Computing 29, 183—200 (1982)

Computing

Solving Nonlinear Systems with Least Significant Bit Accuracy

S. M. Rump, Karlsruhe

Received May 12, 1980; revised May 4, 1982

Abstract — Zusammenfassung

Solving Nonlinear Systems with Least Significant Bit Accuracy. We give an algorithm for constructing an inclusion of the solution of a system of nonlinear equations. In contrast to existing methods, the algorithm does not require properties which are difficult to verify such as the non-singularity of a matrix. In fact this latter property is demonstrated by the algorithm itself. The highly accurate computational results are obtained in terms of a residue of first or higher order of the system.

AMS Subject Classifications: 65H10, 65G05, 65D99.

Key words: Automatic verification, existence, uniqueness, inclusion, rounding error, condition number, residue.

Einschließung der Lösung nichtlinearer Gleichungssysteme mit hoher Genauigkeit. Im folgenden wird ein Algorithmus zur Konstruktion einer Einschließung einer Lösung eines nichtlinearen Gleichungssystems angegeben. Im Gegensatz zu bekannten Methoden benötigt der Algorithmus keine schwierig verifizierbaren Voraussetzungen wie etwa die Nichtsingularität einer Matrix. Tatsächlich wird diese Eigenschaft vom Algorithmus automatisch verifiziert. Die Ergebnisse des Algorithmus zeichnen sich durch hohe Genauigkeit aus. Diese wird durch Residuen (eventuell höherer Ordnung) erreicht.

0. Introduction

Let T be one of the sets $\mathbb{R}$ (real numbers), $V\mathbb{R}$ (real vectors with n components) or $M\mathbb{R}$ (real $n\times n$-matrices). In the power set $\mathbb{P}T$ operations are defined by

$$A * B := \{a * b \mid a \in A, b \in B\} \quad \text{for } A, B \in \mathbb{P}T,\ * \in \{+, -, \cdot, /\},$$

($/$ only for $T=\mathbb{R}$). The order relation in $\mathbb{R}$ is extended to $V\mathbb{R}$ and $M\mathbb{R}$ by

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

Now the sets $\mathbb{I}T$ of intervals over $\mathbb{R}$, $V\mathbb{R}$ or $M\mathbb{R}$ are defined by

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

So we have $[A, B] \in \mathbb{P}T$ and $\mathbb{I}T \subseteq \mathbb{P}T$. We consider (see [11]) a rounding $\square : \mathbb{P}T \to \mathbb{I}T$ with the properties

(R) $\forall A \in \mathbb{P}T$: $\quad \square A = \cap \{B \in \mathbb{I}T \mid A \subseteq B\}$

(R1) $\forall A \in \mathbb{I}T$: $\quad \square A = A$

(R2) $\forall A, B \in \mathbb{P}T$: $\quad A \subseteq B \Rightarrow \square A \subseteq \square B$

0010-485X/82/0029/0183/$ 03.60

$(R3)$ $\forall A \in \mathbb{P}T$: $\quad A \subseteq \square A$

$(R4)$ $\forall \phi \neq A \in \mathbb{P}T$: $\quad \square(-A) = -\square(A)$.

$(R1)$, $(R2)$ and $(R3)$ are (together) equivalent to (R). Operations $\boxtimes$ for $* \in \{+, -, \cdot, /\}$ in $\mathbb{I}T$ are defined by (cf. [11])

$$(RG)\ \forall A, B \in \mathbb{I}T: \quad A \boxtimes B := \square(A * B) \quad (= \cap \{C \in \mathbb{I}T \mid A * B \subseteq C\}) \tag{1}$$

By means of semimorphisms (cf. [11]) it can be shown, that the operations in $\mathbb{I}T$ are well-defined (with well-known restrictions for /). In an expression operations of the same priority are to be executed from left to right.

To be perfectly clear take the following example. Let $z \in V\mathbb{R}$, $\mathscr{C} \in \mathbb{I}M\mathbb{R}$ and $X \in \mathbb{I}V\mathbb{R}$. Then e.g. $z + \mathscr{C} \boxdot X$ is well-defined regarding $z \in V\mathbb{R} \subseteq \mathbb{I}V\mathbb{R} \subseteq \mathbb{P}V\mathbb{R}$ with the cannonical embedding $V\mathbb{R}$ in $\mathbb{I}V\mathbb{R}$. Following the rules of priorities first $Y := \mathscr{C} \cdot X$ is computed with the multiplication $\cdot : \mathbb{P}M\mathbb{R} \times \mathbb{P}V\mathbb{R} \to \mathbb{P}V\mathbb{R}$ and rounded with $\square : \mathbb{P}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ with result in $\mathbb{I}V\mathbb{R}$. Then $z + Y$ is computed with $+ : \mathbb{P}V\mathbb{R} \times \mathbb{P}V\mathbb{R} \to \mathbb{P}V\mathbb{R}$ using the cannonical embedding $V\mathbb{R}$ in $\mathbb{P}V\mathbb{R}$ and $\mathbb{I}V\mathbb{R}$ in $\mathbb{P}V\mathbb{R}$. Because $z \in V\mathbb{R}$ and $Y \in \mathbb{I}V\mathbb{R}$ moreover the result is an element of $\mathbb{I}V\mathbb{R}$. So, for instance,

$$z + \mathscr{C} \cdot X \subseteq z + \mathscr{C} \boxdot X = z \boxplus \mathscr{C} \boxdot X.$$

If S denotes a subset of $\mathbb{R}$ (e.g. the set of single-precision floating-point numbers on a computer), we consider the set VS of n-tuples over S and the set MS of n^2-tupels over S. Let U denote one of the sets S, VS or MS. Then intervals over one of these sets U are defined by

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

where T is the corresponding set to U. The order relation in U is defined canonically regarding U as a subset of T. We consider a rounding $\diamondsuit : \mathbb{I}T \to \mathbb{I}U$ having the same properties $(R), \ldots, (R3)$, respectively (cf. [11]). If U is symmetric $(U = -U)$, then $(R4)$ is also satisfied. The operations $\circledast$ for $* \in \{+, -, \cdot, /\}$ in $\mathbb{I}U$ are defined by

$$A \circledast B := \diamondsuit(A \boxtimes B) \quad \text{for } A, B \in \mathbb{I}U.$$

It can be shown, that

$\circledast$ is well-defined

$\circledast$ is effectively implementable on a computer and

$A \circledast B = \cap \{C \in \mathbb{I}U \mid A \boxtimes B \subseteq C\}$ for $A, B \in \mathbb{I}U$.

These important properties are shown by means of algebraic and order isomorphisms $\mathbb{I}V\mathbb{R} \leftrightarrow V\mathbb{I}\mathbb{R}$, $\mathbb{I}MS \leftrightarrow M\mathbb{I}S$ etc. pp. (the operations in $V\mathbb{I}\mathbb{R}$, $M\mathbb{I}S$ etc. are defined componentwise), with the canonical embeddings $U \subseteq T \subseteq \mathbb{I}T \subseteq \mathbb{P}T$ and $\mathbb{I}U \subseteq \mathbb{I}T \subseteq \mathbb{P}T$ and by explicitely giving algorithms for the operations $\circledast$ in all sets S, VS, MS, $\mathbb{I}S$, $V\mathbb{I}S$ and $M\mathbb{I}S$. For the latter purpose a precise arithmetic and Bohlender's algorithm (cf. [3], [11]) are required.

Let A, B be elements of $\mathbb{I}U$, $\mathbb{P}U$, $\mathbb{I}T$ or $\mathbb{P}T$. Then

$$A \subsetneqq B :\Leftrightarrow \quad A \subseteq B \text{ and } A \neq B,$$

where the $\dotplus$-sign is to be understood componentwise. $\mathring{A}$ denotes the topological interior of A. For $A=[a,b]\in\mathbb{I}T$ with $a,b\in T$ the diameter $d(A)$ and the absolute value $|A|$ are elements of T defined by

$$d(A):=b-a\in T \text{ and } |A|:=\max(|a|,|b|)\in T,$$

where the maximum is to be understood componentwise. For $A\in V\mathbb{R}$ (or $M\mathbb{R}$), $|A|\in V\mathbb{R}$ (or $M\mathbb{R}$) is obtained by taking absolute values componentwise.

1. Basic Theorems

In the following we give a theorem for the existence and a theorem for the uniqueness of a fixed point of a function in a given region. The first theorem is cited from [7]:

Theorem 1: *Let the continuous function $f: V\mathbb{R}\to V\mathbb{R}$ and the mapping $F:\mathbb{P}V\mathbb{R}\to\mathbb{P}V\mathbb{R}$ have the property*

$$A\in\mathbb{P}V\mathbb{R} \text{ and } x\in A\to f(x)\in F(A). \tag{2}$$

If

$$F(X)\subseteq X, \tag{3}$$

for an interval vector $X\in\mathbb{I}V\mathbb{R}$, then f has at least one fixed point $x\in X$.

Moreover

$$\hat{x}\in\bigcap_{k\geq 0} F^k(X), \tag{4}$$

where $F^0(X):=X$ and $F^{k+1}(X):=F(F^k(X))$ for $0\leq k\in N$.

Proof: By Brouwer's Fixed Point Theorem every continuous injection of a non-empty, bounded, closed and convex region $X\subseteq\mathbb{P}V\mathbb{R}$ has a fixed point $\hat{x}$ in X. From (2) and (3) obviously

$$f(X):=\{f(x)\,|\,x\in X\}\subseteq F(X)\subseteq X.$$

By induction over k

$$\hat{x}\in F^k(X)\Rightarrow\hat{x}=f(\hat{x})\in F(F^k(X))=F^{k+1}(X) \text{ for } k\geq 0$$

demonstrating the theorem. □

Notice, that apart from having the property (2) F is an arbitrary mapping from $\mathbb{P}V\mathbb{R}$ to $\mathbb{P}V\mathbb{R}$. One way to obtain such a mapping from f itself is to substitute interval operations in the computation of f for the corresponding real operations. We call this process the "interval arithmetic evaluation of f".

Using appropriate functions f and F, theorem 1 leads to algorithms giving inclusions[1] of the solution of certain problems. Before describing these algorithms the theorem for the uniqueness of the solution will be given. For this we need some preliminaries.

[1] By an inclusion of an object we mean an interval which contains that object.

Let $f: V\mathbb{R} \to V\mathbb{R}$ be continuously differentiable. Then for $x, \tilde{x} \in V\mathbb{R}$ there exist $t_1, \ldots, t_n \in \mathbb{R}$ with $0 < t_i < 1$ for $i = 1(1)n$ such that

$$f(x) = f(\tilde{x}) + \begin{pmatrix} f_1'(\tilde{x} + t_1(x - \tilde{x})) \\ \vdots \\ f_n'(\tilde{x} + t_n(x - \tilde{x})) \end{pmatrix} \cdot (x - \tilde{x}). \tag{5}$$

This is a n-dimensional version of the Mean-value Theorem with

$$f_i'(x) := \operatorname{grad} f_i(x) = \frac{\partial f_i}{\partial x_1}(x) \ldots \frac{\partial f_i}{\partial x_n}(x).$$

The following important lemma has been introduced in [15]:

Lemma 2: *Let $Z \in \mathbb{I}V\mathbb{R}$, $\mathscr{C} \in \mathbb{I}M\mathbb{R}$ and $X \in \mathbb{I}V\mathbb{R}$. If*

$$Z + \mathscr{C} \cdot X \subsetneq X, \tag{6}$$

then the spectral radius of every $C \in \mathscr{C}$ is less than one.

Proof: From (6) we obtain

$$d(\mathscr{C} \cdot X) < d(X) \tag{7}$$

(cf. [2]), where the $<$-sign is to be understood componentwise. On the other hand formula (18) on p. 153 in [2] gives

$$d(\mathscr{C} \cdot X) \geq |\mathscr{C}| \cdot d(X).$$

Combining this with (7) yields

$$|\mathscr{C}| \cdot d(X) < d(X).$$

Applying Corollary 3, p. 18 in [18] gives $\rho(|\mathscr{C}|) < 1$. For every $C \in \mathscr{C}$ the Perron-Frobenius Theory yields

$$\rho(C) \leq \rho(|C|) \leq \rho(|\mathscr{C}|) < 1. \qquad \square$$

In [8] a version of lemma 2 in Banach spaces is given using the slightly more stringent condition $Z + \mathscr{C} \cdot X \subseteq \mathring{X}$ instead of (6). The proof of this is similar to the original one in [15], but completely different from the here presented one.

Lemma 3: *Let $f: V\mathbb{R} \to V\mathbb{R}$ be continuous and let $R \in M\mathbb{R}$ be an arbitrary $n \times n$ matrix. Define*

$$g: V\mathbb{R} \to V\mathbb{R},\ g(x) := x - R \cdot f(x). \tag{8}$$

Let $G: \mathbb{I}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ be given such that

$$I \in \mathbb{I}V\mathbb{R} \text{ and } x \in I \Rightarrow g(x) \in G(I). \tag{9}$$

If

$$G(X) \subseteq \mathring{X} \tag{10}$$

for an interval vector $X \in \mathbb{I}V\mathbb{R}$, then there exists a zero $\hat{x} \in X$ of $f: f(\hat{x}) = 0$.

Proof: In every ε-neighbourhood of R there exists a non-singular matrix. (This can be proved by regarding the determinant of R as a polynomial in n^2 variables which is continuous and not identically vanishing. The latter property holds since all coefficients of the polynomial are ± 1.) Therefore, a non-singular matrix $\bar{R}$ exists,

$$y-R\cdot f(y)=y-R\cdot\{f(\tilde{x})+M(y-\tilde{x})\}=$$
$$=\tilde{x}-R\cdot f(\tilde{x})+\{E-R\cdot M\}(y-\tilde{x})\in G(Y)$$

for some matrix $M\in f'(\tilde{x}\cup Y)$. Of course, M depends on y. However, M need not be a Jacobian matrix in the sense that no $x\in\tilde{x}\cup Y$ has to exist with $M=f'(x)$. Regarding (14) and applying theorem 1 the function $g:V\mathbb{R}\to V\mathbb{R}$ defined by $g(x):=x-R\cdot f(x)$ for $x\in V\mathbb{R}$ (which is continuous) has a fixed point $\hat{x}\in X$. Now for any matrix $M\in f'(\tilde{x}\cup X)$

$$\tilde{x}-R\cdot f(\tilde{x})+(E-R\cdot M)(X-\tilde{x})\subseteq G(X)\subsetneqq X \tag{15}$$

follows from (13) and (14). Setting $Z_M:=\tilde{x}-R\cdot f(\tilde{x})-\{E-R\cdot M\}\cdot\tilde{x}\in V\mathbb{R}$ and $C_M:=E-R\cdot M\in M\mathbb{R}$ we get

$$Z_M+C_M\cdot X\subsetneqq X.$$

Now lemma 2 yields that the spectral radius of C_M is less than one and henceforth

the matrix R and every matrix $M\in f'(\tilde{x}\cup X)$ are non-singular.

Thus the fixed point $\hat{x}\in X$ of g is a zero of f. For any $\hat{y}\in X$ with $f(\hat{y})=0$ setting $x=\hat{y}$ and $\tilde{x}=\hat{x}$ in (5) yields

$$0=f(\hat{y})-f(\hat{x})=M\cdot(\hat{y}-\hat{x})$$

for some matrix $M\in f'(\tilde{x}\cup X)$. The non-singularity of every such matrix M implies the uniqueness of the zero $\hat{x}$ of f in X. □

The essential purport of theorem 4 is, that R and $\tilde{x}$ are in no way restricted. In particular the non-singularity of R is demonstrated, if (14) holds (in contrast to [9] and [13], where the non-singularity must be assumed). In practice, proving the non-singularity of a matrix on a computer is a difficult problem.

2. Applications and Improvements

Theorem 4 and its proof can be applied to solve non-linear equations. However, some essential improvements should be introduced before giving an algorithm. It is far better to aim at an inclusion for a correction than at an inclusion of the solution itself directly (cf. [15]).

Corollary 5: *Let $f:V\mathbb{R}\to V\mathbb{R}$ be continuously differentiable, $R\in M\mathbb{R}$ be an arbitrary $n\times n$-matrix and $\tilde{x}\in V\mathbb{R}$ be an arbitrary vector. Define $G:\mathbb{P}V\mathbb{R}\to\mathbb{P}V\mathbb{R}$ for $Y\in\mathbb{P}V\mathbb{R}$ by*

$$G(Y):=-R\cdot f(\tilde{x})+\big(E-R\cdot f'(\tilde{x}\boxplus Y)\big)\cdot Y. \tag{16}$$

Here $f'(V)$ for $V\in\mathbb{P}V\mathbb{R}$ is defined as in theorem 4. If

$$G(X)\subsetneqq X \tag{17}$$

for an interval vector $X\in\mathbb{I}V\mathbb{R}$ holds, then the equation $f(x)=0$ has one and only one solution $\hat{x}\in V\mathbb{R}$ in $\tilde{x}\boxplus X$.

Proof: Obvious.

Applying corollary 5 instead of theorem 4 yields far better inclusions of the solution $\hat{x}$ because a relative error of X in $\tilde{x} \boxplus X$ plays a less important role. Of course, it is possible to determine an inclusion of a correction of higher order yielding $\tilde{x}_1 \boxplus \ldots \boxplus \tilde{x}_n \boxplus X$ as an inclusion of the solution. For details see [15].

With the abbreviations $Z_M := \tilde{x} - R \cdot f(\tilde{x}) - \{E - RM\} \cdot \tilde{x} \in V\mathbb{R}$ and $C_M := E - RM \in M\mathbb{R}$ formula (14) is equivalent to

$$Z_M + C_M \cdot X \subsetneqq X \quad \text{for every } M \in f'(\tilde{x} \cup X).$$

If $X = -X$ this is equivalent to

$$-|X| \leq Z_M + |C_M| \cdot X \leq |X|$$

with at most one equality in each component. $Z_M + |C_M| \cdot x$ is isotone and in this way theorem 4 and corollary 5 can be regarded as an extension of the Kantorovich Lemma.

Theorem 4 and corollary 5 can be extended in several ways. For instance the proofs do not change, if $\mathbb{I}V\mathbb{R}$ is replaced by any subset of the powerset $\mathbb{P}V\mathbb{R}$ consisting of closed, bounded and convex sets. Likewise all assertions remain true upon replacing $\mathbb{R}$ by $\mathbb{C}$. The proofs are similar. Regarding the first remark one may use rectangular or circular complex arithmetic. The hypothesis that f be continuously differentiable may be replaced by weaker conditions. (11) is nothing other than the Newton Method if R is the inverse of the Jacobian matrix $f'(\tilde{x})$. If the matrix is left fixed, we have the simplified Newton Method. In the algorithm $\tilde{x}$ will be an approximation of a zero of f and R will be an approximate inverse of the Jacobian matrix $f'(\tilde{x})$. The important problem of finding a function G satisfying (9) has been solved by (13) and (16). Any function $H: \mathbb{P}V\mathbb{R} \to \mathbb{P}V\mathbb{R}$ with

$$V \in \mathbb{P}V\mathbb{R}: G(V) \subseteq H(V)$$

is suitable because $H(X) \subsetneqq X$ implies immediately $G(X) \subsetneqq X$. This may, for instance, be

$$H(V) := \boxminus R \boxdot f(\tilde{x}) \boxplus \{E \boxminus R \boxdot f'(\tilde{x} \boxplus V)\} \cdot V \tag{18}$$

instead of (16), where every operation is the usual interval operation.

Defining

$$H^*(W) := \diamondminus R \diamondcdot \langle f \rangle(\tilde{x}) \diamondplus \{E \diamondminus R \diamondcdot \langle f' \rangle(\tilde{x} \diamondplus W)\} \diamondcdot W \text{ for } W \in \mathbb{I}VS$$

where $\langle f \rangle$ and $\langle f' \rangle$ denotes the computation of f and f' when replacing every operation by the corresponding operation in $\mathbb{I}S$, then

$$H^*(W) \subsetneqq W \text{ implies } G(W) \subseteq H(W) \subseteq H^*(W) \subsetneqq W$$

and the conclusions of corollary 5 are true. Therefore corollary 5 is applicable on a computer.

The problem remains to find an interval vector X satisfying (17). One may try to take a small interval X around an approximate zero x and examine whether (17) is satisfied. If this is the case, we are finished. If not, we take $X := G(X)$ and continue:

$$\begin{array}{ll} \text{Choose} & X; \\ \textit{repeat} & Y := X;\ X := G(X) \\ \textit{until} & X \subsetneqq Y; \end{array} \tag{19}$$

The pivotal question is: Will this iteration terminate and when and moreover for which problems and under which conditions. We will now give some sufficient conditions for these objectives. However, such a criterion attacks the set of all possible problems and is therefore not adapted to given data. Thus for a given problem it is preferable simply to apply algorithm (19) rather than to check by a general sufficient condition, whether it will terminate or not. By the way, this determination can be performed by the computer.

Lemma 6: *For an arbitrary matrix $A \in M\mathbb{R}$ the following properties are equivalent:*

(i) $\exists\, c \in V\mathbb{R}\ \exists\, X \in \mathbb{I}V\mathbb{R}$: $\quad c \boxplus A \boxdot X \subseteq \mathring{X}$

(ii) $\exists\, Y \in \mathbb{I}V\mathbb{R}$: $Y = -Y$ *and* $A \boxdot Y \subseteq \mathring{Y}$

(iii) $\exists\, y \in V\mathbb{R}$: $0 < y$ $\quad$ *and* $|A| \cdot y < y$.

Proof: (i) $\Rightarrow$ (ii). Consider $Y := X \boxminus m(X)$, where $m(X)$ is the midpoint of X. Then $Y = -Y$ and (cf. [2])

$$A \boxdot Y = A \boxdot (X \boxminus m(X)) = A \boxdot X \boxminus A \boxdot m(X) \subseteq \mathring{X} \boxminus c \boxminus A \boxdot m(X).$$

Abbreviating $v := m(X) \boxminus c \boxminus A \boxdot m(X) \in V\mathbb{R}$ yields

$$A \boxdot Y \subseteq \mathring{Y} \boxplus v.$$

Now (ii) follows immediately from $A \boxdot Y = -A \boxdot Y$.

(ii) $\Rightarrow$ (iii). Take $y = |Y|$ and observe that $d(Y) > 0$.

(iii) $\Rightarrow$ (i). Obvious. □

If one of the equivalent conditions of lemma 6 holds, then by lemma 2 the spectral radii of $|A|$ and A are both less than one. The next lemma states, when this conclusion is reversible. For definitions cf. [18].

Lemma 7: *If for an arbitrary matrix $A \in M\mathbb{R}$ there is a positive eigenvector y of $|A|$, then the conditions (i), (ii), (iii) and*

(iv) $$\rho(|A|) < 1$$

are equivalent.

Proof: It suffices to show (iv) $\Rightarrow$ (iii). However, this is obvious. □

Corollary 8: For a nonnegative, irreducible matrix or for a positive matrix the conditions (i), (ii), (iii) and (iv) are equivalent.

So the existence of an interval vector X satisfying (ii) is equivalent to (iv). Next we deal with the question, whether for *every* $X \in \mathbb{I}V\mathbb{R}$ with $X = -X$ the iteration (19) will terminate.

Lemma 9: *Given a nonnegative, primitive (and therefore irreducible) matrix* $A \in M\mathbb{R}$ *the following conditions are equivalent:*

(a) $\forall\, Y \in \mathbb{I}V\mathbb{R}$ *with* $Y = -Y$ *and* $|Y| > 0\ \exists k \in \mathbb{N} : A^{k+1} \boxdot Y \subseteq A^k \mathbin{\dot{\boxdot}} Y$

(b) $\rho(A) < 1$.

Proof: By Perron-Frobenius Theory there exists a positive eigenvalue λ of A with linear elementary divisor such that $\lambda = \rho(A)$. Since the corresponding eigenvector is positive, by lemma 7 it suffices to prove (b) $\Rightarrow$ (a).

Let $Y \in \mathbb{I}V\mathbb{R}$ with $Y = -Y$ and $|Y| > 0$ be given and $y := |Y| \in V\mathbb{R}$. The eigenvector v of A^T belonging to λ is also positive. So in particular $v^T \cdot y \neq 0$. Applying Theorem 8.3.1, p. 325 in [6] yields therefore that the iteration

$$y^0 := y;\ y^{k+1} := \frac{A \cdot y^k}{\| A \cdot y^k \|} \tag{20}$$

converges for every norm $\| \cdot \|$ to a vector $u \in V\mathbb{R}$ with

$$A \cdot u = \lambda \cdot u \quad \text{and} \quad \| u \| = 1.$$

From (20) we get for $k \geq 0$

$$y^{k+1} = A^{k+1} \cdot y^0 \cdot \left\{ \prod_{i=0}^{k} \| A y^i \| \right\}^{-1}. \tag{21}$$

By assumption $\lambda < 1$ and therefore there is a positive vector $\varepsilon \in V\mathbb{R}$ with

$$\varepsilon + A\varepsilon < (1 - \lambda) u. \tag{22}$$

From the cited Theorem there is a $k \in \mathbb{N}$ with

$$y^k = u + x \quad \text{and} \quad |x| < \varepsilon. \tag{23}$$

So by (20), (22) and (23) we have

$$\begin{aligned} \| A y^k \| \cdot y^{k+1} = A y^k = \lambda u + A x = y^k - u - x + \lambda u + A x < \\ < y^k - (1 - \lambda) u + \varepsilon + A\varepsilon < y^k. \end{aligned} \tag{24}$$

The assertion (a) follows now from a short computation using (21) and the fact that $Y = -Y$. □

Again, lemma 9 remains true if, instead, A is assumed to be positive.

Lemma 10: *Let* $A \in M\mathbb{R}$ *be a nonnegative, primitive matrix and denote* $\lambda = \rho(A)$. *Consider the mapping* $f : \mathbb{I}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ *defined by*

$$f(V) := z \boxplus A \boxdot V \quad \text{for} \quad V \in \mathbb{I}V\mathbb{R} \text{ and some fixed } z \in V\mathbb{R}.$$

Then the following are equivalent:

(A) *For all* $X \in \mathbb{I}VR$ *with* $|z \boxplus A \boxdot m(X) \boxminus m(X)| < (1 - \lambda) \cdot |X \boxminus m(X)|$ *there exists a* $k \in \mathbb{N}$ *with*

$$f^{k+1}(X) \subseteq f^{k}(X) \tag{25}$$

(B) $\rho(A) < 1$.

Proof: We need only to prove (B) $\Rightarrow$ (A). Let $X \in \mathbb{I}V\mathbb{R}$ be given satisfying

$$|z \boxplus A \boxdot m(X) \boxminus m(X)| < (1-\lambda) \boxdot |X \boxminus m(X)|.$$

Let us abbreviate

$$\tilde{x} := m(X),\ Y := X \boxminus \tilde{x},\ y := |Y| \text{ and } v := |z + A\tilde{x} - \tilde{x}|.$$

By assumption, there is an $\alpha \in \mathbb{R}$ with $0 < \alpha < 1-\lambda$ and $v \leq \alpha \cdot y$. Again by Theorem 8.3.1, p. 325 in [6] there is a positive vector $\varepsilon \in V\mathbb{R}$ with

$$(1+\alpha)\varepsilon + A \cdot \varepsilon < (1-\lambda-\alpha) \cdot u, \tag{26}$$

where u is the limit vector of the iteration (20). As in the proof of the preceding lemma there exists a $k \in \mathbb{N}$ with

$$y^k = u + x \text{ and } |x| < \varepsilon.$$

For this k we conclude from (26) that

$$\alpha y^k + \varepsilon + A\varepsilon < \alpha u + (1+\alpha)\varepsilon + A\varepsilon < (1-\lambda)u$$

and therefore

$$y^k - (1-\lambda)u + \varepsilon + A\varepsilon < (1-\alpha)\, y^k. \tag{27}$$

Now as in (24) we get from (27)

$$\| A y^k \| \cdot y^{k+1} < (1-\alpha) \cdot y^k. \tag{28}$$

Therefore by (21) we have

$$A^{k+1} \cdot y < (1-\alpha) \cdot A^k \cdot y$$

and

$$A^k \cdot v + A^{k+1} \cdot y < A^k \cdot y. \tag{29}$$

Using $Y = -Y$ we obtain in the original notation

$$A^k \cdot (z + A\tilde{x} - \tilde{x}) \boxplus A^{k+1} \boxdot (X \boxminus \tilde{x}) \subseteq A^k \boxdot (X \boxminus \tilde{x})$$

and

$$A^k \boxdot z \boxplus A^{k+1} \boxdot X \subseteq A^k \boxdot X.$$

To complete the proof we observe that

$$f^{k+1}(X) = \left(\sum_{i=0}^{k} A^i \right) \cdot z + A^{k+1} \boxdot X. \qquad \square$$

Lemma 10 states, that (25) is always satisfied for some $k \in \mathbb{N}$ if the absolute value of z is not too large compared with $d(X)$.

Theorem 11: *Let $\mathscr{C} \in \mathbb{I}M\mathbb{R}$ be an interval matrix. Then there is always an $0 < \varepsilon \in \mathbb{R}$ such that with the matrix $M \in M\mathbb{R}$ defined by*

$$M = (M_{ij}) \text{ with } M_{ij} := \begin{cases} |\mathscr{C}_{ij}| & \text{if } |\mathscr{C}_{ij}| \neq 0 \\ \varepsilon & \text{if } |\mathscr{C}_{ij}| = 0 \end{cases} \tag{30}$$

the following statements are equivalent:

(I) $\forall\, Y \in \mathbb{I}V\mathbb{R}$ *with* $Y = -Y$ *and* $|Y| > 0\ \exists k \in \mathbb{N} : M^{k+1} \boxdot Y \subseteq M^k \dot{\boxdot} Y$

(II) $\forall\, C$ *with* $|C| \leq |\mathscr{C}|$ *is* $\rho(C) < 1$.

Proof: As mentioned above the determinant of a matrix is a continuous function in the n^2 components. Thus, iff (II) is true then $\rho(M)<1$ for a certain $\varepsilon\in\mathbb{R}$. Because $M>0$ we can apply lemma 9. (I) $\Rightarrow$ (II) follows by Perron-Frobenius Theory:

$\rho(C)\leq\rho(|\mathscr{C}|)\leq\rho(|M|)$ for all $C\in\mathscr{C}$. □

The size of ε can be estimated by Hadamard's estimation of the maximum value of the determinant of a matrix and an estimation on a polynomial minimum root separation (see [16]).

Now an algorithm can be given to prove the non-singularity of a matrix $A\in M\mathbb{R}$:

(1) Compute an approximate inverse $R\in M\mathbb{R}$ of A;
(2) Compute $\mathscr{C}:=\square(E-R\cdot A)$ by interval arithmetic and define $M\in M\mathbb{R}$ by (30).
(3) $Y^0:=([-1,1])$; $k:=-1$;
(4) *repeat* $k:=k+1$; $X:=Y$; Compute $Y:=M\boxdot X$ by interval arithmetic; *until* $Y\subseteq\mathring{X}$ *or* $k>k_{\max}$
(5) *if* $Y\subseteq\mathring{X}$ *then* {A is not singular}

Algorithm 1: Non-Singularity of a matrix

Here $\square(E-R\cdot A)$ is computed with one final rounding in every component using Bohlender's algorithm (cf. [3], [11]). When applying algorithm 1 on a computer $\mathbb{R}$ and the roundings $\square$ are to be replaced by S and $\diamondsuit$. Here $Y\subseteq\mathring{X}$ and $Y\subsetneq X$ are equivalent because all X are symmetric. The condition $Y\subseteq\mathring{X}$ for $X, Y\in\mathbb{I}V\mathbb{R}$ is equivalent to $\lambda X<\lambda Y$ and $\rho Y<\rho X$ where λX, ρX is the vector of left, right components of X, resp. The approximate inverse R may be computed by any floating-point algorithm. The constant $k_{\max}$ might be estimated using the proof of lemma 9. In practice it turnes out, that if $\rho(M)<1$ then $Y\subseteq\mathring{X}$ is satisfied for $k\leq 5$. By the way, the condition $Y\subseteq\mathring{X}$ for $k=0$ is equivalent to $\|M\|_1<1$.

Of course, the algorithm depends on the accuracy of R. Computational remark: only $d(X)$ has to be computed because $X=-X$ for every $k\in\mathbb{N}$.

The following lemma demonstrates what might happen when replacing M by $|\mathscr{C}|$ in algorithm 1.

Lemma 12: *For every starting vector* $0=X^0\neq -X^0$ *there is a nonnegative, irreducible* $\mathscr{C}\in M\mathbb{R}$ *with* $\rho(\mathscr{C})<1$ *and* $X^{k-1}\not\subseteq X^k$ *for every* $k\in\mathbb{N}$, *where* $X^{k+1}:=\mathscr{C}\cdot X^k$ *for* $k\geq 0$.

Proof: We first assume $n=2$. Let $\sigma(X^0)=(x,y)'$. Then define

$$a:=\frac{x}{y}+1,\ b:=(a+1)^{-1} \quad \text{if } y\neq 0$$

$$b:=\frac{y}{x}+1,\ a:=(b+1)^{-1} \quad \text{if } y=0.$$

Take

$$\mathscr{C}:=\begin{pmatrix}0 & a\\ b & 0\end{pmatrix}\in M\mathbb{R}\subseteq\mathbb{I}M\mathbb{R}.$$

Short computation shows $\rho(\mathscr{C})<1$ and

$$\mathscr{C}^{2k}\cdot\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}a^k b^k x\\a^k b^k y\end{pmatrix} \text{ and } \mathscr{C}^{2k+1}\cdot\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}a^{k+1} b^k y\\a^k b^{k+1} x\end{pmatrix} \text{ for } k\geq 0.$$

Then for $\underline{v}=(x,y)'$ the statements

$$\mathscr{C}^{2k+1}\cdot v\subseteq\mathscr{C}^{2k}\cdot v \text{ and } \mathscr{C}^{2k+1}\cdot v\subseteq\mathscr{C}^{2k+1}\cdot v$$

are individually equivalent to the statement

$$(ay\geq x \quad or \quad bx\geq y).$$

By definition of a and b this statement is false. Similar examples are easily constructed for $2<n\in\mathbb{N}$. □

After this excursion into the question of whether an $X\in\mathbb{I}V\mathbb{R}$ satisfying (17) can be found and whether (19) terminates for some or any initial X we give an important improvement for algorithmic application on a computer.

Definition 13: Let $I\in\mathbb{I}S$ and $0<\varepsilon\in S$. The ε-inflation $I\circ\varepsilon$ is defined by

$$I\circ\varepsilon:=\begin{cases} I\diamondplus[-\varepsilon,\varepsilon]\diamondtimes d(I) & \text{if } d(I)\neq 0\\ [\mathrm{pred}(\mathrm{pred}(a)),\ \mathrm{succ}(\mathrm{succ}(a))] & \text{if } d(I)=0 \text{ and } I=[a,a],\ a\in S.\end{cases}$$

Here pred(a) resp. succ(a) are the predecessor resp. successor of a in the floating-point screen S, so definition 13 is machine dependent. E.g. if $d(I)=0$ and $I=[x,x]$, then the left and right hand bounds of $I\circ\varepsilon$ differ by 4 in the last digit of the mantissa and $m(I\circ\varepsilon)=x$. The ε-inflation for interval vectors is defined componentwise. Introducing the ε-inflation the inner loop of algorithm 1 becomes

repeat $k:=k+1$; $X:=Y\circ\varepsilon$;
Compute $Y:=M\diamondtimes X$ by interval arithmetic
until $Y\subseteq\mathring{X}$ *or* $k>k_{\max}$;

In practice, $\varepsilon=0.1$ turned out to be a suitable value.

The ε-inflation has been introduced and discussed in [15]. It reduces the number of "interval iterations" (see step 3 in the succeeding algorithm 2) significantly; in fact with using the ε-inflation this number is almost always 1.

3. The Algorithm

Now an algorithm verifying existence and uniqueness (in a certain domain) of the solution of a system of non-linear equations will be described. Suppose, $f\colon V\mathbb{R}\to V\mathbb{R}$ (continuously differentiable) is given.

(1) Use your favourite floating-point algorithm to compute an approximate solution $\tilde{x}$ of $f(x)=0$.

(2) Compute an approximate inverse R of the Jacobian matrix $f'(\tilde{x})$ by any floating-point algorithm.

(3) $Y:=([0])$; $k:=0$; $\varepsilon:=0.1$;
Compute $Z:=\diamondsuit f\diamondsuit(\tilde{x})$ with interval arithmetic;
Compute $Z:=-R\diamondsuit Z$ with interval arithmetic;
repeat $k:=k+1$; $Y:=Y\circ\varepsilon$; $X:=Y$;
Compute $D:=\diamondsuit f\diamondsuit(\tilde{x}\oplus X)$ with interval arithmetic;
Compute $Y:=Z\oplus\diamondsuit\{E-R\cdot D\}\diamondsuit X$ with interval arithmetic;
until $Y\subsetneqq X$ or $k>10$;

(4) *if* $Y\subsetneqq X$ *then* {It has been verified, that there is one and only one solution of $f(x)=0$ in $\tilde{x}\oplus Y$}

Algorithm 2: Zeros of non-linear systems

In the preceding number some necessary and sufficient conditions have been given in order that algorithm 2 yield an inclusion. These are a priori criteria. Algorithm 2 should be regarded as a "sufficient criterion". This criterion can be checked on a computer. In this sense it is an "a posteriori criterion".

In all interval computations the precise scalar product according to Bohlender's algorithm is used where possible in order to achieve sharp results (cf. [3], [11]). Especially $E-R\cdot D$ is computed with one final rounding in every component.

When implementing algorithm 1 resp. 2 it is preferable (to save memory) to use an Einzelschrittverfahren in step (4) resp. step (3). All statements made remain true according to the following lemma:

Lemma 14: *Let $Z\in\mathbb{P}V\mathbb{R}$, $\mathscr{C}\in\mathbb{P}M\mathbb{R}$ and $X\in\mathbb{P}V\mathbb{R}$. Denote the i-th row of $\mathscr{C}$ by $\mathscr{C}_i$, the i-th component of Z resp. X by Z_i resp. X_i and define recursively for $Y\in\mathbb{I}V\mathbb{R}$*

$$Y_i:=Z_i+\mathscr{C}_i\cdot(Y_1,\ldots,Y_{i-1},X_i,\ldots,X_n)' \quad \textit{for } i=1(1)n. \tag{31}$$

If

$$Y\subsetneqq X, \tag{32}$$

then for every $C\in\mathscr{C}$, we have $\rho(C)<1$.

Proof: Set $C:=|\mathscr{C}|$ and assume C to be irreducible. Then

$$Z_i+C_i\cdot(Y_1,\ldots,Y_n)'\subseteq Z_i+C_i\cdot(Y_1,\ldots,Y_{i-1},X_i,\ldots,X_n)=Y_i \tag{33}$$

and therefore using (19), p. 154 in [2] we obtain

$$Z+C\cdot Y\subseteq Y \text{ and } C\cdot d(Y)=d(C\cdot Y)\leq d(Y).$$

Since C is irreducible we conclude that (33) is a proper inclusion for $i=1$ and therefore

$$C_1\cdot d(Y)=d(C_1\cdot Y)<d(Y_1).$$

Thus by the Corollary on p. 21 in [18] we have $\rho(C)<1$ and the assertion follows by Perron-Frobenius Theory.

If C is reducible there is a permutation matrix P with

$$PCP'=\begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1m} \\ & R_{22} & \ldots & R_{2m} \\ & & \ldots\ldots & \\ & & & R_{mm} \end{pmatrix},$$

where each R_{ii} is a square matrix and either irreducible or a 1×1 null matrix, $i=1(1)m$. The eigenvalues of C are the eigenvalues of the R_{ii}, $i=1(1)m$. If $R_{11}=(0)$, then $\rho(R_{11})=0$. Suppose $R_{11}\neq(0)$ and therefore that R_{11} is irreducible. Set $M:=R_{11}$, say M is a $k\times k$ matrix and the first k components of $P\cdot(1,2,\ldots,n)'$ are $i_1,\ldots,i_k$. W.l.o.g. we assume $i_1<i_j$, $j=2,\ldots,k$. Then

$$(PCP'\cdot Y)_1=\sum_{j=1}^{k} M_{1j}\cdot Y_{ij}+S\subsetneqq\sum_{j=1}^{k} M_{1j}\cdot X_{ij}+S\subseteq Y_{i_1},$$

where S is a sum of first rows of R_{1j} multiplied by the corresponding components of Y, $j=1(1)m$. As in the first part of the proof we get

$$M\cdot(Y_{i_1},\ldots,Y_{ik})\leq d(Y_{i_1},\ldots,Y_{ik})$$

with strict inequality in the first row. Because $M=R_{11}$ is irreducible we get $\rho(R_{11})<1$. The same proof is applicable to any R_{ii}, $i=2(1)m$. □

A few remarks on applications to some special problems should be added. For details see [15] and [4]. Application on a computer is immediately obtained in replacing the roundings $\square$ by $\diamondsuit$.

Given a system of linear equations $Ax=b$ (16) is written (cf. [15])

$$H(X):=R\boxdot\square(b-A\tilde{x})\boxplus\square\{E-RA\}\boxdot X.$$

Here $\tilde{x}$ is an approximate solution of $Ax=b$ and R an approximate inverse of A. Notice, that the Jacobian matrix is A and is constant. Therefore $E-RA$ has to be computed only once. The computing time of algorithm 2 using this fact is 6 times the computing time of the Gaussian Algorithm. There are no a priori restrictions.

To compute inclusions of eigenvectors and eigenvalues of a matrix A we use the nonlinear system (cf. [15])

$$\begin{aligned}(A-\lambda E)x&=0\\ e_k'x-\zeta\quad&=0.\end{aligned}$$

Here e_k' is the transposed k-th unit vector, i.e. $e_k'x$ the k-th component of x. ζ is a fixed but arbitrary real number.

In order to improve corollary 5 we introduce the following sharper definition of H

$$H\begin{pmatrix}X\\ \Lambda\end{pmatrix}:=-R\boxdot\begin{pmatrix}\square((A-\tilde{\lambda}E)\cdot\tilde{x})\\ 0\end{pmatrix}\boxplus\square\left(E-R\cdot\begin{pmatrix}A-\tilde{\lambda}E & \boxminus\tilde{x}\boxminus X\\ e_k' & 0\end{pmatrix}\right)\cdot\begin{pmatrix}X\\ \Lambda\end{pmatrix}$$

for $X\in\mathbb{IVR}$, $\Lambda\in\mathbb{IR}$. Nevertheless using this new definition we can prove the uniqueness of the eigenvector and the corresponding eigenvalue in the resp. sets $\tilde{x}\boxplus X$ and $\tilde{\lambda}\boxplus\Lambda$ separately (cf. [15]).

Special algorithms can be given for computing inclusions of the (real or complex) zeros of a polynomial. To compute the inclusion of a solution of $p(x)=0$ for a polynomial $p\in\mathbb{R}[x]$ one can use for arbitrary $\tilde{x}$, $r\in\mathbb{R}$ instead of (16) the following definitions of G:

$$G(X):=-p(\tilde{x})\boxslash p'(X) \tag{34}$$

or

$$G(X):=-r\boxdot p(\tilde{x})\boxplus\square\{1-r\cdot p'(X)\}\boxdot X. \tag{35}$$

If $G(X) \subseteq \mathring{X}$ using definition (34) or $0 \in X$ and $G(X) \subsetneqq X$ using definition (35) for an $X \in \mathbb{I}\mathbb{R}$, then $\tilde{x} \boxplus X$ contains one and only one real zero of $p(x)$. These results remain true for complex $\tilde{x}, r$ as follows analogously using complex interval arithmetic. There are several different algorithms for computing inclusions of real or complex zeros (singly or simultaneously) using Frobenius matrices. For details see [4].

There are special applications involving the computation of inclusions of the values of functions in several variables consisting of $+, -, \cdot, /, (,)$ for any values of the variables. For further details cf. [19]. Example:

$$\text{Compute } 4x^4 - y^4 + 2y^2 \text{ for } x = 470832 \text{ and } y = 665857.$$

If in a formula an algebraic expression like $\sqrt{2}$ occurs it may be exactly represented by another equation $y^2 - 2 = 0$. Thus values of algebraic polynomials can be computed (see [5]). If $p \in Q(\alpha)[x]$ with $\Psi(\alpha) = 0$ for $\Psi \in Q[x]$ then one writes

$$\Psi(x) = 0$$
$$p(x) - y = 0.$$

All statements made remain true if the data consists of intervals. Given $\mathscr{E}: \mathbb{I}VR \to \mathbb{I}V\mathbb{R}$, $\mathscr{E}': \mathbb{I}V\mathbb{R} \to \mathbb{I}M\mathbb{R}$, an interval vector $X \in \mathbb{I}V\mathbb{R}$ and an arbitrary $n \times n$-matrix R, if we have

$$-R \boxdot \mathscr{E}(\tilde{x}) \boxplus \square \{E - R \cdot \mathscr{E}'(\tilde{x} \boxplus X)\} \boxdot X \subsetneqq X,$$

then every continuously differentiable function $f: V\mathbb{R} \to V\mathbb{R}$ with

$$x \in X \Rightarrow \{f(x) \in \mathscr{E}(x) \quad \text{and} \quad f'(x) \in \mathscr{E}'(X)\}$$

has one and only one zero $\hat{x}$ in $\tilde{x} \boxplus X$. Moreover the matrix R and all matrices in $\mathscr{E}'(X)$ are non-singular. These statements are obvious regarding the proof of theorem 4. E.g., for sets of systems of linear equations $\{Ax = b \mid A \in \mathscr{A}, b \in \mathscr{b}\}$ with $\mathscr{A} \in \mathbb{I}M\mathbb{R}$, $\mathscr{b} \in \mathbb{I}V\mathbb{R}$ this means, that the solution of every system $Ax = b$ with $A \in \mathscr{A}$ and $b \in \mathscr{b}$ is uniquely determined and is an element of $\tilde{x} \boxplus X$. All matrices $A \in \mathscr{A}$ and R are non-singular.

All statements made remain true taking the field of complex numbers instead of reals. In fact, the used mean-value theorem remains true (in the form we need it) in the complex space $\mathbb{C}^n$ (see [4]).

4. Computational Results

The following examples are computed using a computer having 12 decimal digits for the mantissa and an optimal floating-point arithmetic (see [11]).

To solve a very ill-conditioned problem we take a linear system with the Hilbert 9×9-matrix. Instead of taking $h_{ij} := 1/(i+j-1)$ we multiply all components with the l.c.m. of all denominators. Solving a linear system with right hand sides e_i (i-th unit vector) we obtain the columns of the inverse of the Hilbert matrix. Using algorithm 2 we compute inclusions for all components of the inverse, where the distance of the left and right bounds is always one unit of the last decimal place, i.e. we obtain the smallest possible intervals. Two iterations are necessary in step 3. The sum norm of $E - RA$ is greater than one, so the non-singularity of R and A cannot be proved a priori.

However, with a floating-point algorithm it is no problem to compute "an inverse" of a singular matrix. Algorithm 2 gives notice, that inclusion is not possible and the problem should be handled very carefully. Consider

$$\begin{aligned} -8392848 \cdot x \quad & -3566221 \cdot y \quad -3799934 \cdot z = -15759003 \\ 1699109 \cdot x \quad & +3679519 \cdot y \quad +2370515 \cdot z = \quad 7749143 \\ -6693739 \cdot x \quad & + \quad 113298 \cdot y \quad -1429419 \cdot z = -8009860. \end{aligned}$$

Applying a double-precision floating point residue iteration to the approximation given by Gaussian elimination yields $x^0 = x^1 = \ldots$, the iteration remains constant. This suggests the best possible condition whereas in fact (add first and second row) it is the worst!

As well as "finding a solution" where none exists a bad approximation may be passed as a good one. Consider the following matrix:

$$A := \begin{pmatrix} 941664.000002 & 665857 \\ 665857 & 470832 \end{pmatrix}.$$

To invert this matrix we use the Gauss-Jordan procedure. This yields the following "approximate inverse":

$$\begin{pmatrix} -166\,666 & 235\,702 \\ 235\,702 & -333\,333 \end{pmatrix}$$

rounded to six figures (after 12-decimal-digit computation). The true inverse is

$$\begin{pmatrix} -8\,071\,037 & 11\,414\,170 \\ 11\,414\,170 & -16\,142\,074 \end{pmatrix}$$

and is computed up to the last figure by algorithm 2. For more examples see [15].

For examples concerning arithmetic expressions see [19].

The following examples for non-linear systems were computed on the UNIVAC 1108 of the University of Karlsruhe. The mantissa length on this machine is $8\frac{1}{2}$ decimal digits. The following examples were treated:

1. Example 4 (Rosenbrock) in [1]
2. Example 7 in [1] for $n = 10, 20, 50, 100$
3. The example in [13]
4. Problem 1 in [14] for $n = 10, 20, 50$
5. Problem 2 in [14] for $n = 10, 20, 50$
6. Problem 3 (Brown) in [14] for $n = 10, 20, 50$ for initial approximations
   a) $(0.5, \ldots, 0.5)$ b) $(1 - n^{-1}, \ldots, 1 - n^{-1}, 2)$
7. Example 1 (Rosenbrock) in [12]
8. Example 2 (Brown) in [12] for initial approximations
   a) $(1, -0.5)$; b) $(1.5, 0.8)$; c) $(-0.5, 1)$
9. Example 3 (Brown) in [12] for initial approximations
   a) $(0.1, 2)$; b) $(1, 0)$; c) $(0, 2)$

10. Example 4 (Brown/Conte) in [12]
11. The example in [10] for initial approximations
    a) (0.7, 0.1, 0.8, 0.1); b) (1, 1, 1, 1)

We first tried to improve the approximations using a floating-point Newton iteration. Then we applied algorithm 2. The columns of the table refer to:

1. Number of the problem
2. n: Number of functions and variables
3. Newton-steps: Number of floating-point Newton-iterations; here a^* indicates that the floating-point iteration didn't converge
4. interval steps: $k \cong$ how often step 3 in algorithm 2 has been executed
5. succeeded: $X \subsetneqq Y$ in algorithm 2, i.e. the computer verified the existence and uniqueness of a zero of the non-linear system in $X := \tilde{x} \boxplus Y$
6. digits guaranteed: least number of digits, for which the left and right bounds of X coincide; here an additional l.s.b.a. means "least significant bit accuracy", i.e. between all left and right bounds of the inclusion there is no other floating-point number.

Problem	n	Newton-steps	interval steps	succeeded	digits guaranteed
1.	2	5	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
2.	10	8	2	yes	$8\frac{1}{2}$ (l.s.b.a.)
	20	6	2	yes	$8\frac{1}{2}$ (l.s.b.a.)
	50	6	3	yes	8
	100	7	3	yes	8
3.	2	3	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
4.	10	4	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	20	6	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	50	6	1	yes	8
5.	10	3	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	20	4	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	50	4	1	yes	8
6. a)	10	15*	10	no	
6. b)	10	8	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	20	8	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
	50	11	2	yes	8
7.	2	2	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
8. a)	2	15*	10	no	
8. b)	2	3	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
8. c)	2	6	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
9. a)	2	15*	10	no	
9. b)	2	4	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
9. c)	2	8	2	yes	$8\frac{1}{2}$ (l.s.b.a.)
10.	3	7	2	yes	$8\frac{1}{2}$ (l.s.b.a.)
11. a)	4	2	1	yes	$8\frac{1}{2}$ (l.s.b.a.)
11. b)	4	3	1	yes	$8\frac{1}{2}$ (l.s.b.a.)

The examples show, that if the initial approximation is not too bad algorithm 2 verifies existence and uniqueness within an inclusion, whose left and right bounds coincide to 8 decimal places when computing with $8\frac{1}{2}$ decimal digits accuracy. It never occured that algorithm 2 did not succeed in verifying existence and uniqueness after starting with a reasonable approximation.

References

[1] Abbott, J. P., Brent, R. P.: Fast local convergence with single and multistep methods for nonlinear equations. Austr. Math. Soc. *19*, 173–199 (1975).
[2] Alefeld, G., Herzberger, J.: Einführung in die Intervallrechnung. Mannheim-Wien-Zürich: Bibl. Inst. 1974.
[3] Bohlender, G.: Floating-point computation of functions with maximum accuracy. IEEE Trans. on Computers *1977*, 621.
[4] Böhm, H.: Berechnung von Schranken für Polynomwurzeln mit dem Fixpunktsatz von Brouwer. Interner Bericht des Inst. f. Angew. Math., Universität Karlsruhe, 1980.
[5] Collins, G. E.: Quantifier elimination for real closed fields by cylindrical algebraic decomposition (Lecture Notes in Computer Science, Vol. 33), pp. 134–183. Berlin-Heidelberg-New York: Springer 1975.
[6] Gastinel, N.: Lineare numerische Analysis. Braunschweig: F. Vieweg & Sohn 1972.
[7] Kaucher, E., Rump, S. M.: Generalized iteration methods for bounds of the solution of fixed-point operator equations. Computing *24*, 131–137 (1980).
[8] Kaucher, E., Rump, S. M.: E-methods for fixed point equations. Computing *28*, 31–42 (1982).
[9] Krawczyk, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing *4*, 187–210 (1969).
[10] Köberl, D.: The solution of non-linear equations by the computation of fixed points with a modification of the sandwich method. Computing *25*, 175–178 (1980).
[11] Kulisch, U., Miranker, W.: Computer arithmetic in theory and practice. Academic Press 1981.
[12] Martinez, J. M.: Solving nonlinear simultaneous equations with a generalization of Brent's method. *BIT 20*, 501–510 (1980).
[13] Moore, R. E.: A test for existence of solutions for non-linear systems. SIAM J. Numer. Anal. *4* (1977).
[14] Moré, J. J., Cosnard, M. Y.: Numerical solution of nonlinear equations. ACM Trans. on Math. Software, *5*, 64–85 (1979).
[15] Rump, S. M.: Kleine Fehlerschranken bei Matrixproblemen. Dr.-Dissertation, Inst. f. Angew. Math., Universität Karlsruhe, Februar 1980.
[16] Rump, S. M.: Polynomial minimum root separation. Math. of Comp. *33*, 327–336 (1979).
[17] Rump, S. M., Kaucher, E.: Small bounds for the solution of systems of linear equations. Computing, Suppl. 2. Wien-New York: Springer 1980.
[18] Varga, R. S.: Matrix iterative analysis. Englewood Cliffs, N.J.: Prentice-Hall 1962.
[19] Rump, S. M., Böhm, H.: Least significant bit evaluation of arithmetic expressions in single precision (to appear in Computing).

S. M. Rump
Institut für Angewandte Mathematik
Universität Karlsruhe
Kaiserstrasse 12
D-7500 Karlsruhe 1
Federal Republic of Germany

Printed in Austria
Druck: Paul Gerin, A-1021 Wien