} \to 0$ with probability one for every $\epsilon > 0$. Both results are given for the multidimensional case in Theorems 5.1 and 5.3. The modified choice of Y_n in the scheme $X_{n+1} = X_n - a_n Y_n$ is described in Lemma 3.1. The proofs are similar to those used by Dupač (1957) and are based on Chung's (1954) lemmas and, in Theorem 5.3, on a modification of one of these lemmas. The result of Theorem 5.3 is new also for the usual Kiefer-Wolfowitz procedure. The main and very simple idea, however, is in Lemma 3.1; it will suggest, to a reader acquainted with Dupač's Theorem 3 and its proof, the consequences elaborated in Theorem 5.1.

2. Introduction. The results concerning the speed of stochastic approximation methods show a difference between the Robbins-Monro (RM) procedure and the Kiefer-Wolfowitz (KW) procedure. One has $E\delta_n^2 = O(n^{-1})$ under rather general conditions for the RM procedure. (See the review of Schmetterer (1961) for all results mentioned here without reference.) For the KW procedure the situation is more complicated and worse. The behaviour of δ_n depends on how the minimized function behaves in the neighborhood of the point θ of minimum. Thus for any one-dimensional function f which has a third derivative f''' in a neighborhood of θ and satisfies some other regularity conditions, a suitable choice of the constants a_n, c_n of the KW procedure gives $E\delta_n^2 \leqq C_2 n^{-2/3}$ for some constant C_2. If $f'''(\theta) \neq 0$ then for any a_n, c_n, conversely, $E\delta_n^2 \geqq C_1 n^{-2/3}$ for some positive C_1 and all n greater than some n_0 (Dupač (1957); his results have been generalized to the multidimensional case by Sakrison (1962)). If f is approximately even in a neighborhood of θ, then a suitable adjustment of the c_n's and a_n's gives higher speed (see Remark 3.4). This result, however, is more of theoretical than practical interest because the optimal values of c_n, a_n depend on the quantitative assumption of local eveness of f.

The possibility of taking at every step more observations than necessary has already been investigated by Burkholder (1956), Block (1957), Cochran and

Received 11 February 1966; revised 2 August 1966.

Davis (1965) (the latter two papers being restricted to the RM procedure) but it has never been used to give higher asymptotic speed for the KW procedure. It may be said that the modification goes a step towards more traditional methods (such as those of Box and Wilson (1951)) which, at every step, explore the function f in greater detail than the original KW procedure. It retains, however, the strict determinism of the KW process, thus making possible a rigorous study of its properties.

The greater complexity of the modified KW procedure opens many questions which will be mentioned briefly in Remark 3.2.

Throughout the paper it is assumed that f is a real function defined on k-dimensional Euclidean space R^k; θ is a point in R^k, $C(\epsilon)$ the closed sphere $\{x; \|x - \theta\| \leqq \epsilon\}$. It is supposed that f has a derivative D on R^k; i.e. $D^{(i)}(x) = \partial f(x)/\partial x^{(i)}$, and if it has a Hessian at x, it will be denoted by $H(x)$, $H^{(ij)}(x)$ being $\partial^2 f(x)/(\partial x^{(i)}\partial x^{(j)})$. The vector of the jth derivatives of f at x along the individual coordinates is denoted by $D_j(x)$, i.e. $D_j^{(i)}(x) = \partial^j f(x)/\partial(x^{(i)})^j$. For $i = 1, 2, \cdots, p$, $e_{i,p}$ are the vectors in R^p such that $e_{i,p}^{(j)} = \delta_{ij}$ for $j = 1, \cdots, p$ where δ_{ij} is the Kronecker symbol. If $p = k$, the subscript p may be dropped and $e_{i,k}$ written simply as e_i. The vector of the first differences of f at x with step c is denoted by $d(x, c)$, its ith coordinate being $f(x + ce_i) - f(x - ce_i)$. The notations $g(n) = O(h(n))$ and $g(n) = O^{-1}(h(n))$ mean lim sup $|g(n)/h(n)| < +\infty$ and lim inf $|g(n)/h(n)| > 0$, respectively. Relations between random variables are meant with probability one. If T is a random vector, E_T denotes conditional expectation given T. Throughout the paper a_n and c_n are positive numbers converging to zero. Components of vectors and matricies are referred to by superscripts in an obvious way.

We shall consider k-dimensional random vectors $X_1, X_2, \cdots, Y_1, Y_2, \cdots$ satisfying the relation

$$X_{n+1} = X_n - a_n Y_n . \tag{2.1}$$

We shall write $\mathbf{X}_n$ for $[X_1, X_2, \cdots, X_n]$; similarly for $\mathbf{Y}_n$ and so on.

3. The choice of Y_n. The stochastic approximation of a minimum of f is based on the choice of Y_n such that $E_{\mathbf{X}_n} Y_n$ is approximately $D(X_n)$. In the usual KW procedure, $E_{\mathbf{X}_n} Y_n = (2c_n)^{-1} d(X_n, c_n)$; of course $Y_n^{(i)}$ is usually $(2c_n)^{-1}$ times the difference between estimates of $f(X_n + c_n e_i)$ and $f(X_n - c_n e_i)$. The following lemma describes a more general possibility of constructing Y_n.

LEMMA 3.1. *Let s be an even positive integer, $\epsilon > 0$, and let D_{s+1} exist on $C(2\epsilon)$ and be bounded there. Let u_i be numbers, $0 < u_1 < \cdots < u_m \leqq 1$, $m = s/2$,*

$$U = \|u_j^{2i-1}\|_{i,j=1}^{m}, \qquad v = \tfrac{1}{2} U^{-1} e_{1,m} .$$

(It is well known from elementary algebra that U is non-singular.) Set $v_i = v^{(i)}$ for $i = 1, 2, \cdots, m$.

Let $Y_{n,i}$, $i = 1, 2, \cdots, m$, be k-dimensional random vectors such that the $m \times k$ componets $Y_{n,i}^{(j)}$ are conditionally independent (given $\mathbf{X}_n$) and

$$E_{\mathbf{X}_n} Y_{n,i} = d(X_n, c_n u_i),\ E_{\mathbf{X}_n}[Y_{n,i}^{(j)} - d^{(j)}(X_n, c_n u_i)]^2 \leqq 2\sigma^2. \tag{3.1.1}$$

Set $Y_n = c_n^{-1} \sum_{i=1}^m v_i Y_{n,i}$, $Z_n = Y_n - E_{\mathbf{X}_n} Y_n$, *and* $M_n(x) = c_n^{-1} \sum_{i=1}^m v_i\, d(x, c_n u_i)$. *Then there exists a number* K *such that for every* n *the following assertions hold:*

$$E_{\mathbf{X}_n} Y_n = M_n(X_n); \tag{3.1.2}$$

for every $x \in R^k$ *there are* ξ_i *such that*

$$M_n(x) = 2 \sum_{i=1}^m v_i u_i D(\xi_i),\ \|\xi_i - x\| < c_n\ ; \tag{3.1.3}$$

for every $x \in C(\epsilon)$

$$M_n(x) = D(x) + Q_n(x) c_n^s \tag{3.1.4}$$

where $\|Q_n(x)\| \leqq K$ *if* $c_n < \epsilon$;

$$E_{\mathbf{X}_n} \|Z_n\|^2 \leqq 2k\sigma^2 c_n^{-2} \|v\|^2 = \tfrac{1}{2}\sigma^2 c_n^{-2} [(UU')^{-1}]^{(1,1)}. \tag{3.1.5}$$

PROOF. (3.1.2) and (3.1.5) are obvious, (3.1.3) follows from the fact that $d(x, c) = 2cD(\xi)$ with $\|\xi - x\| < c$. To prove (3.1.4) we may assume $x \in C(\epsilon)$, $c_n < \epsilon$. Fix a j and put $h(c) = \sum_{i=1}^m v_i\, d^{(j)}(x, cu_i)$. The function h has derivatives up to the order $s + 1$ for $|c| \leqq \epsilon$, the pth derivative h_p being

$$h_p(c) = \sum_{i=1}^m v_i u_i^p [D_p^{(j)}(x + cu_i e_j) - (-1)^p D_p^{(j)}(x - cu_i e_j)].$$

Hence, using the relation $Uv = \tfrac{1}{2} e_{1,m}$, it follows that $h_p(0) = 0$ for $p = 0, 2, 3, \cdots, s$ and $h_1(0) = D^{(j)}(x)$. Since $M_n^{(j)}(x) = c_n^{-1} h(c_n)$, (3.1.4) holds with

$$Q_n^{(j)}(x) = ((s + 1)!)^{-1} \sum_{i=1}^m v_i u_i^{s+1} [D_{s+1}^{(j)}(x + \delta u_i e_j) + D_{s+1}^{(j)}(x - \delta u_i e_j)]$$

and with positive $\delta = \delta_j(x) < c_n$. Since D_{s+1} is bounded on $C(2\epsilon)$, the supremum of $\|Q_n(x)\|$ over all $x \in C(\epsilon)$ and $c_n < \epsilon$ is finite.

REMARK 3.2. Neither the question of how to choose the vector u nor the possibility of varying s with n are considered here. Other questions seem to be interesting too: Can one obtain from the additional observations some further information concerning σ and H, and then use this information to improve the choice of a_n, c_n in the subsequent steps? This cannot improve the order of convergence, but it may well (i) dispose of the conditions $2aK_0 > \beta$ and $2a\lambda_0 > \beta_0$ for the optimal $\alpha = 1$ in Theorems 5.1 and 5.2; (ii) guarantee that, under more general conditions on f, X_n does not converge to a local maximum instead of minimum (in the one-dimensional case a method avoiding unwanted extrema was proposed by Fabian (1964), but also shown by Vosiková (1964) to be quite slow near the point of minimum); and finally (iii) improve the speed in the sense that for small σ the speed is comparable with that of good deterministic methods.

EXAMPLE 3.3. Consider the design $u = [\frac{1}{2}, 1]$ leading to $v = [\frac{4}{3}, -\frac{1}{6}]$. If $k = 1$ then $Y_n = c_n^{-1} \cdot [(\frac{4}{3}) Y_{n,1} - (\frac{1}{6}) Y_{n,2}]$ where $Y_{n,1}$ and $Y_{n,2}$ are estimates of values $f(x + \frac{1}{2}c_n) - f(x - \frac{1}{2}c_n)$, $f(x + c_n) - f(x - c_n)$. By (3.1.5) the variance of Y_n will then be at most $65 c_n^{-2} \sigma^2 / 18$.

REMARK 3.4. Suppose f is one-dimensional, D_p exists and is bounded in a neighborhood of θ for a $p \geqq 2$, $D_2(\theta) \neq 0$ and $s \leqq p - 1$. Then it is possible to show that f is $(s - 1)$-locally even in Burkholder's (1956) sense if and only if $D_i(\theta) = 0$ for every odd i, $3 \leqq i \leqq s$. An analogous condition of Sacks (1958) is equivalent to the requirement $D_i(\theta) = 0$ for every $3 \leqq i \leqq s$.

Under some additional assumptions, Burkholder (1956) proved that if f is $(s - 1)$-locally even then for every $\beta < \frac{1}{2}s/(s + 1)$ there are a_n, c_n such that $n^\beta(X_n - \theta)$ is asymptotically normal. Sacks (1958) proved a slightly stronger result under his condition of local evenness. This order of convergence corresponds to that obtained in Theorems 5.1 and 5.3 for our choice of Y_n which eliminates the effect of $D_i(\theta)$ for $i = 3, 5, \cdots, s - 1$.

REMARK 3.5. If M_n is as in Lemma 3.1, f has bounded partial derivatives up to order $s + 2$ on $C(2\epsilon)$, $D_{s+1}(\theta) \neq 0$, and x_n is a sequence of numbers converging to θ, then

$$M_n(x_n) = D(x_n) + c_n{}^s N(x_n) + O(c_n{}^{s+1}) \tag{3.5.1}$$

for a vector valued continuous function N on $C(\epsilon)$, $N(\theta) \neq 0$. This is easily seen by expanding the function h, defined in the proof of Lemma 3.1, up to order $s + 2$ and using the boundedness of D_{s+2} on $C(2\epsilon)$. One then obtains (3.5.1) with $N(x) = 2((s + 1)!)^{-1} \sum_{i=1}^m v_i u_i{}^{s+1} D_{s+1}(x)$. $N(\theta)$ is non-zero because the $(s/2)$-dimensional vectors $u_i{}^3, u_i{}^5, \cdots, u_i{}^{s+1}$, $i = 1, 2, \cdots, m = s/2$, are linearly independent.

COROLLARY 3.6. *Under the assumptions of Lemma* 3.1 *for a positive* K_0 *and every* $x \in R^k - C(\epsilon/2)$ *let*

$$g'(x)D(x) \geqq K_0 \|g(x)\|^2 \tag{3.6.1}$$

where g *is a mapping from* $R^k - C(\epsilon/2)$ *to* R^k, $\|g\|$ *is bounded from below by a positive constant and such that* $D(\xi)\|g(x)\|^{-1}$ *converges to* $D(x)\|g(x)\|^{-1}$ *uniformly in* $x \in R^k - C(\epsilon/2)$ *as* $\xi \to x$. *Then to every* $\eta > 0$ *there is an* n_0 *such that, for every* $n \geqq n_0$ *and every* $x \in R^k - C(\epsilon)$,

$$g'(x)M_n(x) \geqq (1 - \eta)K_0 \|g(x)\|^2. \tag{3.6.2}$$

PROOF. If n_0 is sufficiently large and $n \geqq n_0$ then $c_n < \epsilon/2$ and $\|\xi - x\| < c_n$ implies $\|D(\xi) - D(x)\| < \eta \|g(x)\| K_0(2 \sum_{i=1}^m |u_i v_i|)^{-1}$ for $x \in R^k - C(\epsilon)$. Then, with $\|\xi_i - x\| < c_n$, (3.1.3) yields $g'(x)M_n(x) \geqq g'(x)D(x) - \|g(x)\| 2 \cdot \sum_{i=1}^m |u_i v_i| \, \|D(\xi_i) - D(x)\| \geqq g'(x)D(x) - \eta K_0 \|g(x)\|^2$ and (3.6.2) follows from (3.6.1).

REMARK 3.7. The condition $D(\xi)\|g(x)\|^{-1} \to D(x)\|g(x)\|^{-1}$ uniformly in $x \in R^k - C(\epsilon/2)$ for $\xi \to x$ is, of course, weaker than the uniform continuity of D because of the further assumption that $\inf \{\|g(x)\|; x \in R^k - C(\epsilon/2)\} > 0$.

4. Preparatory relations and results.

LEMMA 4.1. *Let* f, Y_n *be as in Lemma* 3.1, a_n, c_n *be positive,* $a_n \to 0$, $c_n \to 0$, *let there be positive numbers* K_0, K_1 *and a subset* $C \subset R^k$ *such that* $C \supset C(\epsilon)$ *and that*

for every $x \in C$,

$$K_0 \|x - \theta\|^2 \leqq (x - \theta)'D(x),\ \|D(x)\| \leqq K_1 \|x - \theta\|. \tag{4.1.1}$$

Let $D(\xi)\|x - \theta\|^{-1}$ *converge to* $D(x)\|x - \theta\|^{-1}$ as $\xi \to x$, *uniformly for* $x \in C - C(\epsilon/2)$. *Then for every* $\eta > 0$ *there are positive numbers* n_0 *and* Q *such that, for* $n \geqq n_0$ *and for* $X_n \in C$,

$$\|X_{n+1} - \theta\|^2 \leqq \|X_n - \theta\|^2[1 - a_n(2 - \eta)K_0] + Qa_nc_n^{2s} + U_n + V_n^2 \tag{4.1.2}$$

where U_n *and* V_n *are random variables satisfying*

$$\begin{aligned} &E_{\mathfrak{U}_{n-1}}U_n = 0,\ E_{\mathfrak{U}_{n-1}}U_n^2 \leqq 16k\sigma^2 \|v\|^2 a_n^2c_n^{-2} \\ &\qquad \cdot[\|X_n - \theta\|^2(1 + \eta) + \eta a_nc_n^{2s}], \\ &E_{\mathfrak{V}_{n-1}}V_n^2 \leqq 2k\sigma^2 \|v\|^2 a_n^2c_n^{-2}. \end{aligned} \tag{4.1.3}$$

If $C = R^k$ *and* $E\|Z_n\|^2 \geqq dc_n^{-2}$, $d > 0$, *then*

$$E\|X_{n+1} - \theta\|^2 \geqq [1 - a_nQ]E\|X_n - \theta\|^2 - a_nc_n^{2s}Q + a_n^2c_n^{-2}d. \tag{4.1.4}$$

PROOF. Without loss of generality one may assume that $\theta = 0$, $\eta < \frac{1}{8}$. We assume also that $n \geqq n_0$ for some n_0 such that $c_n < \epsilon/2$ (we shall occasionally increase n_0 to meet further requirements). Thus if $x \in C - C(\epsilon)$ and $\|\xi_i - x\| < c_n$ then $\|\xi_i\| \leqq 2\|x\|$. Hence (3.1.3) and (4.1.1) imply that $\|M_n(x)\| \leqq K_3\|x\|$ for all $x \in C - C(\epsilon)$ with $K_3 = 4K_1 \sum_{i=1}^m |u_iv_i|$. Combining this with (3.1.4) and (4.1.1) one obtains for all $x \in C$

$$\|M_n(x)\| \leqq K_4\|x\| + Kc_n^s \tag{4.1.5}$$

with $K_4 = \max\{K_1, K_3\}$, and, adjusting n_0 if necessary,

$$a_n^2\|M_n(x)\|^2 \leqq \tfrac{1}{4}(a_nK_0\eta\|x\|^2 + \eta a_nc_n^{2s}) \tag{4.1.6}$$

for all $x \in C$.

As a further step we shall establish

$$x'M_n(x) \geqq (1 - \eta/8)K_0\|x\|^2 - Kc_n^s\|x\|. \tag{4.1.7}$$

For $x \in C - C(\epsilon)$ this follows from Corollary 3.6, applied to $g(x) = x$; for $x \in C(\epsilon)$ it is a consequence of (3.1.4) and (4.1.1). Finally, from (4.1.7) and (4.1.5) one obtains, applying the inequality $\|x\| \leqq d\|x\|^2 + d^{-1}$ with $d = K_0/(8Kc_n^s)$, that

$$(1 - \eta/4)K_0\|x\|^2 - K_5c_n^{2s} \leqq x'M_n(x) \leqq K_5\|x\|^2 + K_5c_n^{2s} \tag{4.1.8}$$

for a suitably chosen K_5.

Now from the relation $X_{n+1} = X_n - a_nM_n(X_n) - a_nZ_n$ if follows that

$$\|X_{n+1}\|^2 = \|X_n\|^2 - 2a_nX_n'M_n(X_n) + a_n^2\|M_n(X_n)\|^2 + U_n + V_n^2 \tag{4.1.9}$$

where $U_n = -2a_nZ_n'(X_n - a_nM_n(X_n))$, $V_n^2 = a_n^2\|Z_n\|^2$. Now using (4.1.6) and (4.1.8) one gets (4.1.2) with $Q = 2K_5 + \eta$. Concerning U_n one clearly has

$E_{\mathbf{X}_n} U_n = 0$. Now $\|X_n - a_n M_n(X_n)\|^2 \leqq 2\|X_n\|^2 + 2a_n^2 \|M_n(X_n)\|^2$ which is not greater than $2\|X_n\|^2(1 + O(a_n)) + \eta a_n c_n^{2s}$ for $X_n \in C$ according to (4.1.6). Because $\mathbf{U}_{n-1}$ is functionally dependent on $\mathbf{X}_n$, the second relation in (4.1.3) holds. The inequality for $E_{\mathbf{V}_{n-1}} V_n^2$ in (4.1.3) follows immediately from (3.1.5).

If $C = R^k$ and $E\|Z_n\|^2 \geqq d c_n^{-2}$, then (4.1.4) follows easily from (4.1.9) and (4.1.8) with a suitable positive Q.

The following lemma will be used in the proof of Theorem 5.1. The lemma summarizes Chung's (1954) Lemmas 1 to 4, Lemma 1 being included in a weaker form than the original. In 1964, Dupač and Vosiková observed and communicated to the present author that the original proof of Chung's Lemma 1 is incorrect (on page 466 the inequality following the sentence "Similarly but more roughly, for some $c_3 > 0 \cdots$" need not hold).

LEMMA 4.2 (Chung). *Let b_n, A_n, D_n, α, β, B be real numbers, let $a_0 = \liminf A_n$ and $a_1 = \limsup A_n$ be finite and*

$$b_{n+1} = b_n(1 - A_n n^{-\alpha}) + B n^{-\alpha-\beta} + D_n \tag{4.2.1}$$

for all sufficiently large n and $0 < \alpha \leqq 1$, $0 < \beta$, $0 < B$. Set $C_i = a_i$ if $\alpha < 1$ and $C_i = a_i - \beta$ if $\alpha = 1$.

If the D_n's are non-positive and $C_0 > 0$ then

$$\limsup n^\beta b_n \leqq B/C_0\,; \tag{4.2.2}$$

if they are non-negative and $C_1 > 0$ then

$$\liminf n^\beta b_n \geqq B/C_1\,. \tag{4.2.3}$$

PROOF. Let ϵ be a positive number, $\epsilon < C_0$ if C_0 is positive. Put $\xi_n = n^\beta b_n$. Observe that $(1 + n^{-1})^\beta = 1 + \beta_n n^{-1}$ with $\beta_n \to \beta$ and $(1 + n^{-1})^\beta(1 - A_n n^{-\alpha}) = 1 + \beta_n n^{-1} - A_n n^{-\alpha} + O(n^{-\alpha-1}) = 1 - \gamma_n n^{-\alpha}$ with $C_0 - \epsilon < \gamma_n < C_1 + \epsilon$ for all $n \geqq n_0$ and some n_0. Multiplying (4.2.1) by $(n+1)^\beta = n^\beta(1 + n^{-1})^\beta$ and writing $Q_n = (n+1)^\beta D_n$ one obtains, with $B_n \to B$,

$$\xi_{n+1} = \xi_n(1 - \gamma_n n^{-\alpha}) + B_n n^{-\alpha} + Q_n\,. \tag{4.2.4}$$

Now suppose $D_n \leqq 0$, $C_0 > 0$ and let n_1 be such that $n_1 \geqq n_0$, $B_n < B + \epsilon$, $\gamma_n n^{-\alpha} < 1$ for all $n \geqq n_1$. Let $n \geqq n_1$. If $\xi_n \geqq (B + 2\epsilon)/(C_0 - \epsilon)$ then $\xi_{n+1} \leqq \xi_n - \epsilon n^{-\alpha}$. Since the right hand side of (4.2.4) is an increasing function of ξ_n, the last inequality implies that $\xi_n \leqq (B + 2\epsilon)/(C_0 - \epsilon)$ if $\xi_n \leqq (B + 2\epsilon)/(C_0 - \epsilon)$. This shows that $\limsup \xi_n \leqq (B + 2\epsilon)/(C_0 - \epsilon)$. Since ϵ was arbitrarily small and positive, (4.2.2) holds. The proof of (4.2.3) is entirely analogous.

LEMMA 4.3. *Let b_n be numbers satisfying*

$$b_{n+1} \leqq b_n(1 - An^{-1}) + B_n n^{-\beta} \tag{4.3.1}$$

where $A > \beta$, $\beta > 0$ and $\sum_{n=1}^\infty B_n$ converges. Then

$$\limsup n^\beta b_n < +\infty. \tag{4.3.2}$$

PROOF. Put $\xi_n = n^\beta b_n$ so that, as in the preceding proof,

$$\xi_{n+1} \leqq (1 - \gamma_n{}^{n-1})\xi_n + (1 + \beta_n n^{-1})B_n$$

with $\gamma_n \to A - \beta$, $\beta_n \to \beta$. For n sufficiently large, $0 < \gamma_n n^{-1} < 1$, $\beta_n B_n \leqq \gamma_n$ and $\xi_{n+1} \leqq \xi_n + B_n$ if $\xi_n \geqq 1$; and $\xi_{n+1} \leqq 1 + B_n$ if $\xi_n \leqq 1$. Hence $\xi_{n+1} \leqq \max\{1, \xi_n\} + B_n$ for all sufficiently large n which implies $\limsup \xi_n < +\infty$.

5. Two theorems. To make possible a simple appraisal of our choice of Y_n, the first situation we consider is analogous to that of Dupač's (1957) Theorem 4. This theorem gives a rather definitive result in the one-dimensional case on the asymptotic behaviour of $E|X_n - \theta|^2$ for the class of functions whose first derivative lies between two straight lines and whose third derivative exists and is bounded. Of course in our theorem we must suppose f has derivatives up to order $s + 1$ with $s \geqq 2$. Boundedness of the third derivative is not assumed but we do suppose that the continuity condition from Corollary 3.6 is satisfied for $g(x) = x - \theta$. The theorem is formulated directly for the multidimensional case. Conditions on α and γ are weaker than in Dupač's paper in that they do not imply $\sum a_n c_n < +\infty$, $\sum a_n{}^2 c_n{}^{-2} < +\infty$. The proof of the positive part is similar to that of Dupač with the difference that no preliminary result on the convergence of $E\|X_n - \theta\|^2 \to 0$ is needed.

THEOREM 5.1. *Let f, Y_n be as in Lemma 3.1, let $D(\xi)\|x - \theta\|^{-1}$ converge to $D(x)\|x - \theta\|^{-1}$ as $\xi \to x$, uniformly in $x \in R^k - C(\epsilon/2)$, let there be two positive numbers K_0 and K_1 such that, for every $x \in R^k$,*

$$K_0\|x - \theta\|^2 \leqq (x - \theta)'D(x), \quad \|D(x)\| \leqq K_1\|x - \theta\|, \tag{5.1.1}$$

and let

$$a_n = an^{-\alpha}, \qquad c_n = cn^{-\gamma}, \tag{5.1.2}$$

$a > 0$, $c > 0$, $0 < \alpha \leqq 1$, $0 < \gamma < \alpha/2$, *and* $2K_0 a > \beta$ *if* $\alpha = 1$, *where*

$$\beta = \min\{2s\gamma, \alpha - 2\gamma\}. \tag{5.1.3}$$

Then

$$E\|X_n - \theta\|^2 = O(n^{-\beta}). \tag{5.1.4}$$

This result cannot be improved within the class of functions considered, in the sense that to any a_n, c_n satisfying (5.1.2) there exist f and Y_n satisfying all the conditions stated above and such that

$$\limsup n^\beta E\|X_n - \theta\|^2 > 0. \tag{5.1.5}$$

REMARK 5.2. Set

$$\gamma_\alpha = \tfrac{1}{2}\alpha/(s + 1), \qquad \beta_\alpha = \alpha s/(s + 1); \tag{5.2.1}$$

then, for α fixed, β is minimal and equal to β_α if $\gamma = \gamma_\alpha$. When α is optional too,

β is maximal for $\alpha = 1$ and $\gamma = \frac{1}{2}(s+1)^{-1}$; this choice gives $\beta = s/(s+1)$. Unfortunately, for $\alpha = 1$, a has to be greater than $\beta(2K_0)^{-1}$, a number which is usually unknown.

PROOF OF THEOREM 5.1 Conditions of Lemma 4.1 are satisfied with $C = R^k$. Choose $\eta > 0$ so that $(2-\eta)aK_0 > \beta$ if $\alpha = 1$ and $\eta = 1$ otherwise. Because of (5.1.3) it follows from (4.1.2) and (4.1.3) that for n sufficiently large, the values $b_n = E_{X_n}\|X_n - \theta\|^2$ satisfy the relation (4.2.1) for some $C_0 > 0$ and for non-positive D_n's. According to Lemma 4.2, (5.1.4) holds.

To prove the second part of the theorem, assume first of all that $\gamma > \gamma_\alpha$. If $E\|Z_n\|^2 \geqq dc_n^{-2}$, $d > 0$ then according to Lemma 4.1 the inequality (4.1.4) holds for sufficiently large n. Since $\alpha + 2s\gamma > 2\alpha - 2\gamma = \alpha + \beta$, (4.2.1) holds with some $C_1 > 0$ and with non-negative D_n's and (5.1.5) follows according to Lemma 4.2.

Secondly assume that $\gamma \leqq \gamma_\alpha$, $\sigma^2 = 0$, and, without loss of generality, $\theta = 0$. Suppose f has bounded partial derivatives of order $s+2$ on $C(\epsilon)$, $D_{s+1}(0) \neq 0$. One has $\beta = 2\gamma s$, $X_n \to 0$ by the first and already proved part of the theorem. According to Remark 3.5 one obtains that

$$X_{n+1} = X_n - an^{-\alpha}(D(X_n) + n^{-\beta/2}M_n)$$

with $M_n \to c^s N(0) = M$ (say); $M \neq 0$.

In the one-dimensional case we may assume, without loss of generality, that $M < 0$. Then, according to (5.1.1) and with $aK_0 \leqq A_n \leqq aK_1$,

$$X_{n+1} \geqq X_n(1 - A_n n^{-\alpha}) - \tfrac{1}{2}aMn^{-\alpha-\beta/2}$$

for sufficiently large n, and Lemma 4.2 gives $\liminf n^{\beta/2}X_n > 0$ which implies (5.1.5).

In the multidimensional case we shall suppose $\|X_n\| \leqq K_3 n^{-\beta/2}$ for $n \geqq n_0$ and show that for a suitably chosen positive K_3 this leads to a contradiction. Put $\xi_n = n^{\beta/2}X_n$. Then $\|\xi_n\| \leqq K_3$ and $\|D(X_n)\| \leqq K_1K_3n^{-\beta/2}$ for $n \geqq n_0$. Choose j such that $M^{(j)} \neq 0$, assume without loss of generality that $M^{(j)} < 0$ and set $Q = -\frac{1}{4}aM^{(j)}$. For sufficiently large n

$$\xi_{n+1}^{(j)} \geqq (1+n^{-1})^{\beta/2}[\xi_n^{(j)} + n^{-\alpha}(2Q - aK_1K_3)].$$

Since $(1+n^{-1})^{\beta/2}\xi_n^{(j)} \geqq \xi_n^{(j)} - \frac{1}{2}\beta n^{-1}|\xi_n^{(j)}| \geqq \xi_n^{(j)} - \beta n^{-\alpha}K_3$, it follows that

$$\xi_{n+1}^{(j)} = \xi_n^{(j)} + n^{-\alpha}(2Q - 2aK_1K_3 - \beta K_3) \geqq \xi_n^{(j)} + n^{-\alpha}Q$$

for n sufficiently large, where K_3 is so chosen that $K_3(2aK_1 + \beta) < Q$. Hence $\xi_n^{(j)} \to +\infty$ which is the desired contradiction. This completes the proof of the second part of the theorem.

THEOREM 5.3. *Let f and Y_n be as in Lemma* 3.1 *and let the following additional conditions hold*:

(i) *The Hessian H exists, is bounded in norm by a constant K_1 on R^k and at θ it is positive definite and continuous*; $D(\theta) = 0$.

(ii) *To every $\epsilon > 0$ there is a $\rho(\epsilon) > 0$ such that $f(x) - f(\theta) \geqq \rho(\epsilon)$ and $\|D(x)\| \geqq \rho(\epsilon)$ for $x \in R^k - C(\epsilon)$.*

(iii) $a_n = an^{-1}$, $c_n = cn^{-\gamma}$, $a > 0$, $c > 0$, $0 < \gamma < \frac{1}{2}$ *and* $2\lambda_0 a > \beta_0$ *where* λ_0 *is the smallest eigenvalue of* $H(\theta)$

$$\beta_0 = \min\{2s\gamma, 1 - 2\gamma\}. \tag{5.3.1}$$

Then $n^\beta(X_n - \theta) \to 0$ *with probability one for every* $\beta < \beta_0/2$.

PROOF. Assume again that $\theta = 0$. From the boundedness of H it follows that D is uniformly continuous. Thus Corollary 3.6 may be applied to $g = D$, yielding $D'(x)M_n(x) \geqq \frac{1}{2}\|D(x)\|^2$ for every $x \varepsilon R^k - C(\epsilon)$ and all sufficiently large n. This together with (3.1.4) and (3.1.5) makes it possible to prove $X_n \to 0$ with probability one, either by a trivial modification of Blum's (1954) proof of his Theorem 3, or by a direct application of its generalization given by Fabian ((1960), Theorem 5.2).

By the assumptions, $D(x) = x'H(\xi(x))$ where $\|\xi(x)\| < \|x\|$. Choose η positive and such that $a(2 - 3\eta)\lambda_0 > \beta_0$. By continuity of H at θ, there is an ϵ_0 such that for $\|x\| < \epsilon_0$, $x'D(x) = x'H(\xi(x))x > (1 - \eta)\lambda_0\|x\|^2$. Of course $\|D(x)\| \leqq \|x\| \sup_{\xi \varepsilon R^k} \|H(\xi)\|$ for all x. We may assume $\epsilon_0 = \epsilon$. Then all the conditions of Lemma 4.1 are satisfied with $K_0 = (1 - \eta)\lambda_0$, $C = C(\epsilon)$, so that (4.1.2) and (4.1.3) hold for $X_n \varepsilon C(\epsilon)$ and all sufficiently large n. Now suppose that $\|X_n\|n^{\Delta/2}$ is bounded, i.e. lim sup $\|X_n\|n^{\Delta/2}$ is finite (with probability one), for a Δ such that $0 \leqq \Delta < \beta_0$. This is surely true at least for $\Delta = 0$. Then

$$E_{U_{n-1}}U_n^2 \leqq Qn^{2\gamma-2}[n^{-\Delta} + n^{-1-2\gamma s}] \leqq Qn^{-1-\beta_0-\Delta}$$

with Q possibly depending on elementary events. Choose δ such that $0 < \delta < \frac{1}{4}(\beta_0 - \Delta)$, put $\beta = \frac{1}{2}(\beta_0 + \Delta) - \delta$; thus $\beta - \Delta = \frac{1}{2}(\beta_0 - \Delta) - \delta > \frac{1}{4}(\beta_0 - \Delta)$ and $-\beta_0 - \Delta + 2\beta = -2\delta$. Therefore

$$\sum_{n=1}^{\infty} E_{U_{n-1}}(n^\beta U_n)^2 \leqq \sum_{n=1}^{\infty} Qn^{-1-2\delta} < +\infty$$

with probability one, and the sharper form of the Borel-Cantelli lemma (Dubins and Freedman (1965), Lemma (10)), implies that $\sum_{n=1}^{\infty} n^\beta U_n$ is a convergent series with probability one; similarly $\sum_{n=1}^{\infty} n^\beta V_n^2 < +\infty$. The remaining term in (4.1.2) multiplied by n^β is $O(n^{-1-2\gamma s+\beta})$ and is also summable because $\beta < \beta_0 \leqq 2\gamma s$. Hence with probability one, the conditions of Lemma 4.3 are satisfied for $b_n = \|X_n(\omega)\|^2$, so that lim sup $n^{\beta/2}\|X_n\| < +\infty$. By induction this is true for every $\beta < \beta_0$, which proves the theorem.

REMARK 5.4. The restriction to $\alpha = 1$ in Theorem 5.3 is made because for $\alpha < 1$ the method of proof is not efficient and gives a weak result in comparison with Theorem 5.1.

REFERENCES

BLOCK, H. D. (1957). Estimates of error for two modifications of the Robbins-Monro stochastic approximation process. *Ann. Math. Statist.* **28** 1003–1010.

BLUM, J. R. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Statist.* **25** 737–744.

BURKHOLDER, D. L. (1956). On a class of stochastic approximation processes. *Ann. Math. Statist.* **27** 1044–1059.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *J. Roy. Statist. Soc. B* **13** 1–45.

Chung, K. L. (1954). On a stochastic approximation method. *Ann. Math. Statist.* **25** 463–483.

Cochran, W. G. and Davis, M. (1965). The Robbins-Monro method for estimating the median lethal dose. *J. Roy. Statist. Soc. B* **27** 28–44.

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.

Dupač, V. (1957). O Kiefer-Wolfowitzově aproximační metodě. *Časopis Pěst. Mat.* **82** 47–75.

Fabian, V. (1960). Stochastic approximation methods. *Czech. Math. J.* **10** 123–159.

Fabian, V. (1964). A new one-dimensional stochastic approximation method for finding a local minimum of a function. *Trans. Third Prague Conf. Inform. Theor., Statist. Dec. Functions, Random Processes*. Czechoslovak Academy of Sciences. Prague. 85–105.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405.

Sakrison, D. (1962). Application of stochastic approximation methods to system optimization. Technical Report 391, Massachusetts Inst. Technology.

Schmetterer, L. (1961). Stochastic approximation. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 587–609. Univ. of California Press.

Vosiková, M. (1964). Asymptotické vlastnosti Kiefer-Wolfowitzovy aproximační metody. Thesis, Charles University, Prague.