DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 730

November 1983

WEAK CONVERGENCE OF STOCHASTIC APPROXIMATION PROCESSES WITH RANDOM INDICES[1]

by

Edward W. Frees

WEAK CONVERGENCE OF STOCHASTIC APPROXIMATION PROCESSES WITH RANDOM INDICES[1]

by

Edward W. Frees

Department of Statistics
1210 West Dayton Street
University of Wisconsin
Madison, Wisconsin 53706

Running Head: Stochastic approximation

November 1983

1980 AMS subject classification: Primary 62L20; Secondary 62L15, 60F17.

Keywords and phrases: strong approximations, mode estimation, stopping times.

---

[1]This research was supported in part by the National Science Foundation (Grant MCS-8100748).

ABSTRACT

Conditions are given for weak convergence through random indices of a general stochastic approximation process which includes the Robbins-Monro and Kiefer-Wolfowitz processes. For a particular index, a sequential fixed-width bounded length confidence interval for the parameter being estimated is established. As an example, an optimal recursive estimator and confidence interval for the mode of a distribution function is constructed.

§1. Introduction and Summary

In a stochastic approximation (SA) procedure, stopping time variables or stopping rules are important because they give the researcher a reasonable criterion for halting the procedure with a certain amount of confidence (see Stroup and Braun (1982), Sielken (1973) and Farrell (1962)). Moreover, stopping times are useful in developing confidence regions for parameters of interest. Proof of weak convergence through a general random index gives results for stopping times as simple corollaries.

Consider a general SA algorithm,

$$X_{n+1} = X_n - n^{-1}\{f(X_n) + n^{\tau-1/2}\beta_n + n^{\tau}\varepsilon_n\}, \tag{1.1}$$

for locating the assumed unique root of $f(x)=0$, say θ, where $f(\cdot)$ is a measurable function such that $f:R \to R$. In (1.1), β_n and ε_n are random variables and τ is a fixed constant such that $0 \leq \tau \leq 1/2$. Heuristically, one may think of ε_n as the random error at the n^{th} stage due to the fact that $f(X_n)$ can only be measured with error. β_n may be thought of as the bias in the measurement of $f(X_n)$, e.g., if $f(\cdot)$ is a probability density function. The algorithm (1.1) is a special case of the algorithm studied by Kushner (1977) and Ljung (1978). It contains the well-known Robbins-Monro (RM) (1951) and Kiefer-Wolfowitz (KW) (1952) procedures for finding roots and maxima of a regression function, respectively.

Motivation for studying a general algorithm has been given by Ruppert (1982). Ruppert considered a similar algorithm to estimate the (assumed) unique θ such that $f(\theta)=0$,

$$X_{n+1} = X_n - n^{-1}\{f(X_n) + n^{-2\tau}\beta_n + n^{\tau}\varepsilon_n\} \tag{1.2}$$

where $0 \leq \tau \leq 1/4$. As argued by Ruppert, an important reason for considering algorithms of the form (1.2) is that a conventional restrictive assumption in SA that the errors $\{\varepsilon_n\}$ be martingale differences need not be made. Ruppert showed that for large n, the estimator X_n can be almost surely (strongly) approximated by a weighted average of $\varepsilon_1, \ldots, \varepsilon_n$, where X_n and the errors ε_n are k-dimensional random vectors. This representation and strong approximations for sums of dependent random variables immediately give several important results about the large sample behavior of X_n. We consider only the case k=1. We choose a modified version of (1.2) since Ruppert's representation is not directly applicable to statistical problems that have non-zero bias terms (β_n) infinitely often and $\tau < 1/6$, as in the example in §4. The algorithm (1.1) is in a form more related to the easily accessible algorithm given by Fabian (1968).

After a short preliminary section on notation and assumptions, results which describe the large sample behavior of the recursive estimator X_n are given in §3. The main result is weak convergence via random indices (Theorem 3.2). This result subsumes the earlier work on stopping rules in SA (see Stroup and Braun (1982) and references cited therein). A fixed-width bounded length confidence interval for the parameter θ is an easy corollary of Theorem 3.2. This corollary is the result given by Sielken (1973) and McLeish (1976) for the RM process. Sielken's method was to verify Anscombe's (1952) uniform continuity in probability condition. McLeish showed convergence of a randomized version of the RM process in D[0,1) (the space of right-continuous real functions on [0,1) having left-hand limits). The methods in this paper are different. Via strong approximation techniques, we prove convergence of a randomized version of a process derived from (1.1) in $D[0,\infty)$, a more convenient function space for stopping rules.

As an example, §4 shows to how develop a fixed-width bounded length confidence interval for the mode of a distribution function. Using SA to estimate the mode was introduced by Burkholder (1956) under restrictive conditions and more generally by Fritz (1973). As a preliminary result, we show that our proposed estimator when suitably standardized converges to a nondegenerate distribution. This preliminary result is new and should be of independent interest since it gives a method of constructing a mode estimator which has a rate of convergence similar to the optimal rate of convergence of density estimators (see Muller and Gasser (1979)). The example also demonstrates why it is important to consider a general SA algorithm since the mode estimation problem is neither an RM nor a KW procedure. The final section gives the results of an investigation of finite sample properties of the mode estimator proposed in §4 via a Monte-Carlo study.

§2 Notation and Assumptions

Assume that random elements are defined on a fixed probability space (Ω,F,P) unless otherwise specified. Endow $D[0,\infty)$ with Stone's (1963) extension of Skorokhod's (1956) J_1-topology and use $\Rightarrow_W$ to denote convergence in this topology. Also, use $=_D$ to denote equivalence in distribution. Let $[\cdot]$ be the greatest integer function. Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables (r.v.). We write $X_n=0(Y_n)$ if there exists a r.v. Z such that $|X_n|/|Y_n|\leq Z$ almost surely (a.s.) for all n. If $|X_n|/|Y_n| \to 0$ a.s., we write $X_n=o(Y_n)$. Let $0_p(\cdot)$ and $o_p(\cdot)$ be the corresponding symbols for relationships in probability.

The asymptotic coverage probability is fixed at $1-\alpha$, where $\alpha\varepsilon(0,1/2)$. We will define a sequence of random intervals $\{I_n\}$ and a stopping rule N_d so that N_d is the first n such that $\text{length}(I_n)\leq 2d$ and $\lim_{d\to 0} P(\theta\varepsilon I_{N_d})=1-\alpha$. The basic assumptions about the relationships defined in (1.1) are stated below.

A1. Let $\eta>0$ and $G>1/2-\tau$. Assume that $f(x) = G(x-\theta) + O(|x-\theta|^{1+\eta})$.

A2. Let $\rho>0$ and $\beta\in R$. Assume that $\beta_n = \beta + o(n^{-\rho})$.

A3. $\lim_{n\to\infty} X_n = \theta$ a.s.

A4. There exists a probability space which contains a sequence $\{e_n\}$ and a standard Brownian Motion B on $[0,\infty)$ such that

$$\{e_n;\ n=1,2,\ldots\} =_D \{\varepsilon_n;\ n=1,2,\ldots\}$$

and for some $\sigma, \varepsilon > 0$ depending on $\{\varepsilon_n\}$

$$\sum_{k\leq t} e_k = \sigma B(t) + O(t^{1/2-\varepsilon}).$$

A5. Let $N_n/m_n = N + o_p(1)$, where N is a positive random variable, $\{m_n\}$ are integers going to infinity and $\{N_n\}$ is a sequence of random variables.

A6. Let $\{\beta_n'\}$, $\{\sigma_n\}$ and $\{G_n\}$ be sequences of random variables such that $\beta_n' = \beta + o(1)$, $\sigma_n = \sigma + o(1)$ and $G_n = G + o(1)$.

Remarks: In A6, we allow estimators of β that do not converge to β as quickly as those in A3. These will be useful in the construction of confidence intervals. The monograph of Phillip and Stout (1975) gives several sets of sufficient conditions so that the sequence of random variables $\{\varepsilon_n\}$ can be redefined on a richer probability space without changing its distribution so that A4 holds. Thus, our formulation includes stochastic approximation processes with certain types of dependent noise and even nonstationary noise (see Phillip and Stout, 1975, Chapter 8). For notational convenience, we follow Strassen (1967) and adopt the phrase "without loss of generality" when redefining a sequence of random variables on a possibly richer probability space containing a Brownian Motion process (see Csorgo and Revesz, 1981, Remark 2.2.1). For example, assumption A4 would be written as

A4'. Without loss of generality, there exists a standard Brownian Motion process B on $[0,\infty)$ and a $\sigma, \varepsilon>0$ such that

$$\sum_{k\leq t} \varepsilon_k = \sigma B(t) + O(t^{1/2-\varepsilon}).$$

§3 SA Convergence Results

Theorem 3.2 contains the main result, weak convergence via random indices of the process $\{W_n\}$ (defined in (3.1) below) derived from (1.1). To prove that result we need:

Theorem 3.1

Consider the algorithm defined in (1.1) and assume A4. Let

$$(3.1) \qquad W_n(t) = [nt]^{1/2-\tau}(X_{[nt]+1}-\theta) + \beta/(G-1/2+\tau) \quad n=1,2,\ldots$$

Then, without loss of generality, there exists a standard Brownian Motion process B and a sequence of Gaussian processes $\{Z_n\}$ defined on $[0,\infty)$ such that

$$(3.2) \qquad W_n(t) = Z_n(t) + O(n^{-\varepsilon})$$

where $Z_n(t) = -Z(nt)$ and

$$(3.3) \qquad Z(t) = \{2(G-1/2+\tau)\}^{-1/2}\ \sigma\, t^{-(G-1/2+\tau)} B(t^{2(G-1/2+\tau)}).$$

Hence,

$$(3.4) \qquad W_n(t) \Rightarrow_W Z(t).$$

The proof of Theorem 3.1 uses the following two lemmas.

Lemma 3.1 (Ruppert, 1982, Lemma 4.1)

Assume $a>-1/2$ and A4. Then there exists a standard Brownian Motion B_a and an $\varepsilon'>0$ such that

$$\sum_{k\leq t} k^a e_k = \sigma B_a(t^{2a+1}(2a+1)^{-1}) + O(t^{a+1/2-\varepsilon'}).$$

If $a<-1/2$, then $\lim_{t\to\infty} \sum_{k\leq t} k^a e_k$ exists and

(3.5) $\quad \lim_{t\to\infty} \sum_{k\leq t} k^a e_k$ is finite a.s.

Lemma 3.2

Consider the algorithm defined in (1.1) and assume A4. Then, there exists $\varepsilon>0$ such that

$$n^{1/2-\tau}(X_{n+1}-\theta) = -\beta/(G-1/2+\tau) - n^{-1/2}\sum_{k=1}^{n}(k/n)^{G+\tau-1}\varepsilon_k + O(n^{-\varepsilon}).$$

Proof: For the one-dimensional case, (3.5) and assumption A4 imply Ruppert's (1982) assumption A4. The remainder of the proof is similar to Ruppert's Theorem 3.1. ‡

Proof of Theorem 3.1: To prove (3.2), first let $\gamma=G-1/2+\tau$. From Lemmas 3.1 and 3.2, with $a=G-1+\tau=\gamma-1/2$,

$$W_n(t) = -[nt]^{-\gamma}\sum_{k\leq[nt]} k^{G+\tau-1}\varepsilon_k + O(n^{-\varepsilon})$$

$$= -[nt]^{-\gamma}\{\sigma B([nt]^{2a+1}(2a+1)^{-1}) + O([nt]^{a+1/2-\varepsilon'})\} + O(n^{-\varepsilon}).$$

$$= -[nt]^{-\gamma}\,\sigma B([nt]^{2\gamma}(2\gamma)^{-1}) + O(n^{-\varepsilon}).$$

Since $B([nt]^{2\gamma}(2\gamma)^{-1}) = B((nt)^{(2\gamma)}(2\gamma)^{-1}) + O(1)$ and $[nt]^{-\gamma} = (nt)^{-\gamma} + O(n^{-1-\gamma})$, we have

(3.6) $$W_n(t) = -(nt)^{-\gamma}\,\sigma B((nt)^{2\gamma}(2\gamma)^{-1}) + O(n^{-\varepsilon})$$

which proves (3.2). Since $-B(t) =_D B(t)$ and $a^{-1/2} B(at) =_D B(t)$, (3.4) is immediate from (3.2). ‡

The following lemma is a modification of a result that can be found in Csorgo and Revesz (1981, Theorem 7.2.2). However, the method of proof is different and can be extended to more general situations.

Lemma 3.3

Suppose $\{S(t): t\geq 0\}$ is an element of $D[0,\infty)$ defined on (Ω,F,P). Assume A5, that there exists a Gaussian process $G(t)=t^{-a/2}B(t^a)$ for some $a>0$, that $B(\cdot)$ is a Brownian motion on $[0,\infty)$ and that

$$(3.7) \qquad n^{-1/2}\ |S(nt) - G(nt)| = o(1).$$

Then, $\{n^{-1/2}S(n.)\}$ is Renyi-mixing, i.e.,

$$(3.8) \qquad P(n^{-1/2}\ S(nt)\ \varepsilon\ A,\ t\geq 0|E) \Rightarrow_W P(G(t)\ \varepsilon\ A,\ t\geq 0)$$

for each $A,E\ \varepsilon\ F$ such that $P(E)>0$. Thus,

$$(3.9) \qquad N_n^{-1/2}S(N_n t) \Rightarrow_W G(t).$$

Proof: Fix $T>0$ and consider only $0\leq t\leq T$. Now, from (3.7),

$$P\{\lim_{n\to\infty}\sup_t |n^{-1/2}S(nt) - n^{-1/2}G(nt)| = 0\} = 1.$$

This gives, for any set E having positive probability,

$$P\{\lim_{n\to\infty}\sup_t |n^{-1/2}S(nt)-n^{-1/2}G(nt)| = 0|E\} = 1.$$

This implies, for $A\ \varepsilon\ F$,

$$(3.10) \qquad |P(n^{-1/2}S(nt)\varepsilon A, 0\leq t\leq T|E) - P(n^{-1/2}G(nt)\varepsilon A, 0\leq t\leq T|E)| \to 0.$$

It is easy to show that $n^{-1/2}G(n.)$ is Renyi-mixing, i.e., (3.11) $P\{n^{-1/2}G(nt)\varepsilon A,\ 0\leq t\leq T|E\} \Rightarrow_W P(G(t)\varepsilon A,\ 0\leq t\leq T)$.

See for example, Csorgo and Revesz (1981, Lemma 7.2.2 and subsequent remarks). Since T is arbitrary, (3.10) and (3.11) are sufficient for (3.8). (3.8) and Theorem 4 of Durrett and Resnick (1976) give (3.9). ‡

Theorem 3.2

Consider the sequence of stochastic processes $\{W_n\}$ defined in (3.1) and assume A1-A5. Then,

$$W_{N_n}(t) \Rightarrow_W Z(t).$$

Proof: A direct application of Lemma 3.3 to Theorem 3.1. Define, with $\gamma = G-1/2+\tau$,

$$G(t) = -t^{1/2-\gamma}\sigma B(t^{2\gamma}(2\gamma)^{-1})$$

$$S(nt) = (nt)^{1/2} W_n(t)$$

From (3.6), (3.7) is satisfied. Thus, from Lemma 3.3,

$$N_n^{-1/2} S(N_n t) = t^{1/2} W_{N_n}(t) \Rightarrow_W G(t).$$

Using $t^{-1/2} G(t) =_D Z(t)$ completes the proof. ‡

Remarks: It is not very difficult to show that the algorithm (1.1) with assumptions A1-A4 contains the Robbins-Monro process. See Ruppert (1982) or Ljung (1978) for proofs under different assumptions. Thus, Theorem 3.2 contains the main result of Stroup and Braun (1982, Theorem 5).

To emphasize the importance of Theorem 3.2, we now construct a sequential confidence region for the main parameter of interest, θ. Let z_α be the $(1-\alpha)^{th}$ quantile of the standard normal distribution. For random variables β_n', σ_n and G_n and for each d>0, define

$$(3.12)\quad N_d = \inf\{n\geq 1:\ d\geq z_{\alpha/2}\sigma_n n^{1(1/2-\tau)}\{2(G_n-1/2+\tau)\}^{-1/2}\}$$
$$= \infty \text{ if no such } n \text{ exists}$$

and

$$(3.13)\quad I_n = [X_n+n^{-(1/2-\tau)}(\beta_n'/(G_n-1/2+\tau)-z_{\alpha/2}\sigma_n\{2(G_n-1/2+\tau)\}^{-1/2}),$$
$$X_n+n^{-(1/2-\tau)}(\beta_n'/(G_n-1/2+\tau)+z_{\alpha/2}\sigma_n\{2(G_n-1/2+\tau)\}^{-1/2})].$$

Remark: We may define the sequence $\{n_d\}$ where $n_d = \inf\{n\geq 1:\ d\geq z_{\alpha/2}\ \sigma\ n^{-(1/2-\tau)}\{2(G-1/2+\tau)^{-1/2}\}$. It is immediate that under the assumptions of Theorem 3.1 that $\lim_{d\to 0} P(\theta\varepsilon I_{n_d})=1-\alpha$. Further, from Lemma 3.6 of McLeish (1976), we get $\lim_{d\to 0} N_d/n_d=1$ a.s.

<u>Corollary 3.3</u>

Consider the algorithm defined in (1.1) and assume A1-A4 and A6. Then,

$$\lim_{d\to 0} P(\theta\varepsilon I_{N_d}) = 1-\alpha.$$

<u>Proof:</u> From (3.12) and A6, we have,

$$d\ N_d^{1/2-\tau}\{2(G-1/2+\tau)\}^{1/2}/\sigma - z_{\alpha/2} = o(1) \text{ as } d\to 0.$$

Thus, A5 is satisfied. From Theorem 3.2 and the Continuous Mapping Theorem at t=1, we have

$$N_d^{1/2-\tau}(X_{N_d}-\theta) \Rightarrow_W N(-\beta/(G-1/2+\tau),\sigma^2/\{2(G-1/2+\tau)\}).\ \ddagger$$

Remark: The applicability of Theorem 3.2 and Corollary 3.3 for the RM process was cited in §1 and for the mode estimation problem is in §4. These results can also be used in conjunction with estimating the <u>optimal replacement time</u> in an age replacement policy, an important problem in resource allocation. Frees and Ruppert (1983) showed that a stochastic approximation estimator of the optimal replacement time has desirable asymptotic properties such as strong consistency and weak convergence to a limiting distribution (cf., Frees and Ruppert for definitions of an age replacement policy and the optimal replacement time). Corollary 3.3 can be used to determine at what stage of the estimation process the estimator is reasonably close to the optimal replacement time with high probability. This tool may be useful to the experimenter as a cost-saving device.

§4 Mode Estimation Via SA

Estimating the mode of a distribution function is a delicate problem that has drawn the attention of many researchers (cf., Eddy (1980, 1982) and Hall (1982)). As pointed out by Fritz (1973), SA is a natural vehicle to use in mode estimation. In this section we give a variation of Fritz's SA mode estimator, prove its optimality and then construct sequential fixed-width bounded length confidence intervals for the mode. Proofs of these properties appear at the end of the section.

Let $f(\cdot)$ be the probability density function of some distribution function $F(\cdot)$. When it exists, use $f^{(s)}(\cdot)$ for the s^{th} derivative of f, $s=0,1,2,\ldots$ Assume that the mode of F, $\theta = \sup_x f(x)$, is unique and finite.

Let B_o be the class of all Borel-measurable real-valued functions $k(\cdot)$, where $k(\cdot)$ is bounded and equals zero outside $[-1,1]$. Let r be a fixed positive integer. For $0 \leq s \leq r$, define

$$M_s = \left\{k \in B_o : \int_{-1}^{1} y^j/j!\, k(y)dy = \begin{bmatrix} 1 & j=s \\ 0 & j\neq s \end{bmatrix} j=0,\ldots,r-1\right\}.$$

M_s is a class of kernel functions used to estimate $f^{(s)}$ which is similar to a class used by Singh (1977).

Take X_1 to be an arbitrary random variable with finite second moments. Let $F_n = \sigma(X_1, Z_1,\ldots,Z_{n-1})$, the sigma field generated by past events at the n^{th}

stage. We assume there exist sequences of positive random variables $\{a_n\}$ and $\{c_n\}$ where a_n, c_n are F_n-measurable. Let $\{Z_n\}$ be an i.i.d. sequence of random variables such that Z_1 has density f. Define E_{F_n} to be the conditional expectation given F_n. Use $k \in M_1$ to define the estimator of $f^{(1)}(\cdot)$ by

(4.1) $$f_n^{(1)}(x) = k((Z_n - x)/c_n)/c_n^2.$$

Define the estimator of θ at the n^{th} stage, X_{n+1}, recursively by

(4.2) $$X_{n+1} = X_n + a_n f_n^{(1)}(X_n).$$

A list of the basic assumptions is collected below.

B1. For some integer $r>1$, assume $f(x)$, $f^{(r)}(x)$ exist for each x and are bounded on the entire real line.

B2. For each $x \neq \theta$, $(x-\theta)f^{(1)}(x) < 0$.

B3. For some $A,C>0$ and $\tau=3/(2(2r+1))$, $a_n n \to A$ $c_n n^{2\tau/3} \to C$ a.s.

B4. $f^{(r)}(x)$ is continuous at θ.

B5. Assume that $G > 1/2-\tau = (r-1)/(2r+1)$ where $G = -Af^{(2)}(\theta)$.

B6. For some $d>0$, $f^{(r)}(x) = f^{(r)}(\theta) + O(|x-\theta|^d)$ for each x.

<u>Lemma 4.1</u>

Consider the algorithm defined in (4.2) and assume B1-B3. Then,

(4.3) $$\lim_{n\to\infty} X_n = \theta \text{ a.s.}$$

The proof is a modification of Fritz (1973, Theorem 2). See also Frees (1983, Theorem 4.1). The weak convergence of the mode estimator is now stated.

Theorem 4.1

Consider the algorithm defined in (4.2) and assume B1-B5. Define $\beta = -A\, C^{r-1} f^{(r)}(\theta) \int_{-1}^{1} y^r/r!\, k(y)dy$, $\sigma^2 = A^2 C^{-3} f(\theta) \int_{-1}^{1} k^2(y)dy$ and $\{W_n\}$ as in (3.1). Then, for Z(t) defined in (3.3), we have

$$(4.4) \qquad W_n(t) \Rightarrow_W Z(t).$$

Remarks: Using the Continuous Mapping Theorem at t=1 gives the asymptotic normality of the mode estimator in (4.2) when suitably standardized. The rate of convergence, $1/2-\tau = (r-1)/(2r+1)$, is better than the rate recently given by Eddy (1980, Corollary 2.2) when the same number of bounded derivatives is assumed.

An immediate extension is to find the θ such that $f^{(p)}(\theta)=\alpha$ for nonnegative integer p and fixed, known constant α. For asymptotic normality of a multivariate mode estimator see Frees (1983, Theorem 4.4), in which a different method of proof (due to Fabian (1968)) is used.

To get a version of Corollary 3.3, we introduce estimators of β_n', σ_n and G_n that satisfy A6. Let $k_o \in M_o$, $k_2 \in M_2$, $k_r \in M_r$ and define $f_n^{(s)}(t) = k_s((Z_n - t)/c_n)/c_n^{s+1}$ for s=0,2,r. Now, define

$$(4.5) \qquad \beta_n' = -AC^{r-1} \int_{-1}^{1} y^r/r!\, k(y)dy \,\{n^{-1} \sum_{j=1}^{n} f_j^{(r)}(X_j)\}$$

$$(4.6) \qquad \sigma_n^2 = A^2 C^{-3} \int_{-1}^{1} k^2(y)dy \,\{n^{-1} \sum_{j=1}^{n} f_j^{(0)}(X_j)\}$$

$$(4.7) \qquad G_n = -A \,\{n^{-1} \sum_{j=1}^{n} f_j^{(2)}(X_j)\}.$$

Corollary 4.2

Consider the algorithm defined in (4.2) and assume B1-B6. Use the estimators in (4.5)-(4.7) to construct $\{N_d\}$ and $\{I_n\}$ as in (3.12) and (3.13). Then

$$\lim_{d\to 0} P(\theta \varepsilon I_{N_d}) = 1-\alpha.$$

The remainder of this section contains the proofs of Theorem 4.1 and Corollary 4.2. We use the following result due to Strassen. (Recall the convention stated in the remark following the assumptions in §2.)

Lemma 4.2 (Strassen, 1967, Theorem 4.4)

Let $\{\varepsilon_n, F_{n+1}\}$ be a martingale difference sequence. Define $Y_n = \sum_{k\leq n} E_{F_n} \varepsilon_n^2$, $S(Y_n) = \sum_1^n \varepsilon_n$, and S(t) by linear interpolation. Let h be a nonnegative, nondecreasing function on $[0,\infty)$ such that $t^{-1}h(t)$ is nonincreasing. If

(4.8) $\quad Y_n \to \infty$ a.s. and

(4.9) $\quad \sum_1^\infty E_{F_n} \{\varepsilon_n^2 I(\varepsilon_n^2 > h(Y_n))\}(h(Y_n))^{-1} < \infty$ a.s.,

then, without loss of generality, there exists a Brownian Motion process B on $[0,\infty)$ such that

(4.10) $\quad S(t)-B(t) = o((\log t)(t\ h(t))^{1/4})$.

Proof of Theorem 4.1: We need to satisfy the assumptions of Theorem 3.1. A1 is true by B1, B4 and B5. Define

(4.11) $\quad \varepsilon_n = a_n n^{1-\tau}\{E_{F_n} f_n^{(1)}(X_n) - f_n^{(1)}(X_n)\}$

(4.12) $\quad \beta_n = a_n n^{3/2-\tau} \{f^{(1)}(X_n) - E_{F_n} f_n^{(1)}(X_n)\}$.

Thus, from (4.2) with $f^*(\cdot)=-f^{(1)}(\cdot)$,

(4.13) $$X_{n+1} = X_n - n^{-1}\{a_n n\, f^*(X_n) + n^{\tau-1/2}\beta_n + n^{\tau}\varepsilon_n\}.$$

Some easy calculations show that A2 is true with β defined in the statement of the theorem. Note that some of the bias term is in $a_n n\, f^*(X_n)$ $((a_n n-A)f^*(X_n))$ but is asymptotically negligible. Lemma 4.1 satisfies A3. The main work of the theorem is to satisfy A4 for which we will Lemma 4.2.

Let $\{\varepsilon_n\}$ as in (4.11) and define Y_n as in Lemma 4.2. Since $k\varepsilon M_1$, we have that $E_{F_n} f_n^{(1)}(X_n)$ is bounded. Further, for p>0, by change of variables,

(4.14) $$E_{F_n}(f_n^{(1)}(X_n))^p = c_n^{-2p} \int k^p((s-X_n)/c_n)\, f(s)ds$$

$$= c_n^{-2p+1} \int_{-1}^{1} k^p(y)\, f(X_n+c_n y)dy.$$

Thus, with p=2, $n^{-2\tau}E_{F_n}(f_n^{(1)}(X_n))^2 \to C^{-3}f(\theta) \int k^2(y)dy$ a.s. by Lemma 4.1, B1, B3 and the Dominated Convergence Theorem. Further,

(4.15) $$E_{F_n}\varepsilon_n^2 \to A^2C^{-3}f(\theta) \int_{-1}^{1} k^2(y)dy = \sigma^2 \text{ a.s.}$$

This satisfies (4.8) since $Y_n/n\to\sigma^2$ a.s.

To satisfy (4.9), use a conditional version of Holder's inequality. Let p,q be nonnegative constants such that $2/p + 1/q = 1$. Then,

$$(h(Y_n))^{-1}\, E_{F_n}\{\varepsilon_n^2 I(\varepsilon_n^2 > h(Y_n))\}$$

$$\leq (h(Y_n))^{-1}(E_{F_n}\varepsilon_n^p)^{2/p}[E_{F_n}(I(\varepsilon_n^2 > h(Y_n)))]^{1/q}$$

$$\leq (h(Y_n))^{-1}(E_{F_n}\varepsilon_n^p)^{2/p}[E_{F_n}\varepsilon_n^p]^{1/q}[h(Y_n)]^{-p/(2q)}$$

$$= E_{F_n}\varepsilon_n^p (h(Y_n))^{-p/2}.$$

Now, by B3 and (4.14), $E_{F_n}(f_n^{(1)}(X_n))^p = 0(n^{2\tau(p-2)})$ and with (4.12),

(4.16) $\quad E_{F_n}\varepsilon_n^p = 0(n^{\tau(p-2)})$.

For some $0<\varepsilon<1/2-\tau$, let $h(t)=t^{1-\varepsilon}$. Since $Y_n=0(n)$, we have $wh(Y_n))^{-p/2}\; E_{F_n}\varepsilon_n^p = 0(n^{\tau(p-2)-p/2(1-\varepsilon)})$. Requiring that $0<\varepsilon<1/2-\tau$ implies that

$\tau(p-2)-(p/2)(1-\varepsilon)<-1$

and thus (4.9) is satisfied. We thus have (4.10) which is sufficient for A4. ‡

The proof of Corollary 4.2 is similar to the proof of Corollary 3.3 using Theorem 4.1, Theorem 3.2 and the following:

Lemma 4.3.

Consider the algorithm defined in (4.2) and assume B1-B6. Then, the estimators β_n', σ_n and G_n defined in (4.5)-(4.7) satisfy A6.

Proof: We prove only $\beta_n'=\beta + o(1)$ as the others are similar. Sufficient for this is

(4.17) $\quad n^{-1}\sum_{j=1}^{n} f_j^{(r)}(X_j) \to f^{(r)}(\theta)$ a.s.

Now, $$E_{F_n} f_n^{(r)}(X_n) = \int k_r((s-X_n)/c_n)/c_n^{r+1} f(s)ds$$

$$= c_n^{-r} \int_{-1}^{1} k_r(y)\, f(X_n + c_n y)dy$$

$$= c_n^{-r} \int_{-1}^{1} k_r(y)\, (c_n y)^r/r!\, f^{(r)}(\eta_n(y))dy$$

where $\|\eta_n(y)-X_n\| \leq c_n y$, by a Taylor-series expansion and since $k_r \varepsilon M_r$. By B6,

$$(4.18) \quad E_{F_n} f_n^{(r)}(X_n) = \int_{-1}^{1} y^r/r!\, k_r(y) \; \{f^{(r)}(\theta)+o(|\eta_n(y)-\theta|^d)\}dy$$

$$\rightarrow f^{(r)}(\theta)$$

by Lemma 4.1 and the Dominated Convergence Theorem. By Kronecker's lemma, we have

$$(4.19) \quad n^{-1}\sum_{j=1}^{n} E_{F_j} f_j^{(r)}(X_j) \rightarrow f^{(r)}(\theta) \text{ a.s.}$$

It is not hard to find a $t \varepsilon (0,2]$ such that

$$(4.20) \quad \sum j^{-t} E_{F_j} |f_j^{(r)}(X_j) - E_{F_j} f_j^{(r)}(X_j)| < \infty \text{ a.s.}$$

(4.19) and (4.20) are sufficient for (4.17) by Theorem 5 of Chow (1965). ‡

§5 Monte-Carlo Simulation

In a simple example, finite sample properties of the procedure introduced in §4 were investigated. The Gamma distribution was used with location and scale parameters 11 and 10, respectively, which produces a mean of 1.1 and standard deviation of .3317. The distribution function and parameters determine a unique θ=1.

To estimate the derivatives of the density function, we used Legendre polynomials to calculate simple, polynomial kernel estimators. For rates r=3 and r=4, these kernel estimators are given in Table 1 below. By using r=4 and Legendre polynomials, the mode estimator is asymptotically unbiased (β=0). This fact had a great impact on the performance of the stopping rules which is described below.

Table 1 Inserted Here

Let A and C be positive constants, K_A and K_C be nonnegative constants. As first suggested by Dvoretsky (1956), we used $\{a_n\}$ and $\{c_n\}$ of the form $a_n = A(n+K_A)^{-1}$ and $c_n = C(n+K_C)^{-2\tau/3}$. Taking K_A to be positive improved the convergence properties of the procedure in finite samples. Choice of the positive parameters A and C is constrained only by B5 (for the above Gamma distribution and r=3 this implies $A > (-2/7)/f^{(2)}(\theta) = .024$). One criterion for selection, introduced by Abdelhamid (1973), is to choose the A and C that minimize the asymptotic mean square error. For r=3, some easy calculations show that these optimal values are A=.08 and C=.7. Unfortunately, these optimal values are usually not available to the experimenter a priori. Tables 2-4 give results for some alternative parameters A and C (and various choices of K_A, K_C and the initial estimator X_1). In these tables, the criteria for choosing among alternative estimators based on different parameters are the expected bias and mean square error standardized by the sample size.

Another important consideration in finite samples is the choice of the appropriate confidence level (α) and confidence interval width (2d). Based on the above parametrization of the model, it is easy to calculate factors in the asymptotic mean and variance of $n^{(r-1)/(2r+1)}(X_n-\theta)$ as given in Theorem 4.1. For r=4, these are β=0.0, σ^2=.21885 and G=1.00088. Assume that these values are known, that $n^{1/3}(X_n-\theta)$ is distributed normally (not only asymptotically) and that it is desired to have a 95% confidence interval for θ with error no more than .10 (which implies $z_{\alpha/2}$=1.96 and d=.10). Based on the remarks preceding Corollary 3.3, we require a sample size at least

$$(5.1) \qquad n_d \geq z_{\alpha/2} d^{-1} \sigma / \{2(G-1/2)\}^{1/2} = 499.7.$$

Even in this simplified version, a large sample size is required for moderately precise results. Because of the large sample nature of the problem, we imposed the somewhat arbitrary constraint that the procedure not be stopped before n=200.

At the 95% level of confidence, Tables 5-7 give the results of using the stopping rule in the SA algorithm for selected values of A, C, K_A, K_C, X_1 and d. The proportion of successes column (PROPN SUCC) is the number of times $\theta \varepsilon I_{N_d}$ divided by the number of times the algorithm was actually stopped. Similarly, averages for the stopping time, N_d, and the bias at N_d were computed only over the runs stopped.

The following remarks summarize the results of the simulation study, the details of which can be found in Tables 2-7. Each simulation is based on 500 independent trials.

(1). The performance of the mode estimator can be characterized as typical of a stochastic approximation estimator. From Tables 2-4, we see that the effects of the initial estimator X_1 still persist for large n. The mode estimator performed better when the initial estimator was less than the true mode than when X_1 was greater than the mode, an important point in practice. The performance was dramatically enhanced by taking K_A to be positive, but the introduction of positive K_C only slowed the convergence of the algorithm.

(2). For r=4, the stopping rule performed well even in cases in which the mode estimator did poorly. Criteria for judging the performance of the stopping rules were proportion of successes and the average bias at the stopping time. When the convergence of the algorithm was extremely poor, the stopping rule performed slightly worse. Thus, the stopping rule should not be relied on as the sole criterion for halting the procedure when the experimenter has no knowledge of the distribution function.

(3). For r=3, the stopping rule performed well on the basis of the average bias at the stopping time. However, the performance was markedly poorer than r=4 when judged by the proportion of successes criterion. This is due to the fact that to construct the confidence interval we needed to estimate the third derivative of the density at the mode. Even the sample sizes we used were not large enough to get accurate estimators of the bias. Thus, the confidence intervals were much poorer than the asymptotically unbiased case (r=4).

Tables 2-7 Inserted Here

REFERENCES

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

Anscombe, F. (1972). Large-sample theory of sequential estimation. Proc. Cambridge Phil. Soc. 48, 600-607.

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

Chow, Y.S. (1965). Local convergence of martingales and the law of large numbers. Ann. Math. Statist. 36, 552-558.

Csorgo, M. and Revesz, P. (1981). Strong Approximations in Probability and Statistics. Akademiai Kiado, Budapest.

Durrett, R.T. and Resnick, S.I. (1977). Weak convergence with random indices. Stoch. Proc. Appl. 5, 213-220.

Dvoretsky, A. (1956). On stochastic approximation. Proc. Third Berkeley Symp. Math. Statist. Probab., 1 (J. Neyman, ed.), 39-55. Univ. California Press.

Eddy, W. (1980). Optimal kernel estimators of the mode. Ann. Statist. 8, 870-882.

Eddy, W. (1982). The asymptotic distributions of kernel estimators of the mode. Z. Wahrscheinlichkeitstheorie Verw. Gebiete. 59, 279-290.

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.), 439-460.

Farrell, R.H. (1962). Bounded length confidence intervals for the zero of a regression function. Ann. Math. Statist. 33, 237-247.

Frees, E.W. (1983). On Construction of Sequential Age Replacement Policies via Stochastic Approximation. Ph.D. Dissertation, Department of Statistics, University of North Carolina. Chapel Hill, North Carolina.

Frees, E.W. and Ruppert, D. (1983). Sequential nonparametric age replacement policies. Institute of Statistics Mimeo Series #1535. Chapel Hill, North Carolina.

Fritz, J. (1973). Stochastic approximation for finding local maxima of probability densities. Studia Sci. Math. Hungar. 8, 309-322.

Hall, P. (1982). Asymptotic theory of Grenander's mode estimator. Z. Wahrscheinlichkeitstheorie Verw. Gebiete. 60, 315-334.

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

Kushner, H.J. (1977). General convergence results for stochastic approximations via weak convergence theory. J. Math. Anal. and Appl. 61, 490-503.

Lai, T.L. and Robbins, H. (1979). Adaptive designs and stochastic approximation. Ann. Statist. 7, 1196-1221.

Ljung, L. (1978). Strong convergence of a stochastic approximation confidence. Ann. Statist. 6, 680-696.

McLeish, D.L. (1976). Functional and random central limit theorems for the Robbins-Monro process. J. Appl. Prob. 13, 148-154.

Muller, H.G. and Gasser, T. (1979). Optimal convergence properties of kernel estimates of derivatives of a density function. Smoothing Techniques for Curve Estimation. (T. Gasser and M. Rosenblatt, eds.), 144-154. Springer-Verlag, Berlin.

Phillip, W. and Stout, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. Amer. Math. Soc. Mem. No. 161, Amer. Math. Soc., Providence, RI.

Renyi, A. (1958). On mixing sequences of sets. Acta Math. Acad. Sci. Hungar. 9, 215-228.

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

Ruppert D. (1982). Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise. Ann. Statist. 10, 178-187.

Sielken, R.L. (1973). Stopping times for stochastic approximation procedures. Z. Wahrscheinlichkeitstheorie Verw. Geb. 26, 67-75.

Singh, R.S. (1977). Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. Ann. Statist. 5, 394-399.

Skorohod, A.V. (1956). Limit theorems for stochastic processes. Theor. Probab. Appl. 1, 261-290.

Stone, C.J. (1963). Weak convergence of stochastic processes defined on semi-infinite time intervals. Proc. Amer. Math. Soc. 14, 694-696.

Strassen, V. (1967). Almost sure behaviour of sums of independent random variables and martingales. Proc. Fifth Berkeley Symp. Statist. Probab. (Ed. L. LeCam et al.). Los Angeles: University of California Press, v. 2, 315-343.

Stroup, D.F. and Braun, I. (1982). On a stopping rule for stochastic approximation. Z. Wahrscheinlichkeitstheorie Verw. Geb. 60, 535-554.

TABLE 1 -- KERNEL FUNCTIONS

r = 3

$k_0(y) = .375\ (3-5y^2)\ I(-1\leq y\leq 1)$

$k\ (y) = k_1(y) = 1.5y\ I(-1\leq y\leq 1)$

$k_2(y) = 3.75(3y^2-1)\ I(-1\leq y\leq 1)$

$k_3(y) = 26.25(5y^3-3y)\ I(-1\leq y\leq 1)$

r = 4

$k_0(y) = .375(3-5y^2)\ I(-1\leq y\leq 1)$

$k\ (y) = k_1(y) = 1.875(5y-7y^3)\ I(-1\leq y\leq 1)$

$k_2(y) = 3.75(3y^2-1)\ I(-1\leq y\leq 1)$

$k_4(y) = 0$

TABLE 2 -- PERFORMANCE OF ESTIMATORS

r = 4   A = .08   C = .7

	X_1	K_A	K_C	Stage of Algorithm 250	500	2000	∞
$E(X_n-\theta)$	1.5	0	0	.126	.103	.072	0
$n^{1/3}MSE_n$				3.14	4.25	7.77	.164
$E(X_n-\theta)$		5		.093	.061	.024	0
$n^{1/3}MSE_n$				.907	.809	.517	.164
$E(X_n-\theta)$		10		.118	.075	.029	0
$n^{1/3}MSE_n$				1.02	.829	.481	.164
$E(X_n-\theta)$	1.5	0	5	.191	.172	.139	0
$n^{1/3}MSE_n$				6.22	9.07	19.5	.164
$E(X_n-\theta)$			10	.205	.185	.149	0
$n^{1/3}MSE_n$				7.29	10.7	23.4	.164
$E(X_n-\theta)$	.5	0	0	-.070	-.070	-.067	0
$n^{1/3}MSE_n$				2.84	4.31	10.3	.164
$E(X_n-\theta)$	.5	5	0	-.024	-.011	.000	0
$n^{1/3}MSE_n$				.233	.189	.160	.164
$E(X_n-\theta)$	1.33	0	0	.070	.055	.034	0
$n^{1/3}MSE_n$				1.42	1.79	3.00	.164
$E(X_n-\theta)$	1.66	0	0	.237	.208	.161	0
$n^{1/3}MSE_n$				7.16	10.2	21.0	.164

TABLE 3 -- PERFORMANCE OF ESTIMATORS

$r = 4 \quad K_A = 0 \quad K_C = 0$

	A	C	X_1	Stage of Algorithm 250	500	2000	∞
$E(X_n - \theta)$	.08	.7	1.5	.126	.103	.072	0
$n^{1/3}MSE_n$				3.14	4.25	7.77	.164
$E(X_n - \theta)$	.05			.137	.112	.073	0
$n^{1/3}MSE_n$				2.06	2.54	3.72	.148
$E(X_n - \theta)$	.10			.143	.122	.092	0
$n^{1/3}MSE_n$				4.56	6.44	13.3	.186
$E(X_n - \theta)$	.08	.5	1.5	.252	.243	.214	0
$n^{1/3}MSE_n$				13.2	19.9	45.7	.450
$E(X_n - \theta)$		1.3		.077	.058	.033	0
$n^{1/3}MSE_n$				.269	.239	.197	.026
$E(X_n - \theta)$		1.5		.112	.086	.051	0
$n^{1/3}MSE_n$				.526	.492	.428	.017
$E(X_n - \theta)$	.15	1.0	1.5	.046	.037	.027	0
$n^{1/3}MSE_n$				.910	1.25	2.47	.085
$E(X_n - \theta)$	.15	1.0	.5	-.004	-.008	-.012	0
$n^{1/3}MSE_n$				1.02	1.54	3.72	.085
$E(X_n - \theta)$	.05	.5	.5	-.152	-.140	-.119	0
$n^{1/3}MSE_n$				4.56	6.69	14.9	.401

TABLE 4 -- PERFORMANCE OF ESTIMATORS

$r = 3 \quad K_A = 0 \quad K_C = 0$

	A	C	X_1	Stage of Algorithm 250	500	2000	∞
$E(X_n - \theta)$	.08	.7	1.5	.066	.047	.024	0
$n^{2/7}MSE_n$				.151	.117	.077	.043
$E(X_n - \theta)$			.5	-.008	.002	.009	0
$n^{2/7}MSE_n$				.033	.028	.034	.043
$E(X_n - \theta)$			1.33	.047	.035	.020	0
$n^{2/7}MSE_n$				.087	.075	.059	.043
$E(X_n - \theta)$			1.66	.120	.082	.038	0
$n^{2/7}MSE_n$				4.85	.355	.176	.043
$E(X_n - \theta)$			2.00	.662	.604	.475	0
$n^{2/7}MSE_n$				11.3	14.6	22.4	.043
$E(X_n - \theta)$	.08	1.0	1.5	.134	.101	.056	0
$n^{2/7}MSE_n$				.443	.377	.257	.087
$E(X_n - \theta)$	.10	.7		.047	.033	.017	0
$n^{2/7}MSE_n$				.094	.076	.055	.045
$E(X_n - \theta)$	.15	1.0	1.5	.062	.046	.028	0
$n^{2/7}MSE_n$				.108	.088	.077	.069
$E(X_n - \theta)$	.15	1.0	.5	.026	.029	.025	0
$n^{2/7}MSE_n$				.032	.043	.064	.069

TABLE 5 -- PERFORMANCE OF STOPPING RULES

$r = 4 \quad A = .08 \quad C = .7$

X_1	K_A	K_C	d	PROPN STOPPED	PROPN SUCC	AVG N_d	AVG $(X_{N_d}-\theta)$
1.5	0	0	.100	.848	.986	909.6	.012
			.075	.578	.972	1513.	.004
	5		.100	.884	.980	976.4	.019
			.075	.542	.963	1538.	.007
	10		.100	.850	.981	1047.	.028
			.075	.476	.958	1581.	.012
1.5	0	5	.100	.788	.982	910.8	.011
			.075	.544	.982	1491.	.003
		10	.100	.758	.982	895.5	.018
			.075	.524	.954	1499.	.008
.5	0	0	.100	.902	.998	857.3	.003
			.075	.656	.979	1494.	-.002
.5	5	0	.100	.976	.988	865.6	-.004
			.075	.724	.983	1510.	-.001
1.33	0	0	.100	.910	.982	870.1	.010
			.075	.658	.973	1482.	.003
1.66	0	0	.100	.716	.980	932.9	.015
			.075	.456	.969	1505.	.005

TABLE 6 -- PERFORMANCE OF STOPPING RULES

$r = 4 \quad K_A = 0 \quad K_C = 0$

A	C	X_1	d	PROPN STOPPED	PROPN SUCC	AVG N_d	AVG $(X_{N_d}-\theta)$
.08	.7	1.5	.100	.848	.986	909.6	.012
			.075	.578	.972	1513.	.004
.05	.7	1.5	.100	.748	.960	880.9	.029
			.075	.572	.948	1328.	.018
.10			.100	.822	.985	1023.	.008
			.075	.386	.964	1662.	.002
.08	.5	1.5	.100	.242	.884	1477.	.039
			.075	0	0	0	0
	1.3		.100	1.000	.998	650.4	.052
			.075	1.000	.978	954.4	.044
			.050	.848	.917	1777.	.034
	1.5		.100	.998	1.000	1246.	.061
			.075	.980	.953	1550.	.056
			.050	.012	.500	1941.	.046
.15	1.0	1.5	.100	.980	.996	542.8	.017
			.075	.980	.996	1083.	.013
.15	1.0	.5	.100	.974	.998	538.0	.015
			.075	.974	.994	1078.	.012
.05	.5	.5	.100	.522	.954	1202.	-.011
			.075	.126	.889	1612.	-.007

TABLE 7 -- PERFORMANCE OF STOPPING RULES

$r = 3$ $K_A = 0$ $K_C = 0$

A	C	X_1	d	PROPN STOPPED	PROPN SUCC	AVG N_d	AVG $(X_{N_d}-\theta)$
.08	.7	1.5	.100	.994	.505	274.9	.064
			.075	.988	.508	379.5	.057
			.050	.912	.575	868.3	.035
		.5	.075	.998	.577	283.8	-.009
			.050	.972	.568	744.7	.003
		1.33	.075	.996	.516	329.5	.043
			.050	.938	.574	797.3	.029
		1.66	.075	.956	.446	522.2	.082
			.050	.818	.538	989.4	.050
		2.00	.075	.100	.240	1134.	.140
			.050	.050	.440	1371.	.093
.08	1.0	1.5	.075	1.000	.282	283.1	.131
			.050	1.000	.344	508.4	.103
.10	.7	1.5	.075	.996	.594	384.4	.038
			.050	.916	.600	1005.	.024
.15	1.0	1.5	.075	1.000	.590	223.0	.065
			.050	1.000	.652	518.0	.045
.15	.10	.5	.075	1.000	.648	208.9	.023
			.050	1.000	.648	475.1	.028