The Annals of Statistics
1987, Vol. 15, No. 3, 1115–1130

MULTIVARIATE ADAPTIVE STOCHASTIC APPROXIMATION[1]

BY C. Z. WEI

University of Maryland

Herein we study a multivariate version of the adaptive stochastic approximation developed recently by Lai and Robbins. An adaptive procedure which involves a Venter-type estimate of the Jacobian of the response function is proposed and shown to be asymptotically efficient from both the estimation and the control points of view.

1. Introduction. Consider the regression model

$$Y_n = f(X_n) + \varepsilon_n, \qquad n = 1, 2, \dots, \tag{1.1}$$

where $f\colon R^p \to R^p$ is a Borel function, Y_n is a $p \times 1$ observable random vector and ε_n are i.i.d. $p \times 1$ random vectors with mean vector zero and covariance matrix Σ. Let $Y_0 = f(\theta)$ be a known optimal response vector and θ be the unknown desired factor level which is to be estimated. The classical Robbins–Monro (1951) stochastic approximation, when $p = 1$, is as follows: Initialize X_1 and then choose X_n by the recursion

$$X_{n+1} = X_n - a_n(Y_n - Y_0),$$

where $\{a_n\}$ is a sequence of positive real numbers such that $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n^2 < \infty$. After n observations, the estimate of θ is X_{n+1}. By minimizing the asymptotic variance of $(X_n - \theta)$, it is known [Chung (1954) and Sacks (1958)] that the asymptotically optimal choice of a_n is $(n(\partial f/\partial\theta))^{-1}$. However, in practice $\partial f/\partial\theta$ is usually known. This raises the question of estimating $\partial f/\partial\theta$ and also leads us to consider the adaptive stochastic approximation procedure

$$X_{n+1} = X_n - \frac{A_n}{n}(Y_n - Y_0), \tag{1.2}$$

where A_n^{-1} is an estimate of $\partial f/\partial\theta$ based on the data already observed. Venter (1967) proposed a modified Robbins–Monro procedure (cf. Section 4) with a strongly consistent estimate A_n^{-1}. Since then an extensive literature was devoted to the Robbins–Monro procedure and its generalization [cf. Nevelson and Hasminskii (1973a) and Kushner and Clark (1978)].

Recently, however, Lai and Robbins (1978a, 1979) started investigating the adaptive stochastic approximation not only from the *estimation* (of θ) but also from the *control* point of view. As pointed out by Lai and Robbins (1979), in applications where X_n is the dosage level given to the nth patient and Y_n is the corresponding response level, minimum asymptotic variance of $(X_n - \theta)$ is of

Received May 1986; revised January 1987.
[1]Research supported by the National Science Foundation under Grant DMS-8404081.
AMS 1980 *subject classification*. Primary 62L20.
Key words and phrases. Adaptive stochastic approximation, Robbins–Monro process, asymptotically efficient.

interest only for future patients, and the *cost* [defined as $\Sigma_1^n(X_i - \theta)(X_i - \theta)'$] to the patients already treated should also be taken into consideration. Under smoothing conditions on the response function f, the minimization of $\Sigma_1^n(X_i - \theta)(X_i - \theta)'$ is asymptotically equivalent to the minimization of $\Sigma_1^n Y_i Y_i'$ which arises in adaptive control problems in econometrics [Lai and Robbins (1982)] and in the feedback control schemes for linear dynamic systems [Goodwin, Ramadge and Caines (1982) and Lai and Wei (1982)].

For the linear regression function $f(X) = B(X - \theta)$, where B is a $p \times p$ known matrix, if we use the least-squares estimator,

$$\theta_n^* = \overline{X}_n - B^{-1}\overline{Y}_n \quad \left(= \theta - B^{-1}\bar{\varepsilon}_n\right), \tag{1.3}$$

to estimate θ, then irrespective of how the levels X_i are chosen, whether preassigned or sequentially determined,

$$E\left[(\theta_n^* - \theta)(\theta_n^* - \theta)'\right] = \frac{1}{n}B^{-1}\Sigma(B^{-1})' \tag{1.4}$$

and

$$\sqrt{n}\,(\theta_n^* - \theta) \to N\left(0, B^{-1}\Sigma(B^{-1})'\right) \quad \text{in distribution.} \tag{1.5}$$

It follows from (1.3) and (1.4) that the expected cost of the adaptive procedure,

$$X_{n+1} = \overline{X}_n - B^{-1}\overline{Y}_n \quad \left(= X_n - \frac{1}{n}B^{-1}Y_n\right), \tag{1.6}$$

at stage n is of the order of $\log n$, i.e.,

$$E\left[\sum_1^n (X_i - \theta)(X_i - \theta)'\right] = (\log n)B^{-1}\Sigma(B^{-1})' + O(1).$$

In ignorance of B, it is natural to try using a stochastic approximation scheme of the form (1.2). Of course, we want A_n^{-1} in (1.2) to be a strongly consistent estimator of B. For the regression model (1.1), under the assumptions that $p = 1$ and $B = \partial f/\partial\theta$, Lai and Robbins (1979) established some sufficient conditions on A_n^{-1} to ensure such a scheme has the following asymptotically optimal properties:

$$\sqrt{n}\,(X_n - \theta) \to N\left(0, B^{-1}\Sigma(B^{-1})'\right) \quad \text{in distribution;} \tag{1.7}$$

$$\begin{aligned} &P\Big[\text{set of limit points of} \\ &\left\{(X_n - \theta)\left[n/(2\log\log n)\right]^{1/2}: n \geq 3\right\} = K\Big] = 1, \\ &\text{where } K = \left\{X: X'B^{-1}\Sigma(B^{-1})'X \leq 1\right\}; \end{aligned} \tag{1.8}$$

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{j=1}^n (X_j - \theta)(X_j - \theta)' = B^{-1}\Sigma(B^{-1})' \quad \text{a.s.} \tag{1.9}$$

Furthermore, they showed that a modified least-squares estimate of B as well as a modified Venter's estimate satisfy the required conditions.

In this paper we shall consider the multivariate versions of the theory developed by Lai and Robbins. Blum (1954) is the first one who established the strong consistency of a multivariate version of the Robbins–Monro procedure,

$$X_{n+1} = X_n - a_n(Y_n - Y_0),$$

under the assumption that for all $\varepsilon > 0$,

$$\inf\{(X - \theta)'(f(X) - Y_0): \varepsilon < \|X\| < \varepsilon^{-1}\} > 0. \tag{1.10}$$

This is a quite restrictive assumption for it implies that we know how to adjust the direction of X after observing $f(X)$. Although, for the case $p = 1$, there are only two possible directions; for $p > 1$, there are infinite possible directions to be chosen. A theory which does not require (1.10) seems more desirable. Nevelson and Hasminskii (1973b) (using a Venter-type estimate) showed that an adaptive stochastic approximation of the form (1.2) has the asymptotically optimal property (1.7) under the less restrictive assumption

$$\frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial y}\right)' + \frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x}\right)' \tag{1.11}$$

is positive definite uniformly with respect to x, $y \in R^p$.

Condition (1.11) is satisfied by any linear function $f(X) = BX + C$ with $\det(B) \neq 0$. A recent attempt along this line is due to Ruppert (1985). However, neither Nevelson and Hasminskii nor Ruppert had considered the problem from the control point of view. In fact, the procedures proposed by Nevelson and Hasminskii (1973b) are rather inefficient in the sense that the associated costs grow algebraically instead of logarithmically (cf. Remark 4 of Section 4). In Section 3, sufficient conditions are imposed on A_n so that the optimal properties (1.7)–(1.9) can be achieved. In Section 4, a generalized Venter estimate is proposed and the corresponding procedure is shown to be optimal under the assumption (1.11). For the sake of completeness, a short discussion on the optimal choice of A in the stochastic approximation scheme,

$$X_{n+1} = X_n - \frac{A}{n} Y_n, \tag{1.12}$$

is also included in Section 2. Note that Y_0 is assumed to be zero in (1.12). This can be done without loss of generality and will be assumed throughout the sequel.

2. The optimal choice of A in the stochastic approximation scheme (1.12). Let B be a nonsingular $p \times p$ matrix, θ a $p \times 1$ vector and ε_n a sequence of i.i.d. $p \times 1$ random vectors with $E\varepsilon_i = 0$, $E\varepsilon_i\varepsilon_i' = \Sigma$. In order to find the root of the linear function $B(X - \theta)$ based on the observations,

$$Y_n = B(X_n - \theta) + \varepsilon_n, \qquad n = 1, 2, \dots,$$

we consider stochastic approximation procedures of the form

$$X_{n+1} = X_n - \frac{A}{n} Y_n,$$

where A is a $p \times p$ nonsingular matrix. In view of the identities (1.3) and (1.6), the asymptotic variance of $(X_n - \theta)$ is of the order $1/n$ when $A = B^{-1}$. In order that the bias of X_n (as an estimate of θ) be negligible relative to the covariance matrix of X_n, it is natural to restrict to the class of matrices A for which

$$\lim n^{1/2}E(X_n - \theta) = 0.$$

As shown by Nevelson and Hasminskii (1973b), this is true iff all eigenvalues of the matrix $C = I/2 - AB$ have negative real parts. We denote this class of matrices by $\mathscr{D}$. It is known [Nevelson and Hasminskii (1973b)] that if $A \in \mathscr{D}$, then

$$n^{1/2}(X_n - \theta) \to N(0, \sigma(A)) \quad \text{in distribution},$$

where

$$\sigma(A) = \int_0^\infty e^{Ct}A\Sigma A'e^{C't}\,dt. \tag{2.1}$$

In order to get the optimum choice of A, we have to "minimize" the matrix $\sigma(A)$ in the sense of the following

THEOREM 1. *If $A \in \mathscr{D}$, then $\sigma(A) - \sigma(B^{-1})$ is nonnegative definite.*

Before proving Theorem 1, we quote a lemma from Daleckii and Krein (1974).

LEMMA 1. *Let G, H, Y be $p \times p$ matrices such that all eigenvalues of G and H have negative real parts. Then the equation*

$$GX + XH = Y$$

has a unique solution

$$X = -\int_0^\infty e^{Gt}Ye^{Ht}\,dt.$$

PROOF OF THEOREM 1. Let

$$A \in \mathscr{D}, \qquad C = I/2 - AB, \qquad D = \sigma(A) - \sigma(B^{-1}).$$

Apply Lemma 1 to the case $G = C$, $H = C'$. Then by (2.1),

$$C\sigma(A) + \sigma(A)C' = -A\Sigma A'. \tag{2.2}$$

Now,

$$\begin{aligned}\sigma(B^{-1}) &= \int_0^\infty e^{-(I/2)t}B^{-1}\Sigma(B^{-1})'e^{-(I/2)t}\,dt \\ &= B^{-1}\Sigma(B^{-1})'.\end{aligned} \tag{2.3}$$

In view of (2.2) and (2.3),

$$\begin{aligned}CD + DC' &= -A\Sigma A' - (I/2 - AB)B^{-1}\Sigma(B^{-1})' \\ &\quad - B^{-1}\Sigma(B^{-1})'(I/2 - B'A') \\ &= -A\Sigma A' - B^{-1}\Sigma(B^{-1})' + A\Sigma(B^{-1})' + B^{-1}\Sigma A' \\ &= -(A - B^{-1})\Sigma(A - B^{-1})'.\end{aligned}$$

By Lemma 1, again,

$$D = \int_0^\infty e^{Ct}(A - B^{-1})\Sigma(A - B^{-1})' e^{C't}\, dt$$

is nonnegative definite. □

REMARK 1. Nevelson and Hasminskii (1973b) have proved Theorem 1 by using the Cramér–Rao inequality. Since Lemma 1 is true even for the Banach space case [cf. Daleckii and Krein (1974)], our algebraic approach has the advantage that it can be generalized to the corresponding case where the order of the operators is defined. For possible applications along this line, we refer to Walk (1977).

3. Some lemmas. Throughout the sequel, we denote the norm of a $p \times p$ matrix A by

$$\|A\| = \sup\{\|AX\|\colon X \in R^p, \|X\| = 1\}.$$

LEMMA 2. *Let B be a nonsingular $p \times p$ matrix and A_n, B_n two sequences of random $p \times p$ matrices such that $B_n \to B$ a.s. and $A_n \to B^{-1}$ a.s. Let Z_n be an arbitrary sequence of random p-dimensional vectors and X_1, X_1^* two arbitrary random p-dimensional vectors. Suppose X_n, X_n^* are defined recursively by*

$$X_{n+1} = X_n - (A_n/n)(B_n X_n + Z_n), \tag{3.1}$$

$$X_{n+1}^* = X_n^* - (I/n)X_n^* - (A_n Z_n)/n, \quad \text{for } n \geq 1. \tag{3.2}$$

Suppose that

$$(n/\log\log n)^{1/2} X_n^* = O(1) \quad a.s. \tag{3.3}$$

Then

$$(n/\log\log n)^{1/2}(X_n - X_n^*) = o(1) \quad a.s. \tag{3.4}$$

Furthermore, if we add the assumptions

$$\|B_n - B\|^2 = o(1/\log\log n), \tag{3.5}$$

$$\|A_n - B^{-1}\|^2 = o(1/\log\log n) \quad a.s., \tag{3.6}$$

then

$$\|X_n - X_n^*\| = o(1/n). \tag{3.7}$$

PROOF. Let $C_n = A_n B_n$, $D_n = I - A_n B_n$, $Y_n = X_n - X_n^*$, for $n \geq 0$. By (3.1) and (3.2),

$$\begin{aligned} Y_{n+1} &= Y_n - (C_n/n)(X_n - X_n^*) + (I - C_n)X_n^*/n \\ &= (I - C_n/n)Y_n + (I - C_n)X_n^*/n \\ &= (I - C_n/n) \cdots (I - C_k/k)Y_k \qquad (3.8) \\ &\quad + \sum_{m=k}^{n-1} (I - C_n/n) \cdots (I - C_{m+1}/m + 1)D_m/mX_m^* + D_n X_n^*/n. \end{aligned}$$

Since $B_n \to B$, $A_n \to B^{-1}$ a.s., $P(\Omega_0) = 1$, where Ω_0 denotes the event $\{C_n \to I\}$. For each $w \in \Omega_0$, there exists $N = N(w)$ and $d_n = d_n(w)$ such that $\lim d_n = 1$ and

$$d_n < 1, \qquad \|I - C_n/n\| \le 1 - d_n/n, \quad \text{for } n \ge N. \tag{3.9}$$

Let

$$\gamma_n = \left[(1 - d_n/n) \cdots (1 - d_N/N)\right]^{-1}$$

and

$$\delta_n = \|D_n\|, \quad \text{for } n \ge N.$$

Then, in view of (3.8), (3.9) and the trigonometric inequality,

$$\|Y_{n+1}\| \le \gamma_n^{-1}\|Y_N\| + \sum_{k=N}^{n} \gamma_n^{-1}\gamma_k\delta_k\|X_k^*\|/k. \tag{3.10}$$

Note that $d_n \to 1$ implies that γ_n is a regularly varying sequence with exponent 1 and consequently [cf. Bojanic and Seneta (1973)],

$$n^{1/2}\gamma_n^{-1}\|Y_n\| = o(1). \tag{3.11}$$

Now, in view of (3.3),

$$\begin{aligned} &\gamma_n^{-1} \sum_{k=N}^{n} \gamma_k\delta_k\|X_k^*\|/k \\ &\quad \le a\gamma_n^{-1} \sum_{k=N}^{n} \gamma_k(\log\log k/k^3)^{1/2}\delta_k, \quad \text{for some } a > 0. \end{aligned} \tag{3.12}$$

In order to show (3.4), by (3.10)–(3.12), we need only show that

$$g_n = (n/\log\log n)^{1/2}\gamma_n^{-1} \sum_{k=N}^{n} \gamma_k(\log\log k/k^3)^{1/2} = O(1),$$

since $\delta_n = o(1)$. This can be demonstrated by the relation

$$0 \le g_n \le n^{1/2}\gamma_n^{-1} \sum_{k=N}^{n} \gamma_k/(k)^{3/2} \to 2$$

[cf. Bojanic and Seneta (1973)]. Now we are going to show (3.7). Let

$$b_n = (\log\log n)^{1/2}\delta_n.$$

By (3.5), (3.6) and the definition of δ_n,

$$\begin{aligned} b_n &\le (\log\log n)^{1/2}\left(\|B\|\,\|B^{-1} - A_n\| + \|B - B_n\|\,\|A_n\|\right) \\ &= o(1). \end{aligned}$$

In view of (3.3) and (3.10),

$$\|Y_{n+1}\|^2 \le 2\gamma_n^{-2}\|Y_N\|^2 + 2e\left(\gamma_n^{-1} \sum_{k=N}^{n} \gamma_k b_k/k^{3/2}\right)^2,$$

for some $e > 0$.

By (3.11) and the same argument which shows (3.4) above,

$$n\|Y_{n+1}\|^2 = o(1).$$

This completes the proof. □

LEMMA 3. *Let $u_1, u_2, \ldots$ be a sequence of i.i.d. random variables such that $Eu_i = 0$, $Eu_i^2 = \sigma^2 < \infty$. Let $\mathscr{F}_n$ be the Borel field generated by $u_1, \ldots, u_n$ ($\mathscr{F}_0$ = trivial σ-field). Let v_n be an $\mathscr{F}_{n-1}$-measurable random variable such that $\lim v_n = 0$ a.s. Then*

$$\lim\left(\sum_1^n v_i u_i\right)/(n \log\log n)^{1/2} = 0 \quad a.s.$$

PROOF. The result follows from Theorem 2 of Lai and Robbins (1978b). □

LEMMA 4. *Let ε_n be a sequence of i.i.d. p-dimensional random vectors such that $E\varepsilon_n = 0$ and $E\varepsilon_n\varepsilon_n' = \Sigma$. Let $\mathscr{F}_n$ be the Borel field generated by $\varepsilon_1, \ldots, \varepsilon_n$ and let A_n be an $\mathscr{F}_{n-1}$-measurable $p \times p$ nonsingular matrix. Let X_1 be an arbitrary p-dimensional random vector and δ_n a sequence of p-dimensional random vectors. For $n \geq 1$, define X_n recursively by*

$$X_{n+1} = (n-1)X_n/n - A_n(\delta_n + \varepsilon_n)/n. \tag{3.13}$$

If

$$(n/\log\log n)^{1/2}\delta_n = o(1) \quad a.s., \tag{3.14}$$

then with probability 1, *the set of cluster points of*

$$\left\{(n/\log\log n)^{1/2} X_n,\ n \geq 3\right\}$$

is

$$\{X \in R^p\colon X'B^{-1}\Sigma(B^{-1})'X \leq 1\}.$$

PROOF. In view of (3.13),

$$\begin{aligned} X_{n+1} &= -(1/n)\sum_{k=1}^n A_k(\delta_k + \varepsilon_k) \\ &= -(1/n)\sum_{k=1}^n A_k\delta_k - (1/n)\sum_{k=1}^n A_k\varepsilon_k. \end{aligned} \tag{3.15}$$

Let

$$a_n = (n/\log\log n)^{1/2}.$$

Since $\sup_k\|A_k\| < \infty$ a.s., it follows from (3.14) that

$$(a_n/n)\sum_{k=1}^n A_k\delta_k = o(1) \quad \text{a.s.} \tag{3.16}$$

Now,

$$(1/n)\sum_{k=1}^{n} A_k\varepsilon_k = (1/n)\sum_{k=1}^{n}(A_k - A)\varepsilon_k + (1/n)\sum_{k=1}^{n} A\varepsilon_k, \tag{3.17}$$

where $A = B^{-1}$. By Lemma 3,

$$(a_n/n)\sum_{k=1}^{n}(A_k - A)\varepsilon_k = o(1) \quad \text{a.s.} \tag{3.18}$$

Furthermore, by a theorem of Kuelbs (1977), the limit set of $(a_n/n)\Sigma_{k=1}^{n}A(-\varepsilon_k)$ is

$$\{X\colon X'A\Sigma A'X \le 1\}.$$

In view of (3.15)–(3.18), the proof is complete. □

Lemma 5. *Let the notation and assumptions be the same as in Lemma* 4.

(i) *If* $n^{1/2}\delta_n = o(1)$ *a.s., then*

$$n^{1/2}X_n \to N\big(0, B^{-1}\Sigma(B^{-1})'\big) \quad \textit{in distribution}. \tag{3.19}$$

(ii) *If* $(n\log\log n)^{1/2}\delta_n = o(1)$ *a.s., then*

$$\lim\left(\sum_{1}^{n} X_kX_k'\right)\Big/\log n = B^{-1}\Sigma(B^{-1})' \quad \textit{a.s.} \tag{3.20}$$

Proof. Let $A = B^{-1}$, $S_n = \Sigma_{k=1}^{n}A_k\varepsilon_k$, $T_n = \Sigma_{k=1}^{n}A_k\delta_k$. In view of (3.15),

$$-(1/n)X_n = S_n + T_n. \tag{3.21}$$

To prove (i), since $n^{1/2}\delta_n = o(1)$ a.s. and $\sup_k\|A_k\| < \infty$, a.s.,

$$n^{-1/2}T_n = o(1) \quad \text{a.s.} \tag{3.22}$$

Now,

$$E(A_k\varepsilon_k\varepsilon_k'A_k' | \mathscr{F}_{k-1}) = A_k\Sigma A_k' \to A\Sigma A' \quad \text{a.s.}$$

By Theorem 1 of Dvoretzky (1977),

$$n^{-1/2}S_n \to N(0, A\Sigma A') \quad \text{in distribution.} \tag{3.23}$$

In view of (3.21)–(3.23),

$$\begin{aligned} n^{1/2}X_n &= n^{-1/2}S_n + n^{-1/2}T_n \\ &= n^{-1/2}S_n + o(1) \to N(0, A\Sigma A') \quad \text{in distribution.} \end{aligned}$$

To prove (ii), by (3.21)

$$\begin{aligned} \sum_{k=1}^{n} X_kX_k' = {} & \sum_{k=1}^{n} S_kS_k'/k^2 + \sum_{k=1}^{n} S_kT_k'/k^2 \\ & + \sum_{k=1}^{n} T_kS_k'/k^2 + \sum_{k=1}^{n} T_kT_k'/k^2. \end{aligned} \tag{3.24}$$

Since $(n \log\log n)^{1/2}\delta_n = o(1)$ and $\sup_k \|A_k\| < \infty$ a.s.,

$$\|T_n\|^2 = o(n/\log\log n) \quad \text{a.s.} \tag{3.25}$$

Hence,

$$\left\| \sum_{k=1}^{n} T_k T_k'/k^2 \right\| \le \sum_{k=1}^{n} \|T_k/k\|^2 = o(\log n) \quad \text{a.s.,} \tag{3.26}$$

and in view of Lemma 4 and (3.25),

$$\begin{aligned} \left\| \sum_{k=1}^{n} S_k T_k'/k^2 \right\| &\le \sum_{k=1}^{n} (\|S_k\|/k)(\|T_k\|/k) \\ &= \sum_{k=1}^{n} O\left((\log\log k/k)^{1/2}\right) o\left((k \log\log k)^{-1/2}\right) \\ &= o(\log n) \quad \text{a.s.} \end{aligned} \tag{3.27}$$

Similarly,

$$\left\| \sum_{k=1}^{n} T_k S_k'/k^2 \right\| = o(\log n) \quad \text{a.s.} \tag{3.28}$$

By (3.24) and (3.26)–(3.28), in order to prove (3.20), we need only show that

$$\lim (1/\log n) \sum_{k=1}^{n} S_k S_k'/k^2 = A\Sigma A' \quad \text{a.s.} \tag{3.29}$$

Now,

$$\begin{aligned} \sum_{k=1}^{n} S_k S_k'/k^2 &= \sum_{k=1}^{n} S_k S_k'/k(k+1) \\ &\quad + \sum_{k=1}^{n} (S_k/k)(S_k/k)'/(k+1) \\ &= I_{n1} + I_{n2} \quad \text{(say).} \end{aligned} \tag{3.30}$$

Since $S_k = o(k)$ a.s. by Lemma 4,

$$\|I_{n2}\| \le \sum_{k=1}^{n} \|S_k/k\|^2/(k+1) = o(\log n). \tag{3.31}$$

We note that

$$\begin{aligned} I_{n1} &= \sum_{k=1}^{n} S_k S_k'/k - \sum_{k=1}^{n} S_k S_k'/(k+1) \\ &= S_1 S_1' + \sum_{k=2}^{n} (S_k S_k' - S_{k-1} S_{k-1}') - S_n S_n'/(n+1) \\ &= J_{n1} + J_{n2} - J_{n3} \quad \text{(say).} \end{aligned} \tag{3.32}$$

In view of (3.30)–(3.32) and Lemma 4,

$$\begin{aligned}\sum_{k=1}^{n} S_k S_k'/k^2 &= I_{n1} + o(\log n)\\ &= J_{n1} + J_{n2} - J_{n3} + o(\log n)\\ &= O(1) + J_{n2} + o(\log n) + o(\log n)\\ &= J_{n2} + o(\log n).\end{aligned}$$

Consequently, in order to show (3.29), it suffices to prove that

$$\lim J_{n2}/\log n = A\Sigma A' \quad \text{a.s.} \tag{3.33}$$

Now,

$$\begin{aligned}J_{n2} &= \sum_{k=2}^{n} A_k \varepsilon_k \varepsilon_k' A_k'/k + \sum_{k=2}^{n} (A_k \varepsilon_k S_{k-1}' + S_{k-1}\varepsilon_k' A_k')/k\\ &= L_{n1} + L_{n2} \quad (\text{say}).\end{aligned} \tag{3.34}$$

Since by Lemma 4,

$$\begin{aligned}E\big(\|A_n \varepsilon_n S_{n-1}'\|^2 | \mathscr{F}_{n-1}\big) &\le \|A_n\|^2 \|S_{n-1}\|^2 E\|\varepsilon_n\|^2\\ &= o(n \log\log n) \quad \text{a.s.,}\end{aligned}$$

it follows from the martingale convergence theorem that

$$\sum_{k=2}^{n} (A_k \varepsilon_k S_{k-1}' + S_{k-1}\varepsilon_k' A_k')/(k \log k)$$

converges a.s. Thus, in view of Kronecker's lemma,

$$L_{n2} = o(\log n) \quad \text{a.s.} \tag{3.35}$$

Let

$$V_n = \sum_{k=1}^{n} A_k \varepsilon_k \varepsilon_k' A_k'.$$

Then

$$\begin{aligned}L_{n1} &= \sum_{k=2}^{n} (V_k - V_{k-1})/k\\ &= \sum_{k=2}^{n-1} (1/k - 1/(k+1)) V_k - V_1/2 + V_n/n\\ &= \sum_{k=2}^{n-1} (1/k+1) V_k/k - V_1/2 + V_n/n.\end{aligned}$$

If it can be shown that

$$V_n/n \to A\Sigma A' \quad \text{a.s.,} \tag{3.36}$$

then

$$L_{n1} = \sum_{k=2}^{n-1} (1/k + 1)(A\Sigma A' + o(1)) + O(1) \tag{3.37}$$
$$= (\log n)A\Sigma A' + o(\log n) \quad \text{a.s.},$$

and in view of (3.34), (3.35) and (3.37), (3.33) is proved. Hence, it remains to show (3.36).

Since $A_k \to A$ a.s., by Theorem 1 of Lai and Robbins (1978b),

$$(V_n)_{ij}/n \to (A\Sigma A')_{ij} \quad \text{a.s.},$$

where $(M)_{ij}$ denotes the (i, j) element of the matrix M. Hence, (3.36) holds, and the proof is complete. □

4. Generalized Venter estimator of the Jacobian matrix and asymptotic properties of multivariate Venter-type stochastic approximation schemes. Throughout the sequel we shall assume that the mean response function f: $R^p \to R^p$ satisfies conditions (4.9) and (4.10). For the adaptive stochastic approximation scheme (4.8) of this section, successive estimates of the desired level θ are constructed at stages $n = 2pm$, $m = 1, 2, \ldots$. This is due to the fact that it is necessary to use at least $2p$ vector observations to estimate the values of the matrix $\partial f/\partial x$. This refinement is a generalization of Venter's scheme in the case $p = 1$ [cf. Venter (1967), Nevelson and Hasminskii (1973b) and Lai and Robbins (1979)].

We now describe a multivariate extension of Venter-type designs. In the following, we shall let e_j denote the unit jth coordinate vector in R^p, i.e., all the components of e_j are zero except for the jth component which is 1. Let c_k and λ_k be two predetermined sequences of positive numbers. At stage m, define X_m recursively by

$$X_m = X_{m-1} - (m-1)^{-1} U_{m-1} Y_{m-1} \qquad (X_1 \text{ being arbitrary}), \tag{4.1}$$

where U_{m-1}^{-1} is the Venter-type estimate [see (4.7)] of the Jacobian matrix $\partial f/\partial \theta$ based on the observations up to stage $m - 1$ and Y_{m-1} is the "fitted" response at the level X_{m-1} [see (4.5)]. Take $2p$ observations $Y_{m1}^*, Y_{m1}^{**}, \ldots, Y_{mp}^*, Y_{mp}^{**}$ at the points

$$X_{mj}^* = X_m + c_m e_j, \tag{4.2}$$

$$X_{mj}^{**} = X_m - c_m e_j, \qquad j = 1, \ldots, p. \tag{4.3}$$

Thus,

$$\begin{aligned} Y_{mj}^* &= f(X_{mj}^*) + \varepsilon_{mj}^*, \\ Y_{mj}^{**} &= f(X_{mj}^{**}) + \varepsilon_{mj}^{**}, \qquad j = 1, \ldots, p. \end{aligned} \tag{4.4}$$

Set

$$Y_m = (1/2p)\sum_{j=1}^{p}\left(Y_{mj}^{*} + Y_{mj}^{**}\right), \tag{4.5}$$

$$W_m = \sum_{k=1}^{m-1}\lambda_k Z(k)\Big/\sum_{k=1}^{m-1}\lambda_k, \qquad W_1 = 0, \tag{4.6}$$

where $Z(k)$ is the matrix whose jth column is

$$Z_j(k) = (2c_k)^{-1}\left(Y_{kj}^{*} - Y_{kj}^{**}\right).$$

Define

$$U_m = \begin{cases} W_m^{-1}, & \text{if } \det W_m \neq 0, \\ I, & \text{if } \det W_m = 0. \end{cases} \tag{4.7}$$

The matrix U_m is used to estimate $(\partial f/\partial\theta)^{-1}$ at stage m.

REMARK 2. The above condition with

$$\lambda_n = 1, \qquad k_1 n^{-\alpha_1} \le c_n \le K_2 n^{-\alpha_2} \tag{4.8}$$

where $\alpha_1 < \frac{1}{2}$ and $\alpha_2 < \frac{1}{4}$ has been described by Nevelson and Hasminskii (1973b). However, this choice of λ_i is rather inefficient (cf. Remark 4).

Since at stage m, the Venter-type design described previously has taken $n = 2mp$ observations, we shall define the cost due to these n observations as

$$K_n = \sum_{k=1}^{m}\left\{\sum_{j=1}^{p}\left(X_{kj}^{*} - \theta\right)\left(X_{kj}^{*} - \theta\right)^{*} + \left(X_{kj}^{**} - \theta\right)\left(X_{kj}^{**} - \theta\right)'\right\}.$$

THEOREM 2. *Assume that*

(4.9) *the function $\partial f/\partial x$ is bounded and satisfies the Hölder condition , i.e., for $x,\ y \in R^p$,*

$$\left\|\frac{\partial f}{\partial x} - \frac{\partial f}{\partial y}\right\| \le \alpha\|x - y\|, \quad \textit{where } \alpha > 0, \textit{ a constant};$$

the symmetric function

$$\frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial y}\right)' + \frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x}\right)' \tag{4.10}$$

is positive definite uniformly w.r.t. $x,\ y \in R^p$;

(4.11) $\{\varepsilon_{mj}^{*}, \varepsilon_{mj}^{**},\ m = 1, 2, \dots,\ j = 1, 2, \dots, p\}$ *are i.i.d. random vectors with mean vector 0, covariance matrix Σ;*

$$c_k = o\left((k \log\log k)^{-1/2}\right); \tag{4.12}$$

$$\sum_{k=1}^{\infty}\left(\lambda_k/c_k\Lambda_k\right)^2 \log\log k < \infty, \qquad \Lambda_k^2/\log\log k \uparrow \infty, \tag{4.13}$$

where $\Lambda_2 = \sum_1^k \lambda_j$.

Let $n = 2mp$ *and* $\hat{\theta}_n = X_m$. *Then* (1.7)–(1.9) *hold with* $\hat{\theta}_n$ *replacing* X_n. *Furthermore*,

$$\lim (1/\log n) K_n = B^{-1}\Sigma(B^{-1})' \quad a.s., \text{ where } B = \frac{\partial f}{\partial \theta}. \tag{4.14}$$

REMARK 3. Condition (4.10) is due to Nevelson and Hasminskii (1973b). It is satisfied by any linear function $f(X) = BX + C$ with $\det B \neq 0$.

PROOF OF THEOREM 2. Without loss of generality we can assume that $\theta = 0$. By using condition (4.10) and the same argument as in Theorem 2.1 of Nevelson and Hasminskii (1973b), we can prove that there is a random variable $\frac{1}{2} > r > 0$ such that

$$\lim m^r X_m = 0 \quad \text{a.s.} \tag{4.15}$$

Set

$$B_m X_m = f(X_m), \qquad B_{mj}^* X_{mj}^* = f(X_{mj}^*), \qquad B_{mj}^{**} X_{mj}^{**} = f(X_{mj}^{**}), \tag{4.16}$$

$$\varepsilon_m = (1/2p) \sum_{j=1}^{p} (\varepsilon_{mj}^* + \varepsilon_{mj}^{**}), \qquad \delta_m = B_m X_m + \varepsilon_m - Y_m. \tag{4.17}$$

Then (4.1) becomes

$$X_m = X_{m-1} - (m-1)^{-1} U_{m-1}(B_{m-1} X_{m-1} + \delta_{m-1} + \varepsilon_{m-1}). \tag{4.18}$$

Note that by (4.9), (4.15)–(4.17),

$$\|B_m - B\| \leq d\|X_m\| = o(m^{-r}) = o(1/\log\log m) \quad \text{a.s.} \tag{4.19}$$

and

$$\begin{aligned} \|\delta_m\| &\leq (1/2p) \sum_{j=1}^{p} \left\| f(X_{mj}^*) + f(X_{mj}^{**}) - 2f(X_m) \right\| \\ &\leq d_1 \sum_{j=1}^{p} \|c_m e_j\| \quad (\text{for some random variable } d_1) \\ &= O(c_m) = o\big((m \log\log m)^{-1/2}\big) \quad \text{a.s., by (4.12).} \end{aligned} \tag{4.20}$$

By (4.6), the jth column of W_m is

$$\begin{aligned} W_m(j) &= \sum_{k=1}^{m-1} \lambda_k (2c_k)^{-1} \big(f(X_{kj}^*) - f(X_{kj}^{**}) \big) / \Lambda_{m-1} \\ &\quad + \sum_{k=1}^{m-1} \lambda_k (2c_k)^{-1} (\varepsilon_{kj}^* - \varepsilon_{kg}^{**}) / \Lambda_{m-1} \\ &= I_{m1}(j) + I_{m2}(j) \quad (\text{say}). \end{aligned} \tag{4.21}$$

Let B_j be the jth column of B. Then in view of (4.9) and (4.16), with probability 1,

$$
\begin{aligned}
&\left(f(X_{kj}^*) - f(X_{kj}^{**})\right)/2c_k \\
&\quad = \left(B_{kj}^* X_{kj}^* - B_{kj}^{**} X_{kj}^{**}\right)/2c_k \\
&\quad = \left(B_{kj}^* - B_{kj}^{**}\right)X_k/2c_k + \left(B_{kj}^* + B_{kj}^{**}\right)e_j/2 \\
&\quad = O\left(\|X_{kj}^* - X_{kj}^{**}\|\,\|X_k\|/c_k\right) + B_j + O\left(\|X_{kj}^*\| + \|X_{kj}^{**}\|\right) \\
&\quad = O(\|X_k\|) + B_j + O(\|X_k\| + c_k) \\
&\quad = B_j + o(k^{-r}) \\
&\quad = B_j + o(1/\log\log k) \quad \text{a.s., by (4.12) and (4.15).}
\end{aligned}
$$

Hence by (4.21),

$$
\begin{aligned}
\left\|I_{m1}(j) - B_j\right\| &\le \sum_{k=1}^{m-1} \lambda_k o(1/\log\log k)/\Lambda_{m-1} \\
&= o(1/\log\log m) \quad \text{a.s.}
\end{aligned} \tag{4.22}
$$

In view of (4.11), (4.13) and Kolmogorov's three series theorem,

$$
\sum_{k=1}^{m-1} \lambda_k (2c_k)^{-1}(\varepsilon_{kj}^* - \varepsilon_{kj}^{**})(\log\log k)^{1/2}/\Lambda_k \quad \text{converges a.s.}
$$

By Kronecker's lemma and (4.21),

$$
\left\|I_{m2}(j)\right\|^2 = o(1/\log\log m) \quad \text{a.s.} \tag{4.23}
$$

Thus in view of (4.21)–(4.23),

$$
\begin{aligned}
\|W_m - B\| &\le \sum_{j=1}^{p} \left\|W_m(j) - B_j\right\| \\
&\le \sum_{j=1}^{p} \left\|I_{m1}(j) - B_j\right\| + \left\|I_{m2}(j)\right\| \\
&= o\left((\log\log m)^{-1/2}\right) \quad \text{a.s.}
\end{aligned} \tag{4.24}
$$

Hence,

$$
\begin{aligned}
\|U_m - B^{-1}\| &\le \|U_m B^{-1}\|\,\|B - W_m\| \\
&= o\left((\log\log m)^{-1/2}\right) \quad \text{a.s.}
\end{aligned} \tag{4.25}
$$

In view of (4.19), (4.25) and Lemma 2, in order to show that (1.7)–(1.9) hold for $\hat{\theta}_n$, we need only show that (1.7)–(1.9) hold for

$$
T_{n+1} = (n-1)/nT_n - (U_n/n)(\delta_n + \varepsilon_n).
$$

This is true in view of (4.20), (4.25), Lemmas 3 and 4. To prove (4.14), we apply

(1.8) and note that

$$\|X_{mj}^{*} - X_m\| = \|X_{mj}^{**} - X_m\| = c_m = o(m^{-1/2}) \quad \text{a.s.} \qquad \square$$

REMARK 4. The choice of c_n, λ_n [see (4.8)] in Nevelson and Hasminskii (1973b) does not satisfy (4.12) and (4.13). The cost of their procedure K_n grows at a larger order than $\log n$ in (4.14). This can be seen from the following consideration. Let $f(X) = B(X - \theta)$. As shown by Nevelson and Hasminskii under (4.8),

$$U_m \to B^{-1} \quad \text{a.s.}$$

Hence, we can apply Lemma 5 with $\delta_m = 0$ to show that

$$\text{(4.26)} \qquad \lim\left(\sum_1^m X_k X_k'\right)\Big/\log m = B^{-1}\Sigma(B^{-1})'/2p.$$

We note that

$$\text{(4.27)} \qquad \sum_1^m c_k^2 \geq k_1 \sum_1^m k^{-\alpha_1} \sim k_1 m^{1-\alpha_1}/(1-\alpha_1), \qquad \alpha_1 < \tfrac{1}{2}.$$

By (4.26) and (4.27),

$$\text{(4.28)} \qquad \lim K_n\Big/\left(\sum_1^m c_k^2\right) = I \quad \text{a.s.}\ (n = 2mp).$$

In view of (4.27) and (4.28), the cost K_n of the Nevelson–Hasminskii procedure grows algebraically instead of logarithmically.

REMARK 5. The choice of c_n, λ_n in Lai and Robbins (1979) for the simple linear model

$$y_i = \beta(x_i - \theta) + \varepsilon_i$$

is $\lambda_n^{1/2} = c_n = n^{-1/2}(\log n)^{-\alpha}$, where $0 < \alpha < \frac{1}{2}$. This choice clearly satisfies (4.12) and (4.13).

Acknowledgment. This is an updated version of the last chapter of my thesis which was completed in 1980 under the supervision of Professor T. L. Lai. For his earlier guidance and long-term cooperation since then, I would like to express my gratitude to him.

REFERENCES

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

BOJANIC, R. and SENETA, E. (1973). A unified theory of regularly varying sequences. *Math. Z.* **134** 91–106.

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

DALECKII, J. L. and KREIN, M. G. (1974). *Stability of Solutions of Differential Equations in Banach Spaces*. Amer. Math. Soc., Providence, R.I.

DVORETZKY, A. (1977). Asymptotic normality for sums of dependent random vectors. In *Multivariate Analysis* (P. R. Krishnaiah, ed.) **4** 23–34. North-Holland, Amsterdam.

GOODWIN, G. C., RAMADGE, P. J. and CAINES, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19** 829–853.

KUELBS, J. (1977). Kolmogorov's law of the iterated logarithm. *Illinois J. Math.* **21** 784–800.

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

LAI, T. L. and ROBBINS, H. (1978a). Adaptive design in regression and control. *Proc. Nat. Acad. Sci. U.S.A.* **75** 586–587.

LAI, T. L. and ROBBINS, H. (1978b). Limit theorems for weighted sums and stochastic approximation processes. *Proc. Nat. Acad. Sci. U.S.A.* **75** 1068–1070.

LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196–1221.

LAI, T. L. and ROBBINS, H. (1982). Iterated least squares in multiperiod control. *Adv. Appl. Math.* **3** 50–73.

LAI, T. L. and WEI, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** 154–166.

NEVELSON, I. A. and HASMINSKII, R. Z. (1973a). *Stochastic Approximation and Recursive Estimation*. Amer. Math. Soc., Providence, R.I.

NEVELSON, M. B. and HASMINSKII, R. Z. (1973b). An adaptive Robbins–Monro procedure. *Automat. Remote Control* **34** 1594–1607.

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

RUPPERT, D. (1985). A Newton–Raphson version of the multivariate Robbins–Monro procedure. *Ann. Statist.* **13** 236–245.

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

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* **39** 135–150.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742