A REPRESENTATION THEOREM FOR GENERALIZED ROBBINS-MONRO PROCESSES AND APPLICATIONS

David Ruppert[1]
Department of Statistics
University of North Carolina at Chapel Hill

Running title: Robbins-Monro Processes

Summary

Many generalizations of the Robbins-Monro process have been proposed for the purpose of recursive estimation. In this paper, it is shown that for a large class of such processes, the estimator can be represented as a sum of possibly dependent random variables, and therefore the asymptotic behavior of the estimate can be studied using limit theorems for sums. Two applications are considered: robust recursive estimation for autoregressive processes and recursive nonlinear regression.

[1] Supported by National Science Foundation Grant MCS81-00748.

1. Introduction. Stochastic approximation was introduced with Robbins and Monro's (1951) study of recursive estimation of the root, θ, to $R(x) = 0$, when the function R from $\mathbb{R}$ (the real line) to $\mathbb{R}$ is unknown, but when for each x, an unbiased estimator of $R(x)$ can be observed. Their procedure lets $\tilde{\theta}_1$ be an initial estimate of θ and defines $\tilde{\theta}_n$ recursively by

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - a_n Y_n ,$$

where a_n is a suitable positive constant and Y_n is an unbiased estimate of $R(\tilde{\theta}_n)$. Blum (1954) extended their work to the multidimensional case: that is, when R maps $\mathbb{R}^p$ (p-dimensional Euclidean space) to $\mathbb{R}^p$.

A situation where the classical Robbins-Monro (RM) procedures are inadequate occurs when there is a different function (now mapping $\mathbb{R}^p$ to $\mathbb{R}$) at each stage of the recursion. Letting R_n be the function at the n^{th} stage, there may be multiple solutions to

$$R_n(x) = 0 , \tag{1.1}$$

but it is assumed that there exists a unique θ solving (1.1) for all n. At the n^{th} stage, we can, for any x in $\mathbb{R}^p$, observe an unbiased estimate of $R_n(x)$. In this context, we will define a generalized Robbins-Monro (GRM) procedure to be a process satisfying

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - a_n Y_n U_n , \tag{1.2}$$

where $\tilde{\theta}_1$ is an arbitrary initial estimate of θ, $\{a_n\}$ is a sequence of positive constants, Y_n is an unbiased estimate of $R_n(\tilde{\theta}_n)$, and U_n is a suitably chosen random vector in $\mathbb{R}^p$.

GRM processes are used by Albert and Gardner (1967) for recursive estimation of nonlinear regression parameters. They assume that one observes a process $\{Y_n: n = 1,2,\ldots\}$ such that $EY(n) = F_n(\theta)$, where F_n is a known map of $\mathbb{R}^p$ to $\mathbb{R}$ and θ is an unknown parameter vector which is to be estimated. If $R_n(x) = F_n(x) - F_n(\theta)$, then θ solves (1.1) for all n.

Another example of a GRM process appears in Campbell (1979), who considers robust recursive estimation for autoregressive processes. Suppose $\theta = (\theta_1,\ldots,\theta_p)'$ is a vector parameter and the process $\{Y_n: n = 1,2,\ldots\}$ satisfies

$$Y_n = \sum_{k=1}^{p} Y_{n-k}\theta_k + u_n ,$$

where $\{u_n\}_{n=-\infty}^{\infty}$ are iid. Assume ψ, which maps $\mathbb{R}$ to $\mathbb{R}$, satisfies $E\psi(s+u_1) = 0$ if and only if $s = 0$. Then for $x = (x_1,\ldots,x_p)'$, let

$$R_n(x) = \psi(Y_n - \sum_{k=1}^{p} Y_{n-k}x_k) .$$

In this case, because the Y's are random for each x, $R_n(x)$ is a random variable. (Albert and Gardner's setup allows $R_n(x)$ to be either random or fixed; the situation is analogous to regression where the independent variables may be random or fixed.) Except in degenerate situations,

$$P\{\text{there exists } x \neq 0 \text{ satisfying } R_n(x) = 0 \text{ for all } n\} = 0 .$$

A third example of a GRM is given by Ruppert (1979, 1981a).

The classical Robbins-Monro process has been studied quite extensively,

but little beyond consistency has been established for GRM procedures. Albert and Gardner (1967) investigate asymptotic distributions for $p = 1$ but state that "owing to its considerable complexity, the question of large-sample distribution for the vector estimates is not examined." Campbell (1979) does not treat asymptotic distributions, and Ruppert (1981a) proves asymptotic normality of $n^{\frac{1}{2}}(\tilde{\theta}_n - \theta)$ (only in his special case) under restrictions which imply that for each fixed x, $\{R_n(x): n = 1,2,...\}$ are iid, that $\{(Y_n - R_n(\tilde{\theta}_n)): n = 1,2,...\}$ are iid, and that $\{R_n(x): n = 1,2,...$ and $x \in \mathbb{R}^p\}$ is independent of $\{(Y_n - R_n(\tilde{\theta}_n)): n = 1,2,...\}$.

The classical studies of stochastic approximation procedures -- for example, Chung (1954), Sacks (1958), and Fabian (1968) -- assume that the sequence $Y_n - R(\tilde{\theta}_n)$ forms a martingale difference sequence, but recent authors, including Ljung (1978) and Kushner and Clark (1978), have felt that this assumption can be unduly restrictive in applications. Ljung (1978) proves consistency, and Kushner and Clark (1978) give a detailed treatment of consistency and asymptotic distributions for stochastic approximation methods, under assumptions considerably weaker than that the errors are martingale differences. Ruppert (1978) proves almost sure representations for multidimensional Robbins-Monro and Kiefer-Wolfowitz (1952) processes with weakly dependent errors. These representations can be used to establish almost sure behavior, i.e. laws of the iterated logarithm and invariance principles, as well as asymptotic distributions. Invariance principles are useful for the study of stopping rules; see Sielken (1973).

For GRM procedures, the results of Kushner and Clark (1978) -- in particular, their theorems 2.4.1, 2.5.1, and 2.5.2 -- are sufficient to prove consistency under weak conditions on $(Y(n) - R_n(\tilde{\theta}(n))$, but they do not appear adequate to handle asymptotic distributions.

In this paper, the techniques of Ruppert (1978) are used to show that GRM processes can be almost surely approximated by a sequence of partial sums of random vectors, and therefore, central limit theorems, invariance principles, laws of the iterated logarithm, and other results for such sums can also be established for GRM processes. The procedures of Campbell (1979) and Albert and Gardner (1967) are used as examples. In future work, these results will be applied to the procedure of Ruppert (1979, 1981a).

2. Notation and assumptions. All random variables are defined on a probability space (Ω,F,P). All relations between random variables are meant to hold with probability 1. $(\mathbb{R}^k,\mathcal{B}^k)$ is k-dimensional Euclidean space with the Borel σ-algebra. Let $X^{(i)}$ be the i^{th} coordinate of the vector X, $A^{(ij)}$ be the i,j^{th} entry of the matrix A, A' be the transpose of A, I_k be the $k\times k$ identity matrix, and $||A||$ be the Euclidean norm of the matrix A, i.e. $||A|| = (\sum_{i=1}^{n}\sum_{j=1}^{m}(A^{(ij)})^2)^{\frac{1}{2}} = \text{trace}(A'A)^{\frac{1}{2}}$. If A is a square matrix, then $\lambda_*(A)$ and $\lambda^*(A)$ denote the minimum and maximum of the real parts of the eigenvalues of A, respectively.

Let I(A) be the indicator of the set A. Throughout, K will denote a generic positive constant.

For a square matrix M, exp(M) is given by the usual series definition:

$$\exp(M) = \sum_{n=0}^{\infty} M^n(n!)^{-1} ,$$

and for $t > 0$,

$$t^M = \exp((\log t)M) .$$

Let p be a fixed integer. For convenience, we will often abbreviate I_p by I.

Notice that $t^I = tI$, $t^M s^M = s^M t^M = (st)^M$, $(t^{-1})^M = t^{-M}$, $Mt^M = t^M M$, and $(d/dt)t^M = Mt^{M-I}$.

We write $X_n = o(Y_n)$ or $X_n << Y_n$ if the random vectors X_n and Y_n satisfy $||X_n|| \le X||Y_n||$ for a (finite) random variable X.

We will be needing the following assumptions.

A1. $M_n(X,\omega)$ $(= M_n(X))$ is a $(\mathbb{R}^p \times \Omega, \mathcal{B}^p \times \mathcal{F})$ to $(\mathbb{R}^p, \mathcal{F})$ measurable transformation, $A_n(\omega)$ $(= A_n)$ is a random matrix, v is a nonnegative Borel function on $\mathbb{R}^p$, θ is in $\mathbb{R}^p$, and B_n is a nonnegative random variable such that

$$v(x) = o(x-\theta) \quad \text{as} \quad x \to \theta \tag{2.1}$$

and

$$||M_n(x) - A_n(x-\theta)|| \le B_n v(x) \tag{2.2}$$

for all x in $\mathbb{R}^p$ and all n.

A2. $\tilde{\theta}_{n+1} = \tilde{\theta}_n - n^{-1}(M_n(\tilde{\theta}_n) + W_n)$, where W_n is a random vector in $\mathbb{R}^p$.

A3. $\tilde{\theta}_n \to \theta$.

A4. A is a $p \times p$ matrix such that $\lambda_*(A) > \frac{1}{2}$ and C_1 is a constant such that

$$\sum_{k=1}^{\infty} k^{-1}(A_k - A) \text{ converges} \tag{2.3}$$

and

$$\lim_{m\to\infty} \sup_{m\le n} \sum_{k=m}^{n} k^{-1}(||A_k - A|| - C_1) \le 0 . \tag{2.4}$$

A5. Suppose

$$\sup_n E(||A_n||^2 + B_n^2) < \infty , \tag{2.5}$$

$$\sum_{k=1}^{\infty} k^{-\frac{1}{2}-\varepsilon} W_k \text{ converges for all } \varepsilon > 0 , \tag{2.6}$$

and for a constant C_2,

$$\lim_{m\to\infty} \sup_{m\le n} \sum_{k=m}^{n} k^{-1}(||B_k|| - C_2) \le 0 . \tag{2.7}$$

A6. There exists a constant C_3 such that for all $n > m$, α real, and matrices Q_i such that $||Q_i|| \le 1$, we have

$$E\{\max||\sum_{i=m}^{k} i^{\alpha} Q_i(A_i - A)||^2 : m \le k \le n\} \le C_3 \sum_{i=m}^{n} i^{2\alpha} \tag{2.8}$$

and

$$E\{\max||\sum_{i=m}^{k} i^{-1} W_i||^2 : m \le k \le n\} \le C_3 \sum_{i=m}^{n} i^{-2} . \tag{2.9}$$

A7. For some $\delta > 0$,

$$v(x) = O(||x||^{1+\delta}) \text{ as } x \to 0 .$$

REMARKS. The relationship between (1.2) and A2 is that $M_n(x) = a^{-1} R_n(x) U_n$ (a is some positive number), $W_n = a^{-1}(Y_n - R_n(\tilde{\theta}_n))U_n$, and $a_n = an^{-1}$. More general a_n sequences could be treated, but they would not lead to better rates for convergence of $\tilde{\theta}_n$ to θ. McLeish (1975) has investigated the convergence of the series $\sum_{n=1}^{\infty} c_n X_n$, where the c_n are constants and $\{X_n\}$ is a sequence of random variables satisfying certain mixing and moment conditions. As we will see in Section 4, his results can be used in the verification of A4 and A5 in some specific case. Note that (2.4) is a weaker assumption than that $\sum_{k=1}^{\infty}(||A_k - A|| - C_1)k^{-1}$ converges (and similarly for (2.7)). Also McLeish's theorem 1.6, which is a generalization of a martingale inequality due to Doob (1953), gives conditions which imply A6.

3. Main results.

LEMMA 3.1. *Let* M *be a* p×p *square matrix such that* $a < \lambda_*(M) \le \lambda^*(M) < b$ *for real numbers* a *and* b. *Then there exists a norm* $||\ ||_*$ *on* $\mathbb{R}^p$ *such that*

$$e^{at}||X||_* \le ||e^{Mt} X||_* \le e^{bt}||X||_*$$

for all real t *and all* X *in* $\mathbb{R}^p$.

PROOF. The proof is given by Hirsch and Smale (1974, page 146). □

LEMMA 3.2. *Assume* A1 *to* A5 *hold. Then* $n^{\varepsilon}(\tilde{\theta}_n - \theta) \to 0$ *for all* $\varepsilon < \frac{1}{2}$.

PROOF. Without loss of generality, we take $\theta = 0$. Fix $\varepsilon < \frac{1}{2}$ and define $\phi_n = (n-1)^{\varepsilon}\tilde{\theta}_n$. Then from (2.2) and A2, we obtain

$$\phi_{n+1} = [I + n^{-1}\{\varepsilon I - A_n + u_n\}]\phi_n - n^{\varepsilon-1} W_n ,$$

where the random matrix u_n satisfies

$$(3.1) \qquad ||u_n|| = O(n^{-1} + n^{-1}||A_n||) + o(B_n) .$$

Define $D_n = A_n - \varepsilon I - u_n$. Suppose $n \geq m$. Then iterating (3.1) back to m and using the definitions

$$D = A - \varepsilon I ,$$

$$\mathcal{D}_m^n = \sum_{k=m}^{n} k^{-1}(D_k - D), \quad K_m^n = \sum_{k=m}^{n} k^{-1} ,$$

and

$$w_m^n = \sum_{k=m}^{n} k^{\varepsilon-1} W_k ,$$

we see that

$$(3.2) \qquad \phi_{n+1} = \phi_m - [K_m^n D\phi_m + \mathcal{D}_m^n \phi_m + D \sum_{k=m}^{n} k^{-1}(\phi_k - \phi_m)$$

$$+ \sum_{k=m}^{n} k^{-1}(D_k - D)(\phi_k - \phi_m) + w_m^n] .$$

Choose t_0 and a norm $||\ ||_*$ such that

$$||(I - Dt)x||_* \leq (I - \lambda_*(D)t/2)||x||_*$$

for all t in $[0, t_0]$. (This can be done by Lemma 3.1 and a simple calculation.) Let

$$||D||_* = \sup\{||Dx||_*: ||x||_* = 1\} .$$

For $x > 0$ and $y > 0$, define

(3.3) $$\rho(x,y) = (x - y(x+2))/(||D||_*(1+x) + y + C_1 x + xy)$$

(C_1 is given by A4) and

(3.4) $$r(x,y) = (\rho(x,y) - y)(\lambda_*(D)/2 - y - ||D||_* x - xy - C_1 x) - (2y + xy) .$$

Choose $\eta > 0$ and $L > 0$ such that

$$t_0/2 > \rho(L,\eta) > \eta \text{ and } r(L,\eta) > 0 .$$

Then using (2.3), (2.4), (2.6), and (2.7), choose $N > \max\{2/t_0, \eta^{-1}\}$ so large that if $n > m \geq N$, then

(3.5) $$||v_m^n||_* \leq \eta(1+K_m^n) ,$$

(3.6) $$\sum_{k=m}^{n} k^{-1}||D_k - D||_* \leq (\eta+C_1)K_m^n + \eta ,$$

and

(3.7) $$||w_m^n||_* < \eta^2 .$$

Define $n(1) = N$ and $n(i+1) = \inf\{k \geq n(i): K_{n(i)}^k \geq \rho(L,\eta)\}$ for $i = 1,2,\ldots$. Note that

(3.8) $$K_{n(i)}^{n(i+1)} \leq t_0$$

since $\rho(L,\eta) \leq t_0/2$ and $(n(i+1))^{-1} \leq N^{-1} \leq t_0/2$. Since $\rho(L,N) > \eta > N^{-1}$,

$n(i+1) > n(i)$ for all $i \geq 1$. Let

$$\eta_i = \sup\{||w^k_{n(i)}||^{\frac{1}{2}} : n(i) \leq k \leq n(i+1)\} . \tag{3.9}$$

By (3.7), $\eta_i < \eta$ for $i \geq 1$, and by (2.6), $\eta_i \to 0$.

We now will prove that for each i and each k in $\{n(i),\ldots,n(i+1)\}$,

$$||\phi_k - \phi_{n(i)}|| \leq L \max\{||\phi_{n(i)}||, \eta_i\} . \tag{3.10}$$

The proof will be by induction on k for each fixed i. Suppose that (3.10) holds for all ℓ less than or equal to some $k < n(i+1)$. Then by (3.2), (3.5), (3.6), and (3.9), and the fact that $K^k_{n(i)} < \rho(L,\eta)$ for $k < n(i+1)$,

$$\begin{aligned}
||\phi_{k+1} - \phi_{n(i)}||_* &\leq K^k_{n(i)}||D||_* \, ||\phi_{n(i)}||_* + \eta(K^k_{n(i)} + 1)||\phi_{n(i)}||_* \\
&\quad + ((C_1 + \eta + ||D||_*)K^k_{n(i)} + \eta) \cdot \sup\{||\phi_\ell - \phi_{n(i)}||_* : n(i) \leq \ell \leq k\} + \eta_i^2 \\
&\leq \{\rho(L,\eta)[||D||_*(1+L) + \eta + C_1 L + \eta L] + \eta(L+2)\} \cdot \max\{||\phi_{n(i)}||_*, \eta_i\} \\
&\leq L \max\{||\phi_{n(i)}||_*, \eta_i\} ,
\end{aligned}$$

which completes the proof of (3.10). It follows from (3.2), (3.5) to (3.8), and (3.10) that

$$\begin{aligned}
&I\{||\phi_{n(i)}||_* > \eta_i\}||\phi_{n(i+1)}||_* \\
&\quad \leq ||\phi_{n(i)}||_*\{1 + 2\eta + \eta L - K^{n(i+1)-1}_{n(i)}[\lambda_*(D)/2 - \eta - ||D||_* L - \eta L - C_1 L]\} .
\end{aligned}$$

By the definition of $n(i+1)$,

$$K_{n(i)}^{n(i+1)-1} \geq \rho(L,\eta) - (n(i+1))^{-1} \geq \rho(L,\eta) - \eta ,$$

and therefore

$$(3.11) \quad I\{||\phi_{n(i)}|| > \eta_i\}||\phi_{n(i+1)}||_* \leq ||\phi_{n(i)}||_*(1 - r(\eta,L)) .$$

Now choose $\gamma_i > 0$ such that $\gamma_i > \eta_i$ and $\sum \gamma_i = \infty$. Then by (3.10) and (3.11), we have

$$||\phi_{n(i+1)}||_* \leq \max\{\gamma_i(L+1), ||\phi_{n(i)}||_* - r(\eta,L)\gamma_i\} .$$

An application of Derman and Sack's (1959) lemma 1, with $a_i = \gamma_i(L+1)$, $b_i = 0$, $\delta_i = 0$, and $c_i = r(\eta,L)\gamma_i$, proves that $||\phi_{n(i)}||_* \to 0$, and then by (3.10) we can conclude that $||\phi_n||_* \to 0$. □

THEOREM 3.1. *Assume* A1 *to* A7. *Then there exists* $\varepsilon > 0$ *such that*

$$n^{\frac{1}{2}}(\tilde{\theta}_{n+1} - \theta) = -n^{-\frac{1}{2}} \sum_{k=1}^{n} (k/n)^{A-I} W_k + O(n^{-\varepsilon}) .$$

PROOF. Again we take $\theta = 0$. By (2.2), A2, and A7,

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - n^{-1}(A\tilde{\theta}_n + \kappa_n\tilde{\theta}_n + \xi_n + W_n) ,$$

where

$$\kappa_n = A_n - A \text{ and } \xi_n = O(B_n||\tilde{\theta}_n||^{1+\delta}) .$$

By A6, $E||\kappa_n||^2 \leq C_3$ for all n, whence

$$\sum_{n=1}^{\infty} n^{-1-\varepsilon} E||\kappa_n|| < \infty$$

for all $\varepsilon > 0$. Thus for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} n^{-1-\varepsilon}||\kappa_n|| < \infty .$$

(Here and elsewhere in the proof, we make use of the fact that for nonnegative random variables X_n, $\sum_1^\infty X_n$ converges if $\sum_1^\infty EX_n$ converges.) Then since by Lemma 3.2, $\tilde{\theta}_n << n^{-\frac{1}{2}+\varepsilon}$ for all $\varepsilon > 0$, all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} n^{-\frac{1}{2}-\varepsilon}||\kappa_n\tilde{\theta}_n|| < \infty .$$

Also, we can prove in a similar fashion that

$$\sum_{n=1}^{\infty} n^{-\frac{1}{2}}||\xi_n|| < \infty .$$

Therefore for each $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} n^{-\frac{1}{2}-\varepsilon}(\kappa_n\tilde{\theta}_n + \xi_n + W_n) \text{ converges}$$

by (2.6). Now applying Theorem 3.1 of Ruppert (1978) with $\tau = 0$, $X_n = \tilde{\theta}_n$, $f(x) = Ax$, $\beta_n = 0$, $\beta = 0$, $e_n = (\kappa_n\tilde{\theta}_n + \xi_n + W_n)$, and $\gamma = \frac{1}{2}$, we obtain

$$n^{\frac{1}{2}}\tilde{\theta}_{n+1} = -n^{-\frac{1}{2}} \sum_{k=1}^{n} (k/n)^{A-I} (\kappa_k\tilde{\theta}_k + \xi_k + W_k) + O(n^{-\varepsilon})$$

for some $\varepsilon > 0$. The proof will be completed by showing that for some

$\varepsilon > 0$,

$$(3.12) \qquad n^{-\frac{1}{2}+\varepsilon/3} \sum_{k=1}^{n} (k/n)^{A-I} (\kappa_k \tilde{\theta}_k + \xi_k) = o(1) .$$

We will assume that $A-I$ has only one eigenvalue, λ. This involves no loss in generality since, by a change of basis, we can put $A-I$ in real canonical form (Hirsch and Smale (1974, page 130)) so that

$$(A-I) = \operatorname{diag}(M_1,\ldots,M_q) , \qquad q \le p ,$$

where the square matrices $M_1,\ldots,M_q$ each have exactly one eigenvalue. Then we can apply the proof q times. By Lemma 3.1, there exists Q_k and Q_k' such that $\sup_k(||Q_k|| + ||Q_k^*||) < \infty$ and

$$k^{A-I} = k^{\lambda+\varepsilon/3} Q_k$$

$$k^{I-A} = k^{-\lambda+\varepsilon/3} Q_k^* .$$

Then (3.12) holds if

$$n^{-\frac{1}{2}-\lambda+2\varepsilon/3} Q_n^* \sum_{k=1}^{n} k^{\lambda+\varepsilon/3} Q_k(\kappa_k \tilde{\theta}_k + \xi_k) = o(1) ,$$

and so by Kronecker's lemma it is sufficient to find $\varepsilon > 0$ such that

$$(3.13) \qquad \sum_{k=1}^{\infty} k^{-\frac{1}{2}+\varepsilon} Q_k \xi_k \text{ converges}$$

and

$$(3.14) \qquad \sum_{k=1}^{\infty} k^{-\frac{1}{2}+\varepsilon} Q_k \kappa_k \tilde{\theta}_k \text{ converges} .$$

Now since $||\xi_k|| << B_n||\tilde{\theta}_n||^{1+\delta}$ and $||\tilde{\theta}_n|| << n^{-\frac{1}{2}+\varepsilon}$ for all $\varepsilon > 0$, (3.13) holds for some $\varepsilon > 0$. It remains to prove (3.14).

Now fix a, ε, and ε' such that

$$a > 3/2 ,$$

$$0 < \varepsilon < \tfrac{1}{4} ,$$

and

$$0 < 2a\varepsilon' < \varepsilon .$$

Define $\phi(n)$ to be the greatest integer less than or equal to n^a, $\eta(i) = \{j: \phi(i) \le j \le \phi(i+1) - 1\}$, and $\psi_k = Q_k \kappa_k k^{-\frac{1}{2}+\varepsilon'}$. Then

$$\sum_{k=1}^{\infty} \psi_k \tilde{\theta}_k = \sum_{k=1}^{\infty} \sum_{j \in \eta(k)} \psi_j \tilde{\theta}'_j .$$

We will prove (3.14) by showing that

$$\sum_{k=1}^{\infty} || \sum_{j \in \eta(k)} \psi_j \tilde{\theta}_j || < \infty . \tag{3.15}$$

Now define

$$z_{1,i} = \sum_{j \in \eta(i)} \psi_j \Big(\sum_{\ell=\phi(i)}^{j-1} \ell^{-1} W_\ell \Big) ,$$

$$z_{2,i} = \sum_{j \in \eta(i)} \psi_j \Big(\sum_{\ell=\phi(i)}^{j-1} \ell^{-1} M_\ell(\tilde{\theta}_\ell) \Big) ,$$

and

$$z_{3,i} = \tilde{\theta}_{\phi(i)} \sum_{j=\eta(i)} \psi_j .$$

Then since $\tilde{\theta}_j = \tilde{\theta}_{\phi(i)} - \sum_{\ell=\phi(i)}^{j-1} \ell^{-1}(M_\ell(\tilde{\theta}_\ell) + W_\ell)$ for $j > \phi(i)$, we will have proved (3.15) if we prove that $\sum_{i=1}^{\infty} ||z_{m,i}|| < \infty$ for $m = 1, 2$, and 3. Now

$$\sum_{i=1}^{\infty} ||z_{1,i}|| \le \sum_{i=1}^{\infty} || \sum_{j\in\eta(i)} \psi_j ||\{ \max_{j\in\eta(i)} || \sum_{\ell=\phi(i)}^{j-1} \ell^{-1} W_\ell ||\}$$

$$\le (\sum_{i=1}^{\infty} i^{-\varepsilon} || \sum_{j\in\eta(i)} \psi_j ||^2)^{\frac{1}{2}} (\sum_{i=1}^{\infty} i^{\varepsilon} \max_{i\in\eta(i)} || \sum_{\ell=\phi(i)}^{j-1} \ell^{-1} W_\ell ||^2)^{\frac{1}{2}}$$

$$< \infty ,$$

since by A6

$$E|| \sum_{j\in\eta(i)} \psi_j ||^2 << i^{-1+2a\varepsilon'} , \tag{3.16}$$

$$2a\varepsilon' < \varepsilon ,$$

$$i^{\varepsilon} E \max_{j\in\eta(i)} || \sum_{\ell=\phi(i)}^{j-1} \ell^{-1} W_\ell ||^2 << i^{\varepsilon-a-1} ,$$

and $\varepsilon < a$ by choice of ε and a.

By (2.1), (2.2), and Lemma 3.2,

$$\ell^{-1} ||M_\ell(\tilde{\theta}_\ell)|| << (||A_\ell|| + B_\ell)\ell^{-3/2+\varepsilon'} .$$

Therefore,

$$\sum_{i=1}^{n} ||z_{2,i}|| << \sum_{i=1}^{n} || \sum_{j\in\eta(i)} \psi_j ||[\sum_{j\in\eta(i)} (||A_j|| + B_j)]i^{a(-3/2+\varepsilon')} . \tag{3.17}$$

By (2.5),

$$E[\sum_{j\in\eta(i)}(||A_j|| + B_j)]^2 << (\#\eta(i))^2 << i^{2(a-1)} ,$$

where #A is the cardinality of the set A. Therefore, using (3.16), the expectation of the i^{th} summand on the RHS of (3.17) is bounded by

$$Ki^{a(-3/2+\varepsilon')}\, i^{-\frac{1}{2}+a\varepsilon'}\, i^{a-1} = Ki^{-a/2-3/2+2a\varepsilon'} .$$

Since $\varepsilon' < \frac{1}{4}$, the expectation of the RHS of (3.17) has a finite limit. Therefore, $\sum_{i=1}^{\infty}||z_{2,i}|| < \infty$.

Since $\sup_i ||i^{(\frac{1}{2}-\varepsilon')a}\,\tilde{\theta}_{\phi(i)}|| < \infty$, $\sum_{i=1}^{\infty}||z_{3,i}||$ converges if the following converges:

$$(3.18)\qquad \sum_{i=1}^{\infty} i^{(\varepsilon'-\frac{1}{2})a}||\sum_{j\in\eta(i)}\psi_j|| .$$

Using (3.16) and $-(a+1)/2 + 2a\varepsilon' < -(a+1)/2 + \varepsilon < -1$, (3.18) converges. □

4. Applications. In this section, we elaborate upon two examples of generalized Robbins-Monro processes that were mentioned briefly in the introduction.

ROBUST RECURSIVE ESTIMATION FOR AUTOREGRESSIVE PROCESSES.

Let $\{Y_n\}$ be a strictly stationary autoregressive process such that

$$Y_n = \sum_{k=1}^{p} Y_{n-k}\,\theta^{(k)} + u_n ,$$

where $p < \infty$ and $\{u_n\}$ is an iid sequence with distribution function F and satisfying $Eu_n^4 < \infty$ and $Eu_n = 0$. There are, of course, many methods, e.g. least squares, of estimating $\theta = (\theta^{(1)},\ldots,\theta^{(p)})'$ based on the sample $(Y_1,\ldots,Y_n)$. If the Y_n are being observed rapidly and we need to update our estimate after receiving each new observation, then recursively defined estimators are very attractive. Campbell (1979) has defined a class of recursive estimators which are robust (that is, relatively insensitive to the tail behavior of the distribution of u_n and to the presence of spurious observations), but she only examined the consistency of these procedures. We will look at a class of procedures including Campbell's, investigate its asymptotic distributions, and find the procedure which minimizes the asymptotic variance-covariance matrix. Let $Z_n = (Y_{n-1},\ldots,Y_{n-p})'$. Define $\tilde{\theta}_n$ recursively by

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - n^{-1}\, g(Z_n)\psi(Z_n'\tilde{\theta}_n - Y_n)DZ_n \,, \tag{4.1}$$

where g and ψ are maps from $\mathbb{R}^p$ to $\mathbb{R}$ and $\mathbb{R}$ to $\mathbb{R}$ respectively, and D is a $p\times p$ matrix. For robustness, we require that

$$\sup_{x\in\mathbb{R}} |\psi(x)| < \infty \tag{4.2}$$

and

$$\sup_{x\in\mathbb{R}^p} ||g(x)x|| < \infty \,.$$

These boundedness assumptions will also simplify later proofs. Campbell uses only $D = I$, but she considers a more general subsequence, $\{a_n\}$, in place of n^{-1} in expression (4.1). We will demonstrate that $D = I$ is asymptotically optimal only under rather special circumstances.

Moreover, as we will see, certain of our recursive estimators are asymptotically equivalent to their nonrecursive competitors, bounded influence regression estimates (Hampel (1978)), so there appears to be no advantage in using other a_n.

First, we will verify that A1 to A7 hold. Let

$$F_n = \sigma(u_k: k \le n - 1) ,$$

$$M_n(x) = [\int_{-\infty}^{\infty} \psi(Z_n'(X-\theta) - u)dF(u)]\, g(Z_n)DZ_n$$

$$= E^{F_n} \psi(Z_n'(X-\theta) - u_n)\, g(Z_n)\, DZ_n ,$$

and

$$W_n = \psi(Z_n'(\tilde{\theta}_n - \theta) - u_n)\, g(Z_n)DZ_n - M_n(\tilde{\theta}_n) .$$

Then (4.1) is equivalent to A2.

We will also assume that:

B1. ψ has a bounded Radon-Nikodym derivative $\dot{\psi}$.

B2. ψ has a bounded second derivative $\ddot{\psi}$ except at a finite collection of points $\{\alpha_1, \ldots, \alpha_J\}$.

B3. $E\psi(-u_1) = 0$.

B4. F has a density f satisfying

$$\int_{-\infty}^{\infty} |f(x) - f(x+\alpha)|dx \le K|\alpha|^{\beta} \text{ for all } \alpha \text{ and some } \beta > 0 .$$

B5. The polynomial $x^p - \sum_{r=1}^{p} x^{p-r}\theta^{(r)}$ has all roots inside the unit circle.

B6. Define $c = \int_{-\infty}^{\infty} \dot{\psi}(-u)dF(u)$, $S_n = g(Z_n)Z_nZ_n'$, $A_n = cDS_n$, $S = ES_n$, and $A = EA_n$. Assume $\lambda_*(A) > \frac{1}{2}$.

All ψ functions in the Princeton robustness study (Andrews (1972)), for example, Huber's

$$\psi(x) = \min(k, |x|)\operatorname{sign} x$$

and Andrews'

$$\psi(x) = \sin \frac{x}{k} \quad \text{if} \quad |x| \leq \pi k$$

$$= 0 \quad \text{otherwise} ,$$

satisfy B1 and B2, and they satisfy B3 if F is symmetric about zero.

Since $\{W_n, \mathcal{F}_{n+1}\}$ is a martingale difference sequence and is uniformly bounded in L_2, (2.6) and (2.9) hold by the martingale convergence theorem and Doob's inequality (Doob (1953, Chapter VII, Theorem 3.4).

Theorem 3.2 of Pham and Tran (1980), B4, and the assumption that $E\mu_4^4 < \infty$ imply that Y_n is strong mixing with mixing coefficients $\alpha(n)$ which are $O(\rho^n)$ as $n \to \infty$ for some ρ in $(0,1)$.

Therefore, if h is a Borel function on $\mathbb{R}^q$ and $V_k = h(Y_k, \ldots, Y_{k+q-1})$, then $\{V_n\}$ is also strong mixing, and its mixing coefficients are $O(\rho^{n/2})$.

Define

$$B_n = ||Z_n||^2 . \tag{4.3}$$

Since $EY_n^4 < \infty$ and $g(x)$ is bounded, A_n (see B6) and B_n are strictly stationary and have finite second moments, so (2.5) holds. Define $C_1 = E||A_n - A||$ and $C_2 = EB_n$. Then $\{A_n\}$, $\{||A_n - A|| - C_1\}$, and $\{B_n - C_2\}$ are mixingales of size $-q$ for all $q > 0$. (See McLeish (1975) for a definition and discussion of mixingales.) Therefore, by Corollary 1.8 of McLeish, (2.3), (2.4), and (2.7) hold. Therefore A5 holds, and since $\lambda_*(A) > \frac{1}{2}$ is assumed, A4 holds. By Theorem 1.6 of McLeish, A6 holds.

By B1, B2, and B4,

$$||M_n(x) - A_n(x-\theta)|| = O(B_n||x - \theta||^2)$$

as $x \to \theta$, so that A1 and A7 hold with $v(x) = ||x - \theta||^2$ and B_n given by (4.3). One can show that $\tilde{\theta}_n \to \theta$, i.e. that A3 holds, using Theorem 2.4.1 of Kushner and Clark (1978) with $X_n = \tilde{\theta}_n - \theta$, $a_n = n^{-1}$, $\xi_n = (Y_{n-1}, \ldots, Y_{n-p}, u_n)$, $h_0 = 0$, $B_n \equiv 0$, and $h(x,y) = Dg(y_1)y_1\psi(y_1'x - y_2)$, where $y' = (y_1'y_2')$, $y_1 \in \mathbb{R}^p$, $y_2 \in \mathbb{R}$, and $x \in \mathbb{R}^p$. Kushner and Clark's assumptions A2.4.2 and A2.4.3' can be verified using the mixing properties of Y_n which we have established. Use $\theta(x) = K||x||$, $g_1(x,y) = 1$, and $g_2(\xi) = ||\xi||$ in A2.4.3'. The assumption that $\tilde{\theta}_n - \theta$ is bounded is established using the fact that

$$E^{\mathcal{F}_n}||\tilde{\theta}_{n+1} - \theta||^2 \leq ||\tilde{\theta}_n - \theta||^2 + n^{-2}K$$

and Theorem 1 of Robbins and Siegmund (1971).

We have now verified A1 to A7, so by Theorem 3.1,

$$n^{\frac{1}{2}}(\tilde{\theta}_n - \theta) = -n^{-\frac{1}{2}} \sum_{k=1}^{n} (k/n)^{cDS-I} DW_k + o(n^{-\varepsilon})$$

for some $\varepsilon > 0$. With this approximation, we could investigate the asymptotic behavior of $\tilde{\theta}_n$ rather thoroughly. Here, however, we will restrict attention to asymptotic normality. Define

$$\Sigma_n = [E\psi(-u_1)][g(Z_n)^2 Z_n Z_n']$$

and

$$\Sigma = E\Sigma_n .$$

Since $\tilde{\theta} \to \theta$ and W_n is bounded uniformly in n,

$$E^{\mathcal{F}_n} W_n W_n' - \Sigma_n \to 0 \tag{4.4}$$

by the bounded convergence theorem. Now $(\Sigma_n - \Sigma)$ is a mixingale of size -q for all $q > 0$, so by Corollary 1.8 of McLeish (1975),

$$\sum_{k=1}^{\infty} k^{-(\frac{1}{2}+\varepsilon)} Q_k(\Sigma_n - \Sigma)Q_k' \text{ converges}$$

for any $\varepsilon > 0$ and any sequence of uniformly bounded matrices Q_k. Thus by (4.4) and the same argument used to conclude (3.12) from (3.13) and

(3.14), we see that

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} (k/n)^{cDS-I} (E^{F_k} W_k W_k')((k/n)^{cDS-I})'$$

$$= \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} (k/n)^{cDS-I} \Sigma_n((k/n)^{cDS-I})'$$

$$= \int_0^1 (t^{cDS-I} D\Sigma D'(t^{cDS-I})' \, dt$$

$$= V(c,D,S) \ , \qquad \text{say} \ .$$

(The integral does not diverge at 0 since $\lambda_*(cDS - I) > -\frac{1}{2}$.) By the martingale central limit theorem of, for example, Brown (1971, Theorem 2),

$$n^{\frac{1}{2}}(\tilde{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, V(c,D,S)) \ .$$

(A functional central limit theorem can also be established by the same theorem of Brown.) The Lindeberg condition is easily established because W_n is uniformly bounded. If ψ and g are fixed so that c and S are fixed, then $D = (cS)^{-1}$ is optimal, that is,

$$V(c,D,S) - V(c,(cD)^{-1},S)$$

is p.s.d. This can be seen by noting that

$$\int_0^1 t^{cDS-I} D\Sigma(cS)^{-1} \, dt = (cDS)^{-1} t^{cDS} \Big|_{t=0}^{t=1} D\Sigma(cS)^{-1}$$

$$= (cS)^{-1} \Sigma(cS)^{-1}$$

so that

$$\int_0^1 (t^{cDS-I} D - (cS)^{-1})\Sigma(t^{cDS-I} D - (cS)^{-1})' \, dt = V(c,D,S) - V(c,(cS)^{-1},S) \ ,$$

and the integrand of the LHS is p.s.d.

Note that

$$(4.5) \qquad V(c,(cS)^{-1},S) = c^{-2}S^{-1}\,\dot{\psi}\,S^{-1}$$

$$= \frac{E\psi^2(-u_1)}{(E\dot{\psi}(-u_1))^2}(Eg(Z_1)Z_1Z_1')^{-1}\,(Eg^2(Z_1)Z_1Z_1')(Eg(Z_1)Z_1Z_1')^{-1} .$$

Also, if $D = (cS)^{-1}$, then

$$n^{\frac{1}{2}}(\tilde{\theta}_n - \theta) = -n^{-\frac{1}{2}} \sum_{k=1}^{n} (cS)^{-1} W_k + o(n^{-\epsilon})$$

$$= n^{\frac{1}{2}} \sum_{k=1}^{n} (cS)^{-1} \psi(u_n)g(Z_n)DZ_n + o_p(1) ,$$

so by Carroll and Ruppert (1981, Theorem 1), $\tilde{\theta}_n$ is asymptotically equivalent to the bounded influence regression estimate, $\tilde{\theta}$, which solves

$$0 = \sum_{i=1}^{n} \psi(Y_i - Z_i'\tilde{\theta})g(Z_i)Z_i = 0 .$$

Of course, c and S will generally not be known *a priori*. One can, however, estimate them by, e.g.,

$$\tilde{c}_n = n^{-1} \sum_{i=1}^{n} \dot{\psi}(Y_i - Z_i'\tilde{\theta}_i)$$

and

$$\tilde{S}_n = n^{-1} \sum_{i=1}^{n} g(Z_i)Z_iZ_i'$$

and then use $(\tilde{c}_n S_n)^{-1}$ for D in (4.1). These estimates can be defined recursively. The asymptotic behavior of the resulting stochastic approximation process is still an open problem, but once consistency is established, then general results of Section 3 should be applicable.

If one is not concerned with robustness, then for any value of ψ, $g(x) \equiv 1$ is optimal, but of course our proof assumes that $xg(x)$ is bounded, so it would need to be modified for this choice of g. For this g and its optimal D, i.e. $D = (c(EZ_1Z_1'))^{-1}$, $S = \sharp$ so that $n^{\frac{1}{2}}\tilde{\theta}_n$ has asymptotic variance-covariance

$$c^{-2}\sharp^{-1} = (E\psi^2(u_1)/(E\dot{\psi}(u_1))^2)(EZ_1Z_1')^{-1} .$$

To show that this g is optimal, we must show that for any real-valued g,

$$(4.6) \qquad (Eg(Z_1)Z_1Z_1)^{-1}\,(Eg^2(Z_1)Z_1Z_1')(Eg(Z_1)Z_1Z_1')^{-1} - (EZ_1Z_1')^{-1}$$

is p.s.d., but (4.6) is the variance-covariance matrix of $[g(Z_1)(Eg(Z_1)Z_1Z_1')^{-1} - (EZ_1Z_1')^{-1}]Z_1$.

The use of $g(x) = 1$ allows outlying Z_n to have great influence on the estimation process, which could be disasterous if any of these Z_n have actually been observed with gross error, or if in some other way the pair (Y_n,Z_n) is an outlier. The optimal choice of g subject to the bound

$$||xg(x)|| \le B \quad \text{for all } x ,$$

where B is a fixed constant, has been considered by Hampel (1978) in a (slightly) different context, bounded influence linear regression.

Regardless of the value of g, we see from (4.5) that ψ should be chosen to minimize $E(\psi^2(u_1))/(E\dot{\psi}(u_1))^2$. If F has a density f

satisfying the typical regularity conditions of maximum likelihood estimation, then of course $\psi = \dot{f}/f$ is optimal. If one believes that f is only "close" to some known f_0, then one should choose ψ robustly, and for this purpose the reader is referred to Huber (1964, 1981) for an introduction to robust estimation.

RECURSIVE NONLINEAR REGRESSION.

Now we consider the nonlinear regression model

$$Y_i = F(v_i,r) + e_i \ ,$$

where F is a known function from $\mathbb{R}^q \times \mathbb{R}^p$ to $\mathbb{R}$, v_i is a known vector of covariates, r is an unknown parameter vector, and e_i is a random noise term. Usually r is estimated using least squares, that is, by finding $\tilde{r}$ which minimizes

$$\sum_{i=1}^{n} (Y_i - F(v_i,\tilde{r}))^2 \ .$$

However, just as with our last example, we may wish to use recursively defined estimators. We will study a particular recursive estimation sequence introduced by Albert and Gardner (1967), and we will show that its asymptotic distribution is the same as least squares, at least under certain regularity conditions. (Both the least squares estimator and Albert and Gardner's recursive estimator can be robustified as in the previous example.)

Define $h(v,w)$ to be the gradient $(\partial/\partial w)F(v,w)$, and $h_n(w) = h(v_n,w)$. Let $\tilde{H}_0$ be a p.d. p×p matrix, let $\tilde{r}_1$ be in $\mathbb{R}^p$, and define $\tilde{H}_n$ and $\tilde{r}_n$ by

$$\tilde{H}_n = n^{-1}(\tilde{H}_0 + \sum_{i=1}^{n} h_i(\eta_i)h_i'(\eta_i)) ,$$

where $\eta_i = \tilde{r}_\ell$ for some $\ell \le i$, and

$$(4.7) \qquad \tilde{r}_{n+1} = \tilde{r}_n - n^{-1}[F(v_n,\tilde{r}_n) - Y_n]\tilde{H}_n^{-1} h_n(\tilde{r}_n) .$$

Notice that

$$(4.8) \qquad \tilde{H}_{n+1} = \tilde{H}_n - n^{-1}[\tilde{H}_n - h_n(\eta_n)h_n'(\eta_n)] ,$$

and if we define $B_n = n\tilde{H}_n$, then

$$B_{n+1}^{-1} = B_n^{-1} - \frac{B_n^{-1} h_{n+1}(\eta_{n+1})h_{n+1}'(\eta_{n+1})B_n^{-1}}{1 + h_{n+1}'(\eta_{n+1})B_n^{-1} h_{n+1}(\eta_{n+1})}$$

(Albert and Gardner (1967, page 111)). Consequently, $B_n^{-1} = n^{-1}\tilde{H}_n^{-1}$ can be calculated recursively, and the inversion of $\tilde{H}_n$ is not necessary at each stage of the recursion. However, (4.8) is useful to our study of asymptotic properties of $\tilde{r}_n$ and $\tilde{H}_n$.

First, we list the assumptions we will use. These assumptions are not the weakest possible, since we intend only to illustrate the content of our general results.

Let col be the function mapping the set of p×p matrices into $\mathbb{R}^{p^2}$ by stacking columns one below the other.

C1. The functions D_1, D_2, D_3, and D_4 map $\mathbb{R}^p$ into $\mathbb{R}^+$, $\mathbb{R}^{p\times p}$, $\mathbb{R}^{p\times p^2}$, and $\mathbb{R}^{p^2\times p}$, respectively. Let H be a p.d. p×p matrix. The following hold for all v, w, and p.d. matrices M:

$$(4.9)\qquad ||[F(v,w) - F(v,r)]M^{-1} h(v,w) - H^{-1} h(v,w)h'(v,w)(w-r)||$$
$$\leq KD_1(v)(||w - r||^2 + ||M - H||^2) ,$$

$$(4.10)\qquad ||M^{-1} h(v,w) - H^{-1} h(v,r) - D_2(v)(w-r) - D_3(v)\mathrm{col}(M-H)||$$
$$\leq KD_1(v)(||w - r||^2 + ||M - H||^2) ,$$

$$(4.11)\qquad ||\mathrm{col}[h(v,w)h'(v,w) - h(v,r)h'(v,r)] - D_4(v)(w-r)||^2$$
$$\leq KD_1(v)||w - r||^2 .$$

C2. Let $\xi_n' = (v_n', e_n)$. Suppose $R > 2$ and $\{\xi_n\}$ is strong mixing with mixing numbers $\{\alpha_n\}$ which are of size $-R(R-2)^{-1}$. McLeish (1975) defines size; here we mention McLeish's remark that $\{\alpha_n\}$ is of size $-q$ if $\{\alpha_n\}$ is monotonely decreasing and $\sum_{n=1}^{\infty} \alpha_n^{\phi} < \infty$ for some $\phi < 1/q$. For all n, the marginal distribution of ξ_n satisfies

$$E(e_n|v_n) = 0 ,$$

and

$$(4.12)\qquad Eh_n(r)h_n'(r) = H \text{ for all } n .$$

Moreover,

$$\sup_n E\{D_1^R(v_n) + ||D_2(v_n)||^{2R} + ||D_3(v_n)||^{2R} + ||D_4(v_n)||^R$$
$$+ ||e_n||^{2R} + ||h_n(r)||^{2R}\} < \infty ,$$

and there exists a $p^2 \times p$ matrix D_4 such that

$$ED_4(v_n) = D_4 \text{ for all } n .$$

C3. $\tilde{r}_n \to r$ and $\tilde{H}_n \to H$.

REMARKS. Equations (4.9) to (4.11) merely state that certain Taylor series expansions are valid. Assumption C3 can be verified using the results of Ruppert (1981b) to prove that $\tilde{r}_n \to r$, and then using (4.12) and a continuity assumption on h to show that $\tilde{H}_n \to H$.

Using C2 and (2.3) of McLeish (1975) with $p = 2$, one sees that $\{D_1(v_n)\}$, $\{e_n D_2(v_n)\}$, $\{e_n D_3(v_n)\}$, $\{||e_n D_2(v_n)||\}$, $\{||e_n D_3(v_n)||\}$, $\{D_4(v_n)\}$, and $\{h_n(r)h_n'(r)\}$, when centered at expectations, are each mixingales of size $-\frac{1}{2}$. Moreover, $E(e_n D_2(v_n)) = E(e_n D_3(v_n)) = 0$. Therefore, the following series converge by Corollary 1.8 of McLeish:

$$\sum_1^\infty n^{-1} e_n D_k(v_n) \quad \text{for } k = 2,3 ,$$

$$\sum_1^\infty n^{-1}(||e_n D_k(v_n)|| - E||e_n D_k(v_n)||) \quad \text{for } k = 2,3 ,$$

$$\sum_1^\infty n^{-1}(D_1(v_n) - ED_1(v_n))$$

$$\sum_1^\infty n^{-1}(D_4(v_n) - D_4) ,$$

$$\sum_1^\infty n^{-1}\{||D_4(v_n) - D_4|| - E||D_4(v_n) - D_4||\} ,$$

$$\sum_1^\infty n^{-(\frac{1}{2}+\delta)} e_n h_n(r) \quad \text{for all } \delta > 0 ,$$

and

$$\sum_1^\infty n^{-(\frac{1}{2}+\delta)}(h_n(r)h_n'(r) - H) \quad \text{for all } \delta > 0 .$$

Define

$$A_n = \begin{pmatrix} H^{-1}h_n(r)h_n'(r) + e_n D_2(v_n) & e_n D_3(v_n) \\ D_4(v_n) & I_{p^2} \end{pmatrix} .$$

By C1 and C3,

$$\begin{pmatrix} \tilde{r}_{n+1} - r \\ \mathrm{col}(\tilde{H}_{n+1} - H) \end{pmatrix} = (I_{p(p+1)} - n^{-1}A_n)\begin{pmatrix} \tilde{r}_n - r \\ \mathrm{col}(\tilde{H}_n - H) \end{pmatrix} - n^{-1}\left[\begin{pmatrix} e_n H^{-1}h_n(r) \\ \mathrm{col}[H - h_n(r)h_n'(r)] \end{pmatrix} + \nu_n\right] ,$$

where $||\nu_n|| = O(D_1(v_n)(||\tilde{r}_n - r||^2 + ||\tilde{H}_n - H||^2))$. Therefore, assumptions A1 to A7 hold with

$$\tilde{\theta}_n = \begin{pmatrix} \tilde{r}_n \\ \mathrm{col}\ \tilde{H}_n \end{pmatrix} , \qquad A = \begin{pmatrix} I_p & 0 \\ D_4 & I_{p^2} \end{pmatrix} ,$$

and $W_n = e_n H^{-1}h_n(r)$. It is easy to see that all eigenvalues of A are 1. Moreover, the $p \times p$ matrix in the top left-hand corner of n^{A-I} is I_p. Therefore, we have shown that C1 to C3 imply the existence of an $\varepsilon > 0$ such that

$$n^{\frac{1}{2}}(\tilde{r}_n - r) = -n^{-\frac{1}{2}}H^{-1}\sum_{k=1}^{n} e_k h_k(r) + O(n^{-\varepsilon}) .$$

REFERENCES

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

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

BROWN, B.M. (1971). Martingale central limit theorems. *Ann. Math. Statist.* 42 59-66.

CAMPBELL, KATHERINE (1979). Stochastic approximation procedures for mixing stochastic processes. Ph.D. Dissertation, The University of New Mexico, Albuquerque, N.M.

CARROLL, RAYMOND J. and RUPPERT, DAVID (1980). Weak convergence of bounded influence estimates with applications. *Institute of Statistics Mimeo Series #1285*, University of North Carolina, Chapel Hill, N.C.

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

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

DOOB, JOSEPH L. (1953). *Stochastic Processes*. John Wiley and Sons, New York.

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

HAMPEL, FRANK R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the American Statistical Association*, Statistical Computing Section.

HIRSCH, MORRIS and SMALE, STEPHEN (1974). *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, New York.

HUBER, PETER J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35 73-101.

HUBER, PETER J. (1981). *Robust Statistics*. John Wiley and Sons, New York.

KIEFER, JACK and WOLFOWITZ, JACOB (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23 462-466.

KUSHNER, H.J. and CLARK, D.S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York.

LJUNG, L. (1978). Strong convergence of a stochastic approximation algorithm. *Ann. Statist.* 6 680-696.

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

PHAM, TUAN D. and TRAN, LANH T. (1980). The strong mixing property of the autoregressive moving average time series model. Technical Report, Department of Mathematics, Indiana University.

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

ROBBINS, HERBERT and SIEGMUND, DAVID (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (J.S. Rustagi, ed.) 233-257. Academic Press, New York.

RUPPERT, DAVID (1978). Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise. *Institute of Statistics Mimeo Series #1203*, University of North Carolina, Chapel Hill, N.C. (To appear in *Ann. Prob.*)

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

RUPPERT, D. (1981a). Stochastic approximation of an implicitly defined function. To appear in *Ann. Statist.*

RUPPERT, DAVID (1981b). Convergence of recursive estimators with applications to nonlinear regression. *Institute of Statistics Mimeo Series #1333*, University of North Carolina, Chapel Hill, N.C.

SACKS, JEROME (1958). Asymptotic distributions of stochastic approximation procedures. *Ann. Math. Statist.* 29 373-405.

SIELKEN, ROBERT L., JR. (1973). Stopping times for stochastic approximation procedures. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* 26 67-75.