References:

/1/ Apostolatos, N., Kulisch, U., Krawczyk, R., Lortz, B., Nickel, K., Wippermann, H.-W.: The Algorithmic Language TRIPLEX ALGOL 60, Num. Math. 11, 175-180 (1968)

/2/ Bohlender, G.: Floating-point Computation of functions with maximum accuracy. IEEE Trans. Comp. C-26, Nr. 7, 621-632 (1977)

/3/ Bohlender, G., Kaucher, E., Klatte, R., Kulisch, U., Miranker, W.L., Ullrich, Ch. und Wolff von Gudenberg, J.: FORTRAN for contemporary numerical computation, Report RC 8348, IBM Thomas J. Watson Research Center 1980 and Computing 26, 277-314 (1981)

/4/ Kulisch, U.: An axiomatic approach to rounded computations, TS Report No. 1020. Mathematics Research Center, University of Wisconsin, Madison, Wisconsin, 1969 und Numer. Math. 19, 1-17 (1971)

/5/ Kulisch, U.: Interval arithmetic over completely ordered ringoids, TS Report No. 1105, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin, 1970

/6/ Kulisch, U.: Grundlagen des Numerischen Rechnens - Mathematische Begründung der Rechnerarithmetik. Reihe Informatik, Band 19, Wissenschaftsverlag des Bibliographischen Instituts Mannheim, 1976

/7/ Kulisch, U., Miranker, W.L.: Computer Arithmetic in Theory and Practice, Academic Press, 1980

/8/ Coonan, J. et al.: A proposed standard for floating-point arithmetic, SIGNUM newsletter, Oct. 1979

/9/ INTEL 12 1586-001: The 8086 family user's manual, Numeric Supplement, July 1980

/10/ Kulisch, U., Miranker, W.L. (Editors): A New Approach to Scientific Computation, Academic Press, 1983

/11/ Kulisch, U., Miranker, W.L.: The Arithmetic of the Digital Computer, IBM Research Report RC 10580, 1984, to appear in SIAM Reviews

/12/ High Accuracy Arithmetic, Subroutine Library, IBM Program Description and User's Guide, Program Number 5664-185, 1984

/13/ Böhm, H.: Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrücke mit garantierter maximaler Genauigkeit. Dissertation, Universität Karlsruhe 1984

Additional References are given in /7/, /10/ and /11/

New Results on Verified Inclusions

Siegfried M. Rump

Abstract. The computational results of traditional numerical algorithms on computers are usually good approximations to the solution of a given problem. However, no verification is provided for some bound on the maximum relative error of the approximation. As can be demonstrated by ill-conditioned examples, those approximations may be drastically wrong. The algorithms based on the inclusion theory (cf.[38]) do have an automatic verification process. Rather than approximations to the solution an inclusion of the solution is computed and the correctness of the bounds together with the existence and uniqueness of the solution within the bounds is automatically verified by the computer without any effort on the part of the user. The computing time of these algorithms is of the order of a comparable, standard floating-point algorithm (such as Gaussian elimination in case of general linear systems).

In the following some new results complementing the inclusion theory are given. One of the main results is that the inclusion sets need not to be convex. Therefore other types of inclusion sets such as torus-sectors can be used. Another main observation is that the new and old theorems can be proved without using fixed point theorems. Moreover improvements of existing theorems of the inclusion theory by means of weaker assumptions are presented.
Another fundamental observation is the following. It is well-known that a real iteration in $\mathbb{R}^n$ with affine iteration function converges if and only if the spectral radius of the iteration matrix is less than one. It can be shown that a similar result holds for our inclusion algorithm: An inclusion will be achieved if and only if the spectral radius of the iteration matrix is less then one. This result is best possible.

It is demonstrated by means of theorems and examples that even for extremely ill-conditioned examples very sharp inclusions of the solution are computed. The inclusions are almost always of least significant bit accuracy, i.e. the left and right bounds of the inclusion are adjacent floating-point numbers.

1. Introduction. Let $\mathbb{R}$ be the set of real numbers, $V\mathbb{R}$ the set of real vectors with n components, $M\mathbb{R}$ the set of real square matrices with n rows, $\mathbb{C}$ the set of complex numbers, $V\mathbb{C}$ the set of complex vectors with n components and $M\mathbb{C}$ the set of complex square matrices with n rows. In the following the letter n is reserved to denote the number of elements of a vector or the number of rows and columns of a square matrix. Vectors with other than n components are, for instance, denoted by $V_{n+1}\mathbb{R}$, non-square matrices for example with ℓ rows and m columns over the complex numbers by $M_{\ell,m}\mathbb{C}$. I denotes the identity matrix.

The operations in the power set PT over T for $T \in \{\mathbb{R}, V\mathbb{R}, M\mathbb{R}, \mathbb{C}, V\mathbb{C}, M\mathbb{C}\}$ are as usual defined by

$$A \in PT,\ B \in PT: \quad A * B := \{a * b \mid a \in A,\ b \in B\} \tag{1}$$

for $* \in \{+,-,\cdot,/\}$ and well-known restrictions for /. Definition (1) applies for inner and outer operations. In case a set (an element of the power set) occurs more than once in a formula special care has to be taken. Consider as an example $f : PV\mathbb{R} \to PV\mathbb{R}$ being defined by $f(x) := Z + \mathbb{C} \cdot X$ for $X, Z \in PV\mathbb{R}$ and $\mathbb{C} \in PM\mathbb{R}$. Then

$$f(f(X)) = Z_1 + \mathbb{C}_1 \cdot (Z_2 + \mathbb{C}_2 \cdot X) \quad \text{for} \quad Z_1 = Z_2 = Z \quad \text{and} \quad \mathbb{C}_1 = \mathbb{C}_2 = \mathbb{C}.$$

Moreover

$$\{z + C \cdot (z + C \cdot x) \mid z \in Z,\ C \in \mathbb{C},\ x \in X\} \subseteq f(f(x)) \tag{2}$$

where in general eguality does not hold in (2).

The order relation $\leq$ in $\mathbb{R}$ is extended to $V\mathbb{R}$ and $M\mathbb{R}$ by

$$A, B \in V\mathbb{R}: \quad A \leq B :\Leftrightarrow A_i \leq B_i \quad \text{for} \quad 1 \leq i \leq n \quad \text{and}$$
$$A, B \in M\mathbb{R}: \quad A \leq B :\Leftrightarrow A_{ij} \leq B_{ij} \quad \text{for} \quad 1 \leq i,j \leq n.$$

The order relation in $\mathbb{C}$ is defined by

$$a + bi,\ c + di \in \mathbb{C}: \quad a + bi \leq c + di :\Leftrightarrow a \leq c \quad b \leq d$$

and similarly in VC and MC.

The set of intervals $\mathbb{I}T$ over T for $T \in \{\mathbb{R}, V\mathbb{R}, M\mathbb{R}, \mathbb{C}, VC, MC\}$ is defined by

$$[A,B] \in \mathbb{I}T :\Leftrightarrow [A,B] := \{x \in T \mid A \leq x \leq B\} \quad \text{for} \quad A, B \in T \quad \text{and} \quad A \leq B.$$

Therefore $\mathbb{I}T \subseteq PT$ where every element of $\mathbb{I}T$ is nonempty, convex, closed and bounded.

By S we denote some finite set of real numbers (which could be regarded as the set of single precision floating-point numbers on some computer). We consider the set of n-tupels VS over S and the set of n^2-tupels MS over S. Similarly for some finite subset CS of $\mathbb{R} \times \mathbb{R}$ we consider VCS and MCS. If U is some set out of S, VS, MS, CS, VCS, MCS and T is the corresponding set in $\mathbb{R}$, $V\mathbb{R}$, $M\mathbb{R}$, $\mathbb{C}$, VC, MC, then intervals over U are defined by

$$[A,B] \in \mathbb{I}U :\Leftrightarrow [A,B] := \{x \in T \quad A \leq x \leq B\} \quad \text{for} \quad A, B \in U \quad \text{and} \quad A \leq B.$$

For example, an interval $[a,b] \in S$ consists of all real numbers x with $a \leq x \leq b$. Usually, interval operations $\diamondsuit$, $* \in \{+, -, \cdot, /\}$ are defined by

$$A, B \in \mathbb{I}T: \quad A \diamondsuit B := \cap \{M \in \mathbb{I}T \mid A * B \in M\} \tag{3}$$

and

$$A, B \in \mathbb{I}U: \quad A \diamondsuit B := \cap \{M \in \mathbb{I}U \mid A * B \in M\} \tag{4}$$

for T and U as above. Using (3) and (4) the operations $\diamondsuit$ are well defined (cf. [4],[28]). Sometimes the strict definition (4) is difficult to realize on computers, e.g. for complex division. Therefore we allow in our further discussions any isotone interval operation $\circledast$, i.e. any operation

$$\circledast : PT \times PT \to \mathbb{I}T \quad \text{with} \quad A, B \in PT \Rightarrow A * B \subseteq A \circledast B \text{ respectively} \tag{5}$$
$$\circledast : PU \times PU \to \mathbb{I}U \quad \text{with} \quad A, B \in PU \Rightarrow A * B \subseteq A \circledast B. \tag{6}$$

There will no confusion be caused by the fact that in $\mathbb{I}T$ and $\mathbb{I}U$ the same symbol $\circledast$ is used. In (5) and (6) elements of the power set over T or U are allowed to define more general interval operations. As follows by (3) and (4), $\diamondsuit$ is best possible for interval arguments. We allow any interval rounding $\square : PT \to \mathbb{I}T$ resp. $\square : PU \to \mathbb{I}U$ satisfying

$$A \in PT: \quad A \subseteq \square A \quad \text{and} \quad A \in PU: \quad A \subseteq \square A.$$

Here $\diamond : \mathbb{PT} \to \mathbb{IT}$ resp. $\diamond : \mathbb{PU} \to \mathbb{IU}$ has to be regarded as an operator rather than a mapping. That means we do not necessarily require

$$\diamond A = \diamond B \quad \text{for} \quad A = B.$$

Instead the essential property for $\diamond$ is $A \subseteq \diamond A$ for $A \in \mathbb{PT}$ resp. $A \in \mathbb{PU}$. The reason for this is that in practical implementations of $\diamond$ it might happen that $\diamond A \neq \diamond B$ for $A = B$ because of performance reasons. An overestimation of A may be allowed to save computing time, but this does not affect the following theorems.

There is always a best rounding $\diamond$ (cf.[4]) satisfying

$$A \in \mathbb{PT} : \diamond A = \cap\{B \in \mathbb{IT} : A \subseteq B\} \quad \text{and}$$

$$A \in \mathbb{PU} : \diamond A = \cap\{B \in \mathbb{IU} : A \subseteq B\}.$$

The rounding $\diamond$ and the operations $\circledast$ over $\mathbb{IU}$ for $* \in \{+,-,\cdot,/\}$ and $U \in \{S, VS, MS, CS, VCS, MCS\}$ can be effectively implemented on computers (cf.[25]).

For an interval $X = [A,B]$ the lower and upper bound is defined by $\inf(X) := A$ and $\sup(X) := B$, or, in short notation $\underline{X} := A$ and $\overline{X} := B$ (where is no confusion with the algebraic closure).

In the following we use beside the ordinary inclusion $\subseteq$ three other types of inclusions:

$$A,B \in \mathbb{PVR}: \quad A \overset{\circ}{\subseteq} B :\Leftrightarrow A \subseteq \overset{\circ}{B},$$

$$A,B \in \mathbb{IIVR}: \quad A \subsetneq B :\Leftrightarrow A \subseteq B \wedge A_i \neq B_i \text{ for all } i,\ 1 \leq i \leq n,$$

$$A,B \in \mathbb{IIVR}: \quad A \overset{\circ}{\subseteq} B :\Leftrightarrow A \subseteq B \wedge A_i \neq B_i \text{ for some } i,\ 1 \leq i \leq n.$$

The definitions apply similarly over S. It is

$$A,B \in \mathbb{IIVR}: \quad A \overset{\circ}{\subseteq} B \Rightarrow A \subsetneq B \Rightarrow A \overset{\circ}{\subseteq} B.$$

For any algorithm performing verification of the results on a computer a precisely defined computer arithmetic is mandatory. Preferably this computer arithmetic should deliver highly accurate or maximum accurate results. An arithmetic delivering always results of least significant bit accuracy even for matrix and vector operations has been given in [8].

Of course, composed operations need not to be of high accuracy. Such highly accurate results for composed operations and even for whole algorithms for solving linear or nonlinear systems are delivered by the inclusion theory as described in the following.

2. Basic new results. In [41] one of the main steps to develop the inclusion theory was to apply Brouwer's Fixed Point Theorem to an affine mapping and to show contraction in order to obtain a zero from the fixed point. The following theorem gives Brouwer's Fixed Point Theorem for affine mappings, the elegant proof of which the author learnt at the University of Karlsruhe.

Theorem 1. Let X be a nonempty, compact, convex subset of the normed vector space E, $C : E \to E$ a linear transformation with continuous restriction $C/_X$, $z \in E$ and $T(X) \subseteq X$ for the affine mapping $Tx := z + Cx$. Then there is a fixed point $\hat{x} \in X$ of $T : T\hat{x} = \hat{x}$.

Proof. Let $x^0 \in X$ and define

$$x^m := \frac{1}{m+1} \cdot \{x^0 + Tx^0 + \ldots + T^m x^0\}, \quad m \geq 1.$$

Then short computation shows

$$Tx^m = \frac{1}{m+1} \cdot \{Tx^0 + T^2 x^0 + \ldots + T^{m+1} x^0\}$$

and by the convexity of X follows $Tx^m \in X$. Therefore

$$Tx^m - x^m = \frac{1}{m+1} \cdot (T^{m+1} x^0 - x^0)$$

and because X is bounded $Tx^m - x^m \to 0$ for $m \to \infty$. Therefore some subsequence (x^{m_n}) of (x^m) converges to some $\hat{x} \in X$ with $T\hat{x} = \hat{x}$. □

To come from a fixed point of the affine mapping to a zero of the original problem (e.g. a system of linear or nonlinear equations) in [37] the contraction of the linear part is shown by assuming a proper inclusion $TX \overset{\circ}{\subseteq} X$ rather than $TX \subseteq X$. Next we will show this contraction for general compact subsets of $\mathbb{R}^n$ without assuming convexity.

Lemma 2. Let $Z \in \mathbb{PVR}$, $\mathbb{C} \in \mathbb{PMR}$ and $\emptyset \neq X \in \mathbb{PVR}$ compact. Then

$$\begin{aligned} &\text{a)} \quad Y = \mathrm{ch}(X) \text{ and } Z + \mathbb{C} \cdot X \overset{\circ}{\subseteq} Y \quad \Rightarrow \quad Z + \mathbb{C} \cdot Y \overset{\circ}{\subseteq} Y. \\ &\text{b)} \quad Y = X - X \text{ and } Z + \mathbb{C} \cdot X \overset{\circ}{\subseteq} X \quad \Rightarrow \quad \mathbb{C} \cdot Y \overset{\circ}{\subseteq} Y. \end{aligned} \tag{7}$$

Remark. All operations are the power set operations, ch(X) denotes the convex hull.

Proof. ad a): Let $z \in Z$ and $C \in \mathbb{C}$. Then $y \in Y \Leftrightarrow \exists x_1, x_2 \in X : y = x_1 + \lambda(x_2 - x_1)$ for $0 \leq \lambda \leq 1$. Then with $x_3 := z + C \cdot x_1$ and $x_4 := z + C \cdot x_2$ follows $z + C \cdot y = x_3 + \lambda(x_4 - x_3) \overset{\circ}{\subseteq} Y$ by $x_3, x_4 \overset{\circ}{\in} Y$ and (7).

ad b): For $z \in Z$, $C \in \mathbb{C}$ is $y \in X - X \Leftrightarrow \exists x_1, x_2 \in X : y = x_2 - x_1$. Therefore $C \cdot y = (z + C \cdot x_1) - (z + C \cdot x_2) \overset{\circ}{\subseteq} Y$. □

The proof in [40] of assertion b) in lemma 2 for convex, compact sets required Brouwer's Fixed Point Theorem by $z + C \cdot \hat{x} = \hat{x} \Leftrightarrow C \cdot (X - \hat{x}) = C \cdot X + z - \hat{x} \subseteq X - \hat{x}$. The assumption of proper inclusion $\overset{\circ}{\subseteq}$ can be sharpened for interval vectors. For this purpose we need the following lemma.

Lemma 3. Let $Z \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\,\mathbb{M}\mathbb{R}$ and $X \in \mathbb{I}\,\mathbb{V}\mathbb{R}$. Then for $Y := X \diamondsuit X$ holds

a) $\quad 0(Z + \mathbb{C} \cdot X) \overset{\circ}{\subseteq} X \quad \Rightarrow \quad \Diamond(\mathbb{C} \cdot Y) \overset{\circ}{\subseteq} Y$

b) $\quad 0(Z + \mathbb{C} \cdot X) \subsetneq X \quad \Rightarrow \quad \Diamond(\mathbb{C} \cdot Y) \overset{\circ}{\subseteq} Y$

c) $\quad 0(Z + \mathbb{C} \cdot X) \subseteq X \quad \Rightarrow \quad \Diamond(\mathbb{C} \cdot Y) \subseteq Y$.

Proof. ad a) and b): Suppose $0(Z + \mathbb{C} \cdot X) \subsetneq X$ and $z \in Z$, $C \in \mathbb{C}$. Obviously $Y = [\underline{X} - \overline{X}, \overline{X} - \underline{X}]$ and $(\Diamond(C \cdot Y))_i = [\underline{\sum}_{j=1}^{n} |C_{ij}| \cdot (\underline{X}_j - \overline{X}_j), \overline{\sum}_{j=1}^{n} |C_{ij}| \cdot (\overline{X}_j - \underline{X}_j)]$. Therefore the proof is finished if $\overline{\sum}_{j=1}^{n} |C_{ij}| \cdot (\overline{X}_j - \underline{X}_j) < \overline{X}_i - \underline{X}_i$ for $1 \leq i \leq n$ holds. Define

$$\overline{y}_j := \begin{cases} \overline{X}_j & \text{if } C_{ij} \geq 0 \\ \underline{X}_j & \text{otherwise} \end{cases} \quad \text{and} \quad \underline{y}_j := \begin{cases} \underline{X}_j & \text{if } C_{ij} \geq 0 \\ \overline{X}_j & \text{otherwise} \end{cases} \tag{8}$$

for some fixed $1 \leq i \leq n$. Then by assumption

$$\underline{X}_i \leq z_i + \underline{\sum}_{j=1}^{n} C_{ij} \overline{y}_j \leq \overline{X}_i \quad \text{and} \quad \underline{X}_i \leq z_i + \underline{\sum}_{j=1}^{n} C_{ij} \underline{y}_j \leq \overline{X}_i , \tag{9}$$

where either both left or both right $\leq$ can be replaced by $<$. The left of the second inequality in (9) implies

$$-z_i - \underline{\sum}_{j=1}^{n} C_{ij} \cdot \underline{y}_j \leq -\underline{X}_i \tag{10}$$

and adding the right of the first inequality of (9) and (10) yields

$$\underline{\sum}_{j=1}^{n} C_{ij} \cdot (\overline{y}_j - \underline{y}_j) < \overline{X}_i - \underline{X}_i \tag{11}$$

because one of both $\leq$ can be replaced by $<$. By definition (8) follows $C_{ij} \cdot (\overline{y}_j - \underline{y}_j) = |C_{ij}| \cdot (\overline{X}_j - \underline{X}_j)$ and therefore demonstrating assertion b) using (11). Assertion a) follows by b).

ad c): Follows by the preceding proof for cases a) and b). □

Consider a system of linear equations $Ax = b$. For an approximate inverse R of A lemma 3 can be applied to the residual function $f(x) := R \cdot b + (I - RA) \cdot x$. If $f(X) \overset{\circ}{\subseteq} X$ for some $X \in \mathbb{I}\,\mathbb{V}\mathbb{R}$, then by Brouwer's Fixed Point Theorem there is a $\hat{x} \in X$ with $f(\hat{x}) = \hat{x} = Rb + (I - RA)\hat{x}$ and therefore $b - A\hat{x} \in \ker R$. If the non-singularity of R could be shown, then the existence of a zero $\hat{x}$ of $Ax - b = 0$ would be demonstrated. In the following we give criteria for the regularity of R and A and show the existence of a fixed point of f and a zero of $Ax - b = 0$ without using Brouwer's Fixed Point Theorem.

Lemma 4. Let $\mathbb{C} \in \mathbb{P}\,\mathbb{M}\mathbb{R}$ and $\emptyset \neq X \in \mathbb{P}\,\mathbb{V}\mathbb{R}$ compact with $0 \in X$. If

$$\mathbb{C} \cdot X \overset{\circ}{\subseteq} X \tag{12}$$

then $\forall C \in \mathbb{C} : \rho(C) < 1$.

Proof. Let $C \in \mathbb{C}$. For $Y := X + iX$ holds $\mathbb{C} \cdot Y \overset{\circ}{\subseteq} Y$ and $0 \overset{\circ}{\in} Y$. Suppose $C \neq (0)$ and let $\lambda \in \mathbb{C}$, $x \in \mathbb{V}C$ be an eigenvalue/eigenvector pair of C. Define

$$T \in \mathbb{P}\mathbb{C} \quad \text{by} \quad T := \{\gamma \in \mathbb{C} \mid \gamma \cdot x \in Y\}.$$

T is not empty because $0 \in Y$ and T is closed and bounded because Y is compact. Therefore there is some $\gamma^* \in C$ with

$$|\gamma^*| = \max_{\gamma \in T} |\gamma| .$$

Then by (12) we have $\gamma^* \neq 0$ and

$$C \cdot (\gamma^* x) = (\gamma^* \lambda) x \in \overset{\circ}{Y}.$$

Because $\gamma^* x \in \partial Y$ and the definition of γ^* follows $|\gamma^*| > |\gamma^* \lambda|$ and therefore $|\lambda| < 1$. □

With these preperations we get the following theorem.

Theorem 5. Let $Z \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\,\mathbb{M}\mathbb{R}$ and $\emptyset \neq X \in \mathbb{P}\,\mathbb{V}\mathbb{R}$ compact. If

$$Z + \mathbb{C}\cdot X \overset{\circ}{\subseteq} X \tag{13}$$

then $\forall C \in \mathbb{C}: \rho(C) < 1$.

Proof. Follows by lemma 2, b) and lemma 4. □

If the inclusion sets are intervals, then weaker conditions suffice to prove a stronger assertion than stated in theorem 5.

Theorem 6. Let $Z \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\,\mathbb{M}\mathbb{R}$ and $X \in \mathbb{I}\,\mathbb{V}\mathbb{R}$. If

$$\square(Z + \mathbb{C}\cdot X) \subsetneq_{*} X \tag{14}$$

then $\forall C \in \mathbb{C}: \rho(C) \leq \rho(|C|) < 1$.

Proof. By lemma 3, b) we have $\mathbb{C}\cdot Y \subseteq \diamondsuit(\mathbb{C}\cdot Y) \overset{\circ}{\subseteq} Y$ for $Y = X \diamondsuit X$. If the i-th component of X is $[\underline{X}_i, \overline{X}_i]$, then $Y_i = [\underline{X}_i - \overline{X}_i, \overline{X}_i - \underline{X}_i]$. Therefore $Y_i = -Y_i$ for $1 \leq i \leq n$ and $|\mathbb{C}|\cdot Y = \mathbb{C}\cdot Y \overset{\circ}{\subseteq} Y$ and the assertion follows by lemma 4 and Perron-Frobenius Theory (cf. [45]). □

Given a general set of vectors $x \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, for $Y = X - X$ holds of course $Y = -Y$, but something similar to $Y_i = -Y_i$ is, in general, not true. In theorem 6 with the weaker assumption $\subsetneq_{*}$ instead of $\overset{\circ}{\subseteq}$ in theorem 5, the stronger assertion holds that the spectral radius of the absolute value of every matrix $C \in \mathbb{C}$ is less than one. This follows from the special structure of the elements of $\mathbb{I}\,\mathbb{V}\mathbb{R}$. Therefore the fact, that operations in $\mathbb{I}\,\mathbb{V}\,\mathbb{S}$ are easy to implement and are fast, has to be paid by the impossibility to show $\rho(|C|) < 1$ where $\rho(|C|) \geq 1$ using (14). With spending more computing time for operations in other structures this deficiency can be eliminated. However, in practical cases the matrices C are of the form $I - RA$ where $R \approx A^{-1}$. In our experience never a case occurred where $\rho(I-RA) < 1$ and $\rho(|I-RA|) \geq 1$. On the other hand there are cases where $\|C\| \geq 1$ for the common norms but $X^{k+1} := Z \diamondsuit C \diamondsuit X^k \overset{\circ}{\subseteq} X^k$ holds true for $k = 2$ or $k = 3$. Later some examples are shown demonstrating this fact.

With the assumptions of theorem 5 $\rho(|C|) < 1$ for all $C \in \mathbb{C}$ is, in general, not true. A counterexample is:

$$C := \begin{Bmatrix} 1 & -0.25 \\ 3 & -1 \end{Bmatrix} \quad \text{with } \rho(C) = 0.5 \text{ and } \rho(|C|) = 1 + \sqrt{0.75} > 1.8 .$$

Define $$X := \{\lambda\cdot\begin{bmatrix}1\\2\end{bmatrix} + \mu\begin{bmatrix}1\\6\end{bmatrix} \mid -1 \leq \lambda,\mu \leq 1\} \in \mathbb{P}\,\mathbb{V}\mathbb{R} .$$

Then $$C\cdot X = \{\lambda\cdot\begin{bmatrix}0.5\\1\end{bmatrix} + \mu\begin{bmatrix}-0.5\\-3\end{bmatrix} \mid -1 \leq \lambda,\mu \leq 1\} \overset{\circ}{\subseteq} X.$$

The vectors $(1,2)'$ and $(1,6)'$ are the eigenvectors of C with eigenvalues 0.5 and -0.5. It is an open problem whether something similar to $\subsetneq_{*}$ can be defined for $X \in \mathbb{P}\,\mathbb{V}\mathbb{R}$ to weaken the assumptions of Theorem 5. Theorem 5 has been proved for compact and convex sets in [37].

The assumptions of theorem 6 can still be weakened. For preparing this we need the following lemma.

Lemma 7. Let $Z \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, $C \in \mathbb{M}\mathbb{R}$ and $X \in \mathbb{I}\,\mathbb{V}\mathbb{R}$. If C is irreducible, $d(X) > 0$ and

$$\square(Z + C\cdot X) \overset{\circ}{\subseteq} X$$

then $\rho(|C|) < 1$.

Proof. By lemma 3, c), $|C|\cdot Y \subseteq C \diamondsuit Y \overset{\circ}{\subseteq} Y$ for $Y = X \diamondsuit X$. The assertion follows by Perron-Frobenius Theory (cf. [45]). □

Theorem 8. Let $Z \in \mathbb{P}\,\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\,\mathbb{M}\mathbb{R}$ and $X \in \mathbb{I}\,\mathbb{V}\mathbb{R}$. If

$$\square(Z + \mathbb{C}\cdot X) \subsetneq_{*} X \quad \text{using Einzelschrittverfahren,} \tag{15}$$

i.e. for $Y_i := \{\square(Z + \mathbb{C}\cdot(Y_1,\ldots,Y_{i-1},X_i,\ldots,X_n)^T)_i$ holds $Y_i \subsetneq_{*} X_i$ for $1 \leq i \leq n$, then for every $C \in \mathbb{C}: \rho(C) \leq \rho(|C|) < 1$.

Proof. Let $z \in Z$ and $C \in \mathbb{C}$, arbitrarily chosen. By Perron-Frobenius Theory (cf. [45], (2.41) page 46) there is a permutation matrix P with

$$PCP^T = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ & R_{22} & \cdots & R_{2n} \\ & & \ddots & \vdots \\ & & & R_{nn} \end{bmatrix}$$

where $R_{ii} = 0$ or R_{ii} irreducible for $1 \le i \le n$. Let $1 \le \nu \le n$ be fixed but arbitrary. If $R_{\nu\nu} = 0$ then $\rho(R_{\nu\nu}) < 1$; suppose $R_{\nu\nu} \neq 0$ and let $P\cdot(1,2,\ldots,n)^T = (\sigma(1),\sigma(2),\ldots,\sigma(n))^T$, $\tau(\sigma(j)) := j$ for $1 \le j \le n$. Let $R_{\nu\nu} \in M_k\mathbb{R}$, $k \le n$ with indices $\ell,\ldots,\ell+k-1$. The indices of A belonging to $R_{\nu\nu}$ are $\tau(\ell) =: \alpha_1,\ldots,\tau(\ell+k-1) =: \alpha_k, \alpha_m := \max(\alpha_1,\ldots,\alpha_k)$. Because of the irreducibility of $R_{\nu\nu}$ there is a path from α_i to α_m for every $1 \le i \le k$ (cf.[45]). Therefore there is a j with $1 \le j \le k$ and $C_{\alpha_j,\alpha_m} \neq 0$. For arbitrary $\overline{y}_\mu, \tilde{y}_\mu \in Y_\mu$, $1 \le \mu \le i-1$ and $\overline{x}_\mu, \tilde{x}_\mu \in X_\mu$, $i \le \mu \le n$ and arbitrary i with $1 \le i \le n$, holds

$$(C\cdot(\overline{y}_1-\tilde{y}_1,\ldots,\overline{y}_{i-1}-\tilde{y}_{i-1},\overline{x}_i-\tilde{x}_i,\ldots,\overline{x}_n-\tilde{x}_n)^T)_i =$$

$$(z+C\cdot(\overline{y}_1,\ldots,\overline{y}_{i-1},\overline{x}_i,\ldots,\overline{x}_n)^T)_i - (z+C\cdot(\tilde{y}_1,\ldots,\tilde{y}_{i-1},\tilde{x}_i,\ldots,\tilde{x}_n)^T)_i \subseteq Y_i - Y_i$$

and therefore $(C\cdot(Y_1-Y_1,\ldots,Y_{i-1}-Y_{i-1},X_i-X_i,\ldots,X_n-X_n)^T)_i \subseteq Y_i - Y_i =$

$$= Y_i \diamondsuit Y_i \subsetneq X_i - X_i .$$

That means for $U := X - X$

$$C \diamondsuit U \subsetneq U \quad \text{using Einzelschrittverfahren.} \tag{16}$$

Let $M := \{\alpha_1,\ldots,\alpha_k\}$ and define

$$C^*_{i\ell} := \begin{cases} C_{i\ell} & \text{if } i \in M,\ \ell \in M \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad U^*_i := \begin{cases} U_i & \text{if } i \in M \\ 0 & \text{otherwise.} \end{cases}$$

Because $0 \in U_i$, $1 \le i \le n$ we have

$$(C^*\cdot U^*)_{\alpha_j} \subseteq (C\cdot U)_{\alpha_j} = (C\cdot(Y_1-Y_1,\ldots,Y_{\alpha_m-1}-Y_{\alpha_m-1},Y_{\alpha_m}-Y_{\alpha_m},\ldots,Y_n-Y_n)^T)_{\alpha_j}$$

$$\tilde{\subseteq} (C\cdot(Y_1-Y_1,\ldots,Y_{\alpha_j-1}-Y_{\alpha_j-1},X_{\alpha_j}-X_{\alpha_j},\ldots,X_{\alpha_m}-X_{\alpha_m},\ldots,X_n-X_n)^T)_{\alpha_j}$$

because $C_{\alpha_j\alpha_m} \neq 0$ and the definition of α_m. Define $C_M \in M_k\mathbb{R}$ resp. $U_M \in V_k\mathbb{R}$ to be the matrix resp. vector out of C resp. U with indices in M. Then $C_M\cdot U_M \tilde{\subseteq} U_M$. But C_M is the permuted $R_{\nu\nu}$ so that by lemma 7 the spectral radius if $|R_{\nu\nu}|$ is less than one. Because ν was arbitrary with $1 \le \nu \le n$ the theorem is proved. □

Theorem 8 provides a sharp test for $\rho(|C|) < 1$. As compared to the general case of compact inclusion sets $X \in \mathbb{P}\,V\mathbb{R}$, for $X \in \mathbb{I}\,V\mathbb{R}$ the weaker inclusion $\subsetneq$ and the Einzelschrittverfahren modus suffices.

Furthermore, as will be shown later, inclusion using Einzelschrittverfahren saves computing time and memory. It is an open problem whether something similar can be defined and be achieved for the general case $X \in \mathbb{P}\,V\mathbb{R}$ compact.

Theorem 9. Let $z \in V\mathbb{R}$, $C \in M\mathbb{R}$ and $X \in \mathbb{I}\,V\mathbb{R}$, $d(X) > 0$. If

$$\diamondsuit(z + C\cdot X) \supseteq X \tag{17}$$

then $\rho(|C|) \ge 1$.

Proof. Let m(X) be the midpoint of X and $Y := X - m(X)$. Then $Y \in \mathbb{I}\,V\mathbb{R}$, $0 \in Y$ and

$$\diamondsuit(z + C\cdot X) = z + C \diamondsuit X = z + C \diamondsuit (m(X)+Y) = z + C\cdot m(X) + C \diamondsuit Y \supseteq X \quad \text{and}$$

$$C \diamondsuit Y \supseteq Y + m(X) - z - C\cdot m(X).$$

Because $Y \in \mathbb{I}\,V\mathbb{R}$ and $Y = -Y$ this implies (abbreviating $v := m(X)-z-C\cdot m(X)$)

$$-|C|\cdot|Y| \le -|Y| + v \le |Y| + v \le |C|\cdot|Y|$$

and therefore

$$|C|\cdot|Y| \ge |Y| . \tag{18}$$

Observing $|Y| > 0$ the proof is finished by applying exercise 2, p.47 in [45]. □

If an iteration $X^{k+1} := \diamondsuit(z + C\cdot X^k)$ is performed, theorem 9 gives a stopping criterion because $X^{k+1} \subsetneq X^k$ would imply $\rho(|C|) < 1$ for every $C \in \mathbb{C}$. Theorem 9 holds true because of the special shape of interval vectors. However, (17) does not imply $\rho(C) > 1$ because of interval dependencies. A counterexample is:

$$C := \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} \quad \text{for } z = 0 \text{ and } X := \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mid -1 \le x_1, x_2 \le 1 \right\}.$$

Then $z + C\cdot X = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mid -2 \le x_1 = x_2 \le 2 \right\}$, but $z + C \diamondsuit X = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mid -2 \le x_1, x_2 \le 2 \right\} \supseteq X$ with $\rho(C) = 0$, where $\rho(|C|) = 2$.

Applying iteration schemes to $f(x) = Ax - b$ the aim is to demonstrate the existence (and maybe uniqueness) of a zero of f(x) within a certain domain. A preparation for this is the following lemma.

Lemma 10. Let $z \in \mathbb{VR}$, $C \in \mathbb{MR}$ and $\emptyset \neq X \in \mathbb{PVR}$ compact. If $\rho(C) < 1$ and

$$z + C \cdot X \subseteq X \tag{19}$$

then there is one and only one $\hat{x} \in \mathbb{VR}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \in X$.

Proof. Define $f: \mathbb{VR} \in \mathbb{VR}$ by $f(x) := z + C \cdot x$, then for $m \in \mathbb{N}$

$$f^m(x) = \sum_{i=0}^{m-1} C^i z + C^m x \quad \text{and} \quad (I-C) \cdot f^m(x) = z + C^m \cdot (x - z - Cx).$$

By (19) $f^m(x) \in X$ for $x \in X$ and therefore $(I-C) \cdot f^m(x) \to z$ for every $x \in X$, $m \to \infty$ because C is convergent. But $I-C$ is invertible and X closed and bounded, therefore $f^m(x) \to (I-C)^{-1} \cdot z$ for $m \to \infty$, i.e. $\hat{x} := (I-C)^{-1} \cdot z \in X$. Now

$$f(\hat{x}) - \hat{x} = z + C \cdot (I-C)^{-1} \cdot z - (I-C)^{-1} \cdot z = 0 \tag{20}$$

Suppose $z + C \cdot \hat{y} = \hat{y}$. Then

$$C \cdot (\hat{x} - \hat{y}) = (z + C \cdot \hat{x}) - (z + C \cdot \hat{y}) = \hat{x} - \hat{y} \tag{21}$$

implying $\hat{x} - \hat{y} = 0$ because $\rho(C) < 1$. □

With the previons theorems 5 and 8 we have the tool to verify $\rho(C) < 1$ by means of conditions (13) and (15), which can be implemented on computers. This leads to the following.

Theorem 11. Let $Z \in \mathbb{PVR}$, $\mathbb{C} \in \mathbb{PMR}$ and $\emptyset \neq X \in \mathbb{PVR}$ compact. If

$$Z + \mathbb{C} \cdot X \overset{\circ}{\subseteq} X \tag{22}$$

then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{VR}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \in \overset{\circ}{X}$.

Proof. Let $z \in Z$ and $C \in \mathbb{C}$ fixed but arbitrary. Then (22) and theorem 5 yields $\rho(C) < 1$ and (19) and by lemma 10 the existence of a $\hat{x} \in X$ with $z + C \cdot \hat{x} = \hat{x}$. Moreover $\hat{x} = z + C \cdot \hat{x} \in Z + \mathbb{C} \cdot X \subseteq \overset{\circ}{X}$ by (22). □

Theorem 12. Let $Z \in \mathbb{PVR}$, $\mathbb{C} \in \mathbb{PMR}$ and $X \in \mathbb{IMR}$. If

$$O(Z + \mathbb{C} \cdot X) \underset{*}{\subseteq} X \quad \text{using Einzelschrittverfahren,} \tag{23}$$

i.e. for $Y_i := \{O(Z + \mathbb{C} \cdot (Y_1, \ldots, Y_{i-1}, X_i, \ldots, X_n)^T\}_i$ holds $Y_i \subsetneq X_i$ for $1 \leq i \leq n$, then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(\mathbb{C}) < 1$ and there is one and only one fixed point $\hat{x} \in \mathbb{VR}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \in Y$ and $Z + \mathbb{C} \cdot Y \subseteq Y$.

Proof. Let $z \in Z$ and $C \in \mathbb{C}$ fixed but arbitrary. Then (23) and theorem 8 implies $\rho(C) < 1$. Moreover

$$\{z + C \cdot (Y_1, \ldots, Y_n)^T\}_i \subseteq \{z + C \cdot (Y_1, \ldots, Y_{i-1}, X_i, \ldots, X_n)^T\}_i = Y_i \text{ for } 1 \leq i \leq n$$

implying $z + C \cdot Y \subseteq Y$. Lemma 10 concludes the proof. □

The next theorem gives another improvement of the inclusion assumption (13) in theorem 5, combined with an iteration scheme.

Theorem 13. Let $Z \in \mathbb{PVR}$, $\mathbb{C} \in \mathbb{PMR}$ and $\emptyset \neq X \in \mathbb{PVR}$, all Z, $\mathbb{C}$ and X being compact. Define $f: \mathbb{PVR} \to \mathbb{PVR}$ by $f(V) := Z + \mathbb{C} \cdot V$ for $V \in \mathbb{PVR}$. If

$$f^{k+m}(X) \overset{\circ}{\subseteq} f^k(X) \quad \text{for some } k, m \in \mathbb{N},\ k \geq 0,\ m \geq 1 \tag{24}$$

then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{VR}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \overset{\circ}{\in} f^k(X)$.

Proof. Let $z \in Z$ and $C \in \mathbb{C}$ fixed but arbitrary. Define $g: \mathbb{VR} \to \mathbb{VR}$ by $g(x) := z + C \cdot x$ and let $Y := f^k(X)$. Y is compact because Z, $\mathbb{C}$ and X are, so with

$$g^m(Y) = \sum_{i=0}^{m-1} C^i \cdot z + C^m \cdot Y \subseteq f^m(Y) \overset{\circ}{\subseteq} Y, \tag{25}$$

theorem 5 implies $\rho(C^m) < 1$ and therefore $\rho(C) < 1$, and $g^m(\hat{x}) = \hat{x}$ for $\hat{x} := (I-C^m)^{-1} \cdot (\sum_{i=0}^{m-1} C^i) \cdot z$. On the other hand

$$(\sum_{i=0}^{m-1} C^i)(I-C) = I - C^m, \tag{26}$$

and by $\rho(C) < 1$ and $\rho(C^m) < 1$ follows $\hat{x} = (I-C)^{-1} \cdot z \in Y$. Now (20) and (21) complete the proof. □

When aiming to compute an inclusion of the solution of a given problem (e.g. a system of linear of nonlinear equations), an iteration scheme is very useful. Since the problem of computing an inclusion is later reduced to verify contraction for an affine mapping, let us consider (24) more closely. Consider the one-dimensional, con-

tracting affine function

$$f(x) := 3 - 0.5 \cdot x. \tag{27}$$

For $X := [1.9,2.1]$ is $f(X) = [1.95, 2.05] \overset{\circ}{\subseteq} X$. Therefore f has one and only one fixed point by theorem 11 and $\hat{x} = (1-(-0.5))^{-1} \cdot 3 = 2$. By the formula used in (25) is

$$f^m(x) = \sum_{i=0}^{m-1} (-0.5)^i \cdot 3 + (-0.5)^m \cdot x = \frac{1-(-0.5)^m}{1-(-0.5)} \cdot 3 + (-0.5)^m \cdot x = 2 + (-0.5)^m \cdot (x-2).$$

For $X^0 := [1.9,2.3]$ and $X^m := f^m(X)$ is

$$X^1 = [1.85,2.05],\ X^2 = [1.975,2.075],\ X^3 = [1.9625,2.0125],\ldots$$

It is easy to show $X^{m+1} \overset{\circ}{\subseteq} X^m$ for every $0 \leq m \in \mathbb{N}$. Therefore the contraction cannot be established by theorem 11. However, $X^2 \overset{\circ}{\subseteq} X^0$ and theorem 13 can be applied. For $X^0 := [3,5]$ is

$$X^1 = [0.5,1.5],\ X^2 = [2.25,2.75] \text{ and } X^m = 2 + (-0.5)^m \cdot [1,3]. \tag{28}$$

Therefore $2 \notin X^m$ for every $0 \leq m \in \mathbb{N}$ and by theorem 13, (24) can never be satisfied. Therefore, although the function f defined by (27) is contracting, no inclusion is possible using any of the theorems so far (for the Einzelschrittverfahren as well since the function is one-dimensional).

Next we seek for methods providing inclusions even under extreme circumstances. The next theorems allows an inclusion for the example above.

Theorem 14. Let $Z \in \mathbb{P}\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\mathbb{M}\mathbb{R}$ and $\emptyset \neq X \in \mathbb{P}\mathbb{V}\mathbb{R}$, all Z,$\mathbb{C}$ and X being compact. Define $f : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ by $f(V) := Z + \mathbb{C} \cdot V$ for $V \in \mathbb{P}\mathbb{V}\mathbb{R}$. If

$$f^{m+1}(X) \overset{\circ}{\subseteq} \bigcup_{i=0}^{m} f^i(X) \quad \text{for some} \quad 0 \leq m \in \mathbb{N}, \tag{29}$$

then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{V}\mathbb{R}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \in \bigcup_{i=0}^{m} f^i(X)$.

Proof. For $U,V \in \mathbb{P}\mathbb{V}\mathbb{R}$ is

$$f(U \cup V) = f(U) \cup f(V).$$

For $Y := \bigcup_{i=0}^{m} f^i(X)$ is $f^{m+1}(X) \overset{\circ}{\subseteq} Y$ and by induction for $k \geq 1$

$$f^{m+k+1}(X) \subseteq f(f^{m+k}(X)) \overset{\circ}{\subseteq} f(Y) \subseteq f^{m+1}(X) \cup \bigcup_{i=0}^{m} f^i(X) \subseteq Y. \tag{30}$$

Therefore

$$f^{m+1}(Y) = \bigcup_{i=0}^{m} f^{m+1}(f^i(X)) = \bigcup_{i=0}^{m} f^{m+i+1}(X) \overset{\circ}{\subseteq} Y.$$

Y is compact because Z,$\mathbb{C}$ and X are and applying theorem 13 completes the proof. □

Corollary 15. Let $Z \in \mathbb{P}\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\mathbb{M}\mathbb{R}$ and $\emptyset \neq X \in \mathbb{P}\mathbb{V}\mathbb{R}$, all Z,$\mathbb{C}$ and X being compact. Define $f : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ by $f(V) := Z + \mathbb{C} \cdot V$ for $V \in \mathbb{P}\mathbb{V}\mathbb{R}$. If ($\underline{\cup}$ denotes the convex union)

$$f^{m+1}(X) \overset{\circ}{\subseteq} \underline{\bigcup_{i=0}^{m}} f^i(X) \quad \text{for some} \quad 0 \leq m \in \mathbb{N},$$

then for every $z \in Z$ and every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{V}\mathbb{R}$ with $z + C \cdot \hat{x} = \hat{x}$. It is $\hat{x} = (I-C)^{-1} \cdot z \in \underline{\bigcup_{i=0}^{m}} f^i(X)$.

Proof. Follows by replacing $\cup$ by $\underline{\cup}$ in the proof of theorem 14 or, by lemma 2,a) and theorem 13. □

Note, that corollary 15 holds true because f is an affine function. Let $A,B \in \mathbb{P}\mathbb{V}\mathbb{R}$, then

$$\begin{aligned} f(A \underline{\cup} B) &= \{f(x) \mid x = a + \lambda (b-a) \text{ for } a \in A,\ b \in B,\ \lambda \in \mathbb{R} \text{ with } 0 \leq \lambda \leq 1\} = \\ &= \{z + C \cdot (a + \lambda(b-a)) \mid a \in A,\ b \in B,\ \lambda \in \mathbb{R} \text{ with } 0 \leq \lambda \leq 1\} = \\ &= \{z + C \cdot a + \lambda \cdot ((z + C \cdot b) - (z + C \cdot a)) \mid a \in A,\ b \in B,\ \lambda \in \mathbb{R} \text{ with } 0 \leq \lambda \leq 1\} = \\ &= \{z + C \cdot a \mid a \in A\} \underline{\cup} \{z + C \cdot b \mid b \in B\} = \\ &= f(A) \underline{\cup} f(B). \end{aligned}$$

If the function f is not affine, there need not be a relation between $f(A \underline{\cup} B)$ and $f(A) \underline{\cup} f(B)$. Consider as examples:

1.) $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x$ and $A := \{1\}$, $B = \{-1\}$. Then

$f(A \underline{\cup} B) = \{f(x) \mid x \in [-1,1]\} = [0,1]$ and

$f(A) \underline{\cup} f(B) = \{1\} \underline{\cup} \{1\}$ with $f(A \underline{\cup} B) \not\subseteq f(A) \underline{\cup} f(B)$.

2.) $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(x,y) = (x^2, y^3)^T$ and $A = \{\binom{0}{0}\}$, $B = \{\binom{1}{1}\}$. Then

$f(A \cup B) = \{f(x,y) \mid 0 \le x = y \le 1\} = \{\binom{x^2}{x^3} \mid 0 \le x \le 1\}$ and

$f(A) \cup f(B) = \{\binom{0}{0}\} \cup \{\binom{1}{1}\} = \{\binom{x}{x} \mid 0 \le x \le 1\}$ with

$f(A) \cup f(B) \subsetneq f(A \cup B)$ and $f(A \cup B) \not\subseteq f(A) \cup f(B)$. In fact

$(f(A) \cup f(B)) \cap f(A \cup B) = \{\binom{0}{0}, \binom{1}{1}\} = f(A) \cup f(B)$ which is the bare minimum the intersection must contain.

For the function f defined in (27) and $X := [3,5]$ is by (28)

$$f^2(X) = [2.25, 2.75] \overset{0}{\subseteq} X \cup f(X) = [3,5] \cup [0.5, 1.5] = [0.5, 5].$$

This implies f to be contractive and gives a (poor) inclusion $[0.5,5]$ for the fixed point $\hat{x} = 2$ of f. In the following chapter theorems are given for improving the quality of an inclusion.

3. Linear systems. Our next aim are theorems for inclusions of the solution of systems of linear equations of those theorems should be verifiable on computers such that inclusions sets can be calculated. In [17] and [39] a number of those theorems are given. Next those theorems will be improved. The inclusion formula of the next theorem occured in [21]. There, Krawczyk supposed $\rho(I-RA) < 1$ a priori. Using the proper inclusion $\overset{0}{\subseteq}$ we have an easy verifaction scheme for the iteration matrix to be convergent.

Theorem 16. Let $A,R \in \mathbb{M}\mathbb{R}$, $b,\tilde{x} \in \mathbb{V}\mathbb{R}$. If for some compact $\emptyset \neq X \in \mathbb{P}\,\mathbb{V}\mathbb{R}$

$$\tilde{x} + R(b - A\tilde{x}) + \{I - RA\} \cdot (X - \tilde{x}) \overset{0}{\subseteq} X, \tag{31}$$

then the matrices A and R are not singular and the uniquely determined solution $\hat{x} := A^{-1} \cdot b$ of $Ax = b$ satisfies $\hat{x} \in \overset{0}{X}$.

Proof. Short computation yields for $x \in \mathbb{V}\mathbb{R}$

$$\tilde{x} + R(b - A\tilde{x}) + \{I - RA\} \cdot (x - \tilde{x}) = R \cdot b + \{I - RA\} \cdot x . \tag{32}$$

Therefore $R \cdot b + \{I-RA\} \cdot X \overset{0}{\subseteq} X$ and by theorem 11 $\rho(I-RA) < 1$ proving the non-singularity of R and A. The fixed point $\hat{x}$ of the function $R \cdot b + \{I-RA\} \cdot x$ satisfies $\hat{x} \in \overset{0}{X}$ and $\hat{x} = \{I-(I-RA)\}^{-1} \cdot R \cdot b = A^{-1} \cdot b$ by theorem 11. □

Obviously (31) can be replaced by $R \cdot b + \{I-RA\} \cdot X \overset{0}{\subseteq} X$, which is, for appropriate X, also verifiable on computers.

One significant improvement of the quality of inclusions is not to include the solution itself but the difference to an approximate solution (cf. [37] and [38]). As shown by (32) the approximation $\tilde{x}$ to $\hat{x}$ does not play any role in (31). If, instead the inclusion shall satisfy $\hat{x} - \tilde{x} \in X$, then the linear system $Ax = b - A\tilde{x}$ has to be solved because

$$A(\hat{x} - \tilde{x}) = A\hat{x} - A\tilde{x} = b - A\tilde{x} .$$

This leads to the well-known residue iteration scheme. The corresponding inclusion theorem is the following.

Theorem 17. Let $A,R \in \mathbb{M}\mathbb{R}$, $b,\tilde{x} \in \mathbb{V}\mathbb{R}$. If for some compact $\emptyset \neq X \in \mathbb{P}\,\mathbb{V}\mathbb{R}$

$$R \cdot (b - A\tilde{x}) + \{I - RA\} \cdot X \overset{0}{\subseteq} X, \tag{33}$$

then the matrices A and R are not singular and the uniquely determined solution $\hat{x} := A^{-1} \cdot b$ of $Ax = b$ satisfies $\hat{x} \in \tilde{x} + X$. If

$$R \cdot (b - A\tilde{x}) + \{I - RA\} \cdot X \cap X = \emptyset , \tag{34}$$

then there is no solution of $Ax = b$ in $\tilde{x} + X$.

Proof. The first assertion follows by applying theorem 16 replacing b by $b - A\tilde{x}$ and regarding (32). Suppose $A\hat{x} = b - A\tilde{x}$. Then

$$R \cdot (b - A\tilde{x}) + \{I - RA\} \cdot (\hat{x} - \tilde{x}) = \hat{x} - \tilde{x} \tag{35}$$

would contradict (34) if $\hat{x} - \tilde{x} \in X$. □

Theorem 17 can be applied on computers if operations extending the power set operations, possibly for special inclusion sets, are available. One possibility is to restrict the inclusion sets to interval vectors, i.e. hyperrectangles or, in the complex number space, to n-dimensional balls or torus-sectors. For the application on computers we formulate theorem 17 for interval vectors over some subset G of $\mathbb{R}$. $\square : \mathbb{P}V\mathbb{R} \to \mathbb{I}VG$ is some isotone rounding, i.e. satisfying $X \in \mathbb{P}V\mathbb{R} \to X \subseteq \square X$.

Theorem 18. Let $G \subseteq \mathbb{R}$ be some subset of $\mathbb{R}$, $A,R \in MG$, $b,\tilde{x} \in VG$ and $X \in \mathbb{I}VG$. If

$$\square(R\cdot(b-A\tilde{x}) + (I-RA)\cdot X) \subsetneq X \quad \text{using Einzelschrittverfahren,} \tag{36}$$

i.e. for $Y_i := \{\square(R(b-A\tilde{x}) + (I-RA)\cdot(Y_1,\ldots,Y_{i-1},X_i,\ldots,X_n)^T\}_i$ holds $Y_i \subsetneq X_i$ for $1 \le i \le n$, then the matrices A and R are not singular and the uniquely determined solution $\hat{x} := A^{-1}\cdot b$ of $Ax = b$ satisfies $\hat{x} \in \tilde{x} + Y$.

Proof. By theorem 12 is $\rho(I-RA) < 1$ and therefore A and R regular, and $\{I-(I-RA)\}^{-1}\cdot R\cdot(b-A\tilde{x}) = A^{-1}\cdot(b-A\tilde{x}) = A^{-1}\cdot b - \tilde{x} \in Y$. □

Assumption (36) can be verified on computers. If G is a set of floating-point numbers and operations $\circledast : \mathbb{I}G \to \mathbb{I}G$ satisfying $A,B \in \mathbb{I}G$: $A * B \subseteq A \circledast B$ for $* \in \{+,-,\cdot,/\}$ are available, then

$$R \boxdot (b \boxminus A \boxdot \tilde{x}) \boxplus (I \boxminus R \boxdot A) \boxdot X \subsetneq X \quad \text{using Einzelschrittverfahren} \tag{37}$$

implies (36) and therefore the assertions of theorem 18 hold true. If a precise scalar product is available, then

$$R \boxdot \square(b-Ax) \boxplus \square(I-RA) \boxdot X \subsetneq X \quad \text{using Einzelschrittverfahren} \tag{38}$$

can be used. Obviously, (38) implies (36) but is much sharper then (37) because the critical parts, the residuals, are enclosed with maximum accuracy.

It occurs frequently, that either the input data of a numerical problem is not exactly representable in a given floating-point screen or, that input data is afflicted with tolerances. In the latter case, one is interested in an inclusion of the set of all possible solutions for all possible combinations of input data. In the first case, the

input data could be enclosed in the input range of the immediate predecessor and successor in the floating-point grid and the resulting problem with input afflicted with tolerances be solved. For input data with tolerances consider the following theorem.

Theorem 19. Let $\mathbb{A} \in \mathbb{P}M\mathbb{R}$, $\mathbb{b} \in \mathbb{P}V\mathbb{R}$, $R \in M\mathbb{R}$ and $\tilde{x} \in V\mathbb{R}$. If for some compact $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$

$$R\cdot(\mathbb{b}-A\tilde{x}) + (I-R\mathbb{A})\cdot X \overset{\circ}{\subseteq} X, \tag{39}$$

then for every $A \in \mathbb{A}$ and for every $b \in \mathbb{b}$ the following is true: the matrices A and R are not singular and the uniquely determined solution $\hat{x} := A^{-1}b$ of $Ax = b$ satisfies $\hat{x} \in \tilde{x} + X$.

Proof. Follows by applying theorem 17 to every $A \in \mathbb{A}$, $b \in \mathbb{b}$. □

For immediate applications on computers the power set operations can be replaced by corresponding isotone operations, which are executable on computers. The input data $\mathbb{A}$ and $\mathbb{b}$, which are afflicted with tolerances, can be replaced by computer representable sets. Usually, one would choose R and $\tilde{x}$ to be point data. An example of such a theorem is the following.

Theorem 20. Let $G \subseteq \mathbb{R}$ be some subset of $\mathbb{R}$, $\mathbb{A} \in \mathbb{I}MG$, $\mathbb{b} \in \mathbb{I}VG$, $R \in MG$ and $\tilde{x} \in VG$. If for some $X \in \mathbb{I}VG$

$$R \boxdot \square(b-A\tilde{x}) \boxplus \square(I-R\mathbb{A}) \boxdot X \subsetneq X \quad \text{using Einzelschrittverfahren,}$$

then for every $A \in M\mathbb{R}$ and $b \in V\mathbb{R}$ with $A \in \mathbb{A}$ and $b \in \mathbb{b}$ the following is true: the matrices A and R are not singular and the uniquely determined solution $\hat{x} := A^{-1}b$ of $Ax = b$ satisfies $\hat{x} \in \tilde{x} \boxplus X$.

Proof. Follows by applying theorem 18 to every $A \in \mathbb{A}$, $b \in \mathbb{b}$. □

Note, that the assertion of theorem 20 applies to every real matrix $A \in M\mathbb{R}$ and every real vector $b \in V\mathbb{R}$ with $A \in \mathbb{A}$, $b \in \mathbb{b}$ and not only to those with elements in G. Moreover, a computer verification holds true without any restriction, e.g. with respect to a given floating-point precision.

In [38] an algorithm is given for computing an inclusion of the solution of a system of linear equations. After computing an approxi-

mate inverse R of A a residual correction method is applied yielding some approximate solution $\tilde{x}$. After computing an inclusion for $C := 0(I-RA)$ an iteration is started. To illustrate the differences between the different iteration schemes for obtaining an inclusion consider the following example. Let

$$A := \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}; \quad A^{-1} = \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix}; \quad R := \begin{pmatrix} -2.85 & 2.1 \\ 1.85 & -1.1 \end{pmatrix}. \text{ Then}$$

$$C := \begin{pmatrix} -0.35 & -0.6 \\ 0.35 & 0.6 \end{pmatrix} \text{ and } \rho(|C|) = 0.95; \quad b := \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{40}$$

The following computational results are obtained on a IBM S/370 processor in single precision, i.e. 6 hexadecimal digits or approximately 7 decimal digits. We apply three different iteration schemes according to theorems 16,17 and 18:

$$\begin{aligned} &1)\ Z := \tilde{x} \oplus R \odot 0(b-A\tilde{x}); \quad X^0 := Z; \quad X^{k+1} := Z \oplus C \odot (X^k \ominus \tilde{x}); \\ &2)\ Z := R \odot b; \qquad\qquad X^0 := Z; \quad X^{k+1} := Z \oplus C \odot X^k; \\ &3)\ Z := R \odot 0(b-A\tilde{x}); \qquad X^0 := Z; \quad X^{k+1} := Z \oplus C \odot X^k; \end{aligned} \tag{41}$$

Then $X^{k+1} \subsetneq X^k$ proves the non-singularity of A and R and $\hat{x} \in X$ resp. $\hat{x} \in x \oplus X$ in cases 1) and 2) resp. 3), $\hat{x} := A^{-1} \cdot b$ according to theorems 16,17 and 18.

In the following table from left to right the number i of the iteration scheme, the number of iterations k and the maximum relative error e for the two inclusions (defined by $\max_{x_1,x_2 \in X} \left|\frac{x_1-x_2}{x_1}\right|$ for an interval X) are displayed.

i	k	e
1	no inclusion	
2	10	$1.5 \cdot 10^{-5}$
3	1	$6.0 \cdot 10^{-8}$

For the first iteration, no inclusion is obtained because the iterates X^k become cyclic. The third scheme enclosing the difference $\hat{x} - \tilde{x}$ is superior with respect to the number of iterations necessary and with respect to the quality of the inclusion. Nevertheless it is

possible that $\rho(|C|) < 1$ but neither of the iteration schemes (41) obtain an inclusion. Consider

$$f(x) := 3 + 0.5 \cdot x. \tag{42}$$

Then by the formula used in (25)

$$f^m(x) = \sum_{i=0}^{m-1} 0.5^i \cdot 3 + 0.5^m \cdot x = \frac{1-0.5^m}{1-0.5} \cdot 3 + 0.5^m \cdot x = 6 + 0.5^m \cdot (x-6).$$

For $X^0 := [4,5]$ is

$$X^1 = [5,5.5]; \quad X^2 = [5.5,5.75]; \quad X^3 = [5.75,5.875] \text{ etc.}$$

It is easy to see that $X^{k+m} \subsetneq X^k$ and $X^{k+m} \not\subseteq \sum_{i=0}^{m-1} X^{k+i}$ for every $k,m \in \mathbb{N}$.

For the purpose of obtaining an inclusion even in those cases, in [37] the ε-inflation has been introduced. An iteration with ε-inflation is

$$\begin{aligned} &\underline{\text{repeat}}\ Y := X \circledcirc \varepsilon;\ X := Z + C \cdot Y \\ &\underline{\text{until}}\ \ X \overset{0}{\subseteq} Y \end{aligned} \tag{43}$$

for some starting interval X and the property $X \in \mathbb{IP}\,\mathbb{VR} : X \subseteq X \circledcirc \varepsilon$. One could, e.g., define $X \circledcirc \varepsilon := X + [-0.5,+0.5]$. Then with $X := [4,5]$ for the example in (42)

$$Y = [3.5,5.5]; \quad X = [4.75,5.75]; \quad Y = [4.25,6.25]; \quad X = [5.125,6.125]$$

with an inclusion [5.125,6.125] and all the verifications. When including the difference to an approximate solution rather than the solution itself this inclusion will be sharpened.

What is observed in the example can be generalized. It can be shown, that an iteration (43) terminates if and only if $\rho(C) < 1$. It is clear from the previous theorems, that if (43) terminates then $\rho(C) < 1$. The other direction is demonstrated in the following.

Lemma 21. Let $Z \in \mathbb{IP}\,\mathbb{VR}$, $C \in \mathbb{MR}$ and $E_i \in \mathbb{IP}\,\mathbb{VR}$, $i \geq 1$. Suppose Z and the E_i to be bounded, $E_{i+1} \subseteq E_i$ and $U_\varepsilon(0) \subseteq E_i$ for all $i \geq 1$ and some fixed $0 < \varepsilon \in \mathbb{R}$. Define $f: \mathbb{IP}\,\mathbb{VR} \to \mathbb{IP}\,\mathbb{VR}$ by

$$X \in \mathbb{IP}\,\mathbb{VR}: \quad f(X) := Z + C \cdot X. \tag{44}$$

For some $X^0 \in \mathbb{IP}\,\mathbb{VR}$ bounded define

$$X^{k+1} := f(X^k) + E_{k+1} \quad \text{for} \quad 0 \le k \in \mathbb{N}. \tag{45}$$

If $\rho(C) < 1$ then there is some $m \in \mathbb{N}$ with

$$f(X^m) \overset{\circ}{\subseteq} X^m . \tag{46}$$

Proof. First we proof be induction

$$X^m = \sum_{i=0}^{m-1} C^i \cdot (Z + E_{m-i}) + C^m \cdot X^0 \quad \text{for} \quad 0 \le m \in \mathbb{N}. \tag{47}$$

(47) is true for $m = 0$. Supposing (47) to hold for $m \in \mathbb{N}$ yields by definition (44) and (45)

$$X^{m+1} = Z + E_{m+1} + C \cdot \sum_{i=0}^{m-1} C^i (Z + E_{m-i}) + C^{m+1} \cdot X^0 = \sum_{i=0}^{m} C^i (Z + E_{m+1-i}) + C^{m+1} \cdot X^0 . \tag{48}$$

Because C is contracting there is an $m \in \mathbb{N}$ satisfying

$$C^m \cdot (Z + E_1) + C^{m+1} \cdot X^0 - C^m \cdot X^0 \overset{\circ}{\subseteq} E_{m+1}$$

because $U_\varepsilon(0) \subset E_i$, $i \ge 1$. This implies

$$Z + \sum_{i=1}^{m-1} C^i \cdot (Z + E_{m+1-i}) + C^m (Z + E_1) + C^{m+1} \cdot X^0 \overset{\circ}{\subseteq} Z + E_{m+1} + \sum_{i=1}^{m-1} C^i \cdot (Z + E_{m+1-i}) + C^m \cdot X^0$$

(for $A, B, C \in \mathbb{P}\,\mathbb{VR}$: $A - B \subseteq C$ implies $A \subseteq B + C$, but the contrary is not true). Therefore

$$Z + C \cdot \{ \sum_{i=0}^{m-1} C^i \cdot (Z + E_{m-i}) + C^m \cdot X^0 \} \overset{\circ}{\subseteq} \sum_{i=0}^{m-1} C^i \cdot (Z + E_{m+1-i}) + C^m \cdot X^0 .$$

Regarding (47) and $E_{i+1} \subseteq E_i$ proofs the lemma. □

Theorem 22. Let $Z \in \mathbb{P}\,\mathbb{VR}$, $C \in \mathbb{MR}$ and $E_i \in \mathbb{P}\,\mathbb{VR}$ for $i \ge 1$, all Z and E_i being compact. Suppose $E_{i+1} \subseteq E_i$ and $U_\varepsilon(0) \subseteq E_i$ for all $i \ge 1$ and some fixed $0 < \varepsilon \in \mathbb{R}$. Define $f : \mathbb{P}\,\mathbb{VR} \to \mathbb{P}\,\mathbb{VR}$ by

$$X \in \mathbb{P}\,\mathbb{VR}: \quad f(X) := Z + C \cdot X.$$

For some compact $X^0 \in \mathbb{P}\,\mathbb{VR}$ define

$$X^{k+1} := f(X^k) + E_{k+1} \quad \text{for} \quad 0 \le k \in \mathbb{N}.$$

Then the following is equivalent:

(1) $f(X^m) \overset{\circ}{\subseteq} X^m$ for some $m \in \mathbb{N}$

(2) $\rho(C) < 1$.

Proof. (1) ⇒ (2) follows by theorem 5 and the fact, that X^m is compact and nonempty. (2) ⇒ (1) follows by lemma 21. □

Theorem 22 shows that whenever in inclusion is possible, namely if the iteration matrix is convergent, an inclusion will be achieved using an iteration scheme (43) with the ε-inflation. This equivalence is a best possible result, it is the equivalent to the same condition for a real iteration

$$x^{k+1} := x^k + R \cdot (b - Ax^k)$$

for the linear system $Ax = b$, $R \approx A^{-1}$. Similar conditions for $\rho(|C|) < 1$ in case $X^0 \in \Pi\,\mathbb{VR}$ are given in [39].

In practice, the ε-inflation would be dependent on the iterate. For more details see [38],[39].

Theorem 22 does not necessarily hold true for sets of matrices. In fact it may happen that for a convex set of matrices $\mathbb{C}$ every $C \in \mathbb{C}$ is convergent but $C_1 \cdot C_2$ for $C_1, C_2 \in \mathbb{C}$ is not. An example is

$$\mathbb{C} := \{A + \sigma(B-A) \mid 0 \le \sigma \le 1\} \text{ for } A := \begin{bmatrix} 0.5 & 0.27 \\ 0.92 & 0.5 \end{bmatrix}, \; B := \begin{bmatrix} 0.17 & 0.6 \\ 0.94 & 0.25 \end{bmatrix}.$$

Then $\rho(A) \approx 0.9984$, $\rho(B) \approx 0.9621$ and $\max_{0 \le \sigma \le 1} \rho(A + \sigma(B-A)) \approx 0.999617$, where the maximum is achieved for $\sigma \approx 0.1266$. On the hand

$$A \cdot B = \begin{pmatrix} 0.3388 & 0.3675 \\ 0.6284 & 0.677 \end{pmatrix} \text{ and } \rho(A \cdot B) \approx 1.0166.$$

Note, that every $A + \sigma(B-A)$, $0 \le \sigma \le 1$ is positive. The condition $\rho(C_1 \cdot C_2) < 1$ for $C_1, C_2 \in \mathbb{C}$ is necessary because

$$Z + \mathbb{C} \cdot X \overset{\circ}{\subseteq} X \quad \text{implies} \quad Z + \mathbb{C} \cdot Z + \mathbb{C} \cdot \mathbb{C} \cdot X \subseteq Z + \mathbb{C} \cdot X \overset{\circ}{\subseteq} X$$

for $Z \in \mathbb{P}\,\mathbb{VR}$, $X \in \mathbb{P}\,\mathbb{VR}$, X compact and $\mathbb{C} \in \mathbb{P}\,\mathbb{MR}$. Then

$$(Z + \mathbb{C} \cdot Z) + (C_1 \cdot C_2) \cdot X \subseteq Z + \mathbb{C} \cdot Z + \mathbb{C} \cdot \mathbb{C} \cdot X \overset{\circ}{\subseteq} X \quad \text{for every} \quad C_1, C_2 \in \mathbb{C}$$

implying $\rho(C_1 \cdot C_2) < 1$ by theorem 5.

Finally we note the generalization of theorem 14 using ε-inflation. This is not clear, because there the inclusion in the convex union of all previous iterates is assumed.

Theorem 23. Let $Z \in \mathbb{P}\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\mathbb{M}\mathbb{R}$ and $\emptyset \neq X^0 \in \mathbb{P}\mathbb{V}\mathbb{R}$, all $Z,\mathbb{C}$ and X^0 being compact. Define $f : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ by $f(V) := Z + \mathbb{C}\cdot V$ and let $g_i : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ with $V \in \mathbb{P}\mathbb{V}\mathbb{R} \to g_i(V) \supseteq V$ for $i \geq 0$. Define

$$X^{k+1} := g_{k+1}(f(X^k)) \quad \text{for} \quad 0 \leq k \in \mathbb{N}.$$

If then

$$f(X^m) \overset{\circ}{\subseteq} \bigcup_{i=0}^{m} X^i \quad \text{for some} \quad 0 \leq m \in \mathbb{N}, \tag{49}$$

then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{V}\mathbb{R}$ with $z + C\cdot\hat{x} \in \hat{x}$. It is $\hat{x} = (I-C)^{-1} \in \bigcup_{i=0}^{m} X^i$.

Proof. Let $Y := \bigcup_{i=0}^{m} X^i$. We first proof by induction

$$f^k(X^{m-k+1}) \overset{\circ}{\subseteq} Y \quad \text{for} \quad 1 \leq k \leq m+1.$$

This is true for $k=1$ by assumption (49). Then

$$f^{k+1}(X^{m-k}) = f^k(f(X^{m-k})) \subseteq f^k(g_{m-k+1}(f(X^{m-k}))) = f^k(X^{m-k+1}) \overset{\circ}{\subseteq} Y. \tag{50}$$

Furthermore

$$f(Y) = \bigcup_{i=0}^{m} f(X^i) \subseteq f(X^m) \cup \bigcup_{i=0}^{m-1} g_{i+1}(f(X^i)) = f(X^m) \cup \bigcup_{i=1}^{m} X^i \subseteq Y, \tag{51}$$

implying $f^i(Y) \subseteq Y$ for $0 \leq i \in \mathbb{N}$ (but not yet $f^k(Y) \overset{\circ}{\subseteq} Y$ for some $k \in \mathbb{N}$). But

$$f^{m+1}(Y) = \bigcup_{i=0}^{m} f^{m+1}(X^i) = \bigcup_{k=1}^{m+1} f^{m+1-k}(f^k(X^{m-k+1})) \overset{\circ}{\subseteq} \bigcup_{k=1}^{m+1} f^{m+1-k}(Y) \subseteq Y$$

by (50) and (51). Y is compact because $Z,\mathbb{C}$ and X^0 are. Therefore application of theorem 13 completes the proof. □

Especially interesting for practical applications is the following corollary.

Corollary 24. Let $Z \in \mathbb{P}\mathbb{V}\mathbb{R}$, $\mathbb{C} \in \mathbb{P}\mathbb{M}\mathbb{R}$ and $\emptyset \neq X^0 \in \mathbb{P}\mathbb{V}\mathbb{R}$, all $Z,\mathbb{C}$ and X^0 being compact. Define $f : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ by $f(Y) := Z + \mathbb{C}\cdot Y$ and let $g_i : \mathbb{P}\mathbb{V}\mathbb{R} \to \mathbb{P}\mathbb{V}\mathbb{R}$ with $V \in \mathbb{P}\mathbb{V}\mathbb{R} \to g_i(V) \supseteq V$ for $i \geq 0$. Define

$$X^{k+1} := g_{k+1}(f(X^k)) \quad \text{for} \quad 0 \leq k \in \mathbb{N}.$$

If then

$$f(X^m) \overset{\circ}{\subseteq} \bigcup_{i=0}^{m} X^i \quad \text{for some} \quad 0 \leq m \in \mathbb{N}, \tag{52}$$

then for every $z \in Z$ and for every $C \in \mathbb{C}$ holds $\rho(C) < 1$ and there is one and only one $\hat{x} \in \mathbb{V}\mathbb{R}$ with $z + C\cdot\hat{x} \in \hat{x}$. It is $\hat{x} = (I-C)^{-1} \in \bigcup_{i=0}^{m} X^i$.

Proof. Follows by replacing $\cup$ by $\underline{\cup}$ in the proof of theorem 23 because f is affine (see the remarks after corollary 15). □

The application of theorem 23 and corollary 24 to intervals of vectors are like shown before. Corollary 24 allows inclusions for iteration matrices C with $\rho(C) < 1$ but $\rho(|C|) \geq 1$. An example is

$$C := \begin{bmatrix} -10 & -9.4 \\ 10.1 & 9.5 \end{bmatrix} \quad \text{with} \quad \rho(C) = 0.6 \quad \text{and} \quad \rho(|C|) \approx 19.5.$$

For $f(X) := C\cdot X$ and $X^0 := (1,-1)^T$ is $f(X^0) = (-0.6,0.6)^T$. With an ε-inflation let $X^1 := ([-0.61,-0.59],[0.59,0.61])^T$. Then

$$f(X^1) = \begin{bmatrix} 0.166 \\ -0.164 \end{bmatrix} \underline{\cup} \begin{bmatrix} 0.354 \\ -0.354 \end{bmatrix} \underline{\cup} \begin{bmatrix} 0.366 \\ -0.366 \end{bmatrix} \underline{\cup} \begin{bmatrix} 0.554 \\ -0.556 \end{bmatrix} \subseteq X^0 \underline{\cup} X^1 \tag{53}$$

proving $\rho(C) < 1$. As follows by theorem 6, $f(X^1) \overset{\circ}{\subseteq} X^0 \underline{\cup} X^1$ cannot be satisfied. In practice, especially for larger n, condition (52) is difficult to verify. To compute the convex hull of the $X^0,\ldots,X^m$ is rather time consuming. On the other hand, $\Diamond(\bigcup_{i=0}^{m} X^i)$ resp. $\square(\bigcup_{i=0}^{m} X^i)$ can be computed very fast. Therefore one might wish to generalize corollary 24 in the way, that (52) could be replaced by

$$f(X^m) \overset{\circ}{\subseteq} \Diamond(\bigcup_{i=0}^{m} X^i) \quad \text{for some} \quad m \in \mathbb{N}. \tag{54}$$

Unfortunately this is not true. Consider the following example:

$$C := \begin{bmatrix} 0.26 & 0.76 \\ 0.76 & 0.26 \end{bmatrix} \quad \text{with} \quad \rho(C) = 1.02.$$

For $X^{k+1} := C\cdot X^k$ and $X^0 := (1,0)^T$ is $X^1 = (0.26,0.76)^T$ and $X^2 = (0.6452,0.3952)^T$ with

$$X^2 \overset{\circ}{\subseteq} (X^0 \underline{\cup} X^1) = \begin{bmatrix} [0.26,1] \\ [0,0.76] \end{bmatrix}.$$

The example shows, that (54) does not imply $\rho(C) < 1$. However, the contraction is very probable if (54) is satisfied as turned out in many practical examples. Therefore, in an algorithm, condition (54) could be tested and, for a final verification,

$$f(Y) \overset{\circ}{\subseteq} Y \quad \text{with } Y := \Diamond\Big(\bigcup_{i=0}^{m} X^i\Big)$$

has to be checked. This method is very effective and inexpensive.

4. Nonlinear systems of equations. In this chapter theorems are developed for the inclusion of the solution of general systems of nonlinear equations, the assumptions of which, again, are verifyable on computers. For this purpose we first linearize a given function $f : \mathbb{VR} \to \mathbb{VR}$, $f \in C^1$ locally. For $f' : \mathbb{VR} \to \mathbb{MR}$ being the Jacobian of f holds

$$f_i(x) = f_i(\tilde{x}) + f_i'(\tilde{x} + \theta_i(x - \tilde{x}))\cdot(x - \tilde{x}) \quad \text{for } 1 \le i \le n,\ 0 < \theta_i < 1, \tag{55}$$

where $f = (f_1, \ldots, f_n)^T$ and f_i' is the i-th row of the Jacobian f'. Defining $f' : \mathbb{P}\,\mathbb{VR} \to \mathbb{P}\,\mathbb{VR}$ by

$$f_{ij}'(X) := \{f_{ij}'(x) \mid x \in X\} \quad \text{for} \quad X \in \mathbb{P}\,\mathbb{VR} \tag{56}$$

yields

$$f(x) \in f(\tilde{x}) + f'(\tilde{x} \cup x)\cdot(x - \tilde{x}) \quad \text{for} \quad \tilde{x}, x \in \mathbb{VR}. \tag{57}$$

Applying (57) to theorem 11 yields the following.

Theorem 25. Let $f : \mathbb{VR} \to \mathbb{VR}$ with $f \in C^1$, $\tilde{x} \in \mathbb{VR}$, $R \in \mathbb{MR}$ and $\emptyset \neq X \in \mathbb{P}\,\mathbb{VR}$ compact be given. If with $f' : \mathbb{P}\,\mathbb{VR} \to \mathbb{P}\,\mathbb{MR}$ defined in (56)

$$\tilde{x} - R\cdot f(\tilde{x}) + \{I - R\cdot f'(\tilde{x} \cup X)\}\cdot(X - \tilde{x}) \overset{\circ}{\subseteq} X, \tag{58}$$

Then R and every $M \in \mathbb{MR}$ with $M \in f'(\tilde{x} \cup X)$ is not singular and there is one and only one $\hat{x} \in X$ with $f(\hat{x}) = 0$. It is $\hat{x} \in \overset{\circ}{X}$.

Proof. Condition (58) implies

$$x - R\cdot\{f(\tilde{x}) + f'(x \cup X)\cdot(x - \tilde{x})\} \overset{\circ}{\subseteq} X \quad \text{for every } x \in X.$$

By (57) and the definition of f' this yields

$$\{x - R\cdot f(x) \mid x \in X\} \overset{\circ}{\subseteq} X. \tag{59}$$

By Brouwer's Fixed Point Theorem this implies the existence of an $\hat{x} \in X$ with $R\cdot f(\hat{x}) = 0$, where (59) shows $\hat{x} \in \overset{\circ}{X}$. By theorem 11 the matrix R and every matrix $M \in f'(\tilde{x} \cup X)$ is not singular implying $f(\hat{x}) = 0$. Suppose $f(\hat{y}) = 0$ for $\hat{y} \in X$. Then by the definition of f' and by (58) there is a $M \in f'(\tilde{x} \cup X)$ with

$$f(\hat{x}) = f(\hat{y}) + M(\hat{x} - \hat{y})$$

implying $M\cdot(\hat{x} - \hat{y}) = 0$ and therefore $\hat{x} = \hat{y}$. □

It has been noted in the literature (cf. [5], [14]) that the set of Jacobians defined in (56) can be weakened. Consider the following lemma.

Lemma 26. Let $f : \mathbb{VR} \to \mathbb{R}$ with $f \in C^1$ and $\tilde{x} \in \mathbb{VR}$. Then for every $x \in \mathbb{VR}$ there are $\theta_j \in \mathbb{R}$, $1 \le j \le n$ with $0 < \theta_j < 1$ and

$$f(x) = f(\tilde{x}) + \sum_{j=1}^{n} \{\tfrac{\partial f}{\partial x_j}(\tilde{x}_1, \ldots, \tilde{x}_{j-1}, \tilde{x}_j + \theta_j(x_j - \tilde{x}_j), x_{j+1}, \ldots, x_n)\cdot(x_j - \tilde{x}_j)\}. \tag{60}$$

Proof. Follows by straightforward calculation. □

Let $f : \mathbb{VR} \to \mathbb{VR}$ with $f \in C^1$ and $f = (f_1, \ldots, f_n)^T$. Then for $\tilde{x} \in \mathbb{VR}$ and the definition of $f' : \mathbb{P}\,\mathbb{VR} \to \mathbb{P}\,\mathbb{VR}$ for $X \in \mathbb{I}\,\mathbb{VR}$ by

$$f_{ij}'(X) := \{\tfrac{\partial f_i}{\partial x_j}(\tilde{x}_1, \ldots, \tilde{x}_{j-1}, \tilde{x}_j \cup x_j, x_{j+1}, \ldots, x_n) \mid x_k \in X_k \text{ for } j \le k \le n\} \tag{61}$$

follows by lemma 26:

$$f(x) \in f(\tilde{x}) + f'(X)\cdot(x - \tilde{x}). \tag{61}$$

Note, that definition (61) depends on $\tilde{x}$. Formula (62) implies that f' defined by (56) can be replaced by definition (61) in theorem 25. However, the uniqueness of the zero $\hat{x}$ of f in X cannot be proved like in theorem 25. There, for some $\hat{x}, \hat{y} \in X$ with $f(\hat{x}) = f(\hat{y}) = 0$ the existence of a matrix $M \in f'(X)$ was assumed with $f(\hat{x}) = f(\hat{y}) + M\cdot(\hat{x} - \hat{y})$. This need not to be true when using (61) instead of (56).

For our purposes, especially when applying Einzelschrittverfahren, another formulation of lemma 26 gives better results. Consider the following lemma.

Lemma 27. Let $f : \mathbb{VR} \to \mathbb{VR}$ with $f \in C^1$ and $\tilde{x} \in \mathbb{VR}$. Then for every $x \in \mathbb{VR}$ there are $\theta_{ij} \in \mathbb{R}$, $1 \le i, j \le n$ with $0 < \theta_{ij} < 1$ and

$$f(x) = f(\tilde{x}) + M\cdot(x-\tilde{x}) \quad \text{with } M \in \mathbb{MR} \text{ and} \tag{63}$$
$$M_{ij} := \frac{\partial f_i}{\partial x_j}(x_1, x_2, \ldots, x_{j-1}, \tilde{x}_j + \theta_{ij}(x_j - \tilde{x}_j), \tilde{x}_{j+1}, \ldots, \tilde{x}_n).$$

Proof. Let i with $1 \le i \le n$ fixed but arbitrary. Then by the -dimensional Mean-Value Theorem

$$f_i(x_1,x_2,\ldots,x_n) = f_i(x_1,\ldots,x_{n-1},\tilde{x}_n) + \frac{\partial f_i}{\partial x_n}(x_1,\ldots,x_{n-1},\tilde{x}_n+\theta_{in}(x_n-\tilde{x}_n))\cdot (x_n-\tilde{x}_n),$$
$$f_i(x_1,\ldots,x_{n-1},\tilde{x}_n) = f_i(x_1,\ldots,\tilde{x}_{n-1},\tilde{x}_n) + \frac{\partial f_i}{\partial x_{n-1}}(x_1,\ldots,x_{n-2},\tilde{x}_{n-1}+\theta_{i(n-1)}\cdot (x_{n-1}-\tilde{x}_{n-1}),\tilde{x}_n)\cdot(x_{n-1}-\tilde{x}_{n-1}),\ldots,$$
$$f_i(x_1,\tilde{x}_2,\ldots,\tilde{x}_n) = f_i(\tilde{x}_1,\ldots,\tilde{x}_n) + \frac{\partial f_i}{\partial x_1}(\tilde{x}_1+\theta_{i1}(x_1-\tilde{x}_1),\tilde{x}_2,\ldots,\tilde{x}_n))\cdot(x_1-\tilde{x}_1)$$

and therefore

$$f_i(x) = f_i(\tilde{x}) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x_1,\ldots,x_{j-1},\tilde{x}_j+\theta_{ij}(x_j-\tilde{x}_j),\tilde{x}_{j+1},\ldots,\tilde{x}_n)\cdot(x_j-\tilde{x}_j)$$

demonstrating the lemma. □

Next the usage of the Einzelschrittverfahren for systems of nonlinear equations will be introduced. Before doing this the technique of computing an inclusion of the difference between an approximate solution and the true solution (like for linear systems theorem 18) will be introduced using definition (63) for f'.

Theorem 28. Let $f : \mathbb{VR} \to \mathbb{VR}$ with $f \in C^1$, $\tilde{X} \in \mathbb{VR}$, $R \in \mathbb{MR}$ and $X \in \mathbb{IVR}$ be given. Let $f' : \mathbb{IVR} \to \mathbb{PMR}$ be defined by

$$f'_{ij}(Y) := \{\frac{\partial f_i}{\partial x_j}(x_1,\ldots,x_{j-1},\overline{x}_j,\tilde{x}_{j+1},\ldots,\tilde{x}_n) \mid x_i \in Y_i \text{ for } 1\le i<j,\ \overline{x}_j \in \tilde{x}_j \underline{\cup} Y_j\} \tag{64}$$

for $Y \in \mathbb{IVR}$ and $1 \le i,j \le n$. If then

$$-R\cdot f(\tilde{x}) + \{I - R\cdot f'(\tilde{x}+X)\}\cdot X \subsetneq X, \tag{65}$$

then R and every matrix $M \in \mathbb{MR}$ with $M \in f'(\tilde{x}+X)$ is not singular and there is an $\hat{x} \in \tilde{x}+X$ with $f(\hat{x}) = 0$.

Proof. By the definition (64) of f', lemma 27 and (65) follows

$$\{x - R\cdot(f(\tilde{x}) + f'(x+X)\cdot x) \mid x \in X\} \subsetneq X$$

and therefore

$$\{x - R\cdot f(\tilde{x}+x) \mid x \in X\} \subsetneq X.$$

Therefore regarding theorem 6 the proof of theorem 25 can be applied. □

Like in the case of systems of linear equations the inclusion of the difference of an approximate solution to the correct solution yields much sharper results than an inclusion of the solution itself. Next the application of the Einzelschrittverfahren is presented combined with the technique of including the residue with respect to an approximate solution.

Theorem 29. Let $f : \mathbb{VR} \to \mathbb{VR}$ with $f \in C^1$, $\tilde{X} \in \mathbb{VR}$, $R \in \mathbb{MR}$ and $x \in \mathbb{IVR}$ be given. Let $f' : \mathbb{IVR} \to \mathbb{PMR}$ be defined by

$$f'_{ij}(V) := \{\frac{\partial f_i}{\partial x_j}(x_1,\ldots,x_{j-1},\overline{x}_j,\tilde{x}_{j+1},\ldots,\tilde{x}_n) \mid x_i \in V_i \text{ for } 1\le i<j, \overline{x}_j \in \tilde{x}_j \underline{\cup} V_j\} \tag{66}$$

for $V \in \mathbb{IVR}$ and $1 \le i,j \le n$. Define recursively $Y \in \mathbb{IVR}$ by

$$Y_i := \bigcirc(-R\cdot f(\tilde{x}) + (I - R\cdot f'(\tilde{x}+Z))\cdot Z)_i \quad \text{for } 1\le i\le n \tag{67}$$

where $Z_i := (Y_1,\ldots,Y_{i-1},X_i,\ldots,X_n)^T \in \mathbb{IVR}$. If then

$$Y_i \subsetneq X_i \quad \text{for} \quad 1 \le i \le n, \tag{68}$$

then the matrix R and every matrix $M \in \mathbb{MR}$ with $M \in f'(\tilde{x}+Y)$ is not singular and there is an $\hat{x} \in \tilde{x}+Y$ with $f(\hat{x}) = 0$. Moreover with

$$W^0 := Y;\ W^{k+1} := \bigcirc(-R\cdot f(\tilde{x}) + (I - R\cdot f'(\tilde{x}+W^k))\cdot W^k)$$

holds

$$\hat{x} \in \bigcap_{k\ge 0} W^k.$$

Proof. By (67), the definition of the Z_i and (68) holds

$$\bigcirc(-R\cdot f(\tilde{x}) + (I - R\cdot f'(\tilde{x}+Y))\cdot X) \subsetneq X \quad \text{using Einzelschrittverfahren (69)}$$

as described in theorem 18 with the rounding $\bigcirc$ defined in chapter 1. Therefore by theorem 12 the matrix R and every matrix $M \in \mathbb{MR}$ with $M \in f'(\tilde{x}+Y)$ is not singular. Furthermore by (67),(69) and theorem 12

$$\{x - R\cdot(f(\tilde{x}) + f'(\tilde{x} + Y)\cdot x) \mid x \in Y\} \subseteq Y$$

and therefore by definition (66) and lemma 27

$$\{x - R\cdot f(\tilde{x} + x) \mid x \in Y\} \subseteq Y. \tag{70}$$

By Brouwer's Fixed Theorem there is an $\hat{y} \in Y$ with $\hat{y} - R\cdot f(\tilde{x} + \hat{y}) = \hat{y}$ implying $f(\tilde{x} + \hat{y}) = 0$ by the non-singularity of R. With $\hat{x} := \tilde{x} + \hat{y}$ the proof is complete observing (69) and (70). □

5. Implementation and examples. Following some implementation hints and numerical examples will be given. Implementation details will be given for systems of linear equations; they apply for systems of nonlinear equations as well.

To calculate an inclusion of a linear system $Ax = b$ first an approximate inverse R is required according to theorem 18. Note, that R can be replaced by LU from an LU-decomposition. The success of the algorithm, i.e. whether an inclusion will be computed or not, depends highly on R. In fact, as theorem 22 shows, an inclusion will be found if and only if $I - RA$ is convergent, However, it is important to note, that the user does not have to know a priori, whether R or A is not singular, he does not have to know in advance, whether the spectral radius of $I - RA$ is less than one. This will be demonstrated by the algorithm automatically a posteriori.

In theorem 18, also an approximate solution $\tilde{x}$ is required. The number of iterations (3. in (41)) necessary depends, of course, on the quality of $\tilde{x}$. However, again no additional information on $\tilde{x}$ such as $\|\hat{x} - \tilde{x}\|$ is required.

Suppose, R and $\tilde{x}$ are given (e.g. computed using some traditional method, cf.[11],[42],[43]), then an algorithm based on theorem 17, on the third iteration in (41) and the ε-inflation would be the following:

```
X := ◇(b - Ax̃); Z := R ⊙ X; C := ◇(I - R·A); X := Z; k := 0;
repeat k := k+1; Y := X ⊛ ε; X := Z ⊕ C ⊙ Y until X ⊆° Y or k > 15;
```

Algorithm 1. Traditional way.
$\diamond : \mathbb{P}\,\mathbb{V}\mathbb{R} \to \mathbb{I}\,\mathbb{V}\,G$ resp. $\diamond : \mathbb{P}\,\mathbb{M}\mathbb{R} \to \mathbb{I}\,\mathbb{M}\,G$ is any isotone rounding, i.e. $A \subseteq \diamond A$. As has been pointed out before, it is of utmost importance to compute the residuals $b - A\tilde{x}$ and $I - RA$

with one rounding. Preferably, a precise scalar product with maximum accuracy introducing only one rounding error is used (cf.[7],[8]). In this case the rounding $\diamond$ is best possible and therefore equal to $\diamondsuit$. X,Y and Z are interval vectors and C is an interval matrix; therefore the additional storage needed is $6n + 2n^2 + O(1)$. Next, the Einzelschrittverfahren can be used according to theorem 18 yielding the following algorithm.

```
X := ◇(b - Ax̃); Z := R ⊙ X; C := ◇(I - R·A); X := Z; k := 0;
repeat k := k + 1; X := X ⊛ ε; incl := true;
       for i := 1 to n do {Q := Z_i ⊖ C_i ⊙ X; if Q ∉° X_i then incl := false;
           X_i := A}
until incl or k > 15;
```

Algorithm 2. Einzelschrittverfahren

Here C_i denotes the i-th row of C, X_i the i-th component of X. Q is an interval and the additional storage needed is $4n + 2n^2 + O(1)$.

In order to avoid the additional $O(n^2)$ storage consider the following. When using interval vectors as subsets of $\mathbb{V}\mathbb{R}$ it has been shown in theorem 12 that

$$Z + \mathbb{C}\cdot X \subsetneq X \quad \text{for} \quad X, Z \in \mathbb{I}\,\mathbb{V}\mathbb{R};\ \mathbb{C} \in \mathbb{I}\,\mathbb{M}\mathbb{R} \tag{71}$$

already implies $\rho(|C|) < 1$ for every matrix $C \in \mathbb{C}$. Therefore, instead of (71),

$$|Z| + |\mathbb{C}|\cdot|X| < |X| \tag{72}$$

can be checked. Necessary and sufficient conditions for an iteration using (72) to stop have been given in [39].

A significant amount of storage can be saved because only the absolute value of X,Z and C are needed. Therefore the additional storage needed essentially halves to $3n + n^2 + O(1)$. Moreover, now the matrix C, which is a point matrix, can be computed in the storage of R. This is possible because $C := I - RA$ can be computed rowwise, storing the intermediate (row) result somewhere. After the first row of C has been computed, the first row of R is no longer needed. This leads to the following algorithm.

```
[x,y] := O(b - Ax̃); z := Δ(|R·[x,y]|);
for i := 1 to n do {x := R_i; R_i := Δ(|I_i - x·A|)}; x := z; k := 0;
repeat k := k + 1; x := x ⧊ ε; incl := true;
       for i:=1 to n do {q := z_i ⧊ R_i ⧋ x; if q≥x_i then incl := false;
                         x_i := q}
until incl or k > 15;
```

Algorithm 3. Checking (72) with Einzelschrittverfahren

Algorithms 1 and 2 imply $\hat{x} \in \tilde{x} \oplus X$ ($\hat{x} := A^{-1} \cdot b$) whereas algorithm 3 implies $\hat{x} \in \tilde{x} \oplus [-x,x]$. The validity of algorithm 3 follows by

$$|Z|+|\mathbb{C}|\cdot|X|<|X| \Rightarrow -|X| < -|Z|-|\mathbb{C}|\cdot|X| \leq |Z|+|\mathbb{C}|\cdot|X| < |X| \Rightarrow$$
$$\Rightarrow [-|Z|,|Z|] + [-|\mathbb{C}|,|\mathbb{C}|]\cdot[-|X|,|X|] \overset{\circ}{\subseteq} [-|X|,|X|] \Rightarrow$$
$$\Rightarrow Z + \mathbb{C}\cdot[-|X|,|X|] \overset{\circ}{\subseteq} [-|X|,|X|].$$

All three algorithms verify the non-singularity of R and A and therefore the unique solvability of the linear system $Ax = b$. The additional storage required for algorithms 1 to 3 is:

algorithm	1	2	3
storage	$2n^2 + 6n + O(1)$	$2n^2 + 4n + O(1)$	$3n + O(1)$

For algorithm 3 the computing reduces as well.

The non-singularity of R and A could also be demonstrated by showing $\| I-RA \| < 1$ for some norm $\| \cdot \| : \mathbb{MR} \to \mathbb{R}$. On the computer, $\|\Delta(|I-RA|)\| < 1$ would be verified. In the following some numerical examples are shown where $\| I-RA \| \geq 1$ for a number of norms but nevertheless an algorithm based on the inclusion theory does verify the contraction and delivers sharp bounds for the solution of the linear system.

The following examples are calculated on an IBM S/370 in single precision (~ 7.5 decimal digits) and double precision (~16.5 decimal digits). The approximate inverse R is computed using Gauß-Jordan algorithm. As examples consider (n is the number of rows)

Hilbert* - matrices $H^*_{ij} := \dfrac{\mathrm{lcm}(1,2,\ldots,2n-1)}{i+j-1}$

Pascal - matrices $P_{ij} := \binom{i+j}{j}$

Pascal* - matrices $P^*_{ij} := \binom{i+j-1}{j}$

Zielke - matrices $Z_{ij} := \dfrac{\binom{n+i-1}{i-1}\cdot n \cdot \binom{n-1}{n-j}}{i+j-1}$ (cf.[50])

$S_{ij}(q) := 1 - q \cdot r_{ij}$ where $r_{ij} \in [0,1]$ randomly.

All matrices except the last have integer entries. Where the precision in use does not suffice to store an entry precisely, the entry rounded to nearest is used.

In the following tables the matrix type and the dimension n is listed. In the column $\| I-RA \|$ the minimum value of sum-, maximum- and Frobenius-norm is listed. k is the number of interval iterations and "verified digits" is the minimum number of decimal digits coinciding of the bounds for the components of the solution using algorithm 3 listed above. In all cases a linear system with the depicted matrix and right hand side $(1,\ldots,1)^T$ is solved. First the single precision results are displayed.

matrix	n	$\| I-RA \|$	k	verified digits
H	7	1.7	2	7.5
P	8	1.2	1	7.5
	9	38	2	7.3
P*	9	3.5	2	7.4
Z	7	1.6	3	7.3
S(10^{-5})	25	0.74	2	7.4

Table 1. Single precision results (~7.5 decimal digits)

As can be seen the number of verified decimal digits a widely independent on the condition of the matrix. This is due to the technique not to compute an inclusion of the solution itself but of the difference of the true solution $\hat{x}$ to an approximate solution $\tilde{x}$. Furthermore, the convergence of I - RA can be shown even if norm estimates fail to show $\rho(I-RA) < 1$. In the next table there are more extreme examples in this respect.

matrix	n	‖ I-RA‖	k	verified digits
P	20	11	1	16.5
	22	210	1	16.5
	24	670	1	16.5
	26	280000	1	16.5
$S(10^{-3})$	50	0.02	1	16.5
	100	0.03	1	16.3
	200	0.49	2	16.4

Table 2. Double precision results (~16.5 decimal digits)

The Pascal 26 x 26 matrix shows the extreme example where the spectral radius of I-RA is estimated by 280000 where in fact the new methods show $\rho(I-RA) < 1$. Also for larger dimensions the inclusions behave very stable.

6. Conclusion. In the preceding chapters the theoretical foundations of the inclusion theory were extended and new proofs were given without using Brouwer's Fixed Point Theorem. Some hints on the implementation were also provided. In [38] the theoretical background and corresponding algorithms are given for other numerical problems such as linear systems with band matrix, symmetric matrix or sparse matrix, for over - and underdetermined linear systems, evaluation and zeros of polynomials, algebraic eigenvalue problems, linear, quadratic and convex programming problems, evaluation of arithmetic expressions and others. Algorithms corresponding to a number of those problems are implemented in the IBM Program Product ACRITH, which is available since March 1984 and with a second Release since early 1985 (cf.[51]).

The key property of the new algorithms is that the verification of the validity of the result is performed automatically by the computer without any effort on the part of the user. The verification includes the existence and uniqueness of a solution within the computed bounds. The input data may be real point or interval data as well as complex point or interval data. Especially if the data is afflicted with tolerances the verification process is of great help. In this case it is

verified that any problem within the tolerances is solvable and the solution of any of the (infinitely many) problems within the tolerances is enclosed within the calculated inclusion interval.

The computing time is of the order of a comparable floating-point algorithm (e.g. Gaussian elimination in case of general linear systems with full matrix), the latter, of course, without the verification of the result.

The computed bounds are of high accuracy, i.e. the difference of the left and right bound of the inclusion of every component is of the order of the relative rounding error unit. By our experience, very often the inclusions are of least significant bit accuracy, i.e. the left and right bound of the inclusion of every component are adjacent floating-point numbers.

7. References

[1] Abbott, J.P., Brent, R.P. (1975). Fast Local Convergence with Single and Multistep Methods for Nonlinear Equations, Austr. Math. Soc. 19 (series B), 173-199.

[2] Alefeld, G., Intervallrechnung über den komplexen Zahlen und einige Anwendungen. Dissertation, Universität Karlsruhe, 1968.

[3] Alefeld, G. and Herzberger, J., "Einführung in die Intervallrechnung". Reihe Informatik, 12. Wissenschaftsverlag des Bibliographischen Instituts Mannheim, 1974.

[4] Alefeld, G. and Herzberger, J., "Introduction to Interval Analysis", Academic Press, New York (1983).

[5] Alefeld, G. (1979). Intervallanalytische Methoden bei nichtlinearen Gleichungen. In "Jahrbuch Überblicke Mathematik 1979", B.I. Verlag, Zürich.

[6] Bauer, F.L. and Samelson, K. Optimale Rechengenauigkeit bei Rechenanlagen mit gleitendem Komma, Z. Angew. Math. Phys. 4, 312-316 (1953).

[7] Bohlender, G., Floating-point computation of functions with maximum accuracy. IEEE Trans. Compute. C-26, No. 7, 621-632 (1977).

[35] Rall, L.B. (1981). Mean value and Taylor forms in interval analysis, SIAM J. Math. Anal. 14, No. 2 (1983).

[36] Reinsch, Ch., Die Behandlung von Rundungsfehlern in der Numerischen Analysis, "Jahrbuch Überblicke Mathematik 1979", Wissenschaftsverlag des Bibliographischen Instituts Mannheim, 43-62 (1979).

[37] Rump, S.M. (1980). Kleine Fehlerschranken bei Matrixproblemen, Dissertation, Universität Karlsruhe.

[38] Rump, S.M. (1983). Solving Algebraic Problems with High Accuracy, Habilitationsschrift, in Kulisch/Miranker: A New Approach to Scientific Computation, Academic Press, New York.

[39] Rump, S.M. (1982). Solving Non-linear Systems with Least Significant Bit Accuracy, Computing 29, 183-200.

[40] Rump, S.M. (1984). Solution of Linear and Nonlinear Algebraic Problems with Sharp, Guaranteed Bounds, Computing Suppl. 5, 147-168.

[41] Rump, S.M. and Kaucher, E., Small bounds for the solution of systems of linear equations, Computing Suppl. 2, 157-164 (1980).

[42] Stoer, J. (1972). Einführung in die Numerische Mathematik I. Heidelberger Taschenbücher, Band 105, Springer-Verlag, Berlin-Heidelberg-New York.

[43] Stoer, J., Bulirsch, R. (1973). Einführung in die Numerische Mathematik II. Heidelberger Taschenbücher, Band 114, Springer-Verlag, Berlin-Heidelberg-New York.

[44] Ullrich, Ch., Zur Konstruktion komplexer Kreisarithmetiken Computing Suppl. 1, 135-150 (1977).

[45] Varga, R.S. (1962). Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, New Jersey.

[46] Walter, W. (1970). Differential and integral Inequalities. Berlin-Heidelberg-New York: Springer.

[47] Wilkinson, J.H., "Rounding Errors in Algebraic Processes". Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[48] Wongwises, P. Experimentelle Untersuchungen zur numerischen Auflösung von linearen Gleichungssystemen mit Fehlerfassung, Interner Bericht 75/1, Institut für Praktische Mathematik, Universität Karlsruhe.

[49] Yohe, J.M., Roundings in floating-point arithmetic, IEEE Trans. Comput. C.12 No. 6, 577-586 (1973).

[50] Zielke, R., Algol-Katalog Matrizenrechnung, Oldenburg Verlag, München, Wien (1972).

[51] ACRITH High-Accuracy Arithmetic Subroutine Library: General Information Manual, IBM Publications, GC33-6163, (1985).

Acknowledgement: The author wants to thank his students of the summer lecture 1985 for several helpful comments.

Address of the author:

Priv.-Doz. Dr. Siegfried M. Rump
IBM Development and Research
Schönaicher-Straße 220

D-7030 Böblingen
Federal Republic of Germany