JOURNAL OF APPROXIMATION THEORY 49, 79–92 (1987)

A Stochastic Remes Algorithm

HARRO WALK

Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-7000 Stuttgart 80, West Germany

Communicated by P. L. Butzer

Received January 14, 1985

We treat the linear Tchebycheff approximation problem for a regression function $f \in C^2[0, 1]$. A stochastic Remes algorithm which uses only estimates of first derivatives is proposed and investigated for almost sure convergence and rate of convergence.

1. INTRODUCTION

Methods of stochastic iteration or stochastic approximation are used for the investigation of functions whose values are observable only with random noise, especially for the estimate of zeros or minimum points of regression functions $f: \mathbb{R}^k \to \mathbb{R}^k$ and $f: \mathbb{R}^k \to \mathbb{R}$, respectively. The Kiefer–Wolfowitz method (see the survey article of Schmetterer [13]) is a modification of the classical Newton method; here, for the correction vector of the recursion formula, the gradient is replaced by a vector of divided differences for noise-contaminated function values with spans tending to zero, the inverse of the Hessian matrix of second partial derivatives is replaced by the identity matrix, and the whole correction vector is provided with a discount factor $\alpha_n \geqslant 0$ in the nth iteration step, because of noise, where $\alpha_n \to 0$ $(n \to \infty)$, $\sum \alpha_n = \infty$; thus the Kiefer–Wolfowitz method can be considered as a stochastic gradient method.

In this paper the Tchebycheff approximation problem of minimizing $\| f - (a_0 h_0 + \cdots + a_N h_N) \|_\infty$ with respect to $(a_0, \ldots, a_N) \in \mathbb{R}^{N+1}$ is treated, where $\| \ \|_\infty$ denotes the maximum norm on $C[0, 1]$ and where the values of $f \in C^2[0, 1]$ are noise-contaminated and the given functions $h_0, \ldots, h_N \in C^2[0, 1]$ satisfy Haar's condition. Pantel [11] investigated a stochastic Remes algorithm of Newton type, i.e., with estimation of second derivatives (adaptation), and obtained for $f \in C^3[0, 1]$, in the form of an invariance principle, a convergence order somewhat weaker than $n^{-1/4}$. In Section 2 below a stochastic Remes algorithm, without estimation of second derivatives, is proposed giving a recursive estimation of the alter-

0021-9045/87 $3.00

nant and thus an estimation of the optimal coefficients $a_0,\dots, a_N$ and of the minimal error which plays a key role in the investigation. This algorithm relates to a certain Newton-type Remes algorithm (see Meinardus [8; (7.26), (7.27)]) as the Kiefer–Wolfowitz method to the classical Newton method. For nearly arbitrary starting points one obtains strong consistency, i.e., almost sure (a.s.) convergence, of the estimation sequences under weak assumption on the noise (Sect. 3) and convergence in distribution formulated by an invariance principle, where under usual assumptions on the noise the order is $n^{-1/4}$, even $n^{-1/3}$ for $f \in C^3[0, 1]$, in the case of the alternant—as for the Kiefer–Wolfowitz process—and $n^{-1/2}$ in the case of the coefficients and the minimal error (Sect. 4).

2. The Model

Let the real-valued functions $f, h_0,\dots, h_N \in C^2[0, 1]$, $N \in \mathbb{N}$, satisfy the Haar condition, e.g., $h_k(t) = t^k$, $t \in [0, 1]$ $(k = 0,\dots, N)$ and f possessing an $(N+1)$th derivative $\neq 0$ on $(0, 1)$. Then, in view of the Tchebycheff approximation of f by linear combinations of $h_0,\dots, h_N$, there exists exactly one alternant $x_0^* < \cdots < x_{N+1}^*$; moreover, $x_0^* = 0$, $x_{N+1}^* = 1$, and the system of equations

$$\left[f(x_i) - \sum_{k=0}^{N} a_k h_k(x_i) \right](-1)^i = \lambda \qquad (i = 0,\dots, N+1), \tag{1}$$

$$f'(x_i) - \sum_{k=0}^{N} a_k h_k'(x_i) = 0 \qquad (i = 1,\dots, N) \tag{2}$$

for $a_0,\dots, a_N$, $\lambda \in \mathbb{R}$, $0 =: x_0 < x_1 < \cdots < x_N < x_{N+1} := 1$ has exactly one solution $a_0^*,\dots, a_N^*$, λ^*, $x_1^*,\dots, x_N^*$, where $a_0^* h_0 + \cdots + a_N^* h_N$ is the best approximation of f, $|\lambda^*| > 0$ the minimal error and $(x_0^*, x_1^*,\dots, x_N^*, x_{N+1}^*)$ the above alternant (see Meinardus [7, Sect. 4; 8, Sects. 4.1 and 6.1]).

With $K := \{x = (x_1,\dots, x_N) \in \mathbb{R}^N;\ 0 < x_1 < \cdots < x_N < 1\}$ let the unique solution of (1) with fixed $x \in K$ be denoted by $A_0(x),\dots, A_N(x)$, $L(x)$. With

$$g_i(x) := \left[f'(x_i) - \sum_{k=0}^{N} A_k(x)\, h_k'(x_i) \right](-1)^i, \qquad x \in K\ (i = 1,\dots, N)$$

and $g = (g_1,\dots, g_N)$, one has on K

$$(\nabla L(x), g(x)) \geqslant 0 \tag{3}$$

and the equivalence

$$(\nabla L(x), g(x)) = 0 \Leftrightarrow \nabla L(x) = 0 \Leftrightarrow g(x) = 0 \Leftrightarrow x = x^* \tag{4}$$

(Meinardus [7, Sect. 5]), where ∇ denotes a gradient. L has either positive or negative sign throughout on K and x^* as the maximum or minimum point, respectively; there holds

$$L(x^*)=\lambda^*, \qquad A_k(x^*)=a_k^*, \qquad \nabla A_k(x^*)=0 \qquad (k=0,\ldots,N). \tag{5}$$

The problem of solving (1), (2) is equivalent to finding the unique extremum point of L which then has to be inserted into $A_0,\ldots,A_N$, L. Both problems can be treated in the deterministic case by Newton–type Remes methods if some regularity conditions are fulfilled and sufficiently good starting values are available ([7, Sect. 5; 8, Sect. 7]). In the theory of stochastic iteration the adaptive methods of Venter type (see Venter [14]; Nevel'son and Has'minskii [9]) correspond to the Newton method and use estimates of first and second derivatives for optimization problems. For the Tchebycheff approximation problem in the case of noise-corrupted functions, Pantel [11] used an adaptive method concerning (1), (2) with some knowledge on the location of the solution. The algorithm proposed in the following treats the extremum problem for L in the stochastic case, but instead of an adaptive method a modification of the gradient–type Kiefer–Wolfowitz method (see Schmetterer [13]) in stochastic iteration is established where ∇L is replaced by g. It has a more simple structure, uses only minor knowledge on the location of the alternant, and yields a better convergence rate.

In $K_\tau := \{x=(x_1,\ldots,x_N)\in\mathbb{R}^N;\quad x_i-x_{i-1}\geqslant\tau \quad (i=1,\ldots,N+1)\}$, $\tau\in(0,1/N)$, the function L is bounded away from zero. Assume that each non-vanishing linear combination of f, $h_0,\ldots,h_N$ has at most $N+1$ zero points in the algebraic sense, which sharpens the above Haar condition somewhat, but is fulfilled in the above example. Then, if $\tau>0$ is sufficiently small, there holds, besides $x^*\in K_\tau$,

$$x+t(g(x)+h)\in K_\tau \qquad \text{for all } x\in K_\tau,\ t\in[0,t_0(\tau)],\ h\in\mathbb{R}^N \text{ with maximum norm } \|h\|\leqslant c_0(\tau). \tag{6}$$

The algorithm described now assumes and uses the knowledge of such a small $\tau>0$.

The recursion sequence $(X_n)_{n\in\mathbb{N}}$ for the estimation of $x^*:=(x_1^*,\ldots,x_N^*)$ consists of random vectors $X_n=(X_{n1},\ldots,X_{nN})$ in K_τ defined on a probability space $(\Omega,\mathfrak{A},P)$. Let $\alpha_n\in(0,1)$, $\delta_n\in(0,\tau)$, $n\in\mathbb{N}$, with

$$\alpha_n\to 0 \quad (n\to\infty), \qquad \sum\alpha_n=\infty, \qquad \delta_n\to 0 \quad (n\to\infty).$$

From (1) with x_i replaced by X_{ni}, where $X_{n0}:=0$, $X_{n,N+1}:=1$, but with $f(X_{ni})-U_{ni}$ instead of $f(X_{ni})$, where the real random variables U_{ni} $(i=0,\ldots,$

$N+1$) describe the contamination of function values, one obtains instead of $A_k(X_n)(k=0,\ldots,N)$ and $L(X_n)$ contaminated real random variables

$$A_k(X_n)-\tilde{U}_{nk}=:\tilde{A}_{nk} \qquad (k=0,\ldots,N) \qquad \text{and} \qquad L(X_n)-\tilde{U}_{n,N+1}=:\tilde{L}_n.$$

The column vector $(U_{n0},\ldots,U_{n,N+1})'$ is transformed into the column vector $(\tilde{U}_{n0},\ldots,\tilde{U}_{n,N+1})'$ by a random $(N+2)\times(N+2)$-matrix $M(X_n)$, where the $(N+2)\times(N+2)$-matrix $M(x)$, $x\in K$, transforms $(f(0), f(x_1),\ldots, f(x_N), f(1))'$ into the solution $(A_0(x),\ldots, A_N(x), L(x))'$ of (1) according to Cramer's rule.

Now the recursion is given by

$$X_{n+1}:=X_n+\alpha_n t_n s_n G_n, \qquad n\in\mathbb{N}, \tag{7}$$

with arbitrary X_1 in K_τ, where the N-dimensional random vector $G_n=(G_{n1},\ldots,G_{nN})$ is defined by

$$G_{ni}:=(-1)^i\left[(2\delta_n)^{-1}(f(X_{ni}+\delta_n)-f(X_{ni}-\delta_n)-V_{ni})-\sum_{k=0}^{N}\tilde{A}_{nk}h_k'(X_{ni})\right]$$

($i=1,\ldots,N$) with stochastic contaminations V_{ni} of function values, and where the random s_n in $\{-1,0,1\}$ fulfills

$$s_n=sgn\,\lambda^* \qquad \text{for} \quad n \text{ sufficiently large a.s.} \tag{8}$$

(see below) and the random t_n in $[0,1]$ is chosen maximum such that X_{n+1} is in K_τ. As usual in stochastic iteration the factors α_n are used to guarantee that the the influence of noise is not too large.

Let $\beta_n:=(1-\alpha_1)^{-1}\cdots(1-\alpha_n)^{-1}$, $\gamma_n:=\alpha_n\beta_n$, which implies because of $\alpha_n\to 0$ the relation

$$\beta_n=1+\gamma_1+\cdots+\gamma_n\uparrow\infty.$$

The special case $\alpha_n=(n+1)^{-1}$ yields $\beta_n=n+1$, $\gamma_n=1$. For the estimation of a_k^* $(k=0,\ldots,N)$, λ^* there is used the sequence of real random variables

$$\bar{A}_{nk}:=\beta_n^{-1}(\gamma_1\tilde{A}_{1k}+\cdots+\gamma_n\tilde{A}_{nk}) \qquad (k=0,\ldots,N),$$
$$\bar{L}_n:=\beta_n^{-1}(\gamma_1\tilde{L}_1+\cdots+\gamma_n\tilde{L}_n),$$

respectively.

3. Almost Sure Convergence

In this section a.s. convergence of X_n, $\bar{A}_{nk}$, $\bar{L}_n$ defined above to x^*, a_k^* $(k=0,\ldots,N)$, λ^*, respectively, shall be investigated.

If one has no prior knowledge on the sign of λ^*, one can construct a sequence (s_n) in $\{-1, 0, 1\}$ with (8) in the following two ways. In the case that the random variables U_{ni} are square integrable with

$$\sum \alpha_n^2 E U_{ni}^2 < \infty \qquad (i = 0,..., N+1) \tag{9}$$

and the relation

$$E(U_{ni} | \mathfrak{A}_n) \to 0 \qquad (n \to \infty) \text{ a.s. } (i = 0,..., N+1) \tag{10}$$

for conditional expectations holds where $\mathfrak{A}_n$ is the σ-algebra in Ω generated by X_1, $U_{1i},..., U_{n-1,i}$ $(i = 0,..., N+1)$, $V_{1i},..., V_{n-1,i}$ $(i = 1,..., N)$, one obtains

$$\beta_n^{-1}(\gamma_1 \tilde{U}_{1,N+1} + \cdots + \gamma_n \tilde{U}_{n,N+1}) \to 0 \qquad (n \to \infty) \quad \text{a.s.}$$

from martingale theory (see [6, p. 53]), noticing $\alpha_n = \gamma_n/\beta_n$, $\beta_n \uparrow \infty$, and the fact that the random matrix $M(X_n)$ has uniformly bounded elements because of X_n in K_τ; setting now

$$s_n := \operatorname{sgn} \beta_n^{-1}(\gamma_1 \tilde{L}_1 + \cdots + \gamma_n \tilde{L}_n),$$

one obtains (8). In the case that one can take a sequence of noise corrupted observations of the values of f on a fixed set $\{x_0,..., x_{N+1}\}$ with $0 \leqslant x_0 < x_1 < \cdots < x_N < x_{N+1} \leqslant 1$ parallel to the stochastic algorithm, then in an analogous manner with s_n as signum of an arithmetic mean of contaminated function values of L, one obtains (8), if the $(N+1)$-dimensional vectors of observation errors have zero expectation vector and fulfill the strong law of large numbers.

From now on (8) shall be assumed. As to the observation errors, let hold the conditions

$$\beta_n^{-1}(\gamma_1 U_{1i} + \cdots + \gamma_n U_{ni}) \to 0 \qquad (n \to \infty) \text{ a.s. } (i = 0,..., N+1), \tag{11}$$

$$\beta_n^{-1}(\gamma_1 |U_{1i}| + \cdots + \gamma_n |U_{ni}|) = O(1) \text{ a.s.} \qquad (i = 0,..., N+1), \tag{12}$$

$$\beta_n^{-1}(\gamma_1 \delta_1^{-1} V_{1i} + \cdots + \gamma_n \delta_n^{-1} V_{ni}) \to 0 \qquad (n \to \infty) \text{ a.s. } (i = 1,..., N), \tag{13}$$

or the conditions (9), (10), together with

$$\sum \alpha_n^2 \delta_n^{-2} E V_{ni}^2 < \infty \qquad (i = 1,..., N) \tag{14}$$

and $E(V_{ni} | \mathfrak{A}_n) = 0$ (or weaker),

$$\delta_n^{-1} E(V_{ni} | \mathfrak{A}_n) \to 0 \qquad (n \to \infty) \text{ a.s. } (i = 1,..., N). \tag{15}$$

It should be mentioned that, e.g., (11) is implied by (9), (10) (compare the consideration on s_n), also by

$$EU_{ni}=0, \qquad EU_{ni}U_{mi}=0 \qquad (n\neq m),$$

$$\sum \alpha_n^2 EU_{ni}^2 (\log n)^2 < \infty \qquad (i=0,\ldots, N+1)$$

(according to a theorem of Rademacher and Mensov, see [12, Sect. 3.2]), further by

$$EU_{ni}=0,$$

$$|EU_{ni}U_{mi}| \leqslant \text{const.}(1+|\beta_n-\beta_m|^q)^{-1}(\beta_n^p+\beta_m^p) \qquad (i=0,\ldots, N+1)$$

for some p, q with $0\leqslant 2p<q<1$ (according to an argument of Cramér and Leadbetter [3, pp. 94–96]). In stochastic iteration theory, Ljung [5] uses an assumption $B1$ which is defined by a recursion, but can be brought into the simple form (11) with another notation, and gives a further sufficient condition for it.

The following theorem yields a.s. convergence of the algorithm.

THEOREM 1. *Under the general assumptions of Section* 2 *and the assumptions* (8) *and* (11–13) *or* (9), (10), (14), (15), *there holds*

$$X_n \to x^*, \qquad \bar{A}_{nk} \to a_k^* \qquad (k=0,\ldots, N), \qquad \bar{L}_n \to \lambda^* \qquad (n\to\infty) \text{ a.s.} \tag{16}$$

Proof. The main part consists in proving $X_n \to x^*$ a.s. Without loss of generality one can assume

$$s_n = \operatorname{sgn} \lambda^* = 1 \qquad \text{for all} \quad n\in\mathbb{N}.$$

The recursion for (X_n) can be written in the form

$$X_{n+1} = X_n + t_n\alpha_n[g(X_n) + W_n' - R(X_n)\, W_n''] \tag{17}$$

with an N-dimensional random vector W_n' and an $(N+2)$-dimensional random vector W_n'' and a function R on K with values in the linear space of $N\times(N+2)$-matrices normed by the maximum sum of absolute values of a row. As to regularity conditions there is only used that g, ∇L are continuous and R satisfies a Lipschitz condition on K_τ. From (11–13) there follows a.s.

$$\beta_n^{-1}(\gamma_1 W_1' + \cdots + \gamma_n W_n') \to 0, \tag{18}$$

$$\beta_n^{-1}(\gamma_1 W_1'' + \cdots + \gamma_n W_n'') \to 0, \tag{19}$$

$$\beta_n^{-1}(\gamma_1 \| W_1''\| + \cdots + \gamma_n \| W_n''\|) = O(1) \tag{20}$$

with maximum norm $\| \ \|$ on $\mathbb{R}^N$. From (9), (10), (14), (15), there follow (18) and

$$\beta_n^{-1}(\gamma_1 R(X_1) W_1'' + \cdots + \gamma_n R(X_n) W_n'') \to 0 \quad \text{a.s.}$$

by the above martingale argument. Because the latter relation is comprehended by (18) together with $R(x) \equiv 0$, it suffices to regard (17) under the conditions (18–20). Neglecting a set of probability measure zero there is used a pathwise consideration in the following, with the same notation for the realizations as for the random variables themselves.

There holds

$$\alpha_n W_n' = Z_n' + \alpha_n U_n'$$

where

$$U_n' := \beta_{n-1}^{-1}(\gamma_1 W_1' + \cdots + \gamma_{n-1} W_{n-1}'),$$
$$Z_n' := U_{n+1}' - U_n'.$$

Choose $\varepsilon \in (0, \max\{\| g(x)\| ; x \in K_\tau\})$ and $n_0 \in \mathbb{N}$ such that

$$\left\| \sum_{k=n}^{\infty} Z_k' \right\| < \frac{\varepsilon}{4} \qquad \text{for} \quad n \geqslant n_0.$$

For the sequence (X_n') defined by

$$X_n' := X_n, \qquad \text{for} \quad n = 1, \ldots, n_0,$$
$$X_{n+1}' := X_n' + t_n' \alpha_n [g(X_n) + U_n' - R(X_n) W_n''] \qquad \text{for} \quad n \geqslant n_0, \tag{21}$$

with maximum $t_n' \in [0, 1]$ such that $X_{n+1}' \in K_\tau$, one obtains

$$\| X_n' - X_n \| \leqslant \varepsilon \qquad \text{for} \quad n \in \mathbb{N}.$$

Let w_1 be the modulus of continuity of g, M_1 a Lipschitz constant of R on K_τ, $R_n := (R(X_n') - R(X_n)) W_n''$, and, by (20),

$$c_1 := M_1 \sup_n \beta_n^{-1} \sum_{k=1}^{n} \gamma_k \| W_k'' \|.$$

Noticing

$$\alpha_n R_n = Z_n'' + \alpha_n U_n''$$

with

$$U_n'' := \beta_{n-1}^{-1}(\gamma_1 R_1 + \cdots + \gamma_{n-1} R_{n-1}),$$
$$Z_n'' := U_{n+1}'' - U_n'',$$

one has

$$X'_{n+1} = X'_n + t'_n[\alpha_n(g(X'_n) - R(X'_n)\, W''_n + Y_n) + Z''_n]$$

with

$$Y_n = U'_n + U''_n - (g(X'_n) - g(X_n)),$$

$$\overline{\lim} \| Y_n \| \leqslant c_1 \varepsilon + w_1(\varepsilon), \qquad \left\| \sum_{k=1}^{n} Z''_k \right\| \leqslant c_1 \varepsilon \qquad (n \in \mathbb{N}).$$

If the sequence (X''_n) is defined by $X''_1 := X'_1$,

$$X''_{n+1} := X''_n + t''_n \alpha_n [g(X'_n) - R(X'_n)\, W''_n + Y_n] \qquad \text{for} \quad n \in \mathbb{N},$$

with maximum $t''_n \in [0, 1]$ such that $X''_{n+1} \in K_\tau$, there is obtained

$$\| X''_n - X'_n \| \leqslant 4 c_1 \varepsilon \qquad \text{for} \quad n \in \mathbb{N}.$$

By partial summation, (19), Lipschitz continuity of R on K_τ, (21), and (20), one obtains

$$U'''_n := \beta_n^{-1} \sum_{k=1}^{n} \gamma_k R(X'_k)\, W''_k \to 0 \qquad (n \to \infty).$$

Regarding

$$\alpha_n R(X'_n)\, W''_n = Z'''_n + \alpha_n U'''_n$$

with

$$Z'''_n := U'''_{n+1} - U'''_n,$$

one has

$$X''_{n+1} = X''_n + t''_n[\alpha_n(g(X''_n) + Y^*_n) - Z'''_n]$$

with

$$Y^*_n := -U'''_{n+1} + Y_n - (g(X''_n) - g(X'_n)),$$

$$\overline{\lim} \| Y^*_n \| \leqslant c_1 \varepsilon + w_1(\varepsilon) + w_1(4 c_1 \varepsilon).$$

Choosing $n_0^* \geqslant n_0$ such that $\| \sum_{k=n}^{\infty} Z'''_k \| < (\varepsilon/4)$ for $n \geqslant n_0^*$, there is defined

$$X'''_n := X''_n \qquad \text{for} \quad n = 1, \ldots, n_0^*,$$

$$X'''_{n+1} := X'''_n + t'''_n \alpha_n (g(X''_n) + Y^*_n) \qquad \text{for} \quad n \geqslant n_0^*,$$

with maximum t_n''' such that $X_{n+1}''' \in K_\tau$. Then $\| X_n''' - X_n'' \| \leqslant \varepsilon$,

$$X_{n+1}''' = X_n''' + t_n''' \alpha_n (g(X_n''') + Y_n^{**}) \qquad \text{for} \quad n \in \mathbb{N}, \tag{22}$$

where

$$\begin{aligned} &Y_n^{**} = Y_n^* - (g(X_n''') - g(X_n'')), \\ &\overline{\lim}\, \| Y_n^{**} \| \leqslant c_1 \varepsilon + 2 w_1(\varepsilon) + w_1(4 c_1 \varepsilon) =: w(\varepsilon), \\ &\| X_n''' - X_n \| \leqslant 2\varepsilon + 4 c_1 \varepsilon \qquad \text{for} \quad n \in \mathbb{N}. \end{aligned} \tag{23}$$

Now, with $c_0(\tau)$, $t_0(\tau)$ from (6), choose $\varepsilon^* > 0$ such that $w(\varepsilon^*) < (1/2)\, c_0(\tau)$ and

$$< \tfrac{1}{4} (\max \{ \| \nabla L(x) \| ; x \in K_\tau \})^{-1} \min \{ (\nabla L(x), g(x)); x \in K_\tau, \| g(x) \| \geqslant \varepsilon \},$$

further $n^* \in \mathbb{N}$ such that

$$\| Y_n^{**} \| < 2 w(\varepsilon^*), \qquad \alpha_n < t_0(\tau) \qquad \text{for} \quad n \geqslant n^*.$$

Thus $t_n''' = 1$ for $n \geqslant n^*$. By (22), a Taylor expansion for $L(X_{n+1}''')$ as usual in optimization theory and uniform continuity of ∇L on K_τ, one obtains, with suitable $n^{**} \geqslant n^*$,

$$\begin{aligned} L(X_{n+1}''') \geqslant L(X_n''') + \tfrac{1}{4} \alpha_n (\nabla L(X_n'''), g(X_n''')) \\ \text{for those } n \geqslant n^{**} \text{ with } \| g(X_n''') \| > \varepsilon. \end{aligned} \tag{24}$$

There holds

$$\begin{aligned} &h(\rho) := \sup \{ \lambda^* - L(x); x \in K_\tau, \| g(x) \| \leqslant \rho \} \to 0 \qquad (\rho \to +0), \\ &p(\sigma) := \sup \{ \| g(x) \| ; x \in K_\tau, \lambda^* - L(x) \leqslant \sigma \} \to 0 \qquad (\sigma \to +0). \end{aligned}$$

Relation (24) yields

$$\| g(X_n''') \| \leqslant \varepsilon \qquad \text{and} \qquad L(X_n''') \geqslant \lambda^* - h(\varepsilon) \qquad \text{infinitely often,}$$

by (3), (4), and $\Sigma \alpha_n = \infty$, but otherwise a monotonicity property, and thus, together with $L(X_{n+1}''') - L(X_n''') \to 0$, the relations

$$\underline{\lim} L(X_n''') \geqslant \lambda^* - h(\varepsilon), \qquad \overline{\lim}\, \| g(X_n''') \| \leqslant p(h(\varepsilon) + \varepsilon)$$

and thus, because of (23),

$$\overline{\lim}\, \| g(X_n) \| \leqslant w_1(2\varepsilon + 4 c_1 \varepsilon) + p(h(\varepsilon) + \varepsilon).$$

Therefore

$$g(X_n) \to 0, \qquad X_n \to x^*, \qquad L(X_n) \to \lambda^* \qquad (n \to \infty). \tag{25}$$

Finally,

$$\bar{L}_n \to \lambda^* \qquad (n \to \infty)$$

shall be shown, also by a pathwise consideration. There holds the representation

$$\bar{L}_n = \beta_n^{-1} \sum_{k=1}^{n} \gamma_k (L(X_k) + r(X_k)\, W_k'')$$

with W_k'' as in (17) and an $1 \times (N+2)$-matrix valued function r on K satisfying a Lipschitz condition on K_τ. Because of (25), it suffices to prove

$$\beta_n^{-1} \sum_{k=1}^{n} \gamma_k r(X_k)\, W_k'' \to 0 \qquad (n \to \infty),$$

but this follows from $r(X_n) - r(x^*) \to 0$, (19), (20). In the same way one obtains

$$\bar{A}_{nk} \to a_k^* \qquad (n \to \infty),\ k = 0, \ldots, N.$$

4. Rate of Convergence

For the estimation sequences (X_n), $(\bar{A}_{nk})(k = 0, \ldots, N)$, $(\bar{L}_n)$ the rate of convergence is investigated in the context of distributional convergence. Under the second order differentiability assumptions of Section 2, a central limit theorem with convergence order $n^{-1/4}$ for X_n is obtained, which in the more general form of an invariance principle (functional central limit theorem) yields a distributional limit theorem with convergence order $n^{-1/2}$ for $\bar{A}_{nk}$, $\bar{L}_n$ also given in the more general form of an invariance principle.

Besides the assumptions of Theorem 1 there is assumed

$$\alpha_n \cong a n^{-1}, \qquad \delta_n \cong d n^{-1/4} \qquad (a > 0,\ d > 0), \tag{26}$$

the stability condition

$$4a \min_{i=1,\ldots,N} \left| f''(x_i^*) - \sum_{k=0}^{N} a_k^* h_k''(x_i^*) \right| > 1, \tag{27}$$

further,

$$\sup_n E U_{ni}^2 < \infty, \tag{28'}$$

$$E(U_{ni} \mid \mathfrak{A}_n) = 0 \qquad (i = 0, \ldots, N+1;\ n \in \mathbb{N}), \tag{29'}$$

with $\mathfrak{A}_n$ as in Section 3, and the Lindeberg–type condition

$$\forall_{s>0} \frac{1}{n} \sum_{r=1}^{n} E(U_{ri}^2 \chi_{[U_{ri}^2 \geqslant sn]} | \mathfrak{A}_r) \xrightarrow{P} 0 \qquad (i=0,\ldots, N+1) \tag{30'}$$

and the analogous conditions (28″), (29″), (30″) for $V_{ni} (i=1,\ldots, N;\ n \in \mathbb{N})$. Assume that the arithmetic mean of the first n conditional covariance matrices of the $(2N+2)$-dimensional random vectors with coordinates $V_{ki}(-1)^{i+1} sgn\, \lambda^*(i=1,\ldots, N)$, $U_{kj}(j=0,\ldots, N+1)$ given $\mathfrak{A}_k$, $k=1,\ldots, n$, converges in probability for $n \to \infty$ to the covariance matrix

$$\begin{pmatrix} S_{\mathrm{I}} & S_{\mathrm{II}} \\ S_{\mathrm{III}} & S_{\mathrm{IV}} \end{pmatrix}, \tag{31}$$

where in an obvious notation S_{I} is an $N \times N$-matrix, S_{II} an $N \times (N+2)$-matrix, $S_{\mathrm{III}} = S_{\mathrm{II}}'$, S_{IV} an $(N+2) \times (N+2)$-matrix.

In this section, X_n, x^* are column vectors. There are used the following notations. Let $A := \operatorname{diag}\{\mu_1,\ldots, \mu_N\}$ with

$$\mu_i := \left| f''(x_i^*) - \sum_{k=0}^{N} a_k^* h_k''(x_i) \right| \qquad (i=1,\ldots, N),$$

and $H_0,\ldots, H_{N+1}$ be the Hessians of $A_0,\ldots, A_N$, L, respectively, at x^*. Further let ζ be an $(2N+2)$-dimensional Brownian motion with the N-dimensional component $\zeta^* = (\zeta_1,\ldots, \zeta_N)'$ and the $(N+2)$-dimensional component ζ^{**}, with $\zeta(0)=0$, $E\zeta(1)=0$ and covariance matrix

$$S := \begin{pmatrix} S_{\mathrm{I}} & S_{\mathrm{II}} M(x^*)' \\ M(x^*) S_{\mathrm{III}} & M(x^*) S_{\mathrm{IV}} M(x^*)' \end{pmatrix}$$

of $\zeta(1)$, and $G = (G_1,\ldots, G_N)'$ be the $N-$dimensional Gaussian Markov process with

$$G_i(0)=0,\ G_i(t) = t^{-a\mu_i + 3/4} \int_{(0,t]} v^{a\mu_i - 3/4}\, d\zeta_i(v),\ t \in (0,1] \qquad (i=1,\ldots, N).$$

Now the following invariance principle (functional limit theorem) can be formulated.

THEOREM 2. *Let the assumptions of Theorem* 1 *together with conditions* (26)–(31) *hold. Then the sequence of random elements* (Z_n) *in* $C_{\mathbb{R}^{2N+2}}[0,1]$ *with maximum norm which are defined by*

$$Z_n(t) := n^{-1/2} R_{[nt]} + (nt - [nt])\, n^{-1/2} (R_{[nt]+1} - R_{[nt]}), \qquad t \in [0,1],$$

with

$$R_n := \begin{pmatrix} n^{3/4}(X_n - x^*) \\ n(\bar{A}_{n0} - a_0^*, \ldots, \bar{A}_{nN} - a_N^*, \bar{L}_n - \lambda^*)' \end{pmatrix}$$

converges in distribution to

$$\begin{pmatrix} G \\ \int_0^{\cdot} t^{-3/2}(G(t)' H_0 G(t)), \ldots, G(t)' H_{N+1} G(t))' \, dt - \zeta^{**} \end{pmatrix}.$$

Proof of Theorem 2. Noticing (16), (26), (27), (29′), (29″), one can take recursion (17), which together with $sgn\, \lambda^* = 1$ is used without loss of generality, into the form

$$X_{n+1} - x^* = X_n - x^* - n^{-1} A_n (X_n - x^*) + n^{-3/4} D_n W_n + n^{-5/4} T_n$$

with $N-$dimensional random column vectors W_n, T_n, random $N \times N$-matrices $A_n \to aA$ a.s., $D_n := (1/2)\, n^{-3/4}\, \alpha_n/\delta_n \to (\alpha/2d)$, $T_n \to 0$ a.s., $a\mu_i > \frac{1}{4}$ $(i = 1, \ldots, N)$, $E(W_n | \mathfrak{A}_n) = 0$. Let

$$U_n := (U_{n0}, \ldots, U_{n,N+1})', \qquad \tilde{U}_n := M(X_n)\, U_n,$$

$$\tilde{Y}_1 := 0, \qquad \tilde{Y}_{n+1} := (n+1)^{1/4}\, n^{-1}(\tilde{U}_1 + \cdots + \tilde{U}_n) \qquad \text{for} \quad n \in \mathbb{N}.$$

Then the above recursion can be supplemented by the recursion

$$\tilde{Y}_{n+1} = \tilde{Y}_n - n^{-1}\,(\tfrac{3}{4} + o(1))\, \tilde{Y}_n + n^{-3/4}(1 + o(1))\, \tilde{U}_n.$$

For the $(2N+2)$-dimensional random vectors $\bar{W}_n$ with components W_n and $\tilde{U}_n$, there holds

$$\sup E \| \bar{W}_n \|^2 < \infty, \qquad E(\bar{W}_n | \mathfrak{A}_n) = 0 \qquad (n \in \mathbb{N})$$

because of (28′), (28″), $E(W_n | \mathfrak{A}_n) = 0$, (29′), and boundedness of M on K_τ; further, there holds a Lindeberg–type condition and that the arithmetic mean of the first n conditional convariance matrices of $\bar{W}_k$ given $\mathfrak{A}_k$, $k = 1, \ldots, n$, converges in probability to S, because of (30′), (30″), and (31), respectively, and $M(X_n) \to M(x^*)$ a.s.

Let $\bar{Z}_n$ be defined analogously to Z_n, but with the $(2N+2)$-dimensional random vector $\bar{R}_n$ having components $n^{3/4}\,(X_n - x^*)$ and $n^{3/4}\,\tilde{Y}_n$ instead of R_n. Now an invariance principle of Berger [1] and Pantel [11, Sect. 8.3] which generalizes results of Fabian [4] and Walk [15] on recursive schemes can be applied and yields convergence of $(\bar{Z}_n)$ in distribution to the random element in $C_{\mathbb{R}^{2N+2}}[0, 1]$ with components G and ζ^{**}. (For an

easily accessible formulation of Berger's result [with a slightly differing assumption on $\bar{W}_n$] see Nixdorf [10, pp. 41–43].)

Finally, from the convergence result on $(\bar{Z}_n)$, one obtains the assertion on (Z_n) regarding (5), by use of Skorokhod's representation theorem (see, e.g., [2, Sect. 3]) which allows one to argue for a.s. convergence instead of distributional convergence.

Remark. For Theorem 2 the evaluation functional obtained by inserting $t=1$ yields distributional convergence of $n^{-1/2}R_n$ to an $(2N+2)$-dimensional random vector and thus the orders of convergence mentioned above. [In the case $f\in C^3[0,1]$ with $\delta_n \cong dn^{-1/6}$, factor 3 instead of 4 in (27), in a similar manner one obtains a functional central limit theorem with convergence order $n^{-1/3}$ for X_n and $n^{-1/2}$ for $(\bar{A}_{n0},..., \bar{A}_{nN}, \bar{L}_n)$, where in the latter case for the limit process no integral term appears.] The reason that for the estimation of the optimal vector $(a_0^*,..., a_N^*, \lambda^*)$ in the Tchebycheff approximation of $f\in C^2[0,1]$ convergence order $n^{-1/2}$ is achieved differently to convergence orders in stochastic iteration for other optimization problems lies in the validity of (5), and in the fact that in the auxiliary linear discrete Tchebycheff approximation problem given by (1) the observation errors U_{ni} $(i=0,..., N+1)$ are not endowed with a convergence rate diminishing factor δ_n^{-1} as the V_{ni} $(i=1,..., N)$.

REFERENCES

1. E. BERGER, Asymptotic behaviour of a class of stochastic approximation procedures, *Probab. Theory Relat. Fields*, in press.
2. P. BILLINGSLEY, "Weak Convergence of Measures: Applications in Probability," Regional Conference Series in Applied Mathematics Vol. 5, SIAM, Philadelphia, 1971.
3. H. CRAMÉR AND M. R. LEADBETTER, "Stationary and Related Stochastic Processes," Wiley, New York, 1967.
4. V. FABIAN, On asymptotic normality in stochastic approximaton, *Ann. Math. Statist.* **39** (1968), 1327–1332.
5. L. LJUNG, Strong convergence of a stochastic approximation algorithm, *Ann. Statist.* **6** (1978), 680–696.
6. M. LOÈVE, "Probability Theory II," 4th ed., Springer-Verlag, New York/Berlin, 1978.
7. G. MEINARDUS, Über Tschebyscheffsche Approximationen, *Arch. Rational Mech. Anal.* **9** (1962), 329–351.
8. G. MEINARDUS, "Approximation of Functions: Theory and Numerical Methods," Springer-Verlag, Berlin/New York, 1967.
9. M. B. NEVEL'SON AND R. Z. HAS'MINSKII, An adaptive Robbins–Monro procedure, *Autom. Rem. Contr.* **34** (1973), 1594–1607.
10. R. NIXDORF, "Stochastische Approximation in Hilberträumen durch endlichdimensionale Verfahren," Mitteilungen Mathem. Seminar Giessen, Vol. 154, 1982.
11. M. PANTEL, "Adaptive Verfahren der stochastischen Approximation," Dissertation, Universität Essen, 1979.

12. P. RÉVÉSZ, "The Laws of Large Numbers," Academic Press, New York, 1968.
13. L. SCHMETTERER, Multidimensional stochastic approximation, *in* "Multivariate Analysis II" (P. R. Krishnaiah, ed.), pp. 443–460, Academic Press, New York, 1969.
14. J. H. VENTER, An extension of the Robbins–Monro procedure, *Ann. Math. Statist.* **38** (1967), 181–190.
15. H. WALK, An invariance principle for the Robbins–Monro process in a Hilbert space, *Z. Wahrsch. Verw. Gebiete* **39** (1977), 135–150.