ON IMPROVING AN APPROXIMATE FACTOR OF A POLYNOMIAL WITH REAL COEFFICIENTS BY AN ITERATIVE PROCESS

by R. D. Bhargava *and* R. N. Soral,* *Indian Institute of Technology, Bombay 76*

(Communicated by R. S. Varma, F.N.I.)

(*Received 1 June 1968*)

Bairstows' (1914) method of improving an approximate quadratic factor of a polynomial of degree $n \geqslant 2$ with real coefficients is generalized to improve an approximate factor of an arbitrary degree $m < n$. The treatment is an algebraic one but the numbers can be easily substituted. As is easily seen the method can have several applications. The algorithm for such a method is given and its convergence is proved in this paper.

§ 1. Consider the polynomial

$$F(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n$$

of degree n with real coefficients $a_0, a_1, \ldots, a_n$ and let an approximate factor of it be

$$P(x) = x^m + p_1x^{m-1} + p_2x^{m-2} + \ldots + p_{m-1}x + p_m$$

where $m < n$. Then the problem is to improve the coefficients $p_1, p_2, \ldots, p_m$ of $P(x)$ so that the modified coefficients are better approximation to the factor of $F(x)$ than the original ones. In other words, if the identity

$$F(x) = (x^m + p_1x^{m-1} + \ldots + p_m)Q(x) + R_1x^{m-1} + R_2x^{m-2} + \ldots + R_m \quad (1)$$

holds, where

$$Q(x) = b_0x^{n-m} + b_1x^{n-m-1} + \ldots + b_{n-m}$$

and the coefficients $b_0, b_1, \ldots, b_{n-m}, R_1, R_2, \ldots, R_m$ depend only upon $p_1, p_2, \ldots, p_m$, then the problem is to find $p_1, p_2, \ldots, p_m$, so that

$$R_i(p_1, p_2, \ldots, p_m) = 0; \quad i = 1, 2, \ldots, m \quad \ldots \quad \ldots \quad (2)$$

For this we use Newton's interative process. That is, we find the small corrections $\Delta p_1, \Delta p_2, \ldots, \Delta p_m$, so that

$$R_i(p_1 + \Delta p_1, p_2 + \Delta p_2, \ldots, p_m + \Delta p_m) = 0; \quad i = 1, 2, \ldots, m \quad \ldots \quad (3)$$

* R. N. Soral has left for U.S.A. recently.

Expanding eqns. (3) in Taylor series and supposing that the squares and products of Δp_i are negligible as compared to Δp_i, we truncate after the first order terms and get a system of m linear simultaneous equations

$$R_i(p_1, p_2, \ldots, p_m) + \sum_{t=1}^{m} \frac{\partial R_i}{\partial p_t} \Delta p_t = 0; \quad i = 1, 2, \ldots, m \qquad (4)$$

in m unknowns $\Delta p_1, \Delta p_2, \ldots, \Delta p_m$. When this system has been solved, the procedure is repeated with the corrected values for $p_1, p_2, \ldots, p_m$. The problem, however, is to find the values of $\partial R_i / \partial p_t$. For this, we proceed as follows. We use the identity (1) to obtain

$$b_k = a_k - \sum_{j=1}^{m} b_{k-j} p_j \qquad (5)$$

where

$$k = 0, 1, 2, \ldots, n-m, \quad b_{-1} = b_{-2} = \ldots = b_{-m} = 0$$

and

$$R_i = a_{n-m+i} - \sum_{k=i}^{m} p_k b_{n-m+i-k}$$

$$= b_{n-m+i} + \sum_{k=1}^{i-1} p_k b_{n-m+i-k} \qquad (6)$$

where $i = 1, 2, \ldots, m$ and $b_{n-m+1}, b_{n-m+2}, \ldots, b_n$ are also defined by eqn. (5) for $k = n-m+1, \ldots, n$.

Differentiating eqn. (5) with respect to p_t we obtain

$$\frac{\partial b_k}{\partial p_t} = -\left(b_{k-t} + \sum_{j=1}^{m} \frac{\partial b_{k-j}}{\partial p_t} p_j \right) \qquad (7)$$

It may be noted that a_k are constants and, therefore, their derivatives are zero. Putting $\frac{\partial b_k}{\partial p_t} = -c_{k-t}$, it is obvious that

$$\frac{\partial b_{k+1}}{\partial p_1} = \frac{\partial b_{k+2}}{\partial p_2} = \ldots = \frac{\partial b_{k+m}}{\partial p_m} = -c_k$$

for all k. Inserting c_k in (7) we find

$$c_k = b_k - \sum_{j=1}^{m} c_{k-j} p_j \qquad (8)$$

where $k = 0, 1, 2, \ldots, n-m$, $c_{-1} = c_{-2} = \ldots = c_{-m} = 0$. We see that c_k are computed from the b_k exactly as the b_k were computed from the a_k. We now differentiate eqn. (6) with respect to p_t and write R_{it} for $\partial R_i/\partial p_t$ to obtain

$$R_{it} = \frac{\partial b_{n-m+i}}{\partial p_t} + \sum_{k=1}^{i-1}\left(p_k \frac{\partial b_{n-m+i-k}}{\partial p_t} + b_{n-m+i\ k}\frac{\partial p_k}{\partial p_t}\right)$$

$$= \begin{cases} \sum_{k=i}^{m} p_k c_{n-m+i-t-k}, & \text{if } t < i \\ -\sum_{k=0}^{i-1} p_k c_{n-m+i-t-k} \quad (p_0 = 1) & \text{if } t \geqslant i \end{cases} \qquad \ldots \quad \ldots \quad (9)$$

Thus the values of R_{it} are known and the system (4) can, therefore, be solved for $\Delta p_1, \Delta p_2, \ldots, \Delta p_m$. The whole procedure is summarized in the following algorithm.

Algorithm—Given an approximate factor

$$P(x) = x^m + p_1 x^{m-1} + \ldots + p_m$$

of the polynomial

$$F(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

determine for each $\gamma = 0, 1, 2, \ldots$ a sequence of coefficients $p_k^{(\gamma)}$ $(k = 1, 2, \ldots, m)$ of $P(x)$ as follows. Determine

(i) the sequence $\{b_k^{(\gamma)}\}$ from

$$b_k^{(\gamma)} = a_k - \sum_{j=1}^{m} b_{k-j}^{(\gamma)} p_j^{(\gamma)}$$

where $k = 0, 1, 2, \ldots, n$ and $b_{-1}^{(\gamma)} = b_{-2}^{(\gamma)} = \ldots = b_{-m}^{(\gamma)} = 0$.

(ii) the sequence $\{c_k^{(\gamma)}\}$ from

$$c_k^{(\gamma)} = b_k^{(\gamma)} - \sum_{j=1}^{m} c_{k-j}^{(\gamma)} p_j^{(\gamma)}$$

where $k = 1, 2, \ldots, n-m$ and $c_{-1}^{(\gamma)} = c_{-2}^{(\gamma)} = \ldots = c_{-m}^{(\gamma)} = 0$.

(iii) the sequence $\{R_i^{(\gamma)}\}$ from

$$R_i^{(\gamma)} = b_{n-m+i}^{(\gamma)} + \sum_{k=1}^{i-1} p_k^{(\gamma)} b_{n-m+i-k}^{(\gamma)}$$

where $i = 1, 2, \ldots, m$.

(iv) the sequence $\{R_{it}^{(\gamma)}\}$ from

$$R_{it}^{(\gamma)} = \begin{cases} \sum_{k=i}^{m} p_k^{(\gamma)} c_{n-m+i-t-k}^{(\gamma)} & \text{if } t < i \\ -\sum_{k=0}^{i-1} p_k^{(\gamma)} c_{n-m+i-t-k}^{(\gamma)} & \text{if } t \geqslant i \end{cases}$$

where $i, t = 1, 2, \ldots, m$. Then

(v) the set $\{p_k^{(\gamma+1)};\ k = 1, 2, \ldots, m\}$ is obtained by $p_k^{(\gamma+1)} = p_k^{(\gamma)} + \Delta p_k^{(\gamma)}$, where $\{\Delta p_k^{(\gamma)};\ k = 1, 2, \ldots, m\}$ is determined by

$$\sum_{t=1}^{m} R_{it}^{(\gamma)} \Delta p_t^{(\gamma)} = -R;\quad i = 1, 2, \ldots, m.$$

§ 2. To prove the convergence, we give below the conditions sufficient to imply that there exists a neighbourhood around a real solution $p = (p_1, p_2, \ldots, p_m)$ of the system (2) such that for an initial choice $p^{(0)} = (p_1^{(0)}, p_2^{(0)}, \ldots, p_m^{(0)})$ in that neighbourhood, the above process converges to the solution.

Let $D(p)$ denote the determinant of the coefficient matrix of system (4), i.e.

$$D(p) = \begin{vmatrix} R_{11} & R_{12} & \ldots & R_{1m} \\ R_{21} & R_{22} & \ldots & R_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ R_{m1} & R_{m2} & \ldots & R_{mm} \end{vmatrix}$$

and let $(s_1, s_2, \ldots, s_m) = s$ be a solution of the system (2). If the function $R_i(p)$ have continuous first partial derivatives in the neighbourhood of the solution $p = s$, and if the determinant $D(p)$ is non-vanishing in the neighbourhood of this solution, then Newton's process of solving the system (2) converges (Householder 1953). Also if $D(s) \neq 0$, then since $D(p)$ is continuous, it follows that $D(p) \neq 0$ for all points p sufficiently close to s. Thus in order to show the uniqueness of the solution of the system (4) and the convergence of the above method we have to show that $D(s) \neq 0$. To prove this for general case seems to be difficult but if the zeros $x_1, x_2, \ldots, x_m$ of the factor $x^m + s_1 x^{m-1} + \ldots + s_m$ of $F(x)$ are simple and distinct the result can be proved as follows. Since $F(x)$ is independent of $p_1, p_2, \ldots, p_m$, differentiating the identity (1) with respect to $p_1, p_2, \ldots, p_m$ respectively, we obtain

$$0 = x^{m-1}Q(x) + P(x)\frac{\partial Q}{\partial p_1} + x^{m-1}R_{11} + x^{m-2}R_{21} + \ldots + R_{m1}$$

$$0 = x^{m-2}Q(x) + P(x)\frac{\partial Q}{\partial p_2} + x^{m-1}R_{12} + x^{m-2}R_{22} + \ldots + R_{m2}$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$0 = Q(x) + P(x)\frac{\partial Q}{\partial p_m} + x^{m-1}R_{1m} + x^{m-2}R_{2m} + \ldots + R_{mm}$$

Setting $p_i = s_i$ and $x = x_i$ $(i = 1, 2, \ldots, m)$ we obtain m systems

$$\left.\begin{aligned} x_i^{m-1}R_{11} + x_i^{m-2}R_{21} + \ldots + R_{m1} &= -x_i^{m-1}Q_i \quad (i = 1, 2, \ldots, m) \\ x_i^{m-1}R_{12} + x_i^{m-2}R_{22} + \ldots + R_{m2} &= -x_i^{m-2}Q_i \quad (i = 1, 2, \ldots, m) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \quad \cdot \quad \cdot \quad \cdot \\ x_i^{m-1}R_{1m} + x_i^{m-2}R_{2m} + \ldots + R_{mm} &= -Q_i \quad (i = 1, 2, \ldots, m) \end{aligned}\right\} \quad (10)$$

each of m linear simultaneous equations in m unknowns with the non-vanishing determinant

$$|A| = \begin{vmatrix} x_1^{m-1} & x_1^{m-2} & \dots & 1 \\ x_2^{m-1} & x_2^{m-2} & \dots & 1 \\ \cdot & \cdot & \cdot & \cdot \\ x_m^{m-1} & x_m^{m-2} & \dots & 1 \end{vmatrix}$$

$$= \prod_{i>j} (x_i - x_j), \quad 1 < j < i \leqslant m$$

$$\neq 0 \text{ (because the zeros } x_1, x_2, \dots, x_m \text{ of } F(x) \text{ are distinct)}$$

of the coefficient matrix A. Here $Q_i = Q(x_i)$ and the derivatives R_{ij} are taken at $p = s$. Now we shall show that the solution vectors of the systems in system (10) are linearly independent. For this we have to show that

$$a_1(R_{11}, R_{21}, \dots, R_{m1}) + a_2(R_{12}, R_{22}, \dots, R_{m2}) + \dots + a_m(R_{1m}, R_{2m}, \dots, R_{mm}) = 0 \quad \dots (11)$$

is true only for $a_1 = a_2 = \dots = a_m = 0$. Let the identity (11) be true of any constants $a_1, a_2, \dots, a_m$. Then

$$\left.\begin{aligned} a_1R_{11} + a_2R_{12} + \dots + a_mR_{1m} &= 0 \\ a_1R_{21} + a_2R_{22} + \dots + a_mR_{2m} &= 0 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \quad \cdot \\ a_1R_{m1} + a_2R_{m2} + \dots + a_mR_{mm} &= 0 \end{aligned}\right\} \quad \dots \quad \dots (12)$$

From (10) and (12) we obtain

$$x_i^{m-1}Q_ia_1 + x_i^{m-2}Q_ia_2 + \dots + Q_ia_m = 0 \quad (i = 1, 2, \dots, m) \quad \dots (13)$$

But $Q_i = Q(x_i) \neq 0$, because the zeros $x_1, x_2, \dots, x_m$ of $P(x)$ are simple zeros of $F(x)$.

Dividing (13) by Q_i we obtain

$$x_i^{m-1}a_1 + x_i^{m-2}a_2 + \dots + a_m = 0 \quad (i = 1, 2, \dots, m) \quad \dots (14)$$

Since the determinant $|A|$ of the coefficient matrix of the above system (14) is non-zero, this system has only trivial solution, i.e. $a_1 = a_2 = \dots = a_m = 0$. Hence the solution vectors $(R_{11}, R_{21}, \dots, R_{m1})$, $(R_{12}, R_{22}, \dots, R_{m2})$, $\dots$, $(R_{1m}, R_{2m}, \dots, R_{mm})$ are linearly independent from which it follows that

$$D(s) = \begin{vmatrix} R_{11} & R_{12} & \dots & R_{1m} \\ R_{21} & R_{22} & \dots & R_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ R_{m1} & R_{m2} & \dots & R_{mm} \end{vmatrix} \neq 0$$

Thus for all p sufficiently close to s the system (4) has a unique solution $\Delta p_1, \Delta p_2, \ldots, \Delta p_m$ and the point $(p_1+\Delta p_1, p_2+\Delta p_2, \ldots, p_m+\Delta p_m)$ is more close to s than p.

Acknowledgement

The authors are grateful to the referee for his kind suggestions.

References

Bairstows, L. (1914). Investigations relating to the stability of the aeroplane. *Reports and Memoranda No. 154 of Advisory Committee for Aeronautics.*

Householder, A. S. (1953). Principles of Numerical Analysis. McGraw-Hill Book Co. Inc., New York.