ON CONVERGENCE OF THE KIEFER-WOLFOWITZ APPROXIMATION PROCEDURE

By J. H. Venter

Potchefstroom University

1. Summary. It is shown that the Kiefer-Wolfowitz procedure for estimating the maximum of a regression surface converges almost surely to the maximum if this is the only stationary point of the surface and some other conditions are satisfied. The result is in a sense stronger than those existing presently.

2. Introduction and preliminaries. Let R^p denote p-dimensional Euclidian space; the elements of R^p will be thought of as column vectors; u_i is the vector with jth component δ_{ij}, the Kronecker delta; $(\cdot, \cdot)$ and $\|\cdot\|$ denote the usual inner product and norm. Let f be a real valued function on R^p. Whenever it exists, we write $D(x) = \sum u_i f_i(x)$ where $f_i(x)$ is the derivative of f in the direction specified by u_i at x; we write $H(x)$ for the matrix with i, jth element $f_{ij}(x)$, the derivative of $f_i(x)$ in the direction specified by u_j.

Suppose f achieves its maximum at $\theta \varepsilon R^p$ and that for each $x \varepsilon R^p$ random variables $Y(x) = f(x) + Z(x)$ with $EZ(x) = 0$ can be observed. The problem is to estimate θ on the basis of the observed $Y(x)$'s at a number of suitably chosen x's. The Kiefer-Wolfowitz (KW) procedure [1], [2], [4], [5], [6], to this effect is as follows: Let $\{a_n\}$ and $\{c_n\}$ be positive sequences, let X_1 be arbitrary and supposing that after the $(n-1)$st step we have an estimate X_n of θ, observe the random variables $Y(X_n \pm c_n u_i)$, $i = 1, \cdots, p$; put

$$\Delta_n = \sum_{i=1}^{p} u_i (2c_n)^{-1} [Y(X_n + c_n u_i) - Y(X_n - c_n u_i)] \tag{1}$$

and take

$$X_{n+1} = X_n + a_n \Delta_n \tag{2}$$

as the next estimate of θ.

Convergence of $\{X_n\}$ to θ in the almost sure (a.s.) sense has been discussed in the references above and elsewhere. Throughout this paper it will be assumed that f has bounded second order partial derivatives. Under this assumption the key condition on f needed by these authors (see e.g. [5], p. 401) essentially amounts to

$$(x - \theta, D(x)) < 0 \quad \text{for all} \quad x \neq \theta. \tag{3}$$

In order to see the geometric implication of this condition, consider the derivative of f along the line through θ in the direction specified by the unit vector $\lambda = \sum \lambda_i u_i$. The equation of the line is $x - \theta = r\lambda$ and the derivative becomes $df(\theta + r\lambda)/dr = \sum \lambda_i f_i(\theta + r\lambda)$. Hence $r\, df(\theta + r\lambda)/dr = \sum r\lambda_i f_i(\theta + r\lambda) =$

Received 29 September 1966.

$(r\lambda, D(\theta + r\lambda))$ and it follows that (3) is equivalent to the condition:

(4) for each $r \neq 0$ and for each unit vector λ $r\, df(\theta + r\lambda)/dr < 0$,

i.e., along each line through θ, f is unimodal with maximum at θ. We will refer to this condition as linear unimodality.

Since the KW procedure may be thought of as an estimated steepest ascent procedure, one would expect that convergence would occur under conditions weaker than linear unimodality. Calling x stationary if $D(x) = 0$, we show below that if θ is the only stationary point of f and some other technical conditions hold, then X_n converges to θ a.s.

3. Main result. By θ being a local maximum of f we will mean that for some $\epsilon > 0$ we have $f(x) < f(\theta)$ for all $x \neq \theta$ in the set $\{x: \|x - \theta\| < \epsilon\}$. $K_0, K_1, K_2, \cdots$, will denote finite positive constants chosen to suit the context in which they appear.

THEOREM. *Let the following conditions hold:*

(a) *f has bounded second order derivatives over R^p;*

(b) *f is bounded above over R^p;*

(c) *θ is a local maximum of f;*

(d) *for each $x \in R^p$, $D(x) \neq 0$ if $x \neq \theta$;*

(e) *$E\|Z(x)\|^2 < K_0$ for all $x \in R^p$;*

(f) *$\sum a_n = \infty$, $\sum a_n c_n < \infty$, $\sum (a_n/c_n)^2 < \infty$, $c_n \to 0$ and $a_n < K_1 c_n^2$ for all n.*

Let $\{X_n\}$ be defined by (1) and (2). Then, with probability one, either $X_n \to \theta$ or $\|X_n\| \to \infty$.

PROOF. Without loss of generality we may take $\theta = 0$. By Taylor's theorem

$$(5)\quad (2c_n)^{-1}\{f(X_n + c_n u_i) - f(X_n - c_n u_i)\} = f_i(X_n) + \tfrac{1}{4}c_n\{f_{ii}(X_n + \phi'_{in} c_n u_i) - f_{ii}(X_n - \phi''_{in} c_n u_i)\}$$

where $0 \leqq \phi'_{in}, \phi''_{in} \leqq 1$. From (1)

$$(6)\quad \Delta_n = D(X_n) + L_n + U_n$$

where

$$(7)\quad L_n = \tfrac{1}{4}c_n \textstyle\sum_i u_i\{f_{ii}(X_n + \phi'_{in} c_n u_i) - f_{ii}(X_n - \phi''_{in} c_n u_i)\},$$

$$(8)\quad U_n = (2c_n)^{-1} \textstyle\sum_i u_i\{Z(X_n + c_n u_i) - Z(X_n - c_n u_i)\}.$$

It follows that

$$(9)\quad E\{U_n \mid X_n\} = 0 \text{ a.s.};$$

$$(10)\quad E\|U_n\|^2 \leqq c_n^{-2} K_2;$$

$$(11)\quad \|L_n\| \leqq c_n K_3;$$

where we have used conditions (a) and (e). Again from Taylor's theorem and (2)

$$f(X_{n+1}) = f(X_n) + a_n(\Delta_n , D(X_n)) + \tfrac{1}{2}a_n^2(\Delta_n , H_n\Delta_n) \tag{12}$$

where $H_n \equiv H(X_n + \phi_n a_n \Delta_n)$ with $0 \leqq \phi_n \leqq 1$. By Schwarz's inequality, condition (a), (6) and (11),

$$|(\Delta_n , H_n\Delta_n)| \leqq \|\Delta_n\|^2 K_4 \leqq 4\{\|D(X_n)\|^2 + c_n^2 K_3^2 + \|U_n\|^2\}K_4 . \tag{13}$$

Similarly

$$|(L_n , D(X_n))| \leqq \|L_n\| \, \|D(X_n)\| \leqq c_n K_3\{1 + \|D(X_n)\|^2\}. \tag{14}$$

Substituting (6) into (12) and applying the bounds derived above together with condition (f), it follows that

$$f(X_{n+1}) = f(\mathrm{X}_n) + a_n A_n \|D(X_n)\|^2 + a_n(U_n , D(X_n)) + B_n \tag{15}$$

where

$$|A_n| \leqq K_5 , \; A_n \to 1 \text{ a.s.} \quad \text{and} \quad \sum E|B_n| < \infty . \tag{16}$$

From (9) $E\{(U_n , D(\mathrm{X}_n)) \mid X_n\} = 0$ a.s. Hence if $\sum a_n^2 E\{(U_n , D(X_n))^2 \mid X_n\} < \infty$ for some sample sequence $\{X_n\}$ then according to Lemma 10 of [3] we may take it that $\sum a_n(U_n , D(X_n))$ converges. Iteration of (15) then shows that if $\sum a_n\|D(X_n)\|^2 = \infty$, we must have $f(X_n) \to \infty$ which is impossible in view of condition (b). Hence we may take it that

$$\sum a_n\|D(X_n)\|^2 < \infty , \tag{17}$$

$$\{f(X_n)\} \text{ converges} \tag{18}$$

for sample sequences such that $\sum a_n^2 E\{(U_n , D(X_n))^2 \mid X_n\} < \infty$. On the other hand, if this series diverges, then according to Lemma 8 of [3] we may assume that

$$\{\textstyle\sum_1^n a_k(U_k , D(X_k))\}/\{\sum_1^n a_k^2 E\{(U_k , D(X_k))^2 \mid X_k\}\} \to 0.$$

Hence using Schwarz's inequality, (10) and condition (f) we get

$$\{\textstyle\sum_1^n a_k(U_k , D(X_k))\}/\{\sum_1^n a_k A_k\|D(X_k)\|^2\} \to 0$$

and now iteration of (15) again shows that if (17) does not hold then $f(X_n) \to \infty$. It follows that (17) and (18) may be taken to be true generally. Further, from (2) and (6)

$$X_{n+1} - X_n = a_n D(X_n) + a_n L_n + a_n U_n . \tag{19}$$

It follows from (9), (10), condition (f) and Lemma 10 of [3] that $\sum a_n U_n$ converges a.s. and hence $a_n U_n \to 0$ a.s.; from (11) $a_n L_n \to 0$ and from (17) $a_n D(X_n) \to 0$ a.s. Hence

$$X_{n+1} - X_n \to 0 \text{ a.s.} \tag{20}$$

Now, supposing that we have a particular sample sequence $\{X_n\}$ for which the results above hold, we have four possibilities, viz.

(i) $0 = \liminf \|X_n\| < \limsup \|X_n\|$;

(ii) $0 < \liminf \|X_n\| \leqq \limsup \|X_n\| < \infty$;
(iii) $0 < \liminf \|X_n\| < \limsup \|X_n\| = \infty$;
(iv) the statement of the theorem holds for $\{X_n\}$.

To rule out case (i), choose ϵ such that $0 < \epsilon < \limsup \|X_n\|$ and such that $f(x) < f(0)$ for all $x \neq 0$ in the set $\{x: \|x\| < \epsilon\}$. Then, due to (20) there must be at least one limit point of $\{X_n\}$ in the set $\{x: \frac{1}{2}\epsilon < \|x\| < \epsilon\}$, say x^*. Let $\{X_{n_k}\}$ be a subsequence converging to x^*. Then $f(x^*) = \lim_{k\to\infty} f(X_{n_k}) = \lim_{n\to\infty} f(X_n)$, due to (18) and continuity of f. There is also a subsequence, $\{X_{n_m}\}$ say, converging to 0. Hence $f(0) = \lim_{m\to\infty} f(X_{n_m}) = \lim_{n\to\infty} f(X_n)$, from which follows the contradiction $f(0) = f(x^*)$.

In case (ii) condition (d) implies that $\liminf \|D(X_n)\| > 0$; this contradicts (17) and condition (f).

In order to rule out (iii) put $d = \liminf \|X_n\|$ and let $c > d$. Let $\epsilon > 0$ satisfy $0 < d - \epsilon < d + \epsilon < c$. Put

$$\alpha = \inf \{\|D(x)\|: d - \epsilon \leqq \|x\| \leqq c\}. \tag{21}$$

By continuity of $D(x)$ and condition (d), $\alpha > 0$. Let $\{X_{n_k}\}$ be a subsequence of $\{X_n\}$ such that $d - \epsilon \leqq \|X_{n_k}\| \leqq d + \epsilon$ for all k. For each k let m_k be the smallest positive integer such that $\|X_{n_k+m_k}\| \geqq c$; in the present circumstances $1 \leqq m_k < \infty$. Iterating (19), taking norms and using the various bounds established above, we get

$$\|X_{n_k+m_k}\| \leqq \|X_{n_k}\| + \textstyle\sum_{j=0}^{m_k-1} a_{n_k+j}\|D(X_{n_k+j})\| + \delta_{n_k}$$

where $\delta_{n_k} \to 0$ as $k \to \infty$. Hence

$$\textstyle\sum_{j=0}^{m_k-1} a_{n_k+j}\|D(X_{n_k+j})\| \geqq c - d - \epsilon - \delta_{n_k}. \tag{22}$$

Iteration of (15) together with arguments similar to those used in establishing (17) and (18) now show that for all k large enough

$$f(X_{n_k+m_k}) \geqq f(X_{n_k}) + \tfrac{1}{2}\alpha\textstyle\sum_{j=0}^{m_k-1} a_{n_k+j}\|D(X_{n_k+j})\| + \delta'_{n_k}$$

where $\delta'_{n_k} \to 0$ as $k \to \infty$. Substituting (22) we find

$$f(X_{n_k+m_k}) \geqq f(X_{n_k}) + \tfrac{1}{2}\alpha(c - d - \epsilon - \delta_{n_k}) + \delta'_{n_k} \tag{23}$$

and letting $k \to \infty$ the left hand side of this inequality tends to $\lim f(X_n)$ according to (18) whereas the right hand side tends to $\lim f(X_n) + \frac{1}{2}\alpha(c - d - \epsilon) > \lim f(X_n)$, yielding a contradiction.

The theorem therefore follows.

4. Discussion. 1. The "usual" choice $a_n = an^{-1}$, $c_n = cn^{-\gamma}$ with $a, c > 0$ and $0 < \gamma < \frac{1}{2}$ satisfies condition (f).

2. In the one dimensional case ($p = 1$) the possibility $|X_n| \to \infty$ can also be ruled out under the conditions of the theorem, as follows. There are three cases, viz. (a) $X_n \to -\infty$, (b) $\liminf X_n = -\infty$, $\limsup X_n = +\infty$, (c) $X_n \to +\infty$. In case (b) there must be a limit point of $\{X_n\}$ at θ due to (20) and the argument

of case (i) above becomes applicable. Further, the assumptions imply that $D(x) < 0$ for $x > \theta$. Hence in case (c), from (19), for all n large enough $X_{n+1} - X_n < a_nL_n + a_nU_n$ and iteration together with the convergence of $\sum a_nL_n$ and $\sum a_nU_n$ show that $\{X_n\}$ is bounded above, contradicting the assumed $X_n \to +\infty$. Similarly for case (a).

3. It seems unlikely that the possibility $\|X_n\| \to \infty$ can be ruled out under the conditions of the theorem in the multi-dimensional case ($p > 1$). We have been able to construct an example satisfying all conditions except boundedness of second order derivatives in which $\|X_n\| \to \infty$.

4. A simple additional condition sufficient for ruling out $\|X_n\| \to \infty$ is

(g) $\lim_{t\to\infty} \inf \{\|D(x)\| : \|x - \theta\| > t\} > 0$,

for this implies $\liminf \|D(X_n)\| > 0$ if $\|X_n\| \to \infty$, which would be impossible in view of (17) and condition (f).

5. Condition (g) is not satisfied by functions such as $f(x) = \exp\{-\|x - \theta\|^2\}$ for which one would expect convergence. Such cases are covered by the following weak linear unimodality condition.

(g′) For some T $(x - \theta, D(x)) \leqq 0$ for all x such that $\|x - \theta\| > T$.

Convergence under this additional condition is proved as follows. Suppose $\theta = 0$ and $\|X_n\| \to \infty$. From (2)

$$\|X_{n+1}\|^2 = \|X_n\|^2 + 2a_n(X_n, \Delta_n) + a_n^2\|\Delta_n\|^2. \tag{24}$$

From (6), (11), Schwarz and (g′)

$$\begin{aligned}(X_n, \Delta_n) &= (X_n, D(X_n)) + (X_n, L_n) + (X_n, U_n) \\ &\leqq c_nK_3\{1 + \|X_n\|^2\} + (X_n, U_n)\end{aligned} \tag{25}$$

for all n large enough. Also from (6) and (11)

$$\|\Delta_n\|^2 \leqq 4\{\|D(X_n)\|^2 + c_n^2K_3^2 + \|U_n\|^2\}$$

and substitution into (24) shows that

$$\|X_{n+1}\|^2 \leqq \{1 + 2a_nc_nK_3 + V_n\|X_n\|^{-1}\}\|X_n\|^2 + F_n \tag{26}$$

where $V_n = 2a_n(X_n, U_n)\|X_n\|^{-1}$ and $\sum |F_n| < \infty$ a.s. Hence $E\{V_n \mid X_n\} = 0$ and $E\{V_n^2 \mid X_n\} \leqq 4a_n^2c_n^{-2}K_2$ according to (9) and (10). By condition (f) and Lemma 10 of [3] it follows that $\sum V_n$ converges and hence also $\sum V_n\|X_n\|^{-1}$ converges. Also $EV_n^2 \leqq 4a_n^2c_n^{-2}K_2$ and hence $\sum V_n^2 < \infty$ which implies $\sum V_n^2\|X_n\|^{-2} < \infty$. Lemma 1b of [7] is applicable and iteration of (26) shows that $\{\|X_n\|\}$ is bounded, a contradiction.

6. The argument just given shows that condition (g) can be weakened as follows

(g″) $\lim_{t\to\infty} \inf \{\|x - \theta\| \, \|D(x)\| : \|x - \theta\| > t\} > 0$,

for, in this case there exists $K_5 > 0$ such that for all n large enough ($\theta = 0$),

$$(X_n, D(X_n)) \leqq \|X_n\| \, \|D(X_n)\| \leqq K_5\|X_n\|^2\|D(X_n)\|^2$$

if $\|X_n\| \to \infty$, and using this in (25) an alternative to (26) with $2K_5a_n\|D(X_n)\|^2$ added to the coefficient of $\|X_n\|^2$ is obtained. (17) ensures that the rest of the argument goes through.

REFERENCES

[1] DŮPAC, V. (1957). On the Kiefer-Wolfowitz approximation method. *Sel. Transl. Math. Statist. Prob.* **4** 43–69.

[2] DVORETZKY, A. (1956). On stochastic approximation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 39–55. Univ. of California Press.

[3] 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.

[4] KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

[5] SACKS, J. (1958). Asymptotic distributions of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405.

[6] SCHMETTERER, L. (1960). Stochastic approximation. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 587–609. Univ. of California Press.

[7] VENTER, J. H. (1966). On Dvoretzky stochastic approximation theorems. *Ann. Math. Statist.* **37** 1534–1544.