Computing, Suppl. 5, 147–168 (1984)

Solution of Linear and Nonlinear Algebraic Problems with Sharp, Guaranteed Bounds

S. M. Rump, Böblingen

Abstract

In this paper new methods for solving algebraic problems with high accuracy are described. They deliver bounds for the solution of the given problem with an automatic verification of the correctness. Examples of such problems are systems of linear equations, over- and underdetermined systems of linear equations, algebraic eigenvalue problems, nonlinear systems, polynomial zeros, evaluation of arithmetic expressions, linear, quadratic and convex programming and others. The new methods apply for these problems over the space of real numbers, complex numbers as well as real intervals and complex intervals.

0. Introduction

In this paper we deal with errors in numerical computation and discuss possibilities for their elimination. The problems we have in mind may contain data exactly representable on a given computer ("point problems") or data afflicted with tolerances. Data which is not exactly representable on a given computer may either be rounded to the smallest enclosing interval; in this case an inclusion of the solution of the interval problem includes the solution of the given problem. Or, the data may be rounded to the nearest representable data; in this case the inclusion of the solution of the rounded point problem need not include the solution of the original problem. Of course, any accuracy claim can only refer to the solution of the given problem, with data represented on a given computer (the "specified problem").

Our aim are algorithms delivering error bounds for the solution of the specified problem with an automatic verification of the existence and uniqueness of the solution within these bounds. If this verification process fails a respective message should be given. Further the aim is to achieve least significant bit accuracy (abbreviated "lsba") for point problems and smallest possible bounds for interval problems.

The key feature of the new algorithms is that error control is performed automatically. This is done without any effort required on the part of the user. Therefore the algorithms can easily be used by non-specialists. Costly reruns, altering of data etc. is not necessary saving human and machine time. The efficiency of the algorithms is, for instance, demonstrated by inverting a Hilbert 21×21 matrix in a 14 hexadecimal digit floating-point system. This is after multiplying with a proper factor, the Hilbert matrix of largest dimension exactly storable in this

floating-point system. The error bounds for all components of the inclusion of the inverse are as small as possible, i.e. the left and right bounds are adjacent in the floating-point system. We call this least significant bit accuracy (lsba). Our experience shows, that the results of our algorithms very often have the lsba-property for every component of the inclusion of the solution.

For verified results with high accuracy a precisely defined computer arithmetic is necessary. Therefore we proceed with a short description of the arithmetic according to the Kulisch/Miranker theory. This arithmetic has the property of maximum accuracy for every single operation including the scalar product. The step from the single operation to a whole algorithm with results of high and verified accuracy is described in the succeeding chapters. It relies strongly on the computation of highly accurate residual values and on their use in iterated defect corrections, as well as on a proper use of interval analysis.

The algorithms have been implemented on a Z80-based minicomputer with 64 k Byte memory, on a UNIVAC 1108 and on IBM System /370. The minicomputer works with a 12 digit decimal mantissa and has been developed at the Institute for Applied Mathematics at the University of Karlsruhe (Prof. Dr. U. Kulisch) and at the Fachbereich Informatik at the University of Kaiserslautern (Prof. Dr. H.-W. Wippermann). So the algorithms are implemented on a decimal, a binary and a hexadecimal computer. All computational results are given for an IBM System /370 machine using ACRITH. This is a program product recently announced by IBM consisting of a collection of algorithms with the mentioned properties and an Online Training Component to get easy and quick access to the routines.

1. Computer Arithmetic

Let T be one of the sets $\mathbb{R}$ (real numbers), $V\mathbb{R}$ (real vectors with n components), $M\mathbb{R}$ (real square matrices with n columns), $\mathbb{C}$ (complex numbers), $V\mathbb{C}$ (complex vectors with n components) or $M\mathbb{C}$ (complex square matrices with n rows and columns). In the following the letter n is reserved to denote the number of components of a vector or the number of rows and columns of a square matrix. If the number of components of a vector is different from n this is denoted by an index, e.g. $V_{n+1}\mathbb{R}$. Non-square matrices with, e.g., l rows and m columns are denoted by $M_{l,m}\mathbb{R}$.

The operations in the power set $\mathbb{P}T$ are as usually defined by

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

with well-known restrictions for /. The order relation $\leqq$ is extended to $V\mathbb{R}$ and $M\mathbb{R}$ by

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

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

$$a+bi,\ c+di \in \mathbb{C}: \ a+bi \leqq c+di :\Leftrightarrow a \leqq c \ . \wedge . \ b \leqq d$$

and similarly in $V\mathbb{C}$ and $M\mathbb{C}$.

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

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

Therefore $[A, B] \in \mathbb{P}T$ and $\mathbb{I}T \subseteq \mathbb{P}T$. Every element of $\mathbb{I}T$ is closed, convex and bounded. The following definitions are taken from the Kulisch/Miranker theory. A detailed description can be found in [KuMi80]. We consider a rounding $\bigcirc: \mathbb{P}T \to \mathbb{I}T$ with the following properties:

$$(R) \quad \forall A \in \mathbb{P}T: \quad \bigcirc A = \cap \{B \in \mathbb{I}T \mid A \subseteq B\}$$
$$(R1) \quad \forall A \in \mathbb{I}T: \quad \bigcirc A = A$$
$$(R2) \quad \forall A, B \in \mathbb{P}T: \quad A \subseteq B \Rightarrow \bigcirc A \subseteq \bigcirc B$$
$$(R3) \quad \forall A \in \mathbb{P}T: \quad A \subseteq \bigcirc A$$
$$(R4) \quad \forall \emptyset \neq A \in \mathbb{P}T: \quad \bigcirc(-A) = -\bigcirc A.$$

The basic operations $\circledast: \mathbb{I}T \times \mathbb{I}T \to \mathbb{I}T$ for $* \in \{+, -, \cdot, /\}$ are defined by

$$(RG) \quad A, B \in \mathbb{I}T: \quad A \circledast B := \bigcirc(A * B) \quad (= \cap \{C \in \mathbb{I}T \mid A * B \subseteq C\}).$$

It can be shown, that the operations $\circledast$ are well defined (with well-known restrictions for /). The operations are executed from left to right respecting the usual priorities and considering the canonical embeddings $T \subseteq \mathbb{I}T \subseteq \mathbb{P}T$ and $\mathbb{R} \subseteq \mathbb{C}$, $V\mathbb{R} \subseteq V\mathbb{C}$ and $M\mathbb{R} \subseteq M\mathbb{C}$.

By S we denote some finite subset of $\mathbb{R}$ which can be regarded as the set of single precision floating-point numbers. We consider the set VS of n-tupels over S, MS of n^2-tupels over S, $\mathbb{C}S$ of pairs over S, $V\mathbb{C}S$ of n-tupels over $\mathbb{C}S$ and $M\mathbb{C}S$ of n^2-tupels over $\mathbb{C}S$. Let U denote one of the sets S, VS, MS, $\mathbb{C}S$, $V\mathbb{C}S$ or $M\mathbb{C}S$. Then intervals over U are defined by

$$[A, B] = :\{x \in T \mid A \leqq x \leqq B\} \quad \text{for } A, B \in U \text{ and } A \leqq B$$

where T is the correspondingset to U. The order relation in U is induced by the order relation in T. We consider a rounding $\diamond: \mathbb{I}T \to \mathbb{I}U$ having the same properties (R), $(R1)$, $(R2)$, $(R3)$, respectively (cf. [KuMi80]). If $U = -U$ then $(R4)$ is also satisfied.

The basic operations $\diamondsuit: \mathbb{I}U \times \mathbb{I}U \to \mathbb{I}U$ for $* \in \{+, -, \cdot, /\}$ are defined by

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

One of the essential results of the Kulisch/Miranker theory is that

$\diamondsuit$ is well defined
$\diamondsuit$ is effectively implementable on computers and
$A \diamondsuit B = \cap \{C \in \mathbb{I}U \mid A \circledast B \subseteq C\}$ for $A, B \in \mathbb{I}U$.

The latter property implies that the operations $\diamondsuit$ for $* \in \{+, -, \cdot, /\}$ are of maximum accuracy. The practical implementation requires a precise scalar product. Finally we consider a rounding $\square: T \to U$ with the properties

$$(R1) \quad \forall A \in U: \quad \square A = A$$
$$(R2) \quad \forall A, B \in T: \quad A \leqq B \Rightarrow \square A \leqq \square B$$
$$(R4) \quad \forall A \in T: \quad \square(-A) = -\square A.$$

The latter property requires $U=-U$. The basic operations in U are defined by

$$(RG)\quad a \boxed{*} b := \square(a * b) \quad \text{for } a,b \in U \text{ and } * \in \{+,-,\cdot,/\}.$$

Again, one of the essential results of the Kulisch/Miranker theory is, that

$\boxed{*}$ is well defined
$\boxed{*}$ is effectively implementable
the operations $\boxed{*}$ are of maximum accuracy.

The latter property can be demonstrated by the following lemma.

Lemma 1.1: *For $a,b,v \in U$,*

$$a \boxed{*} b \leqq v \leqq a * b \Rightarrow v = a \boxed{*} b \text{ and}$$
$$a * b \leqq v \leqq a \boxed{*} b \Rightarrow v = a \boxed{*} b$$

for $* \in \{+,-,\cdot,/\}$.

Proof: Because of symmetry we need only to proof the first assertion:

$$v \leqq a * b \underset{(R2)}{\Rightarrow} \square v \leqq \square(a * b) \underset{(R2, RG)}{\Rightarrow} v \leqq a \boxed{*} b. \qquad \square$$

The practical implementation of the operation $\boxdot$ requires a precise scalar product, the results of which are, under any circumstances, of maximum accuracy (cf. [Bo 77]).

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

$$A \subsetneqq B :\Leftrightarrow A \subseteq B \,.\wedge.\, A \neq B,$$

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

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

where the maximum is to be understood componentwise and the fact is used, that there are algebraic and order isomorphisms $\mathbb{I}V\mathbb{R} \leftrightarrow V\mathbb{I}\mathbb{R}$, $\mathbb{I}M\mathbb{R} \leftrightarrow M\mathbb{I}\mathbb{R}$ etc. For $X \in \mathbb{I}U$ with $X=[A,B]$, $A,B \in U$ we have

$$\inf(X) := A, \quad \sup(X) := B.$$

The "midpoint" of X is defined by (assuming $2 \in S$)

$$\overline{m}(X) := \inf(X) \boxplus (\sup(X) \boxminus \inf(X)) \boxslash 2. \tag{1.1}$$

For floating-point systems the assumption $2 \in U$ is satisfied for all existing machines. We choose definition (1.1) to achieve (cf. [RuBö 83])

$$\inf(X) \leqq \overline{m}(X) \leqq \sup(X)$$

and therefore $\overline{m}(X) \in X$. This is in general not the case when using

$$m^*(X) := (\inf(X) \boxplus \sup(X)) \boxslash 2.$$

I denotes the $n \times n$ identity matrix, I_k the $k \times k$ identity matrix. e_k denotes the k-th unit vector and $y_k := e_k^t \cdot y$ the k-th component of the vector y.

2. Inclusion Methods for Linear Systems

Let a system of linear equations $Ax=b$ for $A \in M\mathbb{R}$, $b \in V\mathbb{R}$ be given. Consider

$$f(x) := x + R(b-Ax), \tag{2.1}$$

where $R \in M\mathbb{R}$ is an approximate inverse of A. Then

$$f(x)=x \Rightarrow b-Ax \in \ker R. \tag{2.2}$$

The residual iteration (also named "Iterated Defect Correction") for $Ax=b$

$$x^{k+1} := f(x^k) = x^k + R(b-Ax^k) \tag{2.3}$$

converges if and only if

$$\rho(I-RA) < 1. \tag{2.4}$$

In this case $\ker R = \{0\}$ and a fixed point of f is a solution of $Ax=b$. Because A is not singular when (2.4) holds this solution is unique.

Theorem 2.1: *Let $A \in M\mathbb{R}$ and $b \in V\mathbb{R}$ be given. If for some $R \in M\mathbb{R}$ and for some norm $\|\cdot\| : M\mathbb{R} \to \mathbb{R}$*

$$\|I-RA\| < 1$$

and for some closed, nonempty $X \in \mathbb{P}V\mathbb{R}$

$$X + R \cdot (b-AX) \subseteq X, \tag{2.5}$$

then the matrices A and R are not singular and the unique solution $\hat{x}=A^{-1}b$ of $Ax=b$ satisfies

$$\hat{x} \in X.$$

Proof: The first part of the theorem is clear from (2.4). The function $f: V\mathbb{R} \to V\mathbb{R}$ defined by (2.1) is continuous and satisfies

$$x \in X \Rightarrow f(x) \in X + R \cdot (b-AX) \subseteq X$$

and

$$x,y \in V\mathbb{R} \Rightarrow \|f(x)-f(y)\| = \|(I-RA)\cdot(x-y)\| \leqq \|I-RA\| \cdot \|x-y\|.$$

So $\hat{x} \in X$ is demonstrated by the fixed point Theorem of Banach. $\square$

The operations in the above theorem are the power set operations. They can be replaced by machine executable operations as shown by the following corollary.

Corollary 2.2: *Let $A \in MS$ and $b \in VS$ be given. If for some $R \in MS$ and for some norm $\|\cdot\| : M\mathbb{R} \to \mathbb{R}$*

$$\|I-RA\| < 1 \tag{2.6}$$

and for some $X \in \mathbb{I}VS$

$$X \diamondsuit R \diamondsuit (b \diamondsuit A \diamondsuit X) \subseteq X,$$

then the matrices A and R are not singular and the unique solution $\hat{x}=A^{-1}b$ of $Ax=b$ satisfies

$$\hat{x} \in X.$$

The *proof* derives from the fact that

$$X + R \cdot (b-AX) \subseteq X \diamondsuit R \diamondsuit (b \diamondsuit A \diamondsuit X)$$

and the preceding theorem. $\square$

With a proper norm like the sum norm, maximum norm, Frobenius norm etc. which can be estimated on computers, Corollary 2.2 is applicable on computers. Condition (2.6) in the corollary could be replaced by the weaker condition $\rho(I-RA)<1$. But this cannot, in general, be verified on computers. To achieve stronger results we seek for some norm-independent condition to verify $\rho(I-RA)<1$.

Consider first the following theorem.

Theorem 2.3: *Let $f: V\mathbb{R}\to V\mathbb{R}$ be a continuous function and let $F: \mathbb{P}V\mathbb{R}\to\mathbb{P}V\mathbb{R}$ be given by*

$$X\in\mathbb{P}V\mathbb{R}: \quad x\in X\Rightarrow f(x)\in F(X). \tag{2.7}$$

If

$$F(X)\subseteq X$$

for a closed, bounded and convex set $\emptyset\neq X\in\mathbb{P}V\mathbb{R}$, then f has at least one fixed point $\hat{x}$ in X. Moreover

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

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

Proof: The first part of the theorem follows from the fixed point Theorem of Brouwer. The second part follows by induction:

$$\hat{x}\in F^k(X)\Rightarrow\hat{x}=f(\hat{x})\in F^{k+1}(X) \quad \text{for } k\geqq 0. \qquad \square$$

Up to now F has been an arbitrary mapping satisfying (2.7). It could, for instance, be obtained by replacing every operation in the computation of f by its respective interval operation. We call this process the "interval arithmetic evaluation" of f.

For the function f from (2.1) we obtain

$$f(X)\subseteq F(X):=X\oplus R\odot(b\ominus A\odot X). \tag{2.8}$$

However,

$$d(X\oplus R\odot(b\ominus A\odot X))=d(X)+d(R\odot(b\ominus A\odot X)),$$

so that (except in trivial cases) $F(X)\subseteq X$ is impossible. Therefore we have to solve two problems:

1. Replace (2.8) by another condition allowing inclusion.
2. Conclude from the fixed point of f to the solution of $Ax=b$.

The first problem can be solved by replacing (2.8) by (cf. [Kr69])

$$F(X):=R\odot b\oplus\{I\ominus R\odot A\}\odot X\subseteq X \tag{2.9}$$

which satisfies (2.7). However, the second aim requires a sharper assumption than (2.9). This is clear from the fact that $R\equiv 0$ would always satisfy assumption (2.9). Actually, a slightly sharper assumption will suffice.

Lemma 2.4: *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*

$$Z\oplus\mathscr{C}\odot X\subsetneqq X, \tag{2.10}$$

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

Proof: From (2.10) we obtain by the theory of interval arithmetic (cf. [AlHe74])

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

On the other hand formula (18) on p. 153 in [AlHe74] gives

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

Therefore

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

Now $\rho(|\mathscr{C}|)<1$ follows by Corollary 3 in [Va62] and by Perron/Frobenius Theory

$$\forall C\in\mathscr{C}: \quad \rho(C)\leqq\rho(|C|)\leqq\rho(|\mathscr{C}|)<1. \qquad \square$$

The preceding lemma applies to $F(X)$ of (2.9):

Theorem 2.5: *Let $A\in M\mathbb{R}$ and $b\in V\mathbb{R}$ be given. If for some $R\in M\mathbb{R}$ and for some $X\in\mathbb{I}V\mathbb{R}$*

$$R\odot b\oplus\{I\ominus R\odot A\}\odot X\subsetneqq X, \tag{2.11}$$

then the matrices A and R are not singular and the unique solution $\hat{x}=A^{-1}b$ of $Ax=b$ satisfies

$$\hat{x}\in X.$$

Proof: From (2.11) it follows for $Z:=R\cdot b$ and $C:=I-R\cdot A$

$$Z+C\cdot X\subseteq R\odot b\oplus\{I\ominus R\odot A\}\odot X\subsetneqq X.$$

Note, that in the definition of C the power set operations were used. By Lemma 2.6 $\rho(I-RA)<1$ and therefore R and A are not singular. $\hat{x}\in X$ follows as in Theorem 2.1. $\square$

For the numerical application there is still one essential disadvantage in using formula (2.11). Here the interval matrix $R\odot A$ has to be subracted from the identity matrix where $R\cdot A$ is supposed to approximate I. Because

$$d(I\ominus R\odot A)=d(I)+d(R\odot A)=d(R\odot A)$$

the results will be poor especially for ill-conditioned matrices. Defining $E\in M_{n,2n}\mathbb{R}$ and $F\in M_{2n,n}\mathbb{R}$ by

$$E_{ij}:=\begin{cases}\delta_{ij} & \text{for } 1\leqq i,j\leqq n\\ -R_{i,j-n} & \text{else}\end{cases} \quad\text{and}\quad F_{ij}:=\begin{cases}\delta_{ij} & \text{for } 1\leqq i,j\leqq n\\ A_{i-n,j} & \text{else}\end{cases} \tag{2.12}$$

we have

$$(E\cdot F)_{ij}=\sum_{\nu=1}^{n}\delta_{i\nu}\delta_{\nu j}+\sum_{\nu=1}^{n}-R_{i\nu}\cdot A_{\nu j}=(I-R\cdot A)_{ij} \tag{2.13}$$

and

$$E\odot F=\bigcirc(I-RA)=I-R\cdot A.$$

The transition to (2.13) implies that in the evaluation of $I-R\cdot A$ in U (where $\bigcirc$ is replaced by $\diamondsuit$) only one final rounding is performed on each component of $I-R\cdot A$. We denote this by $\diamondsuit\{I-R\cdot A\}$. For ill-conditional matrices this improvement is essential:

Theorem 2.6: *Let $A, R \in MS$ and $b \in VS$ be given. Define for $X \in \mathbb{I}VS$*

$$F: \mathbb{I}VS \to \mathbb{I}VS \text{ by } F(X) := R \diamondsuit b \diamondsuit (\diamondsuit \{I - R \cdot A\}) \diamondsuit X \tag{2.14}$$

using (2.13). *If for some $X \in \mathbb{I}VS$*

$$F(X) \subsetneqq X, \tag{2.15}$$

then the matrices A and R are not singular and the unique solution $\hat{x} = A^{-1} b$ of $Ax = b$ satisfies

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

Proof: By $R \cdot b + (I - RA) X \subseteq F(X)$ and Theorem 2.3. □

Theorem 2.6 can be applied on computers. The function F from (2.14) can be evaluated by using (2.13) and the Kulisch/Miranker arithmetic. If some interval vector X can be found satisfying (2.15) then it has been verified that the matrices A and R are not singular, the linear system is therefore uniquely solvable and that the solution lies in X.

A theorem similar to Theorem 2.6 holds for general convex, closed and bounded sets X instead of interval vectors. This is necessary when using other than rectangular interval arithmetics such as complex circular arithmetic, parallel-epiped arithmetic etc. In the general case the proper inclusion $\subsetneqq$ is not well defined. Instead we use $X \subseteq \mathring{Y}$, which is in the case of rectangular interval arithmetic a slightly sharper assumption.

Lemma 2.7: *Let $Z \in \mathbb{P}V\mathbb{R}$, $\mathscr{C} \in \mathbb{P}M\mathbb{R}$ and let $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$ be convex, closed and bounded. If*

$$Z + \mathscr{C} \cdot X \subseteq \mathring{X}, \tag{2.16}$$

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

Proof: Let $z \in Z$ and $C \in \mathscr{C}$ be arbitrarily chosen. Then $f: V\mathbb{R} \to V\mathbb{R}$ defined by $f(x) := z + C \cdot x$ satisfies by (2.16) the assumption of the fixed point Theorem of Brouwer. Therefore there exists a $\hat{x} \in \mathring{X}$ with $\hat{x} = z + C \cdot \hat{x}$. By (2.16) we have

$$C \cdot (X - \hat{x}) = C \cdot X - C \cdot \hat{x} = C \cdot X + z - \hat{x} \subseteq \mathring{X} - \hat{x}. \tag{2.17}$$

Substituting $Y := X - \hat{x}$ we get

$$C \cdot Y \subseteq \mathring{Y}. \tag{2.18}$$

Moreover there is an ε-neighborhood of 0 contained in Y because $\hat{x} \in \mathring{X}$: $U_\varepsilon(0) \subseteq Y$.

Let $U := Y + i \cdot Y$. Then

$$C \cdot U = C \cdot Y + i \cdot C \cdot Y \subseteq \mathring{Y} + i \cdot \mathring{Y} = \mathring{U} \tag{2.19}$$

and a complex ε-neighborhood of 0 is contained in U. If $C = 0$ then $\rho(C) = 0$. Assume $C \neq 0$ and let $\lambda \in \mathbb{C}$ be an arbitrary eigenvalue of C with corresponding eigenvector $v \in V\mathbb{C}$. Define $\Gamma \in \mathbb{P}\mathbb{C}$ by $\Gamma := \{\gamma \in \mathbb{C} \mid \gamma \cdot v \in U\}$. U is closed so Γ is closed and there is a $\gamma^* \in \Gamma$ with $|\gamma^*| = \max_{\gamma \in \Gamma} |\gamma|$. Then by (2.18)

$$C \cdot (\gamma^* v) = \gamma^* \cdot \lambda \cdot v \in \mathring{U}.$$

But $\gamma^* v \in \partial U$ because of the definition of γ^* and therefore

$$|\gamma^*| > |\gamma^* \cdot \lambda| \Rightarrow |\lambda| < 1. \qquad \square$$

The above lemma applies also to complex vectors.

Lemma 2.8: *Let $Z \in \mathbb{P}V\mathbb{C}$, $\mathscr{C} \in \mathbb{P}M\mathbb{C}$ and let $\emptyset \neq X \in \mathbb{P}V\mathbb{C}$ be convex, closed and bounded. If*

$$Z + \mathscr{C} \cdot X \subseteq \mathring{X},$$

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

The *proof* is similar to the one of the previous lemma. □

Next we can develop a theorem similar to Theorem 2.5 for general convex and compact subsets of $V\mathbb{R}$. The application on computers to other than rectangular interval arithmetics is obvious. We give two different proofs for the succeeding theorem to introduce different proving techniques.

Theorem 2.9: *Let $A \in M\mathbb{R}$ and $b \in V\mathbb{R}$ be given. If for some $R \in M\mathbb{R}$ and for some convex, closed and bounded $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$*

$$R \cdot b + \{I - R \cdot A\} \cdot X \subseteq \mathring{X}, \tag{2.20}$$

then the matrices A and R are not singular and the unique solution $\hat{x} = A^{-1} b$ of $Ax = b$ satisfies

$$\hat{x} \in \mathring{X}.$$

Proof 1: By Lemma 2.7 we have $\rho(I - RA) < 1$ and the non-singularity of A and R. The function $f: V\mathbb{R} \to V\mathbb{R}$ defined by $f(x) := x + R(b - Ax)$ satisfies the assumptions of the fixed point Theorem of Brouwer in X. The fixed point $\hat{x}$ of f in $\mathring{X}$ satisfies $b - A\hat{x} \in \ker R$ and because of the non-singularity of R we have $A\hat{x} = b$. □

Proof 2: As in the previous proof we see by the fixed point Theorem of Brouwer the existence of some $\hat{x} \in X$ with $b - A\hat{x} \in \ker R$. By (2.20) we have $\hat{x} \in \mathring{X}$. For some $y \in \ker A$ we have for $f(x) := x + R(b - Ax)$

$$f(\hat{x} + \lambda y) = \hat{x} + \lambda y + R(b - A\hat{x} - \lambda y) = \hat{x} + \lambda y \quad \text{for every } \lambda \in \mathbb{R}.$$

Every $\hat{x} + \lambda y$ is a fixed point of f and for $y \neq 0$ there would be a $\lambda \in \mathbb{R}$ with $\hat{x} + \lambda y \in \partial X$ contradicting (2.20). Therefore $\ker A = \{0\}$. Let $y \in \ker R$. Then

$$f(\hat{x} + \lambda (A^{-1} y)) = \hat{x} + \lambda (A^{-1} y) + R(b - A\hat{x} - A\lambda y) = \hat{x} + \lambda (A^{-1} y) \quad \text{for every } \lambda \in \mathbb{R}.$$

So every $\hat{x} + \lambda (A^{-1} y)$ is a fixed point of f and for $y \neq 0$ there would be a $\lambda \in \mathbb{R}$ with $\hat{x} + \lambda (A^{-1} y) \in \partial X$ contradicting (2.20). □

If we abandon the assertion of the non-singularity of R we can give a third proof of Theorem 2.11. We use the fact that there is always a non-singular matrix $\tilde{R}$ in every ε-neighborhood of R. If ε is small enough, then (2.19) is satisfied replacing R by $\tilde{R}$.

The preceding theorem remains true when replacing $\mathbb{R}$ by $\mathbb{C}$.

3. Implementation of Inclusion Algorithms

Our ultimate goal is the development of algorithms for systems of linear and nonlinear equations with the properties previously mentioned. The theoretical basis are the Theorems 2.6 and 2.9. For the practical implementation we need formulas similar to (2.14) which can be evaluated on computers. Here any operation "enclosing" the power set operations can be used like the interval operations of the Kulisch/Miranker theory or, for instance, a circular arithmetic in the complex space. For the special case of interval operations $\diamondsuit$ *for* $*\in\{+,-,\cdot,/\}$ the condition $\subsetneqq X$ suffices in (2.20) as we saw in Theorem 2.6.

To achieve the properties mentioned in the introduction we need

1. a proper choice of X and
2. result intervals of small diameter.

A first choice of a suitable X will be a small interval around an approximate solution of $Ax=b$. If for this first interval X° condition (2.20) is not satisfied, an iteration may be started:

$$\begin{aligned}&\textbf{repeat}\quad k:=k+1;\ X^{k+1}:=R\diamondsuit b\diamondsuit\diamondsuit\{I-R\cdot A\}\diamondsuit X^k\\ &\textbf{until}\quad X^{k+1}\subseteq\hat{X}^k;\end{aligned}\tag{3.1}$$

In [Ru82] conditions are given when this iteration stops and when not. Moreover, the condition $X^{k+1}\subseteq\hat{X}^k$ can be further weakened.

Theorem 3.1: *Let $A\in M\mathbb{R}$ and $b\in V\mathbb{R}$ be given. Define for some $R\in M\mathbb{R}$ the function $F:\mathbb{P}V\mathbb{R}\to\mathbb{P}V\mathbb{R}$ by*

$$F(X):=Rb+\{I-RA\}\cdot X\quad\text{for } X\in\mathbb{P}V\mathbb{R}\tag{3.2}$$

and define

$$F^\circ(X):=X;\ F^{k+1}(X):=F(F^k(X))\quad\text{for } 0\leqq k\in\mathbb{N}.\tag{3.3}$$

If then for some convex, closed and bounded $\emptyset\neq X\in\mathbb{P}V\mathbb{R}$

$$F^{k+m}(X)\subseteq F^k(X)\quad\text{for some } 1\leqq m\in\mathbb{N}\text{ and } k\in\mathbb{N},\tag{3.4}$$

then the matrices A and R are not singular and the unique solution $\hat{x}=A^{-1}b$ of $Ax=b$ satisfies

$$\hat{x}\in F^{k+m}(X).$$

Proof: Because of $F^{k+m}(X)=F^m(F^k(X))$ it suffices to proof the theorem for $k=0$. With the abbreviations $Z:=R\cdot b$ and $C:=I-RA$ we have

$$F(X)=Z+C\cdot X\text{ and by induction }F^m(X)=\sum_{i=0}^{m-1}C^i\cdot Z+C^m\cdot X.\tag{3.5}$$

By (3.4) and Lemma 2.7 we have $\rho(C^m)<1$ and therefore $\rho(C)<1$. By Theorem 2.3 F^m has a fixed point $\hat{x}\in F^m(X)$. For $\hat{x}$ holds by (3.4)

$$\sum_{i=0}^{m-1}C^i\cdot Z=(I-C^m)\cdot\hat{x}.$$

Multiplying from left by $(I-C)$ yields

$$(I-C^m)\cdot Z=(I-C)(I-C^m)\cdot\hat{x}=(I-C^m)(I-C)\cdot\hat{x}$$

and by the non-singularity of $I-C^m$

$$Z=(I-C)\hat{x}\quad\text{and}\quad Rb=RA\hat{x}\quad\text{and therefore}\quad A\hat{x}=b.$$ □

As demonstrated by the preceding theorem an inclusion in the interior of the last iterative can be replaced by an inclusion in the interior of any of the previous iteratives to achieve an inclusion of the solution. Next we will demonstrate, that even the inclusion in the interior of the convex union of all previous iteratives suffices to achieve an inclusion of the solution ($\sqcup$ denotes the convex union).

Lemma 3.2: *Let an affine function $f:V\mathbb{R}\to V\mathbb{R}$ and some $X\in\mathbb{P}V\mathbb{R}$ be given. Let*

$$Y:=X\sqcup f(X)\sqcup\ldots\sqcup f^k(X)\in\mathbb{P}V\mathbb{R}$$

for some $0\leqq k\in\mathbb{N}$ and assume $f^{k+1}(X)\subseteq Y$. Then $f(Y)\subseteq Y$.

Proof: Because f is affine $f(X\sqcup Y)=f(X)\sqcup f(Y)$ for $X, Y\in\mathbb{P}V\mathbb{R}$. Then

$$f(Y)=f\left(\bigcup_{i=0}^{k}f^i(X)\right)=\bigcup_{i=0}^{k}f^{i+1}(X)\subseteq Y.$$ □

Theorem 3.3: *Let $A\in M\mathbb{R}$ and $b\in V\mathbb{R}$ be given. Define for some $R\in M\mathbb{R}$ the function $F:\mathbb{P}V\mathbb{R}\to\mathbb{P}V\mathbb{R}$ by* (3.2). *If then for some convex, closed and bounded $\emptyset\neq X\in\mathbb{P}V\mathbb{R}$*

$$F^{k+1}(X)\subseteq\text{interior}\left\{\bigcup_{i=0}^{k}F^i(X)\right\}\quad\text{for some } 0\leqq k\in\mathbb{N},\tag{3.6}$$

then the matrices A and R are not singular and the unique solution $\hat{x}=A^{-1}b$ of $Ax=b$ satisfies

$$\hat{x}\in\text{interior}\left\{\bigcup_{i=0}^{k}F^i(X)\right\}.$$

Proof: By induction follows for $m\geqq 1$ using Lemma 3.2

$$F^{k+m+1}(X)=F(F^{k+m}(X))\subseteq F\left(\text{interior}\left\{\bigcup_{i=0}^{k}F^i(X)\right\}\right)\subseteq\text{interior}\left\{\bigcup_{i=0}^{k}F^i(X)\right\}$$

and therefore

$$F^{k+1}\left(\bigcup_{i=0}^{k}F^i(X)\right)\subseteq\text{interior}\left\{\bigcup_{i=0}^{k}F^i(X)\right\}.$$

Now Theorems 3.1 and 2.9 complete the proof. □

To apply Theorem 2.3 to computers we have to replace the power set operations by proper interval operations. However, the convex union of two interval vectors or matrices is not, in general, again an interval vector or matrix. Therefore we define a corresponding "interval convex union" by

$$\textcircled{\scriptsize\sqcup}:\mathbb{I}V\mathbb{R}\times\mathbb{I}V\mathbb{R}\to\mathbb{I}V\mathbb{R}\text{ and }X\textcircled{\scriptsize\sqcup}Y:=\cap\{Z\in\mathbb{I}V\mathbb{R}\mid X\sqcup Y\subseteq Z\}\text{ for }X,Y\in\mathbb{I}V\mathbb{R}.$$

Furthermore

$$F(X):=\diamondsuit\{R\cdot b+(I-RA)\cdot X\}\text{ for }F:\mathbb{I}VS\to\mathbb{I}VS$$

cannot be calculated on computers but some

$$\hat{F}: \mathbb{I}VS \to \mathbb{I}VS \text{ with } F(X) \subseteq \hat{F}(X) \text{ for } X \in \mathbb{I}VS.$$

With these considerations we seek to replace (3.6) by a formula which can be evaluated on computers.

Lemma 3.4: *For $X, Y \in \mathbb{I}V\mathbb{R}$ and $\Lambda \in \mathbb{I}M\mathbb{R}$ with $\Lambda := \operatorname{diag}([0,1])$,*

$$X \bigcirc Y = \{\lambda X + (1-\lambda) Y \mid \lambda \in \Lambda\}, \tag{3.7}$$

and for an affine function $f: V\mathbb{R} \to V\mathbb{R}$,

$$f(X \bigcirc Y) = f(X) \bigcirc f(Y). \tag{3.8}$$

Proof: The first part is clear and the second follows from

$$f(X \bigcirc Y) = \{f(\lambda X + (1-\lambda) Y) \mid \lambda \in \Lambda\} =$$

$$= \{\lambda \cdot f(X) + (1-\lambda) \cdot f(Y) \mid \lambda \in \Lambda\} = f(X) \bigcirc f(Y). \qquad \square$$

Next we need some "inflation" representing rounding errors and overestimations. We indicate this by some operator $G: \mathbb{P}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ with the property

$$X \in \mathbb{P}V\mathbb{R} \Rightarrow X \subseteq G(X). \tag{3.9}$$

Lemma 3.5: *Let an affine function $f: V\mathbb{R} \to V\mathbb{R}$, some operator $G: \mathbb{P}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ satisfying* (3.3) *and some $X \in \mathbb{I}V\mathbb{R}$ be given. Let*

$$Y := X \bigcirc \tilde{f}(X) \bigcirc \ldots \bigcirc \tilde{f}^k(X) \in \mathbb{I}V\mathbb{R}$$

for some $0 \leqq k \in \mathbb{N}$ and $\tilde{f} := G \circ f$. If then $\tilde{f}^k(X) \subseteq Y$, then $f(Y) \subseteq Y$.

Proof: By Lemma 3.4

$$f(Y) = f(X \bigcirc \bigbigcirc_{i=1}^{k} \tilde{f}^i(X)) = f(X) \bigcirc \bigbigcirc_{i=1}^{k} f(\tilde{f}^i(X)) \subseteq Y. \qquad \square$$

Theorem 3.6: *Let $A \in M\mathbb{R}$ and $b \in V\mathbb{R}$ be given. Let $F: \mathbb{I}V\mathbb{R} \to \mathbb{I}V\mathbb{R}$ be given satisfying*

$$X \in \mathbb{I}V\mathbb{R} \Rightarrow Rb + (I - RA) \cdot X \subseteq F(X). \tag{3.10}$$

If then for some $\emptyset \neq X \in \mathbb{I}V\mathbb{R}$

$$F^{k+1}(X) \subsetneqq \bigbigcirc_{i=0}^{k} F^i(X) \quad \text{for some } 0 \leqq k \in \mathbb{N}, \tag{3.11}$$

then the matrices A and R are not singular and the unique solution $\hat{x} = A^{-1} b$ of $Ax = b$ satisfies

$$\hat{x} \in \bigbigcirc_{i=0}^{k} F^i(X).$$

Proof: Follows like Theorem 3.3 by Theorem 3.1 and Theorem 2.5. $\square$

Condition (3.11) is very weak because it is sufficient that $F^{k+1}(X)$ is enclosed in the union of all preceding iterates.

The essential step towards least significant bit accuracy is the following: We represent the intervals X by

$$X = \tilde{x} + Y, \quad Y \in \mathbb{I}V\mathbb{R} \tag{3.12}$$

where $\tilde{x}$ is an approximate solution of $Ax = b$.

Theorem 3.7: *Let $A \in MS$ and $b \in VS$ be given. If for some $R \in MS$ and for some convex, closed and bounded $\emptyset \neq Y \in \mathbb{I}VS$ and some $\tilde{x} \in VS$*

$$R \diamondsuit \{\diamondsuit (b - A\tilde{x})\} \oplus \{\diamondsuit (I - R \cdot A)\} \diamondsuit Y \subsetneqq Y, \tag{3.13}$$

then the matrices A and R are not singular and the unique solution $\hat{x} = A^{-1} b$ of $Ax = b$ satisfies

$$\hat{x} \in \tilde{x} \oplus Y. \tag{3.14}$$

Proof: Replace X by $\tilde{x} + Y$ in the development of Theorem 2.6. $\square$

In practice $\tilde{x}$ will be an approximate solution of $Ax = b$. The notation $\diamondsuit (b - A\tilde{x})$ refers to an analogous method as in (2.13), i.e. there is only one final rounding in each component of the residual. The final inclusion gains in accuracy because an error in Y plays a less important role with respect to $\tilde{x} \oplus Y$. This technique of constructing an inclusion interval Y for the *error* $\hat{x} - \tilde{x}$ of $\tilde{x}$ (cf. (3.12) and (3.14)) rather than an inclusion interval X for $\hat{x}$ itself was introduced in [Ru 80] and can be applied in all similar situations (cf. [Ru83], see also the survey paper at the beginning of this volume). This method leads to a significant shrinking of the result intervals. There is no assumption on $\tilde{x}$ or $\mathbb{R}$ which has to be verified a priori; the only assumption which has to be verified is (3.13) or a similar one.

For a practical implementation a number of further improvements are given in [Ru80] and [Ru83]. One particularly important of these is the ε-inflation, that is before verifying (3.13) the interval Y is inflated. This improves the algorithms significantly and, in some cases, makes an inclusion possible. For details the reader is referred to the literature.

In practice it happens, that $\rho(I - RA) < 1$ but common norm estimates cannot demonstrate the convergence of $I - RA$ whereas iteration (3.1) stops after 2 or 3 steps. Any inclusion algorithm, of course, depends on a suitable inverse R of A. In contrast to other methods here the convergence of $I - RA$ and the correctness of the computed bounds is demonstrated automatically and mathematically verified by the algorithm without any effort on the part of the user.

All the preceding considerations remain valid when applied to complex data.

In practice it is often the case that the input data is not exactly convertible in the given floating-point screen S. In this case an interval matrix A and an interval vector b can be given enclosing the original data (with tolerances). Let a set of matrices $\mathscr{A} \in \mathbb{P}M\mathbb{R}$ and a set of vectors $\mathscr{b} \in \mathbb{P}V\mathbb{R}$ be given. We define the "inclusion of the solutions of $\mathscr{A}x = \mathscr{b}$" by

$$\{x \in V\mathbb{R} \mid \exists A \in \mathscr{A}, b \in \mathscr{b}: Ax = b\}. \tag{3.15}$$

It is possible to compute an inclusion of this set (3.15) according to the following theorem:

Theorem 3.8: *Let $\mathscr{A} \in \mathbb{P}M\mathbb{R}$ and $\mathscr{b} \in \mathbb{P}V\mathbb{R}$ be given. If for some $R \in M\mathbb{R}$ and for some convex, closed and bounded $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$*

$$R \cdot \mathscr{b} + (I - R\mathscr{A}) \cdot X \subseteq \hat{X}, \tag{3.16}$$

then the matrix R and every matrix $A \in \mathscr{A}$ are not singular and for every $A \in \mathscr{A}$, $b \in \mathscr{b}$ the unique solution $\hat{x} = A^{-1} b$ of $Ax = b$ satisfies

$$\hat{x} \in \hat{X}.$$

The *proof* follows from Theorem 2.9. □

Again, we can formulate theorems and lemmata corresponding to the preceding ones for $\mathscr{A} \in \mathbb{I}M\mathbb{R}$, $\ell \in \mathbb{I}V\mathbb{R}$ and for $\mathscr{A} \in \mathbb{I}MS$, $\ell \in \mathbb{I}VS$ with the corresponding operations $\circledast$ and $\circledcirc$, respectively for $* \in \{+, -, \cdot, /\}$. The inclusion of the error of an approximate solution and the ε-inflation can be applied like in the point case. The extension to complex linear systems is similar.

We want to emphasize again, that there is no effort necessary on the part of the user like estimation of spectral radius, verification of the non-singularity of the input matrix etc.

With the preceding considerations algorithms can be formulated for computing a highly accurate inclusion of the solution of a point or interval linear system. Such algorithms with further improvements are given in [Ru80] and [Ru83].

The performance of such algorithms is first demonstrated by the inversion of a Hilbert matrix. We first multiply the components of the Hilbert matrix by the least common multiple of the denominators to obtain integer entries, i.e. we define

$$H^{n^*} \in M_n\mathbb{R} \text{ by } H^{n^*}_{ij} := l\mathrm{cm}(1, 2, \ldots, 2n-1)/(i+j-1).$$

The following results are computed on a IBM S/370 with 14 hexadecimal digits in the mantissa. H^{21^*} is the H^* matrix of largest dimension exactly storable on that computer.

For the right hand side $b = (1, 0, \ldots, 0)^T$ we first computed an approximation $\tilde{x}$ to the solution $\hat{x}$ of $H^{21^*} \cdot x = b$ by Gaussian elimination with extended precision residual correction and obtained, e.g.

$$\tilde{x}_{16} = 0.1245274389638609 \cdot 10^{-7}.$$

With the linear equation solver contained in the subroutine package ACRITH we obtained

$$-0.1086151817859135 \cdot 10^{-1} < \hat{x}_{16} < -0.1086151817859134 \cdot 10^{-1},$$

a result with least significant bit accuracy, with automatic verification of the existence and uniqueness of the solution and with automatic verication of the correctness of the bounds. In fact, we achieved least significant bit accuracy for all 21 components of the inclusion. The approximation obtained in a traditional way is wrong in sign and magnitude.

An example for an interval linear system is the following. Let

$$A := \begin{pmatrix} 100\,000 & 99\,999 \\ 99\,999 & 99\,998 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 200\,000 \\ 200\,000 \end{pmatrix}$$

and consider the linear system $Ax = b$ with the additional information, that the data of b is afflicted with a tolerance of ± 10 in each component. Normally, one try different values for b within the given error margin. Consider

Table 3.1. *Traditional method for linear system with tolerances in data*

b_1	b_2	$\hat{x}_1$	$\hat{x}_2$
200 000	200 000	200 000	−200 000
199 990	199 990	199 990	−199 990
200 010	200 010	200 010	−200 010

For every change of b the (exact) solution $\hat{x}$ differs by the same magnitude in each component and suggests a good condition and the following inclusion of $\hat{x}$:

$$\hat{x} \in \begin{pmatrix} [199\,990, & 200\,010] \\ [-200\,010, & -199\,990] \end{pmatrix}. \tag{3.17}$$

The truth is told by our new method. Solving $Ax = b$ with the interval right hand side

$$\ell = \begin{pmatrix} [199\,990, & 200\,010] \\ [199\,990, & 200\,010] \end{pmatrix}$$

we obtain by the linear equation solver contained in the subroutine package ACRITH

$$\hat{x} \in \begin{pmatrix} [-1799974.5, & 2199974.5] \\ [-2199995.4, & 1799995.4] \end{pmatrix}. \tag{3.18}$$

The exact bounds for the set of solutions of $Ax = b$ for $b \in \ell$ are

$$\hat{x} \in \begin{pmatrix} [-1799970.0, & +2199970.0] \\ [-2199990.0, & +1799990.0] \end{pmatrix}. \tag{3.19}$$

The inclusion (3.18) obtained by ACRITH is only slightly wider than the exact inclusion (3.19). However, the "empirical guess" (3.17) is too small in diameter by a factor of 200 000.

4. Nonlinear Systems

Let a differentiable function $f: V\mathbb{R} \to V\mathbb{R}$ be given. Consider the Newton iteration

$$g(x) := x - R \cdot f(x) \tag{4.1}$$

where $R \in M\mathbb{R}$ is an approximate inverse of the Jacobian $f'(x)$. For a fixed point x of g

$$g(x) = x \Rightarrow f(x) \in \ker R. \tag{4.2}$$

As in the case of linear systems we try to find conditions for the non-singularity of R to verify $f(x) = 0$.

Definition 4.1: Let a differentiable function $f: V\mathbb{R} \to V\mathbb{R}$ with Jacobian f' be given. Then

$$X \in \mathbb{P}V\mathbb{R}: \quad f'(X) := \left\{ \left(\frac{\partial f}{\partial x_1}(\zeta_1), \ldots, \frac{\partial f}{\partial x_n}(\zeta_n) \right)^T \,\middle|\, \zeta_i \in X_i \text{ for } 1 \leqq i \leqq n \right\}.$$

With this definition we see from the Mean Value Theorem that

$$\forall x, \tilde{x} \in V\mathbb{R}\ \exists Q \in f'(\tilde{x} \cup x): \quad f(x) = f(\tilde{x}) + Q \cdot (x - \tilde{x}). \tag{4.3}$$

However, there need not be a $y = \tilde{x} + \lambda(x - \tilde{x})$ where $0 \leqq \lambda \leqq 1$ with

$$f(x) = f(\tilde{x}) + f'(y) \cdot (x - \tilde{x}).$$

Like in the linear case we could assume that an estimate

$$\| I - R \cdot f'(X) \| < 1 \quad \text{for some norm} \quad \| \cdot \| : M\mathbb{R} \to \mathbb{R}$$

is satisfied for $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$ to deduce the non-singularity of R. As in the linear case we can omit the norm estimate:

Theorem 4.2: *Let a differentiable function $f: V\mathbb{R} \to V\mathbb{R}$ be given. If for some $R \in M\mathbb{R}$, some convex, closed and bounded $\emptyset \neq X \in \mathbb{P}V\mathbb{R}$ and some $\tilde{x} \in V\mathbb{R}$*

$$\tilde{x} - R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup X\} \subseteq \hat{X}, \tag{4.4}$$

then the matrix R and every matrix $Q \in f'(\tilde{x} \cup X)$ are not singular and the equation $f(x) = 0$ has one and only one solution $\hat{x}$ satisfying

$$\hat{x} \in X.$$

Proof: For every $x \in X$ the function g defined by (4.1) satisfies

$$g(x) = x - R \cdot f(x) \in x - R \cdot \{f(\tilde{x}) + f'(\tilde{x} \cup x)(x - \tilde{x})\} =$$
$$= \tilde{x} - R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup x)\} \cdot (x - \tilde{x}).$$

Therefore by (4.4)

$$g(X) \subseteq \hat{X}$$

and by the fixed point theorem of Brouwer there is an $\hat{x} \in X$ with $g(\hat{x}) = \hat{x}$. By Lemma 2.7 and (4.4) we get the non-singularity of R and of every matrix $Q \in f'(x \cup x)$ and by (4.2) we get the existence of a zero $\hat{x}$ of f in X. Suppose $\hat{y} \in X$ with $f(\hat{y}) = 0$. Then $g(\hat{y}) = \hat{y}$ and by (4.3) there is a

$$Q \in f'(\hat{x} \cup \hat{y}) \subseteq f'(\tilde{x} \cup X)$$

with

$$f(\hat{y}) = f(\hat{x}) + Q(\hat{y} - \hat{x}) \quad \text{implying} \quad Q \cdot (\hat{y} - \hat{x}) = 0.$$

Because Q is not singular this implies $\hat{y} = \hat{x}$ and the theorem is proved. □

As in the linear case it is preferable to compute an inclusion of the error of an approximate solution to gain in accuracy of the inclusion:

Theorem 4.3: *Let a differentialbe function $f: V\mathbb{R} \to V\mathbb{R}$ be given. If for some $R \in M\mathbb{R}$, some convex, closed and bounded $\emptyset \neq Y \in \mathbb{P}V\mathbb{R}$ and some $\tilde{x} \in V\mathbb{R}$*

$$-R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup (\tilde{x} + Y))\} \cdot Y \subseteq \hat{Y} \tag{4.5}$$

then the matrix R and every matrix $Q \in f'(\tilde{x} \cup (\tilde{x} + Y))$ are not singular and the equation $f(x) = 0$ has one and only one solution $\hat{x}$ satisfying

$$\hat{x} \in \tilde{x} + Y.$$

The *proof* follows by replacing X by $\tilde{x} + Y$ in Theorem 4.2. □

The preceding theorem can be extended to the complex number space (f' is defined similar to Definition 4.1).

Theorem 4.4: *Let a holomorphic function $f: V\mathbb{C} \to V\mathbb{C}$ be given. If for some $R \in M\mathbb{C}$ some convex, closed and bounded $\emptyset \neq Y \in \mathbb{P}V\mathbb{C}$ and some $\tilde{x} \in V\mathbb{C}$*

$$-R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup (\tilde{x} + Y))\} \cdot Y \subseteq \hat{Y},$$

then the matrix R and every matrix $Q \in f'(\tilde{x} \cup (\tilde{x} + Y))$ are not singular and the equation $f(x) = 0$ has one and only one solution $\hat{x}$ satisfying

$$\hat{x} \in \tilde{x} + Y.$$

The *proof* is similar to the one of Theorem 4.2 and uses

Lemma 4.5: *Let $\tilde{z} \in V\mathbb{C}$, $Z \in \mathbb{P}V\mathbb{C}$ and some function $f: \tilde{z} \mathbin{\textcircled{\cup}} Z \to V\mathbb{C}$ be holomorphic. Then*

$$f(z) \in f(\tilde{z}) + f'(\tilde{z} \cup Z) \cdot (Z - \tilde{z}).$$

The *proof* was given by Böhm (cf. [Bö 80]). □

The preceding theorems are applicable on computers by using the operations $\circledast$ instead of $* \in \{+, -, \cdot, /\}$. We give the corresponding version of Theorem 4.3.

Theorem 4.6: *Let a differentiable function $f: V\mathbb{R} \to V\mathbb{R}$ and some function $F: VS \to VS$ and $F': VS \to MS$ with*

$$X \in \mathbb{I}VS \text{ and } x \in X \Rightarrow f(x) \in F(X) \text{ and } f'(X) \subseteq F'(X)$$

be given. If for some $R \in MS$, $Y \in \mathbb{I}VS$ and $\tilde{x} \in VS$

$$-R \diamondsuit F(\tilde{x}) \circledast \{\diamondsuit(I - R \cdot F'(\tilde{x} \mathbin{\textcircled{\cup}} (\tilde{x} \circledast Y)))\} \diamondsuit Y \subseteq \hat{Y}, \tag{4.6}$$

then the matrices R and every matrix $Q \in F'(\tilde{x} \mathbin{\textcircled{\cup}} (\tilde{x} \circledast Y))$ are not singular and the equation $f(x) = 0$ has one and only one solution $\hat{x}$ satisfying

$$\hat{x} \in \tilde{x} \circledast Y.$$

The *proof* follows from (4.6) and Theorem 4.3. □

Theorem 4.4 is valid as well for complex machine numbers from $\mathbb{C}S$ by using the associated interval operations $\circledast$ over $\mathbb{C}S$ for $* \in \{+, -, \cdot, /\}$. A corresponding algorithm for the computation of an inclusion of a solution of a system of nonlinear equations has been given in [Ru 82] and [Ru 83]. As in the case of linear systems an ε-inflation is used in each step of the interval iteration. We do not see an extension of Theorems 3.1 and 3.6 to the nonlinear case.

As has been shown in [Ru 82] the assumption of Theorem 4.3

$$F(Y) \subseteq \hat{Y} \text{ for } F(Y) := -R \cdot f(\tilde{x}) + \{I - R \cdot f'(\tilde{x} \cup (\tilde{x} + Y))\} \cdot Y$$

can be replaced by

$$Z := F(Y);\ F(Z_1, \ldots, Z_k, Y_{k+1}, \ldots, Y_n) \subseteq \hat{Y} \quad \text{for } 0 \leqq k \leqq n-1.$$

This "Einzelschrittverfahren" weakens assumption (4.5). This method is applicable for the complex case and the case of systems of linear equations as well.

Finally we give some examples to demonstrate the performance of the new methods. Consider the examples from [AbBr75]:

Table 4.1. *Examples of nonlinear equations*

1. Boggs:
$$f_1 = x_1^2 - x_2 + 1$$
$$f_2 = x_1 - \cos\left(\frac{\pi}{2} \cdot x_2\right)$$
with $\tilde{x} = (1, 0)$. Solutions are $\hat{x} = (0, 1)$, $\hat{x} = (-1, 2)$ and $\hat{x} = \left(-\frac{\sqrt{2}}{2}, 1.5\right)$.
2. Example 1. with $\tilde{x} = (-1, -1)$.
3. Broyden:
$$f_1 = \frac{1}{2}\sin(x_1 x_2) - x_2/(4\pi) - x_1/2$$
$$f_2 = (1 - 1/(4\pi)) \cdot (e^{2x_1} - e) + e x_2/\pi - 2e x_1 .$$
with $\tilde{x} = (0.6, 3)$ and $\hat{x} = (0.5, \pi)$.
4. Rosenbrock:
$$f_1 = 400 x_1 (x_1^2 - x_2) + 2(x_1 - 1)$$
$$f_2 = 200 x_1 (x_1^2 - x_2)$$
with $\tilde{x} = (-1.2, 1)$ and $\hat{x} = (1, 1)$.
5. Braun:
$$f_1 = 2\sin(2\pi x_1/5) \cdot \sin(2\pi x_3/5) - x_2$$
$$f_2 = 2.5 - x_3 + 0.1 \cdot x_2 \cdot \sin(2\pi x_3) - x_1$$
$$f_3 = 1 + 0.1 \cdot x_2 \cdot \sin(2\pi x_3) - x_3$$
with $\tilde{x} = (0, 0, 0)$ and $\hat{x} = (1.5, 1.809\ldots, 1.0)$.
6. Deist and Sefor:
$$f_i = \sum_{\substack{j=1 \\ j \neq i}}^{6} \cot(\beta_i x_i) \quad \text{for } 1 \leqq i \leqq 6$$
with $\tilde{x} = (75, 75, 75, 75, 75, 75)$, $100\,\beta_i = 2.249, 2.166, 2.083, 2.0, 1.918, 1.835$ for $1 \leqq i \leqq 6$ and $\hat{x} \approx (121.9, 114.2, 93.6, 62.3, 41.3, 30.5)$.

To compute an inclusion of the solution we first apply a Newton iteration and then our new methods according to Theorem 4.6. After 2 to 6 Newton iterations we obtained on an IBM S/370 machine in long format (14 hexadecimal figures in the mantissa which is approximately 16.5 decimal figures) the following result.

We use the short notation of displaying the error margin in the last displayed place of the mantissa. For example, the very last inclusion reads

$$\hat{x}_6 \in [30.5\,026\,656\,940\,32,\ 30.5\,026\,656\,940\,34].$$

As we see the accuracy of the inclusion is at least 14 decimal figures in each component. The slight loss of one to two figures in the inclusion of the sixth examples compared with the previous ones is partly due to the fact, that the cot function has not been included directly but where expressed by cos/sin.

Table 4.2. *Inclusion of the solution of nonlinear systems*

1. $\hat{x}_1 \in$ −1.00 000 000 000 000 0 ±1
   $\hat{x}_2 \in$ 2.00 000 000 000 000 0 ±1
2. Newton iteration not convergent
3. $\hat{x}_1 \in$ 0.500 000 000 000 000 0 ±5
   $\hat{x}_2 \in$ 3.14 159 265 358 979 3 ±2
4. $\hat{x}_1 \in$ 1.00 000 000 000 000 0 ±1
   $\hat{x}_2 \in$ 1.00 000 000 000 000 0 ±1
5. $\hat{x}_1 \in$ 1.50 000 000 000 000 0 ±1
   $\hat{x}_2 \in$ 1.80 901 699 437 494 8 ±2
   $\hat{x}_3 \in$ 1.00 000 000 000 000 0 ±1
6. $\hat{x}_1 \in$ 121. 850 455 344 731 0 ±10
   $\hat{x}_2 \in$ 114. 160 899 365 558 0 ±10
   $\hat{x}_3 \in$ 93.6 487 503 169 383 0 ±50
   $\hat{x}_4 \in$ 62.3 185 704 328 125 0 ±150
   $\hat{x}_5 \in$ 41.3 219 490 821 365 0 ±150
   $\hat{x}_6 \in$ 30.5 026 656 940 330 0 ±100

The following examples show the performance of our new methods for a larger number of unknowns:

Table 4.3. *Examples of nonlinear equations*

7. [AbBr75] Discretization of $3y'' y + y'^2 = 0$, $y(0) = 0$, $y(1) = 20$.
$$f_1 = 3x_1(x_2 - 2x_1) + x_2^2/4$$
$$f_i = 3x_i(x_{i+1} - 2x_i + x_{i-1}) + (x_{i+1} - x_{i-1})^2/4 \qquad 2 \leqq i \leqq n-1$$
$$f_n = 3x_n(20 - 2x_n + x_{n-1}) + (20 - x_{n-1})^2/4$$
with $\tilde{x}_i = 10$ for $1 \leqq i \leqq n$ and $y(t) = 20 \cdot t^{3/4}$.
8. [MoCo79] Discretization of $u''(t) = 0.5 \cdot (u(t) + t + 1)^3$, $u(0) = u(1) = 0$
$$f_i = 2x_i - x_{i-1} + 0.5 \cdot h^2 \cdot (x_i + t_i + 1)^3 \text{ for } 1 \leqq i \leqq n \ (x_0 = x_{n+1} = 0)$$
with $h = 1/(n+1)$, $t_i = i \cdot h$ and $\tilde{x}_i = t_i(t_i - 1)$, $1 \leqq i \leqq n$.

As before we first apply a Newton iteration and then our new methods. In the following Table 4.4 we display from left to right

Problem	the number of the problem
n	the number of unknowns
Newton	the number of Newton iterations starting with $\tilde{x}$
digits	the minimum number of digits guaranteed for each component

we use again IBM S/370 and long format (14 hexadecimal figures accuracy).

Table 4.4. *Inclusion of the solution of nonlinear systems*

Problem	n	Newton	digits
7.	20	7	$16\frac{1}{2}$ (l.s.b.a.)
	50	8	16
	100	7	16
8.	10	5	$16\frac{1}{2}$ (l.s.b.a.)
	20	5	$16\frac{1}{2}$ (l.s.b.a.)
	50	6	16

The accuracy of each of the up to 100 components of the inclusion of the solution is at least 16 decimal figures with $16\frac{1}{2}$ figures precision of the calculation. An additional l.s.b.a. indicates that every component of the inclusion where of least significant bit accuracy.

5. Conclusion

In the preceding chapters we gave the theoretical fundamentals for the development of algorithms to compute inclusions of the solution of systems of linear and nonlinear equations. In [Ru83] the corresponding algorithms have been described. There, moreover, several other problems are treated like linear systems with band, symmetric or sparse matrices, over- and underdetermined linear systems, zeros of polynomials, algebraic eigenvalue problems, linear, quadratic and convex programming problems, the evaluation of arithmetic expressions and others. Algorithms corresponding to a number of these problems have been implemented in the subroutine library of the IBM Program Product ACRITH, which is available on the market since March 1984.

The new methods first perform an automatic verification that the given problem has a solution. Then sharp bounds for a solution are computed with an automatic verification of the correctness of the bounds by the algorithm.

The implemented algorithms based on our new methods have some key properties in common:

every result is automatically verified to be correct by the algorithm;

the computed bounds are of high accuracy, i.e. the error of every component of the result is of the magnitude of the relative rounding error unit;

the solution of the given problem is shown to exist and to be unique within the computed error bounds;

the input data may be afflicted with tolerances;

the computing time is of the same order as a comparable (purely) floating-print algorithm (the latter, of course, satisfies none of the new properties).

Our experience is, that the new algorithms very often achieve bounds with the l.s.b.a. property for every component of the inclusion of the solution.

References

[AbBr75] Abbot, J. P., Brent, R. P. Fast local convergence with single and multistep methods for nonlinear equations. Austr. Math. Soc. *B 19*, 173–199 (1975).

[AlHe74] Alefeld, G., Herzberger, J.: Einführung in die Intervallrechnung. Mannheim-Wien-Zürich: Bibl. Inst. 1974.

[Al79] Alefeld, G.: Intervallanalytische Methoden bei nicht-linearen Gleichungen. In: Jahrbuch Überblicke Mathematik 1979. Zürich: B. I. Verlag 1979.

[Bo71] Boggs, P. T.: The solution of nonlinear systems of equations by A-stable integration techniques. SIAM J. Numer. Anal. *8*, 767–785 (1971).

[Bo77] Bohlender, G.: Floating-point computation of functions with maximum accuracy. IEEE Trans. Comput. *C-26*, 621–632 (1977).

[Bö80] Böhm, H.: Berechnung von Schranken für Polynomwurzeln mit dem Fixpunktsatz von Brouwer. Interner Bericht des Inst. f. Angew. Math., Universität Karlsruhe, 1980.

[Bö83] Böhm, H.: Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrücke mit garantierter, maximaler Genauigkeit, Dr.-Dissertation, Inst. f. Angew. Math., Universität Karlsruhe, 1983.

[Bra72] Branin, F. H.: Widely convergent method for finding multiple solutions of simultaneous nonlinear equations. IBM J. Res. Develop. *16*, 504–522 (1972).

[Bro69] Broyden, C. G.: A new method of solving nonlinear simultaneous equations. Comput. J. *12*, 94–99 (1969).

[DeSe67] Deist, F. H., Sefor, L.: Solution of systems of nonlinear equations by parameter variation. Comput. J. *10*, 78–82 (1967).

[Fo70] Forsythe, G. E.: Pitfalls in computation, or why a Math book isn't enough. Technical Report No. CS147, Computer Science Department, Stanford University, 1–43 1970.

[Kn69] Knuth, D.: The Art of Computer Programming, Vol. 2. Reading, Mass.: Addison-Wesley 1969.

[Kr69] Krawczyk, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing *4*, 187–201 (1969).

[Kö80] 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).

[KuMi81] Kulisch, U., Miranker, W. L.: Computer Arithmetic in Theory and Practice. New York: Academic Press 1981.

[Ku69] Kulisch, U.: Grundzüge der Intervallrechnung (Überblicke Mathematik 2) (Laugwitz, D., Hrsg.), pp. 51–98. Mannheim: Bibliographisches Institut 1969.

[Ku71] Kulisch, U.: An axiomatic approach to rounded computations. Mathematics Research Center. University of Wisconsin, Madison, Wisconsin, TS Report No. 1020, 1–29 (1969); Numer. Math. *19*, 1–17 (1971).

[Ku76] Kulisch, U.: Grundlagen des numerischen Rechnens (Reihe Informatik, 19). Mannheim-Wien-Zürich: Bibliographisches Institut 1976.

[Ma80] Martinez, J. M.: Solving non-linear simultaneous equations with a generalization of Brent's method. BIT *20*, 501–510 (1980).

[Mo66] Moore, R. E.: Interval Analysis. Englewood Cliffs, N. J.: Prentice-Hall 1966.

[Mo77] Moore, R. E.: A test for existence of solution for non-linear systems. SIAM J. Numer. Anal. *4*, 611–615 (1977).

[MoCo79] Moré, J. J., Cosnard, M. Y.: Numerical solution of non-linear equations. ACM Trans. on Math. Software *5*, 64–85 (1979).

[OrRb70] Ortega, J. M. Reinboldt, W. C.: Iterative Solution of Non-linear Equations in Several Variables. New York-San Francisco-London: Academic Press 1970.

[Ru80] Rump, S. M.: Kleine Fehlerschranken bei Matrixproblemen, Dr.-Dissertation, Inst. f. Angew. Math., Universität Karlsruhe, 1980.

[Ru82] Rump, S. M.: Solving non-linear systems with least significant bit accuracy. Computing *29*, 183–200 (1982).

[Ru83] Solving algebraic problems with high accuracy. Habilitationsschrift, Universität Karlsruhe; appeared in: A New Approach to Scientific Computation (Kulisch, U. W., Miranker, W. L., eds.). New York: Academic Press 1983.

[RuBö83] Rump, S. M., Böhm, H.: Least significant bit evaluation of arithmetic expressions in single-precision. Computing *30*, 189–199 (1983).

[RuKa80] Rump, S. M., Kaucher, E.: Small bounds for the solution of systems of linear equations. In: Computing, Suppl. 2. Wien-New York: Springer 1980.
[SRS72] Schwarz, H. R., Rutishauer, H., Stiefel, E.: Matrizen-Numerik. Stuttgart: B. G. Teubner, 1972.
[St72] Stoer, J.: Einführung in die Numerische Mathematik I. (Heidelberger Taschenbücher, Band 105.) Berlin-Heidelberg-New York: Springer 1972.
[StBu73] Stoer, J., Bulirsch, R.: Einführung in die Numerische Mathematik II. (Heidelberger Taschenbücher, Band 114.) Berlin-Heidelberg-New York: Springer 1973.
[Va62] Varga, R. S.: Matrix Iterative Analysis. Englewood Cliffs, N. J.: Prentice-Hall 1962.
[Wi69] Wilkinson, J. H.: Rundungsfehler. Berlin-Heidelberg-New York: Springer 1969.

Priv.-Doz. Dr. S. M. Rump
IBM Entwicklung und Forschung
Schönaicher Strasse 220
D-7030 Böblingen
Federal Republic of Germany

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