1. Introduction. This paper is concerned with the strong convergence of recursive estimators which are generalizations of the Robbins-Monro (1951) stochastic approximation procedure. The Robbins-Monro procedure and its generalizations have been investigated in many contexts, for example recursive nonlinear regression (Albert and Gardner, 1967), recursive maximum likelihood estimation (Fabian, 1978), robust estimation of parameters for autoregressive process (Campbell, 1982), robust estimation of a location parameter (Martin and Masreliez, 1975 and Holst, 1980, 1982), control of physical processes (Comer, 1964 and Ruppert, 1979, 1981), and system identification (Kushner and Clark, 1978, section 2.6). In this paper, we present a rather general convergence theorem which allows the disturbances to be dependent and enter in a non-additive fashion. As examples, the theorem is applied to the nonlinear regression estimator of Albert and Gardner (1967), and to the author's (Ruppert 1979, 1981) Robbins-Monro type procedures for use where the root of the unknown regression function varies with time.

The algorithm studied is

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

where $x_n$ and $v_n$ are random vectors in $\mathbb{R}^p$, $H_n$ is a positive definite matrix, and $\xi_n$ is a random element in a metric space $M$. The standard theory of stochastic approximation applies directly when $\xi_1, \xi_2, \ldots$ are i.i.d. and when $(v_n)$ is a martingale difference sequence. In this paper, we still apply a standard result (Deman and Sacks, 1959), but the application is not quite so direct. In particular, the lemma of Deman and Sacks is applied not to $x_n$ but to a subsequence $x_n(k)$ where $(n(k+1)-n(k)) \to \infty$ as $k \to \infty$. To complete the convergence proof, we then show that

$$\sup_{k \geq 0} \| x_{n(k)} - x_n(k) \| = \lim_{n \to \infty} \| x_{n(k)} - x_n(k) \| = 0$$

To verify the assumptions of Deman and Sacks's lemma, we utilize a result by Ranga Rao (1962) on the relationship between weak and uniform convergence of probability measures.

An alternate approach to convergence is given in Kushner and Clark (1978). They need to assume that the stochastic approximation process is bounded almost surely, but in some applications showing boundedness is nearly as difficult as proving convergence. A major advantage to using uniform convergence arguments is that boundedness need not be assumed. However, as we will point out in this paper, assuming

---

This research was supported by the National Science Foundation through Grants MCS78-01240 and MCS81-00748.

boundedness of the process does allow us to weaken other assumptions.

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 \(\sigma\)-algebra, and let \((M, d)\) be a separable metric space with the Borel \(\sigma\)-algebra \(\mathcal{B}_M\). All functions which we consider between metric spaces are assumed to be Borel measurable. Let a prime denote matrix transposition. For a real matrix \(A\), \(\|A\| = (\text{Trace} A^t A)^{\frac{1}{2}}\). We will need the following assumptions, which are discussed below.

- **A1.** \(h(\cdot, \cdot), h_1(\cdot, \cdot), \text{ and } h_2(\cdot)\) are functions from \(\mathbb{R}^p \times M \to \mathbb{R}^p, \mathbb{R}^p \times M \to \mathbb{R}^{p \times r}, \text{ and } \mathbb{R}^p \to \mathbb{R}^q\) 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 \(\lambda \leq \bar{\lambda}\), all eigenvalues of \(H_n\) are between \(\lambda\) and \(\bar{\lambda}\) for all \(n\).

- **A3.** Suppose \(u\) is a probability measure on \((M, \mathcal{F}_M)\).

- **A4.** Let \(2 \leq r < \infty\). Let \(\xi_1, \xi_2, \ldots\) be a sequence of random vectors in \(M\). Suppose that \(\int g \, d\mu = 0\) and \(g \in L^r(\mu)\) implies that there exists \(c\) such that
  \[
  \mathbb{E} \max_{m \leq \ell \leq n} \left(\sum_{i=m}^{\ell} c_i g(\xi_i)\right)^2 \leq c \sum_{i=m}^{n} c_i^2
  \]
  for all \(n > m\) and all constants \(c_m, \ldots, c_n\).

- **A5.** There exists a nonnegative continuous function \(h_3\) on \(M\) such that (i) \(h_3 \in L^r(\mu)\) and (ii) \(\|h_1(x, \xi)\| \leq h_3(\xi)\) for all \(x\) and \(\xi\).

- **A6.** \(\{h_1(x, \cdot)\}_{x \in \mathbb{R}^p}\) is an equicontinuous family on \(M\), that is, for each \(\xi \in M\) and \(\varepsilon > 0\) there exists \(\delta > 0\) such that \(\xi^* \in M\) and \(d(\xi^*, \xi) < \delta\) implies that \(\|h_1(x, \xi) - h_1(x, \xi^*)\| < \varepsilon\) for all \(x \in \mathbb{R}^p\).

Remark. 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.

If \(h\) satisfies the assumptions below on \(h_1\), then we may simply choose \(h_1 \equiv h\), \(h_2 \equiv 1\), and \(r = 1\). In section 4, we show an example where \(h\) does not satisfy these assumptions and therefore \(h \neq h_1\) and \(h_2 \neq 1\). In any application, there may be many suitable choices of \(h_1\) and \(h_2\). Note that \(h_1\) and \(h_2\) are used only in the proof of convergence; they are not employed in the algorithm itself.

Notation. Define \(F_1(x) = \int h_1(x, \xi) \, d\mu(\xi)\) and \(F(x) = F_1(x) h_2(x) = \int h(x, \xi) \, d\mu(\xi)\).

- **A7.** \(h_2\) is bounded in a neighborhood of \(0\).

- **A8.** For all \(\varepsilon > 0\), \(\inf_{\|x\| > \varepsilon} \min(\|F_1(x)\|, \|F_2(x)\|) > 0\).

- **A9.** Suppose that \(\tilde{V}\) is the gradient of \(V\), (i) \(\inf_{\|x\| > \varepsilon} (V(x) - V(0)) > 0\) for all \(\varepsilon > 0\), and (ii) with \(\tilde{V}\) the Hessian of \(V\), \(\|\tilde{V}(x)\| \leq K\) for some \(K\) and all \(x\).

Remark. The assumption of a bounded Hessian is common in the literature of multi-