Numer. Math. 48, 391–416 (1986)

Numerische
Mathematik
© Springer-Verlag 1986

Componentwise Inclusion and Exclusion Sets for Solutions of Quadratic Equations in Finite Dimensional Spaces

G. Alefeld*

Institut für Angewandte Mathematik, Universität Karlsruhe, Kaiserstr. 12, D-7500 Karlsruhe, Federal Republic of Germany

Summary. In this paper we use interval arithmetic tools for the computation of componentwise inclusion and exclusion sets for solutions of quadratic equations in finite dimensional spaces. We define a mapping for which under certain assumptions we can construct an interval vector which is mapped into itself. Using Brouwer's fixed point theorem we conclude the existence of a solution of the original equation in this interval vector. Under different assumptions we can construct an interval vector such that the range of the mapping has no point in common with this interval vector. This implies that there is no solution in this interval vector. Furthermore we consider an iteration method which improves componentwise error-bounds for a solution of a quadratic. The theoretical results of this paper are demonstrated by some numerical examples using the algebraic eigenvalue problem which is probably the best known example of a quadratic equation.

Subject Classifications: AMS(MOS): 65G10, 65H10; CR: G.1.5.

0. Introduction

In this paper we consider the problem of constructing sets in $\mathbb{R}^m$ in which there exists either at least one solution or no solution of a given quadratic equation in $\mathbb{R}^m$. Such sets are called *inclusion and exclusion sets*, respectively.

After having formulated in detail the problem and having collected some more or less well known formulas in Chap. 1 we list in Chap. 2 some tools from interval analysis. In Chap. 3 we start by proving the fundamental inclusion Theorem 1. Under special assumptions we get a result (Corollary 2) which was proven partly in a series of papers by Yamamoto [24–27]. See also Hoffmann [6].

[*] This paper contains the main results of a talk given by the author on the occasion of the 25th anniversary of the founding of Numerische Mathematik, March 19–21, 1984 at the Technische Universität of Munich, Germany

In Chap. 4 we construct exclusion sets, that is sets which contain no solution of the quadratic equation. To the authors knowledge there do not exist similar results in the literature.

In Chap. 5 we introduce an iteration method by which a solution of a quadratic equation can be computed which is contained in an interval vector. If the interval vector does not contain a solution then this method will break down after a finite number of steps.

In the final Chap. 6 we illustrate some of the results of this paper by using the algebraic eigenvalue problem, the most important case of a quadratic equation in $\mathbb{R}^m$.

In passing we note that some of the results of this paper also hold for more general nonlinear equations. On the other hand quadratic equations possess a series of special properties which can be used with great advantage.

We finally note that there exists a general theory on quadratic equations which was developped by Rall [16]. See also Prenter [14].

1. Quadratic Equations in $\mathbb{R}^m$ and some Preliminaries

Assume that $c=(c_i)\in\mathbb{R}^m$ is a real m-vector, that $A=(a_{ij})$ is a real (m,m) matrix and that $B=(b_{ijk})$ is a bilinear operator from $\mathbb{R}^m\times\mathbb{R}^m$ to $\mathbb{R}^m$. Then the mapping $f\colon \mathbb{R}^m\to\mathbb{R}^m$ defined by

$$f(z)=c+Az+Bz^2, \qquad z\in\mathbb{R}^m \quad (\text{where } Bz^2=Bzz) \tag{1}$$

is called a *quadratic operator*. The equation $f(z)=0$, that is

$$Bz^2+Az+c=0 \tag{2}$$

is called a *quadratic equation* in $\mathbb{R}^m$.

Let $z^0\in\mathbb{R}^m$ and assume that $f'(z^0)$ and $f''(z^0)$ denote the first and second derivatives, respectively. Then for $z\in\mathbb{R}^m$ it follows that

$$f(z)=f(z^0)+f'(z^0)(z-z^0)+\tfrac{1}{2}f''(z^0)(z-z^0)^2 \tag{3}$$

holds.

If $f(z^*)=0$ and if L is a real (m,m) matrix then it follows from (3) that

$$z^*=z^0-Lf(z^0)+\{I_m-Lf'(z^0)-\tfrac{1}{2}(Lf''(z^0))(z^*-z^0)\}(z^*-z^0). \tag{4}$$

Using the real (m,m) matrix L we define for a given quadratic operator the mapping $g\colon \mathbb{R}^m\to\mathbb{R}^m$ by

$$g(z)=z^0-Lf(z^0)+\{I_m-Lf'(z^0)-\tfrac{1}{2}(Lf''(z^0))(z-z^0)\}(z-z^0). \tag{5}$$

We now introduce two Lemmata which are used subsequently.

Lemma 1. *Let $K=(k_{ij})$ be a nonnegative (m,m) matrix and let $u=(u_i)\in\mathbb{R}^m$ be a nonnegative real vector. Then it holds that*

$$Ku \leqq \|u\|_q \kappa_p{}^{1} \tag{6}$$

and for the spectral radius $\rho(K)$ *it holds for* $p>1$, $q>1$, $\frac{1}{p}+\frac{1}{q}=1$, *that*

$$\rho(K) \leqq \|\kappa_p\|_q := \left\{ \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{m} k_{ij}^p \right)^{\frac{1}{p}} \right]^q \right\}^{\frac{1}{q}} \tag{7}$$

where

$$\|u\|_q = \left(\sum_{i=1}^{m} |u_i|^q \right)^{\frac{1}{q}}, \tag{8}$$

and

$$\kappa_p = \left(\left(\sum_{j=1}^{m} k_{ij}^p \right)^{\frac{1}{p}} \right) \in \mathbb{R}^m. \tag{9}$$

If furthermore $H=(h_{ijk})$ *with* $h_{ijk} \geqq 0$ *and if* $u=(u_i)$ *is non-negative then*

$$Hu^2 \leqq \|u\|_q^2 h_p \tag{10}$$

where

$$h_p = \left(\left\{ \sum_{j=1}^{m} \sum_{k=1}^{m} h_{ijk}^p \right\}^{\frac{1}{p}} \right) \in \mathbb{R}^m. \tag{11}$$

If we set $q=\infty$ *for* $p=1$ *and* $p=\infty$ *for* $q=1$ *then* (6) *holds with* $\|u\|_\infty = \max_{1 \leqq i \leqq m} |u_i|$,

$$\kappa_1 = \left(\sum_{j=1}^{m} k_{ij} \right) \in \mathbb{R}^m \tag{9'}$$

and $\|u\|_1 = \sum_{i=1}^{m} |u_i|$,

$$\kappa_\infty = (\max_{1 \leqq j \leqq m} k_{ij}) \in \mathbb{R}^m, \tag{9''}$$

respectively.

Analogously (10) *holds with* $\|u\|_\infty = \max_{1 \leqq i \leqq m} |u_i|$,

$$h_1 = \left(\sum_{j=1}^{m} \sum_{k=1}^{m} h_{ijk} \right) \in \mathbb{R}^m \tag{11'}$$

and $\|u\|_1 = \sum_{i=1}^{m} |u_i|$,

$$h_\infty = (\max_{1 \leqq j,k \leqq m} h_{ijk}) \in \mathbb{R}^m, \tag{11''}$$

respectively. □

[1] Inequalities between vectors are always understood elementwise, that is, if $a=(a_i)$, $b=(b_i)$ are from $\mathbb{R}^m$ then $a \leqq b$ iff $a_i \leqq b_i$, $1 \leqq i \leqq m$

Lemma 2. *Let $H=(h_{ijk})$ be a bilinear operator and let $u=(u_i)$ be a nonnegative vector. Then if $p>1$, $q>1$, $\frac{1}{p}+\frac{1}{q}=1$ it holds that*

$$Hu \leqq \|u\|_q H_p \tag{12}$$

where H_p is the (m,m) matrix

$$H_p=\left(\left(\sum_{k=1}^{m} h_{ijk}^{p}\right)^{\frac{1}{p}}\right). \tag{13}$$

If we set $q=\infty$ for $p=1$ and $p=\infty$ for $q=1$ then (12) *holds with $H_1=\left(\sum_{k=1}^{m} h_{ijk}\right)$ and $H_\infty=(\max_{1\leqq k\leqq m} h_{ijk})$, respectively.* □

2. Tools from Interval Analysis

We assume that the reader has a certain knowledge of the basic facts of interval analysis. See for example [1]. Subsequently we list some details which are used repeatedly.

For two real compact intervals $[a]=[a_1,a_2]$ and $[b]=[b_1,b_2]$ the basic arithmetic operations $*\in\{+,-,\cdot,/\}$ are defined by

$$[a]*[b]=\{a*b \mid a\in[a],\ b\in[b]\}. \tag{1}$$

The so-called subdistributive law holds:

$$[a]([b]+[c])\subseteq[a][b]+[a][c]. \tag{2}$$

Vectors, whose components are intervals, so-called interval vectors, are denoted by $[z],[u],[v],\ldots$. For the components of $[z]$ we write $[z]_i$. Hence we have $[z]=([z]_i)$.

A mapping f from the set of interval vectors into the same set is called *inclusion monotonic* if

$$[u]\subseteq[v] \Rightarrow f([u])\subseteq f([v]) \tag{3}$$

where the inclusion $[u]\subseteq[v]$ is defined via the components. If $B=(b_{ijk})$ is a bilinear operator then the product $B\cdot[u]$ is an interval matrix defined by

$$B\cdot[u]=\left(\sum_{k=1}^{m} b_{ijk}[u]_k\right). \tag{4}$$

The multiplication of an interval matrix and an interval vector is defined by the usual rule for the product of a matrix and a vector. If $A=(a_{ij})$ is a real (m,m) matrix and $B=(b_{ijk})$ is a bilinear operator then the product AB is a bilinear operator defined by

$$AB=\left(\sum_{s=1}^{m} a_{is}b_{sjk}\right). \tag{5}$$

If B is symmetric then the same holds for AB. If A is a real (m,m) matrix and B is a bilinear operator, then for all interval vectors $[x],[y]$ it holds that

$$(AB)[x][y] \subseteq (A(B[x]))[y] \subseteq A(B[x][y]). \tag{6}$$

If $[x]$ is symmetric (that is $[x]=-[x]$) then it holds that

$$(A(B[x]))[y] = A(B[x][y]), \tag{7}$$

but in general

$$(AB)[x][x] \neq (A(B[x]))[x]. \tag{8}$$

The proof of (6)-(8) can be found in [13].

The *width of an interval* $[a]$ is defined to be

$$d[a] = a_2 - a_1. \tag{9}$$

The *absolute value* or modulus *of an interval* $[a]=[a_1,a_2]$ is

$$|[a]| = \max\{|a_1|, |a_2|\}. \tag{10}$$

If $0 \in [a]$ then it holds that

$$|[a]| \leqq d[a]. \tag{11}$$

Furthermoe the following rules hold:

$$d([a] \pm [b]) = d[a] + d[b], \tag{12}$$

$$d([a][b]) \leqq |[a]|\, d[b] + d[a]\, |[b]|, \tag{13}$$

$$d(a[b]) = |a| \cdot d[b], \qquad a \in \mathbb{R}. \tag{14}$$

In [1], p. 16, Theorem 10, it was proven that if $0 \in [a]$ and if for $[b]=[b_1,b_2]$ either $b_1 \geqq 0$ or $b_2 \leqq 0$ then it holds that

$$d([a][b]) = d[a] \cdot |[b]|. \tag{15}$$

We now consider the more general case that $0 \in [a]$ and $0 \in [b]$. Under these assumptions it holds that

$$d([a][b]) \leqq d[a] \cdot d[b]. \tag{16}$$

We omit the details of a proof.

For an *interval vector* $[z]$ the *width* is defined to be the real vector $d[z] = (d[z]_i)$. The *absolute value* $|[z]|$ of an *interval vector* is the real vector $|[z]| = (|[z]_i|)$. Using (12) and (13) one shows that for an interval matrix $[A]$ and an interval vector $[z]$ the relation

$$d([A][z]) \leqq |[A]|\, d[z] + d([A])\, |[z]| \tag{17}$$

holds.

We now state the following: If $0 \in [z]$ (that is if $0 \in [z]_i$, $i=1,\dots,m$) then for the real matrix $A=(a_{ij})$ and the bilinear operator $B=(b_{ijk})$ it holds that

$$d((A + B[z])[z]) \leqq (|A| + |B|\, d[z])\, d[z] \tag{18}$$

where $|A|=(|a_{ij}|)$, $B=(|b_{ijk}|)$.

The proof can be performed by using (15) and (16).

The *center of an interval* $[a]$ is denoted by $m[a]$. It holds that

$$m[a]=\frac{a_1+a_2}{2} \qquad \text{for } [a]=[a_1,a_2]. \tag{19}$$

For an *interval vector* $[z]$ the *center* $m[z]$ is defined to be the real vector

$$m[z]=(m[z]_i). \tag{20}$$

Using the definition of $m[z]$ and $d[z]$ we occasionally represent $[z]$ as

$$[z]=m[z]+\tfrac{1}{2}[-d[z],d[z]]. \tag{21}$$

Using the mapping g, defined by (1.5), we now introduce an interval vector $g([z])$, defined by means of an interval vector $[z]$ and a given real vector z^0:

$$g([z])=z^0-Lf(z^0)+\{I_m-Lf'(z^0)-\tfrac{1}{2}(Lf''(z^0))([z]-z^0)\}([z]-z^0). \tag{22}$$

$g([z])$ has the following fundamental property.

Theorem 1. *If the quadratic equation* (1.2) *has a solution* z^* *in* $[z]$ *then* $z^*\in g([z])$.

Proof. By (1.4) it holds that $z^*=g(z^*)$. Because of $z^*\in[z]$ it follows, using (3), that $z^*=g(z^*)\in g([z])$. □

We close this section with *two remarks* about the interval vector $g([z])$:

1) Because of the subdistributive law (2), one gets a larger interval vector in the set theoretic sense if one is multiplying out in (22) eliminating the braces in this manner.

2) If $f'(z^0)$ is nonsingular then one can choose $L=f'(z^0)^{-1}$. Then $g([z])$ reads

$$g([z])=z^0-Lf(z^0)-\tfrac{1}{2}\{(Lf''(z^0))([z]-z^0)\}([z]-z^0).$$

Note that because of (6), (7) and (8) one has in this case the optimal order of the appearing terms in order to get the smallest interval vector in the set theoretic sense.

3. Inclusion Theorems

In this section we consider the problem of constructing sets in $\mathbb{R}^m$ which contain at least one solution of the quadratic equation (1.2).

Theorem 1. *Suppose that for some interval vector* $[z]$ *it holds that* $g([z])\subseteq[z]$. *If* L *is nonsingular then the quadratic equation* (1.2) *has at least one solution in* $g([z])$.

Proof. We consider the mapping $h\colon \mathbb{R}^m\to\mathbb{R}^m$ defined by

$$h(z)=z-Lf(z),$$

where $f(z)$ is given by (1.1) (or by (1.3)). The mapping h is continuous. From (1.3) we get by using (2.3) that for all $z \in [z]$

$$\begin{aligned} h(z) &= z - Lf(z) \\ &= z^0 - Lf(z^0) + z - z^0 - L(f(z) - f(z^0)) \\ &= z^0 - L(z^0) + (z - z^0) - L\{f'(z^0)(z - z^0) + \tfrac{1}{2} f''(z^0)(z - z^0)^2\} \\ &= z^0 - Lf(z^0) + \{I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))(z - z^0)\}(z - z^0) \in g([z]). \end{aligned}$$

By assumption $g([z]) \subseteq [z]$. Therefore the continuous mapping h, defined on the compact and convex set $[z]$, has as its range a subset of $[z]$. By Brouwer's fixed point theorem h has a fixed point z^* in $[z]$, that is $z^* = h(z^*) = z^* - Lf(z^*)$ holds. Since L is nonsingular the assertion follows. □

In the preceding Theorem $z^0 \in \mathbb{R}^m$ can be chosen as an arbitrary real vector from $\mathbb{R}^m$. If one chooses z^0 to be the center of $[z]$ then one has the following results.

Corollary 1. *Let $z^0 \in \mathbb{R}^m$ and assume that the matrix L is non-singular. For the quadratic operator* (1.1) *we define the matrix $K = (k_{ij})$ by*

$$K = (k_{ij}) = |I_m - Lf'(z^0)|, \tag{1}$$

the bilinear operator $H = (h_{ijk})$ by

$$H = (h_{ijk}) = |Lf''(z^0)| \tag{2}$$

and the vector ε by

$$\varepsilon = (\varepsilon_i) = |Lf(z^0)|. \tag{3}$$

If the inequality (for the vector $\beta \in \mathbb{R}^m$)

$$\varepsilon + K\beta + \tfrac{1}{2} H \beta^2 \leqq \beta \tag{4}$$

has a solution $\beta \geqq 0$ then the quadratic equation (1.2) *has at least one solution in $[z] = z^0 + [-\beta, \beta]$.*

Proof. We have $[z] - z^0 = [-\beta, \beta]$ and $d([z] - z^0) = d[z] = 2\beta$. Hence $g([z])$ can be written as

$$\begin{aligned} g([z]) &= z^0 - Lf(z^0) + \{I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))[-\beta, \beta]\}[-\beta, \beta] \\ &= z^0 - Lf(z^0) + (I_m - Lf'(z^0))[-\beta, \beta] + \tfrac{1}{2}(Lf''(z^0))[-\beta, \beta]^2 \end{aligned}$$

from which it follows that $dg([z]) = 2K\beta + H\beta^2$.

Using this equation and (3) then (4) is equivalent to

$$|Lf(z^0)| + \tfrac{1}{2} d(g[z]) \leqq \tfrac{1}{2} d[z],$$

or, because of

$$mg[z] = z^0 - Lf(z^0), \quad m[z] = z^0,$$

to

$$|m[z] - mg[z]| + \tfrac{1}{2} dg[z] \leqq \tfrac{1}{2} d[z].$$

The last inequality is equivalent to $g([z]) \subseteq [z]$. The assertion now follows from Theorem 1. □

The next result shows that under certain additional assumptions a solution $\beta \geqq 0$ of the inequality (4) can be found explicitly.

Corollary 2. *Let $z^0 \in \mathbb{R}^m$ and let K, H and ε be defined as in Corollary 1. Let κ_p and h_p be defined as in Lemma 1.1. If*

$$\|\kappa_p\|_q < 1, \tag{5}$$

$$(1-\|\kappa_p\|_q)^2 - 2\,\|h_p\|_q\,\|\varepsilon\|_q \geqq 0, \tag{6}$$

$$a_{1/2} = \frac{-(\|\kappa_p\|_q - 1) \mp \sqrt{(\|\kappa_p\|_q - 1)^2 - 2\,\|h_p\|_q\,\|\varepsilon\|_q}}{\|h_p\|_q}, \tag{7}$$

and if for some $a \in [a_1, a_2]$,

$$\beta = \varepsilon + a\,\kappa_p + \tfrac{1}{2} a^2\, h_p \tag{8}$$

then the quadratic equation (1.2) *has at least one solution in*

$$[z] = z^0 + [-\beta, \beta].$$

Proof. Since $\rho(K) < 1$ by (1.7) and (5), it follows by the Perron-Frobenius theory on nonnegative matrices (see [29]) that

$$\rho(I_m - Lf'(z^0)) \leqq \rho(|I_m - Lf'(z^0)|) = \rho(K) < 1.$$

Hence L is nonsingular.

If after multiplying (1.3) by L from the left the bilinear operator $Lf''(z^0)$ vanishes (that is $Lf''(z^0) = 0$) then (1.3) is a linear equation. This case is considered as a trivial one in connection with quadratic equations and it is therefore excluded. We therefore have $h_p \neq 0$ for the vector h_p defined by (1.11). It follows that $\|h_p\|_q \neq 0$. Hence the denominator does not vanish in (7) and a_1 and a_2 are well-defined. Furthermore $a_1 \geqq 0$ and therefore $\beta \geqq 0$ for the vector β defined by (8). Subsequently we use the fact that a_1 and a_2 are the solutions of the quadratic equation

$$\tfrac{1}{2}\,\|h_p\|_q\, a^2 + (\|\kappa_p\|_q - 1)\, a + \|\varepsilon\|_q = 0.$$

Therefore we have for $a \in [a_1, a_2]$ that

$$\tfrac{1}{2}\,\|h_p\|_q\, a^2 + \|\kappa_p\|_q\, a + \|\varepsilon\|_q \leqq a. \tag{9}$$

By Lemma 1.1 we have for $p \geqq 1$, $q \geqq 1$, $\dfrac{1}{p} + \dfrac{1}{q} = 1$ that $K\beta \leqq \|\beta\|_q\, \kappa_p$ and $H\beta^2 \leqq \|\beta\|_q^2\, h_p$.

Therefore the inequality (4) from Corollary 1 certainly holds if $\beta \geqq 0$ is a solution of the inequality

$$\varepsilon + \|\beta\|_q\, \kappa_p + \tfrac{1}{2}\,\|\beta\|_q^2\, h_p \leqq \beta. \tag{10}$$

We now show that β, defined by (8), is a solution of this inequality. Using (9) it follows from (8) that

$$\|\beta\|_q \leqq \|\varepsilon\|_q + a\,\|\kappa_p\|_q + \tfrac{1}{2}a^2\,\|h_p\|_q \leqq a.$$

Hence it follows from (8) that

$$\beta = \varepsilon + a\,\kappa_p + \tfrac{1}{2}a^2\,h_p \geqq \varepsilon + \|\beta\|_q\,\kappa_p + \|\beta\|_q^2\,h_p.$$

The assertion now follows from Corollary 1. □

Some of the results of Corollary 2 were proven by completely different methods by Yamamoto in a series of papers [24–27]. However, he considered only the (under practical aspects most important) cases $p=1,2,\infty$. Furthermore he could only allow $a=a_1$ in (8) which of course gives the smallest inclusion $[z]=z^0+[-\beta,\beta]$.

In Chap. 5 we will prove that the quadratic equation (1.2) has a unique solution in $[z]=z^0+[-\beta,\beta]$ if in (8) the real number a is chosen to be equal to $\dfrac{a_1+a_2}{2}$ where a_1 and a_2 are defined by (7) provided that $a_1 \neq a_2$.

We stress the fact that the real vector z^0 in Theorem 1 can be chosen arbitrarily. Therefore by choosing z^0 appropriately one has a greater chance to have $g([z]) \subseteq [z]$ for some interval vector $[z]$ compared with Corollary 1 or Corollary 2 where z^0 has to be the center of $[z]$.

4. Exclusion Theorems

We start with the following basic result.

Theorem 1. *Let $z^0 \in \mathbb{R}^m$ and let L be an (m,m) matrix. If $[z]$ is some interval vector then the quadratic equation* (1.2) *has no solution in $[z]\backslash g[z])$. ($[z]\backslash g[z]$ denotes the set theoretic difference of the interval vectors $[z]$ and $g([z])$.)*

Proof. By Theorem 1.1 we have for all solutions $z^* \in [z]$ of the quadratic equation (1.2) that $z^* \in g([z])$. Hence in $[z]\backslash g([z])$ exists no solution of (1.2). □

Corollary 1. *Let $z^0 \in \mathbb{R}^m$ and assume that for some real matrix L it holds that*

$$g([z]) \cap [z] = \emptyset \tag{1}$$

for some interval vector $[z]$. ((1) is true if $g([z])_i \cap [z]_i = \emptyset$ for at least one i). Then the quadratic equation (1.2) *has no solution in $[z]$.*

Proof. From (1) it follows that $[z]\backslash g[z]=[z]$. The assertion now follows from Theorem 1. □

Corollary 2. *Let K, H and ε be defined as in Corollary* 3.1. If the vector $\beta \geqq 0$ *is a solution of the inequality*

$$\beta + K\beta + \tfrac{1}{2}H\beta^2 \leqq \varepsilon, \tag{2}$$

where in (2) *the equality-sign is excluded*[2], *then the quadratic equation* (1.2) *has no solution in*

$$[z]=z^0+[-\beta,\beta].$$

Proof. Because of $[z]-z^0=[-\beta,\beta]$ we have $dg([z])=2K\beta+H\beta^2$ (See also the proof of Corollary 3.1). Furthermore

$$|m[z]-mg[z]|=|Lf(z^0)|=\varepsilon.$$

Therefore (2) can be written as

$$\tfrac{1}{2}\{d[z]+dg([z])\}\leqq|m[z]-mg[z]|$$

where the equality-sign is excluded. Hence $g([z])\cap[z]=\emptyset$ holds. The assertion now follows from Corollary 1. □

Under appropriate assumptions one can find solutions of the inequality (2).

Corollary 3. *Assume that* K, H *and* ε *are defined as in Corollary* 3.1. *Let*

$$a=\frac{-(1+\|\kappa_\infty\|_1)+\sqrt{(1+\|\kappa_\infty\|_1)^2+2\,\|\varepsilon\|_1\,\|h_\infty\|_1}}{\|h_\infty\|_1}$$

If $\varepsilon-\frac{1}{2}a^2h_\infty-a\kappa_\infty\geqq 0$, *where the equality-sign is excluded and if* β *is chosen according to*

$$0\leqq\beta\leqq\varepsilon-\tfrac{1}{2}a^2h_\infty-a\kappa_\infty,\qquad \beta\neq\varepsilon-\tfrac{1}{2}a^2h_\infty-a\kappa_\infty, \tag{3}$$

then the quadratic equation (1.2) *has no solution in*

$$[z]=z^0+[-\beta,\beta].$$

Proof. We show that β from above is a solution of (2) where the equality-sign is excluded. Because of $K\beta\leqq\|\beta\|_1\kappa_\infty$ and $K\beta^2\leqq\|\beta\|_1^2h_\infty$ (see Lemma 1.1) it follows that

$$\beta+K\beta+\tfrac{1}{2}H\beta^2\leqq\beta+\|\beta\|_1\kappa_\infty+\tfrac{1}{2}\|\beta\|_1^2h_\infty.$$

Therefore (2) holds if β is a solution of

$$\beta+\|\beta\|_1\kappa_\infty+\tfrac{1}{2}\|\beta\|_1^2h_\infty\leqq\varepsilon \tag{4}$$

where the equality-sign is excluded.

From (3) it follows that

$$\beta+\tfrac{1}{2}a^2h_\infty+a\kappa_\infty\leqq\varepsilon,\qquad \beta+\tfrac{1}{2}a^2h_\infty+a\kappa_\infty\neq\varepsilon.$$

Because of $\beta\geqq 0$, $\frac{1}{2}a^2h_\infty\geqq 0$, $a\kappa_\infty\geqq 0$ we have that

$$\|\beta\|_1+\tfrac{1}{2}a^2\|h_\infty\|_1+a\|\kappa_\infty\|_1=\|\beta+\tfrac{1}{2}a^2h_\infty+a\kappa_\infty\|_1<\|\varepsilon\|_1,$$

or, since a is a solution of the quadratic equation

$$\|\varepsilon\|_1-\tfrac{1}{2}a^2\|h_\infty\|_1-a\|\kappa_\infty\|_1=a,$$

that $\|\beta\|_1<a$.

[2] By this we mean that for at least one component the -sign holds

Therefore if β is chosen according to (3), where the equality-sign is excluded then

$$\varepsilon-\|\beta\|_1\,\kappa_\infty-\tfrac{1}{2}\,\|\beta\|_1^2\,h_\infty\geqq\varepsilon-\tfrac{1}{2}a^2\,h_\infty-a\,\kappa_\infty\geqq\beta$$

where the equality-sign is excluded. □

Corollary 4. *Let K, H and ε be defined as in Corollary* 3.1. *Assume that* $\delta=\min\limits_{1\leqq i\leqq m}\varepsilon_i>0$.

If

$$a_2=\frac{-(1+\|\kappa_1\|_\infty)+\sqrt{(1+\|\kappa_1\|_\infty)^2+2\,\|h_1\|_\infty\,\delta}}{\|h_1\|_\infty},\qquad 0\leqq a<a_2,$$

and

$$\beta=a\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}, \tag{5}$$

then

$$\beta+K\,\beta+\tfrac{1}{2}H\,\beta^2<\varepsilon. \tag{6}$$

Hence by Corollary 2 *the quadratic equation* (1.3)) *has a solution in*

$$[z]=z^0+[-\beta,\beta].$$

Proof. Using β defined by (5) we certainly have that

$$\beta+K\,\beta+\tfrac{1}{2}H\,\beta^2=a\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}+a\left(\sum_{j=1}^{m}k_{ij}\right)+\tfrac{1}{2}a^2\left(\sum_{j=1}^{m}\sum_{k=1}^{m}h_{ijk}\right)$$

$$=a\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}+a\,\kappa_1+\tfrac{1}{2}a^2\,h_1<\varepsilon$$

if

$$a+a\,\|\kappa_1\|_\infty+\tfrac{1}{2}a^2\,\|h_1\|_\infty<\min_{1\leqq i\leqq m}\varepsilon_i=\delta.$$

Since the quadratic

$$\tfrac{1}{2}\,\|h_1\|_\infty\,a^2+\|\kappa_1\|_\infty\,a+a-\delta=0$$

has the zeroes

$$a_1=\frac{-(1+\|\kappa_1\|_\infty)-\sqrt{(1+\|\kappa_1\|_\infty)^2+2\,\|h_1\|_\infty\,\delta}}{\|h_1\|_\infty}<0$$

and

$$a_2=\frac{-(1+\|\kappa_1\|_\infty|+\sqrt{(1+\|\kappa_1\|_\infty)^2+2\,\|h_1\|\,\delta}}{\|h_1\|_\infty}$$

we have

$$\tfrac{1}{2}H\,\beta^2+K\,\beta+\beta<\varepsilon$$

for β defined by (5) if $0\leqq a<a_2$. □

5. Convergence Results

Using Theorem 3.1 we have proven a series of inclusion results in Chap. 3. The basic idea was to prove $g([z]\subseteq[z]$ for some given interval vector $[z]$.

Similarly, using Theorem 4.1 we have proven a series of exclusion results. In this case the basic idea was to prove that $g([z])\cap[z]=\emptyset$ holds for some interval vector $[z]$.

There is, however, a third case besides of $g([z])\subseteq[z]$ or $g([z])\cap[z]=\emptyset$, namely that

$$g([z])\cap[z]\neq\emptyset \quad \text{and} \quad g([z]\nsubseteq[z]. \tag{0}$$

For this last case we cannot make a statement about the existence or nonexistence of a solution of the quadratic equation (1.1) using the results of the preceding Chapters.

We now introduce an iteration method which has the interesting property that under appropriate conditions it will break down after a finite number of steps if (0) holds and if there exists no solution in $[z]$. If, however, there exists a solution in $[z]$ then this method will converge to this solution.

The method introduced in the next theorem can of course also be considered if $g([z])\subseteq[z]$, that is if we know in advance that there exists a solution in $[z]$. We return to this special case later.

Theorem 1. *Let $z^0\in\mathbb{R}^m$ and let $[z]$ be an interval vector. Let K and H be defined as in Corollary 3.1. Assume that the real matrix*

$$M=K+H\,|[z]-z^0| \tag{1}$$

has spectral radius less than one: $\rho(M)<1$. Then for the iteration method

$$\begin{cases}[z]^0=[z]\\ [z]^{k+1}=g([z]^k)\cap[z]^k, \qquad k=0,1,2,\ldots\end{cases} \tag{V}$$

the following holds:

1) *If there exists a zero z^* of the quadratic equation (1.3) in $[z]$ then (V) is well-defined, that is the intersection (which is to be understood component-wise) is never empty. Furthermore $z^*\in[z]^k$, and $\lim\limits_{k\to\infty}[z]^k=z^*$ (which means that the bounds of $[z]^k$ are converging to z^* from below and above). $f'(z^*)$ is nonsingular. z^* is unique in $[z]$.*

2) *If (1.3) has no solution in $[z]$, then there exists a $k_0\geqq 0$ such that $g([z]^{k_0})\cap[z]^{k_0}=\emptyset$ that is (V) is breaking down after a finite number of steps because of empty intersection.*

Proof. Of 1): If (1.3) has a solution z^* in $[z]$, then $z^*=g(z^*)$ by (1.4) and (1.5) and using (2.3) it follows that

$$z^*=g(z^*)\in g([z])\subseteq g([z])\cap[z]=[z]^1.$$

By mathematical induction it follows that $z^*\in g([z]^k)$, and because of $z^*\in[z]^k$ we conclude that $z^*\in g([z]^k)\cap[z]^k=[z]^{k+1}$. Hence $z^*\in[z]^k$ and the well-definedness of (V) is shown.

Using (2.13) it follows from (V) that for $k \geqq 0$

$$\begin{aligned} d[z]^{k+1} &\leqq d g([z]^k) \\ &= d(\{I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))([z]^k - z^0)\}([z]^k - z^0)) \\ &\leqq |I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))([z]^k - z^0)|\, d[z]^k \\ &\quad + d(\{I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))([z]^k - z^0)\})|[z]^k - z^0| \\ &\leqq \{|I_m - Lf'(z^0)| + |Lf''(z^0)|\,|[z]^k - z^0|\}\, d[z]^k \\ &= (K + H|[z]^k - z^0|)\, d[z]^k. \end{aligned}$$

Because of forming the intersection in (V) it holds that $[z]^k \subseteq [z]$, $k=0,1,2,\ldots$, and therefore $[z]^k - z^0 \subseteq [z] - z^0$, that is $|[z]^k - z^0| \leqq |[z] - z^0|$, $k=0,1,2,\ldots$.

Therefore it follows that

$$\begin{aligned} d[z]^{k+1} &\leqq (K + H|[z] - z^0|)\, d[z]^k \\ &= M\, d[z]^k \\ &\leqq M^k\, d[z], \quad k=0,1,2,\ldots. \end{aligned}$$

Since $\rho(M) < 1$ we conclude that $\lim\limits_{k\to\infty} d[z]^k = 0$, and since $z^* \in [z]^k$ it follows that $\lim\limits_{k\to\infty} [z]^k = z^*$.

We now prove that $f'(z^*)$ is nonsingular. From the representation (1.3) of $f(z)$ it follows that

$$f'(z) = f'(z^0) + f''(z^0)(z - z^0),$$

and therefore, for $z = z^*$,

$$Lf'(z^*) = Lf'(z^0) + (Lf''(z^0))(z^* - z^0).$$

Since $z^* \in [z]$ we have

$$\begin{aligned} |I_m - Lf'(z^*)| &= |I_m - Lf'(z^0) - Lf''(z^0)(z^* - z^0)| \\ &\leqq |I_m - Lf'(z^0)| + |Lf''(z^0)|\,|[z] - z^0| \\ &= K + H|[z] - z^0| = M. \end{aligned}$$

From the Perron-Frobenius theory on nonnegative matrices it follows that $\rho(I_m - Lf(z^*)) \leqq \rho(|I_m - Lf(z^*)|) \leqq \rho(M) < 1$.
Hence $\{I_m - (I_m - Lf'(z^*))\}^{-1}$ exists, that is $f'(z^*)^{-1}$ exists. Finally, we have to prove that z^* is unique in $[z]$. This statement follows immediately from the already proven fact that $\lim\limits_{k\to\infty} [z]^k = z^*$.

Of 2): Assume that $g([z]^k) \cap [z]^k \neq \emptyset$ for all $k \geqq 0$. Analogously as in the proof of 1) one shows that $d[z]^{k+1} \leqq M^k d[z]$, $k=0,1,2,\ldots$, from which it follows that $\lim\limits_{k\to\infty} d[z]^k = 0$. From (V) it follows that

$$[z] = [z]^0 \supseteq [z]^1 \supseteq \ldots \supseteq [z]^k \supseteq [z]^{k+1} \supseteq \ldots.$$

Hence there is a real vector $\hat{z} \in [z]^k$ for which $\lim\limits_{k\to\infty} [z]^k = \hat{z}$. Since $g([x])$ is a continuous function of the interval vector $[x]$ and since $g([x]) \cap [x]$ is a continuous function of $[x]$ if the intersection is nonempty (see, for example, [1], p. 128, Corollary 10) it follows from (V) that for $k \to \infty$

$$\hat{z} = g(\hat{z}) \cap \hat{z} = g(\hat{z}),$$

or that

$$\hat{z} = z^0 - Lf(z^0) + \{I_m - Lf'(z^0) - \tfrac{1}{2}(Lf''(z^0))(\hat{z} - z^0)\}(\hat{z} - z^0). \tag{2}$$

Since

$$\begin{aligned} |I_m - Lf'(z^0)| &\leqq |I_m - Lf'(z^0)| + H\,|[z] - z^0| \\ &= K + H\,|[z] - z^0| = M \end{aligned}$$

the nonsingularity of L follows similarly to the nonsingularity of $f'(z^*)$, which was proved in part 1). Therefore (2) can be written as

$$f(z^0) + f'(z^0)(\hat{z} - z^0) + \tfrac{1}{2}f''(z^0)(\hat{z} - z^0)^2 = 0.$$

On the other hand we have by (1.3)

$$f(\hat{z}) = f(z^0) + f'(z^0)(\hat{z} - z^0) + \tfrac{1}{2}f''(z^0)(\hat{z} - z^0)^2$$

from which it follows with the preceding equation that $f(\hat{z}) = 0$. This is a contradiction to the assumption that $f(z) = 0$ has no solution in $[z]$. □

We add some *remarks* concerning the condition $\rho(M) < 1$ for the matrix M defined by (1). The question arises whether by choosing a special z^0 the condition $\rho(M) < 1$ can be replaced by a weaker one.
If we choose $z^0 \in [z]$, then $0 \in [z] - z^0$ and by applying (3.18) we get from (V) for $k = 0$ that

$$d[z]^1 \leqq dg([z]^0) \leqq (K + \tfrac{1}{2}H \cdot d[z]^0)\, d[z]^0.$$

In part 1 of the proof of Theorem 1 we have shown that for arbitrarily chosen $z^0 \in \mathbb{R}^m$

$$d[z]^{k+1} \leqq (K + H\,|[z]^k - z^0|)\, d[z]^k$$

which for $k = 0$ and $z^0 \in [z]^0$ can be written as

$$d[z]^1 \leqq (K + H\,d[z]^0)\, d[z]^0$$

since $|[z]^0 - z^0| \leqq d[z]^0$. Compared with the preceding inequality the term $\frac{1}{2}H\,d[z]^0$ is multiplied by a factor of two and therefore it seems that by choosing $z^0 \in [z]$ we can gain this factor of two. However, in the next steps of method (V) we don't have $z^0 \in [z]^k$, in general, and therefore (3.18) can no longer be applied. Therefore it seems that the spectral radius condition $\rho(M) < 1$ in Theorem 1 cannot be weakened.

In the following Theorem we will see that under natural conditions on z^0, $f'(z^0)$ and L we always can find an interval vector $[z]$ for which the matrix M defined by (1) has spectral radius less than one.

Theorem 2. *Let K and H be defined as in Corollary 3.1. Assume that* $\|\cdot\|_q = \left(\sum_{i=1}^{m} |u_i|^q\right)^{\frac{1}{q}}$, $q \geqq 1$. *If $\|K\| < 1$ for some matrix norm and if $[z]$ is chosen according to*

$$\| |[z]-z^0| \|_q \leqq \frac{1-\|K\|}{\|H_p\|}, \qquad \frac{1}{p}+\frac{1}{q}=1,$$

where H_p is defined in Lemma 1.2 *then $\rho(M)<1$ for the matrix M defined by* (1).

Proof. By Lemma 1.2 we have for all interval vectors $[z]$ that

$$M=K+H|[z]-z^0|\leqq K+H_p\cdot\| |[z]-z^0| \|_q=:M_1.$$

Therefore by the Perron Frobenius theory on nonnegative matrices certainly $\rho(M)<1$ if $\rho(M_1)<1$. The condition $\|M_1\|<1$ is sufficient for $\rho(M_1)<1$. Since

$$\begin{aligned}\|M_1\|&=\|K+H_p\cdot\| |[z]-z^0| \|_q\|\\ &\leqq\|K\|+\|H_p\|\cdot\| |[z]-z^0| \|_q,\end{aligned}$$

we have $\|M_1\|<1$ if

$$\| |[z]-z^0| \|_q \leqq \frac{1-\|K\|}{\|H_p\|}. \quad \square$$

In passing we note that $\|K\|<1$ can always be fullfilled if $f'(z^0)$ is nonsingular and if then L is a sufficiently good approximation to $f'(z^0)^{-1}$.

We now consider the method (V) under more special assumptions.

Theorem 3. *Assume that for some interval vector $[z]$ we have that $g([z])\subseteq[z]$. Then* (V) *is well-defined and it holds that $\lim_{k\to\infty}[z]^k=[z]^*$ where in general $d[z]^*\neq 0$, that is $[z]^*$ is an interval vector. There are no solutions of* (1.3) $[z]\backslash[z]^*$.

Proof. Because of $g([z])\subseteq[z]$ it follows by mathematical induction that $[z]=[z]^0\supseteq[z]^1\supseteq\ldots[z]^k\supseteq\ldots$ and hence $\lim_{k\to\infty}[z]^k=[z]^*$.
Furthermore, by Theorem 1.1, $z^*\in g([z]^0)=[z]^1$ for all solutions z^* of the quadratic equation (1.3). By complete induction we get $z^*\in g([z]^k)=[z]^{k+1}$ for all $k\geqq 0$ and therefore $z^*\in[z]^*$. Hence there are no solutions in $[z]\backslash[z]^*$. $\square$

We now use a simple *example* (see Böhm [3], Ch. 1) to demonstrate that even the stronger condition $g([z]\subset[z]$ is not sufficient for the convergence of the sequence $\{[z]^k\}_{k=0}^{\infty}$ to a real vector: Consider the quadratic operator

$$f(z)=-z+z^2, \qquad z\in\mathbb{R}^1$$

in $\mathbb{R}^1$ and the interval $[z]=[0,2]$.

We choose $z^0=1$ and $L=1$. Then using $[z]^0=[z]$ we get

$$g([z]^0)=1-([z]^0-1)^2=1-[0,1]=[0,1]=[z]^1,$$

hence $g([z]^0)\subset[z]^0$. The following iterates are all equal to $[z]^1$. Therefore $\lim_{k\to\infty}[z]^k=[z]^*=[z]^1$. $\square$

We now discuss the question of which additional assumptions one has to impose besides of $g([z]\subseteq[z]$ in order that $\lim_{k\to\infty}[z]^k=z^*$, where z^* is a real vector. In order to discuss this we now choose $z^0 = m[z]$ and consider the even stronger assumption $g([z]) \subset [z]$ (which is defined as $g([z])_i \subset [z]_i$, $i = 1, 2, \ldots, m$). It then follows that $d[z] = 2\beta > 0$. A similar proof as that of Corollary 3.1 shows that the assumption $g([z]) \subset [z]$ implies that

$$\varepsilon + K\beta + \tfrac{1}{2}H\beta^2 < \beta$$

from which it follows that

$$K\beta + \tfrac{1}{2}H\beta^2 = (K + \tfrac{1}{2}H\beta)\beta < \beta.$$

Since $\beta > 0$ it follows by Corollary 3 in [29], p. 18, that the spectral radius of the matrix

$$M_0 = K + \tfrac{1}{2}H\beta = K + \tfrac{1}{2}H\frac{d[z]}{2} = K + \tfrac{1}{2}H|[z] - z^0|$$

is less than one. For the convergence of (V) to a real vector we need, however, by Theorem 1 the condition $\rho(M) < 1$ which because of

$$M_0 = K + \tfrac{1}{2}H|[z] - z^0| \leqq K + H|[z] - z^0| = M$$

is a stronger condition than $\rho(M_0) < 1$.

The discussion is made more precise in the next theorem.

Theorem 4. *Let* $[z] = z^0 + [-\beta, \beta]$ *where* β *is a solution of the inequality*

$$\varepsilon + K\beta + \tfrac{1}{2}H\beta^2 \leqq \beta$$

and where K, H *and* ε *are defined as in Corollary* 3.1. *If besides of this*

$$\rho(K + H|[z] - z^0|) = \rho(K + H\beta) < 1$$

then (V) *is convergent to the unique solution* z^* *of* (1.3) *in* $[z]$.

Proof. By Corollary 3.1 we have $g[z] \subseteq [z]$ and therefore there exists at least one solution z^* of (1.3) in $[z]$. By Theorem 1 $\lim_{k\to\infty} [z]^k = z^*$. □

The following Corollary shows that by choosing in (3.8) the number a appropriately the spectral radius condition of the preceding theorem holds. Therefore we have the uniqueness result under the assumptions of Corollary 3.2 which was already announced in Chap. 3.

Corollary 1. *Assume that the conditions of Corollary* 3.2 *hold where in* (3.6) *the equality sign is excluded. If in* (3.8) *the real number* a *is chosen to be less than* $(a_1 + a_2)/2$, *where* a_1 *and* a_2 *are defined by* (3.7) *then*

$$\rho(K + H|[z] - z^0|) < 1$$

for

$$[z] = z^0 + [-\beta, \beta],$$

that is the iteration method (V) *is for this* $[z]$ *convergent to the unique solution* z^* *of the quadratic equation* (1.3) *in* $[z]$.

Proof. For $[z]=z^0+[-\beta,\beta]$ it follows, using Lemma 1.2, that

$$K+H|[z]-z^0|\leqq K+H\beta\leqq K+H_p\|\beta\|_q$$

where $p\geqq 1$, $q\geqq 1$, $\frac{1}{p}+\frac{1}{q}=1$.

Hence by the Perron Frobenius theory for nonnegative matrices certainly $\rho(K+H|[z]-z^0|)<1$ if $\rho(K+H_p\|\beta\|_q)<1$.

We use the matrix norm $\|\cdot\|=\sup\limits_{\|u\|_q=1}\|\cdot u\|_q$, $q\geqq 1$.

We then have $\|K\|\leqq\|\kappa_p\|_q$. Since for a nonnegative vector u we have by Lemma 1.1 that $H_p u\leqq\|u\|_q h_p$ it follows, using that the vector norm $\|\cdot\| = \left(\sum\limits_{i=1}^{m}|u_i|^q\right)^{\frac{1}{q}}$ is absolute, that

$$\|H_p\|=\sup_{\|u\|_q=1}\|H_p u\|_q\leqq\|h_p\|_q.$$

Therefore

$$\begin{aligned}\|K+H_p\|\beta\|_q\| &\leqq\|K\|+\|\beta\|_q\|H_p\|\\ &\leqq\|\kappa_p\|_q+\|\beta\|_q\|h_p\|_q.\end{aligned}$$

If we now choose

$$\beta=\varepsilon+a\kappa_p+\tfrac{1}{2}a^2h_p$$

where $a=\frac{a_1+a_2}{2}$, then, as was shown in the proof of Corollary 3.2,

$$\|\beta\|_q\leqq a<\frac{a_1+a_2}{2}=\frac{1-\|\kappa_p\|_q}{\|h_p\|_q}.$$

Therefore

$$\|K+H_p\|\beta\|_q\|\leqq\|\kappa_p\|_q+\|\beta\|_q\|h_p\|_q<1.$$

Hence $\rho(K+H_p\|\beta\|_q)<1$ and all statements follow from the preceding Theorem. □

6. Numerical Examples

As a simple but very important example to the quadratic equation (1.3) we consider the algebraic eigenvalue problem

$$Tx=\lambda x \tag{1}$$

where $T=(t_{ij})$ is a real (n,n) matrix. We assume that the eigenvector $x=(x_i)$ has Euclidean length one:

$$\|x\|_2^2=\sum_{i=1}^{n}|x_i|^2=1. \tag{2}$$

If we set $z^T=(x_1,x_2,\dots,x_n,\lambda)$, then (1) and (2) can be written as a system of nonlinear equations, namely

$$f(z)=\begin{pmatrix}(T-\lambda I)x\\ \frac{1}{2}(1-\|x\|_2^2)\end{pmatrix}=0. \tag{3}$$

It is well known that (3) is a quadratic equation of the form (1.2) where $m=n+1$ and

$$c^T=(0,\ldots,0,\tfrac{1}{2}),\qquad c\in\mathbb{R}^m, \tag{4}$$

$$A=\left(\begin{array}{c|c} T & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline \underbrace{0\ldots0}_{n} & \underbrace{0}_{1}\end{array}\right)\begin{array}{l}\}\,n\\ \}\,1\end{array}, \tag{5}$$

$$B=\tfrac{1}{2}\underbrace{\left(\begin{array}{c|c|c|c|c|c|c} \bigcirc & \begin{matrix}-1\\0\\ \vdots\\0\end{matrix} & \cdots & \bigcirc & \begin{matrix}0\\ \vdots\\0\\-1\end{matrix} & \begin{matrix}-1 & & \bigcirc\\ & \ddots & \\ \bigcirc & & -1\end{matrix} & \begin{matrix}0\\ \vdots\\ 0\end{matrix}\\ \hline \underbrace{-1\ldots0}_{n} & \underbrace{0}_{1} & & \underbrace{0\ldots0\;-1}_{n} & \underbrace{0}_{1} & \underbrace{0\;\ldots\;0}_{n} & \underbrace{0}_{1}\end{array}\right)}_{n+1}\begin{array}{l}\}\,n\\ \}\,1\end{array}. \tag{6}$$

In passing we note that there exists a series of papers starting with Unger [28] in which Newton's method was applied to the nonlinear system (3). See, for example Collatz [5], Krawczyk [7], Rall [15], Rokne [17], Rump [18–20], Symm-Wilkinson [21] and others.

For the mapping (3) we get by using (4, 5) and (6) that

$$f(z)=c+Az+Bz^2,\qquad f'(z)=A+2Bz,\qquad f''(z)=2B.$$

Therefore $f'(z)$ has the matrix representation

$$f'(z)=\left(\begin{array}{cc} T & -x\\ \hline \underbrace{-x^T}_{n} & \underbrace{0}_{1}\end{array}\right)\begin{array}{l}\}\,n\\ \}\,1\end{array} \tag{7}$$

and $f''(z)$ is the bilinear operator defined in (6), multiplied by the factor two.

If we choose $L=f'(z^0)^{-1}$ then the interval vector $g([z])$, defined in (2.22) reads

$$g([z])=z^0-f'(z^0)^{-1}f(z^0)-\tfrac{1}{2}((f'(z^0)^{-1}f''(z^0))([z]-z^0))([z]-z^0).$$

For a given $z^0\in\mathbb{R}^m$ we now use Corollary 3.2 in order to compute an interval vector $[z]=z^0+[-\beta,\beta]$ in which there exists a solution of (3). We have in this case

$$K=0,\qquad H=|f'(z^0)^{-1}f''(z^0)|,\qquad \varepsilon=|f'(z^0)^{-1}f(z^0)|.$$

We now choose $p=1$, $q=\infty$. Then Corollary 3.2 states that if

$$1-2\,\|h_1\|_\infty\,\|\varepsilon\|_\infty>0$$

and $\beta=\varepsilon+\frac{1}{2}a_1^2h_1$ where $a_1=\dfrac{1-\sqrt{1-2\,\|h_1\|_\infty\,\|\varepsilon\|_\infty}}{\|h_2\|_\infty}$, then there exists a solution z^* in

$$[z]=z^0+[-\beta,\beta]. \tag{8}$$

Furthermore, by Corollary 5.1, the iteration method (V) introduced in Theorem 5.1 will converge to this solution.

For $L=f'(z^0)^{-1}$ the iteration method (V) reads

$$[z]^{k+1}=\{z^0-f'(z^0)^{-1}f(z^0)-\tfrac{1}{2}((f'(z^0)^{-1}f''(z^0))([z]^k-z^0))([z]^k-z^0)\}\cap[z]^k.$$

We perform this method in the following manner:

In the k-th step we compute an *interval vector* $[y]^k$ with the following property: For each $z^k\in[z]^k$ the solution z^{k+1} of the linear system

$$f'(z^0)(z^{k+1}-z^0)=-f(z^0)-\tfrac{1}{2}f''(z^0)(z^k-z^0)^2$$

is contained in this interval vector $[y]^k$. This can be done by using an interval algorithm which computes for the given real matrix $f'(z^0)$ and the given right-hand side

$$-f(z^0)-\tfrac{1}{2}(f''(z^0)([z]^k-z^0))([z]^k-z^0)$$

an interval vector which includes all possible solutions. (Note that this is not exactly our proposed method (V) since in general $A^{-1}\cdot[b]\neq IGA(A,[b])$ where $IGA(A,[b])$ denotes the result which is delivered if the Gaussian algorithm is applied to the real matrix A and the right hand side $[b]$. See, for example, [2], Lemma 1, for details. This is not very important in our case because the inclusion of the zero is also guaranteed if (V) is modified as described. Furthermore for an interval vector $[b]$ with $d[b]$ small compared with the nullvector the difference between $A^{-1}\cdot[b]$ and $IGA(A,[b])$ is negligible).

Because of the special structure of $f''(z^0)$ the term

$$-\tfrac{1}{2}(f''(z^0)([z]^k-z^0))([z]^k-z^0)$$

simplifies to the interval vector

$$\begin{pmatrix} ([z]^k_{n+1}-z^0_{n+1})([z]^k_1-z^0_1)\\ ([z]^k_{n+1}-z^0_{n+1})([z]^k_2-z^0_2)\\ \vdots\\ ([z]^k_{n+1}-z^0_{n+1})([z]^k_n-z^0_n)\\ \frac{1}{2}\sum_{i=1}^{n}([z]^k_i-z^0_i)^2 \end{pmatrix}. \tag{9}$$

The residual $f(z^0)$ can be written in the form

$$f(z^0)=\begin{pmatrix} Tx^0-\lambda x^0\\ \frac{1}{2}\left(1-\sum_{i=1}^{n}|x_i|^2\right)\end{pmatrix}.$$

Taking into account the special form (9) of

$$-\tfrac{1}{2}(f''(z^0)([z]^k-z^0))([z]^k-z^0)$$

the *right-hand side interval vector*

$$-f(z^0)-\tfrac{1}{2}(f''(z^0)([z]^k-z^0))([z]^k-z^0) \tag{10}$$

can therefore be computed by using only scalar products.

Using the so-called exact scalar product, introduced by Kulisch (see [8–12]) the individual components of the right-hand side can be computed with maximal available precision.

After having computed $[y]^k$ (this is actually done by using the subroutine LGLSI which is available in the PASCAL *SC* language, see [8–12] for details), we form the intersection with $[z]^k$ getting $[z]^{k+1}$.

All computation was done on an APPLE II*E* using the programming language PASCAL *SC* (see [8–12]). This system uses a decimal number system which has 12 digits in the mantissa of a floating point number. Note that all rounding errors are taken into account using this system. Therefore the bounds computed in the following examples are absolutely safe.

Example 1. The first matrix is taken from [24], p. 196. Let

$$T=\begin{pmatrix} 1 & 1 & 0.5 \\ 1 & 1 & 0.25 \\ 0.5 & 0.25 & 2 \end{pmatrix}$$

and

$$x^0=\begin{pmatrix} -0.721\,207\,180 \\ 0.686\,349\,340 \\ 0.093\,727\,970 \end{pmatrix}, \qquad \lambda_0=-0.016\,647\,302.$$

Then we get for .he interval vector $[z]$ defined in (8);

$$\begin{pmatrix} [-0.721\,207\,3] ; & -0.721\,207\,1] \\ [\ \ 0.686\,349\,2] ; & 0.686\,349\,4] \\ [\ \ 0.093\,727\,96]; & 0.093\,727\,98] \\ [-0.016\,647\,33]; & -0.016\,647\,28] \end{pmatrix}.$$

After two iteration steps of (V) we have the final result

$$\begin{pmatrix} [-0.721\,207\,129\,831 \ ; & -0.721\,207\,129\,830\ \ ] \\ [\ \ 0.686\,349\,287\,710 \ ; & 0.686\,349\,287\,711\ \ ] \\ [\ \ 0.093\,727\,963\,498\,7; & 0.093\,727\,963\,498\,8] \\ [-0.016\,647\,283\,606\,4; & -0.016\,647\,283\,606\,3] \end{pmatrix}$$

Example 2. As a second example we consider the matrix

$$T=\begin{pmatrix} 14 & 9 & 6 & 4 & 2 \\ -9 & -4 & -3 & -2 & -1 \\ -2 & -2 & 0 & -1 & -1 \\ 3 & 3 & 3 & 5 & 3 \\ -9 & -9 & -9 & -9 & -4 \end{pmatrix}$$

introduced in [24], p. 197, which has $\lambda=5$ as a simple eigenvalue. The corresponding normalized eigenvector is $x^T=\frac{1}{\sqrt{2}}(1,-1,0,0,0)$.

We choose

$$x^0=\begin{pmatrix} 0.707\ 106\ 58 \\ -0.707\ 107\ 30 \\ -0.387\ 205\ 54\ E-6 \\ 0.332\ 787\ 40\ E-6 \\ 0.508\ 088\ 97\ E-6 \end{pmatrix}$$

and

$$\lambda_0=4.999\,995\,7.$$

Then, for the interval vector $[z]$ defined in (8) we get

$$\begin{pmatrix} [\ \ 0.707\ 106\ 3; & 0.707\ 106\ 8\ ] \\ [-0.707\ 108; & -0.707\ 106\ ] \\ [-0.78E-6; & 0.12E-10\ ] \\ [-0.9E-11; & 0.67E-6\ ] \\ [-0.4E-10; & 0.11E-5\ ] \\ [0.499\ 999\ 1E+1; & 0.500\ 000\ 1E+1] \end{pmatrix}.$$

After three iteration steps of (V) we have the final result:

$$\begin{pmatrix} [\ \ 0.707\ 106\ 781\ 186; & -0.707\ 106\ 781\ 187\ ] \\ [-0.707\ 106\ 781\ 187; & -0.707\ 106\ 781\ 186\ ] \\ [-0.9E-17; & 0.14E-16\ ] \\ [-0.5E-17; & 0.9E-17\ ] \\ [-0.25E-16; & 0.28E-16\ ] \\ [\ \ 0.499\ 999\ 999\ 999E+1; & 0.500\ 000\ 000\ 001E+1] \end{pmatrix}.$$

Example 3. As a final example we consider a symmetric (14,14) matrix T which was introduced in a paper by Brooker and Sumner [4] and which was reconsidered by Wilkinson [22, 23]. Because of lack of space we do not list the matrix elements. Instead we refer to [4], [22, 23] or to Example 4.12 in the book by R.T. Gregory and D.L. Karney [30].

We choose

$$\lambda_0=1.334\,034\,837\,00$$

and

$$x^0=\begin{pmatrix}9.168\ 195\ 046\ 95E-2\\ 1.945\ 051\ 119\ 20E-1\\ 3.189\ 487\ 364\ 33E-1\\ 3.407\ 954\ 910\ 35E-1\\ 1.457\ 661\ 093\ 15E-1\\ 1.918\ 835\ 439\ 20E-1\\ 1.826\ 099\ 068\ 19E-1\\ 3.382\ 100\ 423\ 35E-1\\ 2.884\ 013\ 952\ 30E-1\\ 2.701\ 082\ 857\ 28E-1\\ 3.233\ 984\ 439\ 34E-1\\ 2.955\ 518\ 490\ 31E-1\\ 3.210\ 872\ 399\ 33E-1\\ 2.809\ 351\ 151\ 29E-1\end{pmatrix}$$

Then for the interval vector $[z]$ defined in (8) we get:

$$\begin{pmatrix}[9.168\ 195\ 044E-2; & 9.168\ 195\ 050E-2]\\ [1.945\ 051\ 119E-1; & 1.945\ 051\ 120E-1]\\ [3.189\ 487\ 363E-1; & 3.189\ 487\ 366E-1]\\ [3.407\ 954\ 910E-1; & 3.407\ 954\ 911E-1]\\ [1.457\ 661\ 092E-1; & 1.457\ 661\ 094E-1]\\ [1.918\ 835\ 438E-1; & 1.918\ 835\ 440E-1]\\ [1.826\ 099\ 067E-1; & 1.826\ 099\ 069E-1]\\ [3.382\ 100\ 420E-1; & 3.382\ 100\ 427E-1]\\ [2.884\ 013\ 950E-1; & 2.884\ 013\ 954E-1]\\ [2.701\ 082\ 857E-1; & 2.701\ 082\ 858E-1]\\ [3.233\ 984\ 439E-1; & 3.233\ 984\ 440E-1]\\ [2.955\ 518\ 489E-1; & 2.955\ 518\ 492E-1]\\ [3.210\ 872\ 396E-1; & 3.210\ 872\ 402E-1]\\ [2.809\ 351\ 150E-1; & 2.809\ 351\ 153E-1]\\ [1.334\ 034\ 836\quad ; & 1.334\ 034\ 838]\end{pmatrix}$$

After two iteration steps of (V) we have the final result:

$$\begin{pmatrix}
[9.168\ 195\ 049\ 16E-2; & 9.168\ 195\ 049\ 17E-2] \\
[1.945\ 051\ 119\ 31E-1; & 1.945\ 051\ 119\ 32E-1] \\
[3.189\ 487\ 365\ 32E-1; & 3.189\ 487\ 365\ 33E-1] \\
[3.407\ 954\ 910\ 38E-1; & 3.407\ 954\ 910\ 39E-1] \\
[1.457\ 661\ 092\ 71E-1; & 1.457\ 661\ 092\ 72E-1] \\
[1.918\ 835\ 438\ 95E-1; & 1.918\ 835\ 438\ 96E-1] \\
[1.826\ 099\ 068\ 59E-1; & 1.826\ 099\ 068\ 60E-1] \\
[3.382\ 100\ 420\ 34E-1; & 3.382\ 100\ 420\ 35E-1] \\
[2.884\ 013\ 953\ 95E-1; & 2.884\ 013\ 953\ 96E-1] \\
[2.701\ 082\ 857\ 18E-1; & 2.701\ 082\ 857\ 19E-1] \\
[3.233\ 984\ 439\ 50E-1; & 3.233\ 984\ 439\ 51E-1] \\
[2.955\ 510\ 489\ 32E-1; & 2.955\ 518\ 489\ 33E-1] \\
[3.210\ 872\ 401\ 74E-1; & 3.210\ 872\ 401\ 75E-1] \\
[2.809\ 351\ 150\ 22E-1; & 2.809\ 351\ 150\ 23E-1] \\
[1.334\ 034\ 836\ 95; & 1.334\ 034\ 836\ 96\ \ \ ]
\end{pmatrix}$$

In [30] also the elements of the tridiagonal matrix $\tilde{T}$ are given which one gets if the Givens method is applied to T using a floating point system with 9 digits in the mantissa. In order to compare the influence of the rounding errors which are introduced by the Givens method we now repeat the computation for the matrix $\tilde{T}$.

In this case we choose

$$\lambda_0 = 1.334\,034$$

and

$$x^0 = \begin{pmatrix}
9.168\ 1E-2 \\
6.467\ 91E-2 \\
7.526\ 87E-2 \\
8.177\ 6E-2 \\
5.504E-3 \\
1.68E-4 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}$$

Then we get for the interval vector $[z]$ defined by (8):

$$\begin{pmatrix}
[\ 9.168\,0\ \ E-2;\ 9.168\,2\ \ E-2] \\
[\ 6.467\,90\ E-1;\ 6.467\,92\ E-1] \\
[\ 7.526\,869E-1;\ 7.526\,871E-1] \\
[\ 8.177\,5\ \ E-2;\ 8.177\,7\ \ E-2] \\
[\ 5.503\ \ E-3;\ 5.505\ \ E-3] \\
[\ 1.67\ \ E-4;\ 1.69\ \ E-4] \\
[\ 3\ \ E-6;\ 5\ \ E-6] \\
[-2\ \ E-7;\ 2\ \ E-7] \\
[-4\ \ E-9;\ 4\ \ E-9] \\
[-2\ \ E-10;\ 2\ \ E-10] \\
[-9\ \ E-12;\ 9\ \ E-12] \\
[-7\ \ E-12;\ 7\ \ E-12] \\
[-8\ \ E-12;\ 8\ \ E-12] \\
[-7\ \ E-12;\ 7\ \ E-12] \\
[\ 1.334\,033\ \ ;\ 1.334\,035\ \ ]
\end{pmatrix}$$

After three iteration steps of (V) we have the final result:

$$\begin{pmatrix}
[9.168\,194\,970\,23E-2;\ 9.168\,194\,970\,24E-2] \\
[6.367\,912\,624\,97E-1;\ 6.467\,912\,624\,98E-1] \\
[7.526\,870\,271\,30E-1;\ 7.526\,870\,271\,31E-1] \\
[8.177\,651\,873\,31E-2;\ 8.177\,651\,873\,32E-2] \\
[5.504\,061\,256\,42E-3;\ 5.504\,061\,256\,43E-3] \\
[1.682\,681\,402\,21E-4;\ 1.682\,681\,402\,22E-4] \\
[4.990\,525\,332\,24E-6;\ 4.990\,525\,332\,43E-6] \\
[1.440\,899\,836\,98E-7;\ 1.440\,899\,837\,01E-7] \\
[3.474\,596\,896\,67E-9;\ 3.474\,596\,896\,74E-9] \\
[1.079\,736\,219\,22E-10;\ 1.079\,736\,219\,26E-10] \\
[1.417\,336\,691\,21E-12;\ 1.417\,336\,691\,27E-12] \\
[2.822\,952\,968\,84E-14;\ 2.822\,952\,968\,97E-14] \\
[6.955\,927\,655\,65E-16;\ 6.955\,927\,656\,02E-16] \\
[4.323\,372\,212\,31E-18;\ 4.323\,372\,212\,58E-18] \\
[1.334\,034\,842\,45\ \ ;\ 1.334\,034\,842\,46\ \ ]
\end{pmatrix}$$

The largest eigenvalue of T – rounded to 17 digits – is

$$\lambda_1 = 1.334\,034\,836\,956\,507\,0.$$

If one compares this value with the last component of the preceding interval vector (which is an inclusion for the largest eigenvalue of $\tilde{T}$) then we conclude that the largest eigenvalues of T and $\tilde{T}$, respectively, differ in the 9-th digit of the mantissa. This is not surprising since – as was mentioned above – T was transformed to $\tilde{T}$ by using 9 digits in the floating point mantissa.

We close this paper with two *final comments:*

a) The including sets for the eigenpairs computed in the numerical examples are very close in the sense that in most cases the individual components are included by two neighbouring machine numbers. A theoretical foundation for this will be given in a future paper in which – using the so-called exact scalar product – the influence of rounding errors on the method (V) is studied in detail.

b) Method (V) will not work for a multiple eigenvalue of the matrix T since in this case the $(n+1, n+1)$ matrix $f'(z^*)$ is necessarily singular. If multiple eigenvalues exist we are in the position to formulate a method which takes into account the multiplicity of an eigenvalue. The investigation of this method will be performed in another paper.

References

1. Alefeld, G., Herzberger, J.: Introduction to Interval Computations. New York: Academic Press 1983
2. Alefeld, G., Platzöder, L.: A quadratically convergent Krawczyk-like algorithm. SIAM J. Numer. Anal. **20**, 210–219 (1983)
3. Böhm, H.: Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrücke mit garantierter maximaler Genauigkeit. Thesis, Universität Karlsruhe, 1983
4. Brooker, R.A., Sumner, F.H.: The method of Lanczos for calculating the characteristic roots and vectors of a real symmetric matrix. Proc. I.E.E. **103**, Part B, Suppl. (1), 114 (1956)
5. Collatz, L.: Functional analysis and numerical mathematics. New York: Academic Press 1966
6. Hoffmann, R.: Fehlerschranken für Näherungen von Eigenwerten und zugehörigen Eigenvektoren. Diplomarbeit. Karlsruhe, 1983. (Not available)
7. Krawczyk, R.: Verbesserungen von Schranken für Eigenwerte und Eigenvektoren von Matrizen. Computing **5**, 100–206 (1970)
8. Kulisch, U., Ullrich, C.: Wissenschaftliches Rechnen und Programmiersprachen. Stuttgart: Teubner 1981
9. Kulisch, U.: Grundlagen des numerischen Rechnens. Reihe Informatik 19. Mannheim: Bibliographische Institut 1976
10. Kulisch, U., Miranker, W.L.: Computer Arithmetic in Theory and Practice. New York: Academic Press 1981
11. Kulisch, U., Wippermann, H.-W.: PASCAL *SC*, Pascal *SC* für wissenschaftliches Rechnen. Gemeinschaftsentwicklung von Inst. f. Angew. Mathematik Universität Karlsruhe (Prof. Dr. U. Kulisch) und Fachbereich Informatik, Universität Kaiserslautern (Prof. Dr. H.-W. Wippermann)
12. Kulisch, U., Miranker, W.L.: A New Approach to Scientific Computation. Notes and Reports in Computer Science and Applied Mathematics 7. New York: Academic Press 1983
13. Platzöder, L.: Einige Beiträge über die Existenz von Lösungen nichtlinearer Gleichungssysteme und Verfahren zu ihrer Berechnung. Thesis, Technische Universität Berlin 1981
14. Prenter, P.M.: On Polynomial Operators and Equations. In: Nonlinear Functional Analysis and Applications. (L.B. Rall, ed.), pp. 361–398. New York: Academic Press 1971
15. Rall, L.B.: Computational Solution of Nonlinear Operator Equations. New York: John Wiley 1969

16. Rall, L.B.: Quadratic equations in Banach spaces. Rend. Circ. Mat. Palermo **10**, 314–332 (1961)
17. Rokne, J.: Fehlererfassung bei Eigenwertproblemen von Matrizen. Computing **7**, 145–152 (1971)
18. Rump, S.: Solving algebraic problems with high accuracy. Habilitationsschrift, Karlsruhe 1982
19. Rump, S.: Computer Demonstration Packages Standard Problems of Numerical Mathematics. In [12], pp. 28–49
20. Rump, S.: Solving Algebraic Problems with high Accuracy. In [12], pp. 53–120
21. Symm, H.J., Wilkinson, J.H.: Realistic error bounds for a simple eigenvalue and its associated eigenvector. Numer. Math. **35**, 113–126 (1980)
22. Wilkinson, J.H.: The calculation of the Eigenvectors of Codiagonal Matrices. Comput. **1**, 90–96 (1958)
23. Wilkinson, J.H.: The evaluation of the zeros of ill-conditioned polynomials. Part II. Num. Math. **1**, 167–180 (1959)
24. Yamamoto, T.: Componentwise error estimates for approximate solutions of systems of equations. Lect. Notes Numer. Appl. Anal. **3**, 1–22 (1981)
25. Yamamoto, T.: Error bounds for computed eigenvalues and eigenvectors. Numer. Math. **34**, 189–199 (1980)
26. Yamamoto, T.: Error bounds for computed eigenvalues and eigenvectors II. Numer. Math. **40**, 201–206 (1981)
27. Yamamoto, T.: Error bounds for approximate solutions of systems of equations (Manuscript 1983)
28. Unger, H.: Nichtlineare Behandlung von Eigenwertaufgaben. Z. Angew. Math. Mech. **80**, 281–282 (1950)
29. Varga, R.S.: Matrix Iterative Analysis. New Jersey: Englewood Cliffs 1962
30. Gregory, R.T., Karney, D.L.: A collection of Matrices for testing computational algorithms. New York: Wiley 1969

Received April 13, 1984/September 28, 1984