COMPUTATIONAL MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 13
PWN–POLISH SCIENTIFIC PUBLISHERS
WARSAW 1984

TWO-SIDED APPROXIMATIONS OF INVERSES, SQUARE ROOTS AND CHOLESKY FACTORS

J. W. SCHMIDT

Department of Mathematics, Technical University of Dresden, Dresden, German Democratic Republic

1. Introduction

For solving the problem $F(x) = 0$ in partially ordered spaces superlinearly convergent methods are known which give a monotonous enclosing of a zero x^* of F,

$$x_0 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant x^* \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_0, \tag{1}$$

see e.g. [1], [3], [5] and [10]. In general these methods are based on the well-known Newton method and on related methods.

This lecture is concerned with the following special matrix problems:

(i) $F(x) = x^{-1} - a = 0$ $\quad (x^* = a^{-1}$, inverse),

(ii) $F(x) = x^2 - a = 0$ $\quad (x^* = a^{1/2}$, square root),

(iii) $F(x) = xx^T - a = 0$ $\quad (x^* = l$, Cholesky factor).

Of course, it is much more appropriate to derive theorems about these concrete problems than to apply theorems referring to the general zero-problem. Furthermore, the original methods mostly have to be adapted to the problem in question.

The aim of the present lecture is to review some recent results in the mentioned matrix problems.

2. Enclosing of inverses

Let R be a complete normed ring with a unit element e and let $K \subset R$ be a cone by which R is partially ordered: $x \leqslant y$ for $x, y \in R$ means that

$y-x \in K$. Let norm and partial ordering be connected by the monotony condition: $0 \leqslant x \leqslant y$ implies $\|x\| \leqslant \|y\|$.

An essential example of a ring possessing the mentioned properties is the set of all real (N, N)-matrices with the natural (elementwise) ordering and the row sum norm.

2.1. For a given $a \in R$ the equation

$$F(x) = x^{-1} - a = 0 \tag{2}$$

is solved by $x^* = a^{-1}$ if the inverse exists. Since the inverses of the Fréchet derivative F' and of a divided difference δF of the operator F are defined by

$$F'(x)^{-1}h = -xhx, \qquad \delta F(x, y)^{-1}h = -yhx,$$

the combined Newton-secant method [12]

$$x_{n+1} = x_n - F'(x_n)^{-1}F(x_n),$$
$$y_{n+1} = y_n - \delta F(x_n, y_n)^{-1}F(y_n)$$

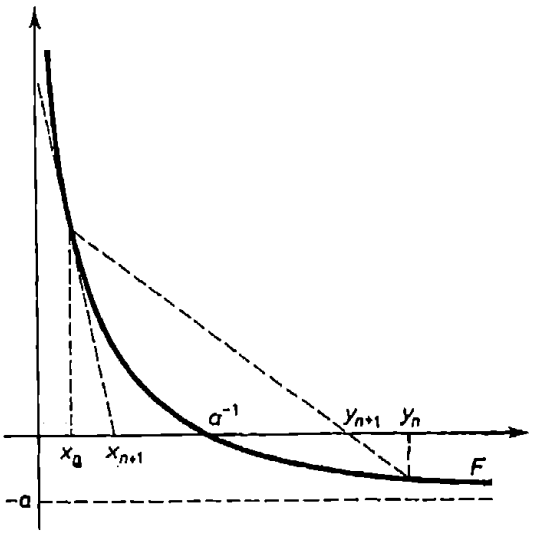

Fig. 1. Method (3) for single-valued functions (i)

applied to (2) immediately leads to the extended Schulz method

$$x_{n+1} = x_n + x_n(e - ax_n),$$
$$y_{n+1} = x_n + y_n(e - ax_n). \tag{3}$$

A first enclosing and convergence theorem concerning method (3) was given by W. Mönch [4] assuming that $a^{-1} \geqslant 0$. Later J. W. Schmidt [7] treated a related method under the condition $a \geqslant 0$. As now will be shown, both the assumptions can be avoided.

THEOREM 1. *Suppose that*

$$x_0 \leqslant y_0, \quad x_0 \leqslant x_1, \quad y_1 \leqslant y_0, \tag{4}$$

$$e - ax_0 \geqslant 0, \tag{5}$$

$$x_0^{-1} \text{ exists and } \|e - ax_0\| < 1. \tag{6}$$

Then the inverse of a is well-defined and the sequences produced by method (3) *ensure the monotonous enclosing*

$$x_0 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant a^{-1} \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_0. \tag{7}$$

Furthermore they converge to a^{-1} *at least with the R-order* 2.

Proof. The enclosing (7) together with $e - ax_n \geqslant 0$ is verified by induction.

To prove the propositions for $n = 1$ set

$$Tx = x_0 + x(e - ax_0) \quad \text{for } x \in R.$$

Because of (6) the operator T is contractive. Furthermore, $x_1 \leqslant y_1$ holds, and the interval $[x_1, y_1]$ is closed because the cone K is so. As easily seen, $Tx \in [x_1, y_1]$ is valid for $x \in [x_1, y_1]$. Therefore, by Banach's fixed-point theorem, T has a unique fixed-point $x^* \in [x_1, y_1]$. Using (6), it follows that x^* is equal to a^{-1}. Thus $x_0 \leqslant x_1 \leqslant a^{-1} \leqslant y_1 \leqslant y_0$ is obtained.

For the step from n to $n+1$ it is seen that

$$x_{n+1} \leqslant x_n + a^{-1}(e - ax_n) = a^{-1},$$

$$y_{n+1} \geqslant x_n + a^{-1}(e - ax_n) = a^{-1}$$

and $e - ax_{n+1} = (e - ax_n)^2 \geqslant 0$. Because of

$$x_{n+1} - x_n = (x_n - x_{n-1})\{(e - ax_{n-1}) + (e - ax_{n-1})^2\},$$

$$y_{n+1} - y_n = (y_n - y_{n-1})\{(e - ax_{n-1}) + (e - ax_{n-1})^2\}$$

the inequalities $x_n \leqslant x_{n+1}$ and $y_{n+1} \leqslant y_n$ are fulfilled. Hence, the enclosing (7) is completely proved.

Finally, the convergence of (x_n) and (y_n) to a^{-1} with the R-order 2 is to be verified. Using $\|e - ax_{n+1}\| \leqslant \|e - ax_n\|^2$, the estimation

$$\|e - ax_n\| \leqslant \|e - ax_0\|^{2^n}$$

can be established. Since $y_{n+1} - x_{n+1} = (y_n - x_n)(e - ax_n)$ and $\|e - ax_n\| < 1$, the inequalities

$$\|y_n - x_n\| \leqslant \|y_0 - x_0\| \quad \text{and} \quad \|y_{n+1} - x_{n+1}\| \leqslant \|y_0 - x_0\| \, \|e - ax_0\|^{2^n}$$

hold, and hence the sequence $(y_n - x_n)$ converges to 0 with the R-order 2. Because of $0 \leqslant a^{-1} - x_n \leqslant y_n - x_n$, $0 \leqslant y_n - a^{-1} \leqslant y_n - x_n$ and the monotony of the norm the proof of the convergence properties is complete.

2.2. For constructing initial elements x_0 and y_0 according to Theorem 1 condition (5) is crucial. Therefore, the following method for solving (2) should be preferred, because no assumption of the type (5) is required,

$$\begin{aligned} m_{n+1} &= m_n + m_n(e - am_n), \\ r_{n+1} &= r_n |e - am_n|. \end{aligned} \tag{8}$$

Here, in addition, R is assumed to be a lattice where the absolute value is defined by $|x| = \sup(x, -x)$ for $x \in R$. With the abbreviations

$$x_n = m_n - r_n, \quad y_n = m_n + r_n \tag{9}$$

from (8) one obtains

$$\begin{aligned} x_{n+1} &= m_n + x_n(e - am_n)^+ - y_n(e - am_n)^-, \\ y_{n+1} &= m_n + y_n(e - am_n)^+ - x_n(e - am_n)^- \end{aligned}$$

where $x^+ = \sup(x, 0)$ and $x^- = \sup(-x, 0)$ for $x \in R$. Now, the following statement is valid.

THEOREM 2. *Assume that*

$$r_0 \geqslant 0, \quad |m_0 - m_1| \leqslant r_0 - r_1, \tag{10}$$

$$m_0^{-1} \text{ exists and } \|e - am_0\| < 1. \tag{11}$$

Then a can be inverted and the sequences (m_n) *and* (r_n) *of method* (8) *provide the monotonous inclusion*

$$|m_n - a^{-1}| \leqslant r_n \leqslant r_{n-1} \leqslant \ldots \leqslant r_0 \tag{12}$$

or, using (9),

$$x_0 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant a^{-1} \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_0. \tag{13}$$

If, in addition,

$$\||e - am_0|\| < 1, \tag{14}$$

the sequences (m_n), (x_n) *and* (y_n) *converge to* a^{-1} *at least with the R-order* 2 *and* (r_n) *converges to* 0 *at least with the same R-order.*

A proof of Theorem 2 is given by J. W. Schmidt [9] and will not be repeated here. For related results with regard to the simplified method

$$\begin{aligned} m_{n+1} &= m_0 + m_n(e - am_0), \\ r_{n+1} &= r_n |e - am_0| \end{aligned} \tag{15}$$

see the earlier paper [8].

Since a^{-1} commutes with a, it is interesting that only the approximations m_n commute with a if m_0 does so.

2.3. In the case where R is the ring of the real (N, N)-matrices, initial matrices can be constructed if m_0 is a sufficiently good approximation to a^{-1}, see [8]. For this, let c be a matrix with positive elements,

$$c_{ik} = (c)_{ik} > 0 \quad \text{for } i, k = 1, \ldots, N,$$

and assume

$$(c - c|e - am_0|)_{ik} > 0 \quad \text{for } i, k = 1, \ldots, N.$$

Then, as is immediately seen, the conditions (10) are satisfied if

$$r_0 = \xi c \quad \text{with} \quad \xi = \max_{i,k=1,\ldots,N} \frac{(|m_0(e - am_0)|)_{ik}}{(c - c|e - am_0|)_{ik}}. \tag{16}$$

Notice that (10) is equivalent to $x_0 \leqslant y_0$, $x_0 \leqslant x_1$ and $y_1 \leqslant y_0$.

2.4. The method (8) is closely related to the interval method of G. Alefeld and J. Herzberger [1]

$$X_{n+1} = X_n \cap \{m(X_n) + X_n(e - am(X_n))\} \tag{17}$$

where X_n is an interval matrix and $m(X_n)$ the corresponding midpoint matrix. Using the interval rules on midpoint and radius matrices, the method without intersection

$$X_{n+1} = m(X_n) + X_n(e - am(X_n))$$

is seen to be a special case of (8). If

$$m(X_\nu) + X_\nu(e - am(X_\nu)) \subset X_\nu \quad \text{for some } \nu,$$

with regard to Theorem 2, the intersection in method (17) can be avoided for all $n \geqslant \nu$.

2.5. The method (15) can be extended to the case where the element a is not given exactly. Now, it is only assumed that

$$\underline{a} \leqslant a \leqslant \bar{a} \tag{18}$$

with known $\underline{a}, \bar{a} \in R$. Let m be an approximation to all corresponding a^{-1} and suppose that elements $\underline{b}, \bar{b}, \underline{h}, \bar{h} \in R$ are available with

$$0 \leqslant \underline{b} \leqslant (e - am)^+ \leqslant \bar{b}, \quad 0 \leqslant \underline{h} \leqslant (e - am)^- \leqslant \bar{h} \tag{19}$$

for all a with (18). In this situation the method

$$
\begin{aligned}
x_{n+1} &= m + x_n^+ \underline{b} - x_n^- \overline{b} - y_n^+ \overline{h} + y_n^- \underline{h}, \\
y_{n+1} &= m + y_n^+ \overline{b} - y_n^- \underline{b} - x_n^+ \underline{h} + x_n^- \overline{h},
\end{aligned} \tag{20}
$$

which is due to J. W. Schmidt [8], has the following properties.

THEOREM 3. *Assume that*

$$x_0 \leqslant y_0, \quad x_0 \leqslant x_1, \quad y_1 \leqslant y_0, \tag{21}$$

$$m^{-1} \text{ exists and } \|\overline{b}\| + \|\overline{h}\| < 1. \tag{22}$$

Then all a with (18) *are invertible and method* (20) *produces the monotonous enclosing*

$$x_0 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant a^{-1} \leqslant y_n \leqslant y_{n-1} \leqslant \cdots \leqslant y_0, \tag{23}$$

valid for all elements a with (18).

For a proof of Theorem 3 see [8] where, in addition, the construction (16) of initial matrices is extended to method (20).

2.6. To demonstrate method (20) let

$$a = \begin{bmatrix} 99 & 99 & 98 \\ 99 & 98 & 98 \\ 98 & 98+\varepsilon & 97+\delta \end{bmatrix}, \qquad \varepsilon, \delta \in [-0.001, +0.001].$$

Set $x^* = \lim x_n$ and $y^* = \lim y_n$. Then, by method (20) one gets the inclusion $x^* \leqslant a^{-1} \leqslant y^*$ for all matrices a in question, where now

$$x^* = \begin{bmatrix} -108.77\ldots & 0.89\ldots & 87.23\ldots \\ 1 & -1 & 0 \\ 87.12\ldots & -0.11\ldots & -109.88\ldots \end{bmatrix},$$

$$y^* = \begin{bmatrix} -87.23\ldots & 1.11\ldots & 108.77\ldots \\ 1 & -1 & 0 \\ 108.88\ldots & 0.11\ldots & -88.12\ldots \end{bmatrix}.$$

The optimal interval enclosing $\overline{x} \leqslant a^{-1} \leqslant \overline{y}$ by means of $\overline{x} = \inf\{a^{-1}\}$ and $\overline{y} = \sup\{a^{-1}\}$ is a little sharper because of

$$\overline{x} = \begin{bmatrix} -108.77\ldots & 0.89\ldots & 89.17\ldots \\ 1 & -1 & 0 \\ 89.08\ldots & -0.11\ldots & -109.88\ldots \end{bmatrix},$$

$$\bar{y} = \begin{bmatrix} -89.17\ldots & 1.11\ldots & 108.77\ldots \\ 1 & -1 & 0 \\ 108.88\ldots & 0.11\ldots & -90.08\ldots \end{bmatrix}.$$

However, here the diameter of the optimal interval $\|\bar{y}-\bar{x}\|$ is considerably greater than the loss of sharpness $\max\{\|\bar{x}-x^*\|, \|\bar{y}-y^*\|\}$.

3. Enclosing of square roots

Let a be a real symmetric, nonnegative definite (N, N)-matrix with the elements $a_{ik} = (a)_{ik}$. Then there exists a unique real symmetric, nonnegative definite matrix $a^{1/2}$ such that

$$a^{1/2}a^{1/2} = a,$$

and $a^{1/2}$ is called the square root of a. No finite method for computing $a^{1/2}$ is known.

3.1. In connection with the enclosing of square roots especially two kinds of partial orderings are of interest. The definite ordering denoted by "$\leqslant$" and generated by the cone of real symmetric, nonnegative definite matrices seems to be more convenient for theoretical manipulation than the natural ordering "$\leqslant$"; the first one is, however, of less numerical utility. Now, by an important theorem of W. Burmeister [2], starting from a given inclusion with respect to the definite ordering, it is always possible to construct an enclosing with respect to the natural ordering. The main step in this direction is the following result.

THEOREM 4. *Let the matrices $r \geqslant 0$ and m be given. Then, the interval of matrices relative to the definite ordering*

$$I = \{x:\ m-r \leqslant x \leqslant m+r\} \tag{24}$$

is optimally enclosed by an interval with respect to the natural ordering whose bounds are given by

$$(\leqslant)\text{-inf}\, I = m-s, \quad (\leqslant)\text{-sup}\, I = m+s \tag{25}$$

where

$$s_{ik} = \sqrt{r_{ii}r_{kk}} \quad \text{for} \quad i, k = 1, \ldots, N. \tag{26}$$

For a proof of this theorem see [2]. Now, Theorem 4 immediately leads to the announced result, see [2].

THEOREM 5. *The monotone enclosing with respect to the definite ordering*

$$x_{n-1} \leqslant x_n \leqslant x^* \leqslant y_n \leqslant y_{n-1} \tag{27}$$

implies the monotone inclusion relative to the natural ordering

$$u_{n-1} \leqslant u_n \leqslant x^* \leqslant v_n \leqslant v_{n-1} \tag{28}$$

where

$$\begin{aligned} m_n &= \tfrac{1}{2}(x_n + y_n), \\ (s_n)_{ik} &= \tfrac{1}{2}\sqrt{(y_n - x_n)_{ii}(y_n - x_n)_{kk}} \quad \text{for} \quad i, k = 1, \ldots, N \end{aligned} \tag{29}$$

and

$$u_n = m_n - s_n, \qquad v_n = m_n + s_n. \tag{30}$$

Proof. Let

$$I_n = \{x\colon x_n \leqslant x \leqslant y_n\} = \{x\colon m_n - r_n \leqslant x \leqslant m_n + r_n\}$$

with $r_n = (y_n - x_n)/2$. Then Theorem 4 gives $(\leqslant)$-inf $I_n = u_n$ and $(\leqslant)$-sup $I_n = v_n$. Since $x^* \in I_n$, it follows that $u_n \leqslant x^* \leqslant v_n$, and because of $I_n \subset I_{n-1}$, using the definition of infimum and supremum, $u_{n-1} \leqslant u_n$ and $v_n \leqslant v_{n-1}$ are obtained.

3.2. It is well known that for a real number $a > 0$ the method

$$y_{n+1} = \tfrac{1}{2}(y_n + a/y_n) \tag{31}$$

provides a quadratically convergent sequence (y_n) with the limit $\sqrt{a}$ for every initial value $y_0 > 0$. Furthermore, using

$$x_n = a/y_n, \tag{32}$$

the monotonous enclosing

$$x_1 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant \sqrt{a} \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_1 \tag{33}$$

holds.

This result can be generalized to the computation of square roots of matrices if the definite ordering is employed, see [2].

THEOREM 6. *Let a and y_0 be real symmetric, positive definite matrices with $ay_0 = y_0 a$. Then the method*

$$x_n = a y_n^{-1}, \qquad y_{n+1} = \tfrac{1}{2}(x_n + y_n) \tag{34}$$

is well-defined, and by the sequences (x_n) and (y_n) the square root $a^{1/2}$ is monotonously enclosed with respect to the definite ordering,

$$x_{n-1} \leqslant x_n \leqslant a^{1/2} \leqslant y_n \leqslant y_{n-1} \quad \text{for } n \geqslant 2. \tag{35}$$

In addition, the sequences converge to $a^{1/2}$ at least with the R-order 2.

Theorem 6, inclusive enclosing (35), is also valid for the method without inversion

$$x_{n+1} = x_n + z_n(a - x_n^2), \quad y_{n+1} = y_n + z_n(a - y_n^2),$$
$$z_{n+1} = z_n + z_n(e - 2z_n y_{n+1}) \tag{36}$$

if $0 \leqslant a \leqslant e$. But now the initial matrices have to be restricted, e.g. to $x_0 = y_0 = e$ and $z_0 = e/2$.

With regard to Theorem 5 the definite enclosing (35) can be transformed into a natural one. The new sequences (u_n) and (v_n) converge to $a^{1/2}$ likewise at least with R-order 2.

3.3. The following numerical example is taken from [2]. Let a be the matrix with the known square root

$$(a^{1/2})_{ik} = (1 + 2|i-k|)^{-1} \quad \text{for } i, k = 1, \ldots, N,$$

and let the initial matrix be $y_0 = 5e$. In general, due to the rounding errors the theoretically assured commutativity of x_n and y_n with a and consequently the enclosing (35) are not preserved during a larger number of steps. Therefore the iteration (34) is stopped if the condition

$$(x_n)_{ii} \leqslant (y_n)_{ii} \quad \text{for } i = 1, \ldots, N$$

necessary for $x_n \leqslant y_n$ is violated. For $N = 30$ the elements in the fifth row and first column are

n	$(x_n)_{51}$	$(y_n)_{51}$	$(u_n)_{51}$	$(v_n)_{51}$
0	0.0893 ...	0.0	–	–
1	0.1279 ...	0.0446 ...	−1.0003 ...	1.1703 ...
2	0.1312 ...	0.0863 ...	−0.3148 ...	0.5324 ...
3	0.1136 ...	0.1087 ...	0.0016 ...	0.2207 ...
4	0.1110 1863 ...	0.1111 9731 ...	0.1000 1492 ...	0.1222 0103 ...
5	0.1111 1425 8912	0.1111 0797 3684	0.1109 4305 9024	0.1112 7917 3572
6	0.1111 1110 5937	0.1111 1111 6299	0.1111 1105 0143	0.1111 1117 2092

Indeed, the elements of x_n and y_n give only an approximate information about the corresponding element of $a^{1/2}$. An inclusion, however, is provided by the elements of u_n and v_n.

4. Enclosing of Cholesky factors

A lower-triangular matrix l with positive diagonal elements is called the *Cholesky factor* of a real symmetric, positive definite matrix a if

$$ll^T = a.$$

The Cholesky factor uniquely exists and can be obtained by the Cholesky method. Here the aim is to give an error estimation for the computed factor by means of a monotonous enclosing with respect to the natural ordering, at first for Stieltjes matrices and then for arbitrary matrices.

4.1. A matrix $a = (a_{ik})$ is defined to be an M-matrix if $a_{ik} \leqslant 0$ for $i \neq k$ and $a^{-1} \geqslant 0$ relative to the natural ordering. A Stieltjes matrix is a real symmetric M-matrix. Stieltjes matrices a are positive definite and the corresponding Cholesky factors l are lower-triangular M-matrices uniquely characterized by

$$l_{ii} > 0,\ l_{ik} \leqslant 0 \quad \text{for } k < i \ (i,\ k = 1, \ldots, N).$$

Applying the Newton type method

$$\begin{aligned} F(y_n) + F'(y_n)(y_{n+1} - y_n) &= 0, \\ F(x_n) + F'(y_n)(x_{n+1} - x_n) &= 0 \end{aligned} \tag{37}$$

to $F(x) = xx^T - a = 0$ the following method is obtained:

$$\begin{aligned} y_n(y_{n+1} - y_n)^T + (y_{n+1} - y_n)y_n^T &= a - y_n y_n^T, \\ y_n(x_{n+1} - x_n)^T + (x_{n+1} - x_n)y_n^T &= a - x_n x_n^T. \end{aligned} \tag{38}$$

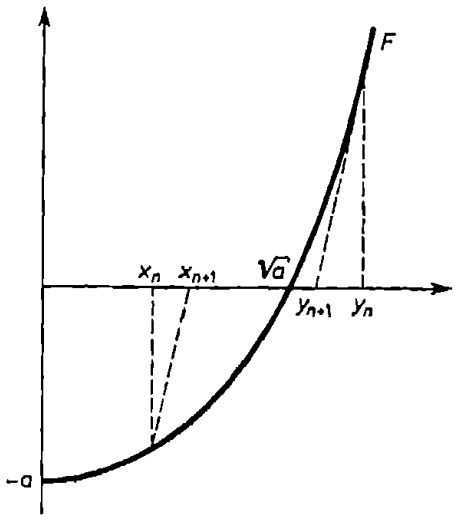

Fig. 2. Method (38) for single-valued functions (iii)

From these linear equations the elements of the lower-triangular matrices y_{n+1} and x_{n+1} can easily be determined, e.g. row by row if $(y_n)_{ii} \neq 0$ for $i = 1, \ldots, N$.

The following result about method (38) is due to J. W. Schmidt and U. Patzke [13].

THEOREM 7. *Let a be a Stieltjes matrix and x_0, y_0 be lower-triangular M-matrices such that*

$$x_0 x_0^T \leqslant a \leqslant y_0 y_0^T. \tag{39}$$

Then, the sequences (x_n) and (y_n) according to (38) are well-defined. They converge to the Cholesky factor l of a at least with the R-order 2 and

$$x_0 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant l \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_0 \tag{40}$$

holds.

The proof [13] is based on three lemmas.

LEMMA 1. *Let x and y be lower-triangular M-matrices. Then*

$$xx^T \leqslant yy^T \text{ implies } x \leqslant y.$$

LEMMA 2. *Let x be a lower-triangular M-matrix and y be a lower-triangular matrix. Then*

$$xy^T + yx^T \geqslant 0 \text{ implies } y \geqslant 0.$$

LEMMA 3. *Let x be a lower-triangular M-matrix, y be a lower-triangular matrix and z be a symmetric matrix. Then*

$$xy^T + yx^T \leqslant z,\ y \geqslant 0 \text{ implies } \|y\| \leqslant \alpha \|z\|,$$

where α is independent of y and z. The norm is defined by

$$\|z\| = \max\{\|z\|_\infty, \|z^T\|_\infty\}$$

where $\|\cdot\|_\infty$ denotes the row sum norm.

Now, a sketch of a proof of Theorem 7 shall be given. At first the enclosing (40) together with $x_n x_n^T \leqslant a \leqslant y_n y_n^T$ is verified by induction.

The assumption (39) in view of Lemma 1 implies the inclusion $x_0 \leqslant l \leqslant y_0$.

In the step from n to $n+1$, because of $l \leqslant y_n \leqslant y_0$ the lower-triangular matrix l_n is an M-matrix. Therefore, the two relations (38) and

$$\begin{aligned} y_n(y_{n+1}-l)^T + (y_{n+1}-l)y_n^T &= (y_n-l)(y_n-l)^T \geqslant 0, \\ y_n(x_{n+1}-l)^T + (x_{n+1}-l)y_n^T &= (y_n-x_n)(x_n-l)^T + (x_n-l)(y_n-l)^T \leqslant 0 \end{aligned}$$

by means of Lemma 2 yield the inequalities $y_{n+1} \leqslant y_n$, $x_{n+1} \geqslant x_n$, $y_{n+1} \geqslant l$, $x_{n+1} \leqslant l$. Finally, the remaining inequalities $y_{n+1}y_{n+1}^T - a \geqslant 0$, $x_{n+1}x_{n+1}^T - a \leqslant 0$ are proved by simple estimations. Therefore (40) holds, and because the Cholesky factor is unique, $\lim x_n = \lim y_n = l$ is true. Using

$$\begin{aligned} &l(y_{n+1}-x_{n+1})^T + (y_{n+1}-x_{n+1})l^T \\ &\qquad \leqslant y_n(y_{n+1}-x_{n+1})^T + (y_{n+1}-x_{n+1})y_n^T = (y_n-x_n)(y_n-x_n)^T, \end{aligned}$$

Lemma 3 ensures that

$$\|y_{n+1}-x_{n+1}\| \leqslant a\,\|y_n-x_n\|^2.$$

Hence the sequences (y_n-x_n) and consequently (x_n) and (y_n) converge at least with the R-order 2.

4.2. The matrix

$$a = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{bmatrix} \Big\} N \text{ rows}$$

is a Stieltjes matrix, and $\operatorname{cond}(a)\to\infty$ for $N\to\infty$.

Let

$$y_0 = \begin{bmatrix} \alpha & & \\ \beta & \alpha & \\ \ddots & \ddots & \ddots \end{bmatrix}, \qquad x_0 = \begin{bmatrix} \gamma & & \\ \delta & \gamma & \\ \ddots & \ddots & \ddots \end{bmatrix}.$$

Then $x_0x_0^T \leqslant a \leqslant y_0y_0^T$ holds if $\alpha^2 \geqslant 2$, $\alpha\beta \geqslant -1$, $\alpha > 0$, $\beta \leqslant 0$ and $\gamma^2+\delta^2 \leqslant 2$, $\gamma\delta \leqslant -1$, $\gamma > 0$, $\delta \leqslant 0$. For $N = 8$ and for $\alpha = 1.42$, $\beta = -0.7$, $\gamma = 1$ and $\delta = -1$ the element of l, e.g. in the eighth row and column, is enclosed by method (38) as follows, see [13]:

n	$(x_n)_{88}$	$(y_n)_{88}$
0	1	1.42
1	1.001 ...	1.181 ...
2	1.020 ...	1.091 ...
3	1.052 ...	1.062 ...
4	1.060 5019 ...	1.060 6654 ...
5	1.060 6601 3921	1.060 6601 7180
6	1.060 6601 7178	1.060 6601 7178

A more general method for the construction of starting matrices according to (39) is described in [13]. There it is assumed that a sufficiently good approximation to l is known.

4.3. Let now a be an arbitrary real symmetric, positive definite (N, N)-matrix. The first equation of the iteration (38)

$$y_ny_{n+1}^T + y_{n+1}y_n^T = a + y_ny_n^T$$

leads to the following method of successive displacements approximating the Cholesky factor l of a:

$$(41)\qquad (y_{n+1})_{kk} = \frac{1}{2}\left\{(y_n)_{kk} + \frac{a_{kk} - \sum_{\nu=1}^{k-1} (y_{n+1})_{k\nu}^2}{(y_n)_{kk}}\right\},$$

$$(y_{n+1})_{ik} = \frac{1}{(y_{n+1})_{kk}}\left\{a_{ik} - \sum_{\nu=1}^{k-1} (y_{n+1})_{i\nu}(y_{n+1})_{k\nu}\right\} \quad \text{for } i > k.$$

Here the elements $(y_{n+1})_{ik}$ of the lower-triangular matrix y_{n+1} are computed, e.g. column by column, $k = 1, \ldots, N$.

U. Patzke [6] modified method (41) under the assumption that a starting matrix y_0 is known with

$$(42)\qquad \operatorname{sign}((y_0)_{ik}) = \operatorname{sign}((l)_{ik}) \quad \text{for } i,\ k = 1, \ldots, N, \quad i > k$$

in order to guarantee monotonous enclosing:

(43) For $k = 1, \ldots, N$ do

$$(y_{n+1})_{kk} = \frac{1}{2(y_n)_{kk}}\left\{a_{kk} - \sum_{\nu=1}^{k-1} (\mu_{k\nu}^{n+1})^2 + (y_n)_{kk}^2\right\},$$

$$(x_{n+1})_{kk} = \frac{1}{(y_{n+1})_{kk}}\left\{a_{kk} - \sum_{\nu=1}^{k-1} (\lambda_{k\nu}^{n+1})^2\right\},$$

$$(y_{n+1})_{ik} = \frac{1}{a_{ik}^{n+1}}\left\{a_{ik} - \sum_{\nu=1}^{k-1} \varrho_{i\nu k}^{n+1}\, \varrho_{k\nu i}^{n+1}\right\} \quad \text{for } i > k,$$

$$(x_{n+1})_{ik} = \frac{1}{\beta_{ik}^{n+1}}\left\{a_{ik} - \sum_{\nu=1}^{k-1} \sigma_{i\nu k}^{n+1}\, \sigma_{k\nu i}^{n+1}\right\} \quad \text{for } i > k,$$

where

$$\mu_{k\nu}^{n+1} = (x_{n+1})_{k\nu}, \quad \lambda_{k\nu}^{n+1} = (y_{n+1})_{k\nu} \quad \text{for } \operatorname{sign}((l)_{k\nu}) \geqslant 0,$$

$$\mu_{k\nu}^{n+1} = (y_{n+1})_{k\nu}, \quad \lambda_{k\nu}^{n+1} = (x_{n+1})_{k\nu} \quad \text{for } \operatorname{sign}((l_{k\nu})) < 0$$

and

$$a_{ik}^{n+1} = (x_{n+1})_{kk}, \quad \beta_{ik}^{n+1} = (y_{n+1})_{kk} \quad \text{for } \operatorname{sign}((l)_{ik}) \geqslant 0,$$

$$a_{ik}^{n+1} = (y_{n+1})_{kk}, \quad \beta_{ik}^{n+1} = (x_{n+1})_{kk} \quad \text{for } \operatorname{sign}((l)_{ik}) < 0$$

and

$$\varrho_{ivk}^{n+1} = (x_{n+1})_{iv}, \quad \sigma_{ivk}^{n+1} = (y_{n+1})_{iv} \quad \text{for } \operatorname{sign}\big((l)_{kv}\big) \geqslant 0,$$

$$\varrho_{ivk}^{n+1} = (y_{n+1})_{iv}, \quad \sigma_{ivk}^{n+1} = (x_{n+1})_{iv} \quad \text{for } \operatorname{sign}\big((l)_{kv}\big) < 0.$$

The following theorem concerning method (43) is proved in [6].

THEOREM 8. *Let a be real symmetric, positive definite and let y_0 be lower-triangular with the property* (42). *If*

$$\begin{aligned} &(x_1)_{kk} > 0 \quad \text{for } k = 1, \ldots, N, \\ &(x_1)_{ik} \geqslant 0 \quad \text{for } i > k \quad \text{and} \quad \operatorname{sign}\big((l)_{ik}\big) \geqslant 0, \\ &(y_1)_{ik} \leqslant 0 \quad \text{for } i > k \quad \text{and} \quad \operatorname{sign}\big((l)_{ik}\big) \leqslant 0, \end{aligned} \tag{44}$$

then the sequences (x_n) and (y_n) are well-defined by (43). *They converge to the Cholesky factor l of a at least with the R-order 2 and l is enclosed monotonously,*

$$x_1 \leqslant \ldots \leqslant x_{n-1} \leqslant x_n \leqslant l \leqslant y_n \leqslant y_{n-1} \leqslant \ldots \leqslant y_1. \tag{45}$$

To show the R-order 2, at first,

$$\|y_{n+1} - x_{n+1}\|_F \leqslant a \|y_n - l\|_F^2$$

with the Frobenius norm $\|\cdot\|_F$ is verified. Then, using $x_{n+1} \leqslant l \leqslant y_{n+1}$, it follows immediately that

$$\|x_{n+1} - l\|_F \leqslant a \|y_n - l\|_F^2,$$

$$\|y_{n+1} - l\|_F \leqslant a \|y_n - l\|_F^2.$$

These estimations imply the R-order 2 of both sequences (x_n) and (y_n), see e.g. [11].

4.4. Let

$$a = \begin{bmatrix} 25 & 5 & -5 & 5 \\ 5 & 17 & 3 & -3 \\ -5 & 3 & 18 & -6 \\ 5 & -3 & -6 & 28 \end{bmatrix}, \quad y_0 = \begin{bmatrix} 3 & 0 & 0 & 0 \\ + & 3 & 0 & 0 \\ - & + & 3 & 0 \\ + & - & - & 3 \end{bmatrix}.$$

Then by the method (43) e.g. for the element in the fourth row and column the following approximations are obtained [6]:

n	$(x_n)_{44}$	$(y_n)_{44}$
1	4.1294 ...	5.7947 ...
2	4.9321 ...	5.0584 ...
3	4.9996 ...	5.0003 ...
4	4.9999 9998	5.0000 0001 ...
5	5.0000 0000 000	5.0000 0000 000

References

[1] G. Alefeld and J. Herzberger, *Einführung in die Intervallrechnung*, Bibliographisches Institut, Mannheim–Wien–Zürich 1974.

[2] W. Burmeister, *Optimal interval enclosing of certain sets of matrices with application to monotone enclosing of square roots*, Computing 25 (1980), 283–295.

[3] N. S. Kurpel and T. S. Kurchenko, *Two-sided methods for the solution of systems of equations* (in Russian), Naukova dumka, Kiev 1975.

[4] W. Mönch, *Monotone Einschließung von positiven Inversen*, Z. Angew. Math. Mech. 53 (1973), 207–208.

[5] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York and London 1970.

[6] U. Patzke, *Ein monoton einschließendes Einzelschrittverfahren für Cholesky-Faktoren*, Beitr. Numer. Math. 11 (1983), 139–146.

[7] J. W. Schmidt, *Monotone Einschließung von Inversen positiver Elemente durch Verfahren vom Schulz-Typ*, Computing 16 (1976), 211–216.

[8] —, *Einschließung inverser Elemente durch Fixpunktverfahren*, Numer. Math. 31 (1978), 313–320.

[9] —, *Monotone Eingrenzung von inversen Elementen durch ein quadratisch konvergentes Verfahren ohne Durchschnittsbildung*, Z. Angew. Math. Mech. 60 (1980), 202–204.

[10] —, *Monotone Einschließungsverfahren bei nichtlinearen Gleichungen*, Mitteilungen Math. Gesellschaft, Heft 1/2 (1982), 62–87.

[11] —, *On the R-order of coupled sequences*, Computing 26 (1981), 333–342.

[12] J. W. Schmidt and H. Leonhardt, *Eingrenzung von Lösungen mit Hilfe der Regula falsi*, Computing 6 (1970), 318–329.

[13] J. W. Schmidt and U. Patzke, *Iterative Nachverbesserung mit Fehlereingrenzung der Cholesky-Faktoren von Stieltjes-Matrizen*, J. Reine Angew. Math. 327 (1981), 81–92.

Presented to the Semester
Computational Mathematics
February 20 – May 30, 1980