RIGOROUS SENSITIVITY ANALYSIS
FOR SYSTEMS OF LINEAR AND NONLINEAR EQUATIONS

SIEGFRIED M. RUMP

ABSTRACT. Methods are presented for performing a rigorous sensitivity analysis of numerical problems with independent, noncorrelated data for general systems of linear and nonlinear equations. The methods may serve for the following two purposes. First, to bound the dependency of the solution on changes in the input data. In contrast to condition numbers, a componentwise sensitivity analysis of the solution vector is performed. Second, to estimate the true solution set for problems whose input data are subject to tolerances. The methods presented are very effective and have the additional property that, owing to an automatic error control mechanism, every computed result is guaranteed to be correct. Examples are given for linear systems, demonstrating that the computed bounds are in general very sharp. Interesting comparisons to traditional condition numbers are given.

0. INTRODUCTION

In the first part of the paper we concentrate on the theoretical results; the practical implementation is discussed in §3.

Let T denote one of the sets $\mathbf{R}$ (real numbers) or $\mathbf{C}$ (complex numbers). Vectors $v \in VT$ and matrices $A \in MT$ consist of n and $n \times n$ components, respectively, throughout this paper. Let S denote one of the sets T, VT, or MT. The power set over one of these sets is denoted by $\mathbf{P}T$, $\mathbf{P}VT$, $\mathbf{P}MT$, respectively.

Unless otherwise stated, operations $+$, $-$, $\cdot$, $/$ are power set operations throughout this paper, defined in the usual way. Sets occurring several times in an expression are treated independently, e.g.

$$Z \in \mathbf{P}S: \quad Z * Z := \{z_1 * z_2 | z_1, z_2 \in Z\} \supseteq \{z * z | z \in Z\}$$

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

The infimum $\inf(z)$ and supremum $\sup(z)$ of nonempty and bounded sets $Z \in \mathbf{P}S$ are defined in the usual way, in case of vectors and matrices componentwise (so that $\inf(A) \in MT$ when $A \in \mathbf{P}MT$). The diameter $d(Z)$ and the radius $r(Z)$ of some nonempty, bounded $Z \in \mathbf{P}S$ are defined by

$$d(Z) := \sup(Z) - \inf(Z) \quad \text{and} \quad r(Z) := 0.5 \cdot d(Z).$$

The diameter of $A \in \mathbf{P}MT$ is the matrix of diameters of its components. Throughout this paper we use partial ordering for complex numbers and for vectors and matrices over those, i.e., for $T \in \{\mathbf{C}, V\mathbf{C}, M\mathbf{C}\}$

$$z_1, z_2 \in T\colon z_1 \leq z_2\colon \Leftrightarrow \operatorname{Re}(z_1) \leq \operatorname{Re}(z_2) \text{ and } \operatorname{Im}(z_1) \leq \operatorname{Im}(z_2).$$

1. Bounds on the sensitivity

In [15, 16] Neumaier gives estimations on the sensitivity of systems of linear equations. He computes sensitivity bounds, together with an inclusion of the solution, using methods described e.g. in [18, 20]. The bounds are sharp but require some additional computational effort. In particular, the solution of the linear system with another right-hand side is needed; for guaranteed estimations on the sensitivity, this solution has to be guaranteed to be correct.

In the following we use ideas of Neumaier to design rigorous bounds on the sensitivity of a linear system, together with an inclusion for the solution, with very little additional computational effort. Similar techniques for systems of nonlinear equations are described in §2. The sensitivity is bounded by estimating the interior of the solution set of a linear system whose data are subject to tolerances.

We start with a lemma which estimates the diameter of sets.

Lemma 1. *Let* $S \in \{\mathbf{R}, V\mathbf{R}, M\mathbf{R}, \mathbf{C}, V\mathbf{C}, M\mathbf{C}\}$, *and let* Q, Z, $\Delta \in \mathbf{P}S$ *be nonempty and bounded subsets of* S *with*

$$(1.1) \qquad\qquad Q \subseteq Z - \Delta.$$

Then

$$(1.2) \qquad\qquad \inf(Z) \leq \inf(Q) + \sup(\Delta)$$

and

$$(1.3) \qquad\qquad \sup(Z) \geq \sup(Q) + \inf(\Delta).$$

Proof. Without loss of generality we prove the real vector case only. The other cases can be dealt with similarly, for example, the complex case by treating real and imaginary parts. Formula (1.1) states

$$(1.4) \qquad\qquad \forall q \in Q \; \exists z \in Z \; \exists \delta \in \Delta\colon q = z - \delta.$$

For every $1 \leq i \leq n$ there is a convergent sequence $\{q^k\}$, $k \in \mathbf{N}$, with $q^k \in Q$ for all $k \in \mathbf{N}$ and

$$(1.5) \qquad\qquad \lim_{k \to \infty} q_i^k = (\inf(Q))_i.$$

By (1.4) there follows

$$\forall k \in \mathbf{N} \; \exists z \in Z \; \exists \delta \in \Delta\colon z = q^k + \delta.$$

Therefore, for fixed i, $1 \leq i \leq n$, one has

$$(1.6) \qquad \forall k \in \mathbf{N} \colon (\inf(Z))_i \leq z_i = q_i^k + \delta_i \leq q_i^k + (\sup(\Delta))_i.$$

Since (1.6) holds for every $k \in \mathbf{N}$ and $i \in \{1, \ldots, n\}$, the assertion (1.2) follows from (1.5). Assertion (1.3) follows similarly. $\square$

In the following we use Lemma 1 to derive an estimation on the infimum and supremum of the solution set of a set of linear systems, the latter being defined by

Definition 2. Let $T \in \{\mathbf{R}, \mathbf{C}\}$, $[A] \in \mathbf{P}MT$, and $[b] \in \mathbf{P}VT$. Then

$$(1.7) \qquad \sum([A], [b]) := \{x \in VT \mid \exists A \in [A] \; \exists b \in [b] \colon Ax = b\}.$$

Definition 2 does not require all, or even any, $A \in [A]$ to be nonsingular; $\sum([A], [b])$ may be empty.

First we mention an outer estimation for $\sum([A], [b])$ given in [18]. All operations are power set operations.

Theorem 3. *Let* $T \in \{\mathbf{R}, \mathbf{C}\}$, $\varnothing \neq [A] \in \mathbf{P}MT$, $\varnothing \neq [b] \in \mathbf{P}VT$, $\tilde{x} \in VT$, $R \in MT$, *and* $\varnothing \neq X \in \mathbf{P}VT$ *compact. Define*

$$(1.8) \qquad Y := \tilde{x} + R \cdot ([b] - [A] \cdot \tilde{x}) + \{I - R \cdot [A]\} \cdot (X - \tilde{x}).$$

If

$$(1.9) \qquad\qquad Y \subseteq \operatorname{int}(X),$$

then R *and every matrix* $A \in [A]$ *is nonsingular,* $\sum([A], [b])$ *is nonempty, and*

$$(1.10) \qquad\qquad \sum([A], [b]) \subseteq Y.$$

Remark. I denotes the identity matrix and $\operatorname{int}(X)$ the interior of X.

Estimations of the set Y described by Theorem 3 are effectively computable, as has been shown in [18, 20]. Note that there are no a priori assumptions on the nonsingularity of R or of matrices in $[A]$. Next we give an upper and lower estimation of the infimum and supremum of $\sum([A], [b])$, using Lemma 1, which is virtually free of cost, together with the outer estimation (1.10).

Theorem 4. *Let* $T \in \{\mathbf{R}, \mathbf{C}\}$, $\varnothing \neq [A] \in \mathbf{P}MT$, $\varnothing \neq [b] \in \mathbf{P}VT$, $\tilde{x} \in VT$, $R \in MT$, *and assume every matrix in* $[A]$ *is nonsingular. Define*

$$(1.11) \qquad \begin{aligned} Q &:= \tilde{x} + R \cdot ([b] - [A] \cdot \tilde{x}), \\ \Delta &:= (I - R \cdot [A]) \cdot \left(\sum([A], [b]) - \tilde{x} \right). \end{aligned}$$

Then

$$(1.12) \qquad\qquad \inf \left(\sum([A], [b]) \right) \leq \inf(Q) + \sup(\Delta),$$

and

$$(1.13) \qquad \sup \left(\sum([A], [b]) \right) \geq \sup(Q) + \inf(\Delta).$$

Remark. All operations in Theorem 4 are power set operations.

Proof. For nonsingular $A \in MT$ and for $b \in VT$ there holds

$$(1.14) \qquad \tilde{x} + R \cdot (b - A\tilde{x}) = A^{-1}b - (I - R \cdot A) \cdot (A^{-1} \cdot b - \tilde{x}).$$

The set $\sum([A], [b])$ is nonempty and

$$
\begin{aligned}
Q &= \{\tilde{x} + R \cdot (b - A \cdot \tilde{x}) \mid A \in [A], \, b \in [b]\} \\
&= \{A^{-1} \cdot b - (I - R \cdot A) \cdot (A^{-1} \cdot b - \tilde{x}) \mid A \in [A], \, b \in [b]\} \\
&\subseteq \sum([A], [b]) - \{(I - R \cdot A) \cdot (A^{-1} \cdot b - \tilde{x}) \mid A \in [A], \, b \in [b]\} \\
&\subseteq \sum([A], [b]) - \{(I - R \cdot A_1) \cdot (A_2^{-1} \cdot b - \tilde{x}) \mid A_1, \, A_2 \in [A], \, b \in [b]\} \\
&= \sum([A], [b]) - (I - R \cdot [A]) \cdot \left(\sum([A], [b]) - \tilde{x} \right) \\
&= \sum([A], [b]) - \Delta,
\end{aligned}
$$

by virtue of Definition 2 and the nonsingularity of every $A \in [A]$. Applying Lemma 1 with $Z := \sum([A], [b])$ finishes the proof. $\square$

The infimum and supremum of $\sum([A], [b])$ are the (componentwise) variations of the set of solutions of $Ax = b$ for $A \in [A]$ and $b \in [b]$.

It may appear to be a vicious circle to estimate $\sum([A], [b])$ by some expression depending on $\sum([A], [b])$. This is not true. Using an outer estimation Y of $\sum([A], [b])$, e.g. some outer estimation computed by using Theorem 3, and defining $\Delta^* := (I - R \cdot [A]) \cdot (Y - \tilde{x})$, yields

$$\Delta = (I - R \cdot [A]) \cdot \left(\sum([A], [b]) - \tilde{x} \right) \subseteq \Delta^*,$$

which implies

$$(1.15) \qquad Q \subseteq \sum([A], [b]) - \Delta^*.$$

Hence, applying Lemma 1 proves estimations (1.12) and (1.13) with Δ replaced by Δ^*.

To obtain very sharp estimations of the infimum and supremum of the set $\sum([A], [b])$, it is necessary that

$$d((I - R \cdot [A]) \cdot (Y - \tilde{x})) \ll d(\tilde{x} + R \cdot ([b] - [A] \cdot \tilde{x})).$$

If $[A]$ and $[b]$ are of small diameter, this is true, in general, because with $R \approx A^{-1}$ for some $A \in [A]$ and $\tilde{x} := R \cdot b$ for some $b \in [b]$ the quantities $I - R \cdot [A]$, $Y - \tilde{x}$, and $[b] - [A] \cdot \tilde{x}$ are of the same small order of magnitude.

It should be mentioned that if, for some source, there is further knowledge on the structure and especially on the interior of $(I - R \cdot [A]) \cdot (Y - \tilde{x})$, estimations (1.12) and (1.13) can be further sharpened.

2. Systems of nonlinear equations

In [18, 20] methods have been given for computing guaranteed bounds for the solution of nonlinear equations involving differentiable real or complex functions in one or more variables. The formulation of such a method requires an appropriate definition of derivative.

Definition 5. Let $T \in \{\mathbf{R}, \mathbf{C}\}$ and $f: D \to VT$, $D \subseteq VT$ continuously differentiable. Let $f': VT \to MT$ denote the Jacobian of f. Then for $X \in \mathbf{P}VT$, $X \subseteq D$,

$$(2.1) \qquad f'(X) := \left\{ \left(\frac{\partial f_1}{\partial x}(\zeta_1), \dots, \frac{\partial f_n}{\partial x}(\zeta_n) \right)^t | \zeta_1, \dots, \zeta_n \in X \right\}.$$

f' coincides with the Jacobian if X consists of one point. f' is chosen such that for $x \sqcup y \subseteq D$

$$(2.2) \qquad \forall x, y \in D \; \exists Q \in f'(x \sqcup y): f(y) = f(x) + Q \cdot (y - x),$$

where $\sqcup$ denotes the convex union and ^{t} denotes transposition (see [20]).

An inclusion of a zero of each individual function out of a set of functions can be effectively calculated according to the following theorem (see [18]):

Theorem 6. *Let* $T \in \{\mathbf{R}, \mathbf{C}\}$, *and let* $[f]$ *be a nonempty set of continuously differentiable functions* $f: D \to VT$, $D \subseteq VT$. *For* $R \in MT$, $\tilde{x} \in VT$, *and compact and convex* $\varnothing \neq X \in \mathbf{P}VT$, *define*

$$(2.3) \qquad Y := \bigcup \{ \tilde{x} - R \cdot f(\tilde{x}) + (I - R \cdot [Q]) \cdot (X - \tilde{x}) | f \in [f] \},$$

where $[Q] := \bigcup \{ f'(\tilde{x} \sqcup X) | f \in [f] \}$.
  If

$$(2.4) \qquad\qquad\qquad Y \subseteq \operatorname{int}(X),$$

then the matrix R *and every matrix* $Q \in [Q]$ *is nonsingular. Furthermore, every nonlinear system* $f \in [f]$ *has exactly one zero in* Y:

$$\forall f \in [f] \; \exists^{1\text{-}1} \hat{x} \in Y: f(\hat{x}) = 0.$$

For the proof cf. [18]. The methods derived in §1 allow us to give inner estimations, i.e., estimations of the infimum and supremum of the set $\{\hat{x} \in Y | \exists f \in [f]: f(\hat{x}) = 0\}$.

Let the assumptions of Theorem 6 be satisfied, especially (2.4) with (2.3) . Then for every $f \in [f]$ there is one and only one $\hat{x}_f \in Y$ with $f(\hat{x}_f) = 0$. Therefore, the set

$$(2.5) \qquad\qquad Z := \{ x \in Y | \exists f \in [f]: f(x) = 0 \}$$

is nonempty. f' is defined in such a way (see (2.2)) that for every $f \in [f]$ there exists a $Q_f \in [Q]$ with

$$(2.6) \qquad\qquad f(\tilde{x}) = f(\hat{x}_f) + Q_f \cdot (\tilde{x} - \hat{x}_f).$$

Q_f is, in general, not uniquely determined. By (2.5) and (2.6) the following is true for every $f \in [f]$:

$$
\begin{aligned}
\tilde{x} - R \cdot f(\tilde{x}) &= \tilde{x} - R \cdot \{f(\hat{x}_f) + Q_f \cdot (\tilde{x} - \hat{x}_f)\} \\
&= \hat{x}_f - \{I - R \cdot Q_f\} \cdot (\hat{x}_f - \tilde{x}) \\
&\subseteq Z - \{I - R \cdot f'(\tilde{x} \underline{\cup} X)\} \cdot (X - \tilde{x}).
\end{aligned}
$$

Defining

$$
\begin{aligned}
(2.7) \qquad \Delta &:= \bigcup \{(I - R \cdot f'(\tilde{x} \underline{\cup} X)) \cdot (X - \tilde{x}) | f \in [f]\}, \\
Q &:= \{\tilde{x} - R \cdot f(\tilde{x}) | f \in [f]\}
\end{aligned}
$$

yields

$$
(2.8) \qquad\qquad\qquad\qquad Q \subseteq Z - \Delta.
$$

Together with Lemma 1, this proves the following theorem.

Theorem 7. *Let* $T \in \{\mathbf{R}, \mathbf{C}\}$, *and let* $[f]$ *be a nonempty set of continuously differentiable functions* $f: D \to VT$, $D \subseteq VT$. *For* $R \in MT$, $\tilde{x} \in VT$, *and compact and convex* $\varnothing \neq X \in PVT$, *define* Y *by* (2.3). *If* $Y \subseteq \text{int}(X)$, *then the following holds true. The set* Z *defined by* (2.5) *is nonempty and*

$$
(2.9) \qquad \inf(Z) \leq \inf(Q) + \sup(\Delta), \qquad \sup(Z) \geq \sup(Q) + \inf(\Delta),
$$

where Q *and* Δ *are defined in* (2.7).

Theorem 7 is applicable on digital computers. In particular, estimations for $f'(\tilde{x} \underline{\cup} X)$ can be computed automatically without calculating the Jacobian explicitly, by computing the *value* of the derivative of an arbitrary function implicitly (see e.g. [4, 12, 14, 17, 21]). Sets of functions can be stored on computers using interval techniques, as will be described in §3.

As in the case of linear systems, (2.9) estimates inner bounds for edges of the smallest hyperrectangle containing Z. The bounds are very sharp as long as $d(\Delta) \ll d(Q)$.

Having bounds for general systems of nonlinear equations, similar estimations of the overestimation of calculated inclusions for other problem areas in numerical analysis, such as eigenvalue problems, polynomial zeros, singular values, etc., can be derived.

3. Implementation on digital computers

In order to implement our methods on a digital computer, a number of problems have to be solved. First, we need an appropriate representation for sets which is simple enough to allow efficient arithmetic operations, and general enough not to be too restrictive for practical applications. Second, an appropriate arithmetic has to be defined, allowing simple and fast execution, with the property that inner and outer estimations of the corresponding power set operations are possible. Third, the arithmetic must handle rounding errors in

an appropriate way to maintain the guarantee of correctness of all results. We want to stress that *any* arithmetic having the above properties is suitable for the following discussions. Here, we will concentrate on a rectangular interval arithmetic. First we discuss this for the set of real or complex numbers, postponing the problems of rounding errors for floating-point numbers.

The set of intervals, i.e., hyperrectangles over real, resp. complex, numbers is denoted by $\mathbf{IR}$, resp. $\mathbf{IC}$. Let $T \in \{\mathbf{R}, \mathbf{C}\}$; then we define

$$[A] \in \mathbf{I}T \Leftrightarrow [A] = \{A \in T | \underline{A} \le A \le \overline{A} \text{ for some } \underline{A}, \overline{A} \in T\}.$$

Obviously, $\underline{A} = \inf([A])$ and $\overline{A} = \sup([A])$. Intervals over vectors, resp. matrices $(\mathbf{I}VT$, resp. $\mathbf{I}MT)$, are defined as vectors, resp. matrices, of intervals.

In contrast with usual definitions, we do not require $\underline{A} \le \overline{A}$ for intervals. This means, for instance in the case of matrices, that some components of an interval matrix may be empty. We do not need, and do not define, operations for those; such interval matrices are needed in results which contain useful information for the nondegenerated components.

The rules of interval arithmetic (see [2, 12]) define an arithmetic which is best possible in the sense that the result interval is the smallest interval containing the result of the power set operation. More precisely, let S_1, S_2, $S_3 \in \{\mathbf{R}, V\mathbf{R}, M\mathbf{R}, \mathbf{C}, V\mathbf{C}, M\mathbf{C}\}$ and let $[A] \in \mathbf{I}S_1$ and $[B] \in \mathbf{I}S_2$ be intervals such that for some fixed but arbitrary operation $* \in \{+, -, \cdot, /\}$

$$A * B \quad \text{for all } A \in [A], B \in [B]$$

is well defined, with result in $\mathbf{P}S_3$. Then the corresponding interval operation $\hat{*}$ is defined by

$$(3.1) \qquad [A] \hat{*} [B] = \bigcap \{[C] \in \mathbf{I}S_3 | A * B \in [C]$$
$$\text{for all } A \in [A], B \in [B]\}.$$

It can be shown (see [2, 12]) that all operations $\hat{*}$ according to (3.1) are well defined and, most important, are effectively computable using the componentwise definition (except for complex division, which we do not need here). For example, the multiplication of two interval matrices $[A]$, $[B] \in \mathbf{I}MT$ with $T \in \{\mathbf{R}, \mathbf{C}\}$ can be performed by using

$$(3.2) \qquad ([A] \hat{\cdot} [B])_{ij} = [A]_{i1} \hat{\cdot} [B]_{1j} \hat{+} \cdots \hat{+} [A]_{in} \hat{\cdot} [B]_{nj},$$

where n is the number of columns of every $A \in [A]$ and rows of every $B \in [B]$. This componentwise definition (3.2) is indeed identical with definition (3.1).

However, in an interval matrix multiplication the components of every matrix occur several times, so that the computed interval matrix is in general an overestimation of the power set operation:

$$[A] \cdot [B] = \{a \cdot b | a \in [A], b \in [B]\} \subseteq [A] \hat{\cdot} [B].$$

This is not the case for multiplying intervals over T, or for performing a dot product of two interval vectors.

For our subsequent considerations it is especially important to notice that the multiplication of an interval matrix by a point vector does not imply an overestimation:

$$(3.3) \qquad [A] \in \mathbf{I}MT, \ \ b \in VT \Rightarrow [A] \cdot b = [A] \hat{\cdot} b .$$

This is true because every component of the interval matrix $[A]$ occurs only once in the process of the multiplication. Addition and subtraction of intervals over scalars, vectors, or matrices are always identical with the power set operations without any overestimation.

For the multiplication of a point matrix $R \in MT$ by an interval vector $[b] \in \mathbf{I}VT$ we have at least

$$(3.4) \qquad \inf(R \hat{\cdot} [b]) = \inf(R \cdot [b]) \quad \text{and} \quad \sup(R \hat{\cdot} [b]) = \sup(R \cdot [b]).$$

This can be shown, for instance, by estimating the multiplication component-wise.

According to the proof of Theorem 4, we need to compute an inner estimation of Q and an outer estimation of Δ and of $\sum([A], [b]) - \Delta$. The latter problem can be solved by replacing every operation in the computation of Δ by its corresponding interval operation:

$$(3.5) \qquad [\Delta] \subseteq (I \hat{-} R \hat{\cdot} [A]) \hat{\cdot} (X \hat{-} \tilde{x}) .$$

The first problem is more difficult to solve. The proof of Lemma 1 shows that for our purposes it suffices to have for fixed, but arbitrary $i \in \{1, \dots, n\}$ a sequence of $q^k \in Q$ with $\lim_{k \to \infty} q_i^k = (\inf(Q))_i$ and a similar sequence for the supremum of Q.

Intervals are closed; therefore, some $\underline{q}, \ \bar{q} \in Q$ with

$$(\underline{q})_i = (\inf(Q))_i \quad \text{and} \quad (\bar{q})_i = (\sup(Q))_i$$

can be found. However, such $\underline{q}, \ \bar{q}$ are effectively computable using interval operations and (3.3). By (3.3), we already know that

$$[b] - [A] \cdot \tilde{x} = [b] \hat{-} [A] \hat{\cdot} \tilde{x} .$$

From this and (3.4), there follows

$$\inf(Q) = \inf(\tilde{x} + R \cdot ([b] - [A]\tilde{x})) = \inf(\tilde{x} \hat{+} R \hat{\cdot} ([b] \hat{-} [A] \hat{\cdot} \tilde{x}))$$

and

$$\sup(Q) = \sup(\tilde{x} + R \cdot ([b] - [A]\tilde{x})) = \sup(\tilde{x} \hat{+} R \hat{\cdot} ([b] \hat{-} [A] \hat{\cdot} \tilde{x})) .$$

Using definition (3.4) for $[\Delta]$ and $[Q] := \tilde{x} \hat{+} R \hat{\cdot} ([b] \hat{-} [A] \hat{\cdot} \tilde{x})$, and carefully following the proof of Lemma 1, demonstrates that

$$(3.6) \qquad \begin{aligned} \inf\left(\sum([A], [b])\right) &\leq \inf([Q]) + \sup([\Delta]), \\ \sup\left(\sum([A], [b])\right) &\geq \sup([Q]) + \inf([\Delta]). \end{aligned}$$

Bounds on the infimum and supremum of $\sum([A],[b])$ are therefore computable using hyperrectangles for the representation of sets, by computing $[Q]$ and $[\Delta]$ using traditional interval operations, and by applying (3.6). If the diameter of $[\Delta]$ gets too large, some or, in extreme cases, all inner estimations on the components of $\sum([A],[b])$ may become empty.

In the following we use the definition of an interval in terms of its bounds for some $A_1,\ A_2 \in T \in \{\mathbf{R}, V\mathbf{R}, M\mathbf{R}, \mathbf{C}, V\mathbf{C}, M\mathbf{C}\}$:

$$[A_1,A_2] := \{A \in T \,|\, A_1 \le A \le A_2\} \in \mathbf{I}T.$$

In this notation, (3.6) can be written as

(3.7)
$$[\inf([Q]) + \sup([\Delta])\,,\, \sup([Q]) + \inf([\Delta])]$$
$$\subseteq \left[\inf\left(\sum([A],[b])\right)\,,\, \sup\left(\sum([A],[b])\right)\right].$$

The final goal is to calculate bounds similar to (3.7) on a computer. This is made difficult by the fact that digital computers only allow the exact representation of a finite set of floating-point numbers to approximate the infinite set of real or complex numbers.

To achieve this goal, we need appropriate rounding procedures from the real numbers $\mathbf{R}$ into a set $\mathbf{F} \subseteq \mathbf{R}$ of floating-point numbers. In order not to exclude certain arithmetics, we state the mathematically necessary properties for these rounding operations and for an appropriate arithmetic. Rounding occurs always together with an arithmetic operator; therefore, we add the rounding symbol to the operator.

Let $\mathbf{F} \subseteq \mathbf{R}$ denote some finite subset of $\mathbf{R}$ (which may be regarded as the set of floating-point numbers on a computer) and $\mathbf{CF} := F + i \cdot F$ be a complex extension of $\mathbf{F}$. Vectors and matrices over $\mathbf{F}$ and $\mathbf{CF}$ are defined as n-tuples, resp. n^2-tuples, forming the sets $V\mathbf{F}$, $M\mathbf{F}$, resp. $V\mathbf{CF}$, $M\mathbf{CF}$.

The set of intervals $\mathbf{IF}$ over $\mathbf{F}$ is defined by

$$[A] \in \mathbf{IF}\colon \Leftrightarrow [A] = [\underline{a},\overline{a}] = \{a \in \mathbf{R} \,|\, \underline{a} \le a \le \overline{a}\}.$$

The corresponding sets $\mathbf{ICF}$, $\mathbf{IVF}$, $\mathbf{IMF}$, $\mathbf{IVCF}$, and $\mathbf{IMCF}$ are defined similarly.

Let $T_1,\ T_2,\ T_3 \in \{\mathbf{F}, V\mathbf{F}, M\mathbf{F}, \mathbf{CF}, V\mathbf{CF}, M\mathbf{CF}\}$ with corresponding sets $S_1,\ S_2,\ S_2 \in \{\mathbf{R}, V\mathbf{R}, M\mathbf{R}, \mathbf{C}, V\mathbf{C}, M\mathbf{C}\}$, respectively. Using the canonical embedding $T_i \subseteq S_i$, $i = 1, 2, 3$, let $* \in \{+, -, /, \cdot\}$ be an operator such that for $A_1 \in T_1$ and $A_2 \in T_2$, the image $A_1 * A_2$ is well defined and $A_1 * A_2 \in S_3$. Let $[A]_1 \in \mathbf{I}T_1$ and $[A]_2 \in \mathbf{I}T_2$ be given. Then $\divideontimes$ is an operator with $\divideontimes\colon \mathbf{I}T_1 \times \mathbf{I}T_2 \to \mathbf{I}T_3$ satisfying

(3.8) $$[A]_1 \divideontimes [A]_2 \subseteq [\inf([A]_1 * [A]_2)\,,\, \sup([A]_1 * [A]_2)],$$

where the last two operations in (3.8) are the power set operations over S_i in the canonical embedding $T_i \subseteq S_i$, $i = 1, 2, 3$. The result may be the empty set for some components. We deliberately do not restrict the operators $\divideontimes$ in

any way except requiring property (3.8). Operators $\stackrel{*}{*}$ give inner estimations on the infimum and supremum of a power set operation.

In a practical implementation it suffices, in principle, to have operators $\stackrel{*}{*}$: **IF** $\times$**IF** $\rightarrow$ **IF** which can be implemented taking advantage of the different rounding modes. Such operations are, for instance, defined in the IEEE 754 standard for binary floating-point arithmetic [5] or in [10, 11]. Operations over **CF** and vector and matrix operations over **F** and **CF** can be defined componentwise using appropriate roundings.

For vector and matrix operations, better results are achieved when using the inner product proposed by Kulisch (see [9, 10, 11]). These inner product algorithms are especially advantageous for point vectors.

For interval operations, i.e., operations with outer roundings, over floating-point numbers we use (without danger of confusion) the same symbol $\stackrel{*}{*}$: $IT_1 \times IT_2 \rightarrow IT_3$ as for those over real numbers. Compared to interval operations over $\{R, VR, MR, C, VC, MC\}$, interval operations over $\{F, VF, MF, CF, VCF, MCF\}$ deliver slightly wider results. They have the property

$$(3.9) \qquad [A]_1 \stackrel{\wedge}{*} [A]_2 \supseteq [\inf([A]_1 * [A]_2), \sup([A]_1 * [A]_2)]$$

for every $[A]_1 \in IT_1$ and $[A]_2 \in IT_2$.

Summarizing the discussion above, and using interval floating-point operations $\stackrel{\vee}{*}$ and $\stackrel{\wedge}{*}$, we can state the following theorem.

Theorem 8. *Let* $F \subseteq R$, $CF \subseteq C$, *and* $T \in \{F, CF\}$ *with operations* $\stackrel{\vee}{*}$ *and* $\stackrel{\wedge}{*}$ *having the properties* (3.8) *and* (3.9), *respectively. Let* $[A] \in IMT$, $[b] \in IVT$, $\tilde{x} \in VT$, $R \in MT$, *and* $X \in IVT$ *be given, where every component of* $[A]$, $[b]$, *and* X *is nonempty. Let*

$$(3.10) \qquad Y := \tilde{x} \stackrel{\wedge}{+} R \stackrel{\wedge}{\cdot} ([b] \stackrel{\vee}{-} [A] \stackrel{\vee}{\cdot} \tilde{x}) \stackrel{\wedge}{+} (I \stackrel{\vee}{-} R \stackrel{\wedge}{\cdot} [A]) \stackrel{\wedge}{\cdot} (X \stackrel{\wedge}{-} \tilde{x}).$$

If $Y \subseteq \text{int}(X)$, *then* R *and every matrix* $A \in MC$ *with* $A \in [A]$ *is nonsingular, the solution set* $\sum([A], [b])$ *defined in Definition 2 satisfies*

$$\sum([A], [b]) \subseteq Y,$$

and for $[Q] := \tilde{x} \stackrel{\vee}{+} R \stackrel{\vee}{\cdot} ([b] \stackrel{\vee}{-} [A] \stackrel{\vee}{\cdot} \tilde{x})$ *and* $[\Delta] := (I \stackrel{\vee}{-} R \stackrel{\wedge}{\cdot} [A]) \stackrel{\wedge}{\cdot} (X \stackrel{\wedge}{-} \tilde{x})$ *there holds*

$$[\inf([Q]) \stackrel{\triangle}{+} \sup([\Delta]), \sup([Q]) \stackrel{\triangledown}{+} \inf([\Delta])]$$
$$(3.11) \qquad \subseteq \left[\inf\left(\sum([A], [b])\right), \sup\left(\sum([A], [b])\right) \right].$$

Here, $\stackrel{\triangle}{+}$ and $\stackrel{\triangledown}{+}$ denote the floating-point addition rounded upwards and downwards, respectively. Components of $[Q]$ may be empty, in which case no lower bound on the sensitivity for this component is given. Computing $[Q]$ is practically free of cost compared to the costs for solving the linear system. $[\Delta]$ has already been computed in (3.10).

In the case of systems of nonlinear equations there is the problem of computing a sharp inner estimation of $\tilde{x} - R \cdot f(\tilde{x})$. This can be done using operations

* for $* \in \{+, -, \cdot, /\}$ or other methods, e.g. [8]. For special nonlinear methods, such as eigenproblems, methods similar to the ones described above can be used.

4. Practical results

In the following we display results for linear systems with well-conditioned and ill-conditioned matrices and varying diameter of matrix and right-hand side. Matrices used are Hilbert matrices, which, in order to be exactly representable on a digital computer, are multiplied by the least common multiple of all denominators:

$$\text{Hilbert} \quad (H_n)_{ij} := (\text{lcm}(1, \ldots, 2n-1))/(i+j-1),$$

and Pascal and Zielke matrices, defined by

$$\text{Pascal} \quad (P_n)_{ij} := \binom{i+j}{i},$$

$$\text{Zielke} \quad (Z_n)_{ij} := \frac{\binom{n+i-1}{i-1} \cdot n \cdot \binom{n-1}{j-1}}{i+j-1}.$$

Zielke matrices in particular are extremely ill-conditioned and have the interesting property that a checkerboard-like distribution of $+/-$ signs over Z_n generates the inverse matrix of Z_n. In the following, randomly generated matrices have components uniformly distributed in $[0, 1]$.

In the following tables we list the "overestimation" δ calculated by (3.11) in percent. In the notation of Theorem 4, and the abbreviation $Z := [\inf([Q]) + \sup([\Delta]), \sup([Q]) + \inf([\Delta])]$, we define

$$(4.1) \quad \delta := \begin{cases} 100 & \text{if } d(Z_i) = 0 \text{ for some } 1 \le i \le n, \\ \max\limits_{1 \le i \le n} \left(1 - \frac{d(Z_i)}{d(Y_i)}\right) \cdot 100 & \text{otherwise.} \end{cases}$$

δ gives the percentage of the inner estimation Z with respect to the outer inclusion Y. Inclusion Y is calculated using Theorem 7 and the inclusion methods described in [18, 20].

It should be mentioned that δ is an upper bound on the true "overestimation" of a computed inclusion. If δ is poor, i.e., near or equal 100%, the true overestimation might still be reasonable.

The first example involves Zielke matrices with tolerances

$$(4.2) \quad M_i := Z_i \cdot (1 \pm 10^{-1.4i}) \quad \text{for } i = 5, \ldots, 10.$$

The right-hand side b is randomly chosen (denoted by rand) with proper dimension:

$$b_i := \text{rand} \cdot (1 \pm 10^{-8+2i}) \quad \text{for } i = 1, \ldots, 5.$$

The computations are done on an IBM 3090 using double precision equivalent to 14 hex or 16 to 17 decimal digits in the mantissa.

TABLE 4.1

Overestimation for Zielke matrices (4.2)

$\delta[\%]$	$b1$	$b2$	$b3$	$b4$	$b5$
M_5	3.0	3.0	3.1	3.2	3.2
M_6	3.3	3.3	3.4	3.5	3.4
M_7	3.8	3.8	3.9	3.9	3.8
M_8	4.4	4.4	4.4	4.5	4.4
M_9	5.2	5.2	5.3	5.3	5.4
M_{10}	6.4	6.4	6.5	6.6	6.5

Obviously, there is only a small dependency on the diameters of the right-hand side. The difference between inner and outer inclusion is less than 7%. This is an excellent value, since even a value of 90% suffices for the purpose of estimating the magnitude of the sensitivity.

For different diameters of the matrix, things change. In Table 2 we have treated in the first row the matrices

$$(4.3) \qquad\qquad Z_5 \cdot (1 \pm 10^{-10+i}) \quad \text{for } i = 0(1)4,$$

and in the second row the matrices

$$(4.4) \qquad\qquad Z_{10} \cdot (1 \pm 10^{-15+i}) \quad \text{for } i = 0(1)4,$$

using Theorem 7 and randomly chosen right-hand sides of relative diameter 10^{-2}. The results for δ are shown in Table 4.2. An entry -1.0 indicates that no inclusion Y using Theorem 7 and the methods described in [18, 20] could be computed. Finer methods proved that in these cases there was indeed a singular matrix within the tolerance matrix. An entry 0.0 indicates that δ was less than 0.05%.

Table 4.2 reveals an almost linear dependence of the overestimation δ on the diameter of the matrix of the linear system. δ gets big when the diameter of the matrix gets so big that nearly singular matrices are enclosed. We omit corresponding tables for Hilbert and Pascal matrices, because they look very similar, in fact almost identical.

TABLE 4.2

Overestimation for Zielke matrices (4.3) and (4.4)

$\delta[\%]$	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$Z_5 \cdot (1 \pm 10^{-10+i})$	0.0	0.0	0.3	3.0	30.4
$Z_{10} \cdot (1 \pm 10^{-15+i})$	0.7	6.5	62.8	-1.0	-1.0

For random right-hand sides of proper dimension and of constant relative

diameter 10^{-2}, and for a linear system with matrices

$$(4.5) \qquad\qquad R_n \cdot (1 \pm 10^{-i}) \quad \text{for } i = 5(-1)2,$$

with random matrices R_n of dimension $10(10)50$, we obtain the percentage values of δ shown in Table 4.3. The entry 100.0 indicates that an inclusion Y was computed according to Theorem 7 but at least one component of the inner estimation Z was empty.

TABLE 4.3
Overestimation for random matrices (4.5)

$\delta[\%]$	$i = 5$	$i = 4$	$i = 3$	$i = 2$
$R_{10} \cdot (1 \pm 10^{-i})$	0.1	2.8	12.2	−1.0
$R_{20} \cdot (1 \pm 10^{-i})$	0.2	1.0	62.8	−1.0
$R_{30} \cdot (1 \pm 10^{-i})$	2.0	6.6	100.0	−1.0
$R_{40} \cdot (1 \pm 10^{-i})$	3.7	13.6	100.0	−1.0
$R_{50} \cdot (1 \pm 10^{-i})$	2.4	28.7	100.0	−1.0

Tables 4.1–4.3 indicate that there is a small area where the matrix of the linear system does not contain singular matrices, but nearly singular matrices where the overestimation δ is poor.

Below we display estimations of the sensitivity of a linear system with respect to perturbations of the input data. Let a linear system $A \cdot x = b$ be given with $A \in MT$, $b \in VT$ for $T \in \{\mathbf{R}, \mathbf{C}\}$, and A invertible. The diameter of $\sum([A], [b])$ for $[A] = A \cdot (1 \pm \varepsilon)$ and $[b] = b \cdot (1 \pm \varepsilon)$ for small $\varepsilon > 0$ gives a componentwise measure of the sensitivity of $A^{-1} \cdot b$ with respect to small changes in A and b. In the following tables we display the maximum of

$$(4.6) \qquad\qquad F := r(Y)/\varepsilon \geq r\left(\sum([A], [b])\right)/\varepsilon,$$

using the upper bound Y of $\sum([A], [b])$, cf. Theorem 8. The quality of Y is estimated by δ from (4.1). The quantity F bounds the maximum factor by which an ε-perturbation of A and b is amplified in terms of variations in the solution. An algorithm based on Theorem 7 gives F for every component independently. In the following we solve the linear system $Ax = I$, i.e., compute a full inverse of A. The algorithm yields n^2 amplification factors F, the largest of which, namely f, is displayed. Next to f we display the condition number c defined by $c := s_1/s_n$ where $s_1 \geq \cdots \geq s_n > 0$ are the singular values of A.

In the following table we choose $\varepsilon := 10^{-15}$, one order of magnitude above the relative rounding error unit. Matrices used are Hilbert matrices H_n, Pascal matrices P_n, and Zielke matrices Z_n of different dimensions.

TABLE 4.4

Sensitivity of linear systems vs. traditional condition number

	$f(H_n)$	$c(H_n)$	$f(P_n)$	$c(P_n)$	$f(Z_n)$	$c(Z_n)$
$n = 5$	2.0e5	4.8e5	1.5e4	6.3e4	1.9e5	7.9e5
6	5.3e6	1.5e7	1.4e5	9.3e5	5.3e6	3.7e7
7	1.5e8	4.8e8	1.2e6	1.4e7	1.5e8	1.8e9
8	4.4e9	1.5e10	1.1e7	2.0e8	4.4e9	9.4e10
9	1.3e11	4.9e11	9.5e7	3.0e9	2.6e11	5.0e12
10	4.0e12	1.6e13	8.4e8	4.5e10	8.0e12	2.7e14

The uncertainty δ of f in all cases of Table 4.4 is less than 1%, i.e., all estimations of f are correct to two figures and, by the general principle underlying the methods, are guaranteed to be correct.

In the examples above, the condition number $c(\cdot)$ is an overestimation of the true sensitivity $f(\cdot)$. Of course, the condition number is not usually computed in this way because a singular value decomposition is too expensive for the sole purpose of giving an estimation of the condition of the matrix.

The next example, Table 4.5, shows a significant underestimation of the sensitivity $r(\sum([A], [b]))/\varepsilon$ by the condition number c, owing to the equilibration effect of norms. Let R_n be a random matrix with n rows and columns, ε is again 10^{-15}.

TABLE 4.5

Sensitivity analysis of linear systems with random matrix
vs. traditional condition number

n	$f(R_n)$	$c(R_n)$
10	5.4e3	5.3e1
30	1.7e5	2.0e2

Table 4.5 shows that the traditional definition of the quotient of largest and smallest singular value underestimates the true sensitivity of the linear system by 2, resp. 3, orders of magnitude. Moreover, the traditional method only considers the matrix of the linear system, not the right-hand side. The true sensitivity depends significantly on the right-hand side.

The reason why the componentwise estimations f defined in (4.6) are much larger than c is that some individual components of the inverse are much more responsive to small perturbations in the input data than others. In the second example, with the matrix R_{30}, the componentwise estimation of f (which is provided by Theorem 7) is

$$\leq \text{1e3} \quad \text{for 95\% of all components,}$$
$$\leq \text{1e4} \quad \text{for 98\% of all components,}$$

and only in two cases greater than 1e5.

In the last example with random matrices, the greater sensitivity of an individual component of the inverse occurred exactly for those components being of significantly smaller absolute value compared to all others. This need not be the case, as has been demonstrated by the examples for Hilbert, Pascal, and Zielke matrices. The new methods allow a componentwise sensitivity analysis.

5. Conclusion

Methods have been described for the computation of inner and outer bounds of the solution set of linear and nonlinear systems whose data are subject to tolerances. The bounds computed are sharp and guaranteed to be correct. It turns out that in most cases the differences of inner and outer bounds is negligible. This difference becomes larger only in extreme cases where (in the examples) a singular matrix is very close to the set of matrices of a linear system.

A criticism of inclusion algorithms for data subject to tolerances was that correct bounds for the solution set are computed and all experiences showed that those bounds are sharp, but the degree of sharpness could not be estimated (see [6]). The present theorems and practical results fill this gap.

The inner estimations come virtually free of cost, together with outer estimations. They allow a sensitivity analysis of problems, with the additional advantage that instead of a single number estimating the condition of the problem in question, a whole sensitivity matrix can be computed, estimating variations of individual components of the solution for perturbations in the input data.

The estimation of the sensitivity of the linear system is guaranteed to be correct and reflects the true sensitivity of the linear system, i.e., of the matrix in combination with the particular right-hand side. It has been shown by means of examples that traditional condition numbers do not necessarily reflect the true sensitivity of individual components of a solution.

The methods described can be implemented very effectively on digital computers. No special computer arithmetic is necessary; a state of the art arithmetic, e.g., described in the IEEE 754 binary floating-point standard, suffices. Especially all kinds of computer arithmetic allowing the representation of sets on computers are suitable; in our implementation we used a rectangular real or complex arithmetic. A computer implementation for nonlinear systems is somewhat more involved and will be described later.

Bibliography

1. ACRITH, High-Accuracy Arithmetic Subroutine Library, Program Description and User's Guide, IBM Publications, Document Number SC 33-6164-3, 1986.
2. G. Alefeld and J. Herzberger, *Introduction to interval computations*, Academic Press, New York, 1983.
3. H. Bauch, K.-U. Jahn, D. Oelschlägel, H. Süsse, and V. Wiebigke, *Intervallmathematik, Theorie und Anwendungen*, Mathematisch-Naturwissenschaftliche Bibliothek, Band 72, Teubner, Leipzig, 1987.
4. W. Baur and V. Strassen, *The complexity of partial derivatives*, Theoret. Comput. Sci. **22** (1983), 317–330.

5. IEEE 754 Standard for Floating-Point Arithmetic, 1986.

6. W. Kahan and E. LeBlanc, *Anomalies in the IBM ACRITH package*, Proc. 7th Sympos. on Computer Arithmetic (Kai Hwang, ed.), Urbana, Illinois, 1985.

7. R. Krawczyk, *Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlershranken*, Computing **4** (1969), 187–201.

8. R. Krawczyk and A. Neumaier, *Interval slopes for rational functions and associated centered forms*, SIAM J. Numer. Anal. **22** (1985), 604–616.

9. U. Kulisch, *Grundlagen des numerischen Rechnens* (Reihe Informatik, 19), Bibliographisches Institut, Mannheim, Wien and Zürich, 1976.

10. U. Kulisch and W. L. Miranker, *Computer arithmetic in theory and practice*, Academic Press, New York, 1981.

11. ____, (eds.), *A new approach to scientific computation*, Academic Press, New York, 1981.

12. R. E. Moore, *Interval analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1966.

13. ____, *A test for existence of solutions to nonlinear systems*, SIAM J. Numer. Anal. **4** (1977), 611–615.

14. ____, *Methods and applications of interval analysis*, SIAM, Philadelphia, 1979.

15. A. Neumaier, *Overestimation in linear interval equations*, SIAM J. Numer. Anal. **24** (1987), 207–214.

16. ____, *Rigorous sensitivity analysis for parameter-dependent systems of equations* (to appear).

17. L. B. Rall, *Automatic differentiation: Techniques and applications*, Lecture Notes in Computer Science, no. 120, Springer-Verlag, Berlin, Heidelberg, and New York, 1981.

18. S. M. Rump, *Solving algebraic problems with high accuracy*, in A New Approach to Scientific Computation (U. Kulisch and W. L. Miranker, eds.), Academic Press, New York, 1981, pp. 51–120.

19. ____, *Solving non-linear systems with least significant bit accuracy*, Computing **29** (1982), 183–200.

20. ____, *New results on verified inclusions*, in Accurate Scientific Computations (W. L. Miranker and R. Toupin, eds.), Lecture Notes in Computer Science, no. 235, Springer-Verlag, 1986, pp. 31–69.

21. B. Speelpennig, *Compiling fast partial derivatives of functions given by algorithms*, Ph. D. thesis, University of Illinois, Urbana, Illinois, 1980.

22. P. Wongwises, *Experimentelle Untersuchungen zur numerischen Auflösung von linearen Gleichungs-systemen mit Fehlererfassung*, Interner Bericht 75/1, Institut für Praktische Mathematik, Universität Karlsruhe, 1975.

INFORMATIK III, TECHNISCHE UNIVERSITÄT, EISSENDORFER STRASSE 38, 2100 HAMBURG 90, WEST GERMANY