Applications of Mathematics

Günter Mayer

Epsilon-inflation with contractive interval functions

Applications of Mathematics, Vol. 43 (1998), No. 4, 241–254

Persistent URL: http://dml.cz/dmlcz/134388

EPSILON-INFLATION WITH CONTRACTIVE INTERVAL FUNCTIONS

Günter Mayer, Rostock

Dedicated to Prof. Dr. Gerhard Heindl on the occasion of his 60th birthday

(Received November 15, 1996)

Abstract. For contractive interval functions $[g]$ we show that $[g]([x]_\varepsilon^{k_0}) \subseteq \operatorname{int}([x]_\varepsilon^{k_0})$ results from the iterative process $[x]^{k+1} := [g]([x]_\varepsilon^k)$ after finitely many iterations if one uses the epsilon-inflated vector $[x]_\varepsilon^k$ as input for $[g]$ instead of the original output vector $[x]^k$. Applying Brouwer's fixed point theorem, zeros of various mathematical problems can be verified in this way.

Keywords: epsilon-inflation, P-contraction, contraction, verification algorithms, interval computation, nonlinear equations, eigenvalues, singular values

MSC 2000: 65F05, 65F10, 65F15, 65G05, 65G10, 65H10, 65H15, 65L05

1. Introduction

If G denotes a nonempty convex, compact subset of $\mathbb{R}^n$ and if t is a continuous self-mapping of G then Brouwer's fixed point theorem guarantees that t has at least one fixed point in G. Often G is an interval vector and t is a function which is defined and continuous in an open superset D of G. Assume that with t an interval function $[g]$ is associated such that the *inclusion property*

$$t(x) \in [g]([x]) \tag{1}$$

holds for all $x \in [x]$ and for all $[x] \subseteq D$. If

$$[g]([x]) \subseteq [x] \quad \text{(or, more strongly, } [g]([x]) \subseteq \operatorname{int}([x])\text{)} \tag{2}$$

is valid for some interval vector $[x] \subseteq D$ then t has a fixed point x^* in $[x]$ by the above mentioned Brouwer's fixed point theorem, since (1) and (2) guarantee the self-mapping property of t.

A simple choice of $[g]$ is the interval arithmetic evaluation of t (cf. [2]) which guarantees (1). But often $[g]$ is chosen in a more sophisticated way. In order to find a vector $[x]$ which satisfies (2) one usually starts with an approximation $\tilde{x}$ of a fixed point x^* of t and one iterates by

$$(3) \qquad [x]^0 := [\tilde{x}, \tilde{x}], \quad [x]^{k+1} := [g]([x]^k_\varepsilon), \quad k = 0, 1, \ldots$$

until (2) holds for some $[x] = [x]^k_\varepsilon$ with $k \leqslant k_{\max}$. Here $k_{\max}$ is a given bound for the number of iterates and $[x]^k_\varepsilon$ is any interval vector which contains $[x]^k$ in its interior. Usually, $[x]^k_\varepsilon$ is called the *ε-inflation* of $[x]^k_\varepsilon$. This name stems from the fact that the construction of $[x]^k_\varepsilon$ normally depends on a parameter $\varepsilon > 0$. A simple example is $[x]_\varepsilon := [x] + \varepsilon[-1, 1](1, \ldots, 1)^T$, further possibilities can be found e.g. in [9]. The iteration (3) does not always end up with (2) as the example $[g]([x]) := 2[x]$, $\tilde{x} := 1$ shows for an arbitrary ε-inflation. But often it helps as in the case $g([x]) := \frac{1}{2}[x]$, $\tilde{x} := 1$ if one chooses the ε-inflation from above with $\varepsilon := 0.1$ whence $[x]^4 \subseteq [x]^3_\varepsilon$.

It is an open question in which situations (3) ends up with (2) for some $[x] = [x]^k_\varepsilon$ in at most $k_{\max}$ steps. For contractive interval functions $[g]$, in particular for functions $[g]$ of the form

$$(4) \qquad [g]([x]) := t(\tilde{x}) + \{t'(\tilde{x}) + [H]([x])\}([x] - \tilde{x}),$$

we will at least be able to show that (3) results in (2) after *finitely* many steps of iterations. In (4) the vector $\tilde{x}$ is a fixed vector from D; $[H]$ is an interval matrix function for which we require the Lipschitz condition

$$(5) \qquad \|q\left([H]([x]), [H]([y])\right)\| \leqslant \kappa \|q([x], [y])\|$$

and the value

$$(6) \qquad [H](\tilde{x}) = O;$$

q denotes the Hausdorff distance; κ is a positive constant which is independent of $[x]$ but which may depend on $\tilde{x}$; $\|\cdot\|$ denotes any monotone vector norm and the corresponding operator norm for matrices, respectively. Functions $[g]$ as in (4) occur, when involving second derivatives in order to compute zeros of a function f; in particular, they arise when verifying eigenpairs, singular values, and solutions of quadratic systems (cf. Section 4). For example, when verifying and enclosing zeros of functions $f\colon\ D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, $f = (f_i) \in C^2(D)$, one often transforms the problem $f(x) = 0$ into the fixed point problem

$$x = t(x) := x - Cf(x), \quad C \in \mathbb{R}^{n\times n} \text{ nonsingular.} \tag{7}$$

The interval function $[g]$ from (4) reads then

$$[g]([x]) := \tilde{x} - Cf(\tilde{x}) + \{I - Cf'(\tilde{x}) + [H]([x])\}([x] - \tilde{x}) \tag{8}$$

with $[H](x) := f''([x]\underline{\cup}\tilde{x})([x] - \tilde{x})$, for example, where $f''(x)y$ is defined by

$$f''(x)y := \left(y^T\left(\frac{\partial^2 f_i(x)}{\partial x_l \partial x_k}\right)\right)_{i=1,\ldots,n} \in \mathbb{R}^{n\times n}$$

with the Hessian $\left(\frac{\partial^2 f_i(x)}{\partial x_l \partial x_k}\right) \in \mathbb{R}^{n\times n}$ of f_i and with the convex hull $[x]\underline{\cup}\tilde{x}$ of $[x]$ and $\tilde{x}$.

The technique and the name ε-inflation have been introduced in [13]. Remarks concerning its practical applicability can be found e.g. in [5] and [6]. Theoretical considerations have been done in [8], [9], [11], [15] and [16]. The idea of replacing a starting interval $[x]^0$ by another one with a larger diameter, say $[\hat{x}]^0$, was already used in [4]. But $[\hat{x}]^0 \supseteq [x]^0$ was not required there. Our paper generalizes the results of [8], [9] and [11] where P-contractivity was assumed. Note that each P-contraction is a contraction but not vice versa; see [9] for a counterexample. Our present paper deals with *contractive* functions; it uses an access which is different from that in [10], where quantitative aspects played the crucial role.

2. Preliminaries

By $\mathbb{I}\mathbb{R}$, $\mathbb{I}\mathbb{R}^n$, $\mathbb{I}\mathbb{R}^{n\times n}$ we denote the set of intervals, the set of interval vectors with n components and the set of $n \times n$ interval matrices, respectively. By 'interval' we always mean a real compact interval. We write interval quantities in brackets with the exception of degenerate interval quantities which we identify with the element which they contain. Examples are the identity matrix I, its i-th column $e^{(i)}$ and the vector $e = (1, 1, \ldots, 1)^T$. With $[z] \in \mathbb{I}\mathbb{R}^n$ we define the subset $I([z]) := \{[x] \mid [x] \subseteq [z]\}$ of $\mathbb{I}\mathbb{R}^n$. We apply the notation $[x] = ([x]_i) = [\underline{x}, \overline{x}] = ([\underline{x}_i, \overline{x}_i]) \in \mathbb{I}\mathbb{R}^n$ simultaneously without further reference, and we proceed similarly with the elements of $\mathbb{I}\mathbb{R}$ and $\mathbb{I}\mathbb{R}^{n\times n}$. By $\mathrm{int}([a])$ we denote the topological *interior* of an interval $[a]$ and by $\check{a}$ we mean its *midpoint*. We define the *absolute value* $|[a]|$ by $|[a]| := \max\{|\underline{a}|, |\overline{a}|\}$, the *diameter* $d([a])$ by $d([a]) := \overline{a} - \underline{a}$ and the *distance* $q([a], [b])$ by $q([a], [b]) :=$

$\max\{|\underline{a}-\underline{b}|, |\overline{a}-\overline{b}|\}$. For interval vectors and interval matrices these items are applied entrywise. Continuity in $\mathbb{IR}$, $\mathbb{IR}^n$ and $\mathbb{IR}^{n\times n}$ is to be understood with respect to q. If $g(x)$ is an expression for some function g, we write $g([x])$ for the interval arithmetic evaluation of this expression (cf. [2]), assuming that $g([x])$ exists. Note that we distinguish between $g([x])$ and $[g]([x])$, where $[g]$ means any interval function. For details on interval arithmetic we refer to [2] or [12].

By $\varrho(A)$ we denote the spectral radius of $A \in \mathbb{R}^{n\times n}$; $A \geqslant 0$ means $a_{ij} \geqslant 0$ for $i, j = 1, \ldots, n$, and $x > 0$ is used for $x \in \mathbb{R}^n$ if $x_i > 0$, $i = 1, \ldots, n$.

As in [2], we define $[g]\colon \mathbb{IR}^n \to \mathbb{IR}^n$ to be a P-*contraction* if there is a matrix $P \in \mathbb{R}^{n\times n}$ with $P \geqslant 0$, $\varrho(P) < 1$ such that

$$q([g]([x]), [g]([y])) \leqslant Pq([x], [y]) \tag{9}$$

for all $[x], [y] \in \mathbb{IR}^n$. If $[g]$ fulfils (9) only for all $[x]$, $[y] \subseteq [z]$ with a given $[z] \in \mathbb{IR}^n$, we call $[g]$ a P-*contraction on* $[z]$. Similarly, we define $[g]\colon \mathbb{IR}^n \to \mathbb{IR}^n$ to be a *contraction* (*with respect to some vector norm* $\|\cdot\|$) if there is a real constant $\alpha \in (0,1)$ such that

$$\|q\left([g]([x]), [g]([y])\right)\| \leqslant \alpha\|q([x], [y])\| \tag{10}$$

holds for all $[x], [y] \in \mathbb{IR}^n$. If $[g]$ fulfils (10) only on $I([z])$ for a given $[z] \in \mathbb{IR}^n$, we call $[g]$ a *contraction on* $[z]$ (*with respect to some vector norm* $\|\cdot\|$).

A vector norm $\|\cdot\|$ on $\mathbb{R}^n$ is termed *monotone* if $|x| \leqslant |y|$ implies $\|x\| \leqslant \|y\|$ for all $x, y \in \mathbb{R}^n$.

If the same symbol $\|\cdot\|$ is used for vectors and matrices then we always assume that the matrix norm is the operator norm generated by the vector norm $\|\cdot\|$. Throughout our paper, $\|\cdot\|_\infty$ denotes the maximum norm when applied to vectors, and the row sum norm when applied to matrices; μ, ν denote positive constants such that

$$\mu\|x\|_\infty \leqslant \|x\| \leqslant \nu\|x\|_\infty. \tag{11}$$

3. Results

We start our results with a theorem which is well-known for P-contractions (cf. [2] and [8], [9]) and which we formulate now for contractive mappings. In Theorems 3.1–3.4 the function $[g]$ need not necessarily be defined by (4).

Theorem 3.1. *Let* $[g]\colon \mathbb{IR}^n \to \mathbb{IR}^n$ *be a contraction with respect to a monotone norm* $\|\cdot\|$. *Then each sequence of iterates* $[x]^{k+1} := [g]([x]^k)$, $k = 0, 1, \ldots$ *converges to the same limit* $[x]^*$, *which is the unique fixed point of* $[g]$.

If

$$[g](x) \in \mathbb{R}^n \tag{12}$$

holds for all $x \in \mathbb{R}^n$, then $[x]^*$ is a degenerate interval vector.

If a function $t\colon \mathbb{R}^n \to \mathbb{R}^n$ satisfies the inclusion property (1) *for all $x \in [x]$ and all $[x] \in \mathbb{IR}^n$, then $[x]^*$ contains all the fixed points of t. If, in addition, t is continuous, then it has at least one fixed point in $[x]^*$.*

If (12) *and* (1) *hold, then t is a contraction. It has a unique fixed point which can be identified with $[x]^*$.*

The assertions hold analogously if $\mathbb{R}^n$ is replaced by $[z]$ and if $\mathbb{IR}^n$ is replaced by $I([z])$ for a fixed vector $[z] \in \mathbb{IR}^n$.

Proof. Since $(\mathbb{IR}^n, \|q(\cdot,\cdot)\|)$ is a complete metric space, the existence and uniqueness of $[x]^*$ follow from Banach's fixed point theorem.

Assume now that (1) holds and that $[x]^*$ does not contain some fixed point y^* of t. Start the iterative process $[x]^{k+1} := [g]([x]^k)$ with $[x]^0 := y^*$. Then $y^* = t(y^*) \in [g](y^*) = [g]([x]^0) = [x]^1$ and, by induction, $y^* \in [x]^k$, $k = 0, 1, \ldots$. Therefore, $y^* \in \lim\limits_{k\to\infty} [x]^k = [x]^*$, which contradicts our assumption. Hence $[x]^*$ contains all fixed points of t. Since $t(x) \in [g]([x]^*) = [x]^*$ for all $x \in [x]^*$, Brouwer's fixed point theorem guarantees at least one fixed point of t in $[x]^*$, provided that t is continuous.

Let now (12) and (1) hold simultaneously. Then, clearly, $[g](x) = t(x)$ for all $x \in \mathbb{R}^n$, and the contractivity of $[g]$ and the monotonicity of $\|\cdot\|$ imply

$$\begin{aligned}\|t(x) - t(y)\| &= \big\||t(x) - t(y)|\big\| = \|q(t(x), t(y))\| = \|q([g](x), [g](y))\| \\ &\leqslant \alpha\|q(x, y)\| = \alpha\big\||x - y|\big\| = \alpha\|x - y\|,\end{aligned}$$

where α is the contraction constant of $[g]$. Hence t is a contraction. □

Theorem 3.2. *Let $[z]^c \in \mathbb{IR}^n$ be a fixed vector and let $[g]\colon I([z]^c) \to \mathbb{IR}^n$ be a contraction on $[z]^c$ with respect to a monotone vector norm $\|\cdot\|$. Let $[z]$ be a vector such that $[z]^c \supseteq [z] + \frac{\|d([z])\|}{\mu(1-\alpha)}[-1, 1]e$, where α is the contraction constant and where μ is from (11). Choose $[x]^0 \subseteq [z]$ and assume $[x]^1 := [g]([x]^0) \subseteq [z]$. Then the iterates $[x]^{k+1} := [g]([x]^k)$ are defined for $k = 0, 1, \ldots$, i.e., they are all contained in $[z]^c$. They converge to a vector $[x]^* \subseteq [z]^c$ which is independent of $[x]^0$.*

P r o o f. Since $\|\cdot\|$ is a monotone norm we get

$$\begin{aligned}
\mu\|q([x]^{k+1},[x]^0)\|_\infty &\leqslant \|q([x]^{k+1},[x]^0)\| \leqslant \Big\|\sum_{i=0}^{k} q([x]^{i+1},[x]^i)\Big\| \\
&\leqslant \sum_{i=1}^{k} \|q([g]([x]^i),[g]([x]^{i-1}))\| + \|q([x]^1,[x]^0)\| \\
&\leqslant \alpha \sum_{i=1}^{k} \|q([x]^i,[x]^{i-1})\| + \|q([x]^1,[x]^0)\| \leqslant \ldots \leqslant \Big(\sum_{i=0}^{\infty}\alpha^i\Big)\|q([x]^1,[x]^0)\| \\
&= \frac{1}{1-\alpha}\|q([x]^1,[x]^0)\| \leqslant \frac{1}{1-\alpha}\|\overline{z}-\underline{z}\| = \frac{1}{1-\alpha}\|d([z])\|.
\end{aligned}$$

Therefore,

$$[x]^{k+1} \subseteq [x]^0 + \frac{\|d([z])\|}{\mu(1-\alpha)}[-1,1]e \subseteq [z]^c, \tag{13}$$

in particular, $[x]^k$ exists for all $k \in \mathbb{N}$. Since

$$\begin{aligned}
\mu|\underline{x}_i^{k+m} - \underline{x}_i^m| &\leqslant \mu\|q([x]^{k+m},[x]^k)\|_\infty \leqslant \|q([g]([x]^{k-1+m}),[g]([x]^{k-1}))\| \\
&\leqslant \alpha\|q([x]^{k-1+m},[x]^{k-1})\| \leqslant \ldots \leqslant \alpha^k\|q([x]^m,[x]^0)\| \leqslant \frac{\alpha^k}{1-\alpha}\|d([z])\|
\end{aligned}$$

for all $m = 0,1,\ldots$, and since an analogous inequality holds for the upper bounds, the sequences $\{\underline{x}^k\}$, $\{\overline{x}^k\}$ converge to limits $\underline{x}^*$ and $\overline{x}^*$, respectively, with $\underline{x}^* \leqslant \overline{x}^*$. Therefore, $\lim\limits_{k\to\infty}[x]^k = [\underline{x}^*,\overline{x}^*] =: [x]^*$ with $[x]^* \subseteq [z]^c$ by (13). Uniqueness follows from $\|q([x]^*,[y]^*)\| = \|q([g]([x]^*),[g]([y]^*))\| \leqslant \alpha\|q([x]^*,[y]^*)\|$ for two different fixed points $[x]^*$, $[y]^*$ of $[g]$. $\square$

Theorem 3.3. *Let $[g]\colon \mathbb{IR}^n \to \mathbb{IR}^n$ be a contraction with respect to a monotone norm $\|\cdot\|$ and with a contraction constant α. Iterate by inflation according to*

$$\left\{\begin{aligned}
[x]^0 &:= \tilde{x}, \\
[x]^k_\varepsilon &:= [x]^k + [\delta]^k \\
[x]^{k+1} &:= [g]([x]^k_\varepsilon)
\end{aligned}\right\}\quad k = 0,1,\ldots, \tag{14}$$

where $[\delta]^k \in \mathbb{IR}^n$ are given vectors which converge to some limit $[\delta]$. If $[\delta]$ contains 0 in its interior then there is an integer $k_0 = k_0([x]^0_\varepsilon)$ such that

$$[g]([x]^{k_0}_\varepsilon) \subseteq \operatorname{int}([x]^{k_0}_\varepsilon) \tag{15}$$

holds.

Proof. Let $[s]([x]) := [g]([x]) + [\delta]$. Then

$$\|q([s]([x]),[s]([y]))\| = \|q([g]([x]),[g]([y]))\| \leqslant \alpha\|q([x],[y])\|, \tag{16}$$

hence $[s]$ is a contraction. By Theorem 3.1 it has a unique fixed point $[x]^*$ which satisfies

$$[x]^* = [g]([x]^*) + [\delta]. \tag{17}$$

Assume for the moment that

$$\lim_{k\to\infty} [x]^k_\varepsilon = [x]^* \tag{18}$$

holds for the sequence in (14). By the continuity of $[g]$ we have

$$\lim_{k\to\infty} [g]([x]^k_\varepsilon) = [g]([x]^*)\ . \tag{19}$$

Since $0 \in \operatorname{int}([\delta])$, equation (17) implies $[g]([x]^*) \subseteq \operatorname{int}([x]^*)$. Together with (18) and (19) this yields (15) for all sufficiently large integers k_0.

We prove now the assumption (18). With the usual rules for q we obtain

$$\begin{aligned}\|q([x]^k_\varepsilon,[x]^*)\| &= \|q([g]([x]^{k-1}_\varepsilon) + [\delta]^k, [g]([x]^*) + [\delta])\| \\ &\leqslant \alpha\|q([x]^{k-1}_\varepsilon,[x]^*)\| + \|q([\delta]^k,[\delta])\| \\ &\leqslant \alpha^2\|q([x]^{k-2}_\varepsilon,[x]^*)\| + \alpha\|q([\delta]^{k-1},[\delta])\| + \|q([\delta]^k,[\delta])\| \\ &\leqslant \ldots \leqslant\ \alpha^k\|q([x]^0_\varepsilon,[x]^*)\| + \sum_{i=0}^{k-1}\alpha^i\|q([\delta]^{k-i},[\delta])\|.\end{aligned} \tag{20}$$

Fix $\theta > 0$ and choose the integer m such that $\alpha^i \leqslant \theta$ for all $i \geqslant m$. Since $\lim_{k\to\infty}[\delta]^k = [\delta]$, there are a constant $\gamma > 0$ and an integer $k' > m$ with $\|q([\delta]^i,[\delta])\| \leqslant \gamma,\ i = 0,1,\ldots$, and $\|q([\delta]^{k-i},[\delta])\| \leqslant \theta,\ k \geqslant k',\ i = 0,1,\ldots,m-1$. For $k \geqslant k'$ we thus get with (20)

$$\begin{aligned}\|q([x]^k_\varepsilon,[x]^*)\| &\leqslant \theta\|q([x]^0_\varepsilon,[x]^*)\| + \sum_{i=0}^{m-1}\alpha^i\theta + \alpha^m\sum_{i=m}^{k-1}\alpha^{i-m}\gamma \\ &\leqslant \theta\Big\{\|q([x]^0_\varepsilon,[x]^*)\| + \frac{1}{1-\alpha} + \frac{\gamma}{1-\alpha}\Big\}.\end{aligned}$$

Since the expression in braces is independent of θ, m and k, and since θ can be chosen arbitrarily small, (18) holds. $\square$

Relying on Theorem 3.2 one can also formulate a local version of Theorem 3.3. For simplicity, we restrict ourselves to the case $[\delta]^k = [\delta]$, $k = 0, 1, \ldots$.

Theorem 3.4. *Let $[z]^0 \in \mathbb{IR}^n$ be a fixed vector and let $[g]\colon I([z]^0) \to \mathbb{IR}^n$. Assume that $[z]$, $[z]^c \subseteq [z]^0$ and $[\delta] \in \mathbb{IR}^n$ possess the following properties:*

i) $0 \in \operatorname{int}([\delta])$,

ii) *$[g]$ is contractive with respect to a monotone norm $\|\cdot\|$ on*

$$[z]^c \supseteq [z] + \frac{\|d([z])\|}{\mu(1-\alpha)}[-1,1]e,$$

where α is the contraction constant and μ is the constant from (11). *If $[x]_\varepsilon^0 \subseteq [z]$ and $[x]_\varepsilon^1 \subseteq [z]$ hold for the iterates from* (14) *with $[\delta]^k := [\delta]$, then there is an integer $k_0 = k_0([x]_\varepsilon^0)$ such that* (15) *is true. In particular, t from* (1) *has a fixed point in $[x]_\varepsilon^{k_0}$.*

P r o o f. Since $[s]([x]) := [g]([x]) + [\delta]$ fulfils (16) for all $[x], [y] \subseteq [z]^c$, the function $[s]$ is a contraction on $[z]^c$. By Theorem 3.2 there is a vector $[x]^* \subseteq [z]^c$ which satisfies

$$\lim_{k\to\infty} [x]_\varepsilon^k = [x]^* = [s]([x]^*) = [g]([x]^*) + [\delta]. \tag{21}$$

Since $0 \in \operatorname{int}([\delta])$, this yields

$$[g]([x]^*) \subseteq \operatorname{int}([x]^*), \tag{22}$$

and the assertion follows from (19), (22) and from the first equality in (21). $\square$

We want to apply now Theorem 3.4 to the function $[g]$ from (4) when $[H]$ satisfies (5) and (6) with $\|\cdot\| := \|\cdot\|_\infty$. (The choice of the maximum norm is not a severe restriction since by the norm equivalence in $\mathbb{R}^n$ the norm in (5) can be replaced by any norm, if the constant κ is changed appropriately.) To this end let $[z]^0 \in \mathbb{IR}^n$ denote a fixed interval vector for which $[g]$ is defined and which contains $\tilde{x}$ in its interior. Following the lines in [11], p. 101, one can show that $[g]$ satisfies the Lipschitz condition

$$\|q\left([g]([x]),[g]([y])\right)\|_\infty \leqslant \beta\|q([x],[y])\|_\infty, \quad [x],[y] \subseteq [z]$$

for each fixed $[z] \subseteq [z]^0$ with

$$\beta := \big\||t'(\tilde{x})|\big\|_\infty + 2\kappa\big\||[z]-\tilde{x}|\big\|_\infty.$$

(This even holds for any monotone norm.)

For the remaining part of this section we assume that $\|t'(\tilde{x})\|_\infty$ is sufficiently small, $\tilde{x}$ is a sufficiently good approximation of a fixed point x^* of t, $[\delta] \in \mathbb{I}\mathbb{R}^n$ is a given vector of sufficiently small diameter which contains 0 in its interior, and $[x]^k$, $k = 0, 1, \ldots$, is defined by (14) with $[\delta]^k := [\delta]$.

Then $[g]$ is a contraction on

$$[z] := \tilde{x} + [\delta][-1,1] + \left\{\|\tilde{x} - t(\tilde{x})\|_\infty + \left(\|t'(\tilde{x})\|_\infty + \kappa\big\||[\delta]|\big\|_\infty\right)\big\||[\delta]|\big\|_\infty\right\}[-1,1]e,$$

and $[x]^0_\varepsilon \subseteq [z]$. From

$$\begin{aligned}\big\||[H]([x])|\big\| &= \big\||[H]([x]) - [H](\tilde{x})|\big\| = \|q([H]([x]), [H](\tilde{x}))\| \\ &\leqslant \kappa\|q([x], \tilde{x})\| = \kappa\big\||[x] - \tilde{x}|\big\|.\end{aligned}$$

we get

$$\begin{aligned}[x]^1 &:= [g]([x]^0_\varepsilon) = t(\tilde{x}) + \{t'(\tilde{x}) + [H](\tilde{x} + [\delta])\}\,[\delta] \\ &\subseteq \tilde{x} + (t(\tilde{x}) - \tilde{x}) + \{|t'(\tilde{x})| + |[H](\tilde{x} + [\delta])|\}\,|[\delta]|[-1,1]e \\ &\subseteq \tilde{x} + \|t(\tilde{x}) - \tilde{x}\|_\infty[-1,1]e + \left\{\|t'(\tilde{x})\|_\infty + \big\||[H](\tilde{x} + [\delta])|\big\|_\infty\right\}\big\||[\delta]|\big\|_\infty[-1,1]e \\ &\subseteq \tilde{x} + \|t(\tilde{x}) - \tilde{x}\|_\infty[-1,1]e + \left\{\|t'(\tilde{x})\|_\infty + \kappa\big\||[\delta]|\big\|_\infty\right\}\big\||[\delta]|\big\|_\infty[-1,1]e.\end{aligned}$$

Hence $[x]^1$ and $[x]^1_\varepsilon$ are also contained in $[z]$. By our assumptions we can assume that $\beta < 0.1$ and that $\|d([z])\|_\infty < \frac{0.1}{4\kappa}$. Let $\alpha := \frac{1}{2}$. By virtue of $[z]^c := [z] + \frac{\|d([z])\|_\infty}{1-\alpha}[-1,1]e = [z] + 2\big\||d([z])|\big\|_\infty[-1,1]e$ we obtain $\big\||[z]^c - \tilde{x}|\big\|_\infty \leqslant \big\||[z] - \tilde{x}|\big\|_\infty + 2\|d([z])\|_\infty$. Hence

$$\begin{aligned}\tilde{\beta} &:= \|t'(\tilde{x})\|_\infty + 2\kappa\big\||[z]^c - \tilde{x}|\big\|_\infty \leqslant \|t'(\tilde{x})\|_\infty + 2\kappa\big\||[z] - \tilde{x}|\big\|_\infty + 4\kappa\|d([z])\|_\infty \\ &= \beta + 4\kappa\|d([z])\|_\infty \leqslant 0.1 + 0.1 \leqslant 0.5 = \alpha,\end{aligned}$$

and $[g]$ is a contraction on $[z]^c$ with contraction constant $\tilde{\beta}$ and therefore also with the contraction constant α. Now Theorem 3.4 applies with $\mu = 1$.

In order to use this result for the particular situations of Section 4 we assume now that t is given by (7) with $[g]$ from (8). If C from (7) approximates $f'(\tilde{x})^{-1}$ sufficiently well then $\|t'(\tilde{x})\|_\infty = \|I - Cf'(\tilde{x})\|_\infty$ is certainly small. If, in addition, $\tilde{x}$ is a sufficiently good approximation of a zero of f then $t(\tilde{x}) \approx \tilde{x}$. Hence the 'essential' assumptions above are fulfilled and Theorem 3.4 can be applied. We state this result as a separate corollary:

Corollary 3.1. *Let* $[g]$ *be defined as in* (4) *with* $t(x) := x - Cf(x)$ *and with* $[H]$ *satisfying* (5) *and* (6) *with respect to* $\|\cdot\|_\infty$. *Assume that* $f'(\tilde{x})^{-1}$ *exists and that*

C is nonsingular and approximates $f'(\tilde{x})^{-1}$ sufficiently well. If $\tilde{x}$ is a sufficiently good approximation of a zero x^* of f and if the inflation $[\delta]$ is sufficiently small and contains 0 in its interior, then the inflation procedure (14) with $[\delta]^k := [\delta]$ stops with $[x]^{k+1} \subseteq \operatorname{int}([x]^k_\varepsilon)$ after finitely many steps.

Note that Corollary 3.1 guarantees success in ε-inflation only if some input parameters are sufficiently good. Unfortunately it neither predicts the minimal number k_0 of iterates which are necessary to fulfill (2), nor specifies by a measure what 'sufficiently' really means. In this respect further work has to be done.

If one computes C as an approximate inverse of $f'(\tilde{x})$ one normally does not know exactly whether $f'(\tilde{x})$ or C are nonsingular. This can be guaranteed, however, a posteriori, if one assumes $[H]$ to be inclusion monotone, i. e., $[H]([x]) \subseteq [H]([y])$ for $[x] \subseteq [y]$, and if (2) can be checked for some k_0 for which $\tilde{x} \in [x]^{k_0}$ still holds—for example for $k_0 = 0$. The proof is based on the following argument:

Since $t'(\tilde{x}) = I - Cf'(\tilde{x})$ in the situation of Corollary 3.1, one gets by standard rules for the diameter (cf. [2] or [12])

$$\begin{aligned} d([x]^{k_0}) &> d([g]([x]^{k_0}_\varepsilon)) \geqslant d([g]([x]^{k_0})) \geqslant |t'(\tilde{x}) + [H]([x]^{k_0})| d([x]^{k_0}) \\ &\geqslant |t'(\tilde{x}) + [H](\tilde{x})| d([x]^{k_0}) = |t'(\tilde{x})| d([x]^{k_0}) = |I - Cf'(\tilde{x})| d([x]^{k_0}). \end{aligned}$$

Therefore, $d([x]^{k_0}) > 0$ and $\varrho(I - Cf'(\tilde{x})) < 1$ by Corollary 3.2.3 and Proposition 3.2.4 in [12], for example. If C or $f'(\tilde{x})$ are singular then 1 would be an eigenvalue of $I - Cf'(\tilde{x})$, which contradicts $\varrho(I - Cf'(\tilde{x})) < 1$.

4. Examples

In this section we will apply Corollary 3.1 to various algorithms for verifying and enclosing solutions of mathematical problems.

Example 4.1. (The algebraic eigenproblem for a simple real eigenvalue)

We consider first the algebraic eigenproblem $Av = \lambda v$. Apparently, each real eigenpair (v^*, λ^*) of $A \in \mathbb{R}^{n\times n}$ can be viewed as a zero of the function $f(x) := \begin{pmatrix} Av - \lambda v \\ v_{i_0} - \zeta \end{pmatrix}$ if the eigenvector v^* is normalized by $v^*_{i_0} = \zeta \neq 0$ in a component i_0 and if $x := (v^T, \lambda)^T$. Let $(\tilde{v}, \tilde{\lambda})$ be an approximation of (v^*, λ^*), where λ^* is an algebraic simple eigenvalue of A. In [14] the interval function

$$[g]([x]) := \tilde{x} - Cf(\tilde{x}) + \left\{ I_{n+1} - C \begin{pmatrix} A - \tilde{\lambda} I_n & -[v]) \\ (e^{(i_0)})^T & 0 \end{pmatrix} \right\} ([x] - \tilde{x}) \tag{23}$$

was applied with $[x] := ([v]^T, [\lambda])^T \in \mathbb{IR}^{n+1}$ in order to verify and to enclose $x^* := ((v^*)^T, \lambda^*)^T$. With $t(x) = x - Cf(x)$ as in Corollary 3.1 one gets

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

In [7] it was mentioned that for degenerate interval vectors $[x] \equiv x$ the expression $[g](x)$ from (23) is the complete Taylor expansion of $t(x)$ at $\tilde{x} := (\tilde{v}^T, \tilde{\lambda})^T$ even if $\tilde{x} \notin [x]$. Therefore, $t(x) \in [g]([x])$ holds trivially for all $x \in [x]$. With

$$[H]([x]) := C \begin{pmatrix} O & [v] - \tilde{v} \\ O & 0 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} \tag{24}$$

the function $[g]$ has the form (4). The property (6) can be seen at once, the Lipschitz condition (5) follows from

$$\begin{aligned} \|q([H]([x]), [H]([y]))\| &\leqslant \left\| |C| q\left(\begin{pmatrix} O & [v] - \tilde{v} \\ O & 0 \end{pmatrix}, \begin{pmatrix} O & [w] - \tilde{v} \\ O & 0 \end{pmatrix} \right) \right\|_\infty \\ &\leqslant \|C\|_\infty \|q([x], [y])\|_\infty, \end{aligned}$$

where $[x] = ([v]^T, [\lambda])^T \in \mathbb{IR}^{n+1}$ and $[y] = ([w]^T, [\mu])^T \in \mathbb{IR}^{n+1}$. Therefore, Corollary 3.1 applies with $\kappa := \|C\|_\infty$. It shows that under appropriate circumstances concerning the approximations C, $\tilde{x}$ and the inflation $[\delta]$, the iteration (14) ends up with the subset property (2), which guarantees an eigenpair of A in the final iterate $[x]^{k_0}$. □

The arguments in Example 4.1 apply without difficulties also to the generalized algebraic eigenproblem $Av = \lambda Bv$, where A, B are matrices from $\mathbb{R}^{n\times n}$. We leave the details to the reader.

Example 4.2. (Two-dimensional invariant subspaces)

In order to verify and to enclose a basis of a two-dimensional subspace of $\mathbb{R}^n$ which is invariant with respect to a linear mapping given by a matrix $A \in \mathbb{R}^{n\times n}$, Alefeld and Spreuer start in [3] with the function

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

where $x = (u^T, m_{11}, m_{21}, v^T, m_{12}, m_{22})^T \in \mathbb{R}^{2n+4}$, $i_1 \neq i_2 \in \{1, \ldots, n\}$ and $\varepsilon\theta - \zeta\eta \neq 0$. Taking into account the normalizations, 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. Again we set $t(x) := x - Cf(x)$ with a nonsingular matrix $C \in \mathbb{R}^{(2n+4)\times(2n+4)}$, and we choose $\tilde{x} = (\tilde{u}^T, \tilde{m}_{11}, \tilde{m}_{21}, \tilde{v}^T, \tilde{m}_{12}, \tilde{m}_{22})^T$ as an approximation of x^*. Then

$$t'(\tilde{x}) = I_{2n+4} - C \begin{pmatrix} A - \tilde{m}_{11} I_n & -\tilde{u} & -\tilde{v} & -\tilde{m}_{21} I_n & 0 & 0 \\ (e^{(i_1)})^T & 0 & 0 & 0 & 0 & 0 \\ (e^{(i_2)})^T & 0 & 0 & 0 & 0 & 0 \\ -\tilde{m}_{12} I_n & 0 & 0 & A - \tilde{m}_{22} I_n & -\tilde{u} & -\tilde{v} \\ 0 & 0 & 0 & (e^{(i_1)})^T & 0 & 0 \\ 0 & 0 & 0 & (e^{(i_2)})^T & 0 & 0 \end{pmatrix}$$

and

$$[H]([x]) = C \begin{pmatrix} O & [u] - \tilde{u} & [v] - \tilde{v} & O & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ O & 0 & 0 & O & [u] - \tilde{u} & [v] - \tilde{v} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{I}\mathbb{R}^{(2n+4)\times(2n+4)}$$

for $[g]$ from (4). Using the ususal rules for q one again easily verifies (5) and (6) which are the crucial assumptions for Corollary 3.1. $\square$

E x a m p l e 4.3. (The singular value problem)

Let $((u^i)^T, (v^i)^T, \sigma_i)^T \in \mathbb{R}^{m+n+1}$ be a vector which gathers a singular value σ_i of a rectangular matrix $A \in \mathbb{R}^{m\times n}$ and the corresponding i-th columns u^i, v^i of the orthogonal matrices $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{m\times m}$ of the singular value decomposition

$$A = V\Sigma U^T = V \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}) U^T.$$

Then this vector is obviously a zero 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 such 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 $\tilde{x} = (\tilde{u}^T, \tilde{v}^T, \tilde{\sigma})^T$ and let $C \in \mathbb{R}^{(m+n+1)\times(m+n+1)}$ be nonsingular. Similar to the development in [7] (cf. also [1]) we use $[g]$ from (4) with $t(x) := x - Cf(x)$,

$$t'(\tilde{x}) = I - C \begin{pmatrix} A & -\tilde{\sigma} I_m & -\tilde{v} \\ -\tilde{\sigma} I_n & A^T & -\tilde{u} \\ 2\tilde{u}^T & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(m+n+1)\times(m+n+1)}$$

and

$$[H]([x]) := C \begin{pmatrix} O & O & [v] - \tilde{v} \\ O & O & [u] - \tilde{u} \\ ([u] - \tilde{u})^T & 0 & 0 \end{pmatrix} \in \mathbb{I}\,\mathbb{R}^{(m+n+1)\times(m+n+1)},$$

in order to verify x^*. Again, $[g](x)$ is the complete Taylor expansion of $t(x) := x - Cf(x)$ at $x = \tilde{x}$. As in Example 4.1 one easily checks that (5) and (6) hold for $[H]$. Thus Corollary 3.1 applies. □

We finally mention that Corollary 3.1 also applies to quadratic systems of the form $t(x) := b + Ax + T(x, x) = x$, where $b, x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ and where $T\colon\ \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a bilinear operator. The details are left to the reader.

Acknowledgement. I want to thank cordially my colleague A. Frommer, University of Wuppertal, who asked me several times whether results in [9] also hold for contractive mappings and who thus gave the impetus for writing the present paper.

References

[1] *G. Alefeld*: Rigorous Error Bounds for Singular Values of a Matrix Using the Precise Scalar Product. Computerarithmetic (E. Kaucher, U. Kulisch and Ch. Ullrich, eds.). Teubner, Stuttgart, 1987, pp. 9–30.

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

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

[4] *O. Caprani and K. Madsen*: Iterative methods for interval inclusion of fixed points. BIT *18* (1978), 42–51.

[5] *K. Grüner*: Solving the Complex Algebraic Eigenvalue Problem with Verified High Accuracy. Accurate Numerical Algorithms, A Collection of Research Papers, Research Reports ESPRIT, Project 1072, DIAMOND, Vol. 1 (Ch. Ullrich and J. Wolff von Gudenberg, eds.). Springer, Berlin, 1989, pp. 59–78.

[6] *S. König*: On the Inflation Parameter Used in Self-Validating Methods. Contributions to Computer Arithmetic and Self-Validating Numerical Methods (Ch. Ullrich, ed.). Baltzer, IMACS, Basel, 1990, pp. 127–132.

[7] *G. Mayer*: Result Verification for Eigenvectors and Eigenvalues. Topics in Validated Computations (J. Herzberger, ed.). Elsevier, Amsterdam, 1994, pp. 209–276.

[8] *G. Mayer*: Über ein Prinzip in der Verifikationsnumerik. Z. angew. Math. Mech. *75* (1995), S II, S 545–S 546..

[9] *G. Mayer*: Epsilon-inflation in verification algorithms. J. Comp. Appl. Math. *60* (1995), 147–169.
[10] *G. Mayer*: On a unified representation of some interval analytic algorithms. Rostock. Math. Kolloq. *49* (1995), 75–88.
[11] *G. Mayer*: Success in Epsilon-Inflation. Scientific Computing and Validated Numerics (G. Alefeld, A. Frommer and B. Lang, eds.). Akademie Verlag, Berlin, 1996, pp. 98–104.
[12] *A. Neumaier*: Interval Methods for Systems of Equations. Cambridge University Press, Cambridge, 1990.
[13] *S. M. Rump*: Kleine Fehlerschranken bei Matrixproblemen. Thesis, Universität Karlsruhe, 1980.
[14] *S. M. Rump*: Solving Algebraic Problems with High Accuracy. A New Approach to Scientific Computation (U. W. Kulisch and W. L. Miranker, eds.). Academic Press, New York, 1983, pp. 53–120.
[15] *S. M. Rump*: New Results in Verified Inclusions. Accurate Scientific Computation, Lecture Notes in Computer Science Vol. 235 (W. L. Miranker and R. A. Toupin, eds.). Springer, Berlin, 1986, pp. 31–69.
[16] *S. M. Rump*: On the solution of interval linear systems. Computing *47* (1992), 337–353.

Author's address: *Günter Mayer*, Fachbereich Mathematik, Universität Rostock, Universitätsplatz 1, D-18051 Rostock, Germany, e-mail: `guenter.mayer@mathematik.uni-rostock.de`. ∎