Rigorous error bounds for singular values of a matrix using the precise scalar product

G. Alefeld

Summary. Assume that there are given a real approximation σ to a simple singular value and real vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ as approximations to the corresponding right and left singular vectors of a real (m,n) matrix A. Then we consider the problem of computing rigorous errorbounds for these approximations. Furthermore we consider an iteration method which improves these bounds iteratively.

0. Introduction

It is well known (see [3],[5], for example) that if A is a real (m,n) matrix with $\operatorname{rank}(A) = r$ then there is an orthogonal (m,m) matrix V and an orthogonal (n,n) matrix U such that

$$V^T A U = \Sigma \tag{1}$$

where (in the case $m \geq n$)

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ & & & & \sigma_{r+1} \\ & & & & & \ddots \\ & & & & & & \sigma_n \\ & & & & & & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \Big\} m$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{n}$$

with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$. The σ_i, $i = 1(1)n$, are called the singular values of the matrix A. The columns u_i, $i = 1(1)n$, of the matrix U are called the right singular vectors of A. Correspondingly the v_i, $i = 1(1)m$, are called the left singular vectors of A. The representation $A = V \Sigma U^T$ of A, which follows from (1) is called the singular value decomposition of A. If σ_i is a singular value of A then σ_i^2 is an eigenvalue both of A^TA and of AA^T. Furthermore the vectors u_i and v_i are eigenvectors of A^TA and AA^T, respectively. A singular value σ_i is called simple if σ_i^2 is a simple eigenvalue of both AA^T and A^TA.

The singular value decomposition has a series of important applications in numerical analysis. See [5], for example.

Assume now that there are given a real approximation σ to a simple singular value and real vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ as approximations to the corresponding right and left singular vectors of a real (m,n) matrix A. Then we consider the problem of computing rigorous errorbounds for these approximations.

The problem of improving approximations which have been delivered by one of the well known algorithms for computing the singular value decomposition (see [5], for example) was already treated in [4].

In this paper we first present a nonlinear system of $m+n+2$ unknowns whose solution delivers an exact singular value and the exact components of the right and left singular vectors belonging to this singular value. This system was already considered in [4].

We then show that for sufficiently accurate approximations we can compute intervals in which the exact singular value and the components of the singular vectors are contained.

Starting with these inclusions we present an iterative method which improves these intervals repeatedly and which (theoretically, that is if no rounding errors are involved) is convergent to the exact values.

Finally we perform a detailed analysis of the proposed method if it is performed on a computer using a floating point system. For this discussion we assume that the so-called precise scalar product introduced by U. Kulisch and W. Miranker is available. See [6] and [7], for example. The precise scalar product exists as a part of the programming language PASCAL-SC. See [6], for example. This programming language is meanwhile implemented on a series of microcomputers like the APPLE IIe and the IBM personal computers. The precise scalar product can nowadays also be used worldwide on all IBM machines. See [6] for details.

In the last chapter we present some numerical examples. The presentation of the material assumes that the reader has a certain basic knowledge of interval analysis. See [1], for example.

1. Computing bounds and their improvement

Let there be given a real (m,n) Matrix A, let σ_1 and σ_2 be approximations to the same simple singular value of A and assume that $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ are approximations to the corresponding right and left singular vectors. (Usually $\sigma_1 = \sigma_2$ but we do not necessarily assume equality.) Then we consider the nonlinear system

$$A(u + y) = (\sigma_1 + \mu_1)(v + z) \tag{1}$$

$$A^T(v + z) = (\sigma_2 + \mu_2)(u + y) \tag{2}$$

$$(u + y)^T(u + y) = 1 \tag{3}$$

$$(v + z)^T(v + z) = 1 \tag{4}$$

consisting of $m+n+2$ equations for the unknown real scalars μ_1 and μ_2 and the unknown vectors $y \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$.

If this system has a solution μ_1, μ_2, y and z then it follows that

$$(v + z)^T A (u + y) = \sigma_1 + \mu_1$$

and

$$(u + y)^T A^T (v + z) = \sigma_2 + \mu_2$$

and therefore $\sigma_1 + \mu_2 = \sigma_2 + \mu_2$. From this it follows that if $\sigma := \sigma_1 + \mu_1 = \sigma_2 + \mu_2 \geq 0$ then σ is a singular vector of A and $u + y$ and $v + z$ are the corresponding right and left singular vectors of A .

In order to solve the equations (1)-(4) we rewrite them as

$$\begin{aligned} Ay - \sigma_1 z - \mu_1 v &= \sigma_1 v - Au + \mu_1 z \\ -\sigma_2 y + A^T z - \mu_2 u &= \sigma_2 u - A^T v + \mu_2 y \\ 2u^T y &= 1 - u^T u - y^T y \\ 2v^T z &= 1 - v^T v - z^T z. \end{aligned}$$

Defining the $(m+n+2, m+n+2)$ matrix B by

$$B = \begin{bmatrix} A & -\sigma_1 I_m & -v & 0 \\ -\sigma_2 I_n & A^T & 0 & -u \\ 2u^T & 0 & 0 & 0 \\ \underbrace{0}_{n} & \underbrace{2v^T}_{m} & \underbrace{0}_{1} & \underbrace{0}_{1} \end{bmatrix} \begin{matrix} \} m \\ \} n \\ \} 1 \\ \} 1 \end{matrix} \qquad (5)$$

and the vectors w and r by

$$w = \begin{bmatrix} y \\ z \\ \mu_1 \\ \mu_2 \end{bmatrix} \in \mathbb{R}^{m+n+2} \quad \text{and} \quad r = \begin{bmatrix} \sigma_1 v - Au \\ \sigma_2 u - A^T v \\ 1 - u^T u \\ 1 - v^T v \end{bmatrix} \in \mathbb{R}^{m+n+2} \qquad (6)$$

then these equations can be written as

$$Bw = r + f(w) \tag{7}$$

where $f: \mathbb{R}^{m+n+2} \to \mathbb{R}^{m+n+2}$ and

$$f(w) = \begin{bmatrix} \mu_1 \cdot z \\ \mu_2 \cdot y \\ -y^T y \\ -z^T z \end{bmatrix} . \tag{8}$$

We now show that the matrix B is nonsingular provided σ_1, σ_2 , u and v are sufficiently accurate approximations to a simple singular value and to the corresponding right and left singular vectors. By continuity arguments this follows from the following Theorem 1.

Theorem 1. Assume that σ is a simple singular value of the real (m,n) matrix A and let u and v be the corresponding right and left singular vectors of A belonging to σ . Then the matrix

$$\begin{bmatrix} A & -\sigma I_m & -v & 0 \\ -\sigma I_n & A^T & 0 & -u \\ 2u^T & 0 & 0 & 0 \\ 0 & 2v^T & 0 & 0 \end{bmatrix} \tag{9}$$

is nonsingular. (This matrix is obtained from B by replacing the approximations by the exact values.)

Proof. We have to show that the linear system

$$Ay - \sigma z - v\mu_1 = 0 \tag{10}$$

$$-\sigma y + A^T z - u\mu_2 = 0 \tag{11}$$

$$2u^T y = 0 \tag{12}$$

$$2v^T z = 0 \tag{13}$$

has only the trivial solution $\mu_1 = \mu_2 = 0$, $y = 0$, $z = 0$. By definition of σ, u and v we have the additional equations

$$Au = \sigma v \tag{14}$$

$$A^T v = \sigma u \tag{15}$$

$$u^T u = 1 \tag{16}$$

$$v^T v = 1 \ . \tag{17}$$

Multiplying (10) by v^T from the left, using (15) and (17) and finally taking into account (12) and (13) we obtain

$$0 = v^T Ay - \sigma v^T z - v^T v \mu_1 =$$

$$= \sigma u^T y - \sigma v^T z - \mu_1 =$$

$$= - \mu_1 .$$

Similarly it follows that $\mu_2 = 0$. Therefore (10) and (11) read

$$Ay - \sigma z = 0$$

$$A^T z - \sigma y = 0$$

from which it follows that

$$A^T Ay - \sigma A^T z = A^T Ay - \sigma^2 y = 0 \ . \tag{18}$$

If $y \neq 0$ then <u>either</u> $y = t \cdot u$, $t \neq 0$, which because of (12) contradicts (16), <u>or</u> $y \neq t \cdot u$. In this case it follows from (18) that the matrix $A^T A$ has two linearly independent eigenvectors u and y belonging to the simple eigenvalue σ^2 of $A^T A$. Therefore $y = 0$. Similarly it follows that $z = 0$. □

The equation (7) is the starting point for computing bounds for the unknown terms in the equations (1)-(4). Assume that σ_1, σ_2, y and z are sufficiently good approximations such that the matrix B defined by (5) is nonsingular. Assume that L is an approximation to the inverse of B or the exact inverse of B itself. Then the equation (7) can be rewritten as

$$w = Lr + (I-LB)w + L \cdot f(w). \tag{19}$$

We now determine an interval vector $[w] = ([w]_i)$ for which

$$Lr + (I-LB)w + L \cdot f(w) \in [w] \tag{20}$$

for all $w \in [w]$ holds. Using Brouwer's fixed point theorem it then follows that the equation (19) has at least one solution w^* in $[w]$. (The application of Brouwer's fixed point theorem can in this and similar cases be avoided. It is planned to discuss this in a more general setting).

In order to compute an interval vector $[w]$ for which (20) holds we try to find $[w]$ in the form

$$[w] = [-\beta,\beta]e \tag{21}$$

where $0 < \beta \in \mathbb{R}$ and $e = (1,\ldots,1)^T \in \mathbb{R}^{m+n+2}$.

Because of the inclusion monotony of interval arithmetic (see [1], Chapter 1, Theorem 5) it holds for $w \in [w]$ that

$$Lr + (I-LB)w+L\cdot f(w) \in Lr+(I-LB)[w]+L\cdot f([w]) =: [k] .$$

Therefore (20) holds if

$$[k] = Lr + (I-LB)[w] + L \cdot f([w]) \subseteq [w] . \tag{22}$$

This inclusion holds iff for the centers $m[w]$ and $m[k]$ and the diameters $d[w]$ and $d[k]$ of $[w]$ and $[k]$ the inequality

$$|m[w] - m[k]| + \frac{1}{2} d[k] \leq \frac{1}{2} d[w] \tag{23}$$

holds. The definition of m and d can be found in [1], Chapter 10, for example.

From (21) it follows that

$$m[w] = 0 \ . \tag{24}$$

Furthermore

$$m[k] = Lr \tag{25}$$

and

$$d[w] = 2\beta \cdot e \ . \tag{26}$$

Finally using (21) it follows that

$$d[k] = d\left[Lr + (I - LB)[w] + L \cdot f[w]\right] \tag{27}$$

$$= 2\beta |I - LB| e + 2\beta^2 |L| \cdot \begin{bmatrix} e^m \\ e^n \\ n \\ n \end{bmatrix}$$

where $e^i = (1,\ldots,1)^T \in \mathbb{R}^i$, $i = m,n$. Hence (23) holds iff

$$|Lr| + \beta |I - LB| e + \beta^2 |L| \cdot \begin{bmatrix} e^m \\ e^n \\ n \\ m \end{bmatrix} \leq \beta e \ . \tag{28}$$

Defining

$$\varrho = \| Lr \|_\infty \tag{29}$$

$$\kappa = \| I - LB \|_\infty \tag{30}$$

$$\ell = \| \, |L| \cdot \begin{bmatrix} e^m \\ e^n \\ n \\ m \end{bmatrix} \|_{\infty} \qquad (31)$$

then (28) clearly is valid if

$$\varrho + \kappa\beta + \beta^2\ell \leq \beta$$

or

$$\ell\beta^2 + (\kappa - 1)\beta + \varrho \leq 0 \qquad (32)$$

holds. Assume that $\kappa < 1$ and $(\kappa-1)^2 - 4\varrho\ell \geq 0$. Then the quadratic

$$\ell\beta^2 + (\kappa - 1)\beta + \varrho = 0$$

has the positive zeros

$$\beta_{1/2} = \frac{1 - \kappa \mp \sqrt{(1-\kappa)^2 - 4\varrho\ell}}{2\ell} \qquad (33)$$

Hence we have the proof for the following result.

Theorem 2. Let ϱ, κ and ℓ be defined by (29), (30) and (31). Assume that $\kappa < 1$ and that $(1-\kappa)^2 - 4\varrho\ell \geq 0$. If then $\beta \in [\beta_1, \beta_2]$ where β_1 and β_2 $(\beta_1 \leq \beta_2)$ are defined by (33) then the equation (19) has at least one solution w^* in the interval vector $[w] = ([w]_i)$ where $[w]_i = [-\beta, \beta]$. □

A solution w^* of the equation (19) is of course a solution of (7) if L is nonsingular. Under the assumption $\kappa < 1$ of the preceding Theorem 2 this is always the case since it follows from $\|I - LB\|_{\infty} = \kappa < 1$ that the inverse of $I-(I - LB) = LB$ exists.

We now consider the following iteration method:

$$\left.\begin{aligned} [w]^0 &= [-\beta,\beta]e \\ [w]^{k+1} &= h([w]^k), \quad k = 0,1,2,\ldots . \end{aligned}\right\} \tag{34}$$

where

$$h([w]) = Lr + (I - LB)[w] + L \cdot f([w]) . \tag{35}$$

This method computes a sequence $\{[w]^k\}_{k=0}^{\infty}$ of interval vectors.

The following result concerning (34) holds.

Theorem 3. Assume that $\kappa < 1$, $(\kappa - 1)^2 - 4\varrho\ell > 0$ and let β_1, β_2 be defined by (33). If then β satisfies $\beta_1 \leq \beta < (\beta_1 + \beta_2)/2$ then (34) delivers a sequence of interval vectors $\{[w]^k\}_{k=0}^{\infty}$ with the properties

$$w^* \in [w]^k , \quad k = 0,1,2,\ldots \tag{36}$$

and

$$\lim_{k\to\infty} [w]^k = w^* \tag{37}$$

where w^* is the unique solution of the equation (19) in $[w]^0$.

Proof. By Theorem 2 there exists at least one solution $w^* \in [w]^0$ of (19). If $w^* \in [w]^k$ which is the case for $k = 0$ then it follows by the inclusion monotony of interval arithmetic (see [1], Chapter 1, Theorem 5) that

$$w^* = Lr + (I - LB)w^* + L \cdot f(w^*)$$

$$\in Lr + (I - LB)[w]^k + L \cdot f([w]^k) = [w]^{k+1} .$$

Hence (36) holds.

In the proof of Theorem 2 we have shown that $[w]^1 = h([w]^0) \subseteq [w]^0$ holds. Assume now that

$$[w]^0 \supseteq [w]^1 \supseteq \ldots \supseteq [w]^{k-1} \supseteq [w]^k$$

holds for some $k \geq 1$ which is the case for $k = 1$. Then using inclusion monotonicity again it follows that

$$[w]^k = h([w]^{k-1}) \supseteq h([w]^k) = [w]^{k+1} .$$

Therefore $\{[w]^k\}_{k=0}^{\infty}$ forms a nested sequence of interval vectors. Hence there exists an interval vector $[w]^*$ for which $[w]^* = \lim_{k\to\infty} [w]^k$. We show that $d[w]^* = 0$ holds.
Because of $w^* \in [w]^k$ it then follows that (37) holds and that w^* is unique in $[w]^0$.
In order to prove $d[w]^* = 0$ we use the notation

$$d_k = \| d[w]^k \|_{\infty}$$

and take into account that for two intervals $[a]$ and $[b]$ the inequality

$$d([a] \cdot [b]) \leq |[a]| \cdot d[b] + d[a] \cdot |[b]|$$

holds where $|\cdot|$ denotes the absolute value of an interval. See [1], Chapter 2. Then by the definition of $f(w)$ we get that

$$d(f([w]^k)) = \begin{bmatrix} d([\mu_1]^k \cdot [z]^k) \\ d([\mu_2]^k \cdot [y]^k) \\ d(-([y]^k)^T \cdot [y]^k) \\ d(-([z]^k)^T \cdot [z]^k) \end{bmatrix}$$

$$\leq \begin{bmatrix} |[\mu_1]^k| d[z]^k + d[\mu_1]^k |[z]^k| \\ |[\mu_2]^k| d[z]^k + d[\mu_2]^k |[y]^k| \\ 2 \sum_{i=1}^{n} |[y]_i^k| d[y]_i^k \\ 2 \sum_{i=1}^{m} |[z]_i^k| d[z]_i^k \end{bmatrix} .$$

Since $[w]^k \subseteq [w]^o$ it follows that

$$d(f([w]^k)) \leq \begin{bmatrix} d_k \cdot \beta \cdot e^m + \beta \cdot d_k \cdot e^m \\ d_k \cdot \beta \cdot e^n + \beta \cdot d_k \cdot e^n \\ 2n\beta d_k \\ 2m\beta d_k \end{bmatrix} = 2 \cdot \beta \cdot d_k \cdot \begin{bmatrix} e^m \\ e^n \\ n \\ m \end{bmatrix} .$$

Using this inequality we get from (34), (35) that

$$d[w]^{k+1} = |I - LB| \cdot d[w]^k + |L| \cdot d(f([w]^k))$$

$$\leq d_k \, |I - LB| \, e + 2 \cdot \beta \cdot d_k \cdot |L| \cdot \begin{bmatrix} e^m \\ e^n \\ n \\ m \end{bmatrix} .$$

Using the definition of κ and ℓ it follows from this inequality that

$$d_{k+1} \leq (\kappa + 2\beta\ell) \, d_k . \qquad (38)$$

Since $\beta < \frac{\beta_1 + \beta_2}{2} = \frac{1 - \kappa}{2\ell}$ it follows that $\kappa + 2\beta\ell < 1$.

Hence from (38) we have $\lim_{k \to \infty} d([w]^k) = d([w]^*) = 0$ and the Theorem is proved. □

2. Rounding error analysis

In order to make the discussion following more easy to understand we start with some general remarks:

1. From the basic properties of interval arithmetic it follows that for a real nonsingular Matrix B and an interval vector $[z]$ it holds that

$$\{B^{-1}z \mid z \in [z]\} \subseteq B^{-1} \cdot [z] \tag{1}$$

where on the right-hand side the product $B^{-1} \cdot [z]$ is formed by following the rules of interval arithmetic. See [1], Chapter 10.

If on the other hand $IGA(B,[z])$ denotes an interval vector which results if Gaussian elimination is applied to a system with B as coefficient matrix and with $[z]$ as right-hand side then it also holds that

$$\{B^{-1}z \mid z \in [z]\} \subseteq IGA(B,[z]). \tag{2}$$

(Note that $IGA(B,[z])$ is not unique even if no rounding errors occur. It is strongly dependent on the fashion how pivoting is performed.)

In [2], Lemma 1 it was shown that always

$$B^{-1} \cdot [z] \subseteq IGA(B,[z]) \tag{3}$$

where in general the proper inclusion sign holds. The computation of $B^{-1} \cdot [z]$ requires approximately three times the amount of work which is necessary to obtain $IGA(B,[z])$. On the other hand (3) shows that the set of real vectors $\{B^{-1}z \mid z \in [z]\}$ is in the set theoretic sense usually better included by $B^{-1}[z]$ then by $IGA(B,[z])$.

Consider now the case that we are choosing $L := B^{-1}$ in (1.35). Since B^{-1} is a point matrix the distributive law

holds (see [1], Chapter 10) and therefore

$$h([w]) = B^{-1} r + B^{-1} f([w])$$

$$= B^{-1} \{r + f([w])\} .$$

Because of (3) we then have

$$h([w]) = B^{-1} \cdot [z] \subseteq IGA(B,[z])$$

where $[z] := r + f([w])$. In other words: The inclusion (1.22) which guarantees the existence of a solution surely holds for $h([w]) = B^{-1}\{r + f([w])\}$ if it holds for $IGA(B, r + f([w]))$ but not necessarily vice versa.

Furthermore using (3) it follows by complete induction that for an interval vector $[w]^0 = [-\beta,\beta]e$ determined by Theorem 1.3 the iteration method

$$\left.\begin{aligned} [w]^0 &= [-\beta,\beta]e \\ [w]^{k+1} &= IGA(B, r + f([w]^k)),\ k = 0,1,2,\ldots \end{aligned}\right\} \qquad (4)$$

delivers a sequence of interval vectors which in each step gives a cruder inclusion of w^* than (1.34), (1.35) with $L := B^{-1}$. Since we are interested in computing close bounds in as few steps as possible we use (1.34), (1.35) with $L := B^{-1}$ (instead of (4) which with respect to the computational amount would be favourable.)

We are now going to discuss how the method (1.34), (1.35) behaves if it is performed using a floating point system.

Let $b > 1$ be the basis of the number system and let t_1 be the mantissa length of a single length floating point number. For the discussion following we make the <u>assumptions</u> (a)-(c):

(a) For two machine intervals $[a]$ and $[b]$ it holds that

$$fl([a]*[b]) = [(1-\varepsilon_1)([a]*[b])_1, (1+\varepsilon_2)([a]*[b])_2] \qquad (5)$$

where

$$* \in \{+,-,\times,/\},$$

$$[a]*[b] = [([a]*[b])_1,\ ([a]*[b])_2],$$

$$|\varepsilon_1|,\ |\varepsilon_2| \leq \varepsilon = b^{1-t_1}.$$

7
$f\ell(\cdot)$ denotes the result of a machine interval operation taking into account rounding of the bounds outwards to the next machine number. (5) states that the lower bound $([a]*[b])_1$ of the exact result is rounded downwards to the next machine number (if rounding is necessary at all). The analogue holds for the upper bound. Note that (5) (and also (6) below) do not hold in general in the underflow range.

(b) We assume that the so-called precise scalar product proposed by U. Kulisch and W.L. Miranker (see [6], [7]) is available on the computer:

For two interval vectors $[x] = ([x]_i)$ and $[y] = ([y]_i)$ which have machine intervals as components it holds that

$$f\ell\left(\sum_{i=1}^{n} [x]_i \cdot [y]_i\right) = [(1 - \varepsilon_1)\sigma_1,\ (1 + \varepsilon_2)\sigma_2] \qquad (6)$$

where

$$\sum_{i=1}^{n} [x]_i \cdot [y]_i = [\sigma_1, \sigma_2]$$

is the exact scalar product and again

$$|\varepsilon_1|\ ,\ |\varepsilon_2| \leq b^{1-t_1}.$$

Normally the precision of the precise scalar product is comparable to double length accumulation and rounding to single length after completion. If, however severe cancellation of terms occurs (which is usually the case if one has to compute residuals) then it is much more accurate.

From (6) it follows that

$$fl\left(\sum_{i=1}^{n}[x]_i[y]_i\right) \subseteq \sum_{i=1}^{n}[x]_i\cdot[y]_i + \varepsilon[-1,1]\left|\sum_{i=1}^{n}[x]_i\cdot[y]_i\right|$$

where again $\varepsilon = b^{1-t_1}$.

(c) We assume that the given matrix A is exactly representable on the machine. The same we assume for σ_1, σ_2, u and v. If we have a binary machine then the matrix B (see (1.5)) is also exactly representable. If this is not the case then $2u^T$ and $2v^T$ can in general not be computed exactly and rounding errors have to be taken into account. This means that on the machine we have a matrix $\overline{B}$ which differs from B in the elements where the vectors $2u^T$ and $2v^T$ are located.

We assume that there is known an interval matrix $[BI]$ which is exactly representable on the machine such that

$$B^{-1} \in \overline{[BI]} \subseteq B^{-1} + \varepsilon\cdot[-1,1]\cdot|B^{-1}| + [-1,1]\cdot\overline{E} \qquad (8)$$

where $\varepsilon = b^{1-t_1}$. If B^{-1} is well conditioned with respect to small changes of the vectors $2u^T$ and $2v^T$ in the matrix B then $\overline{E}$ is a real nonnegative matrix whose elements are exactly representable on the machine. If it is even true that $B = \overline{B}$ then $\overline{E} = 0$. The relation (8) then states that the inverse of B can be included on the machine by an interval matrix $\overline{[BI]}$ whose elements have bounds which differ by at most the distance of two neighbouring machine numbers. Using the precise scalar product (6) such a close inclusion of the inverse of B can be computed even for very badly conditioned matrices B. See [8]. If B is not exactly representable ($B = \overline{B}$) then using the precise scalar product one can compute an interval matrix $\overline{[BI]}$ for which (8) holds with

$\|\overline{E}\| \approx \varepsilon\cdot\|B^{-1}\|\cdot\|R_B\|$, $\varepsilon = b^{1-t_1}$. R_B is the matrix which one obtains from B by setting all elements equal to zero

with the exception of $2u^T$ and $2v^T$.

With $\overline{\cdot}$ we denote terms which are computed by (1.34) and (1.35) on the machine. For example $\overline{[w]}^k$ is the interval vector actually computed on the machine. Furthermore we define

$$[x]^k := r + f(\overline{[w]}^k).$$

The components of $\overline{[x]}^k$ are computed using the precise scalar product. For example, in order to compute the first m components of $\overline{[x]}^k$ we form the vectors

$$(\sigma_1, a_{i1}, \ldots, a_{in}, \overline{[u_1]}^k)^T$$

and

$$(v_1, -u_1, \ldots, -u_n, [z]_i^k)^T$$

each with $n + 2$ components and compute the precise scalar product. Analogously we proceed for the remaining $n + 2$ components of $\overline{[x]}^k$. Because of (7) we then have

$$\overline{[x]}^k \subseteq [x]^k + \varepsilon \cdot [-1,1] \cdot |[x]^k| \,. \tag{9}$$

The interval vector $\overline{[w]}^{k+1}$ is computed by

$$\overline{[w]}^{k+1} = f\ell(\overline{[BI]} \cdot \overline{[x]}^k)$$

where the right hand side is again computed by using the precise scalar product. Using (7) and (9) we then have for the actually computed iterates

$$\overline{[w]}^{k+1} = f\ell\,(\overline{[BI]} \cdot \overline{[x]}^k)$$

$$\subseteq f\ell((B^{-1} + \varepsilon \cdot [-1,1] \cdot |B^{-1}| + [-1,1] \cdot \overline{E}) \cdot ([x]^k + \varepsilon \cdot [-1,1] \cdot |[x]^k|))$$

$$\subseteq (B^{-1}+\varepsilon\cdot[-1,1]\cdot|B^{-1}|+[-1,1]\cdot\overline{E})\cdot([x]^k+\varepsilon\cdot[-1,1]\cdot|[x]^k|)+$$

$$+\ \varepsilon\cdot[-1,1]\cdot|B^{-1}+\varepsilon\cdot[-1,1]\cdot|B^{-1}|+[-1,1]\cdot\overline{E}|\ \times$$

$$\times\ |[x]^k+\varepsilon\cdot[-1,1]\cdot|[x]^k||$$

$$\subseteq B^{-1}\cdot[x]^k+[-1,1]\cdot(3\varepsilon+3\varepsilon^2+\varepsilon^3)\cdot|B^{-1}|\cdot|[x]^k|+$$

$$+\ [-1,1]\cdot\varepsilon\cdot(1+\varepsilon)\cdot\overline{E}\cdot|[x]^k|$$

$$+\ [-1,1]\cdot(1+\varepsilon)\cdot\overline{E}\cdot|[x]^k|.$$

From this we obtain

$$d\overline{[w]}^{k+1} \leq |B^{-1}|\cdot d[x]^k+2(3\varepsilon+3\varepsilon^2+\varepsilon^3)\cdot|B^{-1}|\cdot|[x]^k|+$$

$$+\ 2(1+\varepsilon)^2\cdot\overline{E}\cdot|[x]^k|.$$

We now assume (without serious restriction) that the interval vector $[w]^0 = [-\beta_1,\beta_1]e$ computed by using Theorem 1.3 is exactly representable on the machine. (If this is not the case then it is sufficient to round upwards all operations when computing β_1 by formula (1.33).) It therefore holds that $\overline{[w]}^0 = [w]^0$. By forming intersections after each iteration step in (1.34) it is guaranteed that all iterates are contained in $\overline{[w]}^0 = [w]^0$. Defining

$$\overline{d}_k = \|d\overline{[w]}^k\|_\infty$$

$$\delta = 2\beta_1\ell = 1-\sqrt{1-4\varrho\ell} < 1$$

$$\varepsilon = b^{1-t_1}$$

$$\tilde{\varepsilon} = \|\overline{E}\|_\infty$$

$$s = 2(2+3\varepsilon+\varepsilon^2)\cdot\|B^{-1}\|_\infty\cdot\|\,|[x]^0|\,\|_\infty$$

$$\tilde{s} = 2(1+\varepsilon)^2\cdot\|\,|[x]^0|\,\|_\infty$$

we get

$$\overline{d}_{k+1} \leq \delta\, \overline{d}_k + \varepsilon s + \tilde{\tilde{\varepsilon}}\tilde{\tilde{s}} \qquad (10)$$

from the last inequality. For the proof of (10) it is sufficient to note that similarly as in the proof of Theorem 1.3 it follows that

$$|B^{-1}|d[x]^k \leq \overline{d}_k \cdot 2 \cdot \beta_1 \cdot |B^{-1}| \cdot e \ .$$

Defining $\tau := \varepsilon s + \tilde{\tilde{\varepsilon}}\tilde{\tilde{s}}$ we get from (10) that

$$\overline{d}_{k+1} \leq \delta^{k+1}\overline{d}_o + \left\{ \delta^k + \delta^{k-1} + \ldots + \delta + 1 \right\} \tau$$

and therefore

$$d_{k+1} \leq \delta^{k+1}\, \overline{d}_o + \frac{\tau}{1-\delta}$$

$$= \delta^{k+1}\, \overline{d}_o + \frac{\varepsilon s + \tilde{\tilde{\varepsilon}}\tilde{\tilde{s}}}{1 - \delta} \ .$$

This inequality can be interpreted as follows:
Since by assumption $\delta < 1$ the first term tends to zero for $k \to \infty$. The reachable precision on the computer is therefore determined by the second term. For sufficiently accurate approximations σ_1, σ_2, u and v and if $\|B^{-1}\|_\infty$ is not large we have $\delta < 1$. See the definition of δ in dependence of ϱ and ℓ. Therefore under these circumstances the denominator of the second term is not much smaller than one. Under the same assumptions $|[x]^o|$ has small components. If finally the inverse of B is well conditioned with respect to small changes in B of order $\varepsilon = b^{1-t_1}$ in those elements where the vectors $2u^T$ and $2v^T$ are located then also $\tilde{\varepsilon}$ will be small. Under these assumptions it is therefore guaranteed that we get small diameters for the components of the interval vectors computed on the machine.

3. Numerical example

We consider the (5.3) matrix

$$A = \begin{bmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{bmatrix}$$

for which the matrix Σ from (0.1) reads

$$\begin{bmatrix} 35.127 \ldots & 0 & 0 \\ 0 & 2.465 \ldots & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .$$

We want to include the largest singular value and the corresponding right and left singular vector. We assume that we have given approximations which have relative precision of 10^{-9}:

$$\sigma_1 = \sigma_2 \approx 0.351\ 272\ 233 \times 10^2 \ ;$$

$$u \approx \begin{bmatrix} 0.201\ 664\ 911 \\ 0.516\ 830\ 501 \\ 0.831\ 996\ 092 \end{bmatrix} ;$$

$$v \approx \begin{bmatrix} 0.354\ 557\ 057 \\ 0.398\ 696\ 370 \\ 0.442\ 835\ 683 \\ 0.486\ 974\ 996 \\ 0.531\ 114\ 309 \end{bmatrix} .$$

Using these values we get for β_1 defined by (1.33):

$$\beta_1 = 0.335\ 748\ 344\ 331 \times 10^{-9} .$$

After one step of (1.34), (1.35) with $L = B^{-1}$ we get the following inclusions:

$$\sigma_1 = \sigma_2 \in [\ 0.351\ 272\ 233\ 33\tfrac{5}{6}]$$

$$u \in \begin{bmatrix} [0.201\ 664\ 911\ 19\tfrac{2}{3}] \\ [0.516\ 830\ 501\ 39\tfrac{2}{3}] \\ [0.831\ 996\ 091\ 59\tfrac{1}{2}] \end{bmatrix}$$

$$v \in \begin{bmatrix} [0.354\ 557\ 057\ 03\tfrac{7}{8}] \\ [0.398\ 696\ 369\ 99\tfrac{8}{9}] \\ [0.442\ 835\ 682\ 9\tfrac{59}{60}] \\ [0.486\ 974\ 995\ 92\tfrac{1}{2}] \\ [0.531\ 114\ 308\ 88\tfrac{2}{3}] \end{bmatrix}$$

($[0\ \ldots\ xx\tfrac{7}{8}]$ denotes an interval with lower bound $0\ \ldots\ xx7$ and upper bound $0\ \ldots\ xx8$) .
In all computed examples the bounds were similarly close. The example has been computed using an IBM-PC . In PASCAL-SC a floating point number has 12 decimal digits in the mantissa.

References

[1] Alefeld, G.; Herzberger, J.: Introduction to Interval Computation. Academic Press, New York, 1983

[2] Alefeld, G.; Platzöder, L.: A quadratically convergent Krawczyk-like algorithm. SIAM J. Numer. Anal. 20, 1983

[3] Bunse, W.; Bunse-Gerstner,A.: Numerische lineare Algebra. Teubner Stuttgart 1985. Studienbücher Mathematik

[4] Dongarra, J.J.: Improving the accuracy of computed singular values. SIAM J. Sci. Stat. Comput., 4 (1983) 712-719

[5] Golub, G.H.; van Loan C.F.: Matrix Computations. The Johns Hopkins University Press. Baltimore Maryland, 1983

[6] Kulisch, U.; Miranker, W.L.: A New Approach to Scientific Computation. Notes and Reports in Computer Science and Applied Mathematics. Academic Press, 1983

[7] Kulisch, U.; Miranker, W.L.: Computer Arithmetic in Theory and Practise, Academic Press, New York, 1981

[8] Rump, S.M.: Solving Algebraic Problems with High Accuracy. In [6], 53-120