The Annals of Statistics
1985, Vol. 13, No. 1, 236–245

A NEWTON-RAPHSON VERSION OF THE MULTIVARIATE ROBBINS-MONRO PROCEDURE

BY DAVID RUPPERT[1]

University of North Carolina, Chapel Hill

Suppose that f is a function from $\mathbb{R}^k$ to $\mathbb{R}^k$ and for some θ, $f(\theta) = 0$. Initially f is unknown, but for any x in $\mathbb{R}^k$ we can observe a random vector $Y(x)$ with expectation $f(x)$. The unknown θ can be estimated recursively by Blum's (1954) multivariate version of the Robbins-Monro procedure. Blum's procedure requires the rather restrictive assumption that infimum of the inner product $(x - \theta)'f(x)$ over any compact set not containing θ be positive. Thus at each x, $f(x)$ gives information about the direction towards θ. Blum's recursion is $X_{n+1} = X_n - a_n Y_n$ where the conditional expectation of Y_n given $X_1, \cdots, X_n$ is $f(X_n)$ and $a_n > 0$. Unlike Blum's method, the procedure introduced in this paper does not necessarily attempt to move in a direction that decreases $\| X_n - \theta \|$, at least not during the initial stage of the procedure. Rather, except for random fluctuations it moves in a direction which decreases $\|f\|^2$, and it may follow a circuitous route to θ. Consequently, it does not require that $(x - \theta)'f(x)$ have a constant signum. This new procedure is somewhat similar to the multivariate Kiefer-Wolfowitz procedure applied to $\|f\|^2$, but unlike the latter it converges to θ at rate $n^{-1/2}$. Deterministic root finding methods are briefly discussed. The method of this paper is a stochastic analog of the Newton-Raphson and Gauss-Newton techniques.

1. Introduction. This paper is concerned with a multivariate version of a problem first studied by Robbins and Monro (1951). Supose that f is an unknown function from $\mathbb{R}^k$ to $\mathbb{R}^k$, and that for any x in $\mathbb{R}^k$ we can observe a random vector $Y(x)$ with expectation $f(x)$. Let α in $\mathbb{R}^k$ be known, and suppose there is a unique θ such that $f(\theta) = \alpha$. The goal is to estimate θ.

For example, suppose that $k = 2$ and x gives the doses of two drugs that affect blood chemistry. The concentrations of two chemicals in the blood are measured after administration of the drugs, and $f(x)$ gives the expected concentrations as a function of the doses. If α gives the ideal concentrations of the two blood components, then θ gives the correct doses.

By choosing the appropriate measurement scales, we can without loss of generality assume that $\alpha = 0$.

Blum's (1954) version of the Robbins-Monro (RM) process begins with an initial estimate X_1 of θ. Given $X_1, \cdots, X_n$, one observes Y_n, such that $E_n(Y_n) = f(X_n)$ where E_n denotes conditional expectation given by $X_1, \cdots, X_n$. Then X_n

Received September 1983; revised July 1984.
[1] This research was supported by National Science Foundation Grant MCS-8100748.
AMS 1980 *subject classifications.* Primary 62L20.
Key words and phrases. Root finding, stochastic approximation, asymptotic normality, asymptotic efficiency, Gauss-Newton algorithm.

is updated by the recursion

$$X_{n+1} = X_n - a_n Y_n \tag{1.1}$$

where a_n is a suitably chosen positive sequence converging to 0. Convergence of X_n to θ is proved under the assumption that for each $\varepsilon > 0$.

$$\inf\{(x - \theta)^t f(x): \varepsilon < \| x \| < \varepsilon^{-1}\} > 0. \tag{1.2}$$

The importance of (1.2) can be easily seen as follows. We will suppose that $\sup\{\text{Var } Y(x): x \in \mathbb{R}^k\} < \infty$, though a somewhat weaker assumption is possible. From (1.1) and the fact that $E_n(Y_n) = f(X_n)$, we have

$$E_n(\| X_{n+1} - \theta \|^2) = \| X_n - \theta \|^2 - 2a_n(X_n - \theta)^t f(X_n) + a_n^2 E_n(\| Y_n \|^2).$$

If we could ignore the term of order a_n^2, then $\| X_n - \theta \|^2$ would be a positive supermartingale and would converge a.s. The term of order a_n^2 can be handled using a theorem on "almost" positive supermartingales (Robbins and Siegmund, 1971, which also appears in Ruppert, 1981). The theorem also can be used to show that

$$\sum_{n=1}^{\infty} a_n(X_n - \theta)^t f(X_n) \quad \text{converges a.s.} \tag{1.3}$$

The sequence $\{a_n\}$ is chosen so that

$$\sum_{n=1}^{\infty} a_n = \infty, \tag{1.4}$$

and (1.2)–(1.4) imply that $X_n \to \theta$ a.s. Authors using condition (1.2) or something quite similar include Sacks (1958, assumption (A1*)), Schmetterer (1968, assumption (4.15)), and Walk (1977, assumption (2a)). (Walk's paper concerns the RM process in a general Hilbert space.) Blum's original work uses assumptions stronger than (1.2).

Unfortunately, (1.2) is a rather restrictive assumption, implying that at each x, $f(x)$ "points away from θ." Clearly we can replace (1.2) by

$$\inf\{-(x - \theta)^t f(x): \varepsilon < \| x \| < \varepsilon^{-1}\} > 0; \tag{1.2′}$$

this requires only that we change (1.1) to $X_{n+1} = X_n + a_n Y_n$. An example of a function satisfying neither (1.2) nor (1.2′) is

$$f(x_1, x_2) = (\exp(-x_1^2) - 1, \exp(-x_2^2) - 1)^t.$$

An alternative to the multivariate RM procedure would be to apply the multivariate Kiefer-Wolfowitz (KW) procedure to minimize $\| f(x) \|^2$. (For the moment assume that the conditional variance of $Y(x)$ is independent of x, so that $E_n(\| Y_n \|^2) = \| f(X_n) \|^2 + \text{constant}$. Otherwise, the KW procedure will find the minimizer of $E \| Y(x) \|^2$, not necessarily the solution to $f(x) = 0$.) The KW procedure will tend to follow the negative gradient of $\| f(x) \|^2$. Thus, the KW procedure does not attempt to move *directly* towards θ, in the sense of attempting to decrease $\| X_n - \theta \|$. However, except for random fluctuations, it does move downhill, that is, in a direction decreasing $\| f(x) \|^2$. Therefore, under mild

conditions X_n will converge to a local minimum of $\| f(x) \|^2$. If θ is the only local minimum, then $X_n \to \theta$.

Unfortunately, the rate of convergence to θ of the KW method is slower than the $n^{-1/2}$ rate of RM method, though modifications of the Kiefer-Wolfowitz method can produce rates arbitrarily close to $n^{-1/2}$ if f has derivatives of sufficiently high order (Fabian, 1971).

In this paper, we propose a new multivariate RM process which in some ways behaves as the Kiefer-Wolfowitz method applied to $\| f \|^2$, but which possesses the $n^{-1/2}$ rate of convergence even when f has only two derivatives.

Let $D(x)$ be the $k \times k$ derivative of $f(x)$. Then, the derivative of $\| f(x) \|^2$ is

$$2\, D^t(x) f(x).$$

The Hessian matrix of $\| f(x) \|^2$ is

$$H(x) = 2[D^t(x)D(x) + \textstyle\sum_{i=1}^{k} H^{(i)}(x) f^i(x)], \tag{1.5}$$

where f^i is ith coordinate of f and $H^{(i)}$ is the Hessian of f^i. Thus $H(\theta) = 2\, D^t(\theta)D(\theta)$. The recursive procedure that is introduced here is

$$X_{n+1} = X_n - \mathbf{a} n^{-1} B_n D_n^t f_n, \tag{1.6}$$

where $\mathbf{a} > 0$, B_n is an estimate of $[D^t(\theta)D(\theta)]^{-1}$, D_n is an estimate of $D(X_n)$, and f_n is an estimate of $f(X_n)$.

Blum's multivariate RM procedure uses one observation to construct f_n and thereby to update X_n to X_{n+1}. Our procedure uses $(2k)(m_n)$ observations to construct D_n and $[n^\gamma]$ observations to construct f_n, where $[\cdot]$ is the greatest integer function, $\gamma > 0$, $m_n \to \infty$, and $m_n n^{-\gamma} \to 0$. We let $m_n \to \infty$ sufficiently fast (see below), so that the conditional variance of D_n, given $X_1, \cdots, X_n$, converges to 0. We require that $m_n n^{-\gamma} \to 0$ so that among the totality of observations used to construct both D_n and f_n, the proportion used in estimating D converges to 0. These properties insure full asymptotic efficiency (see Section 5).

Comparing (1.1) and (1.6), one sees that the procedure introduced here differs from Blum's in that f_n is premultiplied by $(B_n D_n^t)$. The factor D_n^t serves to rotate f_n, and the expectation of $(D_n^t f_n)$ is a descent direction for the function $\| f(x) \|^2$. Thus, D_n^t is the key to obtaining consistency when condition (1.2) is not imposed. The factor B_n is needed to obtain asymptotic efficiency, but could be omitted without sacrificing consistency. Also, Blum's a_n is set equal to $\mathbf{a} n^{-1}$ here. This choice of $\{a_n\}$ is asymptotic efficient for a particular $\mathbf{a}$ (see Corollary 3.3), so more general sequences $\{a_n\}$ are not considered.

The estimator f_n is simply the mean of $[n^\gamma]$ observations with conditional expectation, given $X_1, \cdots, X_n$, equal to $f(X_n)$. The ith column of D_n is constructed as follows. Let $e(i)$ be the ith column of the $k \times k$ identity matrix. Let $c_n > 0$ be a constant and let $Y(n, i, 2)$ and $Y(n, i, 1)$ each be the mean of m_n observations with conditional expectation equal to $f(X_n + c_n e(i))$ and $f(X_n - c_n e(i))$, respectively. Then, the ith column of D_n is

$$D_n^{(i)} = [Y(n, i, 2) - Y(n, i, 1)]/(2c_n).$$

We choose m_n and c_n so that $c_n \to 0$ and $m_n^{-1}c_n^{-2} \to 0$. With this choice of c_n and m_n, the bias of D_n as an estimate of $D(X_n)$ converges to 0, and as mentioned above the conditional variance of D_n also converges to 0.

B_n is constructed as follows. Let $\underline{\eta}_n$ and $\bar{\eta}_n$ be positive sequences such that $\underline{\eta}_n \downarrow 0$, $\bar{\eta}_n \uparrow \infty$, and certain other conditions (see Section 2) are met.

Let

$$C_n = (n-1)^{-1} \sum_{i=1}^{n-1} D_i^t D_i.$$

Let $B_n = C_n^{-1}$ if all eigenvalues of C_n^{-1} lie between $\underline{\eta}_n$ and $\bar{\eta}_n$. Otherwise let B_n be some symmetric matrix whose eigenvalues are all between $\underline{\eta}_n$ and $\bar{\eta}_n$.

The procedure of this paper bears some resemblance to the one-dimensional Venter (1967) procedure. It was shown by Chung (1954) that the univariate Robbins-Monro procedure is asymptotically optimal when $a_n = 1/(nf'(\theta))$. Of course, $f'(\theta)$ will typically be unknown. Venter introduced a consistent estimate b_n of $f'(\theta)$ and showed that asymptotic optimality could be achieved with $a_n = 1/(n\, b_n)$.

Our procedure also estimates f' but at each X_n, not simply at θ. Moreover, the Venter process and the original RM process are consistent under roughly the same circumstances. Our procedure is an attempt to improve the consistency properties of Blum's multivariate RM process. The matrix sequence B_n does, however, play a role for our process which is analogous to that of b_n in the Venter process.

It should be mentioned that Blum's version of the RM process may be preferable to the one introduced here under certain circumstances, namely when either (1.2) or (1.2′) is known to hold and $D^t(x)f(x) = 0$ for some $x \neq \theta$. However, when neither (1.2) nor (1.2′) hold, our procedure, but not necessarily Blum's, at least tends to move in a direction decreasing $\| f(x) \|^2$.

We conclude the introduction with a discussion of deterministic methods of function optimization and root-finding, and their relationships with stochastic approximation. Much of this material is taken from Fletcher (1980). Another good, recent reference is Gill, Murray, and Wright (1981). Let $r(x)$ be a function from $\mathbb{R}^k$ to $\mathbb{R}$ with gradient $\nabla r(x)$ and Hessian $\nabla^2 r(x)$. Newton's method for minimizing $r(x)$ is the recursion

$$x_{n+1} = x_n - [\nabla^2 r(x_n)]^{-1}(\nabla r(x_n)).$$

Motivation for this method and a discussion of its theoretical properties and practical limitations can be found in Fletcher (1980). In practice, the need to supply formulas for the second derivatives of r and to invert $[\nabla^2 r(x_n)]$ can prove burdensome. Quasi-Newton methods replace $[\nabla^2 r(x_n)]^{-1}$ by a matrix H_n such that H_{n+1} can be calculated from H_n by a simple updating formula. This formula utilizes $(x_{n+1} - x_n)$ and $(\nabla r(x_{n+1}) - \nabla r(x_n))$ in a clever manner to obtain information about the inverse Hessian $[\nabla^2 r(x_n)]^{-1}$. Under general conditions $(H_n - [\nabla^2 r(x_n)]^{-1}) \to 0$, although the updating formula does not require explicit expressions for $\nabla^2 r$ or matrix inversions. The well-known Davidon-Fletcher-Powell (DFP) and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods are quasi-Newton procedures which have proved successful in practice.

Now suppose that

$$r(x) = \sum_{i=1}^{n} f_i^2(x) = \| f(x) \|^2,$$

where $f(x) = (f_1(x), \cdots, f_n(x))^t$, $x \in \mathbb{R}^k$, and $n \geq k$. For such "sum of squares" problems, one can bypass the need for second derivative formulas either by the general quasi-Newton methods discussed above or by the Gauss-Newton method which takes advantage of the special structure of $r(x)$. As in (1.5),

$$\nabla^2 r(x) = 2 \sum_{i=1}^{n} \{(\nabla f_i(x))(\nabla f_i(x))^t + (\nabla^2 f_i(x)) f_i(x)\},$$

and

$$\nabla^2 r(x) \approx 2 \sum_{i=1}^{n} (\nabla f_i(x))(\nabla f_i(x))^t \tag{1.7}$$

if $f_i(x) \approx 0$ near the minimum of $r(x)$. The Gauss-Newton method uses the approximation (1.7). Note that

$$\nabla r(x) = 2 D^t(x) f(x)$$

where $D^t(x) = (\nabla f_1(x), \cdots, \nabla f_n(x))$, and the RHS of (1.7) is $2D^t(x)D(x)$. The Gauss-Newton recursion is

$$x_{n+1} = x_n - [D^t(x_n)D(x_n)]^{-1} D^t(x_n) f(x_n). \tag{1.8}$$

It is assumed that $D(x_n)$ is of rank k. When $n = k$, $D(x)$ is invertible, $[D^t(x)D(x)]^{-1} D(x) = D^{-1}(x)$, and (1.8) reduces to

$$x_{n+1} = x_n - D^{-1}(x_n) f(x_n). \tag{1.9}$$

Because $n = k$, there may very well be a solution, x^*, to

$$f_i(x^*) = 0, \quad i = 1, \cdots, k. \tag{1.10}$$

Then x^* minimizes $r(x)$, and the Gauss-Newton procedure (1.9) is a method of solving (1.10). In this context the Gauss-Newton procedure is called the Newton-Raphson method.

The problem studied in this paper is a stochastic version of (1.10), and the algorithm introduced here could be considered a stochastic version of the Newton-Raphson technique. However, when developing the algorithm, consistency was more easily proved by *not* estimating $[D^t(X_n)D(X_n)]^{-1}$ as $(D_n^t D_n)^{-1}$ where D_n is used to estimate $D(X_n)$. Since it is more similar to (1.8) than (1.9), perhaps our procedure is better viewed as a stochastic Gauss-Newton method.

An early method of function minimization, the method of steepest descent, is, in our notation,

$$x_{n+1} = x_n - \nabla r(x_n).$$

Steepest descent makes no use, explicitly or implicitly, of information about second derivatives and its convergence can be exceedingly slow. It's use is no longer recommended. If we omitted B_n in (1.6) then we would have a stochastic steepest-descent algorithm. Except for a Newton-like stochastic algorithm in Fabian (1971), multivariate Kiefer-Wolfowitz procedures found in the literature are analogs of the method of steepest descent.

An indication of the unsatisfactory nature of Blum's multivariate RM procedure may be the fact that it apparently has no deterministic analogs.

2. Notation and assumptions. Let $\mathbb{R}^k$ be k-dimensional Euclidean space. If A is a $k \times \ell$ matrix, let A^{ij} be the i, jth entry of A, and if $\ell = 1$ let $A^i = A^{i1}$. Also, let A^t be the transpose of A, and let $\| A \|^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} (A^{ij})^2$.

All random variables are defined on the same probability space, and all relations between random variables are meant to hold with probability one. Let $\to_{\mathscr{L}}$ denote convergence in law.

We say that $a_n \sim b_n$ if $a_n/b_n \to 1$. The "O" and "o" notation has its usual meaning.

The following assumptions on f will be needed.

F1. (i) $f = (f_1, \cdots, f_k)^t$ is a twice differentiable function from $\mathbb{R}^k$ to $\mathbb{R}^k$, θ is in $\mathbb{R}^k$, and $f(\theta) = 0$.

(ii) $D(x)$ is the derivative of f, i.e., $D^{ij}(x) = (\partial/\partial x_j) f^i(x)$.

(iii) $D(\theta)$ is nonsingular.

(iv) For all $\varepsilon > 0$,

$$\inf\{\| D^t(x) f(x) \| : \varepsilon \le \| f(x) \| \le \varepsilon^{-1}\} > 0.$$

(v) $\sup\{\| D(x) \| : x \in \mathbb{R}^k\} < \infty$.

(vi) Let $H(x)$ be the Hessian of $\| f(x) \|^2$, i.e., $H^{ij}(x) = (\partial^2/\partial x_i \partial x_j) \| f(x) \|^2$. Then $\sup\{\| H(x) \| : x \in \mathbb{R}^k\} < \infty$.

F2. For all $\varepsilon > 0$

$$\inf\{\| f(x) \| : \varepsilon \le \| x - \theta \| \le \varepsilon^{-1}\} > 0.$$

The modified Robbins-Monro algorithm will be described formally by the following assumptions.

A1. (i) $X_{n+1} = X_n - \mathbf{a} n^{-1} B_n D_n f_n$ where $\mathbf{a} > 0$, X_n is in $\mathbb{R}^k$ and B_n and D_n are $k \times k$ random matrices.

(ii) $c_n > 0$, $c_n \downarrow 0$, m_n is an integer, $m_n c_n^2 \uparrow \infty$, $\gamma > 0$, $m_n n^{-\gamma} \to 0$, $\underline{\eta}_n \downarrow 0$, $\bar{\eta}_n^2 n^{-2} = o(\underline{\eta}_n n^{-1})$, $\bar{\eta}_n \uparrow \infty$,

$$\sum n^{-1} \underline{\eta}_n = \infty, \tag{2.1}$$

and

$$\sum n^{-1} \bar{\eta}_n [c_n^2 + n^{-1} \bar{\eta}_n (c_n^2 + c_n^{-2} m_n^{-1} + n^{-\gamma})] < \infty. \tag{2.2}$$

(iii) For any random vector X, let $E_n(X)$ and $\mathrm{Var}_n(X)$ be respectively the condition mean and variance of X given $X_1, \cdots, X_n$. If X is a random matrix, then $\mathrm{Var}_n(X)$ is the conditional variance of X arranged as a column vector. Let $\mathscr{F}_n$ be the σ-algebra generated by $X_1, \cdots, X_n$. Then B_n is $\mathscr{F}_{n-1}$ measurable, and

$E_n(f_n) = f(X_n)$. Define $\bar{D}_n = E_n(D_n)$. There exists a constant K such that

$$\| \bar{D}_n - D(X_n) \| \leq Kc_n^2, \quad \| \operatorname{Var}_n(f_n) \| \leq Kn^{-\gamma}, \quad \text{and}$$

$$\| \operatorname{Var}_n(D_n) \| \leq Km_n^{-1}c_n^{-2} \quad \text{for all } n.$$

Given $\mathcal{F}_n$, f_n and D_n are conditionally independent.

(iv) B_n is symmetric and all eigenvalues of B_n are between $\underline{\eta}_n$ and $\bar{\eta}_n$.

A2. (i) $\mathbf{a} > (1 + \gamma)/2$

(ii) If $X_n \to \theta$, then $B_n \to (D^t(\theta)D(\theta))^{-1}$ and $n^\gamma \operatorname{Var}_n(f_n) \to S$ for some matrix S, and for all $r > 0$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^n E\, I\{\| V_i \|^2 \geq r\, i\} \| V_i \|^2 = 0,$$

where $V_n = n^{\gamma/2}(\bar{D}_n^t f(X_n) - D_n^t f_n)$, and $\bar{D}_n = E_n(D_n)$.

REMARKS ON THE ASSUMPTIONS. F1 (vi) corresponds to the assumption of a bounded Hessian which has been used in the study of the Kiefer-Wolfowitz algorithm, e.g., Fabian (1971, assumption (2.2)). Any substantially weaker condition would probably require that the step sizes, $\mathbf{a}n^{-1}B_nD_nf_n$, in A1 (i) be modified to prevent increasingly large oscillations. Otherwise, X_n might have no finite limit points.

Assumptions F1 (iv) and F2 are also similar to conditions that would be needed if the KW process were applied to $\| f(x) \|^2$. See Fabian (1971, assumption (2.2) equation (1)). They imply that θ is the only local minimum of $\| f(x) \|^2$.

Given the method described in the introduction of constructing B_n, D_n, and f_n, assumptions A1 (iii), A1 (iv) and A2 (ii) are natural. To have $\| E_n(D_n) - D(X_n) \| \leq Kc_n^2$, it is sufficient that the second derivative of f be uniformly bounded. Simple conditions sufficient for A2 (ii) can be found using standard martingale techniques. The assumptions on $\operatorname{Var}_n(f_n)$ and $\operatorname{Var}_n(D_n)$ are reasonable and will be met if the measurements $Y(x)$ have bounded variances and if for any constants $s_1, \cdots, s_q$ one can take measurements $Y(X_n + s_i)$, $i = 1, \cdots, q$, that are conditionally independent given $X_1, \cdots, X_n$.

Condition A1 (ii) is satisfied if $0 < \alpha < \gamma$, $m_n = [n^\alpha]$, $c_n = n^{-\alpha/3}$, $\underline{\eta}_n = (\log(n + 1))^{-1}$, and $\bar{\eta}_n = \log(n + 1)$.

3. Theorems.

THEOREM 3.1. *Assume* F1 *and* A1. *Then* $f(X_n) \to 0$. *If* F2 *also holds, then* $X_n \to \theta$.

THEOREM 3.2. *Assume* F1, F2, A1, *and* A2. *Then*

$$n^{(1+\gamma)/2}(X_n - \theta) \to_{\mathcal{L}} N(0, [\mathbf{a}^2/(2\mathbf{a} - 1 - \gamma)]D^{-1}(\theta)SD^{-t}(\theta))$$

COROLLARY 3.3. *Under* F1, F2, A1, *and* A2, *the choice* $\mathbf{a} = (1 + \gamma)$ *is optimal.*

With this choice,

$$N_n^{1/2}(X_n - \theta) \to_{\mathscr{L}} N(0, D^{-1}(\theta)S\, D^{-t}(\theta))$$

where N_n is the total number of observations needed to construct X_n.

4. Proofs.

PROOF OF THEOREM 3.1. Recall the definition $\bar{D}_n = E_n(D_n)$. Define $d_n = D_n - \bar{D}_n$, and $\varepsilon_n = f_n - f(X_n)$. By F1 (ii), A1 (i), and A1 (iii) there is a ζ in (0, 1) such that

$$\begin{aligned} E_n(\| f(X_{n+1}) \|^2 = \| f(X_n) \|^2 &- 2\mathbf{a}n^{-1} f^t(X_n) D(X_n) B_n D^t(X_n) f(X_n) \\ &- 2\mathbf{a}n^{-1} f^t(X_n)\{\bar{D}_n - D(X_n)\} B_n D^t(X_n) f(X_n) \\ &+ \mathbf{a}^2 n^{-2} E_n\{(B_n D_n^t f_n)^t H(X_n - \zeta \mathbf{a}n^{-1} B_n D_n^t f_n)(B_n D_n^t f_n)\}. \end{aligned} \tag{4.1}$$

By F1 (v) and A1 (iii),

$$| f^t(X_n)\{\bar{D}_n - D(X_n)\} B_n D^t(X_n) f(X_n) | = O(\| f(X_n) \|^2 \lambda_n c_n^2) \tag{4.2}$$

where λ_n is the largest eigenvalue of B_n. By F1 (vi)

$$E_n[(B_n D_n^t f_n)^t H(X_n - \zeta \mathbf{a}n^{-1} B_n D_n^t f_n)(B_n D_n^t f_n)] = O(\lambda_n^2 E_n \| D_n^t f_n \|^2). \tag{4.3}$$

By A1 (iii)

$$\begin{aligned} &E_n \| D_n^t f_n \|^2 \\ &\quad = \| D^t(X_n) f(X_n) \|^2 + f^t(X_n)\{\bar{D}_n \bar{D}_n^t - D(X_n) D^t(X_n)\} f(X_n) \\ &\qquad + f^t(X_n)(E_n d_n d_n^t) f(X_n) + E_n[\varepsilon_n^t(\bar{D}_n \bar{D}_n^t + d_n d_n^t)\, \varepsilon_n] \\ &\quad = \| D^t(X_n) f(X_n) \|^2 + O\{(c_n^2 + c_n^{-2} m_n^{-1}) \| f(X_n) \|^2 + n^{-\gamma}\}. \end{aligned} \tag{4.4}$$

By (2.2), (4.1) to (4.4), and A1 (iv),

$$\begin{aligned} &E_n(\| f(X_{n+1}) \|^2) \\ &\quad \le \| f(X_n) \|^2 \{1 + 2\mathbf{a}n^{-1} \bar{\eta}_n c_n^2 + O(n^{-2} \eta_n^{-2}(c_n^2 + c_n^{-2} m_n^{-1}))\} \\ &\qquad - (2\mathbf{a}n^{-1} \underline{\eta}_n - O(\bar{\eta}_n^2 n^{-2})) \| D^t(X_n) f(X_n) \|^2 + O(\bar{\eta}_n^2 g^{-\gamma-2}) \\ &\quad = \| f(X_n) \|^2 (1 + \mu_n) - (2\mathbf{a}n^{-1} \underline{\eta}_n)(1 + o(1)) \| D^t(X_n) f(X_n) \|^2 + \nu_n \end{aligned}$$

where $\sum \mu_n < \infty$ and $\sum \nu_n < \infty$. Therefore, by Theorem 1 of Robbins and Siegmund (1971), $\lim_{n\to\infty} \| f(X_n) \|$ exists and is finite and

$$\sum_{n=1}^{\infty} n^{-1} \underline{\eta}_n \| D^t(X_n) f(X_n) \|^2 < \infty.$$

Then by F1 (iv) $\| f(X_n) \|^2 \to 0$, and therefore $X_n \to \theta$ if F2 holds. □

PROOF OF THEOREM 3.2. By Theorem 3.1 and A2 (ii), $B_n \to (D^t(\theta) D(\theta))^{-1}$,

$\lambda_n = O(1)$, and $D(X_n) \to D(\theta)$. For each $\eta > 0$

$$f^t(X_n)D(X_n)B_nD^t(X_n)f(X_n) \geq (1 - \eta)\| f(X_n)\|^2$$

for all sufficiently large n. Therefore, by (4.1) to (4.4), for each $\eta > 0$

$$E_n(\| f(X_{n+1})\|^2 \leq \| f(X_n)\|^2\{1 - 2\mathbf{a}(1 - \eta)n^{-1} + O(c_n^2 n^{-1} + n^{-2})\} + O(n^{-2-\gamma})$$

$$\leq \| f(X_n)\|^2(1 - 2\mathbf{a}(1 - 2\eta)n^{-1}) + O(n^{-2-\gamma})$$

for all n sufficiently large. Note that $(n + 1)^{1+\varepsilon} = n^{1+\varepsilon} + n^{\varepsilon}(1 + \varepsilon) + O(n^{\varepsilon-1})$. For each $0 < \varepsilon < \gamma$ and for each $\eta > 0$ and for all large n,

$$(n + 1)^{1+\varepsilon}E_n(\| f(X_{n+1})\|^2 \leq$$

$$(1 - [2\mathbf{a}(1 - \eta) - (1 + \varepsilon)]n^{-1})n^{1+\varepsilon}\| f(X_n)\|^2 + O(n^{-1-(\gamma-\varepsilon)}).$$

Therefore, by A2 (i) and another application of Theorem 1 of Robbins and Siegmund (1971), $\lim_{n\to\infty} n^{1+\varepsilon}\| f(X_n)\|^2$ exists and is finite for all $\varepsilon < \gamma$, whence

$$n^{1+\varepsilon}\| f(X_n)\|^2 \to 0 \tag{4.5}$$

for all $\varepsilon > \gamma$.

There exists a matrix D_n^* which is $\mathscr{F}_n$-measurable such that $f(X_n) = D_n^* X_n$ and $D_n^* \to D(\theta)$. We can now apply Theorem 2.2 of Fabian (1968), which also appears in Ruppert (1981). Fabian's theorem is applied with $\Gamma_n = aB_n\bar{D}_n^t D_n^*$, $\Gamma = aI$, $\alpha = 1$, $\beta = (1 + \gamma)$, $U_n = X_n$, $\Phi_n = aB_n$, $\Phi = \mathbf{a}D^{-1}(\theta)D^{-t}(\theta)$, $V_n = n^{\gamma/2}(\bar{D}_n^t f(X_n) - D_n^t f_n)$, $T_n = T = 0$, $P = I$, $\Lambda = \mathbf{a}I$, and $\sum = \lim_{n\to\infty} E_n(V_n V_n^t)$. Note that $\beta = \beta_+$ and $(\Lambda^{(ii)} + \Lambda^{(jj)} - \beta_+) = (2\mathbf{a} - 1 - \gamma)$ for all i and j. We need to calculate $\sum$ more explicitly.

If V_1, V_2, W_1, and W_2 are random variables possessing finite second moments such that (V_1, V_2) is independent of (W_1, W_2), then

$$\mathrm{Cov}(V_1W_1, V_2W_2) = \mathrm{Cov}(V_1, V_2)\mathrm{Cov}(W_1, W_2) + \mathrm{Cov}(V_1, V_2)(EW_1)(EW_2)$$

$$+ \mathrm{Cov}(W_1, W_2)(EV_1)(EV_2).$$

Applying this fact coordinatewise to $D_n^t f_n$ and using A1 (ii) and A1 (iii), one can show that

$$\| \mathrm{Var}_n(D_n^t f_n) - \bar{D}_n^t(\mathrm{Var}_n(f_n))\bar{D}_n\| \leq L\{\| f(X_n)\|^2 c_n^{-2} m_n + n^{-\gamma}c_n^{-2}m_n\} \tag{4.6}$$

for some constant L. Then by A2 (ii) and (4.5)

$$\sum = \lim_{n\to\infty} n^{\gamma}\mathrm{Var}_n(D_n^t f_n) = D^t(\theta)S\,D(\theta).$$

Finally, to use Fabian's theorem one must verify that

$$\| E_n V_n V_n^t - \sum\| < C \tag{4.7}$$

for a positive constant C. This could be done using (4.6) if $\| f(X_n)\|$ converged to 0 uniformly. By Egorov's Theorem, for each $\varepsilon > 0$, $\| f(X_n)\| \to 0$ uniformly on a set of probability at least $(1 - \varepsilon)$. On the complementary set we may change the definition of V_n so that (4.7) holds. Since the resulting process, say X_n^*, agrees with X_n on a set of probability at least $(1 - \varepsilon)$ and since ε is an arbitrary positive number, we are done. □

PROOF OF COROLLARY 3.3. It is trivial to show that $\mathbf{a}^2/(2\mathbf{a} - 1 - \gamma)$ is minimized in $\mathbf{a}$, subject to the constraint that $\mathbf{a} > (1 + \gamma)/2$, by $\mathbf{a} = (1 + \gamma)$. The corollary then follows because $N_n = \sum_{i=1}^{n} \{[i^\gamma] + m_i\} \sim \sum_{i=1}^{n} [i^\gamma] \sim n^{1+\gamma}/(1 + \gamma)$. □

5. Asymptotic efficiency. We will not treat the subject of efficiency in great detail, but we will study a simple example. Suppose D is nonsingular and known, $f(x) = D(x - \theta)$, and Var $Y(x) = S$ for all x. If $Y(x)$ is normally distributed, then $x - D^{-1}Y(x) \sim N(\theta, D^{-1}SD^{-t})$. Thus, if $Z_1, \cdots, Z_{N_n}$ is any sequence of random variables, then the maximum likelihood estimate of θ based on $Y(Z_1), \cdots, Y(Z_{N_n})$ is

$$\hat{\theta} = N_n^{-1} \sum_{i=1}^{N_n} [Z_i - D^{-1}Y(Z_i)].$$

Also $\hat{\theta} \sim N(\theta, N_n^{-1}D^{-1}S\,D^{-t})$. Therefore, $\hat{\theta}$ and our estimator, X_n, have the same asymptotic distributions.

When deriving the asymptotic distribution of X_n, it is crucial that $\text{Var}_n(D_n) \to 0$ so that $\text{Var}_n(D_n^t f_n) \doteq D^t(\theta)(\text{Var}_n(f_n))D(\theta)$. Then because $\text{Var}_n(D_n^t f_n)$ is determined by $\text{Var}_n(f_n)$, full efficiency is obtained by having $m_n n^{-\gamma} \to 0$, so that the ratio of the number of observations used to construct D_n to the number used to construct f_n converges to 0.

Acknowledgement. I wish to thank Jed Frees and an anonymous referee for their useful comments on an earlier version of this paper.

REFERENCES

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

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

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

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

FLETCHER, R. (1980). *Practical methods of optimization, Vol.* 1, *unconstrained optimization.* Wiley, Chicester.

GILL, P. E., MURRAY, W., and WRIGHT, M. H. (1981). *Practical Optimization.* Academic, London.

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

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statitsics.* (J. S. Rustagi, ed.) 233–257. Academic, New York.

RUPPERT, D. (1981). Stochastic approximation of an implicitly defined function. *Ann. Statist.* **9** 555–566.

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

SCHMETTERER, L. (1969). Multidimensional stochastic approximation. (In *Multivariate Analysis, II,* P. R. Krishnaiah, ed.) Academic, New York.

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

WALK, H. (1977). An invariance principle for the Robbins-Monro Process in a Hilbert Space. *Z. Wahrsch. verw. Gebiete* **31** 135–150.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514