Beiträge zur Numerischen Mathematik
*12* (1984), 7–19

# On improving approximate triangular factorizations iteratively

Götz Alefeld and Jon G. Rokne

Summary. Newton's method applied iteratively to the improvement of an approximate triangular factorization of a matrix is discussed in detail. Particular consideration is given to the effect of rounding errors on the convergence of the iteration. It is shown that on a computer employing fixed length floating-point arithmetic Newton's method converges with an arbitrary starting value after $2n-1$ steps to the same value as that obtained by Gaussian elimination. Finally, a new method is proposed for the iterative improvement of bounds for the elements of the triangular factorization where the effects of the rounding errors are also considered.

## 1. Introduction

It is frequently necessary to solve a system of linear equations $Ax = b$ for a variety of right-hand sides $b$. Since this is often done by factoring $A$ as $(I + L^*)\, U^*$ and then solving the resulting simpler sets of equations, it is important to calculate $L^*$ and $U^*$ as accurately as possible.

With this in mind J. W. Schmidt [3] recently proposed to apply Newton's method in order to correct an approximate factorization of a non-singular matrix $A$. Schmidt also proved directly that Newton's method in this case has second order convergence.

In this paper, we show that Newton's method applied to the correction of an approximate factorization yields the exact factorization after a finite number of steps if no rounding errors occur. This is a special case of a more general result: Let the computation be contaminated by the effects of a rounding procedure. Then it turns out that Newton's method gives the exact same result as Gaussian elimination after at most $2n - 1$ iteration steps. (See Theorem 1 below.) Seemingly nothing can be gained using Newton's method in this context.

A new method is now proposed where the elements of $L^*$ and $U^*$ are enclosed by lower and upper bounds. It is shown that the bounds converge to the exact values assuming no rounding errors occur. At each step it is shown that the widths of the enclosing intervals are approximately squared. After proving this a framework for a detailed discussion of rounding errors in this method is presented.

## 2. The iteration method

Suppose we are given a real nonsingular $n$ by $n$ matrix $A$ with rows already arranged in such an order that it possesses a factorization $A = (I + L^*)\, U^*$, where $I$ denotes the identity, $L^*$ is a strictly lower triangular matrix and $U^*$ is an upper triangular matrix with nonvanishing diagonal elements. Denoting by $L$ a strictly lower triangular matrix and by $U$ an upper triangular matrix then, for given $A$,

$$F(L, U) = A - (I + L)\, U$$

defines a nonlinear mapping from $\mathbb{R}^{n^2}$ to $\mathbb{R}^{n^2}$. Solving the equation $F(L, U) = 0$ is therefore equivalent to computing $L^*$ and $U^*$. Given an approximation $(L^{(m)}, U^{(m)})$ to $(L^*, U^*)$ we can use Newton's method, the general form of which is

$$F'(x^{(m)})\,(x^{(m+1)} - x^{(m)}) + F(x^{(m)}) = 0, \qquad m = 0, 1, 2, \ldots,$$

to solve this equation. Since

$$F'(L^{(m)}, U^{(m)})\,(L, U) = -(I + L^{(m)})\, U - L U^{(m)}$$

we get the following iteration method:

$$-(I + L^{(m)})\,(U^{(m+1)} - U^{(m)}) - (L^{(m+1)} - L^{(m)})\, U^{(m)} + A$$
$$-\,(I + L^{(m)})\, U^{(m)} = 0, \qquad m = 0, 1, 2, \ldots$$

This form has been given in [3].

A simple manipulation gives

$$(I + L^{(m)})\, U^{(m+1)} + (L^{(m+1)} - L^{(m)})\, U^{(m)} = A, \qquad m = 0, 1, 2, \ldots \tag{1}$$

For ease of notation we define

$$\Delta L^{(m+1)} = (\Delta l_{ij}^{(m+1)}) := L^{(m+1)} - L^{(m)}.$$

When we perform (1) on a computer using fixed length floating-point arithmetic of relative machine precision eps than the following formulas result:

$$\left.\begin{array}{l}
\text{For } i = 1(1)n\text{:} \\
\qquad \text{For } k = i(1)n\text{:} \\
\text{(a)} \qquad u_{ik}^{(m+1)} = fl\left(a_{ik} - \sum_{j=1}^{i-1} \Delta l_{ij}^{(m+1)} u_{jk}^{(m)} - \sum_{j=1}^{i-1} l_{ij}^{(m)} u_{jk}^{(m+1)}\right). \\
\qquad \text{For } k = i + 1(1)n \ (i < n)\text{:} \\
\text{(b)} \qquad l_{ki}^{(m+1)} = fl\left\{\frac{1}{u_{ii}^{(m)}}\left(a_{ki} - \sum_{j=1}^{i-1} \Delta l_{kj}^{(m+1)} u_{ji}^{(m)} \right.\right. \\
\qquad\qquad \left.\left. - \sum_{j=1}^{i-1} l_{kj}^{(m)} u_{ji}^{(m+1)} + l_{ki}^{(m)}(u_{ii}^{(m)} - u_{ii}^{(m+1)})\right)\right\}.
\end{array}\right\} \tag{2}$$

Here we have assumed that the $n$ by $n$ matrix $A$ may be represented exactly on the computer under consideration. All values computed by (2) are machine numbers.

Furthermore we assume here and in the sequel that floating point expressions are computed in the "natural order" from left to right. For example

$$fl(a_1b_1 + a_2b_2 + a_3b_3) = fl\big(fl(fl(a_1b_1) + fl(a_2b_2)) + fl(a_3b_3)\big).$$

In proving the next theorem we assume that the following rules hold for the floating-point operations. Let $x$ be a machine number and 0 the floating-point zero. Then

$$fl(x - x) = 0, \qquad fl(0 \cdot x) = fl(x \cdot 0) = 0, \qquad fl(x \pm 0) = x.$$

Theorem 1. *Suppose the $n$ by $n$ matrix $A$ has a triangular factorization $A = (I + L^*)\,U^*$ and that $(\tilde{L}, \tilde{U})$ is an approximation to $(L^*, U^*)$ which has been computed by Crout's variant of Gaussian elimination* (see for example [4], p. 166ff.) *using a computer with relative machine precision eps. Then for all starting values which are sufficiently close to $(\tilde{L}, \tilde{U})$, Newton's method* (2) *performed on the same computer gives exactly the same values after at most $2n - 1$ steps. These values remain unchanged by Newton's method. (Sufficiently close means starting values which guarantee the feasibility of the iteration* (2).)

Proof. The nontrivial elements of the matrices $\tilde{L} = (\tilde{l}_{ij})$ and $\tilde{U} = (\tilde{u}_{ij})$ are computed by the following formulas (see, for example, [4], p. 185ff.):

$$\left.\begin{aligned}
&\text{For } i = 1(1)n\text{:}\\
&\qquad \text{For } k = i(1)n\text{:}\\
&\qquad \tilde{u}_{ik} = fl\left(a_{ik} - \sum_{j=1}^{i-1} \tilde{l}_{ij}\tilde{u}_{jk}\right).\\
&\qquad \text{For } k = i + 1(1)n \ (i < n)\text{:}\\
&\qquad \tilde{l}_{ki} = fl\left(\frac{1}{\tilde{u}_{ii}}\left(a_{ki} - \sum_{j=1}^{i-1} \tilde{l}_{kj}\tilde{u}_{ji}\right)\right).
\end{aligned}\right\} \qquad (2')$$

In the first step of Newton's method we get from (2) (a) for $i = 1$

$$u_{1k}^{(1)} = fl(a_{1k}) = a_{1k} = \tilde{u}_{1k} = u_{1k}^*, \qquad 1 \leqq k \leqq n.$$

This means that for arbitrary starting values $(L^{(0)}, U^{(0)})$, the matrix $U^{(1)}$ has the same elements as $\tilde{U}$ in the first row. $U^{(2)}$ has, in the same manner as $U^{(1)}$, the same values in the first row as $\tilde{U}$. Again this follows from (a) for $i = 1$. We therefore get for the first column of $L^{(2)}$

$$l_{k1}^{(2)} = fl\left(\frac{1}{u_{11}^{(1)}}\left(a_{k1} + l_{k1}^{(1)}(u_{11}^* - u_{11}^*)\right)\right) = fl\left(\frac{a_{k1}}{a_{11}}\right) = \tilde{l}_{k1}, \qquad 2 \leqq k \leqq n.$$

This means that after the second iteration step both the first column of $L^{(2)}$ and the first row of $U^{(2)}$ have the same values as those computed by (2'). We now show the following: If the first $i$ rows of $U^{(m)}$ and the first $i$ columns of $L^{(m)}$ have the same values as those computed by (2'), then $U^{(m+1)}$ has at least in the first $i + 1$ rows (and $L^{(m+1)}$ at least in the first $i$ columns) the same values as those computed by (2').

For $i = 0$ this assertion has been shown above. For $i > 0$ it follows by mathematical induction in the following manner. At first we remark that $L^{(m+1)}$ and $U^{(m+1)}$ have the same elements in the first $i$ columns and first $i$ rows as $L^{(m)}$ and $U^{(m)}$ have, that is these elements are identical to those of $\tilde{L}$ and $\tilde{U}$. This follows immediately from (a) and (b). Therefore using (a) we get for row $i+1$ of $U^{(m+1)}$ that

$$u_{i+1,k}^{(m+1)} = fl\left(a_{i+1,k} - \sum_{j=1}^{i} \Delta l_{i+1,j}^{(m+1)} u_{jk}^{(m)} - \sum_{j=1}^{i} l_{i+1,j}^{(m)} u_{jk}^{(m+1)}\right)$$

$$= fl\left(a_{i+1,k} - \sum_{j=1}^{i} \tilde{l}_{i+1,j}\tilde{u}_{jk}\right) = \tilde{u}_{i+1,k}, \qquad i+1 \leqq k \leqq n.$$

To complete the proof we have to show the following: If for some $m \geqq 0$ the first $i$ rows of $U^{(m)}$ and the first $i-1$ columns of $L^{(m)}$ have the same elements as the corresponding rows and columns of $\tilde{L}$ and $\tilde{U}$, then $L^{(m+1)}$ has at least in the first $i$ columns (and $U^{(m+1)}$ at least in the first $i$ rows) the same elements as $\tilde{L}$ (and $\tilde{U}$) have. For $i = 1$ we have proved this assertion above. For $i > 1$ we again remark that it follows from (a) that $U^{(m+1)}$ has at least in the first $i$ rows the same elements as $\tilde{U}$ has. Similarly, it follows from (b) that $L^{(m+1)}$ has at least in the first $i-1$ columns the same elements as $\tilde{L}$ has. For column $i$ of $L^{(m+1)}$ it follows using (b)

$$l_{ki}^{(m+1)} = fl\left(\frac{1}{u_{ii}^{(m)}}\left(a_{ki} - \sum_{j=1}^{i-1} \Delta l_{kj}^{(m+1)} u_{ji}^{(m)} - \sum_{j=1}^{i-1} l_{kj}^{(m)} u_{ji}^{(m+1)} + l_{ki}^{(m)}(u_{ii}^{(m)} - u_{ii}^{(m+1)})\right)\right)$$

$$= fl\left(\frac{1}{\tilde{u}_{ii}}\left(a_{ki} - 0 + \sum_{j=1}^{i-1} \tilde{l}_{kj}\tilde{u}_{ji} + 0\right)\right) = \tilde{l}_{ki}, \qquad i+1 \leqq k \leqq n.$$

After at most $2n-1$ steps we therefore arrive at the pair $(\tilde{L}, \tilde{U})$. □

We therefore have the negative result that nothing is gained using Newton's method (2) in order to improve an approximate factorization computed by Gaussian elimination.

## 3. Iterative improvement of bounds

In this section we propose a method which improves lower and upper bounds for the elements of $L^*$ and $U^*$ iteratively. We suppose that the elements of $L^*$ and $U^*$ are enclosed in compact intervals $L_{ij}^{(0)}$ and $U_{ij}^{(0)}$:

$$l_{ij}^* \in L_{ij}^{(0)}, \qquad u_{ij}^* \in U_{ij}^{(0)}. \tag{3}$$

Such intervals could, for example, be computed from an error estimation or one could compute such inclusion intervals by simply applying Gaussian algorithm and rounding outwards in a systematic way during the elimination process. (See, for example, [1], pp. 218.) We again denote by $F(L, U)$ the mapping from $\mathbb{R}^{n^2}$ to $\mathbb{R}^{n^2}$ defined by

$$F(L, U) = A - (I + L)\,U.$$

Since

$$F(L^*, U^*) - F(L, U) = -(I + L^*)\,(U^* - U) - (L^* - L)\,U$$

it follows that

$$A - (I + L)\,U = (I + \underline{L^*})\,(U^* - U) + (L^* - L)\,U.$$

Solving alternately for a row of $U^* - U$ and for a column of $L^* - L$ and assuming for the moment that the elements of the underlined matrix $L^*$ are known, we get the following identities after some simplification:

For $i = 1(1)n$:

$$u_{ik}^* = a_{ik} - \sum_{j=1}^{i-1} l_{ij}^* u_{jk} - \sum_{j=1}^{i-1} \underline{l_{ij}^*}(u_{jk}^* - u_{jk}), \qquad i \leqq k \leqq n,$$

$$l_{ki}^* = \frac{1}{u_{ii}} \left\{ a_{ki} - \sum_{j=1}^{i-1} l_{kj}^* u_{ji} - \sum_{j=1}^{i} \underline{l_{kj}^*}(u_{ji}^* - u_{ji}) \right\}, \qquad i < k \leqq n.$$

If we set

$$l_{ij} := l_{ij}^{(0)} \in L_{ij}^{(0)}, \qquad u_{ij} := u_{ij}^{(0)} \in U_{ij}^{(0)}$$

and using

$$\underline{l_{ij}^*} \in L_{ij}^{(0)}, \qquad \underline{l_{kj}^*} \in L_{kj}^{(0)},$$

then because of (3) we get by the property of inclusion monotonicity of interval arithmetic (see, for example, [1], p. 7),

$$u_{ik}^* \in \left\{ a_{ik} - \sum_{j=1}^{i-1} L_{ij}^{(1)} u_{jk}^{(0)} - \sum_{j=1}^{i-1} L_{ij}^{(0)} (U_{jk}^{(1)} - u_{jk}^{(0)}) \right\} \cap U_{ik}^{(0)} =: U_{ik}^{(1)}, \qquad i \leqq k \leqq n,$$

$$l_{ki}^* \in \left\{ \frac{1}{u_{ii}^{(0)}} \left( a_{ki} - \sum_{j=1}^{i-1} L_{kj}^{(1)} u_{ji}^{(0)} - \sum_{j=1}^{i} L_{kj}^{(0)} (U_{ji}^{(0)} - u_{ji}^{(0)}) \right) \right\} \cap L_{ki}^{(0)} =: L_{ki}^{(1)}, \; i < k \leqq n.$$

The systematic repetition of these arguments gives us the following iteration method:

For $i = 1(1)n$:

$$\left.\begin{aligned} & U_{ik}^{(m+1)} = \left\{ a_{ik} - \sum_{j=1}^{i-1} L_{ij}^{(m+1)} u_{jk}^{(m)} - \sum_{j=1}^{i-1} L_{ij}^{(m)} (U_{jk}^{(m+1)} - u_{jk}^{(m)}) \right\} \cap U_{ik}^{(m)}, \\ & i \leqq k \leqq n, \\ & L_{ki}^{(m+1)} = \left\{ \frac{1}{u_{ii}^{(m)}} \left( a_{ki} - \sum_{j=1}^{i-1} L_{kj}^{(m+1)} u_{ji}^{(m)} - \sum_{j=1}^{i} L_{kj}^{(m)} (U_{ji}^{(m+1)} - u_{ji}^{(m)}) \right) \right\} \cap L_{ki}^{(m)}, \\ & i < k \leqq n. \end{aligned}\right\} \tag{4}$$

As for the first step one can show that the following result holds. Details of the proof are omitted.

Theorem 2. *Let the matrix $A$ have a factorization $A = (I + L^*)\,U^*$. If* (3) *holds and if we choose $l_{ik}^{(m)} \in L_{ki}^{(m)}$, $u_{ik}^{(m)} \in U_{ik}^{(m)}$, then for all $m \geqq 0$ we have*

$$u_{ik}^* \in U_{ik}^{(m)}, \qquad l_{ki}^* \in L_{ki}^{(m)}. \quad \square$$

Remarks.

1. It is interesting to note that the feasibility of (4) is guaranteed for starting intervals which have arbitrarily large but finite width. The only divisions occur during the computation of $L_{ki}^{(m+1)}$. Since $0 \notin u_{ii}^* \in U_{ii}^{(m)}$ it is always possible to choose $0 \notin u_{ii}^{(m)} \in U_{ii}^{(m)}$ which guarantees the feasibility of (4).

2. Note that it is not necessary to choose $l_{ki}^{(m)} \in L_{ki}^{(m)}$, $u_{ik}^{(m)} \in U_{ik}^{(m)}$ in order that Theorem 2 is valid. The same is the case for the next theorem. On the other hand this choice is quite natural since $l_{ki}^* \in L_{ki}^{(m)}$, $u_{ik}^* \in U_{ik}^{(m)}$ by assumption.

The following theorem may be verified using essentially the same techniques as in Theorem 1 and we therefore omit the details.

Theorem 3. *Under the assumption of the preceding theorem, method* (4) *gives the exact factorization after at most $2n - 1$ iteration steps.* (Here we again assume that all operations are performed without rounding errors.) $\square$

For the proof of the next theorem and in the sequel we need some additional concepts. For a real compact interval $A = [a_1, a_2]$ we denote

$$d(A) := a_2 - a_1$$

as *diameter* or *width* of $A$.

Similarly

$$|A| := \max\{|a_1|, |a_2|\}$$

is called *absolute value* of $A$.

We list some simple rules, the proofs of which may be found in [1], p. 20ff., for example.

(a) $A \subseteqq B \Rightarrow |A| \leqq |B|$

(b) $d(A) = |A - A|$

(c) $a \in \mathbb{R} \Rightarrow d(a) = 0$

(d) $d(A \pm B) = d(A) + d(B)$

(e) $d(AB) \leqq d(A)\,|B| + |A|\,d(B)$

(f) $a \in \mathbb{R} \Rightarrow d(aB) = d(Ba) = |a|\,d(B)$.

Theorem 4. *Let* $d^{(m)} := \max\limits_{1 \leqq i,j \leqq n} \{\max\{d(L_{ij}^{(m)}), d(U_{ij}^{(m)})\}\}$. *Then it holds for* (4) *that*

$$d^{(m+1)} \leqq \alpha (d^{(m)})^2$$

*with a nonnegative real number $\alpha$, which is independent of $m$: The widths of the intervals are approximately squared in each step.*

Proof. (Mathematical induction) Applying the rules (a)—(f) we get from (4):

$$\left.\begin{aligned}
&\text{For } i = 1(1)n\text{:}\\
&\qquad \text{For } k = i(1)n\text{:}\\
&\qquad d(U_{ik}^{(m+1)}) \leqq \sum_{j=1}^{i-1} d(L_{ij}^{(m+1)})\,|u_{jk}^{(m)}| + \sum_{j=1}^{i-1} d(L_{ij}^{(m)})\,|U_{jk}^{(m+1)} - u_{jk}^{(m)}|\\
&\qquad\qquad + \sum_{j=1}^{i-1} |L_{ij}^{(m)}|\, d(U_{jk}^{(m+1)}).\\
&\qquad \text{For } k = i+1(1)n\ (i < n)\text{:}\\
&\qquad d(L_{ki}^{(m+1)}) \leqq \frac{1}{|u_{ii}^{(m)}|}\left\{\sum_{j=1}^{i-1} d(L_{kj}^{(m+1)})\,|u_{ji}^{(m)}| + \sum_{j=1}^{i} d(L_{kj}^{(m)})\,|U_{ji}^{(m+1)} - u_{ji}^{(m)}|\right.\\
&\qquad\qquad \left. + \sum_{j=1}^{i} |L_{kj}^{(m)}|\, d(U_{ji}^{(m+1)})\right\}.
\end{aligned}\right\} \tag{4'}$$

Now define

$$\left.\begin{aligned}
&\text{For } i = 1(1)n\text{:}\\
&\alpha_{ik} = \begin{cases} \sum_{j=1}^{i-1} (\beta_{ij}\,|U_{jk}^{(0)}| + 1 + |L_{ij}^{(0)}|\,\alpha_{jk}), & k = i(1)n,\\ 0, \text{ otherwise,} \end{cases}\\
&\beta_{ki} = \begin{cases} \left|\frac{1}{u_{ii}^{(0)}}\right| \left\{\sum_{j=1}^{i-1} \beta_{kj}\,|U_{ji}^{(0)}| + \sum_{j=1}^{i} (1 + |L_{kj}^{(0)}|\,\alpha_{ji})\right\}, & k = i+1(1)n,\\ 0, \text{ otherwise,} \end{cases}
\end{aligned}\right\} \tag{4''}$$

and finally

$$\alpha = \max_{1 \leqq i,k \leqq n} \{\max\{\alpha_{ik}, \beta_{ki}\}\}.$$

Using definition (4'') we immediately get from (4') for $i = 1$ that

$$d(U_{1k}^{(m+1)}) \leqq \alpha_{1k}(d^{(m)})^2, \qquad 1 \leqq k \leqq n,$$

and

$$d(L_{k1}^{(m+1)}) \leqq \beta_{k1}(d^{(m)})^2, \qquad 1 < k \leqq n.$$

Assume now that for the first $i-1$ rows and columns

$$\left.\begin{aligned}
d(U_{lk}^{(m+1)}) &\leqq \alpha_{lk}(d^{(m)})^2, \qquad l \leqq k \leqq n\\
d(L_{kl}^{(m+1)}) &\leqq \beta_{kl}(d^{(m)})^2, \qquad l \leqq k \leqq n
\end{aligned}\right\} \tag{4'''}$$

hold ($1 \leqq l \leqq i-1$). This is certainly true for $l = 1$. Then we get from (4') and (4'')

$$d(U_{ik}^{(m+1)}) \leqq \sum_{j=1}^{i-1} \beta_{ij}\,|U_{jk}^{(0)}|\,(d^{(m)})^2 + \sum_{j=1}^{i-1} (d^{(m)})^2 + \sum_{j=1}^{i-1} |L_{ij}^{(0)}|\,\alpha_{jk}(d^{(m)})^2 = \alpha_{ik}(d^{(m)})^2,$$

$i \leqq k \leqq n.$

Similarly,

$$d(L_{ki}^{(m+1)}) \leqq \left|\frac{1}{U_{ii}^{(0)}}\right| \left\{ \sum_{j=1}^{i-1} \beta_{kj} |U_{jk}^{(0)}| (d^{(m)})^2 + \sum_{j=1}^{i} (d^{(m)})^2 + \sum_{j=1}^{i} |L_{kj}^{(0)}| \alpha_{ji} (d^{(m)})^2 \right\}$$
$$= \beta_{ki} (d^{(m)})^2, \qquad i < k \leqq n.$$

From these relations the assertion

$$d^{(m+1)} \leqq \alpha (d^{(m)})^2$$

follows. []

## 4. Discussion of rounding errors

Method (4) performed on a computer using finite precision arithmetic may behave quite different from the theoretical behaviour. In order to discuss this point in detail we first list some preliminaries.

We assume that the four arithmetic operations for intervals are performed in such a manner that the lower endpoint of the exact result is rounded downwards to the next machine number and rounding is performed in the analogous way for the right endpoint of the exact result (if rounding is necessary at all). For details see, for example, [2]. If * denotes one of the arithmetic operations $+$, $-$, $\times$, $/$ for real intervals and if

$$A*B = [(A*B)_1, (A*B)_2],$$

then we have for the actually computed interval

$$fl(A*B) = [(1 - \varepsilon_1)(A*B)_1, (1 + \varepsilon_2)(A*B)_2]. \tag{5}$$

Here we have $|\varepsilon_1|/2$, $|\varepsilon_2|/2 \leqq eps$ (= machine precision) and the signs of $\varepsilon_1$ and $\varepsilon_2$ have to be chosen in such a manner that

$$-\varepsilon_1 (A*B)_1 \leqq 0, \qquad \varepsilon_2 (A*B)_2 \geqq 0.$$

Therefore we can also write

$$fl(A*B) = A*B + [-\varepsilon_1 (A*B)_1, \varepsilon_2 (A*B)_2] \subseteqq A*B + \tilde{\varepsilon}[-|A*B|, |A*B|]$$
$$= A*B + |A*B|\,[-\tilde{\varepsilon}, \tilde{\varepsilon}], \tag{6}$$

where $\tilde{\varepsilon} = \max(|\varepsilon_1|, |\varepsilon_2|) \leqq 2eps$. We therefore have the following inequality for the width of the computed interval:

$$d\big(fl(A*B)\big) \leqq d(A*B) + 2\tilde{\varepsilon}\,|A*B|.$$

It shows that the absolute value of $A*B$ is essentially responsible for the difference between the exact width and the computed one.

Before coming back to the iteration method (4) we consider the following problem:

Let there be given machine intervals (that means real intervals the endpoints of which are machine numbers), say,

$$C_0, A_0, B_0, D_0, A_1, B_1, D_1, \ldots, A_{n-1}, B_{n-1}, D_{n-1}$$

and a machine number $a_n$.

The expression

$$R_n = \frac{1}{a_n}\{C_0 - A_0(B_0 - D_0) - A_1(B_1 - D_1) - \cdots - A_{n-1}(B_{n-1} - D_{n-1})\}$$

now has to be computed.

Theoretically we can use the following algorithm:

$$(S)\begin{cases} S_0 := C_0, \\ S_i := S_{i-1} - A_{i-1}(B_{i-1} - D_{i-1}), \qquad 1 \leqq i \leqq n, \\ R_n := S_n/a_n. \end{cases}$$

In practice, however, we are actually performing the following operations:

$$(\bar{S})\begin{cases} \bar{S}_0 := S_0 := C_0, \\ \bar{S}_i := fl\big(\bar{S}_{i-1} - fl(A_{i-1} \cdot fl(B_{i-1} - D_{i-1}))\big), \qquad 1 \leqq i \leqq n, \\ \bar{R}_n := f_l(\bar{S}_n/a_n). \end{cases}$$

For the following considerations we are defining $\varepsilon$ as $\varepsilon := 2\,\mathrm{eps}$.

Assume for the moment that $\bar{S}_0 = S_0 = C_0, \bar{S}_1, \ldots, \bar{S}_{n-1}$ have already been computed. Then we have from (6)

$$\begin{aligned} fl(B_{n-1} - D_{n-1}) &\subseteqq B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|\,[-\varepsilon, \varepsilon], \\ fl\big(A_{n-1} fl(B_{n-1} - D_{n-1})\big) &\subseteqq A_{n-1}(B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|\,[-\varepsilon, \varepsilon]) \\ &\quad + |A_{n-1}(B_{n-1} - D_{n-1} \\ &\quad + |B_{n-1} - D_{n-1}|\,[-\varepsilon, \varepsilon])\,|\,[-\varepsilon, \varepsilon] \\ &\subseteqq A_{n-1}(B_{n-1} - D_{n-1}) \\ &\quad + |A_{n-1}|\,|B_{n-1} - D_{n-1}|\,[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2] \end{aligned}$$

and therefore

$$\left.\begin{aligned} \bar{S}_n &\subseteqq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) \\ &\quad - |A_{n-1}|\,|B_{n-1} - D_{n-1}|\,[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2] \\ &\quad + \big|\bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) \\ &\quad - |A_{n-1}|\,|B_{n-1} - D_{n-1}|\,[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2]\big|\,[-\varepsilon, \varepsilon] \\ &\subseteqq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) + |\bar{S}_{n-1}|\,[-\varepsilon, \varepsilon] \\ &\quad + |A_{n-1}|\,|B_{n-1} - D_{n-1}|\,[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]. \end{aligned}\right\} \tag{7}$$

Using mathematical induction we now show that

$$\left.\begin{aligned}\overline{S}_n \subseteqq S_n + [-\varepsilon, \varepsilon]\sum_{i=0}^{n-1} |\overline{S}_i| \\ + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\sum_{i=0}^{n-1} |A_i|\,|B_i - D_i|\end{aligned}\right\} \tag{8}$$

holds. For $n = 1$ we have from (7) using $\overline{S}_0 = S_0 = C_0$

$$\begin{aligned}\overline{S}_1 &\subseteqq \overline{S}_0 - A_0(B_0 - D_0) + |\overline{S}_0|\,[-\varepsilon, \varepsilon] \\ &\quad + |A_0|\,|B_0 - D_0|\,[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3] \\ &= S_1 + [-\varepsilon, \varepsilon]\,|\overline{S}_0| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\,|A_0|\,|B_0 - D_0|\end{aligned}$$

and therefore the assertion holds for $n = 1$. If (8) holds for some $n \geqq 1$, then replacing $n$ by $n + 1$ in (7) and using $(S)$ we have

$$\begin{aligned}\overline{S}_{n+1} &\subseteqq \overline{S}_n - A_n(B_n - D_n) + [-\varepsilon, \varepsilon]\,|\overline{S}_n| \\ &\quad + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\,|A_n|\,|B_n - D_n| \\ &\subseteqq S_{n+1} + [-\varepsilon, \varepsilon]\sum_{i=0}^{n} |S_i| \\ &\quad + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\sum_{i=0}^{n} |A_i|\,|B_i - D_i|\end{aligned}$$

which is (8) with $n$ replaced by $n + 1$. Employing (6) once more we have the final result

$$\overline{R}_n \subseteqq \frac{\overline{S}_n}{a_n} + \frac{|\overline{S}_n|}{|a_n|}\,[-\varepsilon, \varepsilon]. \tag{9}$$

We are now using (8) and (9) in order to discuss the behaviour of (4) with respect to rounding errors.

Let $U_{ik}^{(m)}$ $(i \leqq k \leqq n)$ and $L_{ki}^{(m)}$ $(i < k \leqq n)$ be given as machine intervals for some $m \geqq 0$ and for $i = 1(1)n$.

Let $U_{ik}^{(m+1)}$ and $L_{ki}^{(m+1)}$ be defined by (4) and define

$$\left.\begin{aligned}\tilde{U}_{ik}^{(m+1)} &:= \left\{a_{ik} - \sum_{j=1}^{i-1} \overline{L}_{ij}^{(m+1)} u_{jk}^{(m)} - \sum_{j=1}^{i-1} L_{ij}^{(m)}(\overline{U}_{jk}^{(m+1)} - u_{jk}^{(m)})\right\} \cap U_{ik}^{(m)}, \\ \tilde{L}_{ki}^{(m+1)} &:= \left\{\frac{1}{u_{ii}^{(m)}}\left(a_{ki} - \sum_{j=1}^{i-1} \overline{L}_{kj}^{(m+1)} u_{ji}^{(m)} - \sum_{j=1}^{i} L_{kj}^{(m)}(\overline{U}_{ji}^{(m+1)} - u_{ji}^{(m)})\right)\right\} \cap L_{ik}^{(m)}\end{aligned}\right\} \tag{10}$$

and

$$\overline{U}_{ik}^{(m+1)} := fl(\tilde{U}_{ik}^{(m+1)}), \qquad \overline{L}_{ki}^{(m+1)} := fl(\tilde{L}_{ki}^{(m+1)}), \tag{11}$$

that is $\overline{U}_{ik}^{(m+1)}$ and $\overline{L}_{ki}^{(m+1)}$ denote the evaluation of (10) using floating point interval arithmetic.

Applying (8) and (9) to (11) we arrive at the following relations:

For $i = 1(1)n$:

For $k = i(1)n$:

$$\overline{U}_{ik}^{(m+1)} \subseteqq \tilde{U}_{ik}^{(m+1)} + [-\varepsilon, \varepsilon]\, r_{ik}^{(m+1)}$$

where

(a)
$$r_{ik}^{(m+1)} := \sum_{j=1}^{2i-2} |\bar{S}_j| + (3 + 3\varepsilon + \varepsilon^2) \sum_{j=1}^{i-1} (|\bar{L}_{ij}^{(m+1)}|\, |u_{jk}^{(m)}| + |L_{ij}^{(m)}|\, |\overline{U}_{jk}^{(m+1)} - u_{jk}^{(m)}|).$$

For $k = i + 1(1)n$:

$$\bar{L}_{ki}^{(m+1)} \subseteqq \tilde{L}_{ki}^{(m+1)} + [-\varepsilon, \varepsilon]\, r_{ki}^{(m+1}$$

where

(b)
$$r_{ki}^{(m+1)} := \frac{1}{|u_{ii}^{(m)}|} \left\{ \sum_{j=1}^{2i-1} |\overline{T}_j| + |\overline{T}_{2i}| + (3 + 3\varepsilon + \varepsilon^2) \times \left( \sum_{j=1}^{i-1} |\bar{L}_{kj}^{(m+1)}|\, |u_{ji}^{(m)}| + \sum_{j=1}^{i} |L_{kj}^{(m)}|\, |\overline{U}_{ji}^{(m+1)} - u_{ji}^{(m)}| \right) \right\}$$

(12)

($\bar{S}_j$ and $\overline{T}_j$ are the actually computed intermediate results if algorithm ($\bar{S}$) is used for the computation of $\overline{U}_{ik}^{(m+1)}$ and $\bar{L}_{ki}^{(m+1)}$ respectively. These intermediate results are also dependent on $i$ and $k$. For ease of notation we suppress this dependency).

Using (12) we now prove that the following relations hold:

For $i = 1(1)n$:

For $k = i(1)n$:

$$\overline{U}_{ik}^{(m+1)} \subseteqq U_{ik}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{ik}^{(m+1)}$$

where

(a)
$$\bar{r}_{ik}^{(m+1)} := \sum_{j=1}^{i-1} \bar{r}_{ij}^{(m+1)} |u_{jk}^{(m)}| + \sum_{j=2}^{i-1} |L_{ij}^{(m)}|\, \bar{r}_{ik}^{(m+1)} + r_{ik}^{(m+1)}.$$

For $k = i + 1(1)n$:

$$\bar{L}_{ki}^{(m+1)} \subseteqq L_{ki}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{ki}^{(m+1)}$$

where

(b)
$$\bar{r}_{ki}^{(m+1)} = \frac{1}{|u_{ii}^{(m)}|} \left\{ \sum_{j=1}^{i-1} \bar{r}_{kj}^{(m+1)} |u_{ji}^{(m)}| + \sum_{j=2}^{i} |L_{kj}^{(m)}|\, \bar{r}_{ji}^{(m+1)} \right\} + r_{ki}^{(m+1)}.$$

(13)

The proof follows by mathematical induction on $i$: For $i = 1$ we have from (4), (10) and (11) that

$$U_{1k}^{(m+1)} = \tilde{U}_{1k}^{(m+1)} = \overline{U}_{1k}^{(m+1)} = a_{1k}, \qquad 1 \leqq k \leqq n,$$

and therefore using (12)

$$\overline{U}_{1k}^{(m+1)} \subseteqq \tilde{U}_{1k}^{(m+1)} + [-\varepsilon, \varepsilon]\, r_{1k}^{(m+1)} = U_{1k}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{1k}^{(m+1)}$$

where $r_{1k}^{(m+1)} = \bar{r}_{1k}^{(m+1)} = 0$. Therefore the first part of (13) holds for $i = 1$. Similarly using the second part of (12) we have

$$\bar{L}_{k1}^{(m+1)} \subseteq \tilde{L}_{k1}^{(m+1)} + [-\varepsilon, \varepsilon]\, r_{k1}^{(m+1)}, \qquad 1 \leqq k \leqq n,$$

where

$$r_{k1}^{(m+1)} = \frac{1}{|u_{ii}^{(m)}|} \left\{ \sum_{j=1}^{2} |\overline{T}_j| + (3 + 3\varepsilon + \varepsilon^2)\, |L_{k1}^{(m)}|\, |\overline{U}_{1i}^{(m+1)} - u_{1i}^{(m)}| \right\}$$

and since

$$L_{k1}^{(m+1)} = \tilde{L}_{k1}^{(m+1)}$$

the second part of (13) also holds for $i = 1$ (with $\bar{r}_{k1}^{(m+1)} = r_{k1}^{(m+1)}$).

Suppose now that (13) holds for the first $i - 1$ rows and columns, respectively. Then we have from (12) for the $i$th row

$$\overline{U}_{ik}^{(m+1)} \subseteq \tilde{U}_{ik}^{(m+1)} + [-\varepsilon, \varepsilon]\, r_{ik}^{(m+1)}.$$

Using (10) and the induction hypothesis it follows that

$$\begin{aligned}
\overline{U}_{ik}^{(m+1)} &\subseteq a_{ik} - \sum_{j=1}^{i-1} \bar{L}_{ij}^{(m+1)} u_{jk}^{(m)} - \sum_{j=1}^{i-1} L_{ij}^{(m)} (\overline{U}_{jk}^{(m+1)} - u_{jk}^{(m)}) + [-\varepsilon, \varepsilon]\, r_{jk}^{(m+1)} \\
&\subseteq a_{ik} - \sum_{j=1}^{i-1} (L_{ij}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{ij}^{(m+1)})\, u_{jk}^{(m)} \\
&\quad - \sum_{j=1}^{i-1} L_{ij}^{(m)} (U_{jk}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{jk}^{(m+1)} - u_{jk}^{(m)}) + [-\varepsilon, \varepsilon]\, r_{ik}^{(m+1)} \\
&\subseteq U_{ik}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{ik}^{(m+1)}.
\end{aligned}$$

The relation

$$\bar{L}_{ki}^{(m+1)} \subseteq L_{ki}^{(m+1)} + [-\varepsilon, \varepsilon]\, \bar{r}_{ki}^{(m+1)}$$

is proven analogously.

The derived relations may be interpreted in the following manner: Suppose that all elements of $L^*$ have absolute value not greater than one. If the widths of the intervals $L_{ij}^{(m)}$ are small, then $|L_{ij}^{(m)}|$ and — because of forming intersections — also $|\bar{L}_{ij}^{(m+1)}|$ is not much greater than one. Under the similar assumption that the widths of the $U_{ik}^{(m)}$ are small it follows by the same reasoning that

$$|\overline{U}_{ik}^{(m+1)} - u_{ik}^{(m)}| \leqq |U_{ik}^{(m)} - U_{ik}^{(m)}| = d(U_{ik}^{(m)})$$

since $\overline{U}_{ik}^{(m+1)} \subseteq U_{ik}^{(m)}$, $u_{ik}^{(m)} \in U_{ik}^{(m)}$, and hence $|\overline{U}_{ik}^{(m+1)} - u_{ik}^{(m)}|$ is small.

Therefore for small widths of the enclosing intervals in the $m$th step, it follows from (12) that the difference $2\varepsilon r_{ik}^{(m+1)}$ between $d(\overline{U}_{ik}^{(m+1)})$ and $d(\tilde{U}_{ik}^{(m+1)})$ is essentially dependent on the behaviour of the size of the elements $|u_{ik}^{(m)}|$ and of the computed intermediate sums. The same is true for the difference $2\varepsilon r_{ki}^{(m+1)}$ between $d(\bar{L}_{ki}^{(m+1)})$ and $d(\tilde{L}_{ki}^{(m+1)})$ with the additional property that small values of $|u_{ii}^{(m)}|$ can make the

estimation of this difference even worse. From (13) we have

$$\left.\begin{aligned} d(\bar{U}_{ik}^{(m+1)}) &\leqq d(U_{ik}^{(m+1)}) + 2\varepsilon\bar{r}_{ik}^{(m+1)}, \\ d(\bar{L}_{ki}^{(m+1)}) &\leqq d(L_{ki}^{(m+1)}) + 2\varepsilon\bar{r}_{ki}^{(m+1)}. \end{aligned}\right\} \tag{14}$$

Assume again that $L^*$ has elements with absolute value not greater than one and that the widths of the $L_{ki}^{(m)}$ are small. Then we conclude from (13) (a) that the behaviour of $\bar{r}_{ik}^{(m+1)}$ is essentially dependent on the size of the elements $|u_{ik}^{(m)}|$. The same is true for $\bar{r}_{ki}^{(m+1)}$ with the additional property that again small values of $|u_{ii}^{(m)}|$ can make $\bar{r}_{ki}^{(m+1)}$ even larger. Since both the first terms of the right-hand sides in (14) are by Theorem 4 approximately the squares of the same terms in the preceding step one can get slowly growing bounds for $d(\bar{U}_{ik}^{(m+1)})$ and $d(\bar{L}_{ki}^{(m+1)})$ if the elements of $U^*$ above the main diagonal are not too large in absolute value, if the elements of $L^*$ are in absolute value not larger than one and if finally the absolute value of the intermediate results are not too large. Otherwise, one has to use higher precision in order that the bounds become more accurate.

## References

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

[2] Miranker, W. L., and U. Kulisch, Computer Arithmetic in Theory and Practice, Research Report RC 7776 (#33658) 7/24/79, Mathematics, IBM Thomas J. Watson Research Center, Yorktown Heights.

[3] Schmidt, J. W., Iterative Nachverbesserung von genäherten LU-Faktorisierungen, Manuskript, Technische Universität, Dresden 1978.

[4] Stoer, J., and R. Bulirsch, Introduction to Numerical Analysis, Springer-Verlag, New York—Heidelberg—Berlin 1980.

Manuskripteingang: 24. 11. 1981

VERFASSER:

Prof. Dr. Götz Alefeld, Institut für Angewandte Mathematik der Universität Karlsruhe
Prof. Dr. Jon G. Rokne, Department of Computer Science, The University of Calgary, Alberta, Canada