The Annals of Statistics
1974, Vol. 2, No. 4, 713-723

STOCHASTIC APPROXIMATION ALGORITHMS FOR CONSTRAINED OPTIMIZATION PROBLEMS[1]

BY HAROLD J. KUSHNER

Brown University

The paper gives convergence theorems for several sequential Monte-Carlo or stochastic approximation algorithms for finding a local minimum of a function $f(\cdot)$ on a set C defined by $C = \{x: q^i(x) \leqq 0, i = 1, 2, \cdots, s\}$. $f(\cdot)$ is unknown, but "noise perturbed" values can be observed at any desired parameter $x \in C$. The algorithms generate a sequence of random variables $\{X_n\}$ such that (for a.a. ω) any convergent subsequence of $\{X_n(\omega)\}$ converges to a point where a certain necessary condition for constrained optimality holds. The techniques are drawn from both stochastic approximation, and non-linear programming.

1. Introduction. For each $x \in R^r$, Euclidean r-space, let $H(y \mid x)$ denote the distribution function of a real-valued random variable Y with mean $f(x) \equiv \int y \, dH(y \mid x)$ and variance bounded by a real number $\hat{\sigma}^2 < \infty$; i.e., $\int [y - f(x)]^2 \, dH(y \mid x) \leqq \hat{\sigma}^2$. For some integer s, let $q^i(\cdot)$, $i = 1, \cdots, s$, denote real-valued functions on R^r and define the set $C \equiv \{x: q^i(x) \leqq 0, i = 1, \cdots, s\}$. Loosely speaking, the problem of the paper is the development of an iterative method for finding an element $\theta \in C$ at which $f(x)$ is a *local minimum* in C; $f(x)$ is unknown, but for any fixed x one or more random variables $Y(x)$ with distribution $H(y \mid x)$ can be obtained. If $\tilde{X}_0, \cdots, \tilde{X}_n$ are the first (not necessarily distinct) $n + 1$ parameter values at which draws from $H(y \mid x)$ are made, write the corresponding random variables as $Y_0(\tilde{X}_0), \cdots, Y_n(\tilde{X}_n)$, and suppose that (w.p. 1)

$$E[Y_n(\tilde{X}_n) \mid \tilde{X}_i, Y_i(\tilde{X}_i), i = 0, \cdots, n-1, \text{ and } \tilde{X}_n] = f(\tilde{X}_n)$$

$$E[(Y_n(\tilde{X}_n) - f(\tilde{X}_n))^2 \mid \tilde{X}_i, Y_i(\tilde{X}_i), i = 0, \cdots, n-1, \text{ and } \tilde{X}_n] \leqq \hat{\sigma}^2 .$$

The aim of the paper is the development of a structure for stochastic optimization algorithms (of the Monte Carlo or stochastic approximation type) which is analogous to that used in non-linear programming. The developed structure is quite versatile, and seems to consider the elements of the problem in a very natural manner from both the theoretical and practical viewpoints.

EXAMPLE 1. Let x denote a vector of parameters of a drug (say the levels of several crucial ingredients) and $\tilde{q}(x)$ the known cost of producing the drug.

Received January 1972; revised September 1973.

[1] This research was supported by grants from the Air Force Office of Scientific Research (Grant No. AF-AFOSR 71-2078), the National Science Foundation (Grant No. GK 31073X), and the Office of Naval Research (Grant No. NONR N00014-67-A-0191-0018).

AMS 1970 *classifications*. 62-45, 90-58, 93-60, 93-70.

Key words and phrases. Sequential Monte Carlo, constrained optimization, constrained stochastic approximation.

Suppose that, at any level x, the drug is either "effective" or "ineffective"; that is, for each fixed x we can only perform an experiment whose outcome $Y(x)$ is 1 if the drug is effective and 0 if it is ineffective. Define $f(x) = -P_x\{Y(x) = 1\}$. Find the value of x which *minimizes* $f(x)$ under the constraint that the cost $\tilde{q}(x)$ is no greater than α; i.e., with constraint $q(\cdot)$ where $q(x) \equiv \tilde{q}(x) - \alpha \leqq 0$. The form of $f(\cdot)$ may not be known, but the value $\tilde{q}(x)$ is assumed to be calculable for any x.

EXAMPLE 2. Let $Z_{n+1} = F(Z_n, u_n, \phi_n)$, $n = 0, 1, \cdots, N-1$, represent the dynamics of a control system with initial condition Z_0, random disturbances $\{\phi_n\}$, and *open loop controls* $\{u_n\}$ (the u_n are vectors with real components, not random variables). For some suitable real-valued functions $g_i(\cdot, \cdot)$, $i = 0, \cdots, N-1$, minimize

$$f(u_0, \cdots, u_{N-1}, Z_0) \equiv E \sum_{i=0}^{N-1} g_i(Z_i, u_i) \tag{1}$$

over the u_i and Z_0, subject to the constraint that the total "fuel" consumed, $\sum_{i=0}^{N-1} |u_i|$, not exceed a given value. For any fixed Z_0 and control sequence $\{u_n\}$, we assume that samples of $\sum_{i=0}^{N-1} g_i(Z_i, u_i)$ can be observed, and that the *exact* amount of fuel consumed can be calculated, but the average cost (1) is usually extremely difficult to calculate.

The paper is concerned with the case where $f(x)$ may have many stationary points. The methods to be discussed yield a sequence of iterates $X_0, X_1, \cdots, X_n, \cdots$ which converge to a point or to a set in C where certain necessary conditions for the *constrained optimality* are satisfied. In this sense, the result is analogous to those usually obtained in deterministic constrained optimization theory, where it is usually proved that an algorithm generates a sequence of points, any convergent subsequence of which converges to a point where a *necessary* condition for optimality holds. The type of constrained stochastic problem discussed here arises frequently in practical problems.

The methods to be discussed are stochastic versions of the basic non-linear programming methods of feasible directions (to be described in the next section). Consider briefly a common form of the deterministic problem of constrained optimization, where $f(x)$ can be calculated. Let X_n denote the nth estimate of a point where a necessary condition for optimality holds. To calculate X_{n+1}, we search in a direction h_n from X_n, where h_n is selected according to a given rule. Define $\lambda_n \equiv \inf\{\lambda : X_n + \lambda h_n \notin C, \lambda \geqq 0\}$. A one-dimensional search for the minimum of $f(x)$ on the segment $l_n^+ \equiv \{x : x = X_n + \lambda h_n, 0 \leqq \lambda \leqq \lambda_n\}$ is made. In any procedure which is to be implemented on a computer, each one-dimensional search procedure cannot continue until a minimum in that direction is found, but must stop in a finite time, yielding the point X_{n+1}, which is not usually the location of the directional minimum. The stopping rule can be such that, for example, the reduction $f(X_{n+1}) - f(X_n)$ is at least a predetermined function of $\nabla f(X_n)$, *the gradient of* $f(\cdot)$ at X_n, or it can be such that X_{n+1} is within ε_n of the minimum of $f(x)$ on l_n^+, where $\varepsilon_n \to 0$ as $n \to \infty$. The rules must be

such that the X_n converge to a desirable point. Since it usually is impossible to guarantee (in the absence of strong convexity conditions) that the X_n converge to the global—or even local minimum—we satisfy ourselves by proving convergence to a point where a necessary condition for optimality holds.

Stochastic counterparts of these rules will be develepod. The procedures to be discussed are fairly general. Within their general type, some are certain to be preferable and, hopefully, future experimental work will point these out. While the Kiefer–Wolfowitz procedure of stochastic approximation (SA) cannot be used in its classical form (say Dvoretsky (1956), Blum (1954)), due to the constraints, it is possible to use truncated forms of SA for the one-dimensional searches. The types of results which such use gives have motivated conditions (A4), (A6) in Section 2. Yet these conditions seem very natural in view of the nature of the problem at hand and the somewhat related conditions which are sometimes used for the deterministic problem (see, e.g., Zangwell (1969), Polak (1971)), and either they or closely related conditions would undoubtedly hold for useful non-SA methods.

Indeed, our point of view towards SA is that its main use is in illuminating the conditions that should be satisfied by other, potentially more useful, minimization methods.

Section 2 discusses two stochastic methods of feasible directions, and Section 3 shows that a form of SA satisfies (A4), (A6). A general discussion of deterministic feasible direction methods can be found in Topkis and Veinott (1967), Zangwell (1969), and Polak (1971).

2. Feasible direction methods. A stochastic version of the feasible direction method of Topkis and Veinott (1967) (see also Polak (1971), pages 150–159) will be developed. It will sometimes be convenient to use the notation $q^0(\cdot)$ for $f(\cdot)$. Assume that

(A1) $f(\cdot)$ has continuous second derivatives.

(A2) The $q^i(\cdot)$, $i = 1, \cdots, s$, have continuous first derivatives and $C = \{x: q^i(x) \leqq 0, i = 1, \cdots, s\}$ is compact and is the closure of its interior.

Let $\hat{x}$ denote any value of x which minimizes $f(\cdot)$ on C, let G denote the hypercube $G = \{h: |h^i| \leqq 1, i = 1, \cdots, r\}$ where $h = (h^1, \cdots, h^r)$, define the function $u_0(\cdot, \cdot)$ on $R^r \times G$ by (where $\langle \cdot, \cdot \rangle$ denotes the inner product in R^r)

$$u_0(x, h) = \max \{\langle \nabla q^0(x), h\rangle; q^i(x) + \langle \nabla q^i(x), h\rangle, i = 1, \cdots, s\}, \tag{2a}$$

and define

$$u(x) = \min_{h \in G} u_0(x, h). \tag{2b}$$

The set G can be replaced by any compact convex set in R^r which contains the origin in its interior. Note that $u(x) \leqq 0$. Under (A1)—(A2), $u(\hat{x}) = 0$ (Polak (1971), page 154).

For any fixed x, the minimum u in the linear program (3) is $u(x)$ and any

minimizing vector h in (3) minimizes (yields $u(x)$) in (2b) (Polak (1971), pages 150–159). The minimizing h may not be unique.

Minimize u subject to

$$\begin{aligned} &-u + \langle \nabla q^0(x), h\rangle \leqq 0\,; \\ &-u + q^i(x) + \langle \nabla q^i(x), h\rangle \leqq 0\,, \qquad i = 1, \cdots, s\,; \\ &-1 \leqq h^i \leqq 1\,. \end{aligned} \tag{3}$$

As is common in the analysis of deterministic algorithms, we are interested in iterative methods which estimate *any* point $\bar{x}$ at which the necessary condition $u(\bar{x}) = 0$ holds. Let X_n denote the nth estimate of a point $\bar{x}$ at which $u(\bar{x}) = 0$. Then the deterministic method uses any value of h which minimizes in (2b) or (3) for $x = X_n$, as the direction[2] h_n. Then (usually) a one-dimensional search (for X_{n+1}, a "better" estimate of $\bar{x}$) is conducted in C along h_n through X_n, until it can be guaranteed that X_{n+1} is either sufficiently close to the minimum in the search direction, or that $f(X_{n+1}) - f(X_n)$ is bounded below zero by some quantity which depends only on the slope $\langle \nabla q^0(X_n), h_n\rangle$. Then a new direction of search is selected, etc. The stochastic scheme is developed somewhat analogously, except, of course, (3) cannot be solved exactly since $\nabla q^0(X_n)$ is unknown, and only "noisy" estimates are obtainable.

Suppose that the random variable h_n (a random vector with values in G) is the random direction of search which is used on the nth cycle (to calculate X_{n+1}) in the stochastic problem. Let $\mathscr{B}_n^+$ and $\mathscr{B}_n$ denote the σ-algebras generated by $\{X_i, h_i; i = 0, \cdots, n\}$ and $\{X_i; h_i, i = 0, \cdots, n-1, \text{ and } X_n\}$, resp.

Define the conditions (A3), (A4) as follows:

(A3) Let $\delta_1(\cdot)$, $\delta_2(\cdot)$ denote some real, non-decreasing and positive functions on $(0, \infty)$, and $n_1(\cdot)$ a positive integer valued non-increasing function on $(0, \infty)$. Let h_n satisfy ($P_{\mathscr{B}}$ denotes the probability conditioned on the σ-algebra $\mathscr{B}$)

$$P_{\mathscr{B}_n}\{u_0(X_n, h_n) \leqq -\delta_1(-u(X_n))\} \geqq \delta_2(-u(X_n)) \tag{4}$$

a.e. on the ω-set where $n \geqq n_1(-u(X_n))$.

(A4) Let

$$E_{\mathscr{B}_n^+} f(X_{n+1}) - f(X_n) \leqq \beta_n \tag{5}$$

where $E\sum_0^\infty |\beta_n| \equiv \bar{\beta} < \infty$. Let $g(\cdot)$ be a real positive non-decreasing function on $(0, \infty)$ and $c_1(\cdot)$ a positive integer-valued non-increasing function on $(0, \infty)$ so that,

$$E_{\mathscr{B}_n^+} f(X_{n+1}) - f(X_n) \leqq -g(-u_0(X_n, h_n)) + \beta_n \tag{6}$$

a.e. on the set where

$$n \geqq c_1(-u_0(X_n, h_n))\,, \qquad u_0(X_n, h_n) < 0\,.$$

[2] h_n may not be a unit vector, but this is unimportant since we could also write the direction as $h_n/|h_n|$ where $|\cdot|$ is the Euclidean norm.

If h_n is chosen according to (A3) and $\lambda_n = 0$, then define $X_{n+1} = X_n$ and choose h_{n+1} according to a rule satisfying (A3).

(A3) relates $u_0(X_n, h_n)$ to $u(X_n)$, and is a relation concerning the choice of direction h_n. (A4) relates, statistically, the improvement per cycle in $Eq^0(X_n)$ to the "directional term" $u_0(X_n, h_n)$. Thus each step of the procedure is broken down into two natural components. The first selects a direction (satisfying (A3)). The second searches in that direction using a method satisfying (A4). (A4) holds when the one-dimensional search procedure is a form of truncated SA, at least if $f(\cdot)$ is convex. See Section 3.

THEOREM 1. *Assume* (A1)—(A4). *Then* $u(X_n) \to 0$ w.p. 1 *and* $u(x) = 0$ *for every accumulation point* x *of* $\{X_n(\omega)\}$ w.p. 1.

PROOF. Define the real-valued function $m_1(\cdot)$ on $(0, \infty)$ by $m_1(\varepsilon) = \max [c_1(\delta_1(\varepsilon)), n_1(\varepsilon)]$. Select an arbitrary real $\varepsilon > 0$. Let n be an integer $\geqq m_1(\varepsilon)$. Then (A3) gives, a.e. on $\{u(X_n) \leqq -\varepsilon\}$,

$$P_{\mathscr{B}_n}\{u_0(X_n, h_n) \leqq -\delta_1(\varepsilon)\} \geqq \delta_2(\varepsilon) .$$

Also, (A4) gives, a.e. on $\{u_0(X_n, h_n) \leqq -\delta_1(\varepsilon)\}$,

$$E_{\mathscr{B}_n} f(X_{n+1}) - f(X_n) \leqq -g(\delta_1(\varepsilon)) + \beta_n .$$

Thus (A3) and (A4) together yield (I_A is the indicator function of the set A)

$$\begin{aligned} E[f(X_{n+1}) - f(X_n)] &\leqq -Eg(\delta_1(\varepsilon))I_{\{u(X_n)\leqq-\varepsilon\}}I_{\{u_0(X_n,h_n)\leqq-\delta_1(\varepsilon)\}} + E\beta_n \\ &\leqq -Eg(\delta_1(\varepsilon))I_{\{u(X_n)\leqq-\varepsilon\}}E_{\mathscr{B}_n}I_{\{u_0(X_n,h_n)\leqq-\delta_1(\varepsilon)\}} + E\beta_n \\ &\leqq -Eg(\delta_1(\varepsilon))\delta_2(\varepsilon)I_{\{u(X_n)\leqq-\varepsilon\}} + E\beta_n . \end{aligned}$$

The second inequality follows from (A4), and the last from (A3). Upon summing the final inequality, we get

$$Ef(X_n) - Ef(X_{m_1(\varepsilon)}) \leqq -E\sum_{i=m_1(\varepsilon)}^{n-1} g(\delta_1(\varepsilon))\delta_2(\varepsilon)I_{\{u(X_i)\leqq-\varepsilon\}} + \tilde{\beta} .$$

Since the left-hand side of the above inequality is uniformly bounded in n, $u(X_n) \leqq -\varepsilon$ only finitely often w.p. 1, and, hence $u(X_n) \to 0$ w.p. 1. Since $u(X_n) \to 0$ w.p. 1, there is a null set N so that for $\omega \notin N$, the continuity of $u(\cdot)$ and the compactness of C imply that $u(x) = 0$ w.p. 1. for any accumulation point x of $\{X_n(\omega)\}$. □

Remarks on the choice of h_n. (A3) requires only that if $u(X_n) < 0$, then with some nonzero probability, $u_0(X_n, h_n)$ is negative, the negativeness and the probability being bounded away from zero (for sufficiently large n) by quantities dependent on $u(X_n)$ but not on X_n, h_n or past data.

As the following argument shows, a purely random choice of h_n can satisfy (A3). Let h be selected according to a uniform distribution on the surface of the box G, with $\{h_n\}$ an independent sequence of such h's. By the uniform continuity of $u(\cdot)$ and $u_0(\cdot, \cdot)$ on C and $C \times G$ resp., there is a positive and

non-decreasing function $\delta_2(\cdot)$ on $(0, \infty)$, such that for any $\varepsilon > 0$, and for all x for which $u(x) \leqq -\varepsilon$

$$P_x\{u_0(x, h) \leqq -\varepsilon/2\} \geqq \delta_2(\varepsilon)\,. \tag{7}$$

(where P_x denotes the probability with x a fixed number). This yields (A3).

It is shown in Kushner (1972a), pages 2–11, 12, that, under certain conditions, (A3) also holds if we let h_n be the minimizing h in (3), where we put $x = X_n$ and a "noisy" finite difference estimate is used for $\nabla q^0(X_n)$.

Another feasible direction method. The method of this section is a stochastic analog of Polak's (1971, pages 159–176) version of a method of Zoutendijk (1960). For any $x \in C$ and $\varepsilon \geqq 0$, define the index set $J_\varepsilon(x) = \{0\} \cup \{i: q^i(x) \geqq -\varepsilon\}$. $J_\varepsilon(x)$ consists of the indices of $q^0(\cdot) = f(\cdot)$ and of the "ε-active" constraints. *Assume* (A1), (A2). Define the real-valued functions $\gamma_0(\cdot, \cdot, \cdot)$ and $\gamma(\cdot, \cdot)$ on $[0, \infty) \times R^r \times G$ and $[0, \infty) \times R^r$ resp., by

$$\gamma_0(\varepsilon, x, h) = \max_{i \in J_\varepsilon(x)} \langle \nabla q^i(x), h\rangle \tag{8}$$

$$\gamma(\varepsilon, x) = \min_{h \in G} \gamma_0(\varepsilon, x, h)\,. \tag{9}$$

If $\hat{x}$ minimizes $f(x)$ in C, then $\gamma(0, \hat{x}) = 0$. Also, for any x and $\varepsilon \geqq 0$, $\gamma(\varepsilon, x)$ is the minimizing γ in the linear program: *minimize* γ *subject to*

$$-\gamma + \langle \nabla q^i(x), h\rangle \leqq 0\,, \qquad i \in J_\varepsilon(x),\ h \in G\,. \tag{10}$$

Again, our interest is in stochastic schemes for determining a sequence $\{X_n\}$ so that (w.p. 1) any accumulation point x of $\{X_n(\omega)\}$ satisfies the necessary condition $\gamma(0, x) = 0$.

A form of the *deterministic* procedure is as follows. Let the nth iterate X_n be given. Fix $\varepsilon_0 > 0$ and $\beta \in (0, 1)$ and solve (10) with $\varepsilon = \varepsilon_0$ and $x = X_n$. Let $h(\varepsilon, X_n)$ denote any vector h which minimizes in (10) (or, equivalently, in (9)). If $\gamma(\varepsilon_0, X_n) \leqq -\varepsilon_0$, then choose X_{n+1} as a value of x which minimizes $f(x)$ on the line segment in C which goes through X_n in direction $h(\varepsilon_0, X_n)$. If $\gamma(\varepsilon_0, X_n) > -\varepsilon_0$, find the least value of k (by solving a sequence of linear programs (10)) for which $\gamma(\beta^k\varepsilon_0, X_n) \leqq -\beta^k\varepsilon_0$; then minimize $f(x)$ in the corresponding direction $h(\beta^k\varepsilon_0, X_n)$. Then any accumulation point x of $\{X_n\}$ satisfies $\gamma(0, x) = 0$. It is not actually necessary to minimize $f(x)$ in each of the one-dimensional searches. As for the previous method, it is sufficient to choose an X_{n+1} so that $f(X_{n+1}) - f(X_n)$ is bounded below zero by some suitable function of the gradient of $f(\cdot)$ at X_n.

It is not required to consider all the constraints at each step, only the "ε active" ones, a definite advantage in the deterministic case, and probably an advantage in the stochastic case also.

For the stochastic method based on the necessary condition $\gamma(0, x) = 0$, let h_n (again) denote the random direction of search on the nth cycle (to compute X_{n+1}). The assumptions (A5), (A6) below will be assumed in lieu of (A3), (A4).

In (A5), (A6), the functions $\delta_i(\cdot)$, $n_1(\cdot)$, $c_1(\cdot)$ and $g(\cdot)$ satisfy the same conditions which we imposed on these functions in (A3), (A4).

(A5) For each $\varepsilon > 0$, let h_n satisfy $P_{\mathscr{A}_n}\{\gamma_0(\varepsilon, X_n, h_n) \leqq -\delta_1(\varepsilon)\} \geqq \delta_2(\varepsilon)$ a.e. on the set where $n \geqq n_1(-\gamma(\varepsilon, X_n))$ and $\gamma(\varepsilon, X_n) \leqq -\varepsilon$.

(A6) For some random sequence β_n' satisfying $E\sum_0^\infty |\beta_n'| = \tilde{\beta}' < \infty$, let the one-dimensional search procedure satisfy (5) for β_n' replacing β_n and also, for each $\varepsilon > 0$,

$$E_{\mathscr{A}_n} f(X_{n+1}) - f(X_n) \leqq -g(-\gamma_0(\varepsilon, X_n, h_n)) + \beta_n' \tag{11}$$

a.e. on the set where

$$n \geqq c_1(-\gamma_0(\varepsilon, X_n, h_n)), \qquad \gamma_0(\varepsilon, X_n, h_n) < 0. \tag{12}$$

If h_n is chosen by a rule satisfying (A5) and $\lambda_n = 0$, then set $X_{n+1} = X_n$ and choose h_{n+1} according to a rule satisfying (A5).

THEOREM 2. *Assume* (A1), (A2), (A5), (A6). *There is a null set N so that for* $\omega \notin N$, *every accumulation point x of the sequence* $\{X_n(\omega)\}$ *satisfies* $\gamma(0, x) = 0$.

PROOF. Let $N_\rho(x)$ denote the set $C \cap \{y: |y - x| < \rho\}$. We will actually prove that if (13) holds, then $\gamma(0, \tilde{x}) = 0$.

$$P\{X_n \in N_\rho(\tilde{x}) \text{ infinitely often}\} > 0, \qquad \text{each } \rho > 0. \tag{13}$$

This will yield the theorem, as follows: C is compact. Define R_1 by $R_1 = \{x: x \in C, \gamma(0, x) = 0\}$. First we show that R_1 is compact. Let $x_n \in R_1$ and $x_n \to x$. If $q^i(x_{n_j}) = 0$ for a subsequence $\{x_{n_j}\}$, then $q^i(x) = 0$. Thus $J_0(x) \supset \bigcap_{N=1}^\infty \bigcup_{n=N}^\infty J_0(x_n)$. Then the continuity of the $\nabla q^i(\cdot)$ yields

$$\begin{aligned} 0 &\geqq \min_G \max_{J_0(x)} \langle \nabla q^i(x), h\rangle = \lim_n \min_G \max_{J_0(x)} \langle \nabla q^i(x_n), h\rangle \\ &\geqq \lim_n \min_G \max_{J_0(x_n)} \langle \nabla q^i(x_n), h\rangle = 0, \end{aligned}$$

which implies $x \in R_1$. Thus R_1 is compact.

Assume that $\gamma(0, \tilde{x}) = 0$ if (13) holds. Then for each $\delta > 0$ there is, for each $x \in C - R_1$, a $\rho_x \in (0, \delta)$ so that $X_n \in N_{\rho_x}(x)$ only finitely often w.p. 1. For $\delta > 0$, let D_δ denote the compact set

$$D_\delta = \{x: \min_{y \in R_1} |x - y| \geqq \delta, x \in C\}.$$

Then there is a finite collection $x_{i_1} \cdots x_{i_\delta}$ (in D_δ) so that D_δ is covered by the sets $N_{\rho_{x_i}}(x_i)$. Thus $X_n \in D_\delta$ only finitely often w.p. 1. Thus, since $\{D_{1/r}, r = 1, 2, \cdots\}$ is a non-decreasing sequence, there can be no accumulation point (w.p. 1) of $\{X_n(\omega)\}$ in $D = \bigcup_{r=1}^\infty D_{1/r}$. This proves the assertion since $C - D = R_1$. Thus, we only need prove that (13) implies that $\gamma(0, \tilde{x}) = 0$.

Let $\tilde{x} \in C$ satisfy $\gamma(0, \tilde{x}) \leqq -v_1 < 0$, and let (13) hold. (The argument from this point to (15) is the same as in the deterministic case. The rest of the proof shows that (13) is inconsistent with $v_1 > 0$; hence $\gamma(0, \tilde{x}) = 0$, as desired.) For each $v > 0$, there is a $\rho_1(v) > 0$ for which $x \in N_{\rho_1(v)}(\tilde{x})$ implies that

$$|\min_G \max_{J_0(\tilde{x})} \langle \nabla q^i(x), h\rangle - \gamma(0, \tilde{x})| < v/2. \tag{14}$$

Next we show that there are some $v_2 > 0$, $\rho_2 > 0$ so that $J_v(x) \subset J_0(\tilde{x})$ for $x \in N_{\rho_2}(\tilde{x})$ and $v \leqq v_2$. (This is equivalent to showing that if $q^i(x) \geqq -v \geqq -v_2$ for $x \in N_{\rho_2}(\tilde{x})$, then $q^i(\tilde{x}) = 0$.) If $q^i(\tilde{x}) = 0$ for all i, we are done. If not, for each $i \notin J_0(\tilde{x})$, there exist (by continuity of $q^i(\cdot)$) $\delta_i > 0$ and $a_i > 0$ such that $q^i(x) \leqq -a_i$ for all $x \in N_{\delta_i}(\tilde{x})$. Let $\rho_2 = \min\{\delta_i\}$, and $v_2 = \min\{a_i\}$. If $i \notin J_0(\tilde{x})$ and $x \in N_{\rho_2}(\tilde{x})$, and if $v \leqq v_2$, then $i \notin J_v(x)$.

Next define $v = \min(v_1, v_2)$, $\rho = \min(\rho_1(v), \rho_2)$. Note that $\gamma(\varepsilon, x)$ is non-increasing as ε decreases, since as ε decreases, the number of "ε active" constraints is non-increasing. Using this fact together with (14) and the (just proved) fact that $J_v(x) \subset J_0(\tilde{x})$ for $x \in N_\rho(\tilde{x})$ yields, for $x \in N_\rho(\tilde{x})$,

$$\begin{aligned}(15)\qquad \gamma(v/2, x) \leqq \gamma(v, x) &= \min_G \max_{J_v(x)} \langle \nabla q^i(x), h\rangle \\ &\leqq \min_G \max_{J_0(\tilde{x})} \langle \nabla q^i(x), h\rangle \leqq -v/2\,.\end{aligned}$$

Since, by (15), $x \in N_\rho(\tilde{x})$ implies that $\gamma(v/2, x) \leqq -v/2$, we have that (13) implies that

$$(16)\qquad P\{\gamma(v/2, X_n) \leqq -v/2 \text{ infinitely often}\} > 0\,.$$

Define $m_1(\cdot)$ by $m_1(\varepsilon) = \max[n_1(\varepsilon), c_1(\delta_1(\varepsilon))]$.

Let $n \geqq m_1(v/2)$. Then (A5) gives, a.e. on $\{\gamma(v/2, X_n) \leqq -v/2\}$,

$$P_{\mathscr{B}_n}\{\gamma_0(v/2, X_n, h_n) \leqq -\delta_1(v/2)\} \geqq \delta_2(v/2)\,.$$

Also (A6) gives, a.e. on $\{\gamma_0(v/2, X_n, h_n) \leqq -\delta_1(v/2)\}$,

$$E_{\mathscr{B}_n} f(X_{n+1}) - f(X_n) \leqq -g(\gamma_0(v/2, X_n, h_n)) + \beta_n'\,.$$

Thus (A5) and (A6) together yield

$$\begin{aligned}&E[f(X_{n+1}) - f(X_n)] \\ &\quad\leqq -Eg(\delta_1(v/2))I_{\{\gamma(v/2, X_n) \leqq -v/2\}} I_{\{\gamma_0(v/2, X_n, h_n) \leqq -\delta_1(v/2)\}} + E\beta_n' \\ &\quad\leqq -Eg(\delta_1(v/2))I_{\{\gamma(v/2, X_n) \leqq -v/2\}} E_{\mathscr{B}_n} I_{\{\gamma_0(v/2, X_n, h_n) \leqq -\delta_1(v/2)\}} + E\beta_n' \\ &\quad\leqq -Eg(\delta_1(v/2))I_{\{\gamma(v/2, X_n) \leqq -v/2\}} \delta_2(v/2) + E\beta_n'\,.\end{aligned}$$

Upon summing the last inequality we get

$$Ef(X_n) - Ef(X_{m_1(v/2)}) \leqq -E\sum_{i=m_1(v/2)}^{n-1} g(\delta_1(v/2))\delta_2(v/2) I_{\{\gamma(v/2, X_n) \leqq -v/2\}} + \bar{\beta}\,.$$

The left-hand side is uniformly bounded in n, since $f(\cdot)$ is bounded on C. Thus $\sum_{i=m_1(v/2)}^{\infty} I_{\{\gamma(v/2, X_n) \leqq -v/2\}} < \infty$ w.p. 1. Hence $\gamma(v/2, X_n) \leqq -v/2$ only finitely often w.p. 1, and, consequently, $X_n \in N_\rho(\tilde{x})$ only finitely often w.p. 1. This contradicts (13). Thus v_2 cannot be greater than zero and, hence $\gamma(0, \tilde{x}) = 0$. □

3. A one-dimensional search procedure which satisfies both (A4) and (A6). Let $\{N_i\}$ denote a sequence of integer-valued random variables, where N_i is non-anticipative with respect to a sequence $\{Z_n, n = 0, 1, \cdots\}$ to be defined below. Let $\{a_{in}\}$, $\{c_{in}\}$ denote positive real sequences such that $a_{in} \to 0$, $c_{in} \to 0$, as $n + i \to \infty$. Suppose $B_1 > 0$ is real, and that for almost all ω the sequences

$\{N_i(\omega)\}$, $\{a_{in}\}$ and $\{c_{in}\}$ satisfy

$$\sum_{i=1}^{\infty} \sum_{n=0}^{N_i-1} a_{in} = \infty , \qquad \sum_{i=1}^{\infty} \sum_{n=0}^{N_i-1} (a_{in} c_{in} + a_{in}^2/c_{in}^2) < \infty \tag{17}$$

and

$$\sum_{n=0}^{N_i-1} a_{in} \geqq B_1 . \tag{18}$$

Define the random function $\phi_i(\cdot)$ on $S_i = [0, \lambda_i]$ by $\phi_i(z) = f(X_i + h_i z)$, and define the random distribution function with parameter z (on S_i) by $H_i(y \mid z) = H(y \mid X_i + h_i z)$. The ith one-dimensional search (sequentially) seeks a local minimum of $\phi_i(z)$ on S_i by taking N_i steps of a type of Kiefer–Wolfowitz procedure with parameters $\{a_{in}, c_{in}, n = 0, 1, \cdots\}$. The iterates of the ith search will be calculated from the formula

$$Z_{n+1}^i = Z_n^i - a_{in} DY(Z_n^i, c_{in}) , \tag{19}$$

where $Z_0^i = 0$ and the real-valued random variable $DY(\cdot, \cdot)$ is defined below. *All observations used in the calculation of* $DY(z, \cdot)$ *must be taken with parameter values* z *on the interval* S_i (so that $X_i + h_i z$ will lie in C), although Z_n^i itself will not necessarily always be on S_i. We define

$$X_{i+1} = X_i + h_i [Z_{N_i}^i]_{S_i} \tag{20}$$

where

$$\begin{aligned} [Z]_{S_i} &= Z \quad \text{if} \quad Z \in S_i \\ &= 0 \quad \text{if} \quad Z < 0 \\ &= \lambda_i \quad \text{if} \quad Z > \lambda_i . \end{aligned}$$

$DY(Z_n^i, c_{in})$ will be a "noisy approximation" to the derivative of $\phi_n(\cdot)$ at Z_n^i, if $Z_n^i \in S_i$, or to the derivative of $\phi_i(\cdot)$ at 0 or λ_i, resp., if $Z_n^i < 0$ or $Z_n^i > 0$, resp.

The definition of $DY(\cdot, \cdot)$.

Case A. Let $Z_n^i \in S_i$ and $c_{in} \leqq \lambda_i$. If $\lambda_i - Z_n^i \geqq c_{in}$, define

$$DY(Z_n^i, c_{in}) = (Y_{2n+1}^i - Y_{2n}^i)/c_{in} , \tag{21}$$

where Y_{2n+1}^i and Y_n^i are drawn from $H_i(y \mid z)$ with parameters $Z_n^i + c_{in}$ and Z_n^i, resp. If $\lambda_i - Z_n^i < c_{in}$, but $Z_n^i \geqq c_{in}$, then use (21) but with Y_{2n+1}^i and Y_{2n}^i drawn from $H_i(y \mid z)$ with parameters Z_n^i, $Z_n^i - c_{in}$, resp. If $\lambda_i - Z_n^i < c_{in}$, $Z_n^i < c_{in}$, use (21) with parameters λ_i, $\lambda_i - c_{in}$, resp.

Case B. Let $Z_n^i \in S_i$ and $c_{in} > \lambda_i$. The actual finite difference interval can now be no greater than λ_i, but if λ_i is "too small too often," the "noise effects" may not be summable. Thus, we use the procedure: Define $r_{in} = \min \{r : c_{in}/r \leqq \lambda_i, r \text{ a positive integer}\}$. Take $2r_{in}^2$ observations $\{Y_{2n+1}^{i,j}; j = 1, \cdots, r_{in}^2\}$ with parameter λ_i and $\{Y_{2n}^{i,j}; j = 1, \cdots, r_{in}^2\}$ with parameter 0. Define

$$DY(Z_n^i, c_{in}) = \frac{1}{r_{in}^2} \sum_{j=1}^{r_{in}^2} (Y_{2n+1}^{i,j} - Y_{2n}^{i,j})/\lambda_i .$$

Note that the variance of $DY(Z_i^n, c_{in})$, conditioned on all past observations, is bounded above by $2\hat{\sigma}^2/c_{in}^2$, which is the same bound that we would get for Case A.

Case C. If $Z_n^i > \lambda_i$, define $DY(Z_n^i, c_{in})$ by $DY(\lambda_i, c_{in})$. If $Z_n^i < 0$, define $DY(Z_n^i, c_{in})$ by $DY(0, c_{in})$.

THEOREM 3. *If we assume* (A1)—(A2) *and* (17) *to* (20), *and that* $f(\cdot)$ *is convex, then* (A4) *and* (A6) *hold.*

The proof, which is straightfoward, but rather tedious, will be omitted. Details can be found in Kushner (1972a, Part 2, Section 3). The proof uses various estimates from Kushner (1972b) for quantities of the type $E_{\mathscr{F}_n^+} f(X_{i+1}) - f(X_i)$, where X_{i+1} is (as here) defined as the terminal iterate of a (truncated) stochastic approximation with initial value X_i. It is not required that $f(\cdot)$ be monomodal.

In Kushner (1972a part 2, Theorem 3) the estimates required for (A4), (A6) were obtained for the three cases:

($\phi_{i,z}(\cdot)$ denotes the derivative with respect to z)

(I) $\phi_{i,z}(0) \leqq 0$, $\quad \phi_{i,z}(\lambda_i) \geqq 0$;

(II) $\phi_{i,z}(0) > 0$, $\quad \phi_{i,z}(\lambda_i) \geqq 0$, $\quad \phi_i(\cdot)$ non-decreasing on S_i;

(III) $\phi_{i,z}(0) \leqq 0$, $\quad \phi_{i,z}(\lambda_i) < 0$.

Then β_i and β_i' of (A4) and (A6) are proportional to

$$E_{\mathscr{F}_n^+}\left[\sum_{n=0}^{N_i-1} (a_{in} c_{in} + a_{in}^2/c_{in}^2)\right] = \tilde{\beta}_i . \tag{22}$$

Also, it is shown that if $\phi_{i,z}(0) \leqq -\varepsilon < 0$ and $\lambda_i \geqq k\varepsilon$ for some real k (independent of i), then there are real $Q_1 > 0$, $k_2 < \infty$, so that for large enough i (depending on ε),

$$E_{\mathscr{F}_n^+} f(X_{n+1}) - f(X_n) \leqq -Q_1\varepsilon^2 + k_2\tilde{\beta}_i . \tag{23}$$

These results are just what is needed to prove (A4), (A6). (I)—(III) do not cover all cases, but they do hold if $f(\cdot)$ is convex. There is an error in the statement of Theorem 4 in Kushner (1972a) (which is Theorem 3 here). (I)—(III) do not always hold under the conditions of $f(\cdot)$ given there, but they do if $f(\cdot)$ is convex. The proof requires no further change.

4. Extensions. A further discussion of the stochastic method of feasible directions, with further algorithms, and some numerical results appear in Kushner and Gavin (1974). A saddle point method for treating the problem where the $q^i(-)$ can only be observed in the presence of noise appears in Kushner and Sanvicente (1973).

REFERENCES

[1] BLUM, J. R. (1954). Multidimensional stochastic approximation procedures. *Ann. Math. Statist.* **25** 737–744.

[2] CANON, M. D., CULLUM, C. D. and POLAK, E. (1970). *Theory of Optimal Control and Mathematical Programming*. McGraw-Hill, New York.

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

[4] KUSHNER, H. J. (1972a). Stochastic approximation type algorithms for the optimization of constrained and multimode stochastic problems. Technical Report 72–1, Center for Dynamical Systems, Brown Univ., Providence.

[5] KUSHNER, H. J. (1972b). Stochastic approximation algorithms for the local optimization of functions with non-unique stationary points. *IEEE Trans. Automatic Control* **AC-16** 646–654.

[6] KUSHNER, H. J. and GAVIN, T. L. (1974). Stochastic approximation type methods for constrained systems: Algorithms and numerical results. To appear in *IEEE Trans. Automatic Control*, July 1974.

[7] KUSHNER, H. J. and SANVICENTE, E. (1973). Stochastic approximation methods for constrained systems with observation noise on the systems and constraints. Submitted to *Automatica*.

[8] POLAK, E. (1971). *Computational Methods in Optimization*. Academic Press, New York.

[9] TOPKIS, D. M. and VEINOTT, A., JR. (1967). On the convergence of some feasible directions methods for nonlinear programming. *SIAM J. Contr.* **5** 268–279.

[10] ZANGWELL, W. I. (1969). *Nonlinear Programming: A Unified Approach*. Prentice Hall, Englewood Cliffs, New Jersey.

[11] ZOUTENDIJK, G. (1960). *Methods of Feasible Directions*. Elsevier, Amsterdam.

DIVISION OF APPLIED MATH.
AND ENGINEERING
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912