ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XII, 2 (1971)

S. PASZKOWSKI (Wrocław)

OPTIMUM CHOICE OF INITIAL APPROXIMATIONS IN INTERPOLATION METHODS OF SOLVING EQUATIONS

0. The equation $f(x) = 0$ may be solved by interpolation methods which are described in detail by Traub ([2], chapter 4). The investigation of the convergence of the approximations obtained by these methods leads to sequences of numbers satisfying certain recurrent inequalities. Such sequences are considered in Section 1. The obtained results are in Section 2 applied to the problem of optimum choice of initial approximations in interpolation methods. It appears that this problem is a generalization of the known problem of the polynomial $x^n + a_1 x^{n-1} + \ldots + a_n$, having its minimal maximum absolute value in the interval $\langle -1, 1\rangle$. That generalization is investigated in Section 3. For the most simple interpolation methods the optimum initial approximations are found.

1. Boundedness and convergence of the sequences satisfying recurrent inequalities. In this section we deal with the sequence $\{d_i\}$ of non-negative numbers satisfying the inequalities

$$(1) \qquad d_i \leqslant \prod_{j=1}^{n} d_{i-j}^{\gamma_j} \qquad (i = n, n+1, \ldots),$$

where $\gamma_1, \gamma_2, \ldots, \gamma_n$ are fixed non-negative numbers and where $\gamma_n \neq 0$. The polynomial

$$(2) \qquad \Gamma(x) = x^n - (\gamma_1 x^{n-1} + \gamma_2 x^{n-2} + \ldots + \gamma_n)$$

has thus exactly one positive zero ξ. Let

$$(3) \qquad \Gamma(x) = (x-\xi)(b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-1}) \qquad (b_0 = 1).$$

In the first part of this section we shall formulate the conditions ensuring that the sequence $\{d_i\}$ is limited in such a way that

$$d_i \leqslant \max\{d_0, d_1, \ldots, d_{n-1}\} \qquad (i = n, n+1, \ldots).$$

THEOREM 1. *If the numbers* $d_0, d_1, \ldots, d_{n-1}$ *satisfy the inequality*

$$\prod_{h=0}^{n-1} d_{n-1-h}^{b_h} < 1, \tag{4}$$

then the successive elements of $\{d_i\}$ *satisfy similar inequalities*

$$\prod_{h=0}^{n-1} d_{i-h}^{b_h} < 1 \quad (i = n, n+1, \ldots). \tag{5}$$

Proof by induction. It is sufficient to show that from the assumptions it follows that

$$\prod_{h=0}^{n-1} d_{n-h}^{b_h} < 1. \tag{6}$$

From (4) and also from (1) for $i = n$ follows

$$\prod_{h=0}^{n-1} d_{n-h}^{b_h} = d_n \prod_{h=1}^{n-1} d_{n-h}^{b_h} \leqslant \Big(\prod_{j=1}^{n} d_{n-j}^{\gamma_j}\Big) \prod_{h=1}^{n-1} d_{n-n}^{b_h} = d_0^{\gamma_n} \prod_{h=1}^{n-1} d_{n-h}^{\gamma_h + b_h}. \tag{7}$$

From (3) and (2) it follows that

$$\gamma_h = \xi b_{h-1} - b_h \quad (h = 1, 2, \ldots, n-1), \qquad \gamma_n = \xi b_{n-1}.$$

Therefore inequality (7) may be written in the form

$$\prod_{h=0}^{n-1} d_{n-h}^{b_h} \leqslant d_0^{\xi b_{n-1}} \prod_{h=1}^{n-1} d_{n-h}^{\xi b_{h-1}} = \Big(\prod_{h=1}^{n} d_{n-h}^{b_{h-1}}\Big)^{\xi} = \Big(\prod_{h=0}^{n-1} d_{n-1-h}^{b_h}\Big)^{\xi}.$$

Hence (6) follows from (4).

THEOREM 2. *If* $\xi \geqslant 1$, *if the coefficients of the polynomial* $\Gamma(x)/(x-\xi)$ *satisfy the inequality*

$$1 \geqslant b_1 \geqslant b_2 \geqslant \ldots \geqslant b_{n-1} > 0, \tag{8}$$

and if inequality (4) *is satisfied, then*

$$d_i \leqslant \max\{d_0, d_1, \ldots, d_{n-1}\} \quad (i = n, n+1, \ldots). \tag{9}$$

Proof. First, we shall prove inequality (9) for $i = n$. Since from (1) it follows that

$$d_n \leqslant \prod_{j=1}^{n} d_{n-j}^{\gamma_j},$$

it suffices to prove that

$$\prod_{j=1}^{n} d_{n-j}^{\gamma_j} \leqslant \max\{d_0, d_1, \ldots, d_{n-1}\}.$$

If the right-hand side expression is equal to d_k $(0 \leqslant k \leqslant n-1)$, this inequality may be stated in the form

$$(10) \qquad d_k^{\gamma_{n-k}-1} \prod_{j=1, j\neq n-k}^{n} d_{n-j}^{\gamma_j} \leqslant 1.$$

The sum of the exponents at the left-hand side equals $\gamma_1+\gamma_2+\ldots+\gamma_n-1$, i.e., in view of (2) and (3), equals

$$(11) \qquad -\Gamma(1) = (\xi-1)(b_0+b_1+\ldots+b_{n-1}).$$

Let us raise both sides of (4) to the power $\xi-1$. Since $\xi-1 \geqslant 0$, from (4) we have

$$(12) \qquad \prod_{h=0}^{n-1} d_{n-1-h}^{(\xi-1)b_h} = \prod_{j=1}^{n} d_{n-j}^{(\xi-1)b_{j-1}} \leqslant 1.$$

Here also the sum of the exponents is equal to (11). In (12) occurs, among others, the factor

$$(13) \qquad d_k^{(\xi-1)b_{n-k-1}} = d_k^{\gamma_1+\gamma_2+\ldots+\gamma_n-1-(\xi-1)(b_0+\ldots+b_{n-k-2}+b_{n-k}+\ldots+b_{n-1})}$$
$$= d_k^{\gamma_{n-k}-1} \prod_{j=1, j\neq n-k}^{n} d_k^{\gamma_j-(\xi-1)b_{j-1}}.$$

The exponents $\gamma_j-(\xi-1)b_{j-1}$ are non-negative. In fact, assumption (8) gives

$$\gamma_j-(\xi-1)b_{j-1} = \xi b_{j-1}-b_j-(\xi-1)b_{j-1} = b_{j-1}-b_j \geqslant 0.$$

But we assume that $d_k \geqslant d_0, d_1, \ldots, d_{n-1}$; therefore, if the factor $d_k^{\gamma_j-(\xi-1)b_{j-1}}$ at the right-hand side of (13) is changed into $d_{n-j}^{\gamma_j-(\xi-1)b_{j-1}}$, then the whole expression will not increase.

Substituting the changed expression in place of $d_k^{(\xi-1)b_{n-k-1}}$ into (12), we obtain inequality (10), q.e.d.

The proof of inequality (9) for $i = n+1, n+2, \ldots$ is the same as the proof for $i = n$ because theorem 1 justifies the use of inequalities (5), analogous to (4).

Example 1. If $\gamma_1 = \gamma_2 = \ldots = \gamma_n > 1/n$, then $\xi > 1$ and inequality (8) is satisfied (see Traub [2], p. 50-51). Therefore, from (4) the assertion (9) of theorem 2 follows.

In the second part of this section we shall formulate the conditions asserting the convergence of $\{d_i\}$ to 0.

THEOREM 3. *The elements d_i satisfy*

$$(14) \qquad d_i \leqslant \prod_{h=0}^{n-1} d_h^{p_{hi}} \quad (i = 0, 1, \ldots),$$

where, for every $h = 0, 1, \ldots, n-1$, the numbers p_{hi} are the solution of the difference equation

$$p_{hi} - \sum_{j=1}^{n} \gamma_j p_{h,i-j} = 0 \quad (i = n, n+1, \ldots) \tag{15}$$

with initial conditions

$$p_{hi} = \delta_{hi} \quad (i = 0, 1, \ldots, n-1). \tag{16}$$

Proof. For $i = 0, 1, \ldots, n-1$ inequalities (14) are obvious, as is seen from the initial conditions (16). Thus it suffices to show that if for any fixed $i \geqslant n$ we have

$$d_{i-j} \leqslant \prod_{h=0}^{n-1} d_h^{p_{h,i-j}} \quad (j = 1, 2, \ldots, n), \tag{17}$$

then (14) is satisfied. That fact follows, however, from (1) and (17). Indeed, we have

$$d_i \leqslant \prod_{j=1}^{n} d_{i-j}^{\gamma_j} \leqslant \prod_{j=1}^{n} \Big(\prod_{h=0}^{n-1} d_h^{p_{h,i-j}}\Big)^{\gamma_j}.$$

The exponent of d_h at the right-hand side is equal to $\sum_{j=1}^{n} \gamma_j p_{h,i-j}$, i.e. to p_{hi}.

The difference equations (15) have only a different notation of the unknowns. Generally, we shall write them in the form

$$P_i - \sum_{j=1}^{n} \gamma_j P_{i-j} = 0 \quad (i = n, n+1, \ldots). \tag{18}$$

The characteristic polynomial (2) is connected with this difference equation. Let $\Gamma(x)$ have, besides of the positive zero ξ, the zeros $\xi_1, \xi_2, \ldots, \xi_k$ with multiplicity equal to $m_1, m_2, \ldots, m_k$ (where $m_1 + m_2 + \ldots + m_k = n-1$), respectively. Since $\gamma_n > 0$, we have $\xi_1, \xi_2, \ldots, \xi_k \neq 0$. It is known that every solution of (18) is given by the formula

$$P_i = C\xi^i + \sum_{j=1}^{k} Q_j(i)\, \xi_j^i,$$

where C is a constant and $Q_j(i)$ is a polynomial of the variable i, of degree at most $m_j - 1$.

Theorem 4. *The solutions p_{hi} of the difference equation* (15) *with initial conditions* (16) *are given by*

$$p_{hi} = c_h \xi^i + \sum_{j=1}^{k} q_{hj}(i)\, \xi_j^i \quad (h = 0, 1, \ldots, n-1;\ i = 0, 1, \ldots), \tag{19}$$

where $q_{hj}(i)$ is a polynomial of the variable i, of degree at most m_j-1, and where

$$c_h = b_{n-1-h}/\Gamma'(\xi) \qquad (h = 0, 1, \ldots, n-1). \tag{20}$$

Proof. After what was said before stating Theorem 4, it suffices to prove formulae (20).

The initial conditions (16) lead to the system of equations

$$c_h \xi^i + \sum_{j=1}^{k} q_{hj}(i)\, \xi_j^i = \delta_{hi} \qquad (i = 0, 1, \ldots, n-1). \tag{21}$$

Let every polynomial $q_{hj}(i)$ be represented as a linear combination of factor polynomials $a_0 + a_1 i + a_2 i(i-1) + \ldots$, where, obviously, the coefficients $a_0, a_1, a_2, \ldots$ depend upon h and j.

Treating these coefficients and c_h as unknowns in (21), we obtain from Cramer's formulae the equality

$$c_h = \det(E_h, \Xi_1, \Xi_2, \ldots, \Xi_k)/\det(\Xi, \Xi_1, \Xi_2, \ldots, \Xi_k). \tag{22}$$

The symbol „det" denotes here a determinant of the matrix composed of the blocks given in parentheses. E_h is the $(h+1)$-st column of the unit matrix of order n, and

$$\Xi_j = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \xi_j & \xi_j & \ldots & 0 \\ \xi_j^2 & 2\xi_j^2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \xi_j^{n-1} & (n-1)\xi_j^{n-1} & \ldots & (n-1)\ldots(n-m_j+1)\xi_j^{n-1} \end{bmatrix}, \quad \Xi = \begin{bmatrix} 1 \\ \xi \\ \xi^2 \\ \cdots \\ \xi^{n-1} \end{bmatrix}.$$

Let us introduce a column matrix X, analogous to Ξ. From (22) it follows that

$$\begin{aligned} &c_0 + c_1 x + \ldots + c_{n-1} x^{n-1} \\ &\qquad = \det(X, \Xi_1, \Xi_2, \ldots, \Xi_k)/\det(\Xi, \Xi_1, \Xi_2, \ldots, \Xi_k). \end{aligned} \tag{23}$$

This equality will remain valid if in every matrix Ξ_j the columns $2, 3, \ldots, m_j$ will be divided by $\xi_j, \xi_j^2, \ldots, \xi_j^{m_j-1}$, respectively. Then

$$\det(X, \Xi_1, \Xi_2, \ldots, \Xi_k) \tag{24}$$

will be a polynomial of degree $n-1$ of the variable x which has zeros $\xi_1, \xi_2, \ldots, \xi_k$ with multiplicity equal to $m_1, m_2, \ldots, m_k$, respectively. This may be seen by differentiating (24) 0, 1, 2, ... times and by substituting $x = \xi_j$. Thus, with an accuracy up to a non-zero constant factor, this polynomial is equal to

$$\Gamma(x)/(x-\xi). \tag{25}$$

The denominator at the right-hand side of (23) is equal to the value of (25) for $x = \xi$, i.e. is equal to $\Gamma'(\xi)$. Hence

$$c_0+c_1x+\ldots+c_{n-1}x^{n-1} = (b_0x^{n-1}+b_1x^{n-2}+\ldots+b_{n-1})/\Gamma'(\xi),$$

q.e.d.

THEOREM 5. *If the roots of the polynomial $\Gamma(x)$ satisfy the inequalities*

$$\xi > 1, \tag{26}$$

$$|\xi_j| < \xi \quad (j = 1, 2, \ldots, k) \tag{27}$$

and if (4) *is satisfied, then the sequence $\{d_i\}$ is convergent to* 0.

Proof. From (19) it follows that

$$p_{hi}-c_h\xi^i = \xi^i\sum_{j=1}^{k} q_{hj}(i)(\xi_j/\xi)^i.$$

From assumption (27) there exists for every $\varepsilon > 0$ an index i_0 such that for $i \geqslant i_0$ we have

$$|p_{hi}-c_h\xi^i| \leqslant \varepsilon\xi^i. \tag{28}$$

Introduce now for positive numbers d the notation

$$d^* = \operatorname{sgn}\log d = \begin{cases} 1 & (d > 1), \\ 0 & (d = 1), \\ -1 & (d < 1). \end{cases}$$

For any r and s such that $s \geqslant |r|$ we have

$$d^r \leqslant d^{d^*|r|} \leqslant d^{d^*s}.$$

Herefrom and from (28) it follows that

$$d_h^{p_{hi}} = d_h^{c_h\xi^i}d_h^{p_{hi}-c_h\xi^i} \leqslant d_h^{c_h\xi^i}d_h^{\overset{*}{d_h}|p_{hi}-c_h\xi^i|} \leqslant d_h^{(c_h+\varepsilon \overset{*}{d_h})\xi^i}.$$

Thus, from (14) and (20) we obtain

$$d_i \leqslant \prod_{h=0}^{n-1} d_h^{(c_h+\varepsilon \overset{*}{d_h})\xi^i} = \prod_{h=0}^{n-1} d_h^{(b_{n-1-h}/\Gamma'(\xi)+\varepsilon \overset{*}{d_h})\xi^i} \tag{29}$$

$$= \Big(\prod_{h=0}^{n-1} d_{n-1-h}^{b_h+\varepsilon \overset{*}{d}_{n-1-h}\Gamma'(\xi)}\Big)^{\xi^i/\Gamma'(\xi)}.$$

If ε is sufficiently small, then from (4) it also follows that

$$\prod_{h=0}^{n-1} d_{n-1-h}^{b_h+\varepsilon d^*_{n-1-h}\Gamma'(\xi)} < 1.$$

Since $\xi > 1$ (assumption (26)) and $\Gamma'(\xi) > 0$, the exponent $\xi^i/\Gamma'(\xi)$ tends to $+\infty$ and the last member of (29) tends to 0. Hence $\{d_i\} \to 0$.

Let us note that from Traub's lemma 3.1 ([2], p. 38) it follows that $\{d_i\}$ tends to 0 under the following assumptions: first, if the numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ are non-negative integers such that

$$(30) \qquad \gamma_1+\gamma_2+\ldots+\gamma_n > 1$$

(i.e. such that (26) holds), and second, if

$$(31) \qquad d_i < 1 \quad (i = 0, 1, \ldots, n-1).$$

Theorem 5 assumes, moreover, that the coefficients γ_j satisfy (27) (however, see Example 2). On the other side, inequality (4), which connects the initial elements $d_0, d_1, \ldots, d_{n-1}$, is weaker than (31).

Example 2. If the numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ are non-negative and if the greatest common divisor of the indices of all positive numbers γ_j is equal to 1, then inequality (27) holds (Ostrowski [1], p. 93, theorem 12.2). As at least two of the numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ are positive, thus, if they all are integers, inequalities (30) and (26) hold. Therefore in the now considered case inequality (4) guarantees the convergence of the sequence $\{d_i\}$ to 0.

Example 3. Let the sequence $\{d_i\}$ of non-negative numbers satisfy the inequality $d_i \leqslant d_{i-1}d_{i-2}$. We then have now $n = 2$, $\gamma_1 = \gamma_2 = 1$,

$$\Gamma(x) = x^2-x-1, \qquad \xi = \tfrac{1}{2}(1+\sqrt{5}), \qquad \Gamma(x)/(x-\xi) = x+\tfrac{1}{2}(\sqrt{5}-1).$$

Directly from that or from the remarks made in Examples 1 and 2 it follows that if $d_1 d_0^{(\sqrt{5}-1)/2} < 1$, then the sequence $\{d_i\}$ is bounded in such a way that $d_i \leqslant \max\{d_0, d_1\}$, and also this sequence converges to 0.

2. Interpolation methods of solving equations. There exist many methods of iterative computation of the root a of the equation $f(x) = 0$. Among them there are interpolation methods ([2], Chapter 4) in which the approximation x_i of the root a is directly expressed by the n preceding approximations $x_{i-1}, x_{i-2}, \ldots, x_{i-n}$. In one of the variants of interpolation methods it is assumed that $x_i = P(0)$, where the polynomial $P(x)$ of degree at most $n-1$ satisfies the conditions

$$P\big(f(x_{i-j})\big) = x_{i-j} \quad (j = 1, 2, \ldots, n).$$

In another variant, the polynomial $P(x)$ is determined by the conditions

$$P(x_{i-j}) = f(x_{i-j}) \qquad (j = 1, 2, \ldots, n)$$

and as x_i one of the zeros of this polynomial is chosen (in particular, for $n = 2$ we have here the so-called Müller's method of equations solving). The behaviour of the successive approximations x_i of the root a may be in interpolation methods characterized as follows. Let δ_i denote the absolute error of the approximation x_i:

$$\delta_i = |a - x_i|.$$

If the approximations $x_0, x_1, \ldots, x_{n-1}$ belong to the interval $I = \langle a-r, a+r\rangle$ in which $f'(x) \neq 0$, then the inequality

$$\delta_n \leqslant M \prod_{j=1}^{n} \delta_{n-j}^{\gamma_j} \tag{32}$$

holds. M is here a positive number depending upon the values of $f(x)$ and of its initial derivatives in the interval I, and γ_j are non-negative integers with $\gamma_n > 0$.

Example 4. For the linear interpolation method given by

$$x_i = x_{i-1} - f(x_{i-1}) \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})} \tag{33}$$

inequality (32) is satisfied for

$$n = 2, \quad \gamma_1 = \gamma_2 = 1, \quad M = \frac{1}{2} \max_{x \in I} \left| \frac{f''(x)}{\big(f'(x)\big)^3} \right| \max_{x \in I} |f'(x)|^2.$$

Assume that

$$q = \sum_{j=1}^{n} \gamma_j > 1.$$

If

$$d_i = M^{1/(q-1)} \delta_i, \tag{34}$$

inequality (32) may be expressed as (1).

Assume that $d_0, d_1, \ldots, d_{n-1}$, given by (34), satisfy inequality (4). Coming back to the errors δ_i, from (4) the inequality

$$\prod_{h=0}^{n-1} (M^{1/(q-1)} \delta_{n-1-h})^{b_h} < 1,$$

may be obtained or, otherwise,

$$\prod_{h=0}^{n-1} \delta_{n-1-h}^{b_h} < M^{-(b_0+b_1+\ldots+b_{n-1})/(q-1)} = M^{-1/(\xi-1)}. \tag{35}$$

Let us assume that the assumptions about the numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ from Section 1 are satisfied. It is so, for instance, for methods in which $\gamma_1 = \gamma_2 = \ldots = \gamma_n$ is a natural number (see Examples 1 and 2). If so, from Theorem 2 and its proof it follows that the approximations x_n, $x_{n+1}, \ldots$ belong to the interval I, and their errors satisfy the inequalities

$$\delta_i \leqslant M \prod_{j=1}^{n} \delta_{i-j}^{\gamma_j} \qquad (i = n, n+1, \ldots),$$

which are similar to (32). It follows from Theorem 5 that the sequence $\{\delta_i\}$ is convergent to 0, i.e. the sequence $\{x_i\}$ is convergent to a.

The quality of the convergence of $\{x_i\}$ depends not only upon ξ (i.e. upon the order of the method), thus indirectly upon the numbers n, $\gamma_1, \gamma_2, \ldots, \gamma_n$, but also upon the choice of initial approximations. Taking Theorem 5 and its proof into account, one may say that the convergence is the better the less the product

$$\prod_{h=0}^{n-1} |a - x_{n-1-h}|^{b_h} = \prod_{h=0}^{n-1} |a - x_h|^{b_{n-1-h}}.$$

From this the following practical hint for the choice of approximations $x_0, x_1, \ldots, x_{n-1}$ follows: they are to be chosen from the interval I containing the root a so as to minimize the expression

$$\max_{a \in I} \prod_{h=0}^{n-1} |a - x_h|^{b_{n-1-h}}. \tag{36}$$

3. Minimization of the expression (36). The search for optimum initial approximations, i.e. such ones for which the expression (36) reaches its minimum, is a generalization of the problem the solution of which are Chebyshev polynomials. In fact, if $b_0 = b_1 = \ldots = b_{n-1} = 1$, then (36) becomes

$$\max_{a \in I} \left| \prod_{h=0}^{n-1} (a - x_h) \right|. \tag{37}$$

If $I = \langle -1, 1 \rangle$, then (37) is minimum if and only if

$$\prod_{h=0}^{n=1} (a - x_h) = 2^{-(n-1)} T_n(a),$$

i.e. if the set $\{x_0, x_1, \ldots, x_{n-1}\}$ is composed of the (arbitrarily ordered) numbers

$$\cos \frac{(2i+1)\pi}{2n} \qquad (i = 0, 1, \ldots, n-1).$$

One may expect that — as in the case of (37) — the expression (36) attains its minimum if and only if the function

$$A(a)=\prod_{h=0}^{n-1}|a-x_h|^{b_{n-1-h}}$$

of variable a has equal values, being the maximum value of $A(a)$ in I, in $n+1$ points of the interval I. This hypothesis is true if we speak of local minimums of (36).

THEOREM 6. *Function* (36) *of the variables* $x_0, x_1, \ldots, x_{n-1}$ *belonging to the interval* I *reaches its local minimums only for such values of these variables for which the function* $A(a)$ *has in* $n+1$ *points of the interval* I *identical values, being equal to the maximum value in the interval* I.

Proof. For the sake of simplicity we shall limit ourselves to the case $x_0<x_1<\ldots<x_{n-1}$ (any other ordering of these variables would cause a different numbering of the exponents $b_0, b_1, \ldots, b_{n-1}$). Since

$$\frac{A'(a)}{A(a)}=\sum_{h=0}^{n-1}\frac{b_{n-1-h}}{a-x_h},\quad \left(\frac{A'(a)}{A(a)}\right)'=-\sum_{h=0}^{n-1}\frac{b_{n-1-h}}{(a-x_h)^2}, \tag{38}$$

the function $A(a)$ decreases from ∞ to 0 in the interval $(-\infty, x_0\rangle$, increases from 0 to ∞ in the interval $\langle x_{n-1}, \infty)$, and has exactly one maximum in every of the intervals

$$\langle x_0, x_1\rangle, \langle x_1, x_2\rangle, \ldots, \langle x_{n-2}, x_{n-1}\rangle.$$

The points in which $A(a)$ reaches the maximum values we denote by $y_1, y_2, \ldots, y_{n-1}$, respectively. To standardize the notation let us assume that $I=\langle y_0, y_n\rangle$ (the function $A(a)$, given in I, has also local maxima at the ends of this interval).

We shall investigate now how the maximum values

$$e_k=A(y_k)\qquad (k=0,1,\ldots,n) \tag{39}$$

depend upon the variables $x_0, x_1, \ldots, x_{n-1}$. Of course, also the points $y_1, y_2, \ldots, y_{n-1}$ depend upon them; only y_0 and y_n are constant. We shall use differentials. We have

$$e_0+de_0=\prod_{h=0}^{n-1}|y_0-x_h-dx_h|^{b_{n-1-h}}=e_0\prod_{h=0}^{n-1}\left|1-\frac{dx_h}{y_0-x_h}\right|^{b_{n-1-h}}$$

$$=e_0\left(1-\sum_{h=0}^{n-1}\frac{b_{n-1-h}dx_h}{y_0-x_h}\right);$$

thus

$$de_0 = -e_0 \sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_0 - x_h}.$$

Analogously,

$$de_n = -e_n \sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_n - x_h}.$$

For $1 \leqslant k \leqslant n-1$ we have

$$e_k + de_k = A(y_k + dy_k) = e_k \prod_{h=0}^{n-1} \left| 1 + \frac{dy_k - dx_h}{y_k - x_h} \right|^{b_{n-1-h}}$$

$$= e_k \left(1 + dy_k \sum_{h=0}^{n-1} \frac{b_{n-1-h}}{y_k - x_h} - \sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_k - x_h} \right).$$

Since $A'(y_k) = 0$, from the first formula of (38) it follows that the first sum equals zero, and

$$de_k = -e_k \sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_k - x_h};$$

therefore this formula is satisfied for all $k = 0, 1, \ldots, n$.

First suppose that not all numbers $e_0, e_1, \ldots, e_n$ are identical. Let the numbers $e_{l_0}, e_{l_1}, \ldots, e_{l_m}$, where $m < n$, be all numbers (39) being equal to $\max\limits_{0 \leqslant k \leqslant n} e_k$. It is then possible to choose the n increments dx_0, $dx_1, \ldots, dx_{n-1}$ as to have $de_{l_0}, de_{l_1}, \ldots, de_{l_m}$ negative, i.e. as to diminish $\max\limits_{0 \leqslant k \leqslant n} e_k$. In that case the parameters $x_0, x_1, \ldots, x_{n-1}$ certainly do not yield the local minimum of (36).

Now, assume that $e_0 = e_1 = \ldots = e_n$. We shall prove that it is not possible to choose $dx_0, dx_1, \ldots, dx_{n-1}$, not all equal zero, as to have $de_k \leqslant 0$, i.e.

$$\sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_k - x_h} \geqslant 0 \quad (k = 0, 1, \ldots, n). \tag{40}$$

The left-hand side of (40) will be denoted by λ_k. The system

$$\sum_{h=0}^{n-1} \frac{b_{n-1-h} dx_h}{y_k - x_h} - \lambda_k = 0 \quad (k = 0, 1, \ldots, n)$$

is interpreted as a system of $n+1$ homogeneous equations with $n+1$ unknowns $b_{n-1}dx_0, b_{n-2}dx_1, \ldots, b_0dx_{n-1}, -1$. The determinant of the systems equals zero:

$$(41)\qquad \begin{vmatrix} \frac{1}{y_0-x_0} & \frac{1}{y_0-x_1} & \ldots & \frac{1}{y_0-x_{n-1}} & \lambda_0 \\ \frac{1}{y_1-x_0} & \frac{1}{y_1-x_1} & \ldots & \frac{1}{y_1-x_{n-1}} & \lambda_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{1}{y_n-x_0} & \frac{1}{y_n-x_1} & \ldots & \frac{1}{y_n-x_{n-1}} & \lambda_n \end{vmatrix} = 0.$$

Expansion of this determinant with respect to the last column leads to an expression form which it may be shown that $\lambda_0 = \lambda_1 = \ldots = \lambda_n = 0$. We shall make use of the formula for a Cauchy determinant

$$\begin{vmatrix} \frac{1}{\eta_0-\xi_0} & \frac{1}{\eta_0-\xi_1} & \ldots & \frac{1}{\eta_0-\xi_m} \\ \frac{1}{\eta_1-\xi_0} & \frac{1}{\eta_1-\xi_1} & \ldots & \frac{1}{\eta_1-\xi_m} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{1}{\eta_m-\xi_0} & \frac{1}{\eta_m-\xi_1} & \ldots & \frac{1}{\eta_m-\xi_m} \end{vmatrix} = \frac{\prod\limits_{0\leqslant h<i\leqslant m}(\xi_h-\xi_i)\prod\limits_{0\leqslant h<i\leqslant m}(\eta_i-\eta_h)}{\prod\limits_{h=0}^{m}\prod\limits_{i=0}^{m}(\eta_i-\xi_h)}.$$

Thus, the coefficient of λ_j in the expansion of the previous determinant is equal to

$$\frac{(-1)^{n+j}\prod\limits_{0\leqslant h<i\leqslant n-1}(x_h-x_i)\prod\limits_{0\leqslant h<i\leqslant n;\,h,i\neq j}(y_i-y_h)}{\prod\limits_{h=0}^{n-1}\prod\limits_{i=0,i\neq j}^{n}(y_i-x_h)}$$

$$= \frac{(-1)^{n+j}\prod\limits_{0\leqslant h<i\leqslant n-1}(x_h-x_i)\prod\limits_{0\leqslant h<i\leqslant n}(y_i-y_h)\prod\limits_{h=0}^{n-1}(y_j-x_h)}{(y_j-y_0)\ldots(y_j-y_{j-1})(y_{j+1}-y_j)\ldots(y_n-y_j)\prod\limits_{h=0}^{n-1}\prod\limits_{i=0}^{n}(y_i-x_h)}.$$

Rejection of factors not depending upon j leads from (41) to the equivalent equality

$$\sum_{j=0}^{n}\frac{\prod\limits_{h=0}^{n-1}(y_j-x_h)}{\prod\limits_{k=0,k\neq j}^{n}(y_k-y_j)}\lambda_j = 0.$$

Since

$$y_0 < x_0 < y_1 < x_1 < \ldots < x_{n-1} < y_n,$$

we have

$$\operatorname{sgn} \prod_{h=0}^{n-1} (y_j - x_h) = (-1)^{n-j}, \qquad \operatorname{sgn} \prod_{k=0, k\neq j}^{n} (y_k - y_j) = (-1)^j.$$

Therefore in the previous equality all coefficients if λ_j have the same sign $(-1)^n$ and this equality holds if and only if $\lambda_0 = \lambda_1 = \ldots = \lambda_n = 0$. Then, however, (40) is a homogeneous system satisfied only if $dx_0 = dx_1 = \ldots = dx_n = 0$.

Example 5. Let $n = 2$. Expression (36) has then the form

$$\max_{a\in I} |a - x_0|^{b_1} |a - x_1|. \tag{42}$$

Theorem 6 will be used while searching for its minimum. Assume for a moment that $x_0 = 0$, $x_1 = 1$. The function $F(a) = |a|^{b_1}|a-1|$ attains in the interval $\langle 0, 1\rangle$ the maximum value $\omega = b_1^{b_1}/(b_1+1)^{b_1+1}$ in the point $b_1/(b_1+1)$.

We find the roots $\mu_0 < 0$ and $\mu_1 > 1$ of the equation $F(a) = \omega$. In a linear transformation of the interval $\langle \mu_0, \mu_1 \rangle$ into the interval I the points 0 and 1 are transformed into x_0 and x_1, respectively. Hence, for $I = \langle c, d\rangle$, we have

$$x_0 = \frac{c\mu_1 - d\mu_0}{\mu_1 - \mu_0}, \qquad x_1 = \frac{d(1-\mu_0) - c(1-\mu_1)}{\mu_1 - \mu_0}.$$

The minimum of (42) is equal

$$b_1^{b_1} \left(\frac{d-c}{(\mu_1 - \mu_0)(b_1+1)} \right)^{b_1+1}.$$

In particular, in Example 3 which corresponds to the linear interpolation method (see Example 4) is $b_1 = \frac{1}{2}(\sqrt{5} - 1)$. Using the above described method, one obtains here the following optimum initial approximations of the root:

$$x_0 = c + .9012106940(d-c), \qquad x_1 = c + .2032153052(d-c), \tag{43}$$

or, symmetrically,

$$x_0 = c + .0987893060(d-c), \qquad x_1 = c + .7967846948(d-c). \tag{44}$$

However, the minimum of (42) is equal to

$$.190563(d-c)^{b_1+1}. \tag{45}$$

Since (35) must be satisfied, the length of the interval I and the number M should satisfy

$$.190563(d-c)^{b_1+1} < M^{-1/(\xi-1)},$$

i.e.

$$M(d-c) < .190563^{1-\xi} \approx 2.7859. \tag{46}$$

The already mentioned theorem of Traub ([2], p. 38) guarantees the convergence of the sequence (33) to the root a provided

$$M|a-x_0| < 1, \quad M|a-x_1| < 1.$$

Knowing about the root only that it belongs to $\langle c, d\rangle$ and assuming any approximations x_0, x_1 from this interval, one should suppose that $M(d-c) < 1$.

It is easy to see that a given choice of initial approximations weakens the conditions which should be imposed on the interval I. However, one cannot say that in the considered example we are allowed to lengthen 2.7859 times the interval I, because that would usually increase M.

Example 6. Assume now that $n = 3$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$ (this is so in Müller's method mentioned at the beginning of Section 2). Then

$$\xi = 1.839286755, \quad b_0 = 1, \quad b_1 = \xi-1, \quad b_2 = 1/\xi.$$

Function (36) of the variables x_0, x_1, x_2 has now not two local minima, as (42), but six ones, for there are six different arrangements of the initial approximations x_0, x_1, x_2. Here are the points in which this function reaches its minimum (and also the minimum values):

$$\begin{aligned} x_0 &= c+.036732638(d-c),\\ x_1 &= c+.373603389(d-c),\\ x_2 &= c+.901124699(d-c),\\ &.0654258762(d-c)^{2.38297577},\\ x_0 &= c+.037549257(d-c),\\ x_1 &= c+.924682335(d-c),\\ x_2 &= c+.415548395(d-c),\\ &.0653290033(d-c)^{2.38297577}, \end{aligned}$$

$$\left\{\begin{aligned} x_0 &= c+.457319437(d-c),\\ x_1 &= c+.072578878(d-c),\\ x_2 &= c+.903027492(d-c),\\ &.0652918941(d-c)^{2.38297577}. \end{aligned}\right. \tag{47}$$

The three remaining minima are obtained by replacing each of the coefficients of $d-c$ in the formulae for x_0, x_1, x_2 by its complement. The optimum approximations are thus given by (47) or by their equivalents obtained in the above sketched method.

Examples 5 and 6 show that the problem of minimizing the function (36) has (at least) two solutions which are mutually symmetric with respect to the centre of interval I. This non-uniqueness may be somewhat misleading in how to choose the optimum initial approximations, and the problem should rather be modified.

Example 7. Consider as an illustration once more the linear interpolation method from Example 4. Assume x_0 and x_1 be the chosen initial approximations. If

$$(48) \qquad |f(x_0)| < |f(x_1)|$$

holds, we may suspect that

$$(48) \qquad |a-x_0| < |a-x_1|,$$

which, of course, is not always reasonable. In case of (49) we have

$$|a-x_0|^{b_1}|a-x_1| > |a-x_1|^{b_1}|a-x_0|$$

since $b_1 < 1$. Thus, an acceleration of the convergence of the linear interpolation method is possible by an (intuitive) renumeration of the approximations in such a way that the first approximation be better than the zero one. This is done by changing the expression (42) into

$$(50) \qquad \max_{a\in I}\min\{|a-x_0|^{b_1}|a-x_1|,\ |a-x_1|^{b_1}|a-x_0|\}.$$

We search its minimum, as in Example 5, and obtain

$$(51) \qquad x_0 = c+.1793718563(d-c), \qquad x_1 = d-.1793718563(d-c).$$

This are the optimum initial approximations which are to be rearranged in case of (48).

In point (51) the expression (50) is equal to

$$.158742737(d-c)^{b_1+1}$$

thus is less than (45). Inequality (46) is now replaced by

$$M(d-c) < 3.1189.$$

Generally, one may advise the following procedure: The approximations $x_0, x_1, \ldots, x_{n-1}$ are determined by searching the minimum of

$$\max_{a\in I}\ \min_{\{k_0,k_1,\ldots,k_{n-1}\}\in K}\ \prod_{h=0}^{n-1}|a-x_{k_h}|^{b_{n-1-h}},$$

where K is the set of all possible permutations $k_0, k_1, \dots, k_{n-1}$ of the numbers $0, 1, \dots, n-1$. Next we change the numeration of the approximations as to have satisfied

$$|f(x_0)| \geqslant |f(x_1)| \geqslant \dots \geqslant |f(x_{n-1})| \tag{52}$$

for the equation being solved.

References

[1] А. М. Островский, *Решение уравнений и систем уравнений*, Москва 1963 (A. M. Ostrowski, *Solution of equations and systems of equations*, New York 1960).

[2] J. F. Traub, *Iterative methods for the solution of equations*, Englewood Cliffs 1964.

NUMERICAL LABORATORY
MATHEMATICAL INSTITUTE
UNIVERSITY OF WROCŁAW

Received on 18. 3. 1970

S. PASZKOWSKI (Wrocław)

OPTYMALNY WYBÓR PRZYBLIŻEŃ POCZĄTKOWYCH W INTERPOLACYJNYCH METODACH ROZWIĄZYWANIA RÓWNAŃ

STRESZCZENIE

W § 1 rozpatruje się ciągi nieujemnych liczb $\{d_i\}$ spełniających rekurencyjne nierówności (1), gdzie $\gamma_1, \gamma_2, \dots, \gamma_n$ są ustalonymi liczbami nieujemnymi ($\gamma_n \neq 0$). Pokazano, kiedy wszystkie elementy ciągu spełniają nierówności (9) (twierdzenie 2) oraz kiedy ciąg jest zbieżny do zera (twierdzenie 5).

W § 2 zastosowano wyniki § 1 do określenia błędów przybliżeń x_k pierwiastka a równania $f(x) = 0$, obliczanych metodami interpolacyjnymi ([2], rozdz. 4). Okazuje się, że przy znajomości przedziału I, w którym zawarty jest pierwiastek a, należy wybrać przybliżenia początkowe $x_0, x_1, \dots, x_{n-1}$ tak, by wyrażenie (36) osiągnęło minimum.

W § 3 podano warunki konieczne na to, by wyrażenie (36) osiągnęło minimum (twierdzenie 6). Znaleziono także najlepsze przybliżenia początkowe (43) dla metody interpolacji liniowej oraz (47) dla metody interpolacji kwadratowej. Sformułowano także zadanie optymalnego wyboru przybliżeń początkowych, gdy przybliżenia $x_0, x_1, \dots, x_{n-1}$ uporządkowane są na początku tak, że spełniają nierówność (52).