COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 10 (1977) 165–173

CONVERGENCE ACCELERATION FOR THE ITERATIVE SOLUTION OF THE EQUATIONS $X = AX + f$

M. MEŠINA

Institut für Kernenergetik, Universität Stuttgart, W. Germany

Received 15 March 1976
Revised manuscript received 15 July 1976

A method is presented for accelerating the convergent iterative procedures of solving the system of linear equations $X = AX + f$. The method is also applicable to divergent iterative schemes if the number of eigenvalues of A that are greater in absolute value than unity is not very large. The method is particularly advantageous if the matrix A has not been explicitly constructed because of extensive storage requirements and if it is not possible to use the alhorithms (such as the Chebyshev and Lanzcos polynomial methods) which are designed with respect to the position of eigenvalues of A in the complex plane.

1. Introduction

Let us consider a system of linear equations in the form

$$X = AX + f, \tag{1}$$

where X und f are M-dimensional, real vectors, and A is a square $M \times M$ matrix whose eigenvalues are all distinct from 1. Then the solution of eq. (1) exists for any vector f and is expressed as

$$X = (I - A)^{-1} f. \tag{2}$$

In many applications A is known only implicity as the result of operating on a given vector (see [1]). A natural iterative procedure for solving (1) is

$$X_{n+1} = AX_n + f, \quad n = 0, 1, \ldots \tag{3}$$

This process fails entirely if the spectral radius of A is greater than 1. In some cases it may be impossible to use a procedure like the Chebyshev or Lanzcos polynomial method to accelerate this iterative scheme; moreover, the transformation of (3) to a new form

$$X_{n+1} = A'X_n + f', \quad n = 0, 1, \ldots, \tag{4}$$

which does converge might be very complicated.

The motivation of the present paper is (a) to introduce an acceleration procedure for eq. (3) if A is not explicitly known and the moduli of all eigenvalues are less than 1, and (b) to supply a

method of solving the problem if some moduli are greater than 1. Moreover, this method can also be used if some eigenvalues of A are complex.

2. Approximate solution of the equations $X = AX + f$

For any M-dimensional vector X_n we introduce two error vectors D_n and U_n:

$$D_n = X_n - X\,, \tag{5}$$

$$U_n = (AX_n + f) - X_n\,. \tag{6}$$

From eq. (2)

$$U_n = (A - I)D_n\,. \tag{7}$$

Let us consider any system of vectors X_j ($j = 0, 1, \ldots m$) which may or may not satisfy eq. (3). We define the vector $\tilde{X}_m$ as a linear combination of the vectors X_j:

$$\tilde{X}_m = \sum_{j=0}^{m} C_j X_j\,. \tag{8}$$

To the vector $\tilde{X}_m$ there correspond two error vectors $\tilde{D}_m$ and $\tilde{U}_m$:

$$\tilde{D}_m = \tilde{X}_m - X\,, \tag{9}$$

$$\tilde{U}_m = A\tilde{X}_m + f - \tilde{X}_m\,. \tag{10}$$

From eq. (2)

$$\tilde{U}_m = (A - I)\tilde{D}_m\,. \tag{11}$$

With the help of (5) and (6) we can write (9) and (10) in the following equivalent form:

$$\tilde{D}_m = \sum_{K=0}^{m} C_K D_K + \left(\sum_{K=0}^{m} C_K - 1\right) X\,, \tag{12}$$

$$\tilde{U}_m = \sum_{K=0}^{m} C_K U_K + \left(1 - \sum_{K=0}^{m} C_K\right) f\,. \tag{13}$$

For our further considerations we shall use the Euclidean norms of the vectors D_j ($j = 0, 1, \ldots m$), $\tilde{D}_m$ and $\tilde{U}_m$. If the coefficients C_K are such that

$$\|\tilde{D}_m\| < \|D_j\|\,, \quad j = 0, 1, \ldots m\,, \tag{14}$$

then $\tilde{X}_m$ is a better approximation for X than any of the vectors X_j. Clearly, if $\|\tilde{D}_m\| = 0$, the vector $\tilde{X}_m = X$. The norm $\|\tilde{D}_m\|$ cannot be minimized directly as a function of the coefficients C_K ($K = 0, 1, \ldots m$) since the vectors D_K ($K = 0, 1, \ldots m$) are unknown. However, we can minimize $\|\tilde{U}_m\|$. From eq. (11)

$$\|\tilde{D}_m\| \leqslant \|(A - I)^{-1}\| \, \|\tilde{U}_m\| . \tag{15}$$

Therefore, if $\|\tilde{U}_m\| = 0$, then also $\|\tilde{D}_m\| = 0$. If the vectors X_j are orthogonal, the minimizing of $\|\tilde{U}_m\|$ leads to a special case of a coarse-mesh method.

3. The minimizing of $\|\tilde{U}_m\|$

Let us define a square $(m + 1)$-dimensional matrix G and a vector g with the same dimension:

$$G_{iK} = (f^t - U_i^t)(f - U_K) , \quad i, K = 0, 1, \ldots m , \tag{16}$$

$$g_i = (f^t - U_i^t) f , \quad i = 0, 1, \ldots m . \tag{17}$$

The from eq. (13)

$$\|\tilde{U}_m\|^2 = C^t G C - C^t g - g^t C + f^t f , \tag{18}$$

where C is the $(m + 1)$-dimensional vector whose components are the coefficients C_K ($K = 0, 1, \ldots m$). The vector C minimizing $\|\tilde{U}_m\|^2$ (and therefore also $\|\tilde{U}_m\|$) can be written

$$C = G^{-1} g . \tag{19}$$

This calculation of the C_K ($K = 0, 1, \ldots m$) with their use in eq. (8) is a generalization of a coarse-mesh method. It works with any system (including a nonorthogonal one) of basis vectors X_K. If $\|U_i\| \ll \|f\|$ for two or more vectors U_i, then with respect to the limited precision of the computer the matrix G can be treated as singular – thus eq. (19) cannot be used. In such a case C must be calculated by another algorithm.

Let us assume that C is subject to the condition

$$C^t e = e^t C = 1 , \tag{20}$$

where e is the $(m + 1)$-dimensional vector with all components equal to 1. From eqs. (12) and (13) we obtain

$$\tilde{D}_m = \sum_{K=0}^{m} C_K D_K , \tag{21}$$

$$\tilde{U}_m = \sum_{K=0}^{m} C_K U_K . \tag{22}$$

Let us define the $(m + 1)$-dimensional matrix $\boldsymbol{H}$ with elements

$$H_{iK} = \boldsymbol{U}_i^{\mathrm{t}} \boldsymbol{U}_K \,, \qquad i, K = 0, 1, \ldots m \,. \tag{23}$$

From eq. (22)

$$\| \widetilde{\boldsymbol{U}}_m \|^2 = \boldsymbol{C}^{\mathrm{t}} \boldsymbol{H} \boldsymbol{C} \,. \tag{24}$$

If the matrix $\boldsymbol{H}$ is regular, the vector $\boldsymbol{C}$ minimizing $\| \widetilde{\boldsymbol{U}} \|^2$ and satisfying eq. (20) is uniquely determined:

$$\boldsymbol{C} = \boldsymbol{H}^{-1} \boldsymbol{e} / \boldsymbol{e}^{\mathrm{t}} \boldsymbol{H}^{-1} \boldsymbol{e} \,. \tag{25}$$

Then the minimal norm $\| \widetilde{\boldsymbol{U}}_m \|$ can be written as

$$\| \widetilde{\boldsymbol{U}}_m \| = 1 / \sqrt{\boldsymbol{e}^{\mathrm{t}} \boldsymbol{H}^{-1} \boldsymbol{e}} \,. \tag{26}$$

If the vectors $\boldsymbol{U}_K$ are lineaely dependent, the matrix $\boldsymbol{H}$ is singular, therefore, eq. (25) cannot be used. In such a case the coefficients C'_K can be found, giving $\widetilde{\boldsymbol{U}}_m = \boldsymbol{o}$. If the sum of such coefficients is different from zero, we can use in eq. (8) the coefficients

$$C_K = C'_K \Big/ \sum_{j=0}^{m} C'_j \,, \qquad K = 0, 1, \ldots m \,. \tag{27}$$

Then $\widetilde{\boldsymbol{X}}_m = \boldsymbol{X}$ since $\widetilde{\boldsymbol{U}}_m$ (and therefore also $\widetilde{\boldsymbol{D}}_m$) is equal to the zero vector.

This method of minimizing $\| \widetilde{\boldsymbol{U}} \|$ can also be considered as a generalization of a coarse-mesh method. The connection with universal algorithms will be given in the next paragraph.

4. Approximate solution of $X = AX + f$ by a component suppression in the error vector $\widetilde{D}$

Up to now we have considered any system of vectors $\boldsymbol{X}_K$ $(K = 0, 1, \ldots m)$. Now assume that the vectors $\boldsymbol{X}_K$ $(K = 1, 2, \ldots m)$ are generated from $\boldsymbol{X}_0$ with the help of the iterative procedure (3). Then these relations hold:

$$\boldsymbol{D}_n = \boldsymbol{A}^n \boldsymbol{D}_0 \,, \tag{28}$$

$$\boldsymbol{U}_n = \boldsymbol{A}^n \boldsymbol{U}_0 \,, \tag{29}$$

$$\boldsymbol{U}_n = \boldsymbol{X}_{n+1} - \boldsymbol{X}_n \,. \tag{30}$$

From eqs. (12), (13), (20) and eqs. (28), (29)

$$\widetilde{\boldsymbol{D}}_m = \mathrm{P}_m(\boldsymbol{A}) \boldsymbol{D}_0 \,, \tag{31}$$

$$\tilde{U}_m = \mathrm{P}_m(A)\, U_0 \,, \tag{32}$$

where

$$\mathrm{P}_m(A) = \sum_{K=0}^{m} C_K A^K \,. \tag{33}$$

For the sake of simplicity we assume that the eigenvalues of A are simple and distinct (generally, they may also be complex and greater than 1 in absolute value). The eigenvalues λ_k and eigenvectors φ_k satisfy

$$A\,\varphi_k = \lambda_k\,\varphi_k \,, \qquad k = 1, 2, \ldots M \,, \tag{34}$$

with

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots |\lambda_M| \,. \tag{35}$$

The vectors D_0 and U_0 can be expanded in terms of the φ_k:

$$D_0 = \sum_{L=1}^{M} d_{0L}\varphi_L \,, \tag{36}$$

$$U_0 = \sum_{L=1}^{M} u_{0L}\,\varphi_L \,. \tag{37}$$

Then from eqs. (31), (32), (33) and (34)

$$\tilde{D}_m = \sum_{L=1}^{M} \mathrm{P}_m(\lambda_L) d_{0L}\,\varphi_L \,, \tag{38}$$

$$\tilde{U}_m = \sum_{L=1}^{M} \mathrm{P}_m(\lambda_L)\, u_{0L}\,\varphi_L \,. \tag{39}$$

The minimizing of $\|\tilde{U}_m\|$ and $\|\tilde{D}_m\|$ as a function of the parameters C_K ($K = 0, 1, \ldots m$) under the restriction (20) leads to a search for polynomials $P_m(\lambda)$ which tend to or are equal to zero for all eigenvalues of the matrix A. If we use as $P_m(A)$ the characteristic polynomial of A, then the vectors $\tilde{D}_m$ and $\tilde{U}_m$ will have all components equal to zero, and $\tilde{X} = X$. However, for large M the use of the characteristic polynomial is practically impossible.

Let us suppose that the vector X_0 has been generated from some initial vector Z_0 with the help of n simple iterations according to eq. (3) (it means $X_0 = Z_n$). To the vector Z_0 correspond the two error vectors $\mathcal{D}_0$ and $\mathcal{U}_0$:

$$\mathcal{D}_0 = \sum_{L=1}^{M} d'_{0L}\varphi_L \,, \tag{40}$$

$$\mathfrak{U}_0 = \sum_{L=1}^{M} u'_{0L} \varphi_L \ . \tag{41}$$

Since $X_0 = Z_n$ from eqs. (40) and (41),

$$D_0 = \sum_{L=1}^{M} d'_{0L} \lambda_L^n \varphi_L \ , \tag{42}$$

$$U_0 = \sum_{L=1}^{M} u'_{0L} \lambda_L^n \varphi_L \ , \tag{43}$$

Let us consider some integer number N such that $M > N \geqslant 1$, and $|\lambda_{N+1}| < |\lambda_N|$, $|\lambda_{N+1}| < 1$. The vectors D_K and U_K ($K = 0, 1, ...$) with respect to (42), (43), (28) and (29) can be written as

$$D_K = \sum_{L=1}^{N} d'_{0L} \lambda_L^{n+K} \varphi_L + \eta_K^D \ , \tag{44}$$

$$U_K = \sum_{L=1}^{N} u'_{0L} \lambda_L^{n+K} \varphi_L + \eta_K^U \ . \tag{45}$$

Since $|\lambda_{N+1}| < 1$, $\lim_{r \to \infty} |\lambda_{N+1}^r| = 0$; therefore for any $\epsilon > 0$ and all K ($K = 0, 1, ...$) there exists an $n > 0$ such that $\| \eta_K^D \| < \epsilon$ and $\| \eta_K^U \| < \epsilon$. Moreover, if at least for one L ($1 \leqslant L \leqslant N$) the coefficients d'_{0L} and u'_{0L} are different from zero, then also $\| \eta_K^D \| < \epsilon \| D_K \|$ and $\| \eta_K^U \| < \epsilon \| U_K \|$. It means that for sufficiently large N the dominantn parts of the vectors D_K and U_K ($K = 0, 1, ...$) belong to the N-dimensional vector space defined by the basis vectors φ_L ($L = 1, 2, ... N$). If the vectors η_K^D and η_K^U ($K = 0, 1, ...$) are exactly equal to zero, then $N + 1$ vectors U_K ($K = 0, 1, ... N$) and, analogously also $N + 1$ vectors D_K must be linearly dependent. We can set $m = N$. The coefficients C'_K giving a linear combination of vectors U_K equal to zero can be redefined with the help of eq. (27). Then $\tilde{X}_m$ will be equal to the exact solution X.

Practically, the vectors η_K^D and η_K^U are always different from zero. Then from eq. (25) the calculated coefficients C_K give a polynomial $P_m(\lambda)$ approximating the annihilation polynomial $(\lambda - \lambda_1)(\lambda - \lambda_2)...(\lambda - \lambda_m)$ for the greatest eigenvalues $\lambda_1, \lambda_2, ... \lambda_m$.

In the proposed algorithm the positions of the greatest eigenvalues of A in the complex plane are taken into account. Therefore (in contrast to other universal algorithms), it is possible to use iterative matrices A with complex eigenvalues or with eigenvalues greater in absolute value than 1. Our method can be considered also as a generalization of Lyusternik's method (see [2]), in which a simple polynomial of first order is used to approximately annihilate the component ψ_1 (corresponding to eigenvalue λ_1) in the error vector.

If the vector $\tilde{X}_m$ is not a sufficiently good approximation to the exact solution of eq. (1), we can set $Z_0 = \tilde{X}_m$, and the calculation of X_0 and $\tilde{X}_m$ can be repeated. Generally, we can define an iteration chain $(n_1, m_1; n_2, m_2; ... n_K, m_K)$. Such a chain means that after $n_1 + m_1$ simple iterations have been executed, the last m_1 will be linearly combined with the help of eqs. (8) and (19) or (25). Then $\tilde{X}_{m_1}$ will be used as an initial vector for the next $n_2 + m_2$ simple iterations (3) etc.

If the distribution of eigenvalues of A in the complex plane is unknown, the optimal iteration chain should be determined empirically. According to our experience, the relation $n_K > m_K$ should always hold.

5. Test results

The proposed method was first used in neutron transport theory for the iterative calculation of interface currents which are generated through a given source distribution in a nuclear fuel element [1], [4]. The dimension of the iterative matrix was about 3000. With this method the number of simple iterations (3) can be reduced 3 to 5 times on the average. Schaerer and Mika [3] have applied this method with success to the solution of the diffusion equation.

We present here some simple examples with iterative matrices having only real eigenvalues. We set $M = 50$ and define the matrix A as

$$A = T^{-1} \mathcal{A} T ,$$

where

$$T_{JK} = J^2 - K + 3000\delta_{JK} , \qquad J, K = 1, 2, \ldots M ,$$

and $\mathcal{A}$ is a diagonal matrix with elements

$$\mathcal{A}_{JJ} = R_1 + (R_2 - R_1)J/M , \qquad J = J_0 + 1, J_0 + 2, \ldots M .$$

Clearly, the diagonal elements of $\mathcal{A}$ are also eigenvalues of the matrix A. The exact solution X has components

$$X_J = -21 + J , \qquad J = 1, 2, \ldots M .$$

The initial iterate X_0 is taken with components

$$X_{0J} = 100\delta_{21,J} \qquad J = 1, 2, \ldots M .$$

Further information about test examples are given in table 1.

We shall compare the efficiency of our method with the Chebyshev polynomial method. In the latter procedure the generated iterative vectors can be written [2] as

$$X_{n+1} = \omega_{n+1}(AX_n + f - X_{n-1}) + X_{n-1} ,$$

where

$$\omega_1 = 1 ,$$

Table 1
Examples of iterative calculations–comparison of various iterative methods

Example	1	2	3	4	5
J_0	2	0	2	3	2
R_1	−0.5	−0.6	−0.6	0.0	−0.5
R_2	0.5	0.6	0.6	0.7	0.5
$\mathcal{A}_{11}$	0.99		1.2	1.2	0.999
$\mathcal{A}_{22}$	−0.99		−1.5	−1.3	−0.99
$\mathcal{A}_{33}$				−2.0	
$\|D_0\|$	1.4637×10^2	1.4637×10^2	1.4637×10^2	1.4637×10^2	1.4637×10^2
$\|D_3\|$	3.591×10^2	6.945×10^1	9.799×10^2	$2.147 \times 10^{+3}$	3.642×10^2
$\|U_3\|$	4.861×10^2	2.055×10^2	1.414×10^3	$3.142 \times 10^{+3}$	4.863×10^2
Number of Chebyshev iterations	104	14			200
$\|\tilde{D}\|$	9.3795×10^{-4}	4.9282×10^{-4}			7.7597×10^{-2}
$\|\tilde{U}\|$	1.3856×10^{-3}	1.3796×10^{-3}			9.7530×10^{-2}
Iteration chain	(13,4; 13,4; 3)	(12,5; 3)	(22,2; 22,2; ...)	(12,4; 12,4; 3)	(12,4; 12,4; 12,4; 12,4; 3)
$\|\tilde{D}\|$	5.8407×10^{-4}	1.2563×10^{-4}	Divergent	3.4552×10^{-6}	6.8668×10^{-5}
$\|\tilde{U}\|$	2.2036×10^{-5}	1.0958×10^{-4}		1.7852×10^{-6}	1.1505×10^{-7}
Iteration chain		(12,2; 8)	(12,4; 12,4; 4)		
$\|\tilde{D}\|$		6.3666×10^{-4}	3.7524×10^{-5}		
$\|\tilde{U}\|$		1.7178×10^{-3}	1.5279×10^{-5}		

$$\omega_{n+1} = 2C_n(1/\rho)/\rho C_{n+1}(1/\rho), \qquad n = 1, 2, \ldots,$$

and

$$C_n(y) = \cos(n \arccos y) \quad \text{for} \quad -1 \leqslant y \leqslant 1.$$

We have used the parameter ρ equal to $|\lambda_1| + 0.00001$.

Table 1 shows results for fice examples. The number of Chebyshev iterations has been given directly. The total number of iterations which are necessary for our method should be calculated as a sum of numbers n_i, m_i characterizing the iteration chain $(n_1, m_1; n_2, m_2; \ldots)$. The choice of this iteration chain has great importance for the efficiency of our method. If the iterative procedure 3 is divergent, the parameters m_i should be greater than the number of eigenvalues that have moduli greater than 1 (see example 3).

6. Conclusion

In agreement with numerical experiments it follows from the theory in section 4 that the proposed method of convergence accelaration is very effective if a group of a few of the eigenvalues of A with largest moduli is well separated from all other eigenvalues. The advantage of the proposed method is the possible use for divergent iterative schemes and also for iterative matrices complex eigenvalues. With respect to the limited precision of the computer, the eigenvalues of A whose moduli are greater than 1 should not be very large (according to our experience not greater

than 6). Our proposed method can be combined also with other acceleration methods. Instead of simple iterations, we can generate a system of iterative vectors for example with the help of Chebyshev polynomials. Such a system of vectors defines a basis in which an optimal linear combination $\tilde{X}$ can be defined by eq. (19 or (25). Numerically the use of (25) seems be more suitable.

Acknowledgement

I am very grateful to R.A. Rosanoff for urging me to pursue this research and to Professor J. Mika for many useful discussions on the subject of this paper.

References

[1] M. Mesina, Dissertation, Univ. Stuttgart, 1975.
[2] D.K. Fadeev and V.N. Fadeeva, Computational methods of linear algebra (Freeman, London, 1963).
[3] D. Schaerer, On a new approach to the residual polynomial method in linear algebra (Computing Center Cyfronet, Swierk, Sept. 1974).
[4] M. Mesina and D. Emendörfer, Transmission probability method for neutron transport calculations in non-uniform reactor lattices, Atomkernenergie 26 (1975) 163–168.