The Annals of Statistics
1973, Vol. 1, No. 5, 851-861

EXTENSIONS OF KESTEN'S ADAPTIVE STOCHASTIC APPROXIMATION METHOD

BY H. J. KUSHNER[1] AND T. GAVIN[2]

Brown University

Kesten proposed a method for adjusting the coefficients of a scalar stochastic approximation process, and proved w.p. 1 convergence. A family of multidimensional processes for function minimization are treated here. Each method consists of a sequence of truncated one-dimensional procedures of the Kesten type. The methods seem to offer a number of advantages over the usual Kiefer-Wolfowitz procedures, and are more natural analogs of the schemes in common use in deterministic optimization theory.

1. Introduction. Over the past twenty years the stochastic approximation method has attracted great attention in both the mathematical statistics and engineering literature (see e.g., [2], [3], [5], [7], [8], [9], [11], [12]). Its attraction to engineering systems optimization lies in the fact that it provides a systematic approach to Monte Carlo optimization, when one has a system with a performance or cost function $f(x)$ depending in a largely unknown way on a vector parameter, and only noise corrupted observations can be taken. Convergence w.p. 1 of the sequence of estimates of the optimum operating point can be guaranteed. Unfortunately, there is a serious disadvantage to the method in that the actual asymptotic rates of convergence, and the initial behavior of the sequence of estimates are very sensitive to the choice of gain sequences $\{a_n\}$ and finite difference sequence $\{c_n\}$.

Kesten (1958) investigated a procedure in which these actual gain and finite difference sequences are allowed to depend on the observed data in a certain way, and proved convergence for a Robbins–Monro and a type of one-dimentional Kiefer–Wolfowitz method. His method is a quite natural and intuitively reasonable procedure. The multidimensional version of his procedure has not been investigated. Furthermore, all the usual forms of the Kiefer–Wolfowitz method for the multidimensional problem are essentially stochastic versions of Newton's method for function minimization, where an estimate of the gradient direction is obtained and one or more search steps are taken in that direction. The most effective *deterministic methods* for hill descending, such as the conjugate gradient or Partan (see, e.g., Wilde and Beightler (1967), pages 304–338, Fletcher and Powell (1963)), do not search along gradient directions and operate something

Received June 1972; revised January 1973.

[1] This author's research was supported by AFOSR 71-2078, NSF GK 31073X, and NONR N00014-67-A-0191-0018.

[2] This author's research was supported by AFOSR 71-2078 and NONR N00014-67-A-0191-0009.

Key words and phrases. Monte-carlo, sequential analysis, adaptive process, stochastic approximation, sequential optimization with noisy observations.

as follows: A direction d_1 and initial point X_1 are selected. We search along the line through X_1 in the directions $\pm d_1$ until a reasonable approximation (denoted by X_2) to the location of the minimum along that line is obtained. A new direction d_2 is selected, etc. The methods differ by the schemes for selecting the d_n, but the gradient directions are almost never selected, since methods are available which yield faster convergence in both the initial and terminal stages of search. (See Fletcher and Powell (1963) for an empirical comparison of several methods).

We follow somewhat the same procedure here. The procedure to be investigated consists of a sequence of one-dimensional search procedures (or cycles as they will be called), each one being a truncated form of a procedure of Kesten's type. Many methods of selections will work. Let X_1 denote an initial point; select a direction d_1. Do an iterative search (a truncated procedure of Kesten's type) on the line through X_1 in directions $\pm d_1$, letting $X_1^1, \cdots, X_{L_1}^1$ denote the iterates in the first cycle, where L_1 is usually random. Then select d_2, define X_2 by $X_2 = X_{L_1}^1$, search along the line through X_2 in the direction $\pm d_2$, and generate the sequence of estimates $X_1^2, \cdots, X_{L_2}^2$ of the location of the minimum of $f(\cdot)$ along the second line; define X_3 by $X_3 = X_{L_2}^2$, etc. The procedure will be described in detail below. The exact method of selecting the $\{d_n\}$ is not important for the convergence proofs, provided that an apparently essential property ((A3) below) holds.

A further difficulty with standard Kiefer–Wolfowitz algorithms is that they require unimodality of $f(\cdot)$. Yet it occurs frequently in applications in control theory, that $f(\cdot)$ is not unimodal. In this paper we prove that the sequence of iterates for a fairly general multidimensional version of Kesten's procedure converge w.p. 1 to a set where the gradient of $f(x)$ is zero for a class of multimodal functions. In this sense, a common current practice in deterministic optimization is followed, where it is usually proved only that the algorithm yields a sequence, any convergent subsequence of which converges to a point where a necessary condition for optimality holds.

Our method of proof differs substantially from that of Kesten, and is somewhat closer in spirit to that of Venter (1967). Kesten did not actually treat a complete Kiefer–Wolfowitz procedure, even in one dimension, since his finite difference intervals were held constant. His convergence theorem is a type of random contraction theorem, a generalization of the type introduced by Dvoretsky (1956). In these proofs it is essential that there be a point about which there is a "contraction." Such methods appear to be inapplicable when there are many stationary points. Venter's method of proof is also closely connected with the uniqueness of the stationary point, and each iteration consists simply of one step in an estimated gradient direction. Owing to the possible non-uniqueness of the stationary point, the general method of selecting the directions for the one-dimensional search cycles (and their random duration), and the "adaptive" method of selecting the coefficients, a somewhat more elaborate method of

proof is required here. Rather than proving convergence directly, we examine the consequences of non-convergence, to obtain the desired contradiction.

2. Description of the process. For each $x \in R^r$, Euclidean r-space, let $H(y \mid x)$ denote a distribution function of a real-valued random variable with finite mean $f(x)$ and uniformly bounded variance $\int [y - f(x)]^2 \, dH(y \mid x) \leqq \hat{\sigma}^2 < \infty$ (for some real $\hat{\sigma}$). The sequence generated by the algorithm converges to the point or set where $f(\cdot)$ is stationary. Let $f_x(\cdot)$ and $f_{xx}(\cdot)$ denote the gradient and Hessian of $f(\cdot)$, if they exist. Next, some terms will be defined, then the method is described and discussed, and some conditions listed. The convergence theorem is proved and discussed in Section 3.

Let the sequence of random variables $X_1, \cdots, X_n, \cdots$ denote the initial points of the sequence of one-dimensional search cycles, and the random variables $d_1, \cdots, d_n, \cdots$, the sequence of directions of iteration. On the nth cycle, we search on the line through X_n in directions $\pm d_n$. Let $X_1^n, \cdots, X_i^n, \cdots$ denote a sequence of random variables generated during the nth search cycle. *Even though the cycle is stopped at a finite time, it is notationally convenient to suppose (which we will do) that X_i^n is defined for all i.*

Let $L_1, \cdots, L_n, \cdots$ denote a sequence of positive integer-valued random variables. Let $\mathscr{B}_n$ and $\mathscr{B}_i^n$ denote the smallest σ-algebras which measure

$$\{X_i^m, m = 1, \cdots, n-1; i = 0, 1, \cdots; d_1, \cdots, d_{n-1}\},$$
$$\{X_i^m, m = 1, \cdots, n-1; i = 0, 1, \cdots; d_1, \cdots, d_n\} \quad \text{and}$$
$$\{X_i^m, m = 1, \cdots, n-1; i = 0, 1, \cdots; d_1, \cdots, d_n; X_0^n, X_1^n, \cdots, X_i^n\},$$

respectively.

That is, $\mathscr{B}_n$ measures all the data up to and including the start of the nth cycle, but not d_n, and $\mathscr{B}_i^n$ measures, in addition, d_n and the first i iterates of the nth cycle. Note that X_n is $\mathscr{B}_n$ measurable.

When we say that a random time M is non-anticipative with respect to the $\{X_n, n = 1, \cdots\}$ or $\{X_i^n, i = 0, \cdots\}$ sequences, we mean the usual, that the ω set $\{M = m\} \in \mathscr{B}_m$ or $\{M = j\} \in \mathscr{B}_j^n$, resp. (i.e., whether or not $M = m$ or $M = j$ can be determined by watching the X_i only up to time m, or watching the X_i^n only up to the first j iterates of the nth cycle, resp.) If τ is a non-anticipative integer-valued random variable we define (in the usual way) $\mathscr{B}_\tau$ as the collection of sets A satisfying, for each n, $\{\tau = n\} \cap A \in \mathscr{B}_n$, and similarly for $\mathscr{B}_i^\tau$ and $\bar{\mathscr{B}}_\tau$. *All the random times used in the sequel will be non-anticipative, or assumed so, without explicit mention.* If $\tau < \infty$ only on a set C, with $P(C) < 1$, then all subsets of $\Omega - C$ are $\mathscr{B}_\tau$, by the above definition.

Let $\{b_i^n, e_i^n\}$ denote real-valued random sequences which are non-anticipative with respect to the $\{X_i^n\}$ (i.e., b_i^n is $\mathscr{B}_i^n$ measurable). Suppose that X_i^n is defined. Define X_{i+1}^n by

$$X_{i+1}^n = X_i^n - d_n b_i^n [Y_{2i+1}^n - Y_{2i}^n]/2e_i^n, \tag{2.1}$$

where Y_{2i+1}^n and Y_{2i}^n are drawn from $H(y \mid x)$ with parameters $X_i^n + d_n e_i^n$ and

$X_i^n - d_n e_i^n$, resp. Thus Y_{2i+1}^n and Y_{2i}^n are noise-corrupted observations on the system at parameter points $X_i^n \pm d_n e_i^n$, resp.

Define

$$DY(X_i^n, e_i^n, d_n) \equiv [Y_{2i+1}^n - Y_{2i}^n]/2e_i^n ,$$
$$\mathrm{Df}(X_i^n, e_i^n, d_n) \equiv [f(X_i^n + d_n e_i^n) - f(X_i^n - d_n e_i^n)]/2e_i^n ,$$
$$\xi_i^n \equiv DY(X_i^n, e_i^n, d_n) - \mathrm{Df}(X_i^n, e_i^n, d_n) .$$

Then (2.1) can be written as

$$X_{i+1}^n = X_i^n - d_n b_i^n(\mathrm{Df}(X_i^n, e_i^n, d_n) + \xi_i^n) . \tag{2.2}$$

Next, we describe the method for determining the b_i^n, e_i^n. The method to be described is only one of a large family of interesting possible methods to which Theorem 1 can be adapted.

Let $N_1, N_2, \cdots, N_n$, denote a sequence of integers. (With a slight bit of extra complication, we could let N_n be a finite valued random variable.) Let $\{a_i^n, c_i^n\}$ denote sequences of positive real numbers satisfying

$$\sum_{n=1}^{\infty} \sum_{i=0}^{N_n-1} a_i^n = \infty , \qquad \sum_{n=1}^{\infty} \sum_{i=0}^{N_n-1} (a_i^n/c_i^n)^2 < \infty ,$$
$$\sum_{n=1}^{\infty} \sum_{i=0}^{N_n-1} a_i^n c_i^n < \infty , \qquad c_i^n \to 0 , \quad a_i^n \to 0 , \quad a_i^n/(c_i^n)^2 \to 0 \tag{2.3}$$
$$\text{as } n + i \to \infty .$$

The following version of Kesten's method will be used. Let $b_1^n = b_0^n = a_0^n$, $e_1^n = e_0^n = c_0^n$. The sequences b_i^n, e_i^n remain fixed as i increases as long as the X_i^n sequence is monotonic as i increases. The b_i^n, e_i^n are changed only when the X_i^n sequence oscillates in the direction d_n. In particular, suppose that, for some $i \geqq j$, $b_i^n = a_j^n$, $e_i^n = c_j^n$. If $i \geqq 1$ and both

$$d_n'(X_{i+1}^n - X_i^n) , \qquad d_n'(X_i^n - X_{i-1}^n)$$

have the same sign, or one or both are zero, then set $b_{i+1}^n = b_i^n = a_j^n$, $e_{i+1}^n = e_i^n = c_j^n$. If, however, neither

$$d_n'(X_{i+1}^n - X_i^n) , \qquad d_n'(X_i^n - X_{i-1}^n)$$

are zero and they are of opposite sign, then set $b_{i+1}^n = a_{j+1}^n$, $e_{i+1}^n = c_{j+1}^n$. At the N_nth change in coefficient, the cycle terminates. That is, let Q_n denote the last value of j for which b_j^n equals $a_{N_n-1}^n$. Then $X_{Q_n+1}^n$ will be calculated, and we define $X_{Q_n+1}^n = X_{n+1} \equiv X_0^{n+1}$; we then select d_{n+1}, and continue.

We require the following additional conditions.

(A1) $f(\cdot)$ is continuous, together with its first and second derivatives, it is bounded from below by a real number B, and there is a real K_0 for which $|y' f_{xx}(x) y| \leqq K_0 |y|^2$, for any vectors y, x, where the Euclidean norm is used.

(A2) Let δ and δ_1 denote arbitrary positive numbers. For arbitrary scalar c, and direction vector d, define

$$\Delta_\delta(c, d) \equiv \{X: |\mathrm{Df}(X, c, d)| \leqq \delta\} .$$

Suppose that

$$P_{\mathscr{B}_i^n}\{DY(X_i^n, e_i^n, d_n) > 0\} \geqq \delta_1,$$
$$P_{\mathscr{B}_i^n}\{DY(X_i^n, e_i^n, d_n) < 0\} \geqq \delta_1,$$

on the set where $X_i^n \in \Delta_\delta(e_i^n, d_n)$.

In other words, if the finite difference estimate $\mathrm{Df}(X_i^n, e_i^n, d_n)$ of the directional derivative is small enough, then there is a minimum nonzero probability that $DY(X_i^n, e_i^n, d_n)$ will be positive, and that it will be negative. Noise need only play a role, where the "slopes" are small, a reasonable assumption. The proof remains valid if (A2) holds only for $n \geqq v$, some random time. When the slope or the Df is large, the noise may not be sufficiently large to allow the possibility of an oscillation or change in value of b_j^n, e_j^n. The conditions should reflect this fact. It is sensible, from an applications point of view, to suppose that the noise plays a role (concerning whether or not the coefficients are changed) where the slopes or the Df are small, but may not play a role where the slopes or Df are large.

Let D_0 denote the set $\{x: f_x(x) = 0\}$. Let $N(\cdot)$ denote a nonnegative continuous real-valued function on $R^r - D_0$.

(A3) For some positive real numbers γ_1, γ_2, let

$$P_{\mathscr{B}_n}\{|d_n' f_x(X_n)| \geqq \gamma_1 |f_x(X_n)|\} \geqq \gamma_2$$

w.p. 1, on the set where $f_x(X_n) \neq 0$ and $n \geqq N(f_x(X_n))$.

Condition (A3) is the only condition to be placed on the choice of the $\{d_n\}$. It is rather weak, and intuitively, says that the direction d_n must not be almost orthogonal to the gradient too often.

For any vectors v_1, v_2, let $\theta[v_1, v_2]$ denote the angle between N_1 and N_2. Then (A3) is equivalent to

$$(*) \qquad P_{\mathscr{B}_n}\{|\cos\theta[d_n, f_x(X_n)]| \geqq \gamma_1\} \geqq \gamma_2,$$

from which it is obvious that if the r-dimensional vector $\tilde{d}_n$ is selected at random from a distribution $Q(\cdot)$ which is not concentrated on a hyperplane in R^r and we define $d_n = \tilde{d}_n/|\tilde{d}_n|$, then (A3) holds.

Under broad conditions, d_n can also be chosen to be an estimate of the gradient direction. Let u_i denote the unit vector in the ith coordinate direction, and $\{k_n\}$ a sequence of non-anticipative (with respect to the $\mathscr{B}_n$) scalar-valued random variables, which tend to 0 as $n \to \infty$. Then, with finite difference interval k_n, we can obtain an estimate of the derivative of $f(x)$ in direction u_i, at $x = X_n$, of the form

$$(**) \qquad \frac{\psi_{n,i}}{k_n} + \frac{[f(X_n + u_n k_n) - f(X_n - u_n k_n)]}{2k_n} = \frac{\psi_{n,i}}{k_n} + \mathrm{Df}(X_n, k_n, u_i) + k_n \tilde{B}_n,$$

where $|\tilde{B}_n| \leqq K_0/2$, $E_{\mathscr{B}_n}\psi_{n,i} = 0$, $E_{\mathscr{B}_n}|\psi_{n,i}|^2 \leqq \sigma^2$ for some real σ. One of the reasons for the introduction of the function $N(\cdot)$ in (A3) is to allow us to ignore the $k_n \tilde{B}_n$ term in (**). That is, if the direction vector obtained from the vector $\tilde{d}_n \equiv \{(\psi_{n,i}/k_n) + \mathrm{Df}(X_n, k_n, u_i), i = 1, \cdots, r\}$ satisfies (*), then (A3) holds for

some suitable function $N(\cdot)$. As $n \to \infty$, the $\phi_{n,i}/k_n$ terms dominate the $\mathrm{Df}(X_n, k_n, u_i)$, and (*) will not hold if the $\phi_n \equiv (\phi_{n,1}, \cdots, \phi_{n,r})$ tend (in probability, conditioned on $\mathscr{B}_n$) to the hyperplane orthogonal to the $f_x(X_n)$, as $n \to \infty$ (unless $E|\phi_n|^2/k_n{}^2 \to 0$). However, if, given $\mathscr{B}_n$, the $\phi_{n,1}, \cdots, \phi_{n,r}$ are independent and satisfy, for real positive σ_i, $0 < \sigma_1{}^2 \leqq E|\phi_{n,i}|^2 \leqq \sigma_2{}^2$, for all n, i, then (*) holds.

(A3) holds for many other schemes. In particular, it holds if we select a random unit vector $\hat{d}_n$ in any way at all, and we define $\{\phi_n\}$ to be a sequence of independent identically distributed random vectors with zero mean whose distribution $Q(\cdot)$ is not confined to a hyperplane, and where ϕ_n is independent of $\mathscr{B}_n$ and $\hat{d}_n$, and define d_n by $d_n = (\hat{d}_n + \phi_n)/|\hat{d}_n + \phi_n|$. To see this, merely note that (with an obvious abuse of notation)

$$P_{\mathscr{B}_n, d_n}\{|\cos\theta[(\hat{d}_n + \phi_n), v_n]| \geqq \gamma_1\} \geqq \gamma_2$$

for some real positive γ_1 and γ_2 and for any $\mathscr{B}_n$ measurable vector v_n. Thus a slight random perturbation of any direction determining method will satisfy (A3).

Although we will not pursue the point in detail, the proof can be modified (at some increase in notational complexity) to yield convergence even if (A3) does not hold on every cycle. The required modifications in the conditions are that (loosely speaking) the ratio of the sum of the $a_i{}^n$, $n \leqq N$, over the cycles for which (A3) does not hold to the sum of $a_i{}^n$, $n \leqq N$, over these cycles for which (A3) does hold, does not tend to infinity, as $N \to \infty$. Thus, assuming the above assertion, we can select the directions d_n by the method: for some integer s, let $d_1, \cdots, d_{s-1}$ be unit vectors in the appropriate estimated gradient directions, let d_s be selected by an (arbitrary) function of $X_1, \cdots, X_{s-1}, d_1, \cdots, d_{s-1}$ (as in some of the deterministic schemes in Wilde and Beightler [13], Chapter 7). Then repeat the method for $d_{s+1}, \cdots, d_{2s}$, etc.

We also require

(A4) There is a real σ for which, for all n, i, $E_{\mathscr{B}_i{}^n}\xi_i{}^n = 0$, $E_{\mathscr{B}_i{}^n}|\xi_i{}^n|^2 \leqq \sigma^2/(e_i{}^n)^2$ w.p. 1.

3. The main theorem.

THEOREM. *Under* (2–3) *and* (A1)—(A4), $X_n \to D_0 \cup \{\infty\}$ w.p. 1 *and* $X_i{}^n \to D_0 \cup \{\infty\}$ w.p. 1.

Remarks and Outline of Proof. The proof will be divided into several parts. First, it will be proved that each $a_i{}^n$ is used only finitely often w.p. 1. Then, in Part 2, an upper bound to the average number of iterations (the number of j's) that both $b_j{}^n = a_i{}^n$ *and* $X_j{}^n \in \Delta_\delta(c_i{}^n, d_n)$ hold simultaneously is obtained. In Part 3, it is proved that $f(X_i{}^n)$ converges w.p. 1 as $n + i \to \infty$, and a very useful representation of the difference $E_{\mathscr{B}_n} f(X_{n+1}) - f(X_n)$ is obtained. The actual convergence result is completed in Part 4. It will be supposed that $\{X_i{}^n\}$ is bounded w.p. 1, for otherwise, the proof implies that $|X_i{}^n| \to \infty$ as $n + i \to \infty$, for any path for which a subsequence of the $\{X_i{}^n\}$ are unbounded. The convergence to ∞, if any, is in the topology of the one point compactification. Thus either

$X_i^n \to D_0$ as $n + i \to \infty$, or for any positive number r, $\limsup |X_i^n| > r$ (w.p. 1).

Conditions guaranteeing $\limsup |X_i^n| < \infty$ w.p. 1 can be introduced (i.e., $f(x) \to \infty$ as $|x| \to \infty$ or $\lim_{|x|\to\infty} \inf |f_x(x)| > 0$, and similar conditions, or the existence of only one finite point at which $f_x(x) = 0$) but it does not seem worth the effort here. The methods of proof suggest a variety of other schemes for adjusting the b_i^n, e_i^n coefficients. For example, we could switch after Q_i^m oscillations, where $\{Q_i^n\}$ is a bounded sequence. Various more complicated methods will also work.

PROOF OF THEOREM.

Part 1. For each i and n, let M_i^n denote the number of j's for which $b_j^n = a_i^n$. Then $M_i^n < \infty$ w.p. 1, i.e., $b_i^n \to 0$ as $n + i \to \infty$.

PROOF. Let $M_i^n = \infty$ for $\omega \in A$. There is a random variable X so that, for almost all $\omega \in A$, $X_j^n \to X$ as $j \to \infty$. Thus, for almost all ω in A, $\mathrm{Df}(X_j^n, a_i^n, d_n) \to \mathrm{Df}(X, a_i^n, d_n)$, and $X_{j+1}^n - X_j^n \to 0$. Hence

$$\mathrm{Df}(X, a_i^n, d_n) + \xi_j^n \to 0\,, \tag{3.1}$$

which implies that $\mathrm{Df}(X, a_i^n, d_n) = 0$. Thus $|\mathrm{Df}(X_j^n, a_i^n, d_n)| \leqq \delta/2$ infinitely often for almost all $\omega \in A$. This fact, together with (A2), implies that A is a null set.

Part 2. For arbitrary but fixed n and i, let $\alpha_1, \cdots, \alpha_\tau$ denote the sequence of values of j for which both $b_j^n = a_i^n$ and $X_j^n \in \Delta_\delta(c_i^n, d_n)$ hold. It follows from (A2) that, w.p. 1,

$$E_{\mathscr{B}_{\alpha_1}^n} \tau = \textstyle\sum_{s=0}^{\infty} P_{\mathscr{B}_{\alpha_1}^n}\{\tau > s\} \leqq 1 + 2/\delta_1\,. \tag{3.2}$$

Part 3. Next, it will be shown that there is a finite valued random time ρ_1 and a sequence of nonnegative random variables $\{\beta_n\}$ satisfying (3.3) and (3.4):

$$E_{\mathscr{B}_{\rho_1}} \textstyle\sum_{n \geqq \rho_1} \beta_n < \infty\,, \tag{3.3}$$

$$E_{\mathscr{B}_m} f(X_{m+1}) - f(X_m) \leqq \beta_m\,, \tag{3.4}$$

w.p. 1, for any random time $m \geqq \rho_1$. *Observe that* (3.3) *and* (3.4) *imply that* $f(X_m)$ *converges* w.p. 1 (by application of the supermartingale convergence theorem, see Kushner (1966)). A similar proof yields that $f(X_i^n)$ converges w.p. 1 as $n + i \to \infty$, but we omit the details.

Define B_i^n by $\mathrm{Df}(X_i^n, e_i^n, d_n) = d_n' f_x(X_i^n) + e_i^n B_i^n$. By (A1), $|B_i^n| \leqq K_0/2$. Equation (2.2), and a truncated Taylor series expansion yield

$$\begin{aligned} E_{\mathscr{B}_{i^n}} f(X_{i+1}^n) - f(X_i^n) \leqq\; & -E_{\mathscr{B}_{i^n}}[b_i^n (d_n' f_x(X_i^n))(\mathrm{Df}(X_i^n, e_i^n, d_n) + \xi_i^n)] \\ & + \frac{(b_i^n)^2}{2} E_{\mathscr{B}_{i^n}} K_0 (\mathrm{Df}(X_i^n, e_i^n, d_n) + \xi_i^n)^2\,. \end{aligned}$$

Using $|d_n' f_x(X_i^n)| \leqq 1 + |d_n' f_x(X_i^n)^2$ and $|\mathrm{Df}(X_i^n, e_i^n, d_n)|^2 \leqq 2|d_n' f_x(X_i^n)|^2 + 2K_0^2 (e_i^n)^2$, we obtain

$$E_{\mathscr{B}_{i^n}} f(X_{i+1}^n) - f(X_i^n) \leqq -b_i^n |d_n' f_x(X_i^n)|^2 q_i^n + u_i^n\,, \tag{3.5}$$

where $q_i^n = (1 - K_0 e_i^n - 2K_0 b_i^n)$ and $u_i^n = K_0[e_i^n + \sigma^2 b_i^n/(e_i^n)^2 + 2K_0^2 b_i^n (e_i^n)^2] b_i^n$. By Part 1 (since $e_j^n \to 0$), there is a random time $\rho_1 < \infty$ w.p. 1, so that for any $n \geqq \rho_1$, $|\mathrm{Df}(X_j^n, e_j^n, d_n)| \leqq \delta$ if $|d_n' f_x(X_j^n)| \leqq \delta/2$, and also $q_i^n \geqq \frac{1}{2}$ and $u_i^n \leqq b_i^n(\delta^2/16)$. Let α_i^n denote the first value of j for which $b_j^n = a_i^n$. Then for any random time $n \geqq \rho_0$, the expectation (given $\mathscr{B}_{\alpha_i^n}^n$) of the number of uses of a_i^n when $|d_n' f_x(X_j^n)| \leqq \delta/2$, is bounded by $(1 + 2/\delta_1)$ w.p. 1, by Part 2.

For any set A, let I_A denote the indicator of A. For $n \geqq \rho_1$, (3.5) can be bounded above by

$$(3.6) \qquad -b_i^n(\delta^2/16) I_{\{|d_n' f_x(X_i^n)| \geqq \delta/2\}} - \tfrac{1}{2} b_i^n I_{\{|d_n' f_x(X_i^n)| < \delta/2\}} |d_n' f_x(X_i^n)|^2 + u_i^n I_{\{|d_n' f_x(X_i^n)| < \delta/2\}} .$$

Define the last term on the right of (3.6) *as* β_i^n, and define $\beta_n = E_{\mathscr{B}_n} \sum_i \beta_i^n$. Then (3.3) and (3.4) hold by (2.3) and Part 2.

Observe that (3.6) and the lower bound B on $f(\cdot)$, imply that for any sequence of random times $m_i \to \infty$ w.p. 1, where $m_i \geqq \rho_1$,

$$\lim_i E_{\mathscr{B}_{m_i}} \sum_{n \geqq m_i} \sum_j b_j^n I_{\{|d_n' f_x(X_j^n)| \geqq \delta/2\}} = 0 .$$

Part 4. Let D, D' denote any open sets satisfying $D' \supset D$ and $\sup_{x,y \in D'} |x - y| \leqq \delta/(8K_0)$ and for some point $\bar{x}$ in D with $f_x(\bar{x}) \neq 0$, and for $x \in D'$, $2|f_x(\bar{x})| \geqq |f_x(x)| \geqq \frac{1}{2}|f_x(\bar{x})|$, and

$$\inf_{x \in D; y \notin D'} |x - y| \equiv d_0 > 0 .$$

For each such D', there is a positive real number $\bar{n}$ so that $\sup_{x \in D'} N(f_x(x)) = \bar{n}$. Define the sequence of random times t_i, t_i^+ (if t_i (or t_i^+) is not defined at some ω, set $t_i = \infty$ (or $t_i^+ = \infty$) there)

$$\begin{aligned} t_1 &= \min \{r\colon X_r = X_0^r \in D,\ r \geqq \rho_1,\ r \geqq \bar{n}\} \\ t_1^+ &= \min \{r\colon X_i^r \notin D',\ \text{for some } i \geqq 0 \text{ and } r \geqq t_1\} \\ t_n &= \min \{r\colon X_r = X_0^r \in D,\ r > t_{n-1}^+\} \\ t_n^+ &= \min \{r\colon X_i^r \notin D',\ \text{for some } i \geqq 0 \text{ and } r \geqq t_n\} . \end{aligned}$$

If we can show that $t_n < \infty$ only finitely often w.p. 1, and that if ever in D, the sequence $\{X_i^n\}$ must eventually leave D' w.p. 1, the theorem will be proved, since for any compact set $\tilde{D}$ not containing any points of D_0, there are a finite number of pairs (D, D') satisfying our conditions, for which the collection of D covers $\tilde{D}$. A slight variation on the following proof yields a similar statement for the $\{X_i^n\}$.

The nth *re-entry* cycle starts with the iterate $X_0^{t_n}$ and ends when we obtain the first subsequent iterate $X_i^{t_n^+}$ which is not in D'.

If $X_j^k \in D'$ for all iterates from $X_0^{t_n}$ up until at least $X_j^{t_n+m}$, then set $J_j^{n,m} = 1$, and zero otherwise. Define $J^{n,m} = 1$, if $X_0^{t_n+s} = X_{t_n+s} \in D'$ for $s = 0, \cdots, m$, and zero otherwise.

Part 4 is divided into three subparts. In the first subpart, it is proved that only the $b_i^{t_n+m} d'_{t_n+m} f_x(X_i^{t_n+m})$ component of $\mathrm{Df}(X_i^{t_n+m}, e_i^{t_n+m}, d_{t_n+m}) + \xi_i^{t_n+m}$ plays a

role, for large n, in the movements of the iterates out of the set D'. In the second subpart, the above fact is used to obtain (3.14), a lower bound to sums of some of the $a_i^{t_n+m}$. In the third subpart, the latter result is used to obtain an estimate of the average change in the function $f(\cdot)$ during a re-entry cycle, which yields the desired conclusions.

In (3.7) *et seq.*, the sums are empty if $t_1 = \infty$, but it may avoid some confusion to carry the $I_{\{t_1<\infty\}}$ term. From Part 3, w.p. 1

$$E_{\mathscr{B}_{t_1}} I_{\{t_1<\infty\}} \sum_{m\geqq t_1} \sum_i (e_i^m b_i^m + (b_i^m/e_i^m)^2) I_{\{|d_m' f_x(X_i^m)|<\delta/2\}} < \infty \tag{3.7}$$

$$E_{\mathscr{B}_{t_1}} I_{\{t_1<\infty\}} \sum_{m\geqq t_1} \sum_i b_i^m I_{\{|d_m' f_x(X_i^m)|\geqq\delta/2\}} < \infty . \tag{3.8}$$

Thus as $n \to \infty$, w.p. 1 (by (3.7), (3.8), and the fact that $e_i^m \to 0$, $b_i^m/(e_i^m)^2 \to 0$)

$$E_{\mathscr{B}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq t_n} \sum_i (e_i^m b_i^m + (b_i^m/e_i^m)^2) \to 0 . \tag{3.9}$$

Now we estimate the effects of the terms in

$$\begin{aligned} X_{i+1}^{t_n+m} = X_i^{t_n+m} &- b_i^{t_n+m} d'_{t_n+m} f_x(X_i^{t_n+m}) \\ &- b_i^{t_n+m} e_i^{t_n+m} B_i^{t_n+m} - b_i^{t_n+m} \xi_i^{t_n+m} . \end{aligned} \tag{3.10}$$

In the *indefinite sum* $\sum_{m,i} b_i^{t_n+m} \xi_i^{t_n+m}$ it is supposed that the terms are summed in precisely the order in which they are obtained by the iteration (2.2). Thus, the sup over the indefinite sum makes sense. The above indefinite sum is a martingale. Using (Doob (1953), Theorem 3.4, page 317), Chebyshev's inequality and (3.9) yields (3.11) w.p. 1, as $n \to \infty$.

$$P_{\mathscr{B}_{t_n}}\{\sup |\textstyle\sum_{m,i} (b_i^{t_n+m} \xi_i^{t_n+m} + b_i^{t_n+m} e_i^{t_n+m} B_i^{t_n+m})| I_{\{t_n<\infty\}} \geqq d_0/2\} \to 0 . \tag{3.11}$$

Assume for the moment that the re-entry cycles have finite duration w.p. 1. Then (3.11) implies that the sums of the terms (second term on r.h.s. of (3.10)) $b_i^{t_n+m} d'_{t_n+m} f_x(X_i^{t_n+m})$ must pull the iterates out of D' for large n. Majorizing the sum of these terms over the re-entry cycle yields that (3.12) holds with a probability (conditioned on $\mathscr{B}_{t_n}$) that approaches 1 as $n \to \infty$.

$$\sum_{m\geqq 0} \sum_i b_i^{t_n+m} J_i^{n,m} |d'_{t_n+m} f_x(X_i^{t_n+m})| I_{\{t_n<\infty\}} \geqq \frac{d_0}{2} I_{\{t_n<\infty\}} . \tag{3.12}$$

The bound $|d'_{t_n+m} f_x(X_i^{t_n+m})| \leqq 2|f_x(\bar{x})|$, for $X_i^{t_n+m} \in D'$, yields

$$\liminf_n E_{\mathscr{B}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i b_i^{t_n+m} J_i^{n,m} \geqq \liminf_n \frac{I_{\{t_n<\infty\}} d_0}{4|f_x(\bar{x})|} \equiv T .$$

Thus (w.p. 1) $T > 0$ if and only if $t_n < \infty$ infinitely often. By (3.8) we can write

$$T \leqq \liminf_n E_{\mathscr{B}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i b_i^{t_n+m} J_i^{n,m} I_{\{|d'_{t_n+m} f_x(X_i^{t_n+m})|\leqq\delta/2\}} . \tag{3.13}$$

Since $J^{n,m} \geqq J_i^{n,m}$ and the average (given $\mathscr{B}_n$) number of times that a_i^n is used *while* $|d_n' f_x(X_j^n)| \leqq \delta/2$ holds is bounded by $1 + 2/\delta_1$, we have

$$T \leqq \liminf_n E_{\mathscr{B}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i a_i^{t_n+m} J^{n,m} (1 + 2/\delta_1) . \tag{3.14}$$

Next, calculate the asymptotic conditional average change in $f(\cdot)$ per re-entry cycle, namely

$$G \equiv \liminf_n E_{\mathscr{F}_{t_n}} I_{\{t_n<\infty\}} f(\tilde{X}^{t_n}) - I_{\{t_n<\infty\}} f(X_{t_n}) ,$$

where we let $\tilde{X}^{t_n}$ denote the first iterate after X_{t_n} which is not in D'. By Part 3,

$$G \leqq -\left(\frac{\delta^2}{16}\right) \liminf_n E_{\mathscr{F}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i b_i^{t_n+m} J_i^{n,m} I_{\{|d'_{t_n+m} f_x(X_i^{t_n+m})|\geqq \delta/2\}} + \limsup_n E_{\mathscr{F}_{t_n}} \sum_{m\geqq t_n} \beta_m , \tag{3.15}$$

where the last term on the right is zero w.p. 1.

Since $\sup_{x,y\in D'} |x-y| \leqq \delta/(8K_0)$,

$$I_{\{|d'_{t_n+m} f_x(X_i^{t_n+m})|\geqq \delta/2\}} J_i^{n,m} \geqq I_{\{|d'_{t_n+m} f_x(X_{t_n+m})|\geqq \delta\}} J_i^{n,m} .$$

Thus, using (3.8),

$$\begin{aligned} G &\leqq -\left(\frac{\delta^2}{16}\right) \liminf_n E_{\mathscr{F}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i b_i^{t_n+m} J_i^{n,m} I_{\{|d'_{t_n+m} f_x(X_{t_n+m})|\geqq \delta\}} \\ &\leqq -\left(\frac{\delta^2}{16}\right) \liminf_n E_{\mathscr{F}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i a_i^{t_n+m} J^{n,m} I_{\{|d'_{t_n+m} f_x(X_{t_n+m})|\geqq \delta\}} . \end{aligned} \tag{3.16}$$

It can be supposed without loss of generality that γ_1 (see condition (A3)) is sufficiently small so that $\delta \geqq \gamma_1 |f_x(X_{t_n+m})|$ for $X_{t_n+m} \in D'$. Using this in (3.16) and taking suitable conditional expectations using (A3), yields

$$G \leqq -\gamma_2 \left(\frac{\delta^2}{16}\right) \liminf_n E_{\mathscr{F}_{t_n}} I_{\{t_n<\infty\}} \sum_{m\geqq 0} \sum_i a_i^{t_n+m} J^{n,m} . \tag{3.17}$$

Observe that if we had used the same type of upper bounding procedure which took us from (3.15) to (3.17), but for fixed n. Then $\sum_n \sum_i a_i^n = \infty$ yields $E_{\mathscr{F}_{t_n}} f(X_{t_n+m}) - f(X_{t_n}) \to -\infty$ on the set $\{t_n < \infty\} \cap \{t_n^+ = \infty\}$, modulo a null set. Thus, as asserted earlier, each re-entry cycle has a finite duration w.p. 1.

Next, (3.14) and (3.17) yied that $G < 0$ on the set A where $t_n < \infty$ infinitely often. Since this contradicts the convergence of $f(X_i^n)$, unless $P(A) = 0$, the proof is concluded.

REFERENCES

[1] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[2] DVORETSKY, A. (1956). On stochastic approximation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 39-55. Univ. of California Press.

[3] ELLIOT, D. F. and SWORDER, D. D. (1970). Applications of a simplified multidimensional stochastic approximation algorithm. *IEEE Trans. Automatic Control AC*-15, 101-104.

[4] FLETCHER, R. and POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Comput. J.* **6** 162-168.

[5] GRAY, K. B. (1964). Applications of stochastic approximation to the optimization of random circuits. *Symp. Appl. Math.* **16** 172-192.

[6] KESTEN, H. (1958). Accelerated stochastic approximation. *Ann. Math. Statist.* **29** 41-59.

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

[8] KUSHNER, H. J. (1963a). A simple iterative procedure for the identification of unknown parameters of a linear time varying system. *Trans. ASME J. Basic Eng. Ser. D* **85** 227–235.

[9] KUSHNER, H. J. (1963b) Adaptive techniques for the optimization of binary detection systems. Convention Record, IEEE International Convention, New York.

[10] KUSHNER, H. J. (1966). A note on the maximum sample excursions of stochastic approximation processes. *Ann. Math. Statist.* **37** 513–515.

[11] SAKRISON, D. S. (1964). A continuous Kiefer–Wolfowitz procedure for random processes. *Ann. Math. Statist.* **35** 590–599.

[12] TSYPKIN, YA. Z. (1966). Adaptation training and self-organization in automatic systems. *Automat. Remote Control* **27** 16–51.

[13] WILDE, D. and BEIGHTLER, C. (1967). *Foundations of Optimization.* Prentice-Hall, Englewood Cliffs.

[14] VENTER, J. H. (1967). On convergence of the Keifer–Wolfowitz approximation procedure. *Ann. Math. Statist.* **38** 1031–1036.

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, R. I. 02912