Computing 47, 337–353 (1992)

Computing

On the Solution of Interval Linear Systems

S. M. Rump, Hamburg

Received January 28, 1991

Abstract — Zusammenfassung

On the Solution of Interval Linear Systems. In the literature efficient algorithms have been described for calculating guaranteed inclusions for the solution of a number of standard numerical problems [3, 4, 8, 11, 12, 13]. The inclusions are given by means of a set containing the solution. In [12, 13] this set is calculated using an affine iteration which is stopped when a nonempty and compact set is mapped into itself. For exactly given input data (point data) it has been shown that this iteration stops if and only if the iteration matrix is convergent (cf. [13]).

In this paper we give a necessary and sufficient stopping criterion for the above mentioned iteration for interval input data and interval operations. Stopping is equivalent to the fact that the algorithm presented in [12] for solving interval linear systems computes an inclusion of the solution. An algorithm given by Neumaier is discussed and an algorithm is proposed combining the advantages of our algorithm and a modification of Neumaier's. The combined algorithm yields tight bounds for input intervals of small and large diameter.

Using a paper by Jansson [6, 7] we give a quite different geometrical interpretation of inclusion methods. It can be shown that our inclusion methods are optimal in a specified geometrical sense. For another class of sets, for standard simplices, we give some interesting examples.

AMS Subject Classifications: 65G10, 65F05

Key words: Interval iteration, linear interval systems, standard simplices.

Zur Lösung linearer Intervallgleichungssysteme. In der Literatur werden eine Reihe effizienter Algorithmen beschrieben zur Berechnung garantierter Einschließungen der Lösung numerischer Standardprobleme [3, 4, 8, 11, 12, 13]. Die Einschließungen werden in Form von Mengen gegeben. In [12, 13] wird diese Menge mit Hilfe einer affinen Transformation berechnet, die stoppt, wenn eine nichtleere kompakte Menge in sich selbst abgebildet wird. Für Punkteingabedaten wurde gezeigt, daß diese Iteration genau dann stoppt, wenn die Iterationsmatrix konvergent ist [13].

In der vorliegenden Arbeit werden notwendige und hinreichende Stop-Bedingungen angegeben für Intervalleingabedaten und Intervalloperationen im reellen und im komplexen. Stoppen heißt hierbei, daß der Algorithmus aus [12] für Intervallgleichungssysteme eine Einschließung liefert. Ein Algorithmus von Neumaier wird diskutiert, und es wird ein Hybrid-Algorithmus vorgeschlagen, der die Vorteile Neumaiers und unseres Algorithmus kombiniert.

Unter Benutzung einer Arbeit von Jansson [6, 7] wird eine interessante geometrische Interpretation von Einschließungsalgorithmen gegeben. Es wird gezeigt, daß die Einschließungsalgorithmen in bestimmtem Sinne optimal sind. Für eine andere Klasse von Mengen, für Standardsimplexe, geben wir einige interessante Beispiele.

0. Introduction

Let T denote one of the sets $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^n$ (real vectors with n components), $\mathbb{C}^n$ (complex vectors with n components), $\mathbb{R}^{n\times n}$ (real square matrices with n rows and columns)

or $\mathbb{C}^{n\times n}$ (complex square matrices with n rows and columns). Throughout this paper the letter "n" is reserved in the prescribed way; only square matrices (which are $n \times n$) will occur. $\mathbb{P}T$ denotes the power set over T.

In the following $* \in \{+, -, \cdot, /\}$ denotes the binary real resp. complex operations. These operations extend in the usual way to power set operations. If $x * y \in T_3$ is defined for $x \in X \in \mathbb{P}T_1$, $y \in Y \in \mathbb{P}T_2$ then

$$X * Y := \{x * y | x \in X, y \in Y\} \in \mathbb{P}T_3 .$$

The set of all n-dimensional resp. n^2-dimensional hyperrectangles parallel to the axis over real resp. complex numbers is denoted by $\mathbb{IR}^n$, $\mathbb{IC}^n$, $\mathbb{IR}^{n\times n}$, $\mathbb{IC}^{n\times n}$, resp. This is one way to represent interval vectors or interval matrices. Intervals are always supposed to be nonempty.

The rounding of an arbitrary set X into the smallest hyperrectangle containing X is denoted by $\diamondsuit: \mathbb{P}T \to \mathbb{I}T$

$$X \in \mathbb{P}T \Rightarrow \diamondsuit(X) := \bigcap \{Y \in \mathbb{I}T | X \subseteq Y\} \in \mathbb{I}T.$$

The set $\diamondsuit(X)$ is well-defined and unique. We define operations $\hat{+}, \hat{-}, \hat{\cdot}, \hat{/}$ over $\mathbb{I}T$ by

$$[X], [Y] \in \mathbb{I}T \Rightarrow [X] \hat{*} [Y] := \diamondsuit([X] * [Y]) \qquad \text{for } * \in \{+, -, \cdot, /\}.$$

This is the smallest hyperrectangle containing the result of the powers set operation. It is uniquely defined and effectively computable (cf. [2, 9, 10, 11]).

With the componentwise order relation $\leq$ for all sets in T (with partial ordering for complex numbers) hyperrectangles are usually described by their bounds. Obviously

$$[X] \in \mathbb{I}T \Leftrightarrow [X] = \{x \in T | \inf([X]) \leq x \leq \sup([X])\}.$$

Therefore we adopt the notation $[\underline{X}, \bar{X}]$ with $\underline{X} = \inf([X])$, $\bar{X} = \sup([X])$ for hyperrectangles and especially

$$\begin{aligned} [X] &= \operatorname{mid}([X]) \pm \operatorname{rad}([X]) \\ &= [\operatorname{mid}([X]) - \operatorname{rad}([X]), \operatorname{mid}([X]) + \operatorname{rad}([X])] \end{aligned} \tag{0.1}$$

where $\operatorname{mid}([X]) = 0.5 \cdot (\inf([X]) + \sup([X]))$ denotes the midpoint of X, $\operatorname{rad}([X]) = 0.5 \cdot (\sup([X]) - \inf([X]))$ the radius of $[X]$. If $[X]$ is a vector or a matrix, then $\operatorname{mid}([X])$ and $\operatorname{rad}([X])$ is a real or complex vector or matrix, respectively. Note that $\operatorname{rad}([X]) \geq 0$. For any $a, b \in T$ with $T \in \{\mathbb{R}, \mathbb{C}, \mathbb{R}^n, \mathbb{C}^n, \mathbb{R}^{n\times n}, \mathbb{C}^{n\times n}\}$ we define similar to (0.1)

$$a \pm b := \{x \in T | a - b \leq x \leq a + b\} \in \mathbb{I}T \qquad \text{for } b \geq 0.$$

If $b_\nu = 0$ for some component of b the hyperrectangle $a \pm b$ is degenerated, the interior is empty.

For sets $X, Y \subseteq T$, $\operatorname{int}(X)$ denotes the interior of X, $X \overset{\circ}{\subseteq} Y$ means $X \subseteq \operatorname{int}(Y)$, $\operatorname{Re}(X)$ denotes the real part, $\operatorname{Im}(X)$ the imaginary part of X. For a real matrix A we define

$|A|$ to be the matrix of absolute values of the components of A, for a complex matrix $|A|$ is $|\mathrm{Re}(A)| + |\mathrm{Im}(A)|$ (cf. [2]). For an interval $[X] \in \mathbb{I}S$, $S \in \{\mathbb{R}, \mathbb{C}\}$ we define $|[X]| = \max\{|x| | x \in [X]\}$ extending componentwise to interval vectors and matrices. For two hyperrectangles $[X]$, $[Y] \in \mathbb{I}S$ the distance q is defined as usual by

$$q([X], [Y]) = \max(|\inf([X]) - \inf([Y])|, |\sup([X]) - \sup([Y])|).$$

For vectors and matrices the distance is defined componentwise. For $A \in S^{n\times n}$, $S \in \{\mathbb{R}, \mathbb{C}\}$ the spectral radius of A is denoted by $\rho(A)$, for $[A] \in \mathbb{I}S^{n\times n}$ we define $\rho([A]) := \max\{\rho(A) | A \in [A]\}$. A_i denotes the ith row of A.

1. Criteria for Convergence of a Matrix

In [12] the following theorem has been proved:

Theorem 1. *Let $S \in \{\mathbb{R}, \mathbb{C}\}$, $C \in S^{n\times n}$, b, $\tilde{x} \in S^n$, $R \in S^{n\times n}$ and $\emptyset \neq X \subseteq S^n$ be compact. If*

$$R \cdot (b - C\tilde{x}) + \{I - RC\} \cdot X \subseteq \mathrm{int}(X) \tag{1.1}$$

then C and R are non-singular and the unique solution $\hat{x}$ of $Cx = b$ satisfies $\hat{x} \in \tilde{x} + X$.

I denote the identity matrix, all operations in (1.1) are power set operations. In a practical application of Theorem 1 one may start an iteration

$$X^{k+1} := R \cdot (b - C\tilde{x}) + \{I - RC\} \cdot X^k$$

for given $X^0 \subseteq S^n$. Clearly,

$$X^{k+1} \subseteq \mathrm{int}(X^k) \tag{1.2}$$

implies all assertions of Theorem 1. In the following conditions will be investigated under which (1.2) is satisfied.

(1.2) can be reduced to an affine iteration

$$X^{k+1} := z + A \cdot X^k \text{ for } z \in S^n, A \in S^{n\times n}. \tag{1.3}$$

In [13] the following theorem has been proved:

Theorem 2. *For $S \in \{\mathbb{R}, \mathbb{C}\}$ let $A \in S^{n\times n}$ be an arbitrary matrix, $z \in S^n$ and $\emptyset \neq X \in \mathbb{P}S^n$ be compact. Then*

$$z + A \cdot X \subseteq \mathrm{int}(X) \textit{ implies } \rho(A) < 1.$$

Therefore a contracting A is necessary for an affine iteration (1.3) to stop with (1.2). But, in general, it cannot be true that (1.2) is satisfied for some $k \in \mathbb{N}$ for *every* starting set X^0 because of two reasons: First, the interior of X^0 must be nonempty because $\mathrm{int}(X^0) = \emptyset$ implies $\mathrm{int}(X^k) = \emptyset$ for every $k \in \mathbb{N}$. Second, (1.2) implies $\hat{x} \in X^0$.

In other words only those sets X^0 already containing $\hat{x}$ are suitable to achieve (1.2). For practical applications this is hardly acceptable.

To overcome those difficulties the so-called ε-inflation has been introduced in [12]. One possible definition for general sets is the following.

Definition 3. For a set $X \subseteq S^n$, $S \in \{\mathbb{R}, \mathbb{C}\}$ the ε-inflation $X \circ \varepsilon$ is defined by

$$X \circ \varepsilon := X + U_\varepsilon(0) \text{ for } 0 < \varepsilon \in \mathbb{R},$$

where $U_\varepsilon(0)$ is some closed and bounded set containing the origin as an interior point.

Obviously, $X \subseteq \text{int}(X \circ \varepsilon)$. An example for $U_\varepsilon(0)$ is the closed ball of radius ε around the origin. Using the ε-inflation we can define an iteration scheme allowing a complete analysis.

Theorem 4. *Let $A \in S^{n \times n}$ be an arbitrary matrix, $\varnothing \neq Z \subseteq S^n$ be a compact set of vectors, $S \in \{\mathbb{R}, \mathbb{C}\}$. For some compact $\varnothing \neq X^0 \subseteq S^n$ let*

$$X^{k+1} := (Z + A \cdot X^k) \circ \varepsilon_k \text{ for } 0 \leq k \in \mathbb{N}, \tag{1.4}$$

where $U_{\varepsilon_{k+1}} \subseteq U_{\varepsilon_k}$ and $U \subseteq U_{\varepsilon_k}$ for every $k \in \mathbb{N}$ and some compact $\varnothing \neq U \subseteq S^n$ with $0 \in \text{int}(U)$. Then the following two conditions are equivalent:

a) $\forall \varnothing \neq X^0 \in S^n$ *compact* $\exists k \in \mathbb{N}: Z + A \cdot X^k \subseteq \text{int}(X^k)$
b) $\rho(A) < 1$.

Proof. See [13].

Theorem 4 is of theoretical interest. In practical implementations general sets can hardly be handled. Therefore we are aiming on obtaining results similar to theorem 4 starting with an interval X^0 and using interval operations in (1.4).

2. Interval Iterations

If the input data are not exactly representable on the computer they may be replaced by the smallest enclosing intervals. Input intervals occur as well if the input data are afflicted with tolerances. In both cases an inclusion of the set of all solutions is to be calculated.

In case of hyperrectangles an ε-inflation should consist of an absolute and a relative part in order to maintain (1.4) for a small value of k. A possible definition which turned out to be very suitable in practical applications is

$$[X] \in \mathbb{I}S: [X] \circ \varepsilon := [J] \diamondsuit [X] \boxplus [E]$$

with a diagonal matrix $[J] \in \mathbb{I}S^{n \times n}$ and $[E] \in \mathbb{I}S^n$ and $1 \in [J_{ii}]$, $0 \in \text{int}([E_i])$ for $1 \leq i \leq n$. In the following we state a theorem similar to Theorem 4 for intervals (hyperrectangles) and the corresponding interval operations $\circledast$. Furthermore, it turned out to be useful to adapt E to the iteration process. Therefore, in the following theorem we use a more general definition of the ε-inflation.

Theorem 5. *Let $[A] \in \mathbb{I}S^{n \times n}$ be an interval matrix, $[Z] \in \mathbb{I}S^n$ be an interval vector, $S \in \{\mathbb{R}, \mathbb{C}\}$. For*

$$f: \mathbb{I}S^n \to \mathbb{I}S^n \text{ with } [Y] \in \mathbb{I}S^n: f([Y]) := [Z] \oplus [A] \diamondsuit [Y]$$

and for $[X^0] \in \mathbb{I}S^n$ *we define the iteration*

$$[X]^{k+1} := J \diamondsuit f([X^k]) \oplus E^k \tag{2.1}$$

with diagonal matrix $[J] \in \mathbb{I}S^{n\times n}$, $[E^k] \in \mathbb{I}S^n$, *for* $0 \le k \in \mathbb{N}$. *Let* $[E^k] \to [E] \in \mathbb{I}S^n$, $0 \in \operatorname{int}([E])$, $1 \in [J]_{ii}$ *for* $1 \le i \le n$ *and* $\rho(|[J]| \cdot |[A]|) < 1$. *Then the following two conditions are equivalent*:

a) $\forall \varnothing \neq [X^0] \in \mathbb{I}S^n \exists k \in \mathbb{N}: f([X^k]) \subseteq \operatorname{int}([X^k])$
b) $\rho(|[A]|) < 1$.

Proof. "⇒" For $S = \mathbb{R}$ this is proved in [13], theorem 6. For $S = \mathbb{C}$ let $[Y] := [X] - [X] = [X] \ominus [X] = \{x_1 - x_2 | x_1, x_2 \in [X]\} \in \mathbb{I}\mathbb{C}$. Then for $A \in [A]$, $z \in [Z]$

$$\begin{aligned} A \cdot [Y] &= \{A \cdot (x_1 - x_2) | x_1, x_2 \in [X]\} \\ &= \{(z + Ax_1) - (z + Ax_2) | x_1, x_2 \in [X]\} \\ &= (z + A \cdot [X]) - (z + A \cdot [X]) \\ &\subseteq \operatorname{int}([X]) - \operatorname{int}([X]) \\ &= \operatorname{int}([Y]). \end{aligned}$$

Since this holds for every $A \in [A]$ we get $[A] \cdot [Y] \subseteq \operatorname{int}([Y])$ and hence $[A] \diamondsuit [Y] \subseteq \operatorname{int}([Y])$. Using $[Y] = \pm \operatorname{rad}([Y])$ we get after short computation

$$\begin{aligned} &\{|\operatorname{Re}([A])| + |\operatorname{Im}([A])|\} \cdot \{\operatorname{Re}(\operatorname{rad}([Y])) + \operatorname{Im}(\operatorname{rad}([Y]))\} \\ &\quad = \operatorname{Re}(\operatorname{rad}([A] \diamondsuit [Y]) + \operatorname{Im}(\operatorname{rad}([A] \diamondsuit [Y])) \\ &\quad < \operatorname{Re}(\operatorname{rad}([Y])) + \operatorname{Im}(\operatorname{rad}([Y])). \end{aligned}$$

By a) the real vector $\operatorname{Re}(\operatorname{rad}([Y])) + \operatorname{Im}(\operatorname{rad}([Y]))$ is positive. Therefore, Perron-Frobenius Theory finishes this part of the proof.

"⇐" Let $g: \mathbb{I}S^n \to \mathbb{I}S^n$ be defined by

$$g([X]) := [J] \diamondsuit ([Z] \oplus [A] \diamondsuit [X]) \oplus [E]$$

for $[X] \in \mathbb{I}S^n$. Then for $[X], [Y] \in \mathbb{I}S^n$ the rules of interval analysis (cf. [2, 10]) imply

$$\begin{aligned} q(g([X]), g([Y])) &\le |[J]| \cdot q([Z] \oplus [A] \diamondsuit [X], [Z] \oplus [A] \diamondsuit [Y]) \\ &\le |[J]| \cdot |[A]| \cdot q([X], [Y]) \end{aligned} \tag{2.2}$$

because $[J]$ is diagonal. By assumption $\sigma := \rho(|[J]| \cdot |[A]|) < 1$ and hence there is some $[X^*] \in \mathbb{I}S^n$ with $g([X^*]) = [X^*]$ (cf. Theorem 1, Chapter 12 in [2]). Now $q(g([X^0]), [X^1]) = q([E], [E^0])$ and by induction follows

$$q(g^k([X^0]), [X^{k+1}]) \le \sum_{i=0}^{k} \sigma^i \cdot q([E], [E^{k-i}])$$

because with (2.2) we have

$$\begin{aligned} q(g^{k+1}([X^0]),[X^{k+2}]) &\le q(g(g^k([X^0])),g([X^{k+1}])) + q(g([X^{k+1}]),[X^{k+2}]) \\ &\le \sigma \cdot q(g^k([X^0]),[X^{k+1}]) + q([E],[E^{k+1}]) \\ &= \sum_{i=0}^{k+1} \sigma^i \cdot q([E],[E^{k+1-i}]). \end{aligned}$$

By assumption $q([E],[E^k]) \to 0$ for $k \to \infty$ and therefore $[X^k]$ and $g^k([X^0])$ have the same limit $[X^*]$ for $k \to \infty$.

Let $0 < \varepsilon^* < \min(|\inf([E])|, |\sup([E])|)$, $\varepsilon^* \in S^n$. Then $0 \in [E]$ and $\pm\varepsilon^* \subseteq [E]$ implies $\operatorname{diam}([X^*]) \ge \operatorname{diam}([E]) > \varepsilon^*$. Let some $\varepsilon \in \mathbb{R}^n$ with $0 < \varepsilon \le \varepsilon^*$ be given. Then there is a $k \in \mathbb{N}$ with

$$q([X^k],[X^*]) < 0.5 \cdot \varepsilon \text{ and } q([Z] \oplus [A] \odot [X^*], [Z] \oplus [A] \odot [X^k]) < 0.5 \cdot \varepsilon. \tag{2.3}$$

Then the first part of (2.3) implies

$$[\bar{X}] := [\inf([X^*]) + 0.5 \cdot \varepsilon, \sup([X^*]) - 0.5 \cdot \varepsilon] \subseteq \operatorname{int}([X^k]). \tag{2.4}$$

Now

$$\begin{aligned} [Z] \oplus [A] \odot [X^k] &\subseteq [Z] \oplus [A] \odot [X^*] \pm 0.5 \cdot \varepsilon \\ &\subseteq [J] \odot ([Z] \oplus [A] \odot [X^*]) \pm 0.5 \cdot \varepsilon \\ &\subseteq [\bar{X}] \\ &\subseteq \operatorname{int}([X^k]) \end{aligned}$$

finishes the proof. □

In a typical application J is a diagonal matrix with identical entries $1 \pm \varepsilon$ in the diagonal. For general sets of matrices $\{A\} \in \mathbb{PR}^{n \times n}$ the generalization of Theorem 5 replacing part b) by

$$\rho(A) < 1 \qquad \text{for all } A \in \{A\}$$

is not true. Part a) implies $\rho\left(\prod_{\nu=1}^{m} A_\nu\right) < 1$ for all $A_\nu \in \{A\}$, $\nu = 1 \ldots m$ and in [13] an example of a set of matrices $\{C + \sigma(D - C) | 0 \le \sigma \le 1\}$ for two matrices C, $D \in \mathbb{R}^{n \times n}$ is given with $\rho(C) < 1$, $\rho(D) < 1$ but $\rho(C \cdot D) > 1$.

The assumption $\rho(|[J]| \cdot |[A]|) < 1$ in Theorem 5 is necessary. Consider

$$[A] := \begin{pmatrix} 0 & 2 \\ 1/8 & 0 \end{pmatrix}, \qquad Z := 0, \qquad [X^0] = \begin{pmatrix} [-1,1] \\ [-1,1] \end{pmatrix},$$

$$[J] := \begin{pmatrix} [-4,4] & 0 \\ 0 & [-4,4] \end{pmatrix}$$

and

$$[E^k] = [E] := \begin{pmatrix} [-1/4,1/4] \\ [-1/4,1/4] \end{pmatrix} \qquad \text{for } k \in \mathbb{N}.$$

Then all succeeding $[X^k]$ are symmetric w.r.t. the origin, i.e. $\diamondsuit[X^k] = [X^k]$. Therefore, $f([X^k]) \subseteq \text{int}([X^k])$ is equivalent to

$$A \cdot X^k < X^k \tag{2.5}$$

for

$$A := \begin{pmatrix} 0 & 2 \\ 1/8 & 0 \end{pmatrix}, \qquad X^0 := \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$X^{k+1} := J \cdot A \cdot X^k + E \qquad \text{with } J := \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}, E := \begin{pmatrix} 1/4 \\ 1/4 \end{pmatrix}.$$

Then short computation yields for $0 \leq k \in \mathbb{N}$

$$X^{2k} = \begin{pmatrix} 7 \cdot 2^{2k-2} - 3/4 \\ 9 \cdot 2^{2k-3} - 1/8 \end{pmatrix} \qquad \text{and} \qquad X^{2k+1} = \begin{pmatrix} 9 \cdot 2^{2k} - 3/4 \\ 7 \cdot 2^{2k-3} - 1/8 \end{pmatrix}$$

and

$$(X^{2k} - A \cdot X^{2k})_1 = -2^{2k-1} - 1/2, \qquad (X^{2k+1} - A \cdot X^{2k})_2 = -2^{2k-2} - 1/32.$$

This shows that (2.5) is not satisfied for any $k \in \mathbb{N}$. It is $\rho(|[J]| \cdot |[A]|) = 2 \geq 1$. In the example it is crucial that A is not primitive.

Using hyperrectangles, i.e. rectangular intervals, is very convenient on digital computers. The operations are simple and fast and can be executed on any computer with a precisely defined computer arithmetic and directed roundings available, e.g. as defined in the IEEE 754 floating-point arithmetic standard (cf. [2, 5, 9, 10, 11]). Using the arithmetic defined by Kulisch [9] with a precise scalar product gives additional advantages, especially in the case of point data or intervals with small diameters.

Working with general sets instead is hardly possible on computers. One way of representing sets being more general than hyperrectangles are simplices. Simplices are representable on digital computers by means of their vertices and are closed under affine mappings. However, operations are fairly expensive: a matrix-vector multiplication costs $0(n^3)$ compared to $0(n^2)$ when using hyperrectangles. Another possibility are standard simplices which will be discussed in Chapter 4.

3. An Inclusion Method without Interval Iteration

In his book [11], page 150, Neumaier proposes the following algorithm for computing an inclusion of the solution set $[A]^H[b] = \{x \in \mathbb{R}^n | Ax = b \text{ for } A \in [A], b \in [b]\}$ of an interval linear system with matrix $[A] \in \mathbb{IR}^{n \times n}$ and right hand side $[b] \in \mathbb{IR}^n$:

Define $$\langle [X] \rangle := \min_{x \in X} |x| \text{ for } [X] \in \mathbb{IR}$$

and the comparison matrix

$$\langle [A] \rangle_{ij} := \begin{cases} \langle [A]_{ij} \rangle & \text{for } i = j \\ -|[A]_{ij}| & \text{otherwise.} \end{cases}$$

Algorithm (Neumaier)

1) Find an approximate inverse $R \approx \mathrm{mid}([A])^{-1}$ and compute $[A'] = R \diamondsuit [A]$, $[b'] = R \diamondsuit [b]$.
2) Find an approximate solution $\tilde{u} > 0$ of $\langle [A'] \rangle \cdot \tilde{u} = |[b']| + (\varepsilon, \ldots, \varepsilon)^T$ for some small $\varepsilon > 0$ and a number $\alpha > 0$ such that $\langle [A'] \rangle \cdot \tilde{u} \geq \alpha \cdot |[b']|$. (If this is not possible we conclude that either $[A]$ was not strongly regular or the precision of the calculation was not high enough).
3) Perform a few (one or two) steps of preconditioned Gauss-Seidel iteration, starting with

$$Z^0 := \alpha^{-1} \cdot \tilde{u} \cdot [-1, 1].$$

Each iterate in step 3 is an enclosure of $[A]^H[b]$.

In order to compare this algorithm with an inclusion algorithm with an interval iteration based on Theorem 1 (cf. [12, 13, 1, 14]) some modifications are necessary. Neumaier's original algorithm assumes A to be strongly regular. We want to avoid any preassumption on A, R or b. Therefore, the algorithm will be modified in a way that no such a priori assumption is necessary. This will also prove the non-singularity of every $A \in [A]$. It can be achieved by assuming $\langle [A'] \rangle \cdot \tilde{u} > \alpha \cdot |[b']|$ in step 2:

Theorem 6. *Let $[A] \in \mathbb{IR}^{n \times n}$, $[b] \in \mathbb{IR}^n$, $R \in \mathbb{R}^{n \times n}$ be given such that some $0 < u \in \mathbb{R}^n$, $0 < \alpha \in \mathbb{R}$ exist with*

$$\langle R \diamondsuit [A] \rangle \cdot u > \alpha \cdot |R \diamondsuit [b]|. \tag{3.1}$$

Then R can be scaled by the diagonal matrix D with $D_{ii} = \beta \cdot (\mathrm{mid}(R \diamondsuit [A])_{ii})^{-1}$ with $0 < \beta \leq 1$ such that $\tilde{R} := D \cdot R$ satisfies $(\mathrm{mid}(\tilde{R} \diamondsuit [A]))_{ii} \leq 1$ for $1 \leq i \leq n$, and for $[X] := \alpha^{-1} \cdot u \cdot [-1, 1]$ holds

$$\tilde{R} \diamondsuit [b] \oplus \{I \ominus \tilde{R} \diamondsuit [A]\} \diamondsuit [X] \subseteq \mathrm{int}([X]). \tag{3.2}$$

Proof. The definition of the comparison matrix $\langle R \diamondsuit [A] \rangle$ and (3.1) imply $0 \notin (\langle R \diamondsuit [A] \rangle)_{ii}$ for $1 \leq i \leq n$. Hence $\tilde{R}$ is well-defined and satisfies $(\mathrm{mid}(\tilde{R} \diamondsuit [A]))_{ii} \leq 1$ and $\langle \tilde{R} \diamondsuit [A] \rangle \cdot u > \alpha \cdot |\tilde{R} \diamondsuit [b]|$. Therefore,

$$\tilde{R}_i \diamondsuit [b] \subseteq \pm |\tilde{R}_i \diamondsuit [b]| \overset{\circ}{\leqq} \pm \alpha^{-1} \cdot (\langle \tilde{R} \diamondsuit [A] \rangle)_i \cdot u = \alpha^{-1} \cdot [+d - e, -d + e] \tag{3.3}$$

with

$$d := + \sum_{\substack{j=1 \\ j \neq i}}^{n} |\tilde{R} \diamondsuit [A]|_{ij} \cdot u_j \text{ and } e := (\langle \tilde{R} \diamondsuit [A] \rangle)_{ii} \cdot u_i. \tag{3.4}$$

Moreover,

$$(I \ominus \tilde{R} \diamondsuit [A])_i \diamondsuit [X] \subseteq \alpha^{-1} \cdot [-d, +d] \pm \alpha^{-1} \cdot |1 - (\tilde{R} \diamondsuit [A])_{ii}| \cdot u_i. \tag{3.5}$$

Adding (3.3) and (3.5) and observing (3.4) yields

$$\text{l.h.s. (3.2)} \overset{\circ}{\leqq} \pm\alpha^{-1} \cdot \{(\langle \tilde{R} \diamondsuit [A] \rangle)_{ii} + |1 - (\tilde{R} \diamondsuit [A])_{ii}|\} \cdot u_i .$$

By the definition of $[X]$ we are finished if we show

$$(\langle \tilde{R} \diamondsuit [A] \rangle)_{ii} + |1 - (\tilde{R} \diamondsuit [A])_{ii}| \leq 1 \tag{3.6}$$

for $1 \leq i \leq n$. With the abbreviation $Y := (\tilde{R} \diamondsuit [A])_{ii}$ for some $1 \leq i \leq n$ it is $Y > 0$ and $\text{mid}(Y) \leq 1$. Therefore, $0 < \inf(Y) \leq 1 \leq \sup(Y)$ and

$$|1 - Y| = \max(1 - \inf(Y), \sup(Y) - 1).$$

Using $\langle Y \rangle = \inf(Y)$ and $\inf(Y) + \sup(Y) = 2 \cdot \text{mid}(Y) \leq 2$ demonstrates (3.6) for every $1 \leq i \leq n$ and therefore finishes the proof. □

Together with Theorem 5 this implies $\rho(|I \diamondsuit \tilde{R} \diamondsuit [A]|) < 1$. Therefore an iteration similar to (2.1) will stop. If, on the other hand, $\rho(|I \diamondsuit \tilde{R} \diamondsuit [A]|) < 1$, then $\tilde{R} \diamondsuit [A]$ is an H-matrix and there are u and α satisfying (3.1) (cf. Proposition 3.7.2 in [11]).

Usually an inclusion algorithm first performs a residual iteration to obtain a reasonably good approximate solution $\tilde{x}$. Then the inclusion algorithm is applied to $Ay = b - A\tilde{x}$ yielding an inclusion for $\Sigma([A], [b]) - \tilde{x}$. To give a fair comparison we modify Neumaier's algorithm in this way. Furthermore, step 2 is changed according to Theorem 6 to prove the non-singularity of every $A \in [A]$. This leads to the following modification of Neumaier's algorithm.

Algorithm A

1) Find an approximate inverse $R \approx \text{mid}([A])^{-1}$, compute $x^0 \approx R \cdot \text{mid}([b])$ and perform a residual iteration yielding $\tilde{x}$, $[A'] := R \diamondsuit [A]$, $[b'] := R \diamondsuit ([b] \diamondsuit [A] \diamondsuit x)$.
2) Find an approximate solution $\tilde{u} > 0$ of $\langle [A'] \rangle \cdot \tilde{u} = |[b']| + (\varepsilon, \ldots, \varepsilon)^T$ for some small number $\varepsilon > 0$ and a number $\alpha > 0$ such that $\langle [A'] \rangle \cdot \tilde{u} > \alpha \cdot |[b']|$. (If this is not possible we conclude that either $[A]$ was not strongly regular or the precision of the calculation was not high enough).
3) Perform a few (one or two) steps of preconditioned Gauss-Seidel iteration, starting with $Z^0 := \alpha^{-1} \cdot \tilde{u} \cdot [-1, 1]$. It has been verified that every $A \in [A]$ is regular and each iterate Z in step 3 satisfies

$$\Sigma([A], [b]) \subseteq \tilde{x} \oplus Z .$$

Algorithm A will be compared with the following Algorithm B given in [12, 13] with the modification that $R \approx \text{mid}([A])^{-1}$ is replaced by $\tilde{R} := D \cdot R$ with $D_{ii} := (\text{mid}(R \diamondsuit [A])_{ii})^{-1}$. According to Theorem 6 this is the best choice. Smaller components D_{ii} still work but increase the spectral radius of $I \diamondsuit \tilde{R} \diamondsuit [A]$.

Algorithm B

1) Find an approximate inverse $R \approx \text{mid}([A])^{-1}$, compute $x^0 := R \cdot \text{mid}([b])$ and perform a residual iteration yielding $\tilde{x}$, $[Z] := R \diamondsuit ([b] \diamondsuit [A] \diamondsuit \tilde{x})$, $[C] := I \diamondsuit R \diamondsuit [A]$. (If $\text{mid}(I \diamondsuit C)_{ii} = 0$ for some $1 \leq i \leq n$ then goto 99). Compute $D_{ii} := (\text{mid}(I \diamondsuit C)_{ii})^{-1}$ and $Z_i := D_{ii} \diamondsuit Z_i$, $C_{ij} := D_{ii} \diamondsuit C_{ij}$ for $1 \leq i, j \leq n$.

2) Define $[X] := [Z]$, $k := 0$ and
repeat $k := k + 1$; $[Y] := [X] \pm \varepsilon$; inclusion := true;
for $i := 1$ *to* n *do*
$\{[X]_i := [Z]_i \boxplus [C]_i \boxdot ([X]_1, \ldots, [X]_{i-1}, [Y]_i, \ldots, [Y]_n)^t$;
inclusion := inclusion *and* $[X]_i \subseteq \text{int}([Y]_i)\}$;
until inclusion *or* $k > 15$;
3) Perform a few (one or two) iterations of the form $[X] := [Z] \boxplus [C] \boxdot [X]$ using Einzelschrittverfahren.
If inclusion *then*
{every $A \in [A]$ is regular and $\Sigma([A], [b]) \subseteq \tilde{x} \boxplus [X]\}$;
stop;
99) Either $|C|$ is not contracting or the precision of the calculation was not high enough.

Note that in step 2 an Einzelschrittverfahren is used (cf. [13]). The discussions above show that either both Algorithms A and B compute an inclusion of $\Sigma([A], [b])$ or not, except when the number of necessary iterations in step 2 in Algorithm B would be greater than 15. In many practical experiments this case did not occur. The price Algorithm A has to pay is the extra solution of a linear system adding some $1/3 \cdot n^3$ operations.

In the following tables we compare Algorithm A with Algorithm B and display the ratio of the diameters of the inclusion of the solution achieved by Algorithm A vs. Algorithm B. Second we display the ratio of computing times. Therefore, a number less than one indicates advantages for Algorithm A. The numbers are rounded to three decimal places.

Our first examples are Hilbert-matrices scaled by $lcm(1, 2, \ldots, 2n-1)$ s.t. all entries are integers, Pascal-matrices P with entries $P_{ij} := \binom{i+j}{j}$ and Boothroyd-matrices B defined by $B_{ij} := n \cdot (i + j - 1)^{-1} \cdot \binom{n+i-1}{i-1} \cdot \binom{n-1}{n-j}$. The system matrix A is transformed to an interval matrix $[A] := A \cdot (1 \pm \varepsilon)$. Results for different values of ε are displayed. It is $[b] := [A] \cdot (+1, -1, +1, \ldots)^T$ and we used an IBM 4361 with 14 hexadecimal digits in the mantissa corresponding to about 17 decimal places.

Let $[X_A]$, $[X_B]$ be the inclusions and t_A, t_B be the computing times for Algorithm A, Algorithm B, respectively. Then

$$d_1 := \min_i \frac{d([X_A]_i)}{d([X_B]_i)}, \qquad d_2 := \max_i \frac{d([X_A]_i)}{d([X_B]_i)} \quad \text{and} \quad t_A/t_B \tag{3.7}$$

is displayed. We have $n = 10$.

Table 1. Comparison Algorithms A, B, $n = 10$

	Hilbert			Pascal			Boothroyd		
ε	d_1	d_2	t_A/t_B	d_1	d_2	t_A/t_B	d_1	d_2	t_A/t_B
$\varepsilon = 0$	1.006	1.008	1.058	1.000	1.000	1.159	1.000	1.013	1.058
$\varepsilon = 10^{-16}$	1.027	1.029	1.058	1.003	1.003	1.159	1.004	1.005	1.058
$\varepsilon = 10^{-14}$	1.000	1.000	1.014	1.000	1.000	1.159	0.996	1.000	1.058
$\varepsilon = 10^{-13}$	0.951	0.983	1.014	1.000	1.000	1.159	0.983	0.995	0.973
$\varepsilon = 10^{-12}$				1.000	1.000	1.159			
$\varepsilon = 10^{-10}$				0.996	1.000	1.159			
$\varepsilon = 10^{-9}$				0.993	0.993	1.058			

Both linear systems with Hilbert and Boothroyd matrices fail for $\varepsilon = 10^{-12}$. Systems with Pascal matrix fail for $\varepsilon = 10^{-8}$. The different ratios in computing time come from the different number of iterations in step 2 of Algorithm B.

The table shows that as long as ε is not too large Algorithm B is a little bit faster than Algorithm A producing similar or even better inclusions. This changes for larger ε. The quality of the inclusions of Algorithm B can be improved to the same quality of those of Algorithm A but with the cost of some extra iterations in step 3.

The next table zooms the behaviour of both algorithms for large ε. We used Hilbert matrices, $n = 10$.

Table 2. Hilbert-matrices for large ε, $n = 10$

ε	d_1	d_2	t_A/t_B
$1.0 \cdot 10^{-13}$	0.951	0.983	1.014
$1.5 \cdot 10^{-13}$	0.977	0.982	0.973
$2.0 \cdot 10^{-13}$	0.938	0.946	0.936
$2.5 \cdot 10^{-13}$	0.885	0.892	0.901
$3.0 \cdot 10^{-13}$	0.759	0.770	0.785

Both algorithms fail for $\varepsilon = 3.5 \cdot 10^{-13}$. So for large diameters in the matrix elements Algorithm A performs better than Algorithm B. For the largest value of ε in table 2 Algorithm A is about 20% faster producing bounds with a 20 to 25% smaller diameter. It should be mentioned that the bounds itself are already of *very* large diameter. In this example, for $\varepsilon = 2.5 \cdot 10^{-13}$, the inclusion of the 7th component is $[-112.5, +114.5]$. There are examples as well where the behaviour of the algorithms is contrary. Consider linear systems with Pascal-matrices for $n = 15$.

Table 3. Pascal matrices for $n = 15$

ε	d_1	d_2	t_A/t_B
0	3.669	484.056	1.059
10^{-16}	1.004	2.299	1.000
10^{-15}	1.000	1.255	1.059
10^{-14}	1.000	1.025	1.000

Here the bounds produced by Algorithm B are always better, sometimes much better than those of Algorithm A requiring the same or less computing time.

For higher dimensions the extra computing time for Algorithm A vs. Algorithm B increases due to the extra $1/3 \cdot n^3$ operations. We display linear systems with matrix $[A] := A \diamondsuit (1 \pm \varepsilon)$ where A has random entries uniformly distributed in $[-1, 1]$, $[b] := [A] \diamondsuit (+1, -1, +1, -1, \ldots)^T$.

Table 4. Random matrices

	$\varepsilon = 10^{-5}$			$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-3}$		
n	d_1	d_2	t_A/t_B	d_1	d_2	t_A/t_B	d_1	d_2	t_A/t_B
20	1.000	1.000	1.163	1.000	1.000	1.135	1.000	1.000	1.135
50	0.999	1.000	1.165	1.000	1.000	1.154	0.988	0.988	1.100
100	0.999	1.000	1.166	1.000	1.000	1.160	0.990	0.990	1.116

Obviously Algorithm B is superior for small ε whereas Algorithm A shows its advantages for larger diameters of $[A]$. The diameter of $[b]$ plays no role at all. We therefore propose to combine both algorithms: If Algorithm B fails to obtain an inclusion after two or three iterations while the diameters of the potential inclusions increase slowly then switch to Algorithm A by computing $\tilde{u}$. This approach combines the advantages of both algorithms because for small diameters it saves computing time whereas the additional $n^3/3$ operations for Algorithm A are only invested if necessary. This approach computed very sharp bounds for the solution. The quality can be measured by the techniques of computing inner inclusions described in [16].

4. Standard Simplices

The special structure of hyperrectangles requires $|A|$ or $|\mathrm{Re}(A)| + |\mathrm{Im}(A)|$ to be convergent in order to allow $f(X^k) \subseteq \mathrm{int}(X^k)$ for some $k \in \mathbb{N}$ (see theorem 5). This is a necessary and sufficient condition. For general sets or general simplices, $f(X^k) \subseteq \mathrm{int}(X^k)$ is equivalent to $\rho(A) < 1$, $A \in S^{n \times n}$, $S \in \{\mathbb{R}, \mathbb{C}\}$. One might try to use other representations of sets in order to omit the assumption $\rho(|A|) < 1$ resp. $\rho(|Re(A)| + |\mathrm{Im}(A)|) < 1$. The representation should be simple enough to allow fast computation of $f(X^k)$ but "general" enough to cover as many matrices as possible.

One such representation is standard simplices:

$$S = \{s_0, \sigma_1, \ldots, \sigma_n\}$$
$$= \left\{ x \in \mathbb{R}^n \mid x = s_0 + \sum_{\nu=1}^{n} \lambda_\nu \sigma_\nu e_\nu, 0 \leq \lambda_\nu \in \mathbb{R}, \sum_{\nu=1}^{n} \lambda_\nu \leq 1 \right\}.$$

In [6, 7] Jansson gave an interesting geometrical approach for the construction of guaranteed error bounds for the solution of a system of linear equations $Ax = b$. For a given standard simplex S he gives a sufficient criterion for $b \in A \cdot S$ in the following way. The matrix S maps S into a general simplex, where the normal vectors

of the supporting hyperplanes are the rows of A^{-1}. Using an approximate inverse R of A he gives the following theorem, which, in some way, estimates the error of R w.r.t. A^{-1} and gives an *inner* estimation of $A \cdot S$. He shows that this estimation is optimal w.r.t. the information given by the approximations R and $\tilde{x}$. This optimality property holds for the general case of convex polyhedrons. It covers also the case of interval vectors.

Theorem 8 (Jansson). *Let $A, R \in \mathbb{R}^{n\times n}$, $C := R \cdot A$ and $b, \underline{x}, \varepsilon \in \mathbb{R}^n$ with $\varepsilon > 0$. If both*

$$R \cdot b > C \cdot \underline{x} + \operatorname{Max}\{(C - \operatorname{diag}(C)) \cdot \operatorname{diag}(\varepsilon)\} \qquad \textit{and} \tag{4.1}$$

$$(\varepsilon^{-1})^T \cdot Rb < (\varepsilon^{-1})^T \cdot C\underline{x} + \operatorname{Min}(\varepsilon^{-1})^T \cdot C \cdot \operatorname{diag}(\varepsilon) \tag{4.2}$$

are satisfied then R and A are nonsingular and the unique solution $\hat{x}$ of $Ax = b$ is contained in the standard simplex $S := \{\underline{x}, \varepsilon_1, \ldots, \varepsilon_n\}$.

Note. $\operatorname{diag}(C) \in \mathbb{R}^{n\times n}$ is the diagonal matrix consisting of the diagonal entries of C; $\operatorname{diag}(\varepsilon) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with $\varepsilon \in \mathbb{R}^n$ in the diagonal; for $M \in \mathbb{R}^{n\times n}$, $\max(M) \in \mathbb{R}^n$ is the column vector consisting of the maximum of the rows of M, and $\varepsilon^{-1} \in \mathbb{R}^n$ is the vector (ε_i^{-1}).

The approach by Jansson and the proof are based on geometrical considerations. It can be shown that with a technical assumption similar to the one used in the previous section this geometrical approach implies the fact that condition (1.1) in Theorem 1 is satisfied for $X = S$.

Theorem 9. *The assumptions* (4.1) *and* (4.2) *of Theorem 7 with R scaled s.t.* $\operatorname{diag}(R \cdot A) = I$ *are equivalent to*

$$R \cdot b + (I - RA) \cdot S \subseteq \operatorname{int}(S). \tag{4.3}$$

Remark. The operations in (4.3) are the power set operations.

Proof. "⇒" By definition $S = ch(\underline{x}, \underline{x} + \varepsilon_1 e_1, \ldots, \underline{x} + \varepsilon_n e_n)$ and therefore

$$\begin{aligned} x \in \operatorname{int}(S) \Leftrightarrow \text{a)}\ & x > \underline{x} \qquad \text{and} \\ \text{b)}\ & (\varepsilon^{-1})^T \cdot x < 1 + (\varepsilon^{-1})^T \cdot \underline{x} \end{aligned} \tag{4.4}$$

(cf. e.g. [6, 7]). We have proved (4.3) if we show conditions a) and b) of (4.4) to be valid for all vertices of $Rb + (I - RA) \cdot S$. By assumption $C^* := C - \operatorname{diag}(C) = RA - I$ and $\operatorname{diag}(RA - I) = 0$. By definition

$$\operatorname{Max}\{C^* \cdot \operatorname{diag}(\varepsilon)\} \geq \{C^* \cdot \operatorname{diag}(\varepsilon)\}_i = C^* \varepsilon_i e_i \tag{4.5}$$

for $1 \leq i \leq n$ and thus (4.1) implies

$$Rb + (I - RA)\underline{x} > \underline{x} + C^* \varepsilon_i e_i \tag{4.6}$$

showing condition a) of (4.4) for the vertices $\underline{x} + \varepsilon_i e_i$. (4.6) holds true for every $1 \leq i \leq n$ and with $(C^* \varepsilon_i e_i)_i = 0$ it follows

$$Rb + (I - RA)\underline{x} > \underline{x}$$

showing condition a) of (4.4) for the vertex $\underline{x}$. Furthermore,

$$(\varepsilon^{-1})^T \cdot (Rb + (I - RA)(\underline{x} + \varepsilon_i e_i)) < 1 + (\varepsilon^{-1})^T \cdot \underline{x} \tag{4.7}$$

$$\Leftrightarrow (\varepsilon^{-1})^T \cdot (Rb - C \cdot (\underline{x} + \varepsilon_i e_i)) < 1 - (\varepsilon^{-1})^T \cdot \varepsilon_i e_i . \tag{4.8}$$

The r.h.s. of (4.8) equals 0 implying

$$(4.7) \Leftrightarrow (\varepsilon^{-1})^T \cdot R \cdot (b - A\underline{x}) < (\varepsilon^{-1})^T \cdot C \cdot \varepsilon_i e_i . \tag{4.9}$$

The r.h.s. of (4.9) follows by (4.2) implying the validity of condition b) of (4.4) for the vertices $\underline{x} + \varepsilon_i e_i$. By assumption $\operatorname{Max}\{(\varepsilon^{-1})^T \cdot C \cdot \varepsilon_i e_i\} \geq 1$ for $1 \leq i \leq n$, hence $\operatorname{Max}\{(\varepsilon^{-1})^T \cdot (I - RA)\varepsilon_i e_i\} \leq 0$ and (4.7) implies

$$(\varepsilon^{-1})^T \cdot (Rb + (I - RA)\underline{x}) < 1 + (\varepsilon^{-1})^T \cdot \underline{x}$$

which finishes the first part of the proof.

"$\Leftarrow$" (4.3) together with (4.4), a) implies $Rb + (I - RA)(\underline{x} + \varepsilon_i e_i) \geq \underline{x}$ for all $1 \leq i \leq n$ and therefore (4.1) follows by using (4.5). (4.3) together with (4.4), b) imply (4.7) and therefore, following the first part of the proof, (4.9) holds for all $1 \leq i \leq n$. Hence (4.2) is true finishing the proof. □

It is well known that (4.3) has the quadratic approximation property (see e.g. [11]). By the previous Theorem 9 and the results of Jansson it follows that (4.3) is optimal in the described geometrical sense.

The following examples will show that w.r.t. the inclusion methods described in [12, 13] standard simplices play a special role.

There are real matrices A which are convergent with $\rho(|A|) \geq 1$ and mapping some standard simplex into itself. On the other hand, there are matrices A the absolute value of which is convergent but A maps *no* standard simplex at all into itself. Consider the case $n = 2$ and a standard simplex $s = \{(a, b)^T, c, d\}$. Then according to (4.4) $A \cdot S \subseteq S$ is equivalent to

$$(a, b)^T \leq A \cdot v_v \text{ and } \frac{x - a}{c} + \frac{y - b}{d} \leq 1 \qquad \text{for } (x, y)^T = A \cdot v_v,\ v = 1, 2, 3$$

and (4.10)

$$v_1 = (a, b)^T, \qquad v_2 = (a + c, b)^T, \qquad v_3 = (a, b + d)^T .$$

As a first example consider

$$A = \begin{pmatrix} 0.9 & -0.05 \\ -0.9 & -0.8 \end{pmatrix}.$$

The eigenvalues of A are $0.05 \pm \sqrt{0.7675}$, those of $|A|$ are $0.85 \pm \sqrt{0.0475}$ implying $\rho(A) < 1 < \rho(|A|)$. However, short computation yields that the standard simplex $S = \{(-2.7, -1.15), 4.9, 5\}$ produces

$$A \cdot v_1 = \begin{pmatrix} -2.3725 \\ 3.35 \end{pmatrix}, \qquad A \cdot v_2 = \begin{pmatrix} 2.0375 \\ -1.06 \end{pmatrix}, \qquad A \cdot v_3 = \begin{pmatrix} -2.6225 \\ -0.65 \end{pmatrix}$$

and satisfies condition (4.10), i.e. $A \cdot S \subseteq S$, in fact $A \cdot S \subseteq \operatorname{int}(S)$.

As a second example consider

$$A = \begin{pmatrix} 0.5 & -0.5 \\ 0.25 & 0.5 \end{pmatrix}.$$

The eigenvalues of $|A|$ are $0.5 \pm \sqrt{0.125}$ implying $\rho(A) \leq \rho(|A|) < 1$.

Let a standard simplex $s = \{(a, b)^T, c, d\}$, $c \neq 0$, $d \neq 0$ be given. Then

$$A \cdot v_1 = \begin{pmatrix} 0.5 \cdot (a - b) \\ 0.25a + 0.5b \end{pmatrix}; \qquad A \cdot v_2 = \begin{pmatrix} 0.5(a + c - b) \\ 0.25(a + c) + 0.5b \end{pmatrix};$$

$$A \cdot v_3 = \begin{pmatrix} 0.5 \cdot (a - b - d) \\ 0.25a + 0.5(b + d) \end{pmatrix}.$$

Assuming (4.10) implies

$$(a, b)^T \leq A \cdot v_3 \Rightarrow a \leq -b - d \qquad \text{and} \tag{4.11}$$

$$(a, b)^T \leq A \cdot v_1 \Rightarrow b \leq 0.5a. \tag{4.12}$$

The condition

$$\frac{x - a}{c} + \frac{y - b}{d} \leq 1 \text{ for } (x, y)^T = A \cdot v_\nu,\ \nu = 1, 2, 3$$

implies for $(x, y)^T = A \cdot v_2$:

$$1 \geq \frac{-a + c - b}{2c} + \frac{a + c - 2b}{4d} \overset{(4.11)}{\geq} \frac{c + d}{2c} + \frac{a + c - 2b}{4d} \overset{(4.12)}{\geq} \frac{c + d}{2c} + \frac{c}{4d} \tag{4.13}$$

Consider the function $f(c, d) = \dfrac{c + d}{2c} + \dfrac{c}{4d}$. The partial derivatives are

$$\frac{\partial f}{\partial c} = -\frac{\partial f}{\partial d} = \frac{c^2 - 2d^2}{4c^2 d}.$$

For $c \neq 0 \neq d$ an extremum of f implies $c = \sqrt{2} \cdot d$ with

$$f(\sqrt{2}d, d) = \frac{(\sqrt{2} + 1)d}{2\sqrt{2}d} + \frac{\sqrt{2}d}{4d} = \frac{1}{2} + \frac{1}{\sqrt{2}} > 1.$$

Since this extreme value is obviously a minimum there is a contradiction to (4.13) A short computation implies immediately that $c = 0$ or $d = 0$ forces $a = b = c = d = 0$, the trivial case.

In other words $A \cdot S \subseteq S$ is, except for the trivial case, not possible although $\rho(|A|) < 1$. That means an iteration (1.4) using hyperrectangles will stop for *any* starting set X^0 whereas *no* standard simplex is mapped into itself by the matrix A. This behaviour becomes clear when looking at the eigenvectors which are $(1, -\sqrt{2}/2)^T$ and $(1, \sqrt{2}/2)^T$.

There might be other representations of sets being suitable for numerical computations and allowing to verify convergence of A even if $\rho(|A|) \geq 1$. At least the standard simplices do not seem to be suitable for general matrices.

5. Conclusion

A constructive method has been given for proving convergence of an interval matrix resp. its absolute value by means of an iteration. It has been shown that the iteration stops if and only if the absolute value of the matrix resp. the sum of absolute values of real and imaginary part is convergent. The criterion is applicable on digital computers with the cost of n^2 operations per iteration step.

The criterion is especially useful in combination with socalled verification algorithms (see [13]) for linear and nonlinear systems of equations and other standard problems in numerical analysis.

For the application to inclusion methods (see [12, 13]) being described for the case of linear systems in theorem 1 this means the following.

The iteration scheme (1.2) is exactly of the form used in Theorems 8 and 9. Therefore an inclusion of the solution of the linear system with matrix $[A]$ and right hand side $[b]$ will be computed

for general sets $X \in \mathbb{P}S^n$ if and only if $\rho(C) < 1$ and

for hyperrectangles $[X] \in \mathbb{IR}^n$ if and only if $\rho(|[C]|) < 1$,

for hyperrectangles $[X] \in \mathbb{IC}^n$ if and only if $\rho(|\mathrm{Re}([C])| + |\mathrm{Im}([C])|) < 1$

where $[C] := I \diamondsuit R \diamondsuit [A]$. In the first case power set operations, in the latter two cases interval operations $\circledast$ for $* \in \{+, -, \cdot, /\}$ are used.

An algorithm based on such an iteration scheme for validated calculation of an inclusion of $\Sigma([A], [b])$ becomes slow then the diameters of $[A]$ are very large. Therefore a combination with a modification of an algorithm proposed by Neumaier has been suggested working very good for small *and* for large diameters of $[A]$.

References

[1] ACRITH High-Accuracy Arithmetic Subroutine Library: General Information Manual, IBM Publications, GC33-6163 (1985).
[2] Alefeld, G., Herzberger, J.: Introduction to interval computations. New York: Academic Press 1983.
[3] Hansen, E.: Interval arithmetic in matrix computations, Part I. SIAM J. Numer. Anal. *2*, 308–320 (1965).
[4] Hansen, E.: Interval arithmetic in matrix computations, Part II. SIAM J. Numer. Anal. *4*, 1–9 (1967).
[5] IEEE 754 Standard for Floating-Point Arithmetic (1986).
[6] Jansson, C.: A geometric approach for computing a posteriori error bounds for the solution of a linear system. Computing *47*, 1–9 (1991).

[7] Jansson, C.: Guaranteed error bounds for the solution of linear systems, Contributions to Computer Arithmetic and Self-Validating Numerical Methods (C. Ullrich editor), J. C. Baltzer AG, Scientific Publishing Co. IMACS, S. 103–110 (1990).
[8] Krawczyk, R.: Newton-Algorithmen zur Besimmung von Nullstellen mit Fehlerschranken. Computing *4*, 187–220 (1969).
[9] Kulisch, U., Miranker, W. L.: Computer arithmetic in theory and practice. New York: Academic Press 1981.
[10] Moore, R. E.: Interval analysis. Englewood Cliffs, New Jersey: Prentice Hall 1966.
[11] Neumaier, A.: Interval methods for systems of equations. Cambridge University Press (1990).
[12] Rump, S. M.: Kleine Fehlerschranken bei Matrixproblemen, Dissertation, Universität Karlsruhe (1980).
[13] Rump, S. M.: New results on verified inclusions, in: Miranker, W. L., R. Toupin (eds.): Accurate scientific computations. Springer Lecture Notes in Computer Science *235*, 31–69 (1986).
[14] Siemens AG: Arithmos (BS2000). Benutzerhandbuch, (1986).
[15] Varga, R. S.: Matrix iterative analysis. Englewood Cliffs, New Jersey: Prentice Hall 1962.
[16] Rump, S. M.: Rigorous sensitivity analysis for systems of linear and nonlinear equations. MATH. of Comp. *54*, *(190)*, 721–736 (1990).

S. M. Rump
Technische Informatik III
TU Hamburg-Harburg
Eissendorferstraße 38
D-W-2100 Hamburg 90
Federal Republic of Germany