On the Iterative Solution of the Matrix Equation $AX^2 - I = 0$

By Pentti Laasonen

Introduction. The general eigenvalue problem

$$(\lambda A - B)x = 0$$

with square matrices A and B often appears in numerical analysis, and it is frequent that A and B are symmetric and A positive definite, so that the problem is equivalent to the following one:

$$(\lambda I - C)y = 0,$$

where

$$C = A^{-\frac{1}{2}}BA^{-\frac{1}{2}},$$

$$y = A^{\frac{1}{2}}x.$$

It is clear that C is symmetric. However, in order to avoid the solution of two special eigenvalue problems, it is desirable to have a way of forming the transforming matrix $A^{-\frac{1}{2}}$ in terms of the known matrix A without a previous determination of the eigenvalues and eigenvectors of A. Indeed, this is possible by an iterative process, based on Newton's algorithm, for any matrix A with real positive eigenvalues.

Newton's method. Two simple algorithms can be presented for the iterative solution of the scalar equation

$$ax^2 - 1 = 0$$

with a positive a; these are

$$x_{i+1} = \tfrac{3}{2}x_i - \tfrac{1}{2}ax_i^3,$$

and

$$x_{i+1} = \tfrac{1}{2}x_i + \frac{1}{2ax_i}.$$

Since the first algorithm does not always converge, the second one, which always converges for a nonvanishing initial value, will be suitably extended to a particular class of matrices. The conditions under which such extension can be accomplished are described in the following theorem.

THEOREM. *Let A denote a real square matrix with real, positive eigenvalues. Then the algorithm*

(1)
$$X^{(0)} = kI$$
$$X^{(i+1)} = \tfrac{1}{2}X^{(i)} + \tfrac{1}{2}(AX^{(i)})^{-1}$$

Received 30 October 1957. The preparation of this paper was sponsored by the Office of Naval Research, U. S. Navy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

where k is a non-zero constant, generates a sequence of matrices which converges to that solution of

$$AX^2 - I = 0 \tag{2}$$

which has positive eigenvalues. Moreover, the rate of convergence is quadratic.

Preliminary remarks. Some simple properties of a class J of Jacobi matrices will be useful in establishing the results of this paper. Such matrices are of the form

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \cdots & \alpha_s \\ 0 & \alpha_0 & \alpha_1 \cdots & \alpha_{s-1} \\ 0 & 0 & \alpha_0 \cdots & \alpha_{s-2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \alpha_0 \end{bmatrix} = [\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_s]. \tag{3}$$

If two such matrices (denoted by bracket expressions as above) are multiplied, then

$$\begin{aligned} (4)\quad & [\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_s]\cdot[\beta_0, \beta_1, \beta_2, \cdots, \beta_s] \\ & = [\alpha_0\beta_0, \alpha_0\beta_1 + \alpha_1\beta_0, \alpha_0\beta_2 + \alpha_1\beta_1 + \alpha_2\beta_0, \cdots, \alpha_0\beta_s + \alpha_1\beta_{s-1} + \cdots + \alpha_s\beta_0], \end{aligned}$$

so the class J is closed under the multiplication; moreover, multiplication of matrices of this class is commutative. Finally, the inverse of a matrix in J, if it exists, is again in J.

The following simple result will also be useful.

LEMMA. *Let*

$$\{\alpha^{(i)}\}, \quad \{\beta^{(i)}\}, \quad \{\epsilon^{(i)}\} \quad (i = 0, 1, 2, \cdots)$$

denote three sequences of real numbers. Then if they are related by the relationship

$$\alpha^{(i+1)} = \epsilon^{(i)}\alpha^{(i)} + \beta^{(i)},$$

and if the sequences $\{\beta^{(i)}\}$ and $\{\epsilon^{(i)}\}$ have limits β and 0, respectively, the sequence $\{\alpha^{(i)}\}$ converges to the limit β.

The *proof* is simple, consisting first of a finite estimate for the upper bound of $\alpha^{(i)}$, whereafter the final conclusion is evident.

Finally, this section concludes with the well known Jordan matrix theorem. This theorem states that any square matrix A can be transformed by a suitable non-singular matrix U into the Jordan normal form

$$UAU^{-1} = D = \begin{bmatrix} \Lambda_1 & & & & \\ & \Lambda_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & \Lambda_m \end{bmatrix}, \tag{5}$$

where all elements outside the diagonal submatrices Λ_μ are zero and these sub-

matrices have the form

$$\Lambda_\mu = \begin{bmatrix} \lambda_\mu & 1 & 0 & & & 0 & 0 \\ 0 & \lambda_\mu & 1 & & & 0 & 0 \\ 0 & 0 & \lambda_\mu & & & 0 & 0 \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ 0 & 0 & 0 & & & \lambda_\mu & 1 \\ 0 & 0 & 0 & & & 0 & \lambda_\mu \end{bmatrix}.$$

Convergence of the algorithm. Define the matrix $Y^{(i)}$ by

$$Y^{(i)} = UX^{(i)}U^{-1}.$$

Evidently the algorithm for $Y^{(i)}$:s is defined by the formula

$$Y^{(i+1)} = \tfrac{1}{2}Y^{(i)} + \tfrac{1}{2}(DY^{(i)})^{-1}.$$

Now, suppose that one of the matrices $Y^{(i)}$, say $Y^{(r)}$, has the special Jacobi form

$$Y^{(r)} = \begin{bmatrix} \Gamma_1^{(r)} & & & & \\ & \Gamma_2^{(r)} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & \Gamma_m^{(r)} \end{bmatrix}, \tag{6}$$

where the square submatrices $\Gamma_\mu^{(r)}$ along the diagonal have the same orders as the corresponding submatrices Λ_μ and, moreover, belong to the class J defined by (3). It is then obvious that this particular character will be maintained by all successive matrices $Y^{(r+1)}$, $Y^{(r+2)}$, $\cdots$, and, moreover, the corresponding submatrices $\Gamma_\mu^{(i)}$ are related by

$$\Gamma_\mu^{(i+1)} = \tfrac{1}{2}\Gamma_\mu^{(i)} + \tfrac{1}{2}(\Lambda_\mu\Gamma_\mu^{(i)})^{-1} \quad (\mu = 1, 2, \cdots, m;\ i = r, r+1, \cdots). \tag{7}$$

Now, transform this equation into the form

$$\Lambda\Gamma^{(i)}(2\Gamma^{(i+1)} - \Gamma^{(i)}) = I, \tag{8}$$

where the subscripts have been deleted. Then use the notations

$$\Gamma^{(i)} = [\alpha_0^{(i)}, \alpha_1^{(i)}, \alpha_2^{(i)}, \cdots, \alpha_s^{(i)}],$$
$$\Lambda = [\lambda, 1, 0, \cdots, 0],$$

substitute into the equation (8) and apply the multiplication rule (4) in order to obtain, for the elements $\alpha_k^{(i)}$ and $\alpha_k^{(i+1)}$, the system of equations

$$\lambda\alpha_0^{(i)}(2\alpha_0^{(i+1)} - \alpha_0^{(i)}) = 1, \tag{9}$$

$$2\lambda\sum_{\nu=0}^{k}\alpha_\nu^{(i)}\alpha_{k-\nu}^{(i+1)} + 2\sum_{\nu=0}^{k-1}\alpha_\nu^{(i)}\alpha_{k-\nu-1}^{(i+1)} - \lambda\sum_{\nu=0}^{k}\alpha_\nu^{(i)}\alpha_{k-\nu}^{(i)} - \sum_{\nu=0}^{k-1}\alpha_\nu^{(i)}\alpha_{k-\nu-1}^{(i)} = 0 \tag{10}$$
$$(k = 1, 2, \cdots, s).$$

The equation (9) gives for $\alpha_0^{(i+1)}$ the formula

$$\alpha_0^{(i+1)} = \tfrac{1}{2}\alpha_0^{(i)} + \frac{1}{2\lambda\alpha_0^{(i)}}.$$

Hence, if the eigenvalue $\alpha_0^{(r)}$ of $\Gamma^{(r)}$ is positive, then the eigenvalues $\alpha_0^{(r+1)}$ $\alpha_0^{(r+2)}$, $\cdots$ converge to the positive value $\lambda^{-\frac{1}{2}}$. The equations (10) may be used for the successive computation of the elements $\alpha_1^{(i+1)}, \alpha_2^{(i+1)}, \cdots, \alpha_s^{(i+1)}$. Indeed, if one solves the equation (10) for the element $\alpha_k^{(i+1)}$, one obtains the expression

$$\alpha_k^{(i+1)} = \left(1 - \frac{\alpha_0^{(i+1)}}{\alpha_0^{(i)}}\right)\alpha_k^{(i)} + \frac{1}{\alpha_0^{(i)}} P(\alpha_0^{(i)}, \alpha_1^{(i)}, \cdots, \alpha_{k-1}^{(i)}; \alpha_1^{(i+1)}, \cdots, \alpha_{k-1}^{(i+1)}),$$

where P is a second order polynomial of its arguments.

An appeal to the lemma mentioned above justifies the conclusion that the elements $\alpha_k^{(i)}$ tend to a definite limit, as i tends to infinity. It is easy to find that the limit values are

$$\lim_{i\to\infty} \alpha_k^{(i)} = (-1)^k \frac{(2k)!}{2^{2k}(k!)^2} \lambda^{-\frac{1}{2}-k}.$$

This implies the conclusion that the sequence of the matrices Y_i is convergent and the limit matrix has the form

$$Y = \begin{bmatrix} \Gamma_1 & & & & \\ & \Gamma_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & \Gamma_m \end{bmatrix},$$

where the diagonal submatrices Γ_μ are Jacobi matrices of the form (3). The equation

$$Y = \tfrac{1}{2}Y + \tfrac{1}{2}(DY)^{-1}$$

is satisfied by this matrix, and therefore this is the solution of

$$DY^2 = I.$$

Similarly, the matrices $X^{(i)}$ tend to a limit matrix X whose transform by U is Y:

$$Y = UXU^{-1}.$$

This matrix X is obviously the solution of

$$AX^2 - I = 0. \tag{2}$$

The assumptions made in this proof (namely, first, that one of the matrices $Y^{(i)}$ has the particular form (6); and second, that the eigenvalues of this matrix are positive) are fulfilled by choosing X_0 to be equal to kI, since in this case also Y_0 is equal to kI. Of course this diagonal matrix fulfills the requirements.

From (7) one easily derives, by using the commutative law for matrices of the class J, the equation

$$\Gamma_\mu^{(i+1)} - \Lambda^{-\frac{1}{2}} = \tfrac{1}{2}(\Gamma_\mu^{(i)})^{-1}(\Gamma_\mu^{(i)} - \Lambda^{-\frac{1}{2}})^2.$$

This equation implies the further equation

$$Y^{(i+1)} - D^{-\frac{1}{2}} = \tfrac{1}{2}(Y^{(i)})^{-1}(Y^{(i)} - D^{-\frac{1}{2}})^2,$$

or, finally,

$$X^{(i+1)} - A^{-\frac{1}{2}} = \tfrac{1}{2}(X^{(i)})^{-1}(X^{(i)} - A^{-\frac{1}{2}})^2.$$

This relationship indicates that the rate of convergence of the approximations $X^{(0)}, X^{(1)}, X^{(2)}, \cdots$ to the solution X is quadratic.

The effect of round-off errors. For the quadratic rate of convergence or for the convergence at all of the applied Newton's method, it is essential that the matrices involved have a similar structure in the sense described above; that is, their transforms by the same matrix A have a similar Jacobi structure (6). A simple counter example suffices to show that the algorithm may diverge if applied to matrices not related in this sense. In fact, this happens even if the initial matrix X_0 is arbitrarily close to the correct solution.

Such an example is provided by the matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & m^{-2} \end{pmatrix}; \quad A^{-\frac{1}{2}} = \begin{pmatrix} 1 & 0 \\ 0 & m \end{pmatrix}.$$

Now, if the initial matrix is taken to be

$$X^{(0)} = \begin{pmatrix} 1 & \epsilon \\ 0 & m \end{pmatrix}, \quad (\epsilon \neq 0)$$

then the algorithm leads to the successive approximations

$$X^{(i)} = \begin{pmatrix} 1 & \left(\frac{1-m}{2}\right)^i \epsilon \\ 0 & m \end{pmatrix}, \quad (i = 1, 2, \cdots).$$

This sequence diverges, if $m > 3$, despite the arbitrary closeness of $X^{(0)}$ to the correct solution. In general, it can be shown that in this sense the algorithm (1), if carried out indefinitely, is not stable whenever the ratio of the largest to the smallest eigenvalue of A exceeds the value 9.

This observation raises the question of convergence of the above method, if it is applied numerically with truncated values. In this case, although one starts with an admissible initial matrix $X^{(0)} = kI$, the method will in general lead to a sequence of approximations which do not possess the required character, due to round-off errors. Therefore, if the method is continued indefinitely, it will probably not converge. For this reason it may be appropriate to compare during the process two successive approximations $X^{(i)}$ and $X^{(i+1)}$ and discontinue the algorithm as soon as their difference no longer decreases (by some proper norm). Due to the quadratic rate of convergence, in most cases this finite process does not involve many steps; thus, the influence of the round-off errors is likely not very great.

This is true, in particular, if the numerical calculations have been made with some additional accuracy, which can be assumed to take care of these truncation errors.

However, it seems desirable to have a method for checking the correctness of an approximation and, if necessary, to improve its accuracy. This is, in fact, possible by the following means.

Let $\bar{X}$ be an approximation for the exact solution $X = A^{-\frac{1}{2}}$ and form the error

$$\Delta = \bar{X}A\bar{X} - I.$$

If δ,

$$\delta = A^{-\frac{1}{2}} - \bar{X}, \tag{11}$$

denotes the correction which has to be added to $\bar{X}$, then A can be expressed in terms of δ and $\bar{X}$ as follows.

$$\begin{aligned} A &= (\bar{X} + \delta)^{-2} = [(I + \bar{X}^{-1}\delta)^{-1}\bar{X}^{-1}]^2 \\ &= (\bar{X}^{-1} - \bar{X}^{-1}\delta\bar{X}^{-1} + \cdots)^2 = \bar{X}^{-2} - \bar{X}^{-2}\delta\bar{X}^{-1} - \bar{X}^{-1}\delta\bar{X}^{-2} + \cdots. \end{aligned}$$

Thus,

$$\Delta = -\bar{X}^{-1}\delta - \delta\bar{X}^{-1} + \cdots.$$

The series expansions, which are convergent provided δ is so small that the eigenvalues of $\bar{X}^{-1}\delta$ are less than unity in absolute value, are to be truncated after the linear terms in δ, the error caused thereby being of second order with respect to δ. An approximation of δ within this accuracy may thus be obtained by solving the equation

$$\bar{X}^{-1}\delta + \delta\bar{X}^{-1} = -\Delta = I - \bar{X}A\bar{X}. \tag{12}$$

Accordingly, the succeeding paragraph is devoted to a study of the solvability and practical methods for solving this equation.

The equation $CV + VC = K$.

THEOREM. *Let C and K be two $(n \times n)$-matrices, all eigenvalues of C being different from zero and of the same sign. Then the equation*

$$CV + VC = K \tag{13}$$

has one and only one solution V.

In order to prove this theorem, transform all matrices involved by a non-singular matrix U,

$$\begin{aligned} D &= UCU^{-1}, \\ L &= UKU^{-1}, \\ W &= UVU^{-1}, \end{aligned}$$

such that $D = (d_{i,j})$ attains any Jacobi form, that is, all elements below the diagonal, $d_{i,j}$ with $i < j$, vanish; the diagonal elements $d_{i,i}$ thus obtained are the eigenvalues of C and therefore of same sign, say positive. Accordingly, the equation is equivalent to the equation

$$DW + WD = L.$$

By setting corresponding elements on both sides of this equation equal to one another and using a self-explanatory notation, one obtains the following system of n^2 equations:

$$d_{i,i}w_{i,j} + \sum_{k=i+1}^{n} d_{i,k}w_{k,j} + \sum_{k=1}^{j-1} w_{i,k}d_{k,j} + w_{i,j}d_{j,j} = l_{i,j} \quad (i, j = 1, 2, \cdots, n).$$

Since the coefficient $d_{i,i} + d_{j,j}$ of $w_{i,j}$ is positive, this equation can be considered as a recurrence formula which determines $w_{i,j}$ in terms of the elements on the left lower side of the line through $w_{i,j}$ and parallel to the diagonal. Accordingly, all elements can be determined successively, beginning from the lower left hand corner.

Of course, the numerical solution of the equation (13) can be achieved either by solving, for the elements of V, the linear system which is obtained by setting corresponding elements on both sides of (13) equal to one another, or by first performing the transformations used in the above proof. However, the first method is quite laborious for an arbitrary C, since the number of simultaneous equations involved is n^2. On the other hand, the second method involves actually the diagonalization of the matrix C.

A third method, based on an iteration process, is certainly preferable. The equation (13) is then written as follows:

$$\left(C + \frac{1}{m}V\right)^2 = C^2 + \frac{1}{m}K + \frac{1}{m^2}V^2, \tag{14}$$

where m is any scaling factor. If m is large enough, then a good approximation for $C + \frac{1}{m}V$ is provided by the matrix $\left(C^2 + \frac{1}{m}K\right)^{\frac{1}{2}}$. Subtract C from this matrix in order to obtain an approximation for $\frac{1}{m}V$. This can now, if necessary, be substituted on the right hand side of (14) and the process continued.

The extraction of the square-root of the matrix on the right hand side of (14) can, of course, be achieved by Newton's method in a way similar to the one used for computing the matrix $A^{-\frac{1}{2}}$. For the equation

$$X^2 - A = 0,$$

where A is a matrix with non-negative eigenvalues, that solution X which has no negative eigenvalues is obtained as the limit matrix from the algorithm

$$X_{i+1} = \tfrac{1}{2}X_i + \tfrac{1}{2}AX_i^{-1}, \tag{15}$$

if the first matrix is $X_0 = kI$ with any positive k.

The determination of the correction δ. In applying the method outlined above to the determination of the correction (11) from the equation (12), observe that the neglected terms are of the order $[(\bar{X}^{-1}\delta)^2]$. Accordingly, the truncated equation (12) should properly be replaced by

$$\bar{X}^{-1}\delta + \delta\bar{X}^{-1} = -\Delta + 0[(\bar{X}^{-1}\delta)^2]$$

and therefore the corresponding equation (14) is actually

$$\left(\bar{X}^{-1} + \frac{1}{m}\delta\right)^2 = \bar{X}^{-2} - \frac{1}{m}\Delta + \frac{1}{m}0[(\bar{X}^{-1}\delta)^2] + \frac{1}{m^2}\delta^2.$$

Now, since the third term of the right hand side of this equation is to be neglected, one concludes that the removal of the last term does not cause any additional inaccuracy, if m is large enough. It is sufficient to choose m of the same order of magnitude as the largest eigenvalue of $\bar{X}$ squared or the reciprocal of the smallest eigenvalue of A. Compute the matrix $\left(\bar{X}^{-2} - \frac{1}{m}\Delta\right)^{\frac{1}{2}}$ by means of the algorithm (15); subtract $\bar{X}^{-1}$ from this result; and finally, multiply by m in order to obtain an approximation for the correction (11). Of course, this procedure may be repeated and thus set up a quadratically convergent algorithm which, moreover, is self-correcting.

Summary. It has been proved that for any real $(n \times n)$-matrix A with only positive eigenvalues the algorithm (1), with an initial matrix $X^{(0)} = kI$, converges quadratically to the matrix $A^{-\frac{1}{2}}$ with positive eigenvalues.

In a numerical case this algorithm, if continued indefinitely, may be divergent due to round-off errors, whose influence may increase in geometrical progression. This makes it necessary to stop the process as soon as the difference between two successive results no longer decreases; of course, it is also desirable to have some additional accuracy in the numerical computation to take care of the round-off errors.

Any approximation to $A^{-\frac{1}{2}}$ sufficiently accurate can be improved successively, the rate of convergence of the procedure being quadratic. Each step, however, involves either the solution of a system of n^2 linear equations or the extraction of the square-root of a matrix, which may be achieved by a quadratically convergent iteration process.

Kasarmikatu 2B11
Helsinki, Finland

On Gauss' Speeding Up Device in the Theory of Single Step Iteration

By Alexander M. Ostrowski

1. In this paper we consider throughout *real* numbers, vectors and matrices. In order to solve the linear system

$$(1) \qquad \sum_{\nu=1}^{n} a_{\mu\nu}x_\nu = y_\mu, \quad a_{\mu\mu} = A_\mu \quad (\mu = 1, \cdots, n)$$

Received in the present version 7 October 1957. This work was performed under a contract of the National Bureau of Standards with the American University and University of California at Los Angeles.