Computing 28, 31—42 (1982)

Computing

E-Methods for Fixed Point Equations f(x) = x

E. Kaucher and **S. M. Rump,** Karlsruhe

Received July 1, 1981; revised September 10, 1981

Abstract — Zusammenfassung

E-Methods for Fixed Point Equations f(x) = x. This paper provides newly implemented [11], [13] and widely applicable methods for computing inclusion (i. e. a containing interval) (Einschließung) of the solution of a fixed point equation $f(x) = x$ as well as automatic verification the existence (Existenz) and uniqueness (Eindeutigkeit) of the solution. These methods make essential use of a new computer arithmetic defined by semimorphisms as developed in [7] and [8]. We call such methods E-Methods in correspondance to the three German words. A priori estimations such as a bound for a Lipschitz constant etc. are not required by the new algorithm. So the algorithm including the a posteriori proof of existence and uniqueness of the fixed point is programmable on computers for linear as well as for non-linear problems. This is a key feature of our results. The computations produced by E-methods deliver answers the components of which have accuracy better than 10^{-t+1} (where t denotes the mantissa length employed in the computer).

Key words: E-method, inclusion, automatic verification.

AMS Subject Classification: 65 H 99.

E-Methoden für Fixpunktgleichungen f(x) = x. Es werden neuartige sehr allgemeine Methoden vorgestellt, die sowohl eine Einschließung der Lösung von Fixpunktgleichungen $f(x) = x$ als auch automatisch die Existenz und gegebenenfalls Eindeutigkeit der Lösung nachweisen. Diese Methoden machen wesentlichen Gebrauch von neuen Rechnerarithmetiken, die charakterisiert sind wie in [2], [7] und [8] entwickelt. Wir nennen solche Methoden E-Methoden in Übereinstimmung mit den drei Anfangsbuchstaben. A-priori-Abschätzungen wie z. B. für Schranken von Lipschitzkonstanten sind nicht mehr notwendig. Daher ist es in eleganter Weise möglich, Algorithmen zu implementieren, die einen automatischen Existenz- und Eindeutigkeitsnachweis für den Fixpunkt von linearen und nichtlinearen Fixpunktgleichungen ermöglichen. Die mit E-Methoden berechneten Lösungen haben i. a. eine relative Genauigkeit, die besser als 10^{-t+1} ist (wobei t die Mantissenlänge des verwendeten Rechners bezeichnet).

1. Introduction

In [5], [10] and [11] methods are introduced, which provide an inclusion (i.e. a containing interval) of the fixed point of an equation. The methods derived in [5] are typically generalizations of those introduced by Moore in [9]. The results presented here both generalize and simplify the methods given in [5], [9] and [10].

The following iteration operator introduced in [6]

$$K(X) := \tilde{x} - R * g(\tilde{x}) + \{E - R * g'(X)\} * (X - \tilde{x}) \tag{1}$$

0010-485X/82/0028/0031/$ 02.40

is used in [9]. Here $\tilde{x} \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n \in \mathscr{C}_X^1$ and $X \in \mathbb{R}^n$ denotes an n-dimensional interval vector and E the $n \times n$ identity matrix. In [9] R is required to be a real non-singular matrix and $\tilde{x} \in X$. Under these conditions the existence of a solution of $g(\hat{x}) = 0$ in X is derived from the property $K(X) \subseteq X$. We will show that it is not necessary to assume R to be non-singular and that $\tilde{x}$ can be chosen arbitrarily (not necessarily $\tilde{x} \in X$). The somewhat more stingent condition $K(X) \overset{\circ}{\subset} X$ (which is almost always satisfied on the computer) is sufficient to show that R is non-singular and that $g(x) = 0$ has exactly one solution $\hat{x} \in X$.

2. Theoretical Preliminaries

Definition 1: Let M_1 and M_2 be closed subsets of the locally convex topological space $\mathscr{M}$. We define the strict inclusion relation as follows

$$M_1 \overset{\circ}{\subset} M_2 :\Leftrightarrow M_1 \subseteq \mathring{M}_2 ; \tag{2}$$

i.e., M_1 lies in the interior $\mathring{M}_2$ of M_2.

An improved form of some fundamental results of [5] and [10] is given in the following theorem.

Theorem 2: *Let $f : Y \to \mathscr{M}$ be a continous mapping and $F : \mathscr{P}\mathscr{M} \to \mathscr{P}\mathscr{M}$ an arbitrary mapping of the power set $\mathscr{P}\mathscr{M}$ into itself such that*

$$x \in Y \Rightarrow f(x) \in F(Y). \tag{3}$$

Let Y be convex and compact. If

$$F(Y) \overset{\circ}{\subset} Y, \tag{4}$$

then there exists a fixed point $\hat{x}$ of f with

$$\hat{x} \in F(Y) \subseteq Y. \tag{5}$$

Moreover

$$\hat{x} \in \bigcap_{i=0}^{\infty} F^i(Y), \tag{6}$$

and $Q(f, Y) \overset{\circ}{\subset} Y$ for the set of fixed points

$$Q(f, Y) := \{x \in Y \mid f(x) = x\} \tag{7}$$

of f in Y. Therefore $Q(f, Y) \cap \partial Y = \emptyset$.

Proof: From (3) and (4) it follows immediately that f maps the convex and compact subset Y of the locally convex space $\mathscr{M}$ into itself. According to the fixed point theorem of Schauder-Tychonoff f has at least one fixed point $\hat{x} \in Y$. With $\hat{x} \in Y = F^0(Y)$ we have by induction

$$\hat{x} \in F^k(Y) \Rightarrow \hat{x} = f(\hat{x}) \in F\left(F^k(Y)\right) = F^{k+1}(Y).$$

The proof of (7) derives from (3) and (4) since for all $x \in Q(f, Y) \overset{\circ}{\subset} Y$ we have

$$x \in Y \Rightarrow x = f(x) \in F(Y) \overset{\circ}{\subset} Y. \qquad \square$$

Theorem 3: *Let f be an affine operator on the topoligical vector-space $\mathcal{M}$. Then under the hypotheses of Theorem 2, f has exactly one fixed point $\hat{x} \in X$.*

Proof: According to Theorem 2 there exist a fixed point $\hat{x} \in Y$. So the linear operator $g(t) := f(t+\hat{x}) - (t+\hat{x})$ has the fixed point $g(0)=0$. Suppose 0 is not the only element of the kernel of g, i.e. $\ker(g) \neq \{0\}$. Then, the $\ker(g)$ is a linear subspace of dimension greater than 0. But $\hat{x} \in Y$ implies $0 \in Y - \hat{x}$, and so $\ker(g) \cap (Y-\hat{x}) \neq \emptyset$. On the other hand for every $s \in \ker(g) \cap (Y-\hat{x})$ the equation

$$0 = g(s) = f(s+\hat{x}) - (s-\hat{x})$$

holds. Thus an $\hat{s} \in \ker(g) \cap (Y-\hat{x})$ would exist with $y = \hat{s} + x \in \partial Y$ being a fixed point of f and an element of $Q(f, Y)$. This contradicts (7) in Theorem 2.

□

Remark 4: Let g, g_1, g_2 be linear operators on the finite-dimensional vectorspace $\mathcal{M}$ and let $g = g_1 \circ g_2$. Then

$$\ker(g) = \{0\} \Leftrightarrow \ker(g_1) = \{0\} \wedge \ker(g_2) = \{0\}.$$

Theorem 5: *Let $\mathcal{M}$ be finite dimensional normed complex space. For arbitrary but fixed $Y \in \mathcal{P}\mathcal{M}$, $y \in \mathcal{M}$, let the operator $F : \mathcal{P}\mathcal{M} \to \mathcal{P}\mathcal{M}$ of Theorem 2 have the following decomposition:*

$$F(Y) = f(\tilde{y}) + \mathcal{L}_{(f,\tilde{y},Y)}(Y-\tilde{y}).$$

Here the set $\mathcal{L}$ is an element of the powerset of the set of linear operators over $\mathcal{M}$. Then under the hypotheses of Theorem 2 (in particular if

$$F(Y) = f(\tilde{y}) + \mathcal{L}_{(f,\tilde{y},Y)}(Y-\tilde{y}) \dot{\subset} Y \tag{8}$$

for a convex and compact $Y \in \mathcal{P}\mathcal{M}$) we have:

The spectral radius of every $l \in \mathcal{L}$ is less than unity:

$$l \in \mathcal{L} \Rightarrow \rho(l) < 1, \text{ abbreviated by } \rho(\mathcal{L}) < 1. \tag{9}$$

If there exists a $\mathcal{K} \subseteq \mathcal{L}_{(f,\tilde{y},Y)}$ with the property

$$\bigwedge_{x,y \in Y} f(x) - f(y) \in \mathcal{K}(x-y),$$

then f has exactly one fixed point in Y.

Proof:

ad (9): Consider the affine operator

$$h(x) := f(\tilde{y}) + l(x-\tilde{y}): \mathcal{M} \to \mathcal{M}$$

for an arbitrary linear operator $l \in \mathcal{L}_{(f,\tilde{y},Y)}$. Then

$$h(x) = f(\tilde{y}) + l(x-\tilde{y}) \in f(\tilde{y}) + \mathcal{L}_{(f,\tilde{y},Y)}(x-\tilde{y}) = F(\{x\}),$$

and due to (8)

$$h(Y) \subseteq F(Y) \dot{\subset} Y. \tag{11}$$

Thus according to Theorem 2 h has a fixed point $\hat{x} = \hat{x}(l) \in Y$. These observations are expressed by the following two equations

$$\begin{aligned} \hat{x} &= h(\hat{x}) = f(\tilde{y}) + l(\hat{x} - \tilde{y}) \\ h(Y) &= f(\tilde{y}) + l(Y - \tilde{y}) \overset{\circ}{\subset} Y. \end{aligned} \tag{12}$$

Substituting $U := Y - \tilde{y}$ and $\hat{y} = \hat{x} - \tilde{y} \in U$ yields

$$\begin{aligned} f(\tilde{y}) - \tilde{y} + l(\hat{y}) &= \hat{y} \\ f(\tilde{y}) - \tilde{y} + l(U) &\overset{\circ}{\subset} U. \end{aligned} \tag{13}$$

Taking the difference of the furmulas in (13) we get

$$l(U - \hat{y}) = l(U) - l(\hat{y}) \overset{\circ}{\subset} U - \hat{y},$$

and with the substitution $V := U - \hat{y}$,

$$0 \in l(V) \overset{\circ}{\subset} V. \tag{14}$$

Thus V contains a neighbourhood of 0. If $l \equiv 0$ then $\rho(l) = 0 < 1$. Let $l \not\equiv 0$ and let $v \in V$ be an arbitrary eigenvector of l which corresponding eigenvalue $\lambda \in \mathbb{C}$. Let

$$\Phi := \{\phi \in \mathbb{C} \mid \phi \cdot v \in V\} \quad \text{and let} \quad |\phi^*| := \max_{\phi \in \Phi} |\phi|$$

with $\phi^* \in \Phi$. Since V is compact (14) yields

$$\lambda \cdot (\phi^* v) = l(\phi^* v) \in V \backslash \partial V.$$

Hence there exists an open neighbourhood of $\lambda \phi^* v$ which lies in $V \backslash \partial V$. Thus a real scalar $\sigma > 1$ exists such that

$$\sigma \cdot (\lambda \phi^* v) \in \partial V.$$

By definition $\sigma \lambda \phi^* \in \Phi$, and by definition of $\phi^* |\sigma| \cdot |\lambda| \cdot |\phi^*| = |\sigma \lambda \phi^*| \leqq |\phi^*|$. Therefore $|\sigma| \cdot |\lambda| \leqq 1$, $|\lambda| < 1$ so that finally $\rho(l) < 1$. Since $l \in \mathscr{L}$ was chosen arbitrarily, we conclude that

$$\rho(\mathscr{L}) := \{\rho(l) \mid l \in \mathscr{L}\} < 1.$$

The uniqueness of the fixed point $\hat{x}(l) \in Y$ (for a fixed l) of h was not used in the preceding proof. For a fixed point y of h we have with $\rho(\mathscr{L}) < 1$:

$$0 = h(\hat{x}) - \hat{x} = f(\tilde{y}) - \hat{x} + l(\hat{x} - \tilde{y})$$

and

$$0 = h(y) - y = f(\tilde{y}) - y + l(y - \tilde{y}).$$

Subtracting these two formulas yields

$$\hat{x} - y = l(\hat{x} - y).$$

Thus $\hat{x} - y$ is an eigenvector of l corresponding to the eigenvalue $\lambda = 1$. This contradicts $\rho(\mathscr{L}) < 1$ and therefore $\hat{x} = y$.

ad (10): According to (8) and Theorem 2 there exists at least one fixed point $\hat{y}$ of f in Y. Supposing f to have two distinct fixed points $a, b \in Y$ with $a \neq b$ leads to

$$a - b = f(a) - f(b) \in \mathscr{K}(a - b).$$

Thus a linear operator $l \in \mathscr{K}$ exists with

$$a - b = l(a - b).$$

Thus $a-b$ is an eigenvector with corresponding eigenvalue $\lambda=1$ contradicting

$$\rho(l)\leqq\rho(\mathscr{K})\leqq\rho(\mathscr{L})<1. \qquad \square$$

The uniqueness property has different applications for the contexts of Theorem 3 and Theorem 5, respectively.

Corollary 5: *The conclusions of Theorem 5 remain valid for real normed vector spaces $\mathscr{M}$ of finite dimension.*

Proof: We can argue as in the proof of Theorem 5 except that we have to prove $\rho(\mathscr{L})<1$ again since (8) is valid only in a real vector space.

Substituting $W:=U-\hat{y}$, $V:=W+iW(i=\sqrt{-1})$ and regarding $0\in l(W)\mathring{\subset}W$, we get

$$l(V)=l(W+iW)=l(W)+il(W)\mathring{\subset}W+iW=V$$

i.e.

$$0\in l(V)\mathring{\subset}V.$$

So (14) holds in the complex vector space $\mathscr{M}+i\mathscr{M}$ and the proof concludes as before. $\square$

Remark 7: The set of linear operators $\mathscr{L}_{(f,\tilde{x},X)}$ in Theorem 5 need not be the complex of Jacobian matrices

$$f'(X):=\{(f_i'(x_{i1},\ldots,x_{in}))\mid x_{ij}\in X_i;\ i,j=1(1)n\}.$$

However, in practice one takes $\mathscr{L}=\mathscr{L}(f,\tilde{x},X):=f'(X)$ or $\mathscr{L}=\mathscr{L}(f,X)\supseteq f'(X)$ to be able to verify condition (10). In this case F becomes

$$F(X):=f(\tilde{x})+f'(X)*(X-\tilde{x}),$$

where

$$f'(X):=\{(f_i'(x))\mid x\in X\}\quad\text{and}\quad\tilde{x}\in X.$$

This property is some kind of "mean value inclusion", since every continously differentiable function f satisfies such a condition for convex $X\in\mathscr{P}\mathscr{M}$.

Note that even from the property (8) (which is easy to verify) we can deduce the interesting and important fact (9) about the spectral radius. From this in turn and in particular for the operator (1), we can deduce that the matrix R as well as every matrix $G\in g'(X)$ is non-singular.

Applying Theorem 5 and Corollary 6 permits the computation of inclusions for the eigenvalues and eigenvectors of a matrix (see [11]) as well as the verification of existence and uniqueness of the solution of boundary value problems in the small. Development of the corresponding E-method is deferred for future work.

For simplicity in the following, we restrict ourselves to systems of linear equations. This causes no loss of generality, since locally systems of non-linear equations have the same behavour as linear systems and can be treated similarly.

3. Application to Systems of Linear Equations

In the following $I\mathbb{R}^n$ denotes the space of n-dimensional interval vectors, R is some fixed but arbitrary $n \times n$-matrix and A a real $n \times n$-matrix. $b, x, \tilde{x}, \hat{x}, y \in \mathbb{R}^n$ are real vectors. E *is the* $n \times n$ identity matrix and $X, Y \in I\mathbb{R}^n$ are interval vectors.

In [5] and [11] solving the linear system $Ax = b$ proceeds by employing the formulas

$$f(y) = y + R \cdot (b - Ay) \tag{15}$$

for f and

$$F(X) = \tilde{x} + R(b - A\tilde{x}) + \{E - RA\} \cdot (X - \tilde{x}) \tag{16}$$

for F (the latter being a special form of (1)). Here in contrast to [6] and [9] $\tilde{x} \in \mathbb{R}^n$ is arbitrarily chosen (cf. [11]). Applying Theorem 3, Remark 4 and Theorem 5 yields the following.

Theorem 8: *Taking* F *as in* (16) *suppose that*

$$F(Y) \overset{\circ}{\subset} Y \tag{17}$$

holds for some $Y \in I\mathbb{R}^n$. *Then the following statements are valid*:

The matrices A *and* R *are non-singular, there exists exactly one solution* $\hat{x} \in \mathbb{R}^n$ *of* $Ax = b$, *moreover* $\hat{x} \in Y$. (18)

Furthermore $\hat{x} \in F^i(Y)$, $0 \leqq i \in \mathbb{N}$. (19)

The spectral radius of $E - RA$ *is less than unity*: $\rho(E - RA) < 1$. (20)

Proof:

ad (18): From Theorem 3 and Remark 4 we deduce that the linear mappings

$$g(t) := -R(b - A(t + \hat{x})) = RAt \quad \text{and} \quad g_1(s) := R \cdot s \quad \text{and} \quad g_2(r) := A \cdot r$$

are regular.

ad (19): See Theorem 1, (7) in [5].

ad (20): Since $f' = E - RA$ is constant the assertion follows from Theorem 5. □

Remark 9: Note that the approximations $\tilde{x}$ and R are not restricted in any way. In particular R is not assumed to be non-singular and $\tilde{x}$ need not be an element of X.

For the algorithmic application of Theorem 8 the function F in (16) has to be computed by interval arithmetic. This means that all operations $+, -, \cdot$ have to be evaluated by following the rules of interval arithmetic (see [1], [7] and [8]). In particular for the matrix-matrix-products and matrix-vector-products which occur, the so-called Bohlender-algorithm (or an equivalent) for computing precisely rounded scalar products should be used (see [2]).

Let $\tilde{F}$ denote the function F computed by interval arithmetic. Then clearly $F(X) \subseteq \tilde{F}(X)$ for every $X \in I\mathbb{R}^n$. Therefore, if an interval Y can be found (on a digital computer) such that, $\tilde{F}(Y) \overset{\circ}{\subset} Y$ we have

$$F(Y) \subseteq \tilde{F}(Y) \overset{\circ}{\subset} Y.$$

This means that (17) holds and the assertions of Theorem 8 are true.

Using Theorem 8 the algorithm proposed in [10] can be improved and simplified. Moreover proving the non-singularity of a matrix or the positive definiteness of symmetric matrix (cf. [11]) is made possible. These proofs are performed "numerically" on a computer; in the first case without computing or approximating the determinant and in the latter case without computing eigenvalues.

For application of (16) to a system of non-linear equations, A may be a convex interval including the complex of Jacobian matrices over the interval X and R may be an approximate inverse of the Jacobian Matrix at the point $\tilde{x}$.

4. Algorithm and Numerical Examples

Having discussed the algorithm which furnishes an interval containing the solution of a system of linear equations, there remains the question of specifying its details. This we now proceed to do. The proper inclusion $\overset{\circ}{\subset}$ for interval vectors $X, Y \in I\mathbb{R}^n$ may be programmed as follows:

$$X \overset{\circ}{\subset} Y :\Leftrightarrow \bigwedge_{i=1}^{n} \{\lambda X > \lambda Y \wedge \nu X < \nu Y\},$$

where λ and ν denote the left and right bounds of intervals, resp.

In the algorithm B_i denotes the i-th row of the matrix B. Further we define an ε-inflation of an interval as follows.

$$I \in I\mathbb{R} \Rightarrow I \circ \varepsilon := \begin{cases} I + 2\varepsilon \cdot d(I) & \text{for } d(I) \neq 0 \\ I + [-\eta, +\eta] & \text{for } d(I) = 0. \end{cases}$$

Here η is the smallest positive floating-point number on the computer being employed. For interval vectors the definition is to be understood componentwise. ε-inflation which is employed in step 5 of the algorithm is responsible for securing the convergence of the interval-iteration. In practice, 0.1 turned out to be a good value for ε (that is a 20% inflation of the diameter of the intervals). In almost all cases one interval iteration in step 5 was required with the choice of $\varepsilon = 0.1$. Moreover the including interval which was obtained was not overly wide.

In the following algorithm the mappings $\square : \mathbb{R}^n \to \mathbb{R}^n$, $\diamondsuit : I\mathbb{R}^n \to I\mathbb{R}^n$, resp. denote the roundings of the exact result to the nearest representable element resp. the smallest including interval of the appropriate date type of the computer (see [8]). Similarly, $\boxed{*} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, $\diamondsuit : I\mathbb{R}^n \times I\mathbb{R}^n \to I\mathbb{R}^n$ for $* \{+, -, \cdot, /\}$ denotes the rounded image of the exact result of the operation $*$ to the nearest representable element resp. the smallest including interval of the appropriate data type (see [8]).

1. Compute $R \approx A^{-1}$ approximately (e.g. using a floating-point Gaussian algorithm).
2. Compute $B := \diamondsuit(E - R \cdot A)$ using interval arithmetic.

3. $x^0 = R \cdot b$; $k := -1$;
   repeat $k := k+1$; $x^{k+1} = x^k \boxplus R \boxdot \square (b - Ax^k)$
   until $|x^{k+1} - x^k| / |x^k| \geqq 10^{-k}/2$ or
   $|x^{k+1} - x^k| / |x^k| < 10^{1-t}$; *
   The last iterate x^{k+1} is named $\tilde{x}$; $r := k+1$;
   $y^0 = R \cdot \square (b - A\tilde{x})$; $k := -1$;
   repeat $k := k+1$; $y^{k+1} = y^k \boxplus R \boxdot \square (b - A\tilde{x} - Ay^k)$
   until $|y^{k+1} - y^k| / |y^k| \geqq 10^{-k}/2$ or
   $|y^{k+1} - y^k| / |y^k| < 10^{2-t}$;
   The last iterate y^{k+1} is named $\tilde{y}$; $r := r+k+1$;

4. Compute $Y^0 := Z := R \diamondsuit \diamondsuit (b - A \cdot \tilde{x} - A \cdot \tilde{y})$;
   using interval arithmetic; $k := -1$;

5. **repeat** $k := k+1$; $Y^{k+1} := Y^k := Y^k \circ \varepsilon$,
   for $i := 1$ to n do $Y_i^{k+1} := \diamondsuit (Z_i + B \cdot Y^k)$
   until $Y^{k+1} \mathring{\subseteq} Y^k$ or $\{k > 2.25 \cdot r - 1\} =: \text{bool}$;
   if bool **then** stop: **

6. The last iterate Y^{k+1} is named Y.
   Now the non-singularity both of the matrices A and R has been verified by the new algorithm.
   Therefore there exists one and only one solution $\hat{x}$ of $Ax = b$ and we have
   $$\hat{x} \in \diamondsuit (x + \tilde{y} + Y).$$

For the precision and quality of the algorithm precise computation of the residues in the rows 9 and 13 from the top play a very important rôle. For instance, the i-th component of $(b - A\tilde{x} - Ax^k)$ is computed as a scalar product of length $2n+1$:

$$(b_i, -A_i, -A_i) \cdot (1, \tilde{x}, y^k).$$

We stress that the steps 4, 5, and 6 contain improvements over the algorithm in [10]. Instead of approximating and providing an including interval for the residue Y, an including interval is obtained for the relative residue $Y := Y - \tilde{y}$. This improvement is possible only if precise scalar products are used. A precise scalar product may be programmed by using Bohlender's algorithm or by using a long accumulator (cf. [2]).

The present procedure is usable on any digital computer provided that the precise scalar product (and therefore interval operations $+, -, \cdot$) are available.

To be perfectly clear we once more underscore the statement that "R and A are not singular" is a direct consequence of Theorem 8.

We say that the non-singularity of R and A have been verified a posteriori and with this therefore the existence and uniqueness of the solution $\hat{x}$ of $Ax = b$ as well as the inclusion $\hat{x} \in \tilde{x} + \tilde{y} + Y$ are assured. We call such methods E-methods. Theorem 8 can

* t is the mantissa length employed in the computer.

** Stopping in step 5 means that either A or R are singular or the accuracy being employed is not sufficient to solve the problem.

be applied because with the automatic error control of interval arithmetic and by Remark 9 of the previous chapter the hypothesis (17) has been proved to be (mathematically) valid by the algorithm.

The following table displays results corresponding to ill-conditioned systems of linear equations.

A	matrix degree n	q	iterations in step 5	$\|E-RA\|_1$	cond $(A)\approx$	$\{\Delta_{\text{rel}}\}_{\max}$
Hilbert	7		2	1.7	$1.3_{10}9$	$7_{10}-17$
Pascal	7		1	0.14	$2.5_{10}7$	$2_{10}-18$
	8		2	1.2	$3.7_{10}8$	$8_{10}-17$
Pascal*	8		1	0.11	$4.0_{10}7$	$2_{10}-18$
	9		3	3.5	$6.0_{10}8$	$1_{10}-18$
S	7		2	1.6	$6.0_{10}9$	$6_{10}-15$
Pascal′	20		1	11	$6.6_{10}18$	$2_{10}-6$
	21		1	5.1	$5.0_{10}18$	$7_{10}-5$
	22		1	210	$1.4_{10}20$	$1_{10}-6$
	23		1	300	$1.4_{10}20$	$1_{10}-6$
	24		1	670	$6.1_{10}20$	$1_{10}-6$
	25		1	2100	$2.3_{10}21$	$2_{10}-6$
	26		1	280000	$8.2_{10}22$	$2_{10}-5$
T	50	10^{-4}	1	0.15	$9_{10}6$	$1_{10}-18$
		10^{-5}	2	1.2	$9_{10}7$	$6_{10}-17$
	100	10^{-4}	2	0.56	$3_{10}7$	$1_{10}-18$
		10^{-5}	4	2.3	$3_{10}8$	$5_{10}-17$
	200	10^{-3}	4	0.49	$6_{10}7$	$8_{10}-17$

In the first column from left to right the type, degree and for the matrices T the value of q is shown. Here the Hilbert-matrices are defined by

$$H=(h_{ij}) \quad \text{with} \quad h_{ij}:=(i+j-1)^{-1},$$

the Pascal-matrices by

$$P'=P=(p_{ij}) \quad \text{with} \quad p_{ij}:=\binom{i+j}{i},$$

the Pascal*-matrices by

$$P^*=(p_{ij}^*) \quad \text{with} \quad p_{ij}^*:=\binom{i+j-1}{i}$$

and the matrix S by

$$s=(s_{ij}) \quad \text{with} \quad s_{ij}=\frac{\binom{n+j-1}{i-1}\cdot n\cdot\binom{n-1}{j-1}}{i+j-1}.$$

The matrices T are given by

$$T=Q-q\cdot U=(t_{ij}) \quad \text{with} \quad t_{ij}=1-q\cdot u_{ij},$$

for fixed q and some random numbers u_{ij} chosen in the interval $[0,1]$. All matrices except the Pascal′-matrices have been computed exactly (the coefficients of the Hilbert matrices were transformed into integers by multiplication of the associated linear system by a suitably large integer). The Pascal′-matrices were computed in double-precision floating point arithmetic and rounded. The right hand side of all systems was taken to be the vector $(1,\ldots,1)$.

In the second column the number of interval iterations in step 5 is displayed, in the third column the norm $\| E-RA \|$ (which is an estimation of the spectral radius) and in the fourth column the approximate condition number of A. In the fifth column the maximum relative error of the inclusion vector is shown. This error is defined as follows.

$$0 \notin X \in I\mathbb{R} \Rightarrow \{\Delta_{\text{rel}}\}_{\max} := \max_{x,y \in X} \left| \frac{y-x}{x} \right| = \frac{d(X)}{\min\limits_{x \in X} |x|}.$$

Note, that the maximum relative error of all components is taken, i.e., every component of the result vector is included with a maximum relative error as displayed. The results were obtained ins single-precision (that is with $8^1/_2$ decimal digit mantissa on the UNIVAC 1108 of the computing center of the University of Karlsruhe) using the precise scalar product of [2]. Double precision arithmetic is used only in the last addition $\tilde{x}+\tilde{y}+Y$ to be able to show the high accuracy of the results. Double precision accuracy without the precise scalar product was employed for the system corresponding to the Pascal'-matrices. In these cases the relatively weak accuracy of the results is obvious from the most right column of the table. Apart from those extreme cases at least 15 decimal digits can be guaranteed by computing in single-precision accuracy, i.e. $8^1/_2$ decimal digits (almost independent of condition number and order of the matrix). In particular systems were treated, were the a priori estimation of the spectral radius is greater than unity and where the construction of a priori bounds (e.g. using Banach's Fixed Point Theorem) is not possible.

The formula (1) occurs in [6] and is used in [12]. However, until now only approximate inverses R could be used for which a spectral estimation $q := \rho(E-RA)$ with $q<1$ were known a priori. This necessary a priori estimation can be omitted. The proof of the non-singularity of matrices A and R is now performed automatically and intrinsically (cf. fifth column in the table).

The algorithm is usable for systems with interval entries instead of real (point) systems. The final statement of the algorithm should read for every matrix $A \in \mathscr{A} \in I M_n \mathbb{R}$ the matrices R and every matrix $A \in \mathscr{A}$ are not singular and for every $A \in \mathscr{A}$ and every $b \in \mathscr{b}$ the linear system $Ax=b$ has one and only one solution $A\hat{x}=b$ and $\hat{x} \in \tilde{x}+\tilde{y}+Y$ holds".

The algorithm is programmed in FORTRAN and is both single and double precision accuracy (for point and interval systems, resp.). The algorithms are installed in the program library of the UNIVAC 1108 of the University of Karlsruhe and are widely used.

Using a precise scalar product (with Bohlender's algorithm or with a long accumulator) should by no means be interpreted as calculating with higher accuracy. Only at one specific point in the algorithm, namely calculation of the residue, is the procedure used for producing a single precision (precisely rounded) result. Performing the final addition $\tilde{x}+\tilde{y}+Y$ in double precision (which may be carried out by employing the precise scalar product) makes the higher accuracy achieved available to the user.

5. Computing Time

The computing time α reguired to compute an approximation $\tilde{x}$ of the solution of an n-th order system using the Gaussian algorithm with r residual iterations is (modulo linear terms)

$$\alpha = n^3/3 + 2rn^2.$$

The computing time β of the algorithm described in section 4 where step 5 is excuted s times is (modulo linear terms)

$$\beta = 2n^3 + n^2(3r+4s+4) \leqq 6\alpha + n^2(-9r+4s+4).$$

In both procedures yield approximations (resp. inclusions) of comparable accuracy, the relative increase in cost σ is

$$\sigma := \alpha/\beta \leqq 6 + \frac{-9r+4s+4}{n/3+2r},$$

Since in our algorithm $s \leqq (9r-4)/r$ holds universally, this last inequality may be expressed as $\sigma = 6 - 0(\frac{1}{n})$. Therefore, the computing time for the inclusion-algorithm is at most six times the computing time of a typical approximate algorithm. This estimate is independent of the order of the linear system.

One appraisal of this factor of six may be obtained by considering the following. Suppose that in order to get "a feeling" about the error of an approximation, a problem is first solved in single precision accuracy and then solved once more in double precision. This double procedure which produces no error estimation requires five times the computing time of the basic (single precision) algorithm.

References

[1] Alefeld, G., Herzberger, J.: Einführung in die Intervallrechnung. Mannheim-Wien-Zürich: Bibliographisches Institut 1974.

[2] Bohlender, G.: Floating-point computation of functions with maximum accuracy. IEEE Trans. on Computers *1977*, 621.

[3] Hansen, E.: Interval arithmetic in matrix computations, Part I. SIAM J. Numer. Anal. *2*, 308 – 320 (1965).

[4] Hansen, E., Smith, R.: Interval arithmetic in matrix computations. Part II. SIAM J. Numer. Anal. *4*, 1 – 9 (1967).

[5] Kaucher, E., Rump, S. M.: Generalized iteration methods for bounds of the solution of fixed operator equations. Computing *24*, 131 – 137 (1980).

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

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

[8] Kulisch, U., Miranker, W. L.: Computer arithmetic in theory and practice. Academic Press 1981.

[9] Moore, R. E.: A test for existence of solutions for nonlinear systems. SIAM J. Numer. Analysis *4* (1977).

[10] Rump, S. M., Kaucher, E.: Small bounds for the solution of systems of linear equations, in: Computing, Suppl. 1. Wien-New York: Springer 1978.

[11] Rump, S. M.: Kleine Fehlerschranken bei Matrixproblemen. Dissertation, Universität Karlsruhe, 1980.

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

[13] Kulisch, U., Wippermann, H.-W.: PASCAL-SC, PASCAL für wissenschaftliches Rechnen, Gemeinschaftsentwicklung von Institut für Angewandte Mathematik, Universität Karlsruhe (Prof. Dr. U. Kulisch), Fachbereich Informatik, Universität Kaiserslautern (Prof. Dr. H.-W. Wippermann).

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