The Annals of Statistics
1978, Vol. 6, No. 4, 854-866

ON ASYMPTOTICALLY EFFICIENT RECURSIVE ESTIMATION[1]

By Václav Fabian

Michigan State University

Stochastic approximation procedures were shown by Sakrison to become asymptotically efficient estimators when used to minimize the Kullback-Leibler information, if certain conditions hold. Further results in this direction were obtained by Nevel'son and Has'minskij. This paper gives, first, alternative conditions for convergence and, secondly, shows that, under weaker conditions, asymptotic optimality is obtained by a modified stochastic approximation procedure. The modified procedure uses a consistent estimate which leads the approximating sequence to a proper local minimum of the Kullback-Leibler information. The conditions under which the procedure is asymptotically optimal are close to or weaker than those for asymptotic optimality of one-step-correction maximum likelihood methods.

1. Introduction. We shall be concerned with an estimation problem, described in the following assumption.

1.1. Assumption. Let m be a positive integer, Θ a subset of the m-dimensional Euclidean space R^m, θ a point in Θ. Suppose $\langle X, \mathbf{X}\rangle$ is a measurable space, ν a measure on $\mathbf{X}$ and for every δ in Θ, Q_δ is a probability measure on $\mathbf{X}$ with a density f_δ with respect to ν. Suppose that $Y, Y_1, Y_2, \cdots$ is a sequence of independent identically distributed random variables on a probability space $\langle \Omega, \mathbf{\Omega}, P\rangle$, and that $PY^{-1} = Q_\theta$.

1.2. Remarks. The problem considered is the estimation of θ on the basis of $Y_1, Y_2, \cdots$.

Define a function K on Θ by

$$K(\delta) = E[\log f_\theta(Y) - \log f_\delta(Y)] \tag{1}$$

so that $K(\delta)$ is the Kullback-Leibler information number for the pair $\langle \theta, \delta\rangle$ (for basic properties of K see Kullback (1959), or, e.g., Bahadur (1971)).

The function K has an absolute minimum, 0, at θ.

Sakrison (1965, 1966) proposed to estimate θ by using a stochastic approximation method. Under some conditions, he obtained an estimate which is asymptotically efficient in the sense that the covariance matrix of the estimate approaches the lower bound given by the Cramér-Rao inequality. Nevel'son and Has'minskij (1972) generalized these results and also proved the convergence in distribution of the normalized estimate.

Received August 1976; revised August 1977.

[1] Research partially supported by NSF Grant No. MPS76-07220.

AMS 1970 *subject classifications*. Primary 62F10, 62F20; Secondary 62L20.

Key words and phrases. Recursive estimation, stochastic approximation, asymptotically efficient, maximum likelihood, Fisher bound.

These results seem to be of considerable importance. In some sense, they are related to methods, in which an estimate is improved by taking a one step correction towards a solution of the maximum likelihood equation. But they are of a form which is easier to use when it is desired to calculate the estimate recursively. The methods are also related to stochastic approximation methods with optimally transformed observations (Abdelhamid (1973); Anbar (1973); Fabian (1973); Obremski (1976)) and in that way they are related to nonparametric asymptotically efficient estimation of a location parameter (see Stone (1975) for such a method, and for references, and Pfanzagl (1974) on the limits of this approach).

The purpose of this paper is to generalize the conditions under which the Sakrison method has its optimal properties. There are two sets of conditions: global conditions on K, to ensure convergence to θ, and local conditions, to establish the asymptotic properties. We shall show that the global conditions can be altered and the local conditions can be weakened. Moreover, if a consistent estimate is available, the procedure may be modified so that *only* the local conditions are used.

Next, we shall agree on some notation and formulate some conditions. Then we shall describe the result of Nevel'son and Has'minskij (Remark 1.6) and our results (Remark 1.7) in additional details.

1.3. NOTATION, CONVENTIONS. $\mathbf{B}_m$ denotes the σ-algebra of all Borel subsets of R^m. Transposition of matrices and vectors is denoted by a prime.

The domain and range of a function h will be denoted by $\mathscr{D}h$ and $\mathscr{R}h$ respectively. If h is a function with $\mathscr{D}h \subset R^u$, $\mathscr{R}h \subset R^v$ and x is an interior point of $\mathscr{D}h$ then a total differential of h at x is a $v \times u$ real valued matrix M such that

$$h(y) = h(x) + M(y - x) + \|y - x\|\varepsilon(y)$$

with a function ε satisfying $\lim_{y\to x} \varepsilon(y) = 0$.

By a (first) derivative of a real valued function h defined on a subset of R^u we mean the vector of the first partial derivatives. By the second derivative we mean the matrix of the second partial derivatives with the (i, j) element $(\partial^2/\partial x_i \, \partial x_j)h$. If the derivative of the function K, defined by (1.2.1), exists, it will be denoted by $\dot{K}$. The functions $\log f_\delta$ will be denoted also by L_δ. The first and second derivatives of $L_\delta(y)$, with respect to δ, will be denoted by $\dot{L}_\delta(y)$ and $\ddot{L}_\delta(y)$, if they exist. Note that $f_\delta(y) > 0$ if $\dot{L}_\delta(y)$ exists. If $\int \dot{L}_\delta \dot{L}_\delta' \, dQ_\delta$ makes sense, it will be denoted by $I(\delta)$. By I we shall denote the function $\delta \rightsquigarrow I(\delta)$ defined on the set of all such δ for which $I(\delta)$ makes sense.

The $m \times m$ identity matrix is denoted by $\mathbf{1}$.

Convergence and equalities among random variables are meant with probability one unless specified otherwise. A sequence of random variables ξ_n has a property eventually if for every ω in a set of probability 1, $\xi_n(\omega)$ has the property for all n greater than an $n_0(\omega)$.

A normal distribution with mean μ and covariance matrix C is denoted by $N(\mu, C)$ and convergence in distribution is denoted by $\rightarrow_{\mathscr{L}}$.

If $Z_1, \cdots, Z_n$ are random vectors then $\sigma(Z_1, \cdots, Z_n)$ denotes the σ-algebra generated by $Z_1, \cdots, Z_n$.

By a convergence in $L_2(P)$ of a sequence of random vectors we mean the $L_2(P)$ convergence in norm of corresponding components.

1.4. Condition. Assumption 1.1 holds with $\Theta = R^m$ and the following requirements are satisfied:

(i) Second derivatives, with respect to δ, of $L_\delta(y)$ and $K(\delta)$ (see (1.2.1)) exist for all δ and all y. Relations (1.2.1) and

$$\int f_\delta \, d\nu = 1 \tag{1}$$

can be twice differentiated under the integral sign.

(ii) The function I is defined on Θ, is finite valued, continuous, and $I(\delta)$ is nonsingular for every δ.

(iii) For every $\delta \neq \theta$,

$$(\delta - \theta)' I(\delta)^{-1} \dot{K}(\delta) > 0 .$$

(iv) With $Z_\delta = \dot{L}_\delta(Y)$,

$$\delta \rightsquigarrow I(\delta)^{-1}(E Z_\delta Z_\delta') I(\delta)^{-1} \tag{2}$$

is continuous and has each component bounded in absolute value by $C[1 + ||\delta||^2]$ for a constant C.

(v) $\theta_1 \in R^m$,

$$\theta_{n+1} = \theta_n + n^{-1} I(\theta_n)^{-1} \dot{L}_{\theta_n}(Y_n) . \tag{3}$$

1.5. Condition. Condition 1.4 holds and there is a neighborhood Θ_0 of θ such that, as $R \rightarrow \infty$,

$$\sup_{\delta \in \Theta_0} E||Z_\delta||^2 \chi_{\{||Z_\delta|| > R\}} \rightarrow 0 . \tag{1}$$

1.6. Remark. Under Conditions 1.4 and 1.5, Nevel'son and Has'minskij (1972, Theorem 5.4, Chapter 8) assert that

$$n^{\frac{1}{2}}(\theta_n - \theta) \rightarrow_{\mathscr{L}} N(0, I(\theta)^{-1}) . \tag{1}$$

This result is restated in our Theorem 5.9, but with some changes. We need to assume, additionally, that $\dot{K}(\theta)$ is a total differential of K at θ. On the other hand, Condition 1.5 can be omitted (also, in (1.4.2) an apparent misprint has been corrected).

We cannot prove the result without the additional assumption on $\dot{K}(\theta)$. Nevel'son and Has'minskij's proof sheds no light on this since it merely refers to the one-dimensional case.

Nevel'son and Has'minskij's theorem has additional assertions under additional assumptions: asymptotic behavior of $n^{\frac{1}{2}}(\theta_n - \theta)$, $n_1^{\frac{1}{2}}(\theta_{n_1} - \theta), \cdots, n_k^{\frac{1}{2}}(\theta_{n_k} - \theta)$ for $n \rightarrow \infty$, $\log n_1/n \rightarrow t_i$ and the convergence of the covariance matrix of $n^{\frac{1}{2}}\theta_n$ to $I(\theta)^{-1}$.

1.7. REMARK. As we said above in Remark 1.4, the purpose of this paper is to generalize the conditions under which the stochastic approximation has the optimal properties.

The most restrictive part of Condition 1.4 is (iii). It means that the square norm $|x|^2 = (x - \theta)'I(\delta)^{-1}(x - \theta)$ should be increasing at δ in the direction $\dot{K}(\delta)$. Thus (1.4.(iii)) will not hold if, roughly speaking, the graph of K has a valley descending in a direction leading away (as measured by $|\ \ |$) from θ. We shall show that this condition can be replaced by an alternative condition (see Theorem 4.8 and Remark 4.10).

Another question, pursued here, is whether any global conditions are necessary. The answer is that if a consistent estimate is available, the stochastic approximation can be modified in such a way that it has the optimal property under local conditions only (Section 3). The idea of using auxiliary estimates has already been used by Has'minskij (1974, Theorem 2). However, his result requires stronger conditions than our results. In particular, some of the properties are required to hold uniformly by Has'minskij, because his modification of the basic procedure is different.

We shall also generalize parts (i) and (v) of Condition 1.4. In particular, we shall consider the recursion given by (1.4.1) with $-\dot{L}_\delta(y)$ replaced by a possibly different function $q(\delta, y)$. In this way we obtain the behavior of $\langle\theta_n\rangle$ if an estimate q of $-\dot{L}$ is used and obtain robustness results. Another point is that $-q$ may be taken as a derivative of L in a weak sense. Nevel'son (1975) studied such estimators in case $X = \Theta = R$ and for q such that $q(\cdot, y)$ is nondecreasing for every y.

The organization of the paper is as follows. Section 2 exhibits and discusses conditions under which stochastic approximation gives asymptotically normal estimates. Section 3 contains results on stochastic approximation supported by auxiliary consistent estimates. Section 4 studies estimation by stochastic approximation without a support and shows alternative conditions for convergence and asymptotic efficiency to those used by Sakrison and Nevel'son and Has'minskij. In Section 5 the assumptions are further discussed and compared with the assumptions under which the Fisher bound was established by Bahadur (1964) (Theorem 5.7), with the assumptions for the behavior of maximum likelihood estimates (Remark 5.8) and with the assumptions used by Nevel'son and Has'minskij (1972) (Theorem 5.2).

2. A lemma.

2.1. REMARK. Here we shall study the asymptotic behavior of recursive estimates assuming, among other things, that the estimate is consistent.

2.2. ASSUMPTION. (i) Assumption 1.1 holds, Θ_1 in a neighborhood of θ, $\Theta_1 \subset \Theta$, and q is a $\mathbf{B}_m \times \mathbf{X}$-measurable function into R^m defined on $\Theta_1 \times X$.

(ii) The random variable $Z_\delta = q(\delta, Y)$ is in $L_2(P)$ for every δ in Θ_1, $Z_\delta \to Z_\theta$ in $L_2(P)$ if $\delta \to \theta$. The function D defined on Θ_1 by $D(\delta) = EZ_\delta$, has value 0 at

θ and a nonsingular total differential H at θ. The covariance matrix $EZ_\theta Z_\theta'$ of Z_θ is denoted by Σ.

(iii) With $\mathscr{F}_n = \sigma(Y_1, \cdots, Y_{n-1})$, Φ_n are $\mathscr{F}_n$-measurable $m \times m$ matrix valued random vectors, θ_n are m-dimensional $\mathscr{F}_n$-measurable random vectors and

(1) $$\Phi_n \to H^{-1} \quad \text{on} \quad \{\theta_n \to \theta\}\,.$$

2.3. Lemma. *Let Assumption* 2.2 *hold and let*

(1) $$\theta_n \to \theta\,.$$

Then the following two assertions are true. If

(2) $$\|\theta_{n+1} - \theta\| \leqq \|\theta_n - \theta - n^{-1}\Phi_n q(\theta_n, Y_n)\|$$

eventually, then

(3) $$n^\beta(\theta_n - \theta) \to 0 \qquad \textit{for every} \quad \beta \in (0, \tfrac{1}{2})\,.$$

If

(4) $$\theta_{n+1} = \theta_n - n^{-1}\Phi_n q(\theta_n, Y_n)$$

eventually, then

(5) $$n^{\frac{1}{2}}(\theta_n - \theta) \to_{\mathscr{L}} N(0, H^{-1}\Sigma H^{-1})\,.$$

Proof. Without loss of generality we may assume that $\theta = 0$. Let Θ_0 be a subset of Θ_1, and a neighborhood of θ. If we define $q_1(\delta, \bullet)$ to be $q(\delta, \bullet)$ for $\delta \in \Theta_0$ and $q(\delta, \bullet) = 0$ for $\delta \in \Theta_1$, all assumptions of the lemma will be satisfied with q replaced by q_1. Thus we may assume that q is defined on $R^m \times X$, $q(\delta, x) = 0$ for $\delta \notin \Theta_0$. Set now

(6) $$D_n = D(\theta_n)\,, \qquad V_n = D_n - q(\theta_n, Y_n)\,.$$

Notice that $D(\delta) - Z_\delta \to -Z_\theta$ strongly in $L_2(P)$ as $\delta \to \theta$. Making Θ_0 smaller, if necessary, we find a constant C such that

(7) $$C > E_{\mathscr{F}_n}\|V_n V_n' - \Sigma\| \to 0\,, \qquad E_{\mathscr{F}_n}\|V_n\|^2 < C\,.$$

We also obtain

(8) $$E\|V_n\|^2\chi_{\{\|V_n\|^2 > \eta n\}} \to 0 \qquad \text{for every} \quad \eta > 0\,.$$

Indeed, the conditional, given $\mathscr{F}_n$, expectation of the random variable in (8) is $b_n(\theta_n)$ with $b_n(\delta) = E\|D(\delta) - Z_\delta\|^2\chi_{\{\|D(\delta) - Z_\delta\|^2 > \eta n\}}$. The uniform integrability of $\|D(\delta_n) - Z_{\delta_n}\|^2$ for $\delta_n \to \theta$ implies $b_n(\theta_n) \to 0$. We obtain $Eb_n(\theta_n) \to 0$ since $b_n(\theta_n) \leqq C$ by (7). Thus (8) holds.

From (1), (2.2.1) and Assumption 2.2.(ii) we obtain

(9) $$\Phi_n \to H^{-1}\,, \qquad \Phi_n D_n = \Gamma_n \theta_n\,, \qquad \Gamma_n \to \mathbf{1}$$

and we can choose Γ_n to be $\mathscr{F}_n$-measurable.

Thus

(10) $$\theta_n - n^{-1}\Phi_n q(\theta_n, Y_n) = (\mathbf{1} - n^{-1}\Gamma_n)\theta_n + n^{-1}\Phi_n V_n\,.$$

To prove the first implication we adopt a technique used in Fabian (1967, proof of Theorem 5.3).

From (2), (9) and (10), eventually,

$$(11) \qquad \|\theta_{n+1}\|^2 \leqq (1 + n^{-1})\|\theta_n\|^2 + 2n^{-1}\theta_n'(\mathbf{1} - n^{-1}\Gamma_n)'\Phi_n V_n + n^{-2}\|\Phi_n V_n\|^2 .$$

By (7) and (9),

$$(12) \qquad \sum_{n=1}^{\infty} n^{2\beta-2}\|\Phi_n V_n\|^2 < +\infty$$

for every $\beta \in (0, \frac{1}{2})$.

Suppose now (3) holds for β equal to a β_1 in $[0, \frac{1}{2})$. Let $\beta_2 \in (\beta_1, \frac{1}{2})$ and set $\beta = \frac{1}{2}(\beta_1 + \beta_2)$. Denoting the middle term on the right-hand side of (11) by W_n, we find

$$E_{\mathscr{F}_n}\|W_n\|^2 \leqq 4n^{-2}\|\Phi_n'(\mathbf{1} - n^{-1}\Gamma_n)\|^2\|\theta_n\|^2 E_{\mathscr{F}_n}\|V_n\|^2 .$$

Because of (9) and (7), the last expression is $o(n^{-2-2\beta_1})$. Multiplied by $n^{4\beta} = n^{2\beta_1+2\beta_2}$ it becomes $o(n^{-2+2\beta_2})$ which is summable.

A generalized Borel–Cantelli lemma (Lemma 10 in Dubins and Freedman, 1965) implies that $\sum_{n=1}^{\infty} n^{2\beta}W_n$ exists and is finite. This, (12), (11) and Lemma 4.3 in Fabian (1967) imply that

$$\limsup n^{2\beta}\|\theta_n\|^2 < +\infty .$$

A complete induction proves the first assertion.

Let us prove the second assertion. From (4) and (10) we obtain

$$\theta_{n+1} = (\mathbf{1} - n^{-1}\Gamma_n)\theta_n + n^{-1}\Phi_n V_n + n^{-\frac{3}{2}}T_n$$

with Γ_n, Φ_n, V_{n-1} being $\mathscr{F}_n$-measurable and with $T_n = 0$ eventually. This, (7) and (8) show that the conditions of Theorem 2.2 in Fabian (1968) are satisfied and the required result follows.

3. Approximation using auxiliary estimates.

3.1. Assumption. Assumption 1.1 holds. For every n, t_n is an m-dimensional random vector, γ_n a positive number, $\eta \in (0, \frac{1}{2})$,

$$(1) \qquad \gamma_n\|t_n - \theta\| \to 0 , \qquad \gamma_n \to +\infty$$

and

$$(2) \qquad \begin{aligned} T_n(\delta) &= \delta \qquad \text{if } \|t_n - \delta\| < \gamma_n^{-1} \vee n^{-\eta} , \\ &= \delta^* \qquad \text{otherwise} \end{aligned}$$

where δ^* denotes the closest point to δ on the sphere $\{u;\, u \in R^m,\, \|u - t_n\| = \gamma_n^{-1}\}$ and $\vee$ denotes maximum.

3.2. Theorem. *Let Assumptions* 2.2 *and* 3.1 *hold. Let*

$$(1) \qquad \theta_{n+1} = T_{n+1}[\theta_n - n^{-1}\Phi_n q(\theta_n, Y_n)] , \quad \textit{eventually}.$$

Then

$$(2) \qquad n^{\beta}(\theta_n - \theta) \to 0 \qquad \textit{for every} \quad \beta \in [0, \tfrac{1}{2})$$

and

$$(3) \qquad n^{\frac{1}{2}}(\theta_n - \theta) \to_{\mathscr{L}} N(0, H^{-1}\Sigma H^{-1}) .$$

PROOF. Since $\|\theta_n - t_n\| < \gamma_n^{-1} \vee n^{-\eta}$, we have $\theta_n \to \theta$ and (2.3.1) holds. Next we establish that, with u_{n+1} denoting the argument of T_{n+1} in (1),

$$(4) \qquad \|\theta_n - \theta\| \leqq \|u_n - \theta\| \quad \text{on} \quad \{\|t_n - \theta\| < \gamma_n^{-1}\} .$$

Notice $\theta_n = u_n$ if $\|t_n - u_n\| \leqq \gamma_n^{-1} \vee n^{-\eta}$. In the opposite case, u_n lies outside the sphere S with center at t_n and radius γ_n^{-1}. The hyperplane perpendicular to the segment $[u_n, \theta_n]$ and intersecting the segment at its center, divides the points δ into those closer to u_n and those closer to θ_n. The hyperplane does not intersect the sphere S and all points in S are closer to θ_n than to u_n. Thus, if θ is in S, then θ is closer to θ_n than to u_n and (4) holds.

Relations (3.1.1), (1) and (4) imply that (2.3.2) holds eventually. By Lemma 2.3, relation (2) holds. But this implies that $\theta_{n+1} = \theta_n - n^{-1}\Phi_n q(\theta_n, Y_n)$ (i.e., (2.3.4)) holds eventually. A new application of Lemma 2.3 yields (3).

4. Approximation without auxiliary estimates.

4.1. REMARK. In this section we shall consider conditions under which it is not necessary to use an auxiliary estimate sequence $\langle t_n \rangle$. Essentially these will be conditions which guarantee that $\theta_n \to \theta$. They will be global conditions and could be formulated as not necessarily implying the asymptotic efficiency, but we thought this would not be worth the necessary restatement of the basic assumptions. Thus in all subsequent results we always start with Assumption 4.2 which then automatically implies the desired result (as described in Condition 4.3) if $\theta_n \to \theta$.

4.2. ASSUMPTION. Assumption 2.2 holds with $\Theta = \Theta_1 = R^m$,

$$(1) \qquad \theta_{n+1} = \theta_n - n^{-1}\Phi_n q(\theta_n, Y_n) .$$

4.3. CONDITION. For every $\beta \in [0, \frac{1}{2})$, $n^\beta(\theta_n - \theta) \to 0$ and

$$(1) \qquad n^{\frac{1}{2}}(\theta_n - \theta) \to_{\mathscr{L}} N(0, H^{-1}\Sigma H) .$$

4.4. LEMMA. *Let Assumption* 4.2 *hold. Let g be a real valued nonnegative function on R^m with the first derivative D_g and the second derivative H_g on R^m, H_g continuous and bounded in norm. Let C be a positive number,* $\eta \in (0, 1)$,

$$(1) \qquad D_g'(\theta_n)\Phi_n D(\theta_n) \geqq (C \log n)^{-1} B(\theta_n) ,$$

$$(2) \qquad E_{\mathscr{F}_n}\|\Phi_n q(\theta_n, Y_n)\|^2 \leqq C \log n[1 + g(\theta_n) + B(\theta_n)]$$

with a nonnegative function B on R^m. Let $g(\theta) = 0$,

$$(3) \qquad \inf\{B(\delta); \varepsilon < g(\delta) < \varepsilon^{-1}\} > 0 ,$$

$$(4) \qquad \inf\{g(\delta); \varepsilon < \|\delta - \theta\|\} > 0$$

for every $\varepsilon > 0$.

Then Condition 4.3 *holds.*

PROOF. Write $z_n = g(\theta_n)$. By the Taylor expansion, enlarging C if necessary,

$$z_{n+1} \leqq z_n - n^{-1}D_g'(\theta_n)\Phi_n q(\theta_n, Y_n) + n^{-2}C\|\Phi_n q(\theta_n, Y_n)\|^2 .$$

Using (1), (2), enlarging C again, and recalling that $E_{\mathscr{F}_n} q(\theta_n, Y_n) = D(\theta_n)$, we obtain

$$E_{\mathscr{F}_n} z_{n+1} \leqq (1 + Cn^{-2}\log n)z_n - [(Cn\log n)^{-1} - Cn^{-2}\log n]B(\theta_n) + Cn^{-2}\log n .$$

Since the sequence $Cn^{-2}\log n$ is summable we obtain from Theorem 1 in Robbins and Siegmund (1971) that $\langle z_n\rangle$ converges to a finite limit and that $\sum_{n=1}^{\infty}[(Cn\log n)^{-1} - Cn^{-2}\log n]B(\theta_n)$ is finite. Then at every ω in an event of probability 1, a subsequence of $B(\theta_n)$ converges to 0, and $\langle g(\theta_n)\rangle$ converges.

By (3), we obtain that a subsequence of $\langle g(\theta_n)\rangle$ converges to 0 at ω, thus $g(\theta_n) \to 0$ at ω. This proves $g(\theta_n) \to 0$ and, by (4), $\theta_n \to \theta$.

The assertion follows now from Lemma 2.3.

4.5. REMARK. Robbins and Siegmund (1971) give numerous applications of their Theorem 1, both within and without the area of stochastic approximation. But the theorem itself is only a slight generalization of a lemma used to establish convergence in stochastic approximation, first in Blum (1954) and then in many other papers (see, e.g., Lemma 3.2 in Fabian (1971)). The convergence result in the preceding lemma would also follow from a slight generalization of Lemma 3.3 in Fabian (1971).

4.6. REMARK. Various choices of the test-function g in Lemma 4.1 give various sufficient conditions for convergence. The next result obtains by taking $g(\delta) = \|\delta - \theta\|^2$.

4.7. THEOREM. *Let Assumption* 4.2 *hold, let A be $\mathbf{B}_m$-measurable,* $\Phi_n = A(\theta_n)$. *On* $R^m - \{\theta\}$ *let*

(1) $$\delta \to (\delta - \theta)'A(\delta)D(\delta)$$

be continuous positive valued. Let

(2) $$E\|A(\delta)Z_\delta\|^2 \leqq C[1 + \|\delta - \theta\|^2]$$

for a constant C and all δ.

Then Condition 4.3 *holds.*

PROOF. Apply Lemma 4.4 with $g(\delta) = \|\delta - \theta\|^2$, $D_g(\delta) = 2(\delta - \theta)$, $B(\delta) = (\delta - \theta)'A(\delta)D(\delta)$. Then condition (4.4.1) holds by the choice of B, (4.4.2) follows from (2), (4.4.3) follows from the properties of B and (4.4.4) is trivial. Thus Theorem 4.7 follows from Lemma 4.4.

4.8. THEOREM. *Let Assumption* 4.2 *hold, let K, defined by* (1.2.1), *have a first derivative equal to D, and a continuous bounded second derivative. Let*

(1) $$\inf\{\|D(\delta)\|; \|\delta - \theta\| > \varepsilon, K(\delta) < \varepsilon^{-1}\} > 0$$

for every $\varepsilon > 0$. For a number C let

(2) $$\inf\{K(\delta); \|\delta - \theta\| > C\} > 0$$

and

$$E\|Z_\delta\|^2 \leqq C[1 + K(\delta) + \|D(\delta)\|^2]\,. \tag{3}$$

Suppose $\Phi_n(\omega)$ *are symmetric, with eigenvalues in* $[(C\log n)^{-1}, C\log n]$, *for all* n *and* ω.

Then Condition 4.3 *holds.*

PROOF. It is enough to verify the assumptions of Lemma 4.4 for $g = K$, $D_g = D$, $B = \|D\|^2$. Note that (4.4.1) and (4.4.2) follow from the properties of Φ_n and from (3). Relations (4.4.3) and (4.4.4) follow from (1), (2) and from the fact that θ is the unique point at which K is zero, and from the continuity of K.

4.9. REMARK. Theorem 4.8 does not require the positiveness of (4.7.1). It requires, however, the boundedness of the second derivative of K. This condition is also unpleasant, and especially so, since we require $\Theta = R^m$. It would be of interest to extend Theorem 4.8 to the case $\Theta \subset R^m$. This may be an easy task if Θ is, e.g., a sphere when a modification similar to that described in Theorem 3.2 would work. For less simple sets Θ we are getting into the area of nonlinear programming. Some results here are known (Fabian, 1965 and Kushner, 1974).

4.10. REMARK. Some additional comments on the two Theorems 4.7 and 4.8. Theorem 4.8 is close to the conditions used by Blum (1954) in his original paper on multidimensional approximation. But Sacks (1958) used a condition analogous to the positivity of (4.7.1). Venter (1967b) pointed out the undesirability of this condition and then apparently unaware of Blum's (1954) results, obtained results weaker than those obtained by Blum.

5. Special cases and comparisons.

5.1. REMARK. The results in preceding sections were obtained under Assumption 2.2 and with the use of an auxiliary consistent estimate (Section 3) or global conditions guaranteeing $\theta_n \to \theta$ (Section 4). Part (iii) of Assumption 2.2 is usually easy to satisfy (see the choice of Φ_n in Theorems 5.2 and 5.7 and also Remark 5.9). Parts (i) and (ii) of Assumption 2.2 are rather general and we shall see that they are satisfied in particular situations considered below.

If only Assumption 2.2, parts (i) and (ii) are required (rather than the stronger Condition 1.4, or the conditions in Theorem 5.7), then there is a question about the meaning of the asymptotic covariance $C = H^{-1}\Sigma H$. Suppose $q(\delta, y)$ is, in some sense, $\dot{L}_\delta(y)$, so that $\Sigma = I(\theta)$. Under weak assumptions, corollaries to Lemma 6 in Le Cam (1970) show that $I(\theta) \leqq H$ and it follows that $C \leqq I(\theta)^{-1}$. If the Fisher bound $C \geqq I(\theta)^{-1}$ holds, we have $C = I(\theta)^{-1}$.

We shall now restate and prove the result of Nevel'son and Has'minskij, with changes explained in Remark 1.6.

5.2. THEOREM. *Let Condition* 1.4 *hold and let* $\dot{K}$ *have a total differential at* θ.

Then

(1) $$n^{\frac{1}{2}}(\theta_n - \theta) \rightarrow_{\mathscr{L}} N(0, I(\theta)^{-1}) .$$

PROOF. Assumption 1.1 is repeated in Condition 1.4. Part (i) of Assumption 2.2 is satisfied with $\Theta_1 = R^m$ and $q(\delta, y) = -\dot{L}_\delta(y)$, since measurability with respect to y and continuity with respect to δ implies joint measurability. We obtain $Z_\delta \in L_2(P)$ from (1.4.(iv)), $Z_\delta \rightarrow Z_\theta$ in $L_2(P)$ because $Z_\delta \rightarrow Z_\theta$ pointwise on Ω if $\delta \rightarrow \theta$ and (1.4.(ii)) together with (1.4.(iv)) imply $E\|Z_\delta\|^2 \rightarrow E\|Z_\theta\|^2$ as $\delta \rightarrow \theta$. Since $D(\delta) = \dot{K}(\delta)$ by (1.4.(i)), we obtain $D(\theta) = 0$. Condition (1.4.(i)) and the assumed existence of a total differential H of $\dot{K}$ imply that $H = \Sigma = I(\theta)$. Thus (2.2.(ii)) holds. Condition (1.4.(ii)) implies that $\Phi_n = I^{-1}(\theta_n)$ satisfy (2.2.(iii)). We have shown that Assumption 2.2 holds and the proof is now completed by an application of Theorem 4.7 with $A = I^{-1}$ and with (4.7.1) and (4.7.2) implied by (1.4.(iii)) and (1.4.(iv)).

Next we shall state a condition used by Bahadur (1964):

5.3. CONDITION. Assumption 1.1 holds with Θ open, $\dot{L}_\delta(y)$, $\ddot{L}_\delta(y)$ exist for all $\delta \in \Theta$, $y \in R^m$, and $\ddot{L}_\delta(y)$ is continuous with respect to δ at θ. Also, $I(\theta)$ is nonsingular,

(1) $$E\dot{L}_\theta(Y) = 0 , \qquad \|I(\theta)\| < +\infty ,$$

(2) $$E\ddot{L}_\theta(Y) = -I(\theta) .$$

There is an X-measurable function M and a neighborhood Θ_0 of θ such that

(3) $$\|\ddot{L}_\delta(y)\| \leqq M(y)$$

for all $\delta \in \Theta_0$, $y \in R^m$, and

(4) $$EM(Y) < +\infty .$$

5.4. REMARK. The validity of the so-called Fisher bound for the asymptotic variance of a sequence of estimation was studied by several authors starting with Le Cam (1953) (see also Pfanzagl (1973) and the references therein). We shall compare our results with conditions used by Bahadur (1964).

Suppose Condition 5.3 is satisfied for every θ in Θ. Write now E_θ for E. Bahadur (1964) showed that if t_n is $(Y_1, \cdots, Y_n)$-measurable for every n and

(1) $$n^{\frac{1}{2}}(t_n - \theta) \rightarrow_{\mathscr{L}} N(0, C(\theta)) \qquad \text{on} \quad (\Omega, \Omega, E_\theta)$$

for every θ then

(2) $$C(\theta) - I(\theta)^{-1}$$

is positive semidefinite for almost all (with respect to the Lebesgue measure) θ in Θ. This gives a precise meaning to a convention under which $\langle t_n \rangle$ is called asymptotically efficient if (1) holds with $C(\theta) = I(\theta)^{-1}$.

We shall show that under little more than Condition 5.3 stochastic approximation methods give asymptotically efficient estimates.

5.5. LEMMA. *Let Condition* 5.3 *hold. Then*

$$EM^2(Y) < \infty \tag{1}$$

implies

$$E\|\dot{L}_\delta(Y)\|^2 \to E\|\dot{L}_\theta(Y)\|^2 \quad \textit{as} \quad \delta \to \theta\,, \tag{2}$$

and (2) *implies parts* (i) *and* (ii) *of Assumption* 2.2 *with* $q(\delta, y) = -\dot{L}_\delta(y)$, *a* Θ_1, *and* $\Sigma = H = I(\theta)$.

PROOF. $D(\theta) = 0$ follows from (5.3.1), $\Sigma = I(\theta)$ from the definition of q. A Taylor expansion, (5.3.3) and (5.3.4) give

$$Z_\delta = Z_\theta - \ddot{L}_\theta(Y)(\delta - \theta) + \|\delta - \theta\|R(\delta) \tag{3}$$

with

$$R_\delta \to 0\,, \qquad E\|R_\delta\| \to 0 \quad \text{as} \quad \delta \to \theta\,. \tag{4}$$

If (1) holds, we have even $R_\delta \to 0$ in $L_2(P)$, $Z_\delta \to Z_\theta$ in $L_2(P)$ and the first assertion of the lemma holds.

Assume (2) holds. Take expectations in (3) to obtain

$$D(\delta) = I(\theta)(\delta - \theta) + \|\delta - \theta\|ER_\delta\,. \tag{5}$$

Thus $I(\theta)$ is a total differential of D at θ and the second assertion holds.

5.6. REMARK. Under conditions of the previous lemma, and if Φ_n are suitably determined we can obtain $n^{\frac{1}{2}}(\theta_n - \theta) \to_{\mathscr{L}} N(0, I(\theta)^{-1})$ either by Theorem 3.2 using auxiliary estimates, or by Theorems 4.7 and 4.8, if certain additional global conditions are satisfied. Let us formulate one of these possible assertions.

5.7. THEOREM. *Let Condition* 5.3 *hold with* $EM^2(Y) < \infty$ *and let* I *be continuous at* θ. *Let* $\theta_1 \in R^m$,

$$\theta_{n+1} = T_{n+1}[\theta_n + n^{-1}\Phi_n \dot{L}_{\theta_n}(Y_n)] \tag{1}$$

with T_n *satisfying Assumption* 3.1 *where*

$$\begin{aligned} \Phi_n &= I(\theta_n)^{-1} \quad && \textit{if} \quad I(\theta_n) \quad \textit{is nonsingular}, \\ &= \mathbf{1} && \textit{if} \quad I(\theta_n) \quad \textit{is singular}. \end{aligned} \tag{2}$$

Then $n^\beta(\theta_n - \theta) \to 0$ *for every* $\beta \in (0, \frac{1}{2})$ *and*

$$n^{\frac{1}{2}}(\theta_n - \theta) \to_{\mathscr{L}} N(0, I(\theta)^{-1})\,. \tag{3}$$

PROOF. Using Lemma 5.5 we establish that Assumption 2.2 holds and the result follows from Theorem 3.2.

5.8. REMARK. We shall now compare the previous result to the known behavior of one-step-maximum likelihood methods.

Let Condition 5.3 hold with I continuous at θ, let $\langle t_n \rangle$ be a sequence of estimates such that $n^{\frac{1}{2}}(t_n - t)$ is a tight sequence and let

$$X_n = t_n + (nI(t_n))^{-1} \textstyle\sum_{j=1}^n \dot{L}_{t_n}(Y_n)\,. \tag{1}$$

Hannan (1976), generalizing slightly Le Cam's (1956) Lemma 6, showed that

$$n^{\frac{1}{2}}(X_n - \theta) \to_{\mathscr{L}} N(0, I(\theta)^{-1}) .$$

The difference in these assumptions and those of Theorem 5.7 is that the one step maximum likelihood method requires more of t_n while the stochastic approximation method requires, e.g., the finiteness of $EM^2(Y)$.

5.9. REMARK. *Choice of* Φ_n. Assuming I continuous and $H = I(\theta)$, we can choose Φ_n as in the preceding theorems, or, if desired, (5.7.2) can be used only for a subsequence $n_1 < n_2 < \cdots$, with $\Phi_n = \Phi_{n_i}$ for $n_i \leqq n < n_{i+1}$.

But even without these assumptions, a choice of consistent estimates Φ_n of H^{-1} is possible and easy since $q(\delta, Y_i)$ are unbiased estimates of $D(\delta)$. In a more complicated situation the matrix of second derivatives of a function is estimated in Fabian (1971, Theorem 2.7); a different method of estimating additional properties of the function, to which stochastic approximation is applied, is due to Venter (1967a). These methods can also be used to obtain suitable Φ_n.

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.

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545-1552.

BAHADUR, R. R. (1971). Some limit theorems in statistics. *Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

BLUM, J. R. (1954). Multidimensional stochastic approximation methods. *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. (1967). Stochastic approximation of constrained minima. *Trans. Fourth Prague Conf. Information Theor., Decision Functions, Random Processes*, 277-290. Academia, Praha.

FABIAN, V. (1967). Stochastic approximation methods with improved asymptotic speed. *Ann. Math. Statist.* **38** 191-200.

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

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

FABIAN, V. (1973). Asymptotically efficient stochastic approximation; the RM case. *Ann. Statist.* **1** 486-495.

HANNAN, J. (1976). Asymptotic theory. Unpublished lecture notes.

HAS'MINSKIJ, R. Z. (1974). Sequential estimation and recursive asymptotically optimal procedures of estimation and observation control. *Proc. Prague Symp. Asymptotic Statist.* (Charles Univ., Prague) **1** 157-178. Charles Univ., Prague.

KULLBACK, S. (1959). *On Information and Sufficiency*. Wiley, New York.

KUSHNER, H. J. (1974). Stochastic approximation algorithms for constrained optimization problems. *Ann. Statist.* **2** 713-723.

LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277-330.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 129–156.

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

NEVEL'SON, M. B. (1975). On the properties of the recursive estimates for a functional of an unknown distribution function. In *Limit Theorems of Probability Theory* (P. Révész, ed.) 227–251. North Holland, Amsterdam.

NEVEL'SON, M. B. and HAS'MINSKIJ, R. Z. (1972). *Stochastic Approximation and Recursive Estimation.* (In Russian.) Nauka, Moskva.

OBREMSKI, T. E. (1976). A Kiefer-Wolfowitz type stochastic approximation procedure. Ph. D. Dissertation. Dept. of Statist. and Probability, Michigan State Univ.

PFANZAGL, J. (1973). Asymptotic optimum estimation and test procedures. *Proc. Prague Symp. Asymptotic Statist.* (Charles Univ., Prague) **1** 201–272. Charles Univ., Prague.

PFANZAGL, J. (1974). Investigating the quantile of an unknown distribution. In *Contributions to Applied Statistics* (dedicated to Arthur Linder; W. J. Ziegler, ed.), 111–126. Birkhauser Verlag, Basel.

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (J. S. Rustagi, ed.), 233–257. Academic Press, New York.

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

SAKRISON, D. J. (1965). Efficient recursive estimation; application to estimating the parameters of a covariance function. *Internat. J. Engrg. Sci.* **3** 461–483.

SAKRISON, D. J. (1966). Stochastic approximation: A recursive method for solving regression problems. *Advances in Communication Systems* **2** 51–106.

STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284.

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

VENTER, J. H. (1967b). On convergence of the Kiefer–Wolfowitz approximation procedure. *Ann. Math. Statist.* **38** 1031–1036.

DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MICHIGAN 48824