CONVERGENCE OF RECURSIVE ESTIMATORS WITH APPLICATIONS TO NONLINEAR REGRESSION

by

David Ruppert

University of North Carolina

ABSTRACT

Strong convergence of a class of recursive estimators is proven using a relationship between weak and uniform convergence of probability measures. Special attention is paid to recursive nonlinear regression, where the independent variables and errors may each be dependent sequences.

AMS 1970 Subject Classification: 62L20.

Key Words and Phrases: Stochastic approximation, recursive estimation, uniform convergence, nonlinear regression, dependent errors.

This research was supported by the National Science Foundation through Grant MCS78-01240.

1. Introduction. This paper is concerned with the strong convergence of recursive estimators which are generalizations of the Robbins-Monro (1951) stochastic approximation procedure. Emphasis is placed on the recursive nonlinear regression estimator developed by Albert and Gardner (1967). However, the main result, Theorem 3.1, is sufficiently general that it should be applicable to other recursive estimators. For example, the author intends to use this result to continue his study (Ruppert (1979, 1981)) of Robbins-Monro type procedures where the root of the unknown regression function varies with time.

Albert and Gardner (1967) investigated nonlinear regression problems where, for $n = 1,2,\ldots$, one observes Y_n such that for a known function F_n, an unknown vector parameter θ, and a mean-zero random variable e_n,

$$Y_n = F_n(\theta) + e_n \ . \tag{1.1}$$

They considered estimators of θ defined by the recursion

$$\hat{\theta}_{n+1} = \hat{\theta}_n + a_n[Y_n - F_n(\hat{\theta}_n)] \ , \tag{1.2}$$

where a_n is a suitably chosen vector. Although it is, of course, possible to use nonlinear least-squares methodology here, Albert and Gardner were interested in situations where the Y_n's are observed sequentially, and one needs to rapidly update one's estimate as each new observation arrives. Besides its use for "on-line" estimation, this recursive nonlinear estimator may be useful when handling large data sets and models with large numbers of parameters. Then, because of its recursive nature, the calculation of the estimator has modest storage requirements.

In their study of "optimal" values of a_n in (1.2), Albert and Gardner used a Taylor series linearization and their calculation of the "optimal" a_n in the linear case and were led to the choice:

$$a_n = B_n b_n \tag{1.3}$$

where $b_n = \dot{F}_n(\eta_n)$ ($\dot{F}_n$ is the gradient of F_n),

$$B_n = (B_0^{-1} + \sum_{j=1}^{n} b_j b_j')^{-1} , \tag{1.4}$$

and η_n is either θ_0, a guessed value of θ, or $\eta_n = \hat{\theta}_m$ for some $m \le n$. Also, B_n can be calculated recursively and without matrix inversions, except possibly for B_0; see their equation (7.45).

Albert and Gardner do not actually prove that the algorithm converges to θ for this value of a_n, but instead they analyze a different algorithm. Let P be a convex set and let $[\ ]_P$ denote the operation of projection into P. Then, they find sufficient conditions for $\tilde{\theta}_n$ defined by

$$\tilde{\theta}_{n+1} = [\tilde{\theta}_n + a_n[Y_n - F_n(\tilde{\theta}_n)]]_P$$

to converge when a_n is given by (1.3). One of these conditions is troublesome: P must lie within the ball of radius R centered at θ. The value of R is not given but can be found by examining their proof. When R is small, one will need good prior knowledge of θ.

Also, the need to project into P complicates the algorithm, perhaps unnecessarily in most applications.

Theorem 2.4.1 of Kushner and Clark (1978) could be used to show that $\hat{\theta}_n$ converges, if one could show that $\sup_n ||\hat{\theta}_n|| < \infty$, but this condition seems difficult to establish.

In this paper, we will suppose that $F_n(\theta) = F(Z_n,\theta)$, where F is a function from $\mathbb{R}^{q-1} \times \mathbb{R}^p$ to $\mathbb{R}$ and Z_n is a known vector. However, we will not require any prior knowledge of θ. Also, we will not require that $e_1, e_2, \ldots$ be independent (or even uncorrelated). Since recursive estimation is often used when the data form a time series, correlated errors should be considered.

The proof of Theorem 3.1 utilizes a result by Ranga Rao (1962) on the relationship between weak and uniform convergence of probability measures. We are able to conclude that certain weighted averages (averaged with respect to n) of functions of Z_n and θ converge uniformly for θ in $\mathbb{R}^p$ (not just on compact subsets). This may be the first use in the recursive estimation literature of uniform convergence of measures, though, of course, use has been made elsewhere in statistical large sample theory.

2. Notation and assumptions. All random variables are defined on a probability space $(\Omega,\mathcal{F},P)$, and all relations between random variables are meant to hold with probability 1. Let $(\mathbb{R}^k,\mathcal{B}^k)$ be k-dimensional Euclidean space with the Borel σ-algebra. All functions which we consider between Euclidean spaces are assumed to be Borel measurable. Let a prime denote matrix transposition. For a real matrix A, $||A|| = (\text{Trace } A'A)^{\frac{1}{2}}$. We will need the following assumptions, which are discussed below.

A1. $h(\cdot,\cdot)$, $h_1(\cdot,\cdot)$, and $h_2(\cdot)$ are functions from $\mathbb{R}^p \times \mathbb{R}^q$ to $\mathbb{R}^p$, $\mathbb{R}^p \times \mathbb{R}^q$ to $\mathbb{R}^{p\times r}$, and $\mathbb{R}^p$ to $\mathbb{R}^r$ respectively such that

$$h(x,\xi) = h_1(x,\xi)h_2(x) \ .$$

A2. For each n, H_n is a $p\times p$ positive definite symmetric random matrix. For positive random variables $\underline{\lambda} \le \bar{\lambda}$, all eigenvalues of H_n are between $\underline{\lambda}$ and $\bar{\lambda}$ for all n.

A3. Suppose μ is a probability measure on $(\mathbb{R}^q,\mathcal{B}^q)$ and E_μ denotes expectation with respect to μ.

A4. Let $\xi_1,\xi_2,\ldots$ be a sequence of random vectors in $\mathbb{R}^q$. Suppose there exists a constant c such that $E_\mu g = 0$ and $E_\mu g^2 = 1$ implies that

$$E \max_{m\le\ell\le n} \Big(\sum_{i=m}^{\ell} c_i\, g(\xi_i)\Big)^2 \le c \sum_{i=m}^{n} c_i^2 \tag{2.1}$$

for all $n > m$ and constants $c_m,\ldots,c_n$.

A5. There exists a nonnegative continuous function h_3 on $\mathbb{R}^q$ such that (i) $E_\mu h_3^2 < \infty$ and (ii) $||h_1(x,\xi)|| \le h_3(\xi)$ for all x and ξ.

A6. $\{h_1(x,\cdot)\}_{x\in\mathbb{R}^p}$ is an equicontinuous family on $\mathbb{R}^q$.

REMARKS. In our application to nonlinear regression, ξ_i will be the vector formed from the error and the independent variables from the i^{th} observation. We are assuming the ξ_i are random, but of course by including degenerate distributions this allows the possibility that the independent variables are chosen by design. If $\xi_1,\xi_2,\ldots$ is a random sequence with a stationary marginal distribution μ, then we can verify A4 in the case of independence using Kolmogorov's inequality, and for weak dependence by a theorem of McLeish (1975, Theorem 1.6).

In Lemma 4.1, A4 is verified when the errors are iid and the independent variables are nonrandom (have degenerate distributions) and are periodic; that is, for some M, the independence variables of the n^{th} and m^{th} observations are equal if $n = m$ modulo M. Thus, the independent variables are selected by a design which repeats a set of M (not necessarily all distinct) values. Then μ is the product of the measure placing mass M^{-1} on each of these values and the error distribution.

The decomposition of h into h_1 and h_2 allows some flexibility in the application of the result of Ranga Rao (1962) on uniform convergence.

NOTATION. Define $\bar{h}_1(x) = \int h(x,\xi)d\mu(\xi)$ and $\bar{h}(x) = \bar{h}_1(x)h_2(x) = \int h(x,\xi)d\mu(\xi)$.

A7. h_2 is bounded in a neighborhood of 0.

A8. For all $\varepsilon > 0$,

$$\inf_{||x||>\varepsilon} \min\{||\bar{h}_1(x)||, ||\bar{h}(x)||\} > 0 .$$

A9. Suppose that $\bar{h}$ is the gradient of V, (i)

$$\inf_{||x||>\varepsilon} (V(x) - V(0)) > 0 \text{ for all } \varepsilon > 0 ,$$

and (ii) with $\ddot{V}$ the Hessian of V,

$$||\ddot{V}(x)|| \leq M \text{ for some } M \text{ and all } x .$$

REMARK. The assumption of a bounded Hessian is common in the literature of multidimensional stochastic approximation. See e.g. Fabian (1971).

A10. There exists a nonnegative function h_4 on $\mathbb{R}^q$ such that $E_\mu h_4^2 < \infty$, and for all ξ, x, and x',

$$||h(x,\xi) - h(x',\xi)|| \le h_4(\xi)||x - x'|| .$$

NOTATION. Suppose $\alpha > 2$, define $n(k)$ to be the integer part of k^α, and define $\rho_k = \sum_{i=n(k)}^{n(k+1)-1} i^{-1}$.

A11. $\sup_{n(k)\le \ell\le n(k+1)-1} ||H_\ell - H_{n(k)}|| = o(1)$ as $k \to \infty$.

3. General results.

LEMMA 3.1. *Suppose* A3 *and* A4 *hold. For* ℓ *in* $\{n(k),\ldots,n(k+1)-1\}$, *define the random probability measure* $\mu_{k,\ell}$ *by*

$$\mu_{k,\ell}(A) = \rho_k^{-1}\{\sum_{i=n(k)}^{\ell} i^{-1} I(\xi_i \in A) + \sum_{i=\ell+1}^{n(k+1)-1} i^{-1} \mu(A)\}$$

for A *in* $\mathcal{B}^q$. *(Here* I(B) *is the indicator function of the set* B.*) Then,* $\mu_{k,\ell}$ *converges weakly to* μ *as* $n(k) + \ell \to \infty$. *(The indices* (k,ℓ) *can be ordered into a sequence according to the magnitude of* $n(k)+\ell$.*)*

PROOF. Suppose $\int g^2\, d\mu < \infty$. Then by A4,

$$E \max_{n(k)\le \ell\le n(k+1)-1} \left(\int g d\mu_{k,\ell} - \int g d\mu\right)^2$$

$$= O(\rho_k^{-2} \sum_{\ell=n(k)}^{n(k+1)-1} i^{-2}) = O(k^{1-\alpha}) .$$

Thus since $\alpha > 2$,

$$\sum_{k=1}^{\infty} E \max_{n(k)\le \ell\le n(k+1)-1} \left(\int g d\mu_{k,\ell} - \int g d\mu\right)^2 < \infty ,$$

whence

$$\max_{n(k)\le \ell\le n(k+1)-1} \left(\int g d\mu_{k,\ell} - \int g d\mu\right) \to 0 \text{ as } k \to \infty .$$

By Theorem 6.6 of Parathasarthy (1968), there exists a sequence of bounded uniformly continuous functions $g_1, g_2, \ldots$ such that, for any measures $\{\nu_n\}_{n=1}^{\infty}$ and ν on $(\mathbb{R}^q, \mathcal{B}^q)$, we have $\nu_n \to \nu$ weakly if and only if $\int g_\ell \, d\nu_n \to \int g_\ell \, d\nu$ as $n \to \infty$ for each ℓ. The lemma follows. □

LEMMA 3.2. *As* $k \to \infty$,

$$\sup_{x \in \mathbb{R}^q} \max_{n(k)\le \ell\le n(k+1)-1} \rho_k^{-1} \left| \sum_{i=n(k)}^{\ell} i^{-1} (h_1(x,\xi_i) - \bar{h}_1(x)) \right| \to 0 .$$

PROOF. The lemma follows from A4, A5, A6, Lemma 3.1, and Theorem 3.2 of Ranga Rao (1962). □

THEOREM 3.1. *Under* A1 *to* A11, $x_n \to 0$ *where* x_n *is defined by the recursion*

$$x_{n+1} = x_n - n^{-1} H_n h(x_n, \xi_n) .$$

PROOF. For $n(k) + 1 \le \ell \le n(k+1) - 1$,

$$
\begin{aligned}
x_{\ell+1} &= x_{n(k)} - \sum_{i=n(k)}^{\ell} i^{-1} H_i \bar{h}(x_{n(k)}) \\
&\quad - \sum_{i=n(k)}^{\ell} i^{-1} H_{n(k)} (h(x_{n(k)}, \xi_i) - \bar{h}(x_{n(k)})) \\
&\quad - \sum_{i=n(k)}^{\ell} i^{-1} (H_i - H_{n(k)}) (h(x_{n(k)}, \xi_i) - \bar{h}(x_{n(k)})) \qquad (3.1) \\
&\quad - \sum_{i=n(k)}^{\ell} i^{-1} H_i (h(x_i, \xi_i) - h(x_{n(k)}, \xi_i)) \\
&= x_{n(k)} - R_{k,\ell} - S_{k,\ell} - T_{k,\ell} - U_{k,\ell} , \qquad \text{say .}
\end{aligned}
$$

By A2,

$$
||R_{k,\ell}|| \le \bar{\lambda}\rho_k ||\bar{h}(x_{n(k)})|| \qquad (3.2)
$$

and

$$
||R_{k,n(k+1)}|| \ge \underline{\lambda}\rho_k ||\bar{h}(x_{n(k)})|| . \qquad (3.3)
$$

By A2 and Lemma 3.2,

$$
||S_{k,\ell}|| = o(\rho_k ||\bar{h}_2(x_{n(k)})||) . \qquad (3.4)
$$

By A11,

$$
||T_{k,\ell}|| = o\Big(\sum_{i=n(k)}^{n(k+1)-1} i^{-1} (||h(x_{n(k)}, \xi_i)|| + ||\bar{h}(x_{n(k)})||) \Big) .
$$

Moreover, by A1, A4, and A5,

$$\sum_{i=n(k)}^{n(k+1)-1} i^{-1}||h(x_{n(k)},\xi_i)||$$

$$\leq ||h_2(x_{n(k)})|| \sum_{i=n(k)}^{n(k+1)-1} i^{-1} h_3(\xi_i)$$

$$= O(\rho_k||h_2(x_{n(k)})||)$$

and

$$\sum_{i=n(k)}^{n(k+1)-1} i^{-1}||\bar{h}(x_{n(k)})|| = O(\rho_k||\bar{h}_2(x_{n(k)})||) .$$

Therefore,

$$||T_{k,\ell}|| = o(\rho_k||\bar{h}_2(x_{n(k)})||) . \tag{3.5}$$

Next, by A4 and A10,

$$||U_{k,\ell}|| = O(\rho_k \max_{n(k)\leq i\leq \ell-1} ||x_i - x_{n(k)}||) . \tag{3.6}$$

By (3.1), (3.2), and (3.4)-(3.6),

$$||x_\ell - x_{n(k)}|| \leq M_k\rho_k(||\bar{h}_2(x_{n(k)})|| + \max_{n(k)\leq i\leq \ell-1} ||x_i - x_{n(k)}||) ,$$

where $M_k = O(1)$. Then by induction,

$$\max_{n(k)\leq \ell\leq n(k+1)-1} ||x_\ell - x_{n(k)}|| \leq \frac{M_k\rho_k||\bar{h}_2(x_{n(k)})||}{1 - M_k\rho_k} \tag{3.7}$$

for all k so large that $M_k\rho_k < 1$. By (3.6) and (3.7),

$$||U_{k,\ell}|| = O(\rho_k^2||\bar{h}_2(x_{n(k)})||) .$$

Now (3.1), (3.4), (3.5), and (3.7) imply that

$$x_{n(k+1)} = x_{n(k)} - \sum_{i=n(k)}^{n(k+1)-1} i^{-1} H_i \bar{h}(x_{n(k)}) + o(\rho_k ||\bar{h}_2(x_{n(k)})||^2 .$$

By A5, $\bar{h}_1(x)$ is bounded, so by A9(ii), there exist $\varepsilon_k \downarrow 0$ such that

$$V(x_{n(k+1)}) \le V(x_{n(k)}) - \underline{\lambda}\rho_k ||\bar{h}(x_{n(k)})||^2 + \varepsilon_k \rho_k ||\bar{h}_2(x_{n(k)})||^2 .$$

Now choose $\alpha_k \downarrow 0$ such that $\varepsilon_k/\alpha_k \to 0$ and $\sum \rho_k \alpha_k^2 = \infty$, then choose $\beta_k \downarrow 0$ such that $||x|| > \beta_k$ implies that $||h_1(x)||^2 > \alpha_k$ and $||\bar{h}(x)||^2 > \alpha_k$, and finally find $\gamma_k \downarrow 0$ such that $||x|| \le \beta_k$ implies that $V(x) - \underline{\lambda}\rho_k ||\bar{h}(x)||^2 + \varepsilon_k \rho_k ||\bar{h}_2(x)||^2 \le \gamma_k + V(0)$. This can be done since $\rho_k \approx \alpha n^{-1}$, and by A7 and A8. Then, for k sufficiently large,

$$V(x_{n(k+1)}) - V(0) \le \max\{\gamma_k, V(x_{n(k)}) - V(0) - \underline{\lambda}\rho_k \alpha_k^2/2\} .$$

Lemma 1 of Derman and Sacks (1959) implies that $V(x_{n(k)}) \to V(0)$, and then $x_{n(k)} \to 0$ by A9(i), whence $x_n \to 0$ by (3.7). □

4. Application to nonlinear regression. In this section, we prove consistency of $\hat{\theta}_n$ given by the assumptions:

B1. Suppose (1.1) holds with $F_n(\theta) = F(Z_n,\theta)$, where F is a known function on $\mathbb{R}^{q-1} \times \mathbb{R}^p$ and the "independent variable," Z_n, is a known element of $\mathbb{R}^{q-1}$.

B2. Suppose $b(Z,\theta)$ is the gradient of $F(Z,\theta)$ with respect to θ and $b_n(\theta) = b(Z_n,\theta)$. Let H_n be a sequence of positive definite, symmetric matrices satisfying A2 and A11. Suppose (1.2) holds with $a_n = n^{-1} H_n b_n$.

B3. Let $\mu = \mu_1 \times \mu_2$, where μ_1 and μ_2 are probability measures on $\mathbb{R}^{q-1}$ and $\mathbb{R}$ respectively. Define $\xi_n' = (z_n', e_n)$. Assume that μ and $\xi_1, \xi_2, \ldots$ satisfy A4.

NOTATION. Define

$$h(z,e,x) = (F(z,x) - F(z,\theta) - e)b(z,x) ,$$

$$V(x) = \tfrac{1}{2} \int (F(z,x) - F(z,\theta) - e)^2 \, d\mu(z,e) , \tag{4.1}$$

and

$$\bar{h}(x) = \int h(z,e,x) d\mu(z,e) .$$

B4. Assume that $\bar{h}(x)$ is the gradient of $V(x)$, i.e. that the RHS of (4.1) can be differentiated under the integral sign. Also assume that h, V, and μ satisfy A1 and A5 to A10, and that for all x, $||b(z,x)|| \leq h_5(z)$ for a nonnegative function such that

$$\int (h_5(z))^4 \, d\mu_1(z) < \infty . \tag{4.2}$$

REMARKS. It is desirable to know when A2 and A11 hold if $H_n = nB_n$ and B_n is defined by (1.4). Define

$$H(x) = \int b(z,x)(b(z,x))' \, d\mu_1(z) .$$

Suppose there exists positive constants $\underline{\lambda}$ and $\bar{\lambda}$ such that, for all x, all eigenvalues of H(x) lie between $\underline{\lambda}$ and $\bar{\lambda}$. Suppose, also, that $||b(z,x)||^4 \leq h_6(z)$, where $\int h_6(z) d\mu_1 < \infty$, that $\{b(\cdot,x)(b(\cdot,x))' : x \in \mathbb{R}^p\}$ is an equicontinuous family on $\mathbb{R}^{q-1}$, and η_ℓ (as in equation (1.4)) is equal to $x_{n(k)}$ if ℓ is in

$\{n(k),\ldots,n(k+1)-1\}$. Then, using a proof like that of Lemma 3.2, one can show that A2 and A11 hold. In the special case of linear regression, $b(z,x)$ does not depend upon x, so the value of η_ℓ is not relevant.

When z_n is a nondegenerate random vector, then B3 essentially implies that z_n and e_n are independent for each n, a condition typically used in regression analysis. Also, A4 can often be verified using the results mentioned at the end of Section 2. If the z_n are degenerate random vectors, then the next lemma may be useful in the verification of A4. It covers the situation where the errors are iid and either (i) the independent variables are selected by repeating some finite design or (ii) the Y_n form a time series with a periodic mean function.

LEMMA 4.1. *Let* M *be a positive integer and let* $\alpha_1,\ldots,\alpha_M$ *be elements of* $\mathbb{R}^{q-1}$. *Suppose* $z_n = \alpha_j$ *if* $n = j$ modulo M. *Let* $\{e_n\}_{n=1}^{\infty}$ *be iid random variables. Define the probability measure* μ *on* $\mathbb{R}^q$ *by*

$$\mu(A) = \frac{1}{M}\sum_{j=1}^{M} P((\alpha_j',e_1)' \in A) ,$$

i.e. μ *is the product of the marginal distribution of* e_1 *with counting measure on* $\{\alpha_1,\ldots,\alpha_M\}$. *Suppose* $Eg^2(\alpha_j,e_1) < \infty$ *for some function* g *from* $\mathbb{R}^q$ *to* $\mathbb{R}$ *and* $j = 1,\ldots,M$. *Then* $\xi_n = (z_n',e_n)$ *and* μ *satisfy* (2.1).

PROOF. Let $I_{n,j}$ be 0 or 1 according to whether $n = j$ modulo M or not. Then, for ℓ in $\{n(k),\ldots,n(k+1)-1\}$,

$$\left|\sum_{i=n(k)}^{\ell} i^{-1}(g(z_i,e_i) - E_\mu g)\right| \le \left|\sum_{j=1}^{M} R_{k,\ell,j}\right| + |S_{k,\ell}| ,$$

where setting $g(\alpha_j) = Eg(\alpha_j, e_1)$,

$$R_{k,\ell,j} = \sum_{i=n(k)}^{\ell} i^{-1}(g(\alpha_j, e_n) - g(\alpha_j))I_{i,j}$$

and

$$S_{k,\ell} = \sum_{j=1}^{M} \sum_{i=n(k)}^{\ell} i^{-1}(g(\alpha_j) - E_\mu g)I_{ij} .$$

Now, since $E_\mu g = \frac{1}{M} \sum_{j=1}^{M} g(\alpha_j)$,

$$\max_{n(k)\le\ell\le n(k+1)} S_{k,\ell}^2 = O(\sum_{i=n(k)}^{n(k+1)-1} i^{-2})$$

by straightforward approximations, and

$$E(\max_{n(k)\le\ell\le n(k+1)} R_{k,\ell,j})^2 = O(\sum_{i=n(k)}^{n(k+1)-1} i^{-2})$$

by Kolmogorov's inequality. □

REMARK. The requirement that $\{e_n\}$ be iid could be weakened considerably, but we will not pursue this matter here.

THEOREM 4.2. *Under* B1 *to* B4, $\hat{\theta} \to \theta$.

PROOF. Without loss of generality, we may take $\theta = 0$. Then the theorem follows from Theorem 3.1. □

EXAMPLE (linear regression). Suppose $Y_i = z_i'\theta + e_i$ and $\xi_n' = (z_n', e_n)$ satisfies B3. Suppose

$$\int (e^2 + ||z||^2) d\mu(z,e) < \infty$$

and define

$$\Sigma_\mu = \int zz' \, d\mu(z,e) \ .$$

Then let $r = 2$,

$$V(\theta) = \tfrac{1}{2}(x-\theta)' \, \Sigma_\mu (x-\theta) \ ,$$

$$h(z,e,x) = (z'(x-\theta) + e)z \ ,$$

$$\bar{h}(x) = \Sigma_\mu (x-\theta) \ ,$$

$$h_1(z,e,x) = [\min\{||x - \theta||^{-1}, 1\} zz'(x-\theta) \quad ze] \ ,$$

$$h_2(x) = [\max(||x - \theta||, 1) \quad 1]' \ ,$$

$$h_3(z,e) = (||z||^4 + ||z||^2 e^2)^{\frac{1}{2}} \ ,$$

and

$$h_4(z,e) = ||z||^2 \ .$$

One may check that Theorem 4.2 applies here. □

REFERENCES

ALBERT, A.E. and GARDNER, L.A., JR. (1967). *Stochastic Approximation and Nonlinear Regression*. The M.I.T. Press, Cambridge, Mass.

DERMAN, C. and SACKS, J. (1959). On Dvoretzky's stochastic approximation theorem. *Ann. Math. Statist.* 30 601-605.

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

MCLEISH, D.L. (1975). A maximal inequality and dependent strong lows. *Ann. Prob.* 3 829-839.

PARATHASARTHY, K.R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

RANGA RAO, R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.* 33 659-680.

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

RUPPERT, D. (1978). Stochastic approximation of an implicitly defined function. *Institute of Statistics Mimeo Series #1164*, University of North Carolina at Chapel Hill. (To appear in *Ann. Statist.*)

RUPPERT, D. (1979). A new dynamic stochastic approximation procedure. *Ann. Statist.* 7 1179-1195.