The Annals of Mathematical Statistics
1969, Vol. 40, No. 1, 299–302

NOTES

STOCHASTIC APPROXIMATION FOR SMOOTH FUNCTIONS

By Václav Fabian

Michigan State University and Czechoslovak Academy of Sciences

1. Summary. The problem of approximating a point θ of minimum of a function $f \varepsilon \mathcal{C}$ (see 2.1) is considered. An approximation procedure of the type described in Fabian (1967) using the design described in Fabian (1968), but with the size of design increasing, achieves the speed

(1) $$E\|X_n - \theta\|^2 = o(t_n^{-1} \log^3 t_n);$$

here X_n is the nth approximation and t_n the number of observations necessary to construct $X_1, X_2, \cdots, X_n$.

2. The result. We shall refer to the two papers mentioned above by using symbols I and II. The k-dimensional Euclidean space will be denoted by R^k, coordinates of matrices and vectors will be denoted using superscripts. All norms are Euclidean. The symbols $o(h_n)$ and $O(h_n)$ mean function sequences (number sequences in particular) such that $h_n^{-1} o(h_n)$ converge uniformly to zero and $|h_n^{-1} O(h_n)|$ are uniformly bounded by a constant. As usual $\log_2$ stands for log log and δ_{ij} is the Kronecker symbol.

2.1. *Functions considered.* The class $\mathcal{C}$ of functions considered contains f if and only if f satisfies the following conditions: Denote, if they exist, by $H(x)$, $D_s(x)$, respectively, the matrix of the second partial derivatives of f at x and the vector of the sth order partial derivatives of f at x with respect to the individual coordinates, so that $D_s^{(j)}(x) = \partial^s f(x)/\partial(x^{(j)})^s$. There exist positive numbers K_0, K_2, K_3, r, a point $\theta \varepsilon R^k$ and a neighborhood N of θ such that for every $x \varepsilon R^k$, the Hessian $H(x)$ exists and

(1) $$D_1(\theta) = 0, \qquad (x - \theta)'D_1(x) \geqq K_0\|x - \theta\|^2, \qquad \|H(x)\| \leqq K_3;$$

for every positive integer s and every $x \varepsilon N$, $D_s(x)$ exists and

(2) $$\|D_s(x)\| \leqq K_2 s! r^s.$$

2.2. *Remark.* Note that if $k = 2$ and f is analytic in a sphere with radius $\rho > r^{-1}$, Cauchy formula implies (2.1.2) for points in the sphere with radius r^{-1} (both spheres with the same center and K_2 suitably chosen).

2.3. *The approximating sequence.* When $f \varepsilon \mathcal{C}$ is given, $X_1, X_2, \cdots$ is supposed to be a sequence of k-dimensional random vectors satisfying the following conditions:

(1) $$X_{n+1} = X_n - a_n Y_n, \qquad EX_1^2 < +\infty$$

Received 16 December 1967.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

with

(2) $a_n = an^{-1}\psi_n, \qquad a > 0, \qquad c_n = (s_n^{-1}n)^{-\frac{1}{2}/(s_n+1)}\psi_n^{-1}, \qquad \psi_n = \log_2^{1/4} n,$

(3) $$s_n = \log n \log_2^{-1} n \quad \text{if} \quad n \geqq e^e,$$

$s_n = \log n$ for $n < e^e$. With $\mathfrak{X}_n = [X_1, X_2, \cdots, X_n]$, $E_{\mathfrak{X}_n} Y_n = M_n(X_n)$ where

(4) $M_n^{(j)}(x) = c_n^{-1} \sum_{i=1}^m v_i[f(x + c_n u_i e_j) - f(x - c_n u_i e_j)], \quad j = 1, 2, \cdots, k,$

(with vectors e_j satisfying $e_j^{(i)} = \delta_{ij}$) and

(5) $$E_{\mathfrak{X}_n} \|Y_n - M_n(X_n)\|^2 \leqq 2k\, c_n^{-2}\sigma^2 \sum_{i=1}^m (v_i^2/n_i).$$

The design $[u_1, \cdots, u_m]$, $[n_1, \cdots, n_m]$ depends on n and is described in 2.4, the v_i's are solutions to equations

(6) $$\sum_{i=1}^m u_i^{2j-1} v_i = \tfrac{1}{2}\delta_{1j}, \qquad j = 1, 2, \cdots, m.$$

2.4. *The choice of design.* We suppose $m = m_n$ is the smallest integer greater or equal to $s_n/2$,

(1) $u_i = \cos([(m - i)/(2m - 1)]\pi), \quad \xi_i = [2m(m - 1) + \tfrac{1}{2}]^{-1} u_i^{-2}(1 - \tfrac{1}{2}\delta_{im}),$

and, with a fixed integer K, n_i is the smallest integer greater or equal to $Km\,\xi_i$. (To avoid notational difficulties we do not indicate the dependence of u_i, ξ_i, n_i, and sometimes of m, on n.) The solution to (2.3.6) is then (see Theorem (II.5.1))

(2) $$v_i = (2m - 1)^{-1}(-1)^{i-1}u_i^{-2}(1 - \tfrac{1}{2}\delta_{im}).$$

2.5. *Remark.* For m and $N = \sum_{i=1}^m n_i$ given, $n_i > 0$, the sum $\sum_{i=1}^m (v_i^2/n_i)$ appearing in (2.3.5) is minimized by the choice of (2.4.1) with $n_i = N\xi_i$ resulting in

(1) $$\sum_{i=1}^m v_i^2/\xi_i = \tfrac{1}{4}(2m - 1)^2, \qquad \sum_{i=1}^m u_i^{-2} = 2m(m - 1) + 1$$

(see Theorem II.5.1 and relation (II.5.1.9)). This choice will not lead in general to integer valued n_i. Under the choice of 2.4, we obtain

(2) $$\sum_{i=1}^m v_i^2/n_i \leqq \tfrac{1}{4}K^{-1}(2m - 1)^2/m, \qquad \sum_{i=1}^m n_i \leqq (K + 1)m$$

(with $\sum_{i=1}^m n_i = (K + 1)m$, without restricting the n_i's to be integers, we would be able to obtain $\frac{1}{4}(K + 1)^{-1}(2m - 1)^2/m$).

With $2k \sum_{i=1}^m n_i$ observations of function values on stage n we can achieve (2.3.4) and (2.3.5) and assume that t_n, the number of observations up to the nth stage, satisfies

(3) $$t_n \leqq k(K + 1) \sum_{j=1}^n (s_j + 2).$$

2.6. *Theorem. If $f \in \mathfrak{C}$ and $X_1, X_2, \cdots$ is the approximation sequence described in* 2.3 *and* 2.4, *θ the unique stationary point of f and t_n the number of observations necessary to construct $X_1, X_2, \cdots, X_n$, then*

(1) $$E\|X_n - \theta\|^2 = o(t_n^{-1} \log^3 t_n).$$

2.7. *Remark.* It is known that there is no procedure for which $E\|X_n - \theta\|^2 \leqq C\,t_n^{-1}$ with C independent of f (see e.g. Schmetterer's (1961) review, Section 4) and there is probably no such with C depending on f, although the latter seems not to have been proved. The present procedure is evidently not asympotically optimal. The choice of s_n was made in accordance to the fact that we need, later in the proof, to have $c_n^2 s_n \to 0$, but it was not a unique possible choice. The sequence c_n was chosen as to make approximately equal the two terms constituting $E\|X_n - \theta\|^2$, the systematic error, and the variance of $a_n Y_n$. The function ψ_n has been used to eliminate the effect of not knowing the values of K_0 and r, appearing in 2.1.

3. Proof. We shall suppose an $f \in \mathcal{C}$ is given and use the notation used in 2.1, relating to our particular f. As above, we shall frequently skip the subscript n in m_n and s_n.

3.1. *Preliminaries.* Set $h(n) = (s^{-1}n)^{s/(s+1)}$, $q(n) = \psi_n^{-4} h(n)$. In a straightforward way it is possible to verify

$$(1) \qquad a_n^2 c_n^{-2} s_n = a^2 n^{-1} q^{-1}(n), \qquad c_n^{4m_n} \leqq h^{-1}(n)\psi_n^{-4m_n},$$

$$(2) \qquad c_n^2 s_n \to 0,$$

$(d/dn) \log q(n) = O(n^{-1})$ and, for $n \geqq e^e$

$$(3) \qquad q(n+1)/q(n) = 1 + O(n^{-1}), \qquad \log q(n) \geqq \log n - 2 \log_2 n.$$

By (2.5.1), $\sum_{i=1}^m u_i^{-2} \leqq 2m^2$ and Schwartz inequality implies $\sum_{i=1}^m u_i^{-1} \leqq 2m^{3/2}$. From (2.4.2), $|u_i v_i| \leqq m^{-1} u_i^{-1}$, $|u_i^{2m+1} v_i| \leqq m^{-1} u_i^{2m-1} \leqq m^{-1}$ and hence, for $m \geqq 3$

$$(4) \qquad \sum_{i=1}^m |u_i v_i| \leqq 2\, s_n^{\frac{1}{2}}, \qquad \sum_{i=1}^m |u_i^{2m+1} v_i| \leqq 1.$$

3.2. *Properties of the* Y_n. Suppose, without loss of generality, that $\theta = 0$. Choose an $\epsilon > 0$ such that $C(2\epsilon) = \{x;\, \|x\| < 2\epsilon\} \subset N$. By Lemma I.3.1 there is n_0 such that for $n \geqq n_0$, $x \in C(\epsilon)$, $M_n(x) = D(x) + c_n^{2m} Q_n(x)$ with $\|Q_n(x)\| \leqq 2k\ [(2m+1)\,!\,]^{-1} \sum_{i=1}^m |v_i u_i^{2m+1}| \sup\{\|D_{2m+1}(y)\|;\, y \in N\}$, which is less or equal to $2k\,K_2\, r^{2m+1}$ because of (2.1.2). Hence and from (3.1.1), because $r^{2m+1}\psi_n^{-2m} \to 0$,

$$(1) \qquad M_n(x) = D(x) + o(h^{-\frac{1}{2}}(n)) \quad \text{all} \quad x \in C(\epsilon).$$

For all x, by (I.3.1.3), $M_n(x) = 2\sum_{i=1}^m v_i u_i D(\xi_i)$ with $\|\xi_i - x\| < c_n$ and, using (2.1.1), (2.3.6), (3.1.4), and (3.1.2), $M_n(x) = D(x) + 2\sum_{i=1}^m v_i u_i [D(\xi_i) - D(x)] = D(x) + O(1) c_n s_n^{1/2} = D(x) + o(1)$. This together with (1) and (2.1.1) implies first

$$(2) \qquad a_n^2\, \|M_n(x)\|^2 \leqq \|x\|^2 o(a_n) + a_n^2 o(h^{-1}(n))$$

and, secondly,

$$(3) \qquad x' M_n(x) \geqq [K_0 - o(1)]\, \|x\|^2 - \|x\|\, o(h^{-1/2}(n)).$$

But $\|x\|o(h^{-\frac{1}{2}}(n)) \leqq o(\|x\|^2) + o(h^{-1}(n))$ and

(4) $$a_n x' M_n(x) \geqq a_n[K_0 - o(1)] \|x\|^2 - n^{-1}o(q^{-1}(n)).$$

Note also that by (2.3.5), (2.5.2) and (3.1.1)

(5) $$a_n^2 E\|Y_n - M_n(X_n)\|^2 = n^{-1}O(q^{-1}(n)).$$

3.3. *Completion of the proof.* By (3.2.2), (3.2.4) and (3.2.5),

$$E\|X_{n+1}\|^2 = [1 - [2K_0 - o(1)]a_n]E\|X_n\|^2 + n^{-1}O(q^{-1}(n)).$$

Using the first part of (3.1.3) we obtain $[1 - [2K_0 - o(1)]a\, n^{-1}\psi_n]q(n + 1)/q(n) \geqq 1 - n^{-1}$ for sufficiently large n, for which then also

$$q(n + 1)\, E\|X_{n+1}\|^2 \leqq (1 - n^{-1})q(n)\, E\|X_n\|^2 + O(n^{-1}).$$

Chung's lemma (see Lemma I.4.2) then implies

(1) $$E\|X_n\|^2 = O(q^{-1}(n)).$$

Next note that $t_n \geqq n$, $t_n \leqq C\, n\, s_n$ with a constant C according to (2.5.3), and, because of (3.1.3), $\log q(n) \geqq \log n - 2 \log_2 t_n$. We have, however, $\log n \geqq \log t_n - \log s_n - \log C$ and, by (2.3.3), $\log s_n \leqq \log_2 t_n - \log_3 t_n$, which gives

$$\log q(n) \geqq \log t_n - 3 \log_2 t_n + \log_3 t_n - \log C$$

and

$$q^{-1}(n) = o(t_n^{-1} \log^3 t_n)$$

which with (1) implies (2.6.1) and completes the proof.

REFERENCES

[1] Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist.* **38** 191–200.

[2] Fabian, V. (1968). On the choice of design in stochastic approximation methods. *Ann. Math. Statist.* **39** 457–465.

[3] Schmetterer, L. (1961). Stochastic approximation. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 587–609. Univ. of California Press.