MATRIX METHODS FOR FINDING ROOTS OF POLYNOMIALS

By

LAJOS LÁSZLÓ

(Received October 6, 1978)

1. Introduction. Present paper deals with a method tracing back investigations connected with polynomials to the field of matrix theory. A well known example of this method is the following.

Let $\{a_i\}_{i=0}^{n-1}$ be complex numbers, $n \in N$. To the polynomial

$$p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \tag{1a}$$

we can join the matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & & & 1 \\ -a_0 & -a_1 & & \ldots & -a_{n-1} \end{bmatrix} \tag{1b}$$

with complex elements, i.e. $A \in C^{n\times n}$. This matrix occurs in the theory of the linear differential equations. The characteristic polynomial of A is a scalar multiple of p, hence the roots of p and the eigenvalues of A are the same. Thus, the theorems of the matrix theory can be applied for the study of the roots of polynomials. Among them the following seems to be the most general and frequently used one.

Theorem (Gerschgorin): Let $A = (a_{ik})_{i=1,\ldots,n}^{k=1,\ldots,n} \in C^{n\times n}$,

$$r_i = \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ij}|, \quad 1 \leqq i \leqq n, \quad K_i = \{z \in C;\ |z - a_{ii}| \leqq r_i\}.$$

Then $K = \bigcup_{i=1}^{n} K_i$ contains all the eigenvalues of A. Moreover, if K can be divided in two disjoint parts $K = K^1 \cup K^2$, where $K^1 = \bigcup_{i\in I_1} K_i$, $K^2 = \bigcup_{i\in I_2} K_i$, $|I_1| = k$, then K^1 contains k, K^2 contains $n-k$ eigenvalues of A.

In the above case the theorem gives that, the roots of p lie in the circle with center in the origo and with radius

$$1 + \max_{0 \leq i \leq n-1} |a_i| \tag{2}$$

This example illustrates the advantage of replacing a polynomial – for a special purpose, of course – by a suitable matrix. The correspondence (1*a*)–(1*b*) is, however, not the most fortunate one.
His *disadvantages* are:

a) The centres of the Gerschgorin-circles are *the same*, hence this matrix cannot be used to localize to disconnected domains the roots of the polynomial, thus (2) gives a rough estimation in some sence.

b) The pre-cited method gives for a polynomial *only one* matrix and this bounds the wide-spread applications of the matrix theory.

A method doing without these disadvantages is described below.

2. Theorem 1. Let $p(\lambda) = \sum_{i=0}^{n} a_i \lambda^i$ be a complex polynomial, $a_n = 1$. Let $\{A_i\}_{i=1}^{n-1}$ be different complex numbers, and

$$A_n = -a_{n-1} - \sum_{i=1}^{n-1} A_i .$$

There exist complex numbers $\{x_i\}_{i=1}^{n-1}$ such that the matrix

$$A = \begin{bmatrix} A_1 & 0 & \dots & 0 & x_1 \\ 0 & A_2 & \dots & 0 & x_2 \\ \vdots & & & & \\ 0 & & \dots & A_{n-1} & x_{n-1} \\ x_1 & x_2 & \dots & x_{n-1} & A_n \end{bmatrix} \tag{3}$$

satisfies the relation

$$p(\lambda) = \det(\lambda I - A) \tag{4}$$

where I is the identity, and the numbers $\{x_i\}_{i=1}^{n-1}$ can be expressed explicity via the formulas

$$x_i = \sqrt{\frac{-p(A_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n-1} (A_i - A_j)}}, \quad 1 \leq i \leq n-1 \tag{5}$$

taking any value of the square root. –

Remark: We describe two proofs. The first one is the direct way to get the result, as for the second one it proves the theorem knowing the formulas (5) very shortly.

PROOF 1. Compare the coefficients in $p(\lambda)$ and $\det(\lambda I - A)$. By fixed $\{A_i\}_{i=1}^n$ we get a system of $n-1$ linear equations with the unknowns $\{x_i^2\}_{i=1}^{n-1}$. (Notice that there is no chance of expressing $\{A_i\}_{i=1}^n$ via $\{x_i^2\}_{i=1}^{n-1}$, because – consider the case $x_i = 0$, $1 \leqq i \leqq n-1$ – it would be equivalent to be able to solve an algebraic equation of degree n).
Expanding the determinant in (4), we obtain

$$p(\lambda) = \prod_{i=1}^{n} (\lambda - A_i) - \sum_{i=1}^{n-1} x_i^2 \prod_{\substack{j=1 \\ j \neq i}}^{n-1} (\lambda - A_j). \tag{6a}$$

The k-th elementary symmetric polynomial of A_l, $l \in H$ will be denoted by $S_k^{(A_l,\, l \in H)}$ (H is a subset of $\{1, 2, \ldots, n\}$). Let $S_k^{(A_l,\, l \in \{1, 2, \ldots, n\})} \equiv S_k$ and for $k < 0$ let $S_k^{(A_l,\, l \in H)} \equiv 0$.
Now, (6a) can be rewritten as

$$\sum_{k=0}^{n} a_k \lambda^k = \sum_{k=0}^{n} \left\{ (-1)^{n-k} S_{n-k} + (-1)^{n-k+1} \sum_{i=1}^{n-1} S_{n-k-2}^{(A_l\, l \neq i,\, n)} x_i^2 \right\} \lambda^k. \tag{6b}$$

Hence

$$\sum_{i=1}^{n-1} (-1)^{n-k} S_{n-k-2}^{(A_l\, l \neq 2,\, n)} x_i^2 = (-1)^{n-k} S_{n-k} - a_k, \quad k = 0, 1, \ldots, n.$$

The coefficients of λ^n and λ^{n-1} are equal in both sides of (6b), in view of the conditions $a_{n-1} = -\sum_{i=1}^{n} A_i = -S_1$ and $a_n = 1$. Therefore the equations for $k = n-1, n$ may be omitted.
In order to apply matrix symbolics we use the following notations:

$$z_k := x_k^2, \quad 1 \leqq k \leqq n-1, \quad b_k = (-1)^k S_{k+2} - a_{n-k-2}, \quad 0 \leqq k \leqq n-2,$$

$$\mathbf{z} := \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \end{bmatrix} \in \mathbf{C}^{n-1}; \quad \mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-2} \end{bmatrix} \in \mathbf{C}^{n-1};$$

$$B := \left((-1)^{n-i} S_{n-i-2}^{(A_l\, l \neq k,\, n)} \right)_{i=n-2,\, n-3,\, \ldots,\, 0}^{k=1,\, 2,\, \ldots,\, n-1} \in \mathbf{C}^{(n-1)\times(n-1)}$$

Then (6a) and (6b) is equivalent to

$$B\mathbf{z} = \mathbf{b} \tag{7}$$

or,

$$\begin{bmatrix} 1 & 1 \dots & 1 \\ -A_2-A_3-\dots-A_{n-1} & & -A_1-A_2-\dots-A_{n-2} \\ \vdots & & \vdots \\ (-1)^n A_2 A_3 \cdot \dots \cdot A_{n-1} & & (-1)^n A_1 A_2 \cdot \dots \cdot A_{n-2} \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_{n-1}^2 \end{bmatrix} =$$

$$= \begin{bmatrix} S_2 - a_{n-2} \\ -S_3 - a_{n-3} \\ \vdots \\ (-1)^n S_n - a_0 \end{bmatrix}$$

Observe, that the i-th column of B consists of the coefficients of

$$q_i(\lambda) = \prod_{\substack{j=1 \\ j \neq i}}^{n-1} (\lambda - A_j).$$

Thus, with the matrix

$$C := \begin{bmatrix} A_1^{n-2} \dots A_1 & 1 \\ A_2^{n-2} \dots A_2 & 1 \\ A_{n-1}^{n-2} \dots A_{n-1} & 1 \end{bmatrix}$$

it holds

$$CB = \begin{bmatrix} q_1(A_1) & 0 & & 0 \\ 0 & q_2(A_2) & & \\ & & & \\ & & & 0 \\ 0 \dots 0 & & & q_{n-1}(A_{n-1}) \end{bmatrix} = \operatorname{diag}[q_i(A_i)],$$

and finally,

$$B^{-1} = \begin{bmatrix} \frac{1}{q_1(A_1)} & 0 & & 0 \\ 0 & \cdot & & \\ & & & \\ & & & 0 \\ 0 \dots 0 & & & \frac{1}{q_{n-1}(A_{n-1})} \end{bmatrix} \cdot C = \operatorname{diag}\left[\frac{1}{q_i(A_i)}\right] \cdot C \tag{8}$$

(Obviously $q_i(A_i) \neq 0$, $1 \leqq i \leqq n-1$, because $\{A_i\}_{i=1}^{n-1}$ are different numbers). Now from (7) and (8) we obtain

$$\mathbf{z} = B^{-1}\mathbf{b} = \operatorname{diag}\left(\frac{1}{q_i(A_i)}\right) C\,\mathbf{b},$$

that is

$$x_i^2 = z_i = \frac{(Cb)_i}{q_i(A_i)}, \quad 1 \leqq i \leqq n-1. \tag{9}$$

Compute $(Cb)_i$, the i-th component of $C\mathbf{b}$. To this end, let $\mathbf{b} = \bar{\gamma} + \bar{\delta}$,

$$\gamma_i := (-1)^i \cdot S_{i+2} \qquad \delta_i := -a_{n-i-2}, \qquad 0 \leqq i \leqq n-2,$$

$$\bar{\gamma} := \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{n-2} \end{bmatrix} \in \mathbf{C}^{n-1}, \quad \bar{\delta} := \begin{bmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_{n-2} \end{bmatrix} \in \mathbf{C}^{n-1}.$$

Then

$$(C\gamma)_i = \sum_{j=2}^{n} (-1)^j A_i^{n-j} S_j = \sum_{j=0}^{n} - \sum_{j=0}^{1} = 0 - A_i^n + S_1 A_i^{n-1},$$

$$(C\delta)_i = -\sum_{j=0}^{n-2} a_j A_j^i = -\sum_{j=0}^{n} + \sum_{j=n-1}^{n} = -p(A_i) + A_i^n + a_{n-1} A_i^{n-1}$$

Consequently, using again the relation $S_1 = -a_{n-1}$, it follows:

$$(Cb)_i = (C\gamma)_i + (C\delta)_i = -p(A_i). \tag{10}$$

By (9) and (10) $x_i^2 = \dfrac{-p(A_i)}{\prod\limits_{\substack{j=1 \\ j \neq i}}^{n-1} (A_i - A_j)}$, which was to be proved.

PROOF 2. To prove (4), we use the formulas (5) for x_i, $1 \leqq i \leqq n-1$. The coefficients by λ^n and λ^{n-1} are at both polynomials the same, thus our proof is ready if we find $n-1$ numbers, for which in (4) – or equivalently in (6) – the equality holds. But it is easy to see, that for this aim the numbers $\{A_i\}_{i=1}^{n-1}$ are suitable.

Combining Theorem 1 and the theorem of Gerschgorin we get immediately:

Theorem 2. Let $p(\lambda) = \sum\limits_{i=0}^{n} a_i \lambda^i$ be a complex polynomial, $a_n = 1$; $\{A_i\}_{i=1}^{n-1}$ be different complex numbers and

$$A_n := -a_{n-1} - \sum_{i=1}^{n-1} A_i.$$

Let

$$r_i := \sqrt[n-1]{\frac{|p(A_i)|}{\prod\limits_{\substack{j=1\\ j\neq i}} |A_i - A_j|}}, \quad 1 \leqq i \leqq n-1, \quad r_n := \sum_{j=1}^{n-1} r_j,$$

$K_i := \{z \in C;\ |z - A_i| \leqq r_i\}$, $1 \leqq i \leqq n-1$. Then $K = \bigcup\limits_{i=1}^{n} K_i$ contains all the roots of p. Moreover, if H can be divided in two disjoint parts, $K = K^1 \cup K^2$ where $K^1 = \bigcup\limits_{i \in I_1} K_i$, $K^2 = \bigcup\limits_{i \in I_2} K_i$, $|I_1| = k$, then K^1 contains k, K^2 contains $n-k$ roots of p.

3. In this point we will demonstrate a practical application of the previous section. We use the above notations, and assume that $p(A_n) = 0$. To construct an iteration process converging to the roots of p, or equivalently, to the eigenvalues of A, we start from the fact that eigenvalues are invariant under similarity transforms and in the obtained matrix we try the elements »pack into the main diagonal«.

Detailing, let $Y \in C^{n \times n}$ be the sum of the unit matrix and the matrix having y_i in the (i, n) positions, $1 \leqq i \leqq n-1$.

Let $t := \sum\limits_{i=1}^{n-1} x_i y_i$. Then

(11)
$$Y^{-1} A Y =$$

$$= \begin{bmatrix} A_1 - x_1 y_1 & -x_2 y_1 & & -x_{n-1} y_1 & x_1 - y_1(t + A_n - A_1) \\ -x_1 y_2 & A_2 - x_2 y_2 & \cdots & -x_{n-1} y_2 & x_2 - y_2(t + A_n - A_2) \\ \vdots & \vdots & & \vdots & \vdots \\ -x_1 y_{n-1} & -x_2 y_{n-1} & \cdots & A_{n-1} - x_{n-1} y_{n-1} & x_{n-1} - y_{n-1}(t + A_n - A_{n-1}) \\ x_1 & x_2 & & x_{n-1} & A_n + t \end{bmatrix}$$

Denote by $f(y_1, y_2, \ldots, y_{n-1})$ the quadratical sum of the moduls of the nondiagonal elements in $Y^{-1} A Y$. The problem:

$$f(y_1, y_2, \ldots, y_{n-1}) \to \min$$

(12)
$$\text{subject to } \sum_{i=1}^{n-1} x_i y_i = 0$$

has an unique solution, which can be computed explicitly, because the restriction of f on the hyperplane $\sum\limits_{i=1}^{n-1} x_i y_i = 0$ is a positiv definite quadratic

form. Let $\tilde{y}_1, \ldots, \tilde{y}_{n-1}$ be the optimal variables. Taking the main diagona of (11) into consideration, the following iteration procedure can be construc ted:

$$A_i^{\text{new}} = A_i - x_i \tilde{y}_i, \quad 1 \leqq i \leqq n-1$$

(13)

$$A_n^{\text{new}} := A_n.$$

Remarks

a) The restriction $t = \sum_{i=1}^{n-1} x_i y_i = 0$ in (12) is necessary, because $p(A_n) = 0$ and, thus A_n need not change.

b) Although $\{\tilde{y}_i\}_{i=1}^{n-1}$ have explicite expressions, the presence ot the Lagrange-multipliers makes (13) difficult to handle. This is the reason why we prefer the following procedure which is easy to manage theoretically too.

Going back to (11), let

$$y_i^* := \frac{x_i}{A_n - A_i}, \quad 1 \leqq i \leqq n-1.$$

This choice has the advantage, that the last column (except his last element) will be zero. Indeed,

$$t = \sum_{i=1}^{n-1} x_i y_i^* = \sum_{i=1}^{n-1} \frac{x_i^2}{A_n - A_i} = 0,$$

this follows from (6), setting $\lambda := A_n$ and dividing by

$$\prod_{j=1}^{n-1} (A_n - A_j).$$

Therefore,

$$x_i - y_i^*(t + A_n - A_i) = x_i - y_i^*(A_n - A_i) = 0.$$

The iteration process obtained in this way is:

$$A_i^{\text{new}} := A_i - x_i y_i^* = A_i - \frac{x_i^2}{A_n - A_i} = A_i - \frac{p(A_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (A_i - A_j)}, \quad 1 \leqq i \leqq n-1$$

(14)

$$A_n^{\text{new}} := A_n.$$

Now, (14) has an interesting geometrical interpretation. Let $\{A_i\}_{i=n}^{n}$ be the approximations of the roots of the polynomial p. Introducing the polynomial

$$q(\lambda) := \prod_{i=1}^{n} (\lambda - A_i)$$

(14) is equivalent to

$$A_i^{\text{new}} := A_i - \frac{p(A_i)}{q'(A_i)} \tag{15}$$

(c. f. the Newton-method!). By standard means can be proven the following theorem.

Theorem 3. Let p be real polynomial having the real and different roots $\{\alpha_i\}_{i=1}^n$. Then (15) is a locally quadratic convergent iteration process, i.e.

1. if $|A_i^{(1)} - \alpha_i|$ are sufficiently small, $1 \leqq i \leqq n-1$ then

$$\lim_r A_i^{(r)} = \alpha_i, \quad 1 \leqq i \leqq n-1$$

where $A_i^{(r)}$ is the $r-$th iterate of α_i.

2. Under the above hypothesis the convergence is of second order, that is

$$\|e^{r+1}\| = 0(\|e^r\|^2),$$

where

$$e_i^r = A_i^r - \alpha_i, \quad 1 \leqq i \leqq n-1.$$

Remark. We assume that $A_i^{(1)}$ are real, $1 \leqq i \leqq n-1$. Then, obviously, $A_i^{(r)}$ are real , $r \in N$.

PROOF. Subtracting α_i from equation (15), $1 \leqq i \leqq n-1$ and setting

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_{n-1} \end{bmatrix} \in \mathbf{R}^{n-1}; \quad \alpha := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \end{bmatrix} \in \mathbf{R}^{n-1};$$

$$g_i(A) := A_i - \alpha_i - \frac{p(A_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (A_i - A_j)}; \quad g := \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{n-1} \end{bmatrix},$$

$g \in \mathbf{R}^{n-1} \to \mathbf{R}^{n-1}$ we obtain: $e^{r+1} = g(A^r)$.

First we prove that $g(\alpha) = 0$ and $g'(\alpha) = 0$.

$$g_i(\alpha) = -\frac{p(\alpha_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\alpha_i - \alpha_j)} = -\frac{p(\alpha_i)}{p'(\alpha_i)} = 0$$

(α_i is of multiplicity 1) $1 \leqq i \leqq n-1$.

Calculating $g'(\alpha)$, let $i \neq k$. Then

$$\frac{\partial g_i(A)}{\partial A_k} = p(A_i) \cdot \frac{-\prod_{j \neq i,k} (A_i - A_j)}{\prod_{j \neq i} (A_i - A_j)^2}, \qquad \frac{\partial g_i(\alpha)}{\partial A_k} = 0$$

On the other hand,

$$\frac{\partial g_i(A)}{\partial A_i} = 1 - \frac{p'(A_i) \prod_{g \neq i} (A_i - A_j) - p(A_i) \frac{\partial}{\partial A_i} \prod_{j \neq i} (A_i - A_j)}{\prod_{j \neq i} (A_i - A_j)^2},$$

$$\frac{\partial g_i(\alpha)}{\partial A_i} = 1 - \frac{p'(\alpha_i)^2}{p'(\alpha_i)^2} = 0.$$

Now, the equations $g(\alpha) = 0$, $g'(\alpha) = 0$ implicate the locally quadratic convergence. We have

$$e^{r+1} = g(A^r) = g(\alpha) + g'(\alpha)\, e^r + \frac{1}{2} \langle g''(\alpha)\, e^r, e^r \rangle + o(\|e^r\|^2) = 0(\|e^r\|^2) \tag{16}$$

because of the boundedness of g'' in a neighbourhood of α.

Therefore there exist positive numbers K and ε, such that

$$\|e^{r+1}\| \leqq K \cdot \|e^r\|^2, \quad \text{if} \quad \|e^r\| < \varepsilon.$$

Let the starting value $A^{(1)}$ will be chosen according to the inequality

$$q := K\|e^1\| < 1.$$

Then $\|e^{r+1}\| \leqq (K \cdot \|e^1\|)^{2^r - 1} < q^{2^r - 1} \cdot \|e^1\|$, and $\lim_r e^r = 0$.

Combining this with (16) our theorem is proved.

Remark. Both (12) and (14) proved to be good machine procedures. The computer experience shows that (14) – and (12) – produce not only local convergence – c. f. Theorem 3 –, but also in some sence global convergence. The starting values $\{A_i^{(1)}\}_{i=1}^{n-1}$ need not be contained in small neighbourhoods of the roots $\{\alpha_i\}_{i=1}^{n-1}$ of p: the procedure is in these cases convergent, too. A future examine is to investigate this problem.