The Annals of Statistics
1981, Vol. 9, No. 3, 555–566

STOCHASTIC APPROXIMATION OF AN IMPLICITLY DEFINED FUNCTION[1]

BY DAVID RUPPERT

University of North Carolina, Chapel Hill

Let S be a set, R the real line, and M a real function on $R \times S$. Assume there exists a real function, f, on S such that $(x - f(s))M(x, s) \geq 0$ for all x and s. Initially neither M nor f are known. The goal is to estimate f. At time n, s_n (a value in S) is observed, x_n (a real number) is chosen, and an unbiased estimator of $M(x_n, s_n)$ is observed. This problem has applications, for example, to process control. In a previous paper the author proposed estimation of f by a generalization of the Robbins-Monro procedure. Here that procedure is generalized and asymptotic distributions are studied.

1. Introduction. This paper considers the following mathematical model for control of a physical process. The process is influenced by two variables, x which is real valued and s which takes values in an abstract space S. At time n, the value s_n of s can be measured but not controlled by the experimenter, and after measuring s_n, the value x_n of x can be chosen by him. The output of the process has conditional expectation, given the past, $M(x_n, s_n)$, where M is a real valued function. In applications, s represents exogenous variables which influence the process. For example, a patient's response to a drug may depend on his age and weight, as well as the dosage. Then, s_n can be the age and weight of the nth patient, and x_n can be the dosage administered to him. In this case, $M(x_n, s_n)$ would be the expected dose-response.

We also assume that the experimenter wants to choose x_n so that the expected output is held equal to a constant, which for convenience may be assumed to be 0. Suppose there exists a function f such that $M(f(s), s) = 0$ for all s. Then, it would be sufficient to choose $x_n = f(s_n)$. However, initially both M and f are unknown. We will be concerned with estimation of f.

To make the problem tractable, we will restrict attention to only a certain class of estimators. Let U be a (known) function from S to R^k. We will consider only estimators in the class

$$\mathscr{G} = \{g : g(s) = \gamma' U(s) \quad \text{for some} \quad \gamma \in R^k\},$$

and we will investigate both the case where f itself is in $\mathscr{G}$ and when f can only be approximated by an element of $\mathscr{G}$.

The author (Ruppert, 1979) has investigated a slightly different formulation of this problem. In that paper, U takes values in an inner product space $(\mathscr{H}, \langle \cdot, \cdot \rangle)$, not necessarily finite dimensional, and $\mathscr{G} = \{g : g(s) = \langle \gamma, U(s) \rangle$ for some $\gamma \in \mathscr{H}^k\}$. Also it is assumed that $f \in \mathscr{G}$, so that $f(s) = \langle \beta, U(s) \rangle$ for some β in $\mathscr{H}$. With $\mathscr{H}$ infinite dimensional, the restriction that f be in $\mathscr{G}$ is not unreasonable, since $\mathscr{G}$ will be a rich class for $\mathscr{H}$ appropriately chosen. The following method of estimating β was proposed. An initial estimate β_1 of β is selected. At time n, β_n is the current estimate of β and the choice $x_n = \beta_n' U(s_n)$ is made. The output, Y_n, is measured and a new estimate is formed:

Received January, 1979; revised November, 1979.

[1] Research supported in part by the National Science Foundation, Grant NSF MCS78-01240.

AMS 1970 *subject classifications.* Primary 62L20; secondary 62J99.

Key words and phrases. Stochastic approximation, asymptotic normality, process control.

$$\beta_{n+1} = \beta_n - a_n Y_n U(s_n), \tag{1.1}$$

where a_n is a suitably chosen positive number.

The Robbins-Monro (1951) process is a special case of (1.1). To see this, suppose S is a singleton, $S = \{s_0\}$, say and let

$$\tilde{M}(x) = M(x, s_0). \tag{1.2}$$

Then the Robbins-Monro procedure for estimating the solution of

$$\tilde{M}(x) = 0,$$

that is, estimation of $f(s_0)$, coincides with (1.1) if we take $k = 1$, $U(s_0) = 1$, and $\beta = f(s_0)$.

When $\{s_n\}_{n=1}^{\infty}$ is a deterministic sequence, Ruppert (1979) showed that under weak conditions

$$n^{-1} \sum_{i=1}^{n} | x_i - f(s_i) | \to 0 \text{ almost surely}, \tag{1.3}$$

and argued that for process control (1.3) is a valuable property. Conditions sufficient for $x_n - f(s_n) \to 0$ almost surely were also given. When $\{s_n\}_{n=1}^{\infty}$ is a stochastic sequence, then $x_n - f(s_n) \to 0$ almost surely under conditions given in Ruppert's (1979) Theorem 4.5. Rates of convergence were also investigated there.

In this paper, it is assumed that $\{s_n\}_{n=1}^{\infty}$ is an independent, identically distributed sequence. With this assumption it is possible to find the asymptotic distribution of β_n when

$$a_n = an^{-\alpha} \quad \text{for some} \quad a > 0 \quad \text{and} \quad \tfrac{1}{2} < \alpha \le 1. \tag{1.4}$$

Instead of considering only procedures of the form (1.1), we treat a wider class, because this class contains procedures which are often more efficient (in a natural sense, to be discussed later) than procedures of the form (1.1). Specifically, let $\mathscr{C}$ be the class of sequences satisfying

$$\beta_{n+1} = \beta_n - n^{-\alpha} h(s_n) Y_n D U(s_n), \tag{1.5}$$

where h is a positive function and D is a symmetric, positive definite matrix.

In Section 3 the properties of the class $\mathscr{C}$ are studied. The asymptotic distributions of sequences β_n satisfying (1.5) are found (cf. Theorem 3.1). In $\mathscr{C}$, the optimal (in a sense to be made more specific later) choice is found (cf. Theorem 3.4). Recall that the motivation for estimating β is to allow x_n to be chosen to keep $M(x_n, s_n)$ at, or close to, zero. Theorem 3.5 gives the asymptotic distribution of $M(x_n, s_n)$ under this optimal choice.

To the best of the author's knowledge, there has been no previous work on the estimation problem posed at the beginning of this introduction. However, in order to compare our procedures with a familiar one, one may restrict M to be of a special form, $M(x, s) = F(s)(x - \beta' U(s))$ where F is a known positive function. Then β can be estimated by least squares. We show that the asymptotic distribution of the optimal procedure in $\mathscr{C}$ is the same as the asymptotic distribution of least squares (cf. remarks following Theorem 3.4).

2. Assumptions, definitions, and notation.

Let $R^{m\times n}$ be the set of all $m \times n$ real matrices. Let tr A denote the trace of the square matrix A and let prime denote transposition. Let $A^{(ij)}$ be the i, jth entry of the matrix A, and let $x^{(i)} = x^{(i1)}$ if x is $m \times 1$.

Define the inner product $[A, B] = \operatorname{tr} A'B$ for A, B in $R^{m\times n}$, and define $\|A\| = [A, A]^{1/2}$.

If $A \in R^{m\times m}$ is symmetric, then $\underline{\lambda}(A)$ and $\bar{\lambda}(A)$ denote the minimum and maximum eigenvalues of A. We say that A is p.d. (p.s.d.) if A is symmetric and $\underline{\lambda}(A)$ is positive (nonnegative). If A is p.s.d., then $A^{1/2}$ is the unique p.s.d. matrix B satisfying $B^2 = A$. A function from $R^{m\times n}$ to $R^{m'\times n'}$ is C^1 if it has a continuous total derivative.

All random variables are defined on a probability space $(\Omega, \mathscr{F}, P)$. All relations between random variables are meant to hold with probability 1. If A is a set, then ${}_{\chi}A$ is the indicator of A. Let x_n and y_n be sequences of random variables. Then we write $x_n = O(y_n)$ if there is a random g satisfying $|x_n| \leq g|y_n|$ for all n. $N(\mu, S)$ will denote a normal random vector with mean μ and variance S, and $\rightarrow_{\mathscr{D}}$ denotes convergence in distribution.

A sequence of random variables ξ_n has a property eventually if, for every ω in a set with probability 1, $\xi_n(\omega)$ has the property for all n greater than some $n_0(\omega)$.

Let $(S, \mathscr{S}, \mu)$ be a probability space, and denote the Borel σ-algebra on R^m by $\mathscr{B}^m$; $\mathscr{B}' = \mathscr{B}$.

Let M, h, and U be measurable tranformations from $(R \times S, \mathscr{B} \times \mathscr{S})$ to $(R, \mathscr{B})$, from $(S, \mathscr{S})$ to $(R, \mathscr{B})$, and from $(S, \mathscr{S})$ to $(R^k, \mathscr{B}^k)$, respectively.

ASSUMPTIONS.

A1. (i) Assume for γ in R^k that

$$H(\gamma) = \int_S h(s)M(\gamma' U(s), s)U(s)\, d\mu(s)$$

exists and has finite entries, and that there exists β in R^k such that (ii) $H(\beta) = 0$, and (iii) for each $\varepsilon > 0$, $\inf_{\varepsilon < \|\gamma - \beta\| < \varepsilon^{-1}} (\gamma - \beta)' H(\gamma) > 0$.

A2. Assume that for some K_4

$$\int_S \| h(s)M(\gamma' U(s), s)U(s) \|^2 \, d\mu(s) \leq K_4 \|\gamma - \beta\|^2.$$

A3. Assume that for some K_5

$$(\gamma - \beta)' H(\gamma) \geq K_5 \|\gamma - \beta\|^2.$$

A4. Assume that for some $0 < \xi < 1$ and K_6

$$\int_S \| h(s)M(\gamma' U(s), s)U(s) \|^2 \, d\mu(s) \leq K_6 \|\gamma - \beta\|^{2\xi}.$$

REMARKS. Conditions implying A1(ii) and A1(iii) are given in Lemmas 3.1 and 3.2. Our main theorem requires that A3 or A4 holds, and under the conditions of either lemma, it is easy to find assumptions on M, h, and U which insure that either A3 or A4 holds, and that A2 holds.

A5. Suppose that

$$F(s) = \frac{\partial}{\partial x} M(x, s) \big|_{x = \beta' U(s)}$$

exists for all s in S, and $\int_S (h^2 + F^2) \|U\|^2 \, d\mu < \infty$.

Define

$$B = \int_S hFUU' \, d\mu$$

and

$$C = \int_S h^2 UU' \, d\mu.$$

A6. Assume that B and C are p.d. Also assume that B is the total derivative of $H(\gamma)$ at $\gamma = \beta$, i.e., $H(\gamma)$ can be differentiated under the integral sign at β.

REMARK. Usually h would have the same sign as F. Then B is p.d. under reasonable conditions on U and μ.

A7. Suppose $\frac{1}{2} < \alpha \leq 1$; set $\theta = 1$ or 0 according as $\alpha = 1$ or $\alpha < 1$.

A8. Let D in $R^{k\times k}$ be symmetric and p.d., and let P be an orthogonal matrix such that

$$P'B^{1/2}DB^{1/2}P = \Lambda$$

where Λ is diagonal.

A9. Assume $\underline{\lambda}(\Lambda) > \frac{1}{2}$ if $\alpha = 1$.

A10. Assume that for some K_7

$$\int_S \| U \|^k \, d\mu \le K_7 \quad \text{for} \quad k = 1, 2, 3, \text{and } 4.$$

A11. Let $\mathscr{F}_n$ and $\mathscr{F}_n^*$ be sub-σ-algebras of $\mathscr{F}$ such that $\mathscr{F}_n \subset \mathscr{F}_n^* \subset \mathscr{F}_{n+1}$.

Remarks. In terms of the practical situation of the introduction, $\mathscr{F}_n$ is the σ-algebra of the past and $\mathscr{F}_n^*$ is the σ-algebra of the past and the determination of s_n. In the following assumption, we formalize the procedure discussed in the introduction. The procedure depends upon h and D. In practice these would be chosen by the statistician. We discuss the choice of h and D later.

A12. Suppose s_n is a measurable transformation from $(\Omega, \mathscr{F}_n^*)$ to $(S, \mathscr{S})$, s_n is μ-distributed and independent of $\mathscr{F}_n$, Y_n is an $\mathscr{F}_{n+1}$-measurable random variable, β_n is a random vector in R^k, and β_1 is $\mathscr{F}_1$ measurable. Suppose that with $\operatorname{Var}^{\mathscr{F}_n^*} Y_n = \sigma_n^2$, $\sigma_n^2 \to \sigma^2 > 0$ and $E(\sigma_n^2 - \sigma^2)^2 \to 0$. Set $U_n = U(s_n)$, $x_n = \beta_n' U_n$ and

$$\sigma_{r,n}^2 = E(Y_n - E^{\mathscr{F}_n^*}Y_n)^2 \, \chi\{(Y_n - E^{\mathscr{F}_n^*}Y_n)^2 \ge rn^\alpha\}$$

for $r > 0$.

A13. Suppose that (i) $E\|\beta_1\|^2 < \infty$; (ii) $\beta_{n+1} = \beta_n - n^{-a}h(s_n)Y_nDU_n$; and (iii) that with $M_n = M(x_n, s_n)$, $E^{\mathscr{F}_n^*}Y_n = M_n$.

A14. Suppose that for all $r > 0$

$$\lim_{n\to\infty} \sigma_{r,n}^2 = 0, \qquad \text{or} \quad \alpha = 1 \quad \text{and} \quad \lim_{n\to\infty} n^{-1} \textstyle\sum_{j=1}^n \sigma_{r,j}^2 = 0.$$

Remarks. The assumption that $\{s_n\}_{n=1}^\infty$ be an i.i.d. sequence is realistic in some applications, e.g., the dose-response example of the introduction if the patients arrive for treatment randomly. Clearly, though, the procedure would have wider applications if it applied to dependent and/or nonidentically distributed s_n.

If $f(s) = \beta'U(s)$, i.e., if f is in $\mathscr{F}$, then Theorem 4.5 of Ruppert (1979) shows that $\beta_n \to \beta$ and $\sup_n n^\alpha E\|\beta_n - \beta\|^2 < \infty$ under fairly weak conditions. However, the techniques used in the present paper to derive asymptotic distributions when $\{s_n\}$ are i.i.d. appear to be inadequate under appreciably weaker conditions on $\{s_n\}$.

3. Asymptotic properties of estimates.

Theorem 3.1. *Suppose* A1, A2, A5–A14, *and either* A3 *or* A4 *hold. Then*

$$n^{\alpha/2}(\beta_n - \beta) \to_{\mathscr{D}} N(0, \sigma^2 B^{-1/2} P \mathscr{M} P' B^{-1/2})$$

where $\mathscr{M}^{(ij)} = (P'B^{1/2}DCDB^{1/2}P)^{(ij)}(\Lambda^{(ii)} + \Lambda^{(jj)} - \theta)^{-1}$.

For the proof we will need some preliminary results which will be stated as theorems.

Theorem 3.2 (Robbins and Siegmund, 1971, Theorem 1). *Let F_n be a nondecreasing sequence of sub-σ-algebras of $\mathscr{F}$. Let z_n, β_n, ξ_n, and ζ_n be F_n-measurable random variables such that $z_n, \zeta_n \ge 0$ and $E^{F_n}z_{n+1} \le z_n(1 + \beta_n) + \xi_n - \zeta_n$. Then* $\lim_{n\to\infty} z_n$ *exists and is finite and $\sum_1^\infty \zeta_n < \infty$ on $\{\sum_1^\infty |\beta_n| < \infty, \sum_1^\infty |\xi_n| < \infty\}$.*

REMARK. Robbins and Siegmund assumed β_n, $\xi_n \geq 0$, but since $z_n(1 + \beta_n) + \xi_n - \zeta_n \leq z_n(1 + |\beta_n|) + |\xi_n| - \zeta_n$, Theorem 3.2 is an immediate consequence of their result.

THEOREM 3.3 (Fabian, 1968, Theorem 2.2). *Suppose F_n is a nondecreasing sequence of sub-σ-algebras of $\mathscr{F}$. Suppose U_n, V_n, T_n are random vectors in R^k, $T \in R^k$, Γ_n and Φ_n are random elements of $R^{k\times k}$, Σ, Γ, Φ, $P \in R^{k\times k}$, Γ is* p.d., *P is orthogonal, and $P'\Gamma P = \Lambda$ diagonal. Suppose Γ_n, Φ_{n-1}, V_{n-1} are F_n-measurable, α, $\beta \in R$ and*

$$\Gamma_n \to \Gamma, \Phi_n \to \Phi, T_n \to 0 \qquad \textit{or} \quad E\|T_n - T\| \to 0, E^{F_n}V_n = 0,$$

and

$$\|E^{F_n}V_nV_n' - \Sigma\| \to 0 \qquad \textit{and} \quad Et'(E^{\mathscr{F}_n}V_nV_n' - \Sigma)t \to 0 \tag{3.1}$$

for all $t \in R^k$.

Suppose with $\sigma_{j,r}^2 = E_X\{\|V_j\|^2 \geq rn^\alpha\}\|V_j\|^2$

$$\lim_{j\to\infty} \sigma_{j,r}^2 = 0 \qquad \textit{for every} \quad r,$$

or

$$\alpha = 1 \text{ and } \lim_{n\to\infty} n^{-1} \textstyle\sum_{j=1}^n \sigma_{j,r}^2 = 0 \qquad \textit{for every} \quad r.$$

Suppose that, with $\lambda = \underline{\lambda}(\Lambda)$, $\beta_+ = \beta$ if $\alpha = 1$, $\beta_+ = 0$ if $\alpha \neq 1$,

$$0 < \alpha \leq 1, \qquad 0 \leq \beta, \qquad \beta_+ < 2\lambda,$$

and

$$U_{n+1} = (I - n^{-\alpha}\Gamma_n)U_n + n^{-(\alpha+\beta)/2}\Phi_nV_n + n^{-(\alpha-\beta)/2}T_n.$$

Then $n^{\beta/2}(U_n - (\Gamma - (\beta_+/2)I)^{-1}T) \to_{\mathscr{D}} N(0, P\mathscr{M}P')$ where $\mathscr{M}^{(ij)} = (P'\Phi\Sigma\Phi'P)^{(ij)}(\Lambda^{(ii)} + \Lambda^{(jj)} - \beta_+)^{-1}$.

REMARK. Fabian stated the theorem with (3.1) replaced by the stronger hypothesis

$$C > \|E^{F_n}V_nV_n' - \textstyle\sum\| \to 0$$

for some constant C. However, his proof goes through with only (3.1).

PROOF OF THEOREM 3.1. Let $z_n = (\beta_n - \beta)'D^{-1}(\beta_n - \beta)$. Then since $\mathscr{F}_n \subset \mathscr{F}_n^*$, we have by A13(ii),

$$E^{\mathscr{F}_n}z_{n+1} = z_n - 2n^{-\alpha}E^{\mathscr{F}_n}(\beta_n - \beta)'U_nh(s_n)M_n + n^{-2\alpha}E^{\mathscr{F}_n}(M_n^2 + \sigma_n^2)(h(s_n))^2U_n'DU_n.$$

By A1(iii),

$$E^{\mathscr{F}_n}(\beta_n - \beta)'U_nh(s_n)M_n = (\beta_n - \beta)'H(\beta_n) \geq 0, \tag{3.2}$$

since s_n is μ-distributed. By A2, A5, and A12, for some K_8 and K_9,

$$E^{\mathscr{F}_n}(M_n^2 + \sigma_n^2)(h(s_n))^2U_n'DU_n \leq K_8(\|\beta_n - \beta\|^2 + 1) \leq K_9(\bar{\lambda}(D)z_n + 1),$$

and by (3.2), for some $K_{10} > 0$,

$$E^{\mathscr{F}_n}z_{n+1} \leq z_n(1 + K_{10}n^{-2\alpha}) - 2n^{-\alpha}(\beta_n - \beta)'H(\beta_n) + K_9n^{-2\alpha}. \tag{3.3}$$

Therefore, by Theorem 3.2, z_n converges to a finite limit and

$$\textstyle\sum_1^\infty n^{-\alpha}(\beta_n - \beta)'H(\beta_n) < \infty. \tag{3.4}$$

If $\lim z_n > 0$, then $\liminf \|\beta_n - \beta\| > 0$ and by A1(iii), $\liminf(\beta_n - \beta)'H(\beta_n) > 0$. Since $\alpha \leq 1$ we obtain a contradiction to (3.4). Therefore $z_n \to 0$ and $\|\beta_n - \beta\| \to 0$.

By (3.2) and (3.3), if $EZ_n < \infty$ then $EZ_{n+1} < \infty$. Therefore, by A13(i), $EZ_n < \infty$ for all n. Taking expectation in (3.3) and applying Theorem 3.2 again, we obtain

$$\lim Ez_n \quad \text{exists and is finite and} \tag{3.5}$$

$$\textstyle\sum_1^\infty n^{-\alpha} E(\beta_n - \beta)' H(\beta_n) < \infty. \tag{3.6}$$

Now (3.5) implies lim sup $E \| \beta_n - \beta \|^2 < \infty$, whence for $0 < \xi < 1$, $\| \beta_n - \beta \|^{2\xi}$ is uniformly integrable. Therefore $E \| \beta_n - \beta \|^{2\xi} \to 0$. If A3 holds, then $(\beta_n - \beta)' H(\beta_n) \geq K_5 \| \beta_n - \beta \|^2$ and then (3.6) implies $E \| \beta_n - \beta \|^2 \to 0$. In summary, so far we have shown that $\| \beta_n - \beta \| \to 0$ and $E \| \beta_n - \beta \|^2 \to 0$, or A4 holds and $E \| \beta_n - \beta \|^{2\xi} \to 0$. By A13(ii),

$$\begin{aligned}(\beta_{n+1} - \beta) &= (\beta_n - \beta) + n^{-\alpha}[-E^{\mathscr{F}_n}(h(s_n)M_n D U_n) \\ &\quad + (E^{\mathscr{F}_n}(h(s_n)M_n D U_n) - h(s_n)M_n D U_n) \\ &\quad + (h(s_n)(M_n - Y_n) D U_n)] \\ &= (\beta_n - \beta) + n^{-\alpha}[-r_{1,n} + r_{2,n} + r_{3,n}], \quad \text{say.}\end{aligned} \tag{3.7}$$

We now have

$$r_{1,n} = DH(\beta_n) = DB_n(\beta_n - \beta), \tag{3.8}$$

where $B_n \to B$ by A6. Let $V_n = r_{1,n} + r_{2,n}$. Then by (3.7) and (3.8),

$$B^{1/2}(\beta_{n+1} - \beta) = (I - B^{1/2} D B_n B^{-1/2}) B^{1/2}(\beta_n - \beta) + n^{-\alpha}(B^{1/2} V_n). \tag{3.9}$$

Note that $E^{\mathscr{F}_n} r_{2,n} = E^{\mathscr{F}_n} r_{3,n} = E^{\mathscr{F}_n}(r_{2,n})(r_{3,n}) = 0$. Also $\bar{\lambda}(\mathscr{D})^{-1} E^{\mathscr{F}_n} \| r_{2,n} \|^2 \leq E^{\mathscr{F}_n} h^2(s_n) M_n^2 \| U_n \|^2$, so for some K_{12}, $E \| r_{2,n} \|^2 \leq K_{12} E \| \beta_n - \beta \|^{2\xi}$ if A4 holds and $E \| r_{2,n} \|^2 \leq K_{12} E \| \beta_n - \beta \|^2$ if A4 doesn't hold. Therefore,

$$E \| r_{2,n} \|^2 \to 0.$$

Since $E^{\mathscr{F}_n} r_{3,n} r'_{3,n} = D(E^{\mathscr{F}_n} \sigma_n^2 h^2(s_n) U_n U'_n) D$,

$$\lim_{n\to\infty} E^{\mathscr{F}_n} V_n V'_n = \lim_{n\to\infty} E^{\mathscr{F}_n} r_{3,n} r'_{3,n} = \sigma^2 DCD, \tag{3.10}$$

and for $t \in R^k$,

$$\begin{aligned} &E \,| t'(E^{\mathscr{F}_n} V_n V'_n - \sigma^2 DCD) t \,| \\ &\quad \leq E \| r_{2,n} \|^2 \| t \|^2 + t' D(E^{\mathscr{F}_n} \sigma_n^2 h^2(s_n) U_n U'_n - \sigma^2 C) Dt \to 0. \end{aligned} \tag{3.11}$$

Define for $r > 0$, $\rho_{r,n}^2 = E \| V_n \|^2 {}_\chi\{ \| V_n \|^2 \geq rn^\alpha \}$. Then since $V_n = r_{2,n} + r_{3,n}$

$$\rho_{r,n}^2 \leq 2E \| r_{2,n} \|^2 + 2E \| r_{3,n} \|^2 {}_\chi\left\{ \| r_{3,n} \| \geq \frac{rn^\alpha}{2} \right\}. \tag{3.12}$$

By A14 we can choose a sequence $0 < k_n \uparrow \infty$ such that with $d_j = k_j^2 E(Y_j - M_j)^2 {}_\chi\{(Y_j - M_j)^2 \geq rn^\alpha 2^{-1}(K_8 \| D \| k_j)^2\}$, $d_n \to 0$ or $\alpha = 1$ and $n^{-1} \sum_{j=1}^n d_j \to 0$. Then

$$E \| r_{3,n} \|^2 {}_\chi\left\{ \| r_{3,n} \|^2 \geq \frac{rn^\alpha}{2} \right\} \leq E \| r_{3,n} \|^2 {}_\chi\{ \| U_n \| \geq k_j \} + K_8^2 \| D \|^2 d_n. \tag{3.13}$$

The first term on the right-hand side of (3.13) goes to 0 so

$$\rho_{r,n} \to 0, \qquad \text{or} \quad \alpha = 1 \qquad \text{and} \quad n^{-1} \textstyle\sum_{j=1}^n \rho_{r,j} \to 0. \tag{3.14}$$

Then by (3.9) to (3.14) and Theorem 3.3, $n^{-\alpha/2} B^{1/2}(\beta_n - \beta) \to_{\mathscr{L}} N(0, \sigma^2 P \mathscr{M} P')$ and Theorem 3.1 follows. □

The following two lemmas deal with Assumption A1; see Remarks after A4 in Section 2.

LEMMA 3.1 *Assume that for a function f,* $(x - f(s))M(x, s) \geq 0$, *that* $M(f(s), s) = 0$ *for all s, and that for some* β *in* R^k, $f(s) = \beta' U(s)$. *Also, assume that for each* $\varepsilon > 0$

$$\delta(\varepsilon) = \inf_{\varepsilon \leq |x - f(s)| \leq \varepsilon^{-1}} |M(x, f(s))| > 0,$$

that $\inf_{s \in S} h(s) > 0$, *and that for some* $\Gamma > 0$,

$$\inf_{x \in R^k} \mu(\Gamma \| x \| \leq | x' U(s) | \leq \Gamma^{-1} \| x \|) \geq \Gamma. \tag{3.15}$$

Then A1(i) *implies* A1(ii) *and* (iii).

PROOF. A1(ii) is trivial. Now fix $\varepsilon > 0$. Then whenever $\varepsilon < \| \gamma - \beta \| < \varepsilon^{-1}$, we have

$$(\gamma - \beta)' H(\gamma) \geq \varepsilon (\inf h(s)) \delta(\varepsilon \Gamma) > 0$$

if $\Gamma \| \gamma - \beta \| \leq |(\gamma - \beta)' U(s)| \leq \Gamma^{-1} \| \gamma - \beta \|$. □

REMARK. If $\int_S UU' \, d\mu$ is p.d., then (3.15) holds (Ruppert (1979), Section 4.6). Also, A6 implies that $\int_S UU' \, d\mu$ is p.d.

LEMMA 3.2 *Suppose that* $M(x, s)$ *is a strictly increasing function of x for each s, and h is a positive function on S. Assume there exists f such that*

$$M(f(s), s) = 0 \quad \textit{for all } s \tag{3.16}$$

for some γ_0,

$$\int_S h(s)(\gamma_0' U(s) - f(s))^2 \, d\mu(s) < \infty \tag{3.17}$$

for some $K > 0$,

$$|M(x, s)| \leq K |x - f(s)| \quad \textit{for all } x \textit{ and } s, \tag{3.18}$$

$$\mu\{s: \gamma' U(s) \neq 0\} > 0 \tag{3.19}$$

for all γ *in* $R^k - \{0\}$, *and*

$$\int_S h(s) \| U(s) \|^2 \, d\mu(s) < \infty. \tag{3.20}$$

Then A1 *holds.*

PROOF. By (3.17), (3.18), and (3.20), A1(i) holds. Define

$$G(x, s) = h(s) \int_{f(s)}^{x} M(y, s) \, dy.$$

For fixed s, $M(x, s)$ is strictly increasing in x, and therefore $G(x, s)$ is a strictly convex function of x (Roberts and Varberg, 1973, Theorem A, page 9).

Now define

$$Q(\gamma) = \int_S G(\gamma' U(s), s) \, d\mu(s).$$

By (3.17) and (3.18), $Q(\gamma) < \infty$ for all γ. Since $G(x, s)$ is strictly convex for each s, (3.19) implies that Q is strictly convex on R^k. Now fix γ_1 and γ_2 in R^k, $\gamma_2 \neq 0$. Since $M(x, s)$ is strictly increasing in x,

$$\lim_{|t| \to \infty} G((\gamma_1 + t\gamma_2)' U(s), s) = +\infty$$

and then (3.19) and Fatou's lemma imply

$$\lim_{|t|\to\infty} Q(\gamma_1 + t\gamma_2) = +\infty.$$

Therefore, for any real a, the convex set

$$L_a = \{\gamma \in R^k\colon Q(\gamma) \le a\}$$

has no directions of recession (see Rockafellar, 1970, page 61) for a definition of direction of recession). By Theorem 27.1f of Rockafellar, the function Q has no directions of recession. Therefore by part d of the same theorem, the set of minimum points of Q is nonempty. Then since Q is strictly convex, Q has a unique minimum point, say β, by Theorem A, page 123, of Roberts and Varberg (1973). Therefore A1(ii) holds.

Using the mean value theorem, the dominated convergence theorem, and (3.16), (3.18), and (3.20), one can show that

$$\nabla \int G(\gamma' U(s), s)h(s)\, d\mu(s) = \int \nabla G(\gamma' U(s), s)h(s)\, d\mu(s),$$

i.e., $\nabla Q(\gamma) = H(\gamma)$. Therefore, by Theorem A, page 98, of Roberts and Varberg (1973),

$$Q(\beta) - Q(\gamma) > (\beta - \gamma)' H(\gamma),$$

which proves A1(iii) since β is the unique minimum point of Q and Q is a finite valued and convex, whence continuous, function so that for all $\varepsilon > 0$

$$\sup_{\varepsilon \le \|\gamma-\beta\| \le \varepsilon^{-1}} Q(\beta) - Q(\gamma) < 0.$$

REMARKS. Under the assumptions of Lemma 3.2, β depends upon h. Let ψ be a strictly decreasing function on $(-\infty, 0]$ and a strictly increasing function on $[0, \infty)$, and assume there exists a minimum point, β_0, of

$$\int_S \psi(M(\beta_0' U(s), s)\, d\mu(s).$$

By the nature of ψ, $\beta_0' U(s)$ is a reasonable approximation to $f(s)$. Then assuming $M'(x, s) = \frac{\partial}{\partial x} M(x, s)$ exists, ψ' exists, and differentiation under the integral sign is permissible, β_0 satisfies

$$\int_S h(s) M(\beta_0' U(s), s) U(s)\, d\mu(s)$$

where

$$h(s) = \psi'(M(\beta_0' U(s), s) M'(\beta_0' U(s), s)/M(\beta_0' U(s), s)$$

(let 0/0 be 0). Then, in principle, if we can only approximate f, we can choose ψ to serve as a criterion for approximation and then select the appropriate h. Of course, in practice there would not be enough information to evaluate h initially, though adaptive procedures which estimate h may be feasible.

The Optimal Choice of D and h. The choice $\alpha = 1$ maximizes the rate of convergence, but choosing h and D is less straightforward. Since the bias $(\beta' U(s) - f(s))$ does not tend to 0 as n increases, while the variance does tend to 0, the choice of h (which determines β) is of paramount importance, at least for large n. Suppose, however, that β does not depend on h, as is true when $\beta' U(s) = f(s)$ for some β (see Lemma 1). Then the choice $h = F$ and $D = A^{-1}$ is optimal in the following sense.

THEOREM 3.4. *Under the conditions of Theorem* 3.1, $(B^{-1/2}P\mathscr{M}P'B^{-1/2} - A^{-1})$ *is* p.s.d. *If* $h = F$ *and* $D = A^{-1}$, *then* $B^{-1/2}P\mathscr{M}P'B^{-1/2} = A^{-1}$.

PROOF. Since $\Lambda = P'B^{1/2}DB^{1/2}P$, we have $P'B^{1/2}(DCD)B^{1/2}P = \Lambda(P'B^{-1/2}CB^{-1/2}P)\Lambda$. Therefore

$$\mathscr{M}^{(ij)} = \Lambda^{(ii)}\Lambda^{(jj)}(\Lambda^{(ii)} + \Lambda^{(jj)} - 1)^{-1}(P'B^{-1/2}CB^{-1/2}P)^{(ij)}.$$

Fix $x \in R^k$. Examination of the partial derivatives of $\sum_{i=1}^k \sum_{j=1}^k \mathscr{M}^{(ij)}x_ix_j$, with respect to $\Lambda^{(ii)}$ subject to $\Lambda^{(ii)} > 1/2$, $i = 1, \ldots, k$, shows that $x'(\mathscr{M} - P'B^{-1/2}CB^{-1/2}P)x$ is nonnegative. Therefore $(\mathscr{M} - P'B^{-1/2}CB^{-1/2}P)$ is p.s.d. and it follows that $(B^{-1/2}P\mathscr{M}P'B^{-1/2} - B^{-1}CB^{-1})$ is p.s.d. Since

$$\int_S (h(s)B^{-1} - F(s)A^{-1})U(s)U'(s)(h(s)B^{-1} - F(s)A^{-1})'\,d\mu(s)$$
$$= B^{-1}CB^{-1} + A^{-1}AA^{-1} - 2(A^{-1}BB^{-1}) = B^{-1}CB^{-1} - A^{-1}$$

is p.s.d., it follows that $B^{-1/2}P\mathscr{M}P'B^{-1/2} - A^{-1}$ is p.s.d.

Finally, if $h = F$ and $D = A^{-1}$, then $A = B = C = D^{-1}$ and consequently $B^{1/2}DB^{1/2} = I$ and we can take $P = I$. It then follows that $\mathscr{M} = I$ and $B^{-1/2}P\mathscr{M}P'B^{-1/2} = A^{-1}$.

REMARKS. Suppose that $M(x, s) = F(s)(x - \beta'U(s))$ and F is known. Then $Y_n - F(s_n)x_n = (-F(s_n)U_n'\beta + (Y_n - E^{\mathscr{F}_n}Y_n)$. Let $z_n = Y_n - F(s_n)x_n, Z_n' = (z_1, \ldots, z_n)$, $w_n = -F(s_n)U_n$, $W_n = (w_1, \ldots, w_n)$, $e_n = Y_n - E^{\mathscr{F}_n}Y_n$, and $E_n' = (e_1, \ldots, e_n)$. Then $Z_n = W_n\beta + E_n$, Z_n and W_n are known, and the least squares estimate of β is $\tilde{\beta}_n = (W_n'W_n)^{-1}W_n'Z_n((W_nW_n)^{-1}$ exists eventually). The covariance of $n^{1/2}(\tilde{\beta}_n - \beta)$ is $n\sigma^2(W_n'W_n)^{-1} \to \sigma^2A^{-1}$, the same asymptotic covariance as our procedure with the optimal choice, $h = F$ and $D = A^{-1}$.

In Ruppert (1978), an adaptive procedure which estimates h and D is proposed. However, this procedure, which is a generalization of one due to Venter (1967), requires that at time n, unbiased estimates of $M(x, s_n)$ can be observed at two distinct values of x. Obtaining two estimates would be feasible, for example, in the dose-response example of the introduction if the drug was administered several times to each patient and the expected dose-response did not vary during the course of treatment. However, procedures which require only one estimate at a time would be preferable, and perhaps these can be developed by generalizing the work of Anbar (1978) and Lai and Robbins (1978, 1979).

Since our goal is to keep $M_n = M(x_n, s_n)$ as close to 0 as possible, and $M(\beta'U(s_n), s_n)$ is our approximation to $M(f(s_n), s_n) = 0$, we now calculate the asymptotic distribution of $\{M_n - M(\beta'U(s_n), s_n)\}$.

THEOREM 3.5. *Suppose the assumptions of Theorem* 3.1 *hold*, $h = F$, *and* $D = A^{-1}$. *Let* Z *be distributed* $N(0, \sigma^2A^{-1})$ *and let* s_0 *be* μ*-distributed and independent of* Z. *Then*

$$(n^{1/2}(\beta_n - \beta), s_n) \to_{\mathscr{D}} (Z, s_0), \tag{3.21}$$

$$n^{1/2}(M_n - M(\beta'U(s_n), s_n)) \to_{\mathscr{D}} F(s_0)(Z'U(s_0)), \tag{3.22}$$

and the asymptotic variance of $n^{1/2}(M_n - M(\beta'U(s_n), s_n))$ *is* $k\sigma^2$.

PROOF. The independence of β_n and s_n yields (3.21). By A5,

$$(M(x_n, s_n) - M(\beta'U(s_n), s_n)) = F(s_n)(\beta_n - \beta)'U(s_n) + o((\beta_n - \beta)'U(s_n))$$

which implies (3.22). Finally, since Z and s_0 are independent,

$$E(F(s_0)Z'U(s_0))^2 = E(Z'(EF^2(s_0)U(s_0)U(s_0)')Z) = EZ'AZ = \operatorname{tr} AEZZ' = k\sigma^2. \qquad \square$$

4. Monte-Carlo Results. Small sample behavior was investigated by simulation, where

$$M(x, s) = (3/2 + s)^{1/2}(x - f(s)),$$

$k = 3$, $U(s) = (1, s, s^2)'$, $f(s) = \beta'_* U(s) + \lambda s^3$ where $\beta_* = (1, 4, 2)'$, $S = [-½, ½]$, and n is the uniform distribution. For all μ, the error, $Y_n - M(x_n, s_n)$, is the sum of 12 independent random variables, each uniformly distributed on $(-½, ½)$. When $\lambda = 0$, Lemma 3.1 is applicable. The assumptions of Lemma 3.2 hold for all λ. Since for s fixed, M is linear in x,

$$F(s) = (3/2 + s)^{1/2}$$

and

$$A^{-1} = 75/7 \begin{bmatrix} 7/50 & 0 & -14/15 \\ \cdot & 4/5 & -8/15 \\ \cdot & \cdot & 104/9 \end{bmatrix} = \begin{bmatrix} 1.5 & 0 & -10 \\ \cdot & 8.571 & -5.714 \\ \cdot & \cdot & 123.8 \end{bmatrix}$$

regardless of h and β. By Lemma 3.2, A1 holds with

$$(\beta - \beta_*) = A^{-1} \int_{-1/2}^{1/2} \lambda s^3 U(s)\, ds = (\lambda/100)(-.36, 14.8, 4.42)'. \tag{4.1}$$

In all simulations, $h(s) = F(s)$ and $D = A^{-1}$ were used. The following conclusions emerge from Table 1, which summarizes the results of four simulations.

(1). Comparing Simulations I and II, we see that the only major effect of having $\lambda = -1$ rather than $\lambda = 0$ is that $E\beta_n^{(2)}$ is reduced by approximately 0.14. This result agrees with the asymptotics given by Lemma 3.2 and Equation (4.1). In particular, when $\lambda = -1$, the Monte-Carlo estimate of EM_n is not significantly different from 0, and the variance of M_n is virtually the same as when $\lambda = 0$. Of course, as $|\lambda|$ is increased, $f(s)$ will be approximated less successfully by a quadratic polynomial and EM_n^2 will increase.

(2). The effect of the starting value, β_1, is very large and persists even when $n = 200$. For example, when $n = 200$, $n\sigma^2(M_n)$ is 3.76 and 6.65, respectively, when $\beta_1 = (1, 4, 2)'$ and $\beta_1 = (0, 0, 0)$ (Simulations I and III).

(3). Although choosing $\alpha < 1$ will reduce the rate of convergence, for finite n, $\alpha = .875$ can be superior to $\alpha = 1$. Comparing Simulations III and IV, where only α differs between the two, one sees that $\alpha = .875$ has lower variances of $\beta_n^{(1)}$, $\beta_n^{(2)}$, $\beta_n^{(3)}$, and M_n for all finite n reported ($n = 25, 50, 200$). (The situation is the same for $n = 100$, but not for $n = 10$.)

For further evidence of the effect of using $\alpha < 1$, see Table 2 where $n\sigma^2(M_n)$ (possibly the most important characteristic for comparison purposes) is shown for $n = 10, 25, 50, 100$, and 200 and $\alpha = .5, .75, .875$, and 1. Note that throughout $h(s) = F(s)$ and $D = A^{-1}$. If, for example, h was changed by multiplication by a positive constant, then the relative performance of the different α might change. Unfortunately, there is very little known about the finite sample behavior of stochastic approximation methods, even in simpler situations than that studied in this paper. I am only aware of a minimax result of Dvoretzky (1956), which, under some restrictions, applies to the one-dimensional Robbins-Monro method. Therefore, there would be little available as a guide in an attempt here to improve the small sample efficiency of asymptotically optimal procedures.

ACKNOWLEDGEMENT. I wish to thank the referee for his suggestions which led to several significant improvements in an earlier draft of this paper.

TABLE 1

A comparison of Monte-Carlo and asymptotic means, variances, and covariances. All finite sample size results are based upon 1000 Monte-Carlo trials.

	Simulation I $\beta_1 = (1, 4, 2)'$ $\alpha = 1, \quad \lambda = 0$				Simulation II $\beta_1 = (1, 4, 2)'$ $\alpha = 1, \quad \lambda = -1$			
	$n = 25$	50	200	∞	25	50	200	∞
$E(\beta_n^{(1)} - 1)$	−.011	−.007	.000	0	−.0082	−.0037	.0035	.0036
$E(\beta_n^{(2)} - 4)$	.005	.011	.011	0	−.143	−.137	−.137	−.148
$E(\beta_n^{(3)} - 2)$	.216	.089	.038	0	.174	.0497	−.0057	.0442
$n\sigma^2(\beta_n^{(1)})$	2.44	1.92	1.63	1.5	2.41	1.91	1.63	1.5
$n\sigma^2(\beta_n^{(2)})$	19.35	14.82	9.86	8.57	19.1	14.7	9.84	8.57
$n\sigma^2(\beta_n^{(3)})$	357	244	161	124	348	240	161	123.8
$n\sigma(\beta_n^{(1)}, \beta_n^{(2)})$	−.673	−.244	−.269	0	−.562	−.204	−.264	0
$n\sigma(\beta_n^{(1)}, \beta_n^{(3)})$	−22.0	−15.4	−11.7	−10	−21.5	−15.1	−11.7	−10
$n\sigma(\beta_n^{(2)}, \beta_n^{(3)})$	−4.37	−8.62	−5.55	−5.71	−5.64	−8.77	−5.52	−5.71
EM_n	−.00273	−.00469	−.00496	0	−.00560	−.00656	.00408	0
$n\sigma^2(m_n)$	7.27	5.25	3.76	3	7.14	5.25	3.75	3

	Simulation III $\beta_1 = (0, 0, 0)'$ $\alpha = 1, \quad \lambda = 0$				*Simulation* IV $\beta_1 = (0, 0, 0)'$ $\alpha = .875, \quad \lambda = 0$			
	25	50	200	∞	25	50	200	∞
$E(\beta_n^{(1)} - 1)$	.004	.003	.002	0	.0052	.0025	.0028	0
$E(\beta_n^{(2)} - 4)$	.040	−.018	.005	0	−.0329	−.0040	.0075	0
$E(\beta_n^{(3)} - 2)$	.017	−.062	.008	0	−.0469	−.0669	.0065	0
$n\sigma^2(\beta_n^{(1)})$	7.85	4.75	2.30	1.5	5.33	2.58	2.00	∞
$n\sigma^2(\beta_n^{(2)})$	81.2	44.4	18.7	8.57	56.6	21.4	11.4	∞
$n\sigma^2(\beta_n^{(3)})$	1882	1010	375	123.8	1123	388	181	∞
$n\sigma(\beta_n^{(1)}, \beta_n^{(2)})$	−9.92	−4.51	−1.54	0	−4.91	−.98	−.29	0
$n\sigma(\beta_n^{(1)}, \beta_n^{(3)})$	−99.1	−52.7	−21.7	−10	−60.2	−22.5	−14.2	∞
$n\sigma(\beta_n^{(2)}, \beta_n^{(3)})$	163	68.2	20.8	−5.71	57.9	6.27	−5.48	∞
EM_n	.0435	.00726	.00409	0	.0168	−.00331	.00559	0
$n\sigma^2(M_n)$	32.3	35.3	6.65	3	19.4	18.9	4.26	∞

TABLE 2

$n\sigma^2(M_n)$ *for* $\lambda = 0$, $\beta_1 = (0, 0, 0)$, *and selected values of* α *and* n. *All values are based upon* 1000 *Monte-Carlo trials.*

	α			
n	1	.875	.75	.5
10	113	113	109	527
25	32.3	19.4	12.2	37.1
50	35.2	18.9	14.7	17.9
100	9.63	5.10	6.92	22.7
200	6.65	4.26	6.58	22.2

REFERENCES

ANBAR, DAN (1978). A stochastic Newton-Raphson method. *J. Statist. Plan. Inference.* **2** 153–163.

DVORETZKY, ARYEH (1956). On stochastic approximation. *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, 1 (Jerzy Neyman, ed.) 39–55. Univ. California Press.

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

LAI, T. L., and ROBBINS, HERBERT (1978). Adaptive design in regression and control. *Proc. Natl. Acad. Sci.* **75** 586–587.

LAI, T. L., and ROBBINS, HERBERT (1979). Local convergence theorems for adaptive stochastic approximation schemes. *Proc. Natl. Acad. Sci.* **76** 3065–3067.

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, New York.

ROBERTS, A. WAYNE, and VARBERG, DALE E. (1973). *Convex Functions.* Academic, New York.

RUPPERT, DAVID (1978). Stochastic approximation of an implicitly defined function. *Institute of Statistics Mimeo Series #*1164. Chapel Hill, North Carolina.

RUPPERT, DAVID (1979). A new dynamic stochastic approximation procedure. *Ann. Statist.* **6** 1179–1195.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
321 PHILLIPS HALL 039 A
CHAPEL HILL, NORTH CAROLINA 27514