Rostock. Math. Kolloq. **49**, 75–88 (1995)

Subject Classification (AMS)
65F15, 65G10, 65H10

Günter Mayer

On a Unified Representation of Some Interval Analytic Algorithms

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. In this article we show that several interval analytic algorithms for verifying solutions x^* of various mathematical problems can be viewed as special cases of some iterative method involving the first and the second derivative of the underlying function. Starting with an approximation of x^* a way is given how to construct interval vectors which contain x^*, and how to improve these interval bounds.

KEY WORDS. Systems of nonlinear equations, second order methods, enclosure methods, verification methods, interval methods for nonlinear systems, algebraic eigenproblem, singular value decomposition, quadratic systems, invariant supspaces

1 Introduction

The solutions of many mathematical problems can be expressed as zeros of some function $f : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$. Among these problems are the algebraic eigenproblem, the generalized eigenproblem, the singular value problem and the generalized singular value problem, see [1]–[4], [8]. For example, let $A,\, B \in \boldsymbol{R}^{n\times n}$, $\zeta \in \boldsymbol{R}\backslash\{0\}$ and $i_0 \in \{1, \dots , n\}$ be given. Then with $x := (v^T, \lambda)^T$, $v \in \boldsymbol{R}^n$, $\lambda \in \boldsymbol{R}$ the zeros $x^* = ((v^*)^T, \lambda^*)^T$ of the function

$$f(x) := \begin{pmatrix} Av - \lambda Bv \\ v_{i_0} - \zeta \end{pmatrix} \tag{1}$$

are obviously eigenpairs of the generalized eigenproblem $Av = \lambda Bv$ with the eigenvector $v^* = (v_i^*)$ being normalized by $v_{i_0}^* = \zeta \neq 0$. We will show how certain algorithms for verifying and enclosing solutions of the above problems can be derived from one verification method for general systems of nonlinear equations. To this end we will consider the interval function

$$[g]([x], \widetilde{x}) := t(\widetilde{x}) + \{t'(\widetilde{x}) + [H]([x], \widetilde{x})\}([x] - \widetilde{x}) \tag{2}$$

for which the function $t : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$ is assumed to be twice continuously differentiable on a given open set D. The matrix $t'(\widetilde{x})$ is the Jacobian of t at some fixed vector $\widetilde{x} \in D$. The function $[H] = [H]([x], \widetilde{x})$ is defined for all interval vectors $[x] \subseteq D$ and has $n \times n$ interval matrices as values. It is supposed to be continuous and inclusion monotone with respect to $[x]$.

For t, $[H]$ and all $[x] \subseteq D$ we require

$$t(x) \in [g]([x], \widetilde{x}) \quad \text{for all } x \in [x] \tag{3}$$

and

$$\| \, |[H]([x], \widetilde{x})| \, \| \leq \gamma \| \, |[x] - \widetilde{x}| \, \|. \tag{4}$$

The constant γ is positive and fixed for all $[x] \subseteq D$; it may depend on $\widetilde{x}$. Throughout the paper $\| \cdot \|$ denotes the maximum norm of a real vector and the row sum norm of a real matrix, respectively; $| \cdot |$ is the absolute value which we shall define for interval quantities in our next section. There, we also show how t and $[H]$ normally are related to the given function f, the zeros of which we are interested in. We shall derive criteria (Theorem 1) for guaranteeing the existence of a fixed point x^* of t which then turns out to be a zero of f. In verifying x^* we will construct an interval vector $[x]^0$ which contains x^* and which in a natural way provides lower and upper bounds for it. Some of our criteria will yield the subset property $[g]([x]^0, \widetilde{x}) \subseteq [x]^0$ which, together with Brouwer's fixed point theorem, forms the basis for many verification algorithms. We also will improve the enclosure $[x]^0$ of x^* by considering the iteration

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}) \cap [x]^k, \quad k = 0, 1, \dots ,$$

where the intersection can be dropped if the above–mentioned subset property holds. Under slight additional assumptions on $[H]$ we will show that x^* is the unique fixed point of t within $[x]^0$, that all the iterates $[x]^k$ contain it as element and that they contract to it for $k \to \infty$.

For the standard choice of t and $[H]$ it will turn out that the function $[g]$ reduces to that in Platzöder [14] and Alefeld [5]. Therefore, parts of our criteria are generalizations of results of these authors.

2 Results

In order to formulate our results we first list some notations needed later on. By $\boldsymbol{IR}$, $\boldsymbol{IR}^n$, $\boldsymbol{IR}^{m \times n}$, respectively, we denote the set of real compact intervals, the set of vectors with n

interval components, and the set of $m \times n$ matrices with entries from $\boldsymbol{IR}$. For degenerate intervals $[a,a]$ we simply write a identifying in this way $\boldsymbol{R}$ with $\{[a,a] \mid a \in \boldsymbol{R}\} \subseteq \boldsymbol{IR}$. We proceed similarly for degenerate interval vectors and degenerate interval matrices. Examples are the null matrix O, the identity matrix I, the i–th column $e^{(i)}$ of I and the vector $e := (1,1,\dots,1)^T$. In order to indicate $I \in \boldsymbol{R}^{n\times n}$ we sometimes write I_n instead of I. As usual, we equip $\boldsymbol{R}^n$ and $\boldsymbol{R}^{m\times n}$, respectively, with the natural semi–ordering '$\leq$' which is defined to hold entrywise. We use the notation $[A] = ([a]_{ij}) \in \boldsymbol{IR}^{m\times n}$ simultaneously without further reference, and we assume the same for the elements of $\boldsymbol{R}^n, \boldsymbol{R}^{m\times n}, \boldsymbol{IR}$ and $\boldsymbol{IR}^n$. For $[a] = [\underline{a}, \overline{a}] \in \boldsymbol{IR}$ we define the *absolute value* $|[a]|$ by $|[a]| := \max\{|a| \mid a \in [a]\} = \max\{|\underline{a}|, |\overline{a}|\}$ and the *diameter* $d([a])$ by $d([a]) := \overline{a} - \underline{a}$, and we denote the *convex hull* of $[a]$, $[b] \in \boldsymbol{IR}$ by $[a]\underline{\cup}[b]$. For interval vectors and interval matrices these terms are applied entrywise. If $f(x)$ is an expression for some function f, we write $f([x])$ for the interval arithmetic evaluation of this expression (cf. [5] assuming that $f([x])$ exists. For further details on interval analysis we refer to [5] or [13].

We start by presenting the main result of our paper. Among others, this result lists criteria for proving the existence of a fixed point x^* of t from (2) and for guaranteeing the subset property

$$[g]([x]^0, \tilde{x}) = t(\tilde{x}) + \{t'(\tilde{x}) + [H]([x]^0, \tilde{x})\}([x]^0 - \tilde{x}) \subseteq [x]^0 \tag{5}$$

for some interval vector $[x]^0$.

Theorem 1 *With D, $[g]$, $[H]$, t, $\tilde{x}$ as in (2) – (4) and with $\gamma > 0$ from (4) choose $r \in \boldsymbol{R}$ with $r > 0$ such that $[x]^0 := \tilde{x} + [-r, r]e \subseteq D$, and define α, β by*

$$\alpha := \|\, t(\tilde{x}) - \tilde{x} \,\|, \qquad \beta := \|\, t'(\tilde{x}) \,\|.$$

Let $\beta < 1$, $\quad \Delta := (1-\beta)^2 - 4\alpha\gamma \geq 0$ and let

$$r^- := (1 - \beta - \sqrt{\Delta})/(2\gamma), \quad r^+ := (1 - \beta + \sqrt{\Delta})/(2\gamma).$$

a) If $r \geq r^-$ then t has at least one fixed point $x^* \in [x]^0$. The iteration

$$[x]^{k+1} := [g]([x]^k, \tilde{x}) \cap [x]^k, \quad k = 0, 1, \dots$$

converges to some interval vector $[x]^*$ with

$$x^* \in [x]^* \subseteq [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

b) If $r \in [r^-, r^+]$ then t has at least one fixed point $x^* \in [x]^0$. In addition, $[g]([x]^0, \tilde{x}) \subseteq [x]^0$ holds and the iteration

$$[x]^{k+1} := [g]([x]^k, \tilde{x}), \quad k = 0, 1, \dots$$

converges to some interval vector $[x]^*$ *with*

$$x^* \in [x]^* \subseteq [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

c) *In addition to* (4), *let* $[H]$ *fulfill*

$$\| \, d\,([H]([x], \widetilde{x})) \, \| \leq \delta \, \| \, d([x]) \, \| \tag{6}$$

for all interval vectors $[x] \subseteq D$ *and for some positive number* δ *which is independent of* $[x]$ *but which may depend on* $\widetilde{x}$. *Define* $\hat{\Delta}$, $\hat{r}^-$, $\hat{r}^+$ *as* Δ, r^-, r^+, *with* γ *being replaced by* $\hat{\gamma} := \max\{\gamma, \delta\}$. *If* $\hat{\Delta} \geq 0$ *and if* $r \in [\hat{r}^-, (\hat{r}^- + \hat{r}^+)/2)$ *then the function* t *has exactly one fixed point* $x^* \in [x]^0$; $[g]([x]^0, \widetilde{x}) \subseteq [x]^0$ *holds, and the iteration*

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}), \quad k = 0, 1, \ldots$$

converges to x^* *with*

$$x^* \in [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

Proof: First we remark that the assumptions $\beta < 1$ and $\Delta \geq 0$ guarantee $0 \leq r^- \leq r^+$.

b) Let $[x]^0 := \widetilde{x} + [-r, r]e$. Then (5) is equivalent to

$$[g]([x]^0, \widetilde{x}) - \widetilde{x} \subseteq [x]^0 - \widetilde{x}. \tag{7}$$

Property (7) certainly holds if

$$|t(\widetilde{x}) - \widetilde{x}| + \left(|t'(\widetilde{x})| + |[H]([x]^0, \widetilde{x})|\right) re \leq re,$$

and this, in turn, is true if

$$\alpha + (\beta + \gamma r)r \leq r, \tag{8}$$

where we used (4). Now (8) can be rewritten as

$$\gamma r^2 + (\beta - 1)r + \alpha \leq 0 \tag{9}$$

with equality for $r = r^-$ and $r = r^+$. Hence (9) is fulfilled for each $r \in [r^-, r^+]$, and by (3)

$$t(x) \in [g]([x]^0, \widetilde{x}) \subseteq [x]^0$$

holds for any $x \in [x]^0$. Therefore, Brouwer's fixed point theorem guarantees that t has at least one fixed point $x^* \in [x]^0$. Since $[H]$ was assumed to be inclusion monotone the iterates $[x]^k$ decrease monotonically with respect to the semi–ordering '$\subseteq$'; hence they are convergent to some limit $[x]^*$, and

$$x^* = t(x^*) \in [g]([x]^k, \widetilde{x}) = [x]^{k+1}$$

holds for $k = 0, 1, \ldots$ by induction.

a) Choose $r' \leq r$ such that $r' \in [r^-, r^+]$. Then b) applies with $[\hat{x}]^0 := \widetilde{x} + [-r', r']e \subseteq [x]^0$ yielding a fixed point $x^* \in [\hat{x}]^0 \subseteq [x]^0$ of t. By the intersection in the definition of $[x]^{k+1}$ the iterates $[x]^k$ decrease monotonically with respect to '$\subseteq$', thus the proof of a) terminates analogously to that of b).

c) Since $\hat{\gamma} \geq \gamma$ we have $\hat{\Delta} \leq \Delta$ and

$$r^- = \frac{2\alpha}{1-\beta+\sqrt{\Delta}} \leq \hat{r}^- \leq \hat{r}^+ \leq r^+ = \frac{2\alpha}{1-\beta-\sqrt{\Delta}} .$$

Therefore, r is contained in $[r^-, r^+]$, and c) is proved by b) with the exception of the uniqueness of x^* and of the degeneracy of $[x]^* = [x^*, x^*]$. In order to show $d([x]^*) = 0$ apply $d(\cdot)$ to the equality $[x]^* = [g]([x]^*, \widetilde{x})$. Then by the subdistributivity of the interval arithmetic and by elementary rules for the diameter (cf. [5] for example) one obtains

$$\begin{aligned}
d([x]^*) &\leq d\left(t'(\widetilde{x})([x]^* - \widetilde{x}) + [H]([x]^*, \widetilde{x})\,([x]^* - \widetilde{x})\right) \\
&= |t'(\widetilde{x})|\, d([x]^*) + d\left([H]([x]^*, \widetilde{x})\,([x]^* - \widetilde{x})\right) \\
&\leq |t'(\widetilde{x})|\, d([x]^*) + d([H]([x]^*, \widetilde{x}))\, |[x]^* - \widetilde{x}| + |[H]([x]^*, \widetilde{x})|\, d([x]^*) \\
&\leq |t'(\widetilde{x})|\, d([x]^*) + d([H]([x]^*, \widetilde{x}))\, |[x]^0 - \widetilde{x}| + |[H]([x]^0, \widetilde{x})|\, d([x]^*)
\end{aligned}$$

Let $d^* := \|d([x]^*)\|$ and apply $\|\cdot\|$ to this inequality in order to get

$$d^* \leq \beta d^* + r\delta d^* + r\gamma d^* \leq \beta d^* + 2r\hat{\gamma} d^*.$$

If $d^* > 0$, we obtain $1 \leq \beta + 2r\hat{\gamma}$ which yields the contradiction

$$r \geq \frac{1-\beta}{2\hat{\gamma}} = \frac{\hat{r}^- + \hat{r}^+}{2}.$$

Therefore, $d^* = 0$, and $x^* \in [x]^*$ implies $[x]^* = [x^*, x^*]$. In particular, this proves uniqueness.

□

Note that if $\widetilde{x}$ is a sufficiently good approximation of a fixed point x^* of t then α will be small and the assumption $\Delta \geq 0$ will certainly be fulfilled provided that $\beta < 1$.

In practical applications, one often chooses

$$[H]([x], \widetilde{x}) := \frac{1}{2}\, t''([x] \underline{\cup} \widetilde{x})([x] - \widetilde{x}) \in \boldsymbol{IR}^{n \times n} \tag{10}$$

where

$$t''(x) : \begin{cases} \boldsymbol{R}^n \times \boldsymbol{R}^n & \to \quad \boldsymbol{R}^n \\ \quad (y, z) & \mapsto \quad t''(x)(y, z) \end{cases} \tag{11}$$

is the second derivative of $t = (t_i)$ at $x \in D$. In (10) we assume, that $t(x)''(y)$ is defined by

$$t''(x)(y) := \begin{pmatrix} y^T t_1''(x) \\ \vdots \\ y^T t_n''(x) \end{pmatrix} \in \boldsymbol{R}^{n \times n} \text{ for } x \in D \text{ and } y \in \boldsymbol{R}^n,$$

and in (11) we define $t(x)''(y,z)$ by

$$t''(x)(y,z) := \left(y^T t_1''(x) z, \ldots, y^T t_n''(x) z\right) \in \boldsymbol{R}^n,$$

with

$$t_i''(x) := \left(\frac{\partial^2 t_i(x)}{\partial x_l\, \partial x_k}\right) = \begin{pmatrix} \dfrac{\partial^2 t_i(x)}{\partial x_1^2} & \dfrac{\partial^2 t_i(x)}{\partial x_2\, \partial x_1} & \cdots & \dfrac{\partial^2 t_i(x)}{\partial x_n\, \partial x_1} \\ \dfrac{\partial^2 t_i(x)}{\partial x_1\, \partial x_2} & \dfrac{\partial^2 t_i(x)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 t_i(x)}{\partial x_n\, \partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial^2 t_i(x)}{\partial x_1\, \partial x_n} & \dfrac{\partial^2 t_i(x)}{\partial x_2\, \partial x_n} & \cdots & \dfrac{\partial^2 t_i(x)}{\partial x_n^2} \end{pmatrix} = (\operatorname{grad} t_i(x))' \in \boldsymbol{R}^{n \times n}$$

being the Hessian associated with $t_i(x)$. Note that k counts the rows while l counts the columns.

The reason behind the choice of $[H]$ according to (10) is the Taylor expansion of t at $\widetilde{x} \in [x]$ which we write in the form

$$t(x) = t(\widetilde{x}) + \{t'(\widetilde{x}) + R(x, \widetilde{x})\}(x - \widetilde{x})$$

with the remainder term $R(x,\widetilde{x}) \cdot (x - \widetilde{x})$, where $R(x, \widetilde{x}) \in \boldsymbol{R}^{n \times n}$. We remind that according to [7], p. 284, and by applying the extended mean value theorem, the entries $r_{ij}(x, \widetilde{x})$ of $R(x, \widetilde{x})$ can be expressed as

$$\begin{aligned} r_{ij}(x,\widetilde{x}) &= (x - \widetilde{x})^T \left(\int_0^1 \frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\widetilde{x} + \tau(x - \widetilde{x}))\,(1-\tau)\, d\tau\right)_{k=1,\ldots,n} \\ &= \int_0^1 (1-\tau)\, d\tau\, (x - \widetilde{x})^T \left(\frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\xi^{(ijk)})\right)_{k=1,\ldots,n} \\ &= \frac{1}{2}\, (x - \widetilde{x})^T \left(\frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\xi^{(ijk)})\right)_{k=1,\ldots,n} \end{aligned}$$

with $\xi^{(ijk)} \in \boldsymbol{R}^n$ between x and $\widetilde{x}$ for $i,\, j,\, k = 1, \ldots n$. Hence

$$R(x, \widetilde{x}) \in \frac{1}{2}\, \left(([x] - \widetilde{x})^T t_i''([x] \underline{\cup} \widetilde{x})\right) = [H]([x], \widetilde{x})$$

and

$$t(x) \in [g]([x], \widetilde{x})$$

for all $x \in [x]$ and $[H]$ from (10). In addition, $[H]([x], \widetilde{x})$ is continuous and inclusion monotone since the function $c([x]) := [x] \underline{\cup} \widetilde{x}$ has these properties and since $[H]([x], \widetilde{x})$ can be interpreted as the interval arithmetic evaluation of the expression $\frac{1}{2}\left((x-\widetilde{x})^T t''(c(x))\right)$; cf.[5] or [13] for details.

Assume now that a function $f : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$ is given and that $[x] \subseteq D$. We are interested in the zeros $x^* \in [x]$ of f. To this end use the transformation

$$t(x) := x - Cf(x), \quad C \in \boldsymbol{R}^{n\times n} \text{ nonsingular and independent of } x. \tag{12}$$

Then the zeros of f are the fixed points of t and vice versa. With $t'(x) = I - Cf'(x)$ and with $[H]$ from (10) we get

$$[g]([x], \widetilde{x}) = \widetilde{x} - Cf(\widetilde{x}) + \left\{ I - Cf'(\widetilde{x}) - \frac{1}{2}(Cf)''([x]\underline{\cup}\widetilde{x})([x]-\widetilde{x}) \right\} ([x]-\widetilde{x}). \tag{13}$$

For $\widetilde{x} \in [x]$ this is just the function k_3 in [5], p. 239, and k_7 in [14], p. 30. Therefore, it is not astonishing that Theorem 1a) and b) reduces to similar results as in [14], § 4, and in [5], § 19. For a comparison take into account the factor $\frac{1}{2}$ in (12). But note that we will also use $[H]$ and $[g]$ in Example 2 in a different meaning as in (10) and (13). This is caused by the possibility of representing the remainder term $R(x, \widetilde{x})(x - \widetilde{x})$ in different ways, as is also shown in the following example.

Example 1 Let $f(x) := \begin{pmatrix} x^2 - 2xy + y^2 \\ 0 \end{pmatrix} \in \boldsymbol{R}^2$, $C := I$, $t(x) := x - Cf(x)$. Then

$$f''(x)(y,z) = \begin{pmatrix} y^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} z \\ 0 \end{pmatrix} = \begin{pmatrix} 2y_1(z_1 - z_2) + 2y_2(-z_1 + z_2) \\ 0 \end{pmatrix}$$

for all x, y, $z \in \boldsymbol{R}^2$, whence

$$f''(x)(y) = \begin{pmatrix} y^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \\ 0 \quad 0 \end{pmatrix} \in \boldsymbol{R}^{2\times 2}. \tag{14}$$

Let

$$A(y) := \begin{pmatrix} y^T \begin{pmatrix} 2 & -4 \\ 0 & 2 \end{pmatrix} \\ 0 \quad 0 \end{pmatrix} \in \boldsymbol{R}^{2\times 2}. \tag{15}$$

Then apparently $f''(x)(y,y) = \begin{pmatrix} 2y_1^2 - 4y_1y_2 + 2y_2^2 \\ 0 \end{pmatrix} = A(y)y$ holds although $f''(x)(y) \neq A(y)$ and therefore $f''(x)(y,z) \neq A(y)z$, in general. So we can represent $t(x)$ by means of (14) as well as by means of (15). Choosing $[H]([x],\tilde{x}) := \frac{1}{2}(Cf)''([x]\underline{\cup}\tilde{x})([x]-\tilde{x})$ and $[H]([x],\tilde{x}) := \frac{1}{2}CA([x]-\tilde{x})$, respectively, yields two different admissible interval functions $[g]$, where the second one differs from (13).

□

If $\tilde{x}$ is a sufficiently good approximation of a zero of f, if $f'(\tilde{x})^{-1}$ exists, and if $C \approx f'(\tilde{x})^{-1}$ then $\alpha \approx 0$, $\beta \approx 0$, $\Delta \approx 1$ and $r^- \approx 0$, $r^+ \approx 1/\gamma$ for the quantities in Theorem 1. In particular, the assumptions $\beta < 1$ and $\Delta \geq 0$ of this theorem are fulfilled.

We want to apply now Theorem 1 with (2) and (12) to various problems of numerical analysis.

Example 2 (The generalized eigenproblem with a simple real eigenvalue)
Consider the generalized algebraic eigenproblem $Av = \lambda Bv$ as in § 1. Let $\tilde{v} \in \boldsymbol{R}^n$ be an approximation of an eigenvector which belongs to an algebraic simple eigenvalue. Let $\tilde{\lambda}$ be an approximation of this eigenvalue and use f from (1), $C \in \boldsymbol{R}^{(n+1)\times(n+1)}$ nonsingular, $\tilde{x} := (\tilde{v}^T, \tilde{\lambda})^T \in \boldsymbol{R}^{n+1}$, and $[v] \in \boldsymbol{IR}^n$. In [15] the interval function

$$[g]([x],\tilde{x}) := \tilde{x} - C\,f(\tilde{x}) + \left\{ I_{n+1} - C \begin{pmatrix} A - \tilde{\lambda}B & -B[v]) \\ (e^{(i_0)})^T & 0 \end{pmatrix} \right\} ([x] - \tilde{x}) \tag{16}$$

was applied in order to verify eigenpairs of the generalized eigenproblem. With $t(x) = x - Cf(x)$ as in (13) one gets

$$t'(\tilde{x}) = I_{n+1} - C \begin{pmatrix} A - \tilde{\lambda}B & -B\tilde{v} \\ (e^{(i_0)})^T & 0 \end{pmatrix}.$$

In [12] it was mentioned that for degenerate interval vectors $[x] \equiv x$ the expression $[g](x,\tilde{x})$ from (16) is the complete Taylor expansion of $t(x)$ at $\tilde{x}$ even if $\tilde{x} \notin [x]$. Therefore, $t(x) \in [g]([x],\tilde{x})$ holds trivially for all $x \in [x]$. Nevertheless $[g]([x],\tilde{x})$ is not in the form (13), i. e., $[H]([x],\tilde{x})$ is not given by (10). In fact, in order to obtain the representation (2) we have to define

$$[H]([x],\tilde{x}) := C \begin{pmatrix} O & B([v] - \tilde{v}) \\ O & 0 \end{pmatrix} \in \boldsymbol{IR}^{(n+1)\times(n+1)}. \tag{17}$$

One can recover this function if one expresses the last, i. e. third, Taylor summand $\frac{1}{2}t''(x)\,(x-\tilde{x}, x-\tilde{x})$ appropriately and if one evaluates this expression interval arithmetically. We want to show that $[H]$ fulfills (4) and (6). From (17) we get

$$|[H]([x], \widetilde{x})| \leq |C| \begin{pmatrix} O & |B|\,|[v] - \widetilde{v}| \\ O & 0 \end{pmatrix}$$

and

$$d([H]([x], \widetilde{x})) = |C| \begin{pmatrix} O & |B| d([v]) \\ O & 0 \end{pmatrix}.$$

With $\gamma := \| \, |C| \begin{pmatrix} |B| \\ 0 \end{pmatrix} \|$ this implies

$$\begin{aligned} \| \, |[H]([x], \widetilde{x})| \, \| &\leq \max_i \sum_{k=1}^{n} |c|_{ik} \sum_{j=1}^{n} |b|_{kj} |[v] - \widetilde{v}|_j \\ &\leq \max_i \sum_{j=1}^{n} \left(\sum_{k=1}^{n} |c|_{ik} |b|_{kj} \right) \| \, |[x] - \widetilde{x}| \, \| \\ &= \gamma \| \, |[x] - \widetilde{x}| \, \| \end{aligned}$$

and, analogously,

$$\| \, d([H]([x], \widetilde{x})) \, \| \leq \gamma \| \, d([x]) \, \|.$$

Therefore, Theorem 1 applies with $\hat{\gamma} := \gamma$.

For practical applications there is also a modification of $[g]$ which one gets by choosing C in the form

$$C = \begin{pmatrix} C_{11} & C_{12} \\ 0 & 1 \\ C_{21} & C_{22} \end{pmatrix} \tag{18}$$

with $C_{11} \in \boldsymbol{R}^{(n-1) \times n}$, $C_{12} \in \boldsymbol{R}^{n-1}$, $C_{21}^T \in \boldsymbol{R}^{n-1}$, $C_{22} \in \boldsymbol{R}$. Theorem 1 then reduces to a result in [4]. For details see [12].

□

Example 3 (The algebraic eigenproblem with a simple real eigenvalue)
Here we start again with f as in (1), where this time we choose $B := I$. The results in Example 2 remain true, of course, and are therefore omitted. They can be found, in [9]–[12] where they have been derived in a different way. For the particular choice of C in (18) they are already contained in [1].

We assumed that the eigenvalue λ^* to be enclosed is an algebraic simple one. This is due to the fact that only in this case $f'(x^*)^{-1}$ exists where $x^* = (v^{*T}, \lambda^*)^T$ is a corresponding eigenpair; cf. Theorem 2 in [12] for details. Thus, for a sufficiently good approximation $\widetilde{x}$ the inverse of $f'(\widetilde{x})$ exists, and $C \approx f'(\widetilde{x})^{-1}$ can be chosen as in the remark preceding the Example 2. □

Example 4 (Two–dimensional invariant subspaces)
In order to enclose double or nearly double eigenvalues, Alefeld and Spreuer verify in [6] a basis of a two–dimensional subspace of $\boldsymbol{R}^n$ which is invariant with respect to the linear mapping given by $A \in \boldsymbol{R}^{n\times n}$. To this end they start with the function

$$f(x) := \begin{pmatrix} Au - m_{11}u - m_{21}v \\ u_{i_1} - \varepsilon \\ u_{i_2} - \zeta \\ Av - m_{12}u - m_{22}v \\ v_{i_1} - \eta \\ v_{i_2} - \theta \end{pmatrix} \in \boldsymbol{R}^{2n+4}$$

where $x = (u^T, m_{11}, m_{21}, v^T, m_{12}, m_{22})^T \in \boldsymbol{R}^{2n+4}$, $i_1 \neq i_2 \in \{1, \dots, n\}$ and $\varepsilon\theta - \zeta\eta \neq 0$. It is obvious that the vectors u^*, v^*, which are part of a zero $x^* = ((u^*)^T, m_{11}^*, m_{21}^*, (v^*)^T, m_{12}^*, m_{22}^*)^T$ of f, form a basis of such an invariant subspace. Note that u^*, v^* are unique within a fixed subspace because of the four normalization conditions which are hidden in f. Again we set $t(x) := x - Cf(x)$ with a nonsingular matrix $C \in \boldsymbol{R}^{(2n+4)\times(2n+4)}$, and we choose $\widetilde{x} = (\widetilde{u}^T, \widetilde{m}_{11}, \widetilde{m}_{21}, \widetilde{v}^T, \widetilde{m}_{12}, \widetilde{m}_{22})^T$ as an approximation of x^*. With

$$B := \begin{pmatrix} A - \widetilde{m}_{11}I_n & -\widetilde{u} & -\widetilde{v} & -\widetilde{m}_{21}I_n & 0 & 0 \\ (e^{(i_1)})^T & 0 & 0 & 0 & 0 & 0 \\ (e^{(i_2)})^T & 0 & 0 & 0 & 0 & 0 \\ -\widetilde{m}_{12}I_n & 0 & 0 & A - \widetilde{m}_{22}I_n & -\widetilde{u} & -\widetilde{v} \\ 0 & 0 & 0 & (e^{(i_1)})^T & 0 & 0 \\ 0 & 0 & 0 & (e^{(i_2)})^T & 0 & 0 \end{pmatrix} \in \boldsymbol{R}^{(2n+4)\times(2n+4)},$$

$$[T] := \begin{pmatrix} O & [u] - \widetilde{u} & [v] - \widetilde{v} & O & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ O & 0 & 0 & O & [u] - \widetilde{u} & [v] - \widetilde{v} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \boldsymbol{IR}^{(2n+4)\times(2n+4)}$$

and

$$[g]([x], \widetilde{x}) := \widetilde{x} - Cf(\widetilde{x}) + \{I_{2n+4} - C(B - [T])\}([x] - \widetilde{x})$$

we again have $t'(\widetilde{x}) = I_{2n+4} - CB$ and $t(x) \in [g]([x], \widetilde{x})$ for all $x, \widetilde{x} \in \boldsymbol{R}^n$. Note that $[g](x, \widetilde{x})$ is again the Taylor expansion of $t(x)$ at $\widetilde{x}$; cf. [12], for example. Therefore, Theorem 1 applies with

$$[H]([x], \widetilde{x}) := C[T]$$

and with

$$\hat{\gamma} := \gamma := \delta := \left\| \, 2|C|(e^T, 0, 0, e^T, 0, 0)^T \, \right\| \leq 2\| \, C \, \|, \quad e := (1, \ldots, 1)^T \in \boldsymbol{R}^n.$$

The results coincide with those in [6] and [12].

□

Example 5 (The singular value problem)

Each rectangular real matrix $A \in \boldsymbol{R}^{m \times n}$ can be represented as

$$A = V \Sigma U^T \iff AU = V\Sigma \iff A^T V = U \Sigma^T$$

with orthogonal matrices $U \in \boldsymbol{R}^{n \times n}$, $V \in \boldsymbol{R}^{m \times m}$ and with a rectangular diagonal matrix $\Sigma \in \boldsymbol{R}^{m \times n}$, where

$$(\Sigma)_{ij} := \begin{cases} 0 & \text{for } i \neq j \\ \sigma_i & \text{for } i = j \end{cases}, \ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_{\min\{m,n\}} = 0.$$

The product $V \Sigma U^T$ is called a *singular value decomposition* of A, the positive values σ_i are called (non–trivial) *singular values* of A. Note that Σ is unique while U and V are not. The i–th singular value and the i–th columns u^i, v^i of U and V – so–called *singular vectors* – can be expressed as the zeros of the function

$$f(x) := \begin{pmatrix} Au - \sigma v \\ A^T v - \sigma u \\ u^T u - 1 \end{pmatrix},$$

where $x := (u^T, v^T, \sigma)^T$. If $x^* = ((u^*)^T, (v^*)^T, \sigma^*)^T$ is a zero of f with $\sigma^* \neq 0$ then

$$(v^*)^T v^* = (v^*)^T \frac{1}{\sigma^*} A u^* = \frac{1}{\sigma^*} \left(A^T v^* \right)^T u^* = (u^*)^T u^* = 1.$$

Let $\widetilde{x} = (\widetilde{u}^T, \widetilde{v}^T, \widetilde{\sigma})^T$. In [12] a slight modification of the interval function

$$[g]([x], \widetilde{x}) := \widetilde{x} - Cf(\widetilde{x}) + \left\{ I_{m+n+1} - C \left(B - \begin{pmatrix} O & O & [v] - \widetilde{v} \\ O & O & [u] - \widetilde{u} \\ ([u] - \widetilde{u})^T & 0 & 0 \end{pmatrix} \right) \right\} ([x] - \widetilde{x})$$

was used with

$$B := \begin{pmatrix} A & -\widetilde{\sigma} I_m & -\widetilde{v} \\ -\widetilde{\sigma} I_n & A^T & -\widetilde{u} \\ 2\widetilde{u}^T & 0 & 0 \end{pmatrix}$$

in order to verify singular values σ and corresponding singular vectors u, v of A. It is an easy task to prove that $[g](x, \widetilde{x})$ is again the complete Taylor expansion of $t(x) := x - Cf(x)$ at $x = \widetilde{x}$ with $t'(\widetilde{x}) = I - CB$. As in Example 2 one easily checks that (4) and (6) hold for

$$[H]([x], \widetilde{x}) := C \begin{pmatrix} O & O & [v] - \widetilde{v} \\ O & O & [u] - \widetilde{u} \\ ([u] - \widetilde{u})^T & 0 & 0 \end{pmatrix}.$$

with

$$\gamma := \delta := \| \, |C| \cdot (1, \ldots, 1, n)^T \, \|.$$

Therefore, Theorem 1 applies with $\hat{\gamma} = \gamma$. Cf. also [2].

□

Example 6 (Quadratic systems)

In this example we consider systems of equations of the form

$$t(x) := b + Ax + T(x, x) = x$$

with $b,\ x \in \boldsymbol{R}^n,\ A \in \boldsymbol{R}^{n \times n}$ and with

$$T(x, y) := \left(\sum_{j=1}^{n} \sum_{k=1}^{n} t_{ijk} x_k y_j \right)_{i=1,\ldots,n}.$$

Note that $t(0) = b,\ t'(0) = A$ and $t''(x)(y, z) = 2T(y, z)$. Let

$$\begin{aligned} [H]([x], 0) &:= T([x]) := \left(\sum_{k=1}^{n} t_{ijk} [x]_k \right) \in \boldsymbol{IR}^{n \times n}, \\ [g]([x], 0) &:= b + (A + T([x]))[x]. \end{aligned}$$

In particular, $\widetilde{x} = 0$ in this example. One easily sees that (4) and (6) hold with

$$\gamma := \delta := \| \left(\sum_{k=1}^{n} |t_{ijk}| \right) \| = \max_{1 \le i \le n} \left\{ \sum_{j=1}^{n} \sum_{k=1}^{n} |t_{ijk}| \right\} \tag{19}$$

Therefore, Theorem 1 applies with $\alpha := \|b\|$, $\beta := \|A\|$, with γ, δ as in (19) and with $\hat{\gamma} = \gamma$. Its results coincide with those in [3] and [12].

□

References

[1] **Alefeld, G. :** *Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts.* Z. Angew. Math. Mech. **67**, 145–152 (1987)

[2] **Alefeld, G. :** *Rigorous error bounds for singular values of a matrix using the precise scalar product.* In: Kaucher, E., Kulisch, U., Ullrich, Ch. (eds.): *Computerarithmetic.* pp. 9–130. Stuttgart 1987

[3] **Alefeld, G. :** *Errorbounds for quadratic systems of nonlinear equations using the precise scalar product.* In: Kulisch, U., Stetter, H. J. (eds.): *Scientific Computation with Automatic Result Verification.* Computing, Suppl. **6**, 59–68 (1988)

[4] **Alefeld, G. :** *Berechenbare Fehlerschranken für ein Eigenpaar beim verallgemeinerten Eigenwertproblem.* Z. Angew. Math. Mech. **68**, 181–184 (1988)

[5] **Alefeld, G.** and **Herzberger, J. :** *Introduction to Interval Computations.* New York 1983

[6] **Alefeld, G.** and **Spreuer, H. :** *Iterative improvement of componentwise errorbounds for invariant subspaces belonging to a double or nearly double eigenvalue.* Computing **36**, 321–334 (1986)

[7] **Heuser, H. :** *Lehrbuch der Analysis. Teil 2.* Stuttgart 1981

[8] **Hoffmann, R. :** *Konstruktion von Fehlerschranken bei der verallgemeinerten Singulärwertzerlegung und ihre iterative Verbesserung.* Thesis, Universität Karlsruhe 1993

[9] **Mayer, G. :** *Enclosures for eigenvalues and eigenvectors.* In: Atanassova, L., Herzberger, J. (eds.): *Computer Arithmetic and Enclosure Methods.* pp. 49–68. Amsterdam 1992

[10] **Mayer, G. :** *Taylor-Verfahren für das algebraische Eigenwertproblem.* Z. Angew. Math. Mech. **73**, T857–T860 (1993)

[11] **Mayer, G. :** *A unified approach to enclosure methods for eigenpairs.* Z. Angew. Math. Mech. **74**, 115–128 (1994)

[12] **Mayer, G. :** *Result verification for eigenvectors and eigenvalues.* In: Herzberger, J. (ed.): *Topics in Validated Computations.* Studies in Computational Mathematics **5**, pp. 209–276. Amsterdam 1994

[13] **Neumaier, A.** : *Interval Methods for Systems of Equations.* Cambridge 1990

[14] **Platzöder, L.** : *Einige Beiträge über die Existenz von Lösungen nichtlinearer Gleichungssysteme und Verfahren zu ihrer Berechnung.* Thesis, Berlin 1981

[15] **Rump, S. M.** : *Guaranteed inclusions for the complex generalized eigenproblem.* Computing **42**, 225–238 (1989)

received: September 11, 1995

Author:

Prof. Dr. G. Mayer
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany
e-mail: guenter.mayer@mathematik.uni-rostock.de