Verification Methods for Dense and Sparse Systems of Equations *

S.M. Rump, Hamburg

In this paper we describe verification methods for dense and large sparse systems of linear and nonlinear equations. Most of the methods described have been developed by the author. Other methods are mentioned, but it is not intended to give an overview over existing methods.

Many of the results are published in similar form in research papers or books. In this monograph we want to give a concise and compact treatment of some fundamental concepts of the subject. Moreover, many new results are included not being published elsewhere. Among them are the following.

A new test for regularity of an interval matrix is given. It is shown that it is significantly better for classes of matrices.

Inclusion theorems are formulated for continuous functions not necessarily being differentiable. *Some* extension of a nonlinear function w.r.t. a point $\widetilde{x}$ is used which may be a slope, Jacobian or other.

More narrow inclusions and a wider range of applicability (significantly wider input tolerances) are achieved by (i) using slopes rather than Jacobians, (ii) improvement of slopes for transcendental functions, (iii) a two-step approach proving existence in a small and uniqueness in a large interval thus allowing for proving uniqueness in much wider domains and significantly improving the speed, (iv) use of an Einzelschrittverfahren, (v) computing an inclusion of the difference w.r.t. an approximate solution.

Methods for problems with parameter dependent input intervals are given yielding inner and outer inclusions.

An improvement of the quality of inner inclusions is described.

Methods for parametrized sparse nonlinear systems are given for expansion matrix being (i) M-matrix, (ii) symmetric positive definite, (iii) symmetric, (iv) general.

A fast interval library having been developed at the author's institute is presented being significantly faster compared to existing libraries.

*in J. Herzberger, editor, Topics in Validated Computations — Studies in Computational Mathematics, pages 63-136, Elsevier, Amsterdam, 1994

A common principle of all presented algorithms is the combination of floating point and interval algorithms. Using this synergism yields powerful algorithms with automatic result verification.

Contents

0. Introduction

Verification methods, inclusion or self-validating methods deliver bounds for the solution of a problem which are verified to be correct. Such verification includes all conversion, rounding or other procedural errors. This is to be sharply distinguished from any heuristic such as, for example, computing in different precisions and using coinciding figures. Such techniques may fail. Consider the following example [77].

$$\begin{aligned} f \;&:=\; 333.75\,b^6 + a^2(11\,a^2b^2 - b^6 - 121\,b^4 - 2) + 5.5\,b^8 + a/(2\,b) \\ &\quad \text{with } a = 77617.0 \text{ and } b = 33096.0 \end{aligned}$$

Even if the powers are executed by successive multiplications in order to avoid transcendental function calls, then on an IBM S/370 computing in single ($\sim$ 7 decimals), double ($\sim$ 16 decimals) and extended ($\sim$ 33 decimals) precision we obtain the following results

single precision	$f = \underline{1.17260}361\ldots$
double precision	$f = \underline{1.17260394005317}847\ldots$
extended precision	$f = \underline{1.17260394005317}863185\ldots$

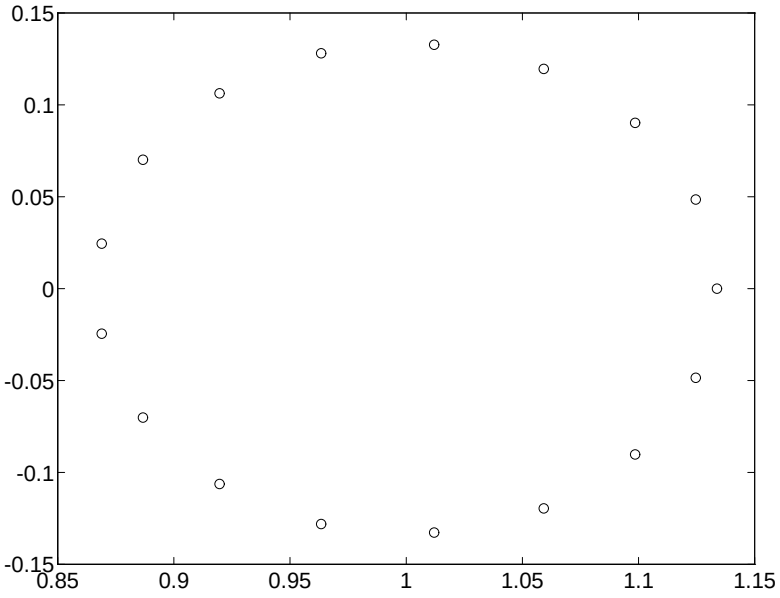
where the coinciding figures are underlined. This might lead to the conclusion that $\widetilde{f} = + 1.1726$ seems at least to be a reasonably good approximation for f. The true value is

$$f = -\,0.827396\ldots$$

Floating point algorithms usually deliver good approximations for the solution of a given problem. However, those results are not *verified to be correct* and may be afflicted with a smaller or larger error. As an example, consider the determination of the eigenvalues of the following matrix A.

$$A = \begin{pmatrix} 1 & & & & -1 & -1 \\ 1 & 1 & & & -1 & 0 \\ & \ddots & \ddots & & \vdots & \vdots \\ & & 1 & 1 & -1 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 2 \end{pmatrix} \in \mathrm{M}_{nn}(\mathbb{R}) \tag{1}$$

Applying $[V, D] = \mathrm{eig}(A)$ from MATLAB [57] which implements EISPACK algorithms [87] delivers the matrix of eigenvectors V and the diagonal matrix D of eigenvalues. The eigenvalues for $n = 17$ are plotted in the complex plane.



No warning is displayed or could be found in the documentation. Checking the residual yields

$$\text{norm}\,(A * V - V * D) = 2.6 \cdot 10^{-14}.$$

Therefore the user might expect to have results of reasonable accuracy. However, checking the rank of V would give $\text{rank}\,(V) = 1$, and indeed

$$A = X^{-1} \cdot \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix} \cdot X \text{ with } X = \begin{pmatrix} 1 & & & 1 \\ & 1 & & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix},\ X^{-1} = \begin{pmatrix} 1 & & & -1 \\ & 1 & & -1 \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}.$$

That means, A from (1) has exactly one eigenvalue equal to 1.0 of multiplicity n. Such errors are rare, but to cite Kahan [44]:

> Kahan: "Significant discrepancies [between the computed and the true result] are very rare, too rare to worry about all the time, yet not rare enough to ignore."

Verification algorithms aim to fill this gap by always producing correct results. One basic tool of verification algorithms is interval analysis. The most basic, *principle property* of interval analysis is the *isotonicity*. This means that for interval quantities $[A]$, $[B]$ in proper interval spaces

$$\forall\ a \in [A] \quad \forall\ b \in [B]: \quad a * b \in [A] * [B] \tag{2}$$

for any suitable operation $*$ (cf. [61], [8], [66]). This leads to the remarkable property that the range of a function f over a box can be rigorously estimated by replacing real operations by their corresponding interval operations during the evaluation of f. This is possible without any further knowledge of f, such as Lipschitz conditions. On the other hand, one also quickly observes that overestimation may occur due to data dependencies.

The main goal of verification algorithms is to make use of this remarkable range estimation and to avoid overestimation where possible. In general, this implies use of floating point arithmetic as much as possible and restriction of interval operations to those specific parts where they are really necessary. This is very much in the spirit of Wilkinson [90], who wrote in 1971:

> Wilkinson: "In general it is the best in algebraic computations to leave the use of interval arithmetic as late as possible so that it effectively becomes an a posteriori weapon." (3)

0.1. Notation

In this paper all operations are power set operations except when explicitly stated otherwise. For example (ρ denotes the spectral radius of a matrix), a condition like

$$\begin{array}{l} Z \in \mathbb{IPIR}^n,\ \boldsymbol{C} \in \mathbb{IPM}_{nn}(\mathbb{IR}),\ X \in \mathbb{IPIR}^n \quad \text{closed and bounded with} \\ \qquad Z + \boldsymbol{C} \cdot X \subseteq \operatorname{int}(X) \quad \Rightarrow \quad \forall\, C \in \boldsymbol{C} : \rho(C) < 1 \end{array} \tag{4}$$

is to read

$$\{\, z + C \cdot x \mid z \in Z,\ C \in \boldsymbol{C},\ x \in X \,\} \subseteq \operatorname{int}(X).$$

Theoretical results are mostly formulated using power sets and power set operations. In a *practical* implementation we mostly use $[Z] \in \mathbb{IIR}^n$, $[C] \in \mathbb{IIM}_{nn}(\mathbb{IR})$, $[X] \in \mathbb{IIR}^n$. Then (4) can be used to verify $\rho(C) < 1$ for all $C \in [C]$ *on the computer* by using the fundamental principle of interval analysis, the isotonicity (0.2). Denote interval operations by $\diamond\!\!\!\!\!\!\diamond$, $* \in \{+, -, \cdot, /\}$. Then

$$\begin{array}{l} [Z] \diamond\!\!\!\!\!\!\diamond [C] \diamond\!\!\!\!\!\!\diamond [X] \subseteq \operatorname{int}(X) \quad \Rightarrow \\ [Z] + [C] \cdot [X] \subseteq [Z] \diamond\!\!\!\!\!\!\diamond [C] \diamond\!\!\!\!\!\!\diamond [X] \subseteq \operatorname{int}(X), \end{array}$$

where the operations in the first part of the second line are power set operations as in (4) using the canonical embeddings $\mathbb{IIR}^n \subseteq \mathbb{IPIR}^n$, $\mathbb{IIM}_{nn}(\mathbb{IR}) \subseteq \mathbb{IPM}_{nn}(\mathbb{IR})$. For $f : \mathbb{IR}^n \to \mathbb{IR}^n$, $X \in \mathbb{IPIR}^n$ we define $f(X) := \{\, f(x) \mid x \in X \,\} \in \mathbb{IPIR}^n$. For more details see standard text books on interval analysis, among them [7], [8], [11], [66], [31], [61], [70]. We also use interval rounding $\diamond$ from the power set $\mathbb{IPS}$ over S to the set of intervals $\mathbb{IIS}$ over S for all suitable S:

$$X \in \mathbb{IPS} \Rightarrow \diamond(X) \in \mathbb{IIS} \quad \text{with} \quad \diamond\,(X) := \bigcap \{\, [Y] \in \mathbb{IIS} \mid X \subseteq [Y] \,\}.$$

$\diamond(X)$ is also called the interval hull of X.

With one exception, in all of the presented theorems given in this paper it is always possible to replace the power set operations by the corresponding interval operations without sacrificing the validity of the assertions. The exception is so-called *inner inclusions*. They allow a sensitivity analysis of parametrized nonlinear systems w.r.t. finite changes in the parameters. This exception is stated explicitly. We prefer using power set operations, because they simplify the proofs, and allow use of the usual symbols for arithmetic operations.

Interval quantities are written in brackets, for example an interval vector $[X] \in \mathbb{IIR}^n$ with components $[X]_i \in \mathbb{IIR}$ or simply X_i. The lower and upper bounds are denoted by $\inf([X]) \in \mathbb{R}^n$ and $\sup([X]) \in \mathbb{R}^n$, where sometimes we also use the notation $[X] = [\underline{X}, \overline{X}]$ with $\underline{X}, \overline{X} \in \mathbb{R}^n$. Absolute value $|\cdot|$, comparison $\leq$, midpoint $\mathrm{mid}([X])$, width $w([X])$ and so forth are always to be understood componentwise. $\mathrm{int}([X])$ denotes the topological interior, I is the identity matrix of proper dimension and

$$[X], [Y] \in \mathbb{IIR}^n : [X] \subsetneqq [Y] \;\Leftrightarrow\; [X] \subseteq [Y] \text{ and } [X]_i \neq [Y]_i \quad \text{for} \quad i = 1 \ldots n.$$

Most of the following results are given for intervals over real numbers. We want to stress that *all results remain true over the domain of complex numbers.*

1. Basic results

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a continuous mapping. Consider the function $g : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$g(x) := x - R \cdot f(x) \tag{5}$$

for $x \in D$ and some fixed matrix $R \in \mathrm{M}_{nn}(\mathbb{R})$. Then g is also continuous, and for convex and compact $\emptyset \neq X \subseteq D$

$$g(X) \subseteq X \quad \text{implies the existence of some} \quad \widehat{x} \in X : g(\widehat{x}) = \widehat{x}$$

by Brouwer's Fixed Point Theorem [33]. If, moreover,

$$R \text{ is regular, then } f(\widehat{x}) = 0.$$

That means, if we can find a suitable set $X \subseteq D$ and could prove that g maps X into itself and that R is regular, then X is verified to contain a zero of f. Therefore, in the following we will first concentrate

– on verification procedures for $g(X) \subseteq X$ and
– on verification of the regularity of R.

Many of the following considerations hold for general closed and bounded and possibly convex sets. Also, many proofs become simpler when using power set operations, whereas

the specialization to interval operations is almost always straightforward by replacing the power set operations by the corresponding interval operations and using the basic principle of isotonicity (2). Therefore, we give the results in a more general setting in order not to preclude specific interval representations such as, for example, circular arithmetic.

$g(X) \subseteq X$ cannot be verified by testing

$$X - R \cdot f(X) = \{\, x_1 - R \cdot f(x_2) \mid x_1, x_2 \in X \,\} \subseteq X$$

unless $R \cdot f(X) \equiv 0$. Therefore we need some expansion of f. For this chapter, we make the following, general assumption:

$$\begin{array}{l} \text{Let } f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n \text{ be continuous, } s_f : D \times D \to \mathrm{M}_{nn}(\mathbb{R}) \text{ such that} \\ x \in D,\ \widetilde{x} \in D \Rightarrow f(x) = f(\widetilde{x}) + s_f(\widetilde{x}, x) \cdot (x - \widetilde{x}). \end{array} \tag{6}$$

Such expansion functions s_f can be computed efficiently by means of slope functions or, if f is differentiable, by automatic differentiation techniques. In Chapter 3 we will discuss such techniques in detail; for the moment we assume that such an s_f satisfying (6) is given.

1.1. Some basic lemmata

It turns out to be superior not to include a zero $\widehat{x}$ of a function itself but the difference w.r.t. some approximate solution $\widetilde{x}$. Note that here and in the following there are no preassumptions on the quality of $\widetilde{x}$. Therefore, we immediately go for inclusions of $\widehat{x} - \widetilde{x}$.

For given nonempty, compact and convex $X \subseteq D$ define $Y := X - \widetilde{x} \subseteq \mathbb{R}^n$. We do not assume $\widetilde{x} \in X$. Then with g from (5)

$$\begin{array}{rl} x \in X \Rightarrow g(x) - \widetilde{x} & = x - \widetilde{x} - R \cdot f(x) \\ & = -R \cdot f(\widetilde{x}) + \{I - R \cdot s_f(\widetilde{x}, x)\} \cdot (x - \widetilde{x}) \\ & \in -R \cdot f(\widetilde{x}) + \{I - R \cdot s_f(\widetilde{x}, \widetilde{x} + Y)\} \cdot Y. \end{array}$$

With the abbreviations

$$z := -R \cdot f(\widetilde{x}) \in \mathbb{R}^n \quad \text{and} \quad \boldsymbol{C} := I - R \cdot s_f(\widetilde{x}, \widetilde{x} + Y) \in \mathbb{IP}\mathrm{M}_{nn}(\mathbb{R}) \tag{7}$$

this means

$$z + \boldsymbol{C} \cdot Y \subseteq Y \quad \Rightarrow \quad g(X) - \widetilde{x} \subseteq Y \quad \Rightarrow \quad g(X) \subseteq X. \tag{8}$$

In other words, $z + \boldsymbol{C} \cdot Y \subseteq Y$ is a sufficient condition for $g(X) \subseteq X$, and our first problem is solved.

For the second problem, the verification of the regularity of R, we give a characterization of the convergence of the iteration matrices $\boldsymbol{C}$ in (8). The following lemma has been given

in [74].

Lemma 1.1. Let $Z \in \mathbb{IPIR}^n$, $\boldsymbol{C} \in \mathbb{IPM}_{nn}(\mathbb{IR})$ and let some closed and bounded $\emptyset \neq X \in \mathbb{IPIR}^n$ be given. Then

$$Z + \boldsymbol{C} \cdot X \subseteq \mathrm{int}(X) \tag{9}$$

implies for every $C \in \boldsymbol{C} : \rho(C) < 1$.

Proof. Let $z \in Z$, $C \in \boldsymbol{C}$ fixed but arbitrary. Then (9) implies $z + C \cdot X \subseteq \mathrm{int}(X)$. Abbreviating $Y := (X + iX) - (X + iX)$ implies

$$\begin{aligned} C \cdot Y &= \{\, C \cdot (x_1 + i\, x_2) - C \cdot (x_3 + i\, x_4) \mid x_\nu \in X \text{ for } 1 \leq \nu \leq 4 \,\} \\ &= z + C \cdot X + i \cdot (z + C \cdot X) - (z + C \cdot X) - i \cdot (z + C \cdot X) \\ &\subseteq \mathrm{int}(Y). \end{aligned} \tag{10}$$

Suppose $C \neq (0)$ and let $\lambda \in \mathbb{C}$, $0 \neq x \in \mathbb{C}^n$ be an eigenvalue/eigenvector pair of C. Define

$$\Gamma \in \mathbb{IPC} \quad \text{by} \quad \Gamma := \{\, \gamma \in \mathbb{C} \mid \gamma \cdot x \in Y \,\}. \tag{11}$$

Then by the definition of Y we have $0 \in Y$ and therefore $\Gamma \neq \emptyset$. Moreover, Y is closed and bounded, hence Γ has this property and there is some $\gamma^* \in \Gamma$ with

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

(11) implies $\gamma^* x \in Y$ and (10) yields $C \cdot (\gamma^* x) = (\gamma^* \lambda) \cdot x \in \mathrm{int}(Y)$, and by the definition of γ^* and Γ, $|\gamma^* \lambda| < |\gamma^*|$. Therefore $|\lambda| < 1$, and since $C \in \boldsymbol{C}$ was chosen arbitrarily the lemma is proved. ■

In a practical implementation we use interval quantities and interval operations. Interestingly enough, if the set X in Lemma 1.1 is replaced by an interval vector, then we can sharpen the result under weaker assumptions. We start with the following lemma which can be found in [76]. The presented proof has been given by Heindl [32].

Lemma 1.2. Let $Z \in \mathbb{IPIR}^n$, $\boldsymbol{C} \in \mathbb{IPM}_{nn}(\mathbb{IR})$ and $[X] \in \mathbb{IIR}^n$ be given. Then

$$\diamond(Z + \boldsymbol{C} \cdot [X]) \subsetneqq [X] \tag{12}$$

implies $\rho(C) < 1$ for every $C \in \boldsymbol{C}$.

Proof. Let $z \in Z$, $C \in \boldsymbol{C}$ fixed but arbitrary and let $[Y] := \diamond(z + C \cdot [X]) \in \mathbb{IIR}^n$. Then $[Y] \subsetneqq [X]$, which means componentwise inclusion but inequality. Thus there is an

ε-perturbation z^* of z with $\diamondsuit(z^* + C \cdot [X]) \subseteq \mathrm{int}([X])$. Lemma 1.1 finishes the proof. ■

If inequality is only required for *some* components of (12), then $C \in \boldsymbol{C}$ must be irreducible to prove $\rho(C) < 1$. Next, we weaken the assumptions even more by introducing an Einzelschrittverfahren and a dependency of the iteration matrices $\boldsymbol{C}$ on the set $[X]$. Let $\boldsymbol{C} : \mathbb{R}^n \to \mathrm{M}_{nn}(\mathbb{R})$ be a mapping. Then for $[X] \in \mathbb{IR}^n$ the set $\boldsymbol{C}_{[X]} := \boldsymbol{C}([X]) = \{\, \boldsymbol{C}(x) \mid x \in [X] \,\}$ is well-defined. We define the following procedure to replace (12):

$$\begin{array}{ll} \text{for} & i = 1 \ldots n \text{ do} \\ & [U] := [Y_1, \ldots, Y_{i-1}, X_i, \ldots, X_n]; \\ & Y_i := \{\diamondsuit(Z + \boldsymbol{C}_{[U]} \cdot [U])\}_i \end{array} \tag{13}$$

Here the $Y_i \in \mathbb{IR}$ and $[U]$ is defined by its components Y_ν, X_μ, the ν-th, μ-th component of Y, X, respectively. Obviously, Y is computed using an Einzelschrittverfahren, where the iteration vector $[U]$ as well as the set of iteration matrices $\boldsymbol{C}_{[U]}$ changes in every step. With these preperations we can state the following lemma.

Lemma 1.3. Let $Z \in \mathbb{IPIR}^n$, $\boldsymbol{C} : \mathbb{R}^n \to \mathrm{M}_{nn}(\mathbb{R})$ be a mapping and, for $S \in \mathbb{IPIR}^n$, set $\boldsymbol{C}_S := \boldsymbol{C}(S) = \{\, \boldsymbol{C}(s) \mid s \in S \,\}$. Let $[X] \in \mathbb{IIR}^n$ and define $[Y] \in \mathbb{IIR}^n$ by (13). If

$$[Y] \subsetneqq [X], \tag{14}$$

then for every $C \in \boldsymbol{C}_{[Y]} \quad \rho(|C|) < 1$ holds.

Proof. In every step of (13), $[U]$ satisfies $[Y] \subseteq [U]$ because of (14). For fixed but arbitrary $z \in Z$, $C \in \boldsymbol{C}_{[Y]}$ we have

$$\forall\ 1 \le i \le n : [U] := [Y_1, \ldots, Y_{i-1}, X_i, \ldots, X_n] \Rightarrow \{\diamondsuit(z + C \cdot [U])\}_i \subseteq Y_i \subsetneqq X_i.$$

Thus

$$w(\{\diamondsuit(z + C \cdot [U])\}_i) = w(\{\diamondsuit(C \cdot [U])\}_i) = \{|C| \cdot w([U])\}_i \le \{w([Y])\}_i < \{w([X])\}_i$$

using some basic facts of interval analysis (see [8]). Thus abbreviating $x := w([X]) \in \mathbb{R}^n$, $y := w([Y]) \in \mathbb{R}^n$ we have $0 \le y < x$ and

$$\forall\ 1 \le i \le n : \{|C| \cdot (y_1, \ldots, y_{i-1}, x_i, \ldots, x_n)^T\}_i \le y_i < x_i. \tag{15}$$

For $0 < \varepsilon_i \in \mathbb{R}$, $1 \le i \le n$ define

$$y_i^* := \left\{|C| \cdot (y_1^*, \ldots, y_{i-1}^*, x_i, \ldots, x_n)^T\right\}_i + \varepsilon_i.$$

For sufficiently small ε_i we still have $y_i^* < x_i$. Hence for $1 \le i \le n$

$$\{|C| \cdot y^*\}_i \le \left\{|C| \cdot (y_1^*, \ldots, y_{i-1}^*, x_i, \ldots x_n)^T\right\}_i = y_i^* - \varepsilon_i < y_i^*$$

or $|C| \cdot y^* < y^*$. Thus a theorem by Collatz [17] implies $\rho(|C|) < 1$. ■

A lemma similar to the preceeding one has been given in [76]. The stronger assertion that even the absolute value of the iteration matrix is contracting is due to the symmetry of interval vectors in *every component* individually w.r.t. their midpoint. There are a number of other conditions proving the contractivity of a matrix, see for example [76].

In an application of Lemma 1.1 or 1.3, we need a set X to check the contractivity conditions. If we restrict our attention, for a moment, to the point case $Z = z \in \mathbb{R}^n$, $\boldsymbol{C} = C \in \mathrm{M}_{nn}(\mathbb{R})$, then $\rho(C) < 1$ implies invertibility of $I - C$, and this yields

$$\widehat{x} := (I - C)^{-1} z \;\Rightarrow\; z + C \cdot \widehat{x} = \widehat{x}.$$

This fixed point is unique, and a fortiori it follows that $\widehat{x} \in X$. In other words the set X must contain the fixed point of the mapping $z + C \cdot x$ (in our applications, this fixed point will be the zero of the function f). But rather than testing a number of sets X potentially containing a fixed point, we would like to *construct* such a set, for example by means of an iteration. We could define

$$X^{k+1} := z + C \cdot X^k \quad \text{for some } X^0. \tag{16}$$

However, $\widehat{x} \notin X^0$ immediately implies $\widehat{x} \notin X^1$ and therefore $\widehat{x} \notin X^k$ for all $k \in \mathbb{N}$. But even if $\widehat{x} \in X^0$, simple examples show that $X^{k+1} \subseteq X^k$ need not be satisfied for any $k \in \mathbb{N}$ using iteration (16). If $[X^0]$ is an interval vector and $\rho(|C|) < 1$, then truly $w([X]) \to 0$. However, we need $[X^{k+1}] \subseteq [X^k]$. Consider

$$z = 0, \quad C = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}, \quad [X^0] = \begin{pmatrix} [-10, & 10] \\ [-1, & 1] \end{pmatrix}.$$

Obviously $\rho(|C|) < 1$. Since $[X^0] = -[X^0]$, we only need to compute the sequence $x^0 := |[X^0]|$, $x^{k+1} := |C| \cdot x^k$ and to check $x^{k+1} < x^k$. We have

$$x^0 = \begin{pmatrix} 10 \\ 1 \end{pmatrix}, \quad x^1 = \begin{pmatrix} 0.5 \\ 5 \end{pmatrix}, \quad x^2 = \begin{pmatrix} 2.5 \\ 0.25 \end{pmatrix}, \quad x^3 = \begin{pmatrix} 0.125 \\ 1.25 \end{pmatrix}, \; \ldots$$

and obviously $x^{k+1} < x^k$ never occurs for any $k \in \mathbb{N}$.

In an application for systems of nonlinear equations, the problem is even more involved, since the iteration matrix C is no longer constant. If we assume C to be constant, we can give a complete overview of the convergence behaviour of the corresponding affine iteration for the case of power set operations as well as for the case of interval operations, if we use the so-called ε-inflation introduced in [74]. Below we give corresponding theorems

from [76] and [81], which are stated without proof. Interval iterations with ε-inflation for nonconstant iteration matrix have been investigated by Mayer [59].

Theorem 1.4. Let $S \in \{\mathbb{IR}, \mathbb{C}\}$ and $C \in M_{nn}(S)$ be an arbitrary matrix, $\emptyset \neq Z \in \mathbb{IPS}^n$ and $\emptyset \neq X^0 \in \mathbb{IPS}^n$ be bounded sets of vectors, and define

$$X^{k+1} := Z + C \cdot X^k + U_{\varepsilon_k}(0) \quad \text{for } k \in \mathbb{IN},$$

where $U_{\varepsilon_{k+1}} \subseteq U_{\varepsilon_k}$ and $U \subseteq U_{\varepsilon_k}$ are bounded for every $k \in \mathbb{IN}$, for some $\emptyset \neq U \in \mathbb{IPS}^n$ with $0 \in \mathrm{int}(U)$. Then the following two conditions are equivalent:

i) $\forall\, \emptyset \neq X^0 \in \mathbb{IPS}^n$ bounded $\quad \exists\, k \in \mathbb{IN}: \quad Z + C \cdot X^k \subseteq \mathrm{int}(X^k)$

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

The operations in the above theorem are the power set operations. A similar theorem does not necessarily hold for sets of matrices $\boldsymbol{C}$. The condition $Z + \boldsymbol{C} \cdot X \subseteq \mathrm{int}(X)$ immediately implies $\rho(C_1 \cdot C_2) < 1$ for all $C_1, C_2 \in \boldsymbol{C}$. However, there are examples of convex sets of matrices $\boldsymbol{C}$ with $\rho(C) < 1$ for all $C \in \boldsymbol{C}$, but $\exists\, C_1, C_2 \in \boldsymbol{C} : \rho(C_1 \cdot C_2) > 1$ (see [76]).

In case of interval operations we can prove a similar theorem for interval matrices. The proof in the real case is given in [76], a slightly more general form of the complex case is proved in [81]. In conjunction with Lemma 1.3 this offers a possibility to verify the regularity of a matrix or a set of matrices. We will need this later.

Theorem 1.5. Let $\mathrm{S} \in \{\mathbb{IR}, \mathbb{C}\}$ and let $[C] \in \mathbb{II}\mathrm{M}_{nn}(\mathrm{S})$, $[Z] \in \mathbb{II}\mathrm{S}^n$ and for $[X^0] \in \mathbb{II}\mathrm{S}^n$ define the iteration

$$[X^{k+1}] := [Z] \boxplus [C] \boxdot [X^k] \boxplus [E^k] \quad \text{for } k \in \mathbb{IN}$$

where $[E^k] \in \mathbb{II}\mathrm{S}^n$, $[E^k] \to [E] \in \mathbb{II}\mathrm{S}^n$ with $0 \in \mathrm{int}([E])$ and all operations are interval operations. Then the following two conditions are equivalent:

i) $\forall\, [X^0] \in \mathbb{II}\mathrm{S}^n \quad \exists\, k \in \mathbb{IN}: \quad [Z] \boxplus [C] \boxdot [X^k] \subseteq \mathrm{int}([X^k])$

ii) $\rho(|[C]|) < 1$.

The absolute value of a complex interval matrix is defined as the sum of absolute values of the real and imaginary part. In practical applications, it may be superior to go from intervals to an absolute value iteration. If $|Z|$ denotes the supremum of $|z|$ for $z \in Z \in \mathbb{IPIR}^n$ and $|\boldsymbol{C}|$ is defined similarly for $\boldsymbol{C} \in \mathbb{IPM}_{nn}(\mathbb{IR})$, then we can state a straightforward application of Lemma 1.3.

Lemma 1.6. Let $Z \in \mathbb{IPIR}^n$, $\boldsymbol{C} : \mathbb{IR}^n \to \mathrm{M}_{nn}(\mathbb{IR})$ be a mapping, and for $S \in \mathbb{IPIR}^n$ let $\boldsymbol{C}_S := \boldsymbol{C}(S) = \{\, \boldsymbol{C}(s) \mid s \in S \,\}$. Let $0 < x \in \mathbb{IR}^n$ and define $y \in \mathbb{IR}^n$ for $1 \leq i \leq n$ by

$$y_i := \{|Z| + |\boldsymbol{C}_{[U]}| \cdot u\}_i \text{ with } u := (y_1, \ldots, y_{i-1}, x_i, \ldots, x_n)^T \text{ and } [U] := [-u, +u].$$

If

$$y < x$$

then for every $C \in \boldsymbol{C}_{[-y,+y]}$ holds $\rho(|C|) < 1$.

In a practical implementation the verification of the assumptions of Lemma 1.6 needs only rounding upwards. Therefore no switching of the rounding mode is necessary. Moreover, only the upper bounds of the interval quantities need to be stored, which is advantageous, especially for large matrices. This gains much more than the factor 2 that might be expected. We will come to this in more detail in Chapter 7.

1.2. Regularity of interval matrices

The preceeding Theorem 1.5 and Lemma 1.6 can be used for the computational verification of the regularity of an interval matrix. An interval matrix $[A]$ is called *regular* if every $A \in [A]$ is regular, whereas an interval matrix is called *strongly regular* if $\mathrm{mid}([A])^{-1} \diamond [A]$ is regular.

Theorem 1.7. Let $[A] \in \mathbb{IM}_{nn}(\mathbb{R})$ be given, $R \in \mathrm{M}_{nn}(\mathbb{R})$ and $0 < x \in \mathbb{R}^n$. Let $C \in \mathrm{M}_{nn}(\mathbb{R})$ with $C := |I - R \cdot [A]|$ and define $x^{(k)}$, $y^{(k)} \in \mathbb{R}^n$ for $k \geq 0$ by

$$y_i^{(k)} := \{C \cdot u\}_i \quad \text{with} \quad u := (y_1^{(k)}, \ldots, y_{i-1}^{(k)}, x_i^{(k)}, \ldots, x_n^{(k)})^T \quad \text{and} \quad x^{(k+1)} := y^{(k)} + \varepsilon$$

for $1 \leq i \leq n$ and some $0 < \varepsilon \in \mathbb{R}^n$. If

$$y^{(k)} < x^{(k)}$$

for some $k \in \mathbb{N}$, then R and every matrix $A \in [A]$ are regular.

Proof. Lemma 1.6 implies $\rho(C) < 1$ and therefore for every $A \in [A]$ $\rho(I - R \cdot A) \leq \rho(|I - R \cdot A|) < 1$. Hence R and every $A \in [A]$ are regular. ■

In fact, Theorem 1.7 verifies strong regularity of $[A]$. Moreover, strong regularity has been the only known simple criterion for checking regularity of an interval matrix (see [66]). All known inclusion algorithms for systems of linear interval equations require strong regularity of the matrix. We will need to prove regularity of an interval matrix in Theorem 2.3 in order to demonstrate uniqueness of a zero of a nonlinear system within a certain domain.

Interestingly enough, at least for theoretical purposes, there is a new criterion to verify regularity of an interval matrix that is *not* necessarily strongly regular.

Theorem 1.8. Let $[A-\Delta, A+\Delta] \in \mathbb{IM}_{nn}(\mathbb{R})$, $0 \leq \Delta \in \mathrm{M}_{nn}(\mathbb{R})$ be an interval matrix. Denote the singular values of A by $\sigma_1(A) \geq \ldots \geq \sigma_n(A)$. Then

$$\sigma_n(A) > \sigma_1(\Delta) \tag{17}$$

implies regularity of $[A-\Delta, A+\Delta]$.

Proof. Every matrix $a \in [A-\Delta, A+\Delta]$ can be expressed in the form $a = A+\delta$ where $\delta \in \mathrm{M}_{nn}(\mathbb{R})$, $|\delta| \leq \Delta$. Let $0 \neq x \in \mathbb{R}^n$. Then

$$\sigma_1(\delta) = \rho\left(\begin{pmatrix} 0 & \delta^T \\ \delta & 0 \end{pmatrix}\right) \leq \rho\left(\begin{pmatrix} 0 & \Delta^T \\ \Delta & 0 \end{pmatrix}\right) = \sigma_1(\Delta)$$

and therefore

$$\|Ax\|_2 \geq \sigma_n(A) \cdot \|x\|_2 > \sigma_1(\Delta) \cdot \|x\|_2 \geq \sigma_1(\delta) \cdot \|x\|_2 \geq \|\delta \cdot x\|_2.$$

Hence $Ax \neq \delta x$ for all $x \neq 0$ implying $a \cdot x \neq 0$ and the regularity of all $a \in [A-\Delta, A+\Delta]$. ■

To check regularity, another criterion *equivalent* to strong regularity can be used, namely

$$\begin{aligned} \rho(|A^{-1}| \cdot \Delta) < 1 \quad &\Leftrightarrow [A-\Delta, A+\Delta] \text{ is strongly regular} \\ &\Rightarrow \text{ all } a \in [A-\Delta, A+\Delta] \text{ are regular.} \end{aligned} \tag{18}$$

This is exactly what Theorem 1.5 checks, by constructing a proper norm. Comparing the two sufficient criterions for regularity of an interval matrix, Theorem 1.8 and (18), there are examples for which either one is satisfied but the other one does not hold. For a better comparison of the two criteria we define the radius of singularity [80], [20].

Definition 1.9. Let $A \in \mathrm{M}_{nn}(\mathbb{R})$, $0 \leq \Delta \in \mathrm{M}_{nn}(\mathbb{R})$. Then the radius of singularity of A w.r.t. perturbations weighted by Δ is defined by

$$\underline{\omega}(A, \Delta) := \inf_{r \in \mathbb{R}} \{[A - r \cdot \Delta, A + r \cdot \Delta] \text{ is singular}\}. \tag{19}$$

If no such r exists we define $\underline{\omega}(A, \Delta) := \infty$.

With the above consideration and Theorem 1.8 we get

Corollary 1.10. For $A \in \mathrm{M}_{nn}(\mathbb{R})$ regular and $0 \leq \Delta \in \mathrm{M}_{nn}(\mathbb{R})$,

$$\underline{\omega}(A, \Delta) \geq \{\rho(|A^{-1}| \cdot \Delta)\}^{-1} \quad \text{and} \quad \underline{\omega}(A, \Delta) \geq \sigma_n(A)/\sigma_1(\Delta). \tag{20}$$

This corollary allows comparison of the two criteria. Consider

$$I) \quad A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \ \Delta = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{with } \underline{\omega}(A, \Delta) = 1.$$

Then $\{\rho(|A^{-1}| \cdot \Delta\}^{-1} = 0.5$ and $\sigma_n(A)/\sigma_1(\Delta) = \frac{1}{2}\sqrt{2} \approx 0.707$.

$$II) \quad A = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}, \ \Delta = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{with } \underline{\omega}(A, \Delta) = 1/3.$$

Then $\{\rho(|A^{-1}| \cdot \Delta\}^{-1} = 1/3 \approx 0.333$
and $\sigma_n(A) \,/\, \sigma_1(\Delta) = \frac{1}{4}(\sqrt{5} - 1) \approx 0.309$.

To see $\underline{\omega}(A, \Delta) = 1/3$ in the second example use

$$A + \frac{1}{3} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} .$$

For a practical application of the second criterion we need a lower bound on $\sigma_n(A)$ and an upper bound on $\sigma_1(\Delta)$. The first can be computed using the methods described in Chapter 6, Lemma 6.4, the latter by using $\sigma_1(\Delta) \leq \|\Delta\|_F$ or $\sigma_1(\Delta) \leq \{\|\Delta\|_1 \cdot \|\Delta\|_\infty\}^{1/2}$. We should stress that computing $\underline{\omega}(A, \Delta)$ is a nontrivial problem. In fact Rohn and Poljak [68] showed that it is NP-hard. A number of useful estimations on the relation between $\underline{\omega}(A, \Delta)$ and $\rho(|A^{-1}| \cdot \Delta)$ are given in [20].

For regular A, condition (17) can be replaced by

$$\|A^{-1}\|^{-1} > \|\Delta\|$$

and any norm satisfying $B \in \mathrm{M}_{nn}(\mathbb{R}) \Rightarrow \|B\| \leq \| \, |B| \, \|$. However, for absolute and consistent matrix norms such as $\|\cdot\|_1$, $\|\cdot\|_\infty$, $\|\cdot\|_F$, this cannot be better than $\rho(|A^{-1}|\cdot\Delta) < 1$ because in this case,

$$\rho(|A^{-1}| \cdot \Delta) \leq \| \, |A^{-1}| \cdot \Delta\| \leq \| \, |A^{-1}| \, \| \cdot \|\Delta\| = \|A^{-1}\| \cdot \|\Delta\| < 1.$$

The 2-norm is not absolute. Therefore, it may yield better results than checking $\rho(|A^{-1}| \cdot \Delta) < 1$. In example I) we have $\| \, |A^{-1}| \, \|_2 = 1$, whereas $\|A^{-1}\|_2 = \frac{1}{2}\sqrt{2}$. This also measures the best possible improvement by

$$\| \, |A^{-1}| \, \|_2 \leq \| \, |A^{-1}| \, \|_F = \|A^{-1}\|_F \leq \sqrt{n} \cdot \|A^{-1}\|_2.$$

There is a class of matrices where this upper bound is essentially achieved. Consider orthogonal A with absolute perturbations, i.e. $\Delta = (1)$. Then, for $x \in \mathbb{R}^n$ with $x = (1)$,

$$|A^{-1}| \cdot \Delta \cdot x = |A^T| \Delta x = \sum_{i,j} |A_{ij}| \cdot x, \quad \text{implying} \quad \rho(|A^{-1}| \cdot \Delta) = \sum_{i,j} |A_{ij}|.$$

On the other hand, $\sigma_n(A) = 1$ and $\sigma_1(\Delta) = n$, implying $\underline{\omega}(A, \Delta) \geq n^{-1}$ by Theorem 1.8. If A is an orthogonalized random matrix, then $|A_{ij}| \lesssim n^{-1/2}$. Hence the ratio between the two estimations on $\underline{\omega}(A, \Delta)$ is

$$(\sigma_n(A)/\sigma_1(\Delta)) \;/\; \rho(|A^{-1}|\Delta)^{-1} \approx n^{-1} \cdot n^2 \cdot n^{-1/2} = \sqrt{n}.$$

In other words, for orthogonal matrices Theorem 1.8 verifies regularity of interval matrices with radius up to a factor of $\sqrt{n}$ larger than (18). The following table shows that this ratio is indeed achieved for orthogonalized random matrices.

$(\sigma_n(A)/\sigma_1(\Delta)) \;/\; \rho(\lvert A^{-1}\rvert \cdot \Delta)^{-1}$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
$\Delta = \lvert A\rvert$	8.3	11.6	18.4	26.0
$\Delta = (1)$	8.3	11.6	18.4	25.9
$\sqrt{n}$	10.0	14.1	22.3	31.6

Table 1.1. Ratio of estimations (20) for $\underline{\omega}(A, \Delta)$, $A^{-1} = A^{-T}$ random, 50 samples each

2. Dense systems of nonlinear equations

With the preparations of the previous chapter we can state an inclusion theorem for systems of nonlinear equations. We formulate the theorem for an inclusion set Y which is an interval vector. A formulation for general compact and convex $\emptyset \neq Y \in \mathbb{IPIR}^n$ is straightforward, following the proof of Theorem 2.1 and using Lemma 1.1.

2.1. An existence test

Theorem 2.1. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function, $R \in \mathbb{R}^{n\times n}$, $[Y] \in \mathbb{IIR}^n$, $\widetilde{x} \in D$, $\widetilde{x} + [Y] \subseteq D$ and let a function $s_f : D \times D \to \mathrm{M}_{nn}(\mathbb{R})$ be given with

$$x \in \widetilde{x} + [Y] \Rightarrow f(x) = f(\widetilde{x}) + s_f(\widetilde{x}, x) \cdot (x - \widetilde{x}). \tag{21}$$

Define $Z := -R \cdot f(\widetilde{x}) \in \mathbb{R}^n$, $\boldsymbol{C} : D \to \mathrm{M}_{nn}(\mathbb{R})$ with $\boldsymbol{C}_x := \boldsymbol{C}(x) = I - R \cdot s_f(\widetilde{x}, x)$ and define $[V] \in \mathbb{IIR}^n$ using the following Einzelschrittverfahren for $1 \leq i \leq n$:

$$V_i := \{\, \diamond(Z + \boldsymbol{C}_{\widetilde{x}+[U]} \cdot [U]\,\}_i \quad \text{with} \quad [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T. \tag{22}$$

If

$$[V] \subsetneqq [Y], \tag{23}$$

then R and every matrix $C \in \boldsymbol{C}_{\widetilde{x}+[V]}$ are regular, and there exists some $\widehat{x} \in \widetilde{x} + [V]$ with $f(\widehat{x}) = 0$.

Remark. The interval vector $[U]$ in (22) is defined individually for every index i (see (13)). For better readability we omit an extra index for $[U]$ and use V_i and $[V]_i$

synonymously.

Proof. Define $g : D \to \mathbb{R}^n$ by $g(x) := x - R \cdot f(x)$ for $x \in D$. The definition (22) of $[V]$ together with (23) yields

$$\diamond(Z + \boldsymbol{C}_{\widetilde{x}+[V]} \cdot [V]) \subseteq [V].$$

Hence, for all $x \in \widetilde{x} + [V]$ we have by (23) and (21)

$$\begin{aligned} g(x) &= x - R \cdot f(x) = x - R \cdot \{f(\widetilde{x}) + s_f(\widetilde{x}, x) \cdot (x - \widetilde{x})\} \\ &= \widetilde{x} - R \cdot f(\widetilde{x}) + \{I - R \cdot s_f(\widetilde{x}, x)\} \cdot (x - \widetilde{x}) \\ &\in \widetilde{x} - R \cdot f(\widetilde{x}) + \{I - R \cdot s_f(\widetilde{x}, \widetilde{x} + [V])\} \cdot [V] \\ &\subseteq \widetilde{x} + Z + \boldsymbol{C}_{\widetilde{x}+[V]} \cdot [V] \\ &\subseteq \widetilde{x} + [V], \end{aligned}$$

that is, g is a continuous mapping of the nonempty, convex and compact set $\widetilde{x} + [V]$ into itself. Thus Brouwer's Fixed Point Theorem implies the existence of some $\widehat{x} \in \widetilde{x} + [V]$ with $g(\widehat{x}) = \widehat{x} = \widehat{x} - R \cdot f(\widehat{x})$, and hence $R \cdot f(\widehat{x}) = 0$. Lemma 1.3 implies the regularity of R and every matrix $C \in \boldsymbol{C}_{\widetilde{x}+[V]}$ which in turn yields $f(\widehat{x}) = 0$ and demonstrates the theorem. ■

Theorem 2.1 implies $\diamond(Z + \boldsymbol{C}_{\widetilde{x}+[V]} \cdot [V]) \subseteq [V]$, not necessarily with $\subsetneq$. The interesting point in using the Einzelschrittverfahren is that the set of iteration matrices $\boldsymbol{C}_{\widetilde{x}+[U]}$ is not fixed but shrinks in every step. Therefore, (23) may be satisfied, whereas $Z + \boldsymbol{C}_{\widetilde{x}+[Y]} \cdot [Y] \subsetneq [Y]$ is not true. So the Einzelschrittverfahren is a convergence accelerator. For examples, see table 2.1.

We want to stress that f is only required to be continuous; no differentiability assumption is required. Also, the only assumption on the function s_f is (21). Moreover, we only conclude existence, and not uniqueness of the zero $\widehat{x}$ within $\widetilde{x} + [V]$. On the other hand, we need the expansion (21) of f only w.r.t. $\widetilde{x}$. Note that we do not require $\widetilde{x} \in \widetilde{x} + [V]$. Those facts are demonstrated in the following simple example. Define

$$f(x) := \begin{cases} |x| \cdot \sin(1/x) & \text{for } x \neq 0 \\ 0 & \text{for } x = 0. \end{cases} \tag{24}$$

f is continuous on the whole real axis. We set $\widetilde{x} := 0.7$, $[Y] := [-3.2, -0.2]$. Note that $[Y]$ should contain the difference of a zero of f and $\widetilde{x}$. Then the slope condition (21) reads

$$x \in [-2.5, 0.5] \Rightarrow f(x) = f(\widetilde{x}) + s_f(\widetilde{x}, x) \cdot (x - \widetilde{x}).$$

One can show (see Figure 2.1) that $[S] := [0.5, 2]$ satisfies

$$x \in [-2.5, 0.5] \Rightarrow f(x) \in f(\widetilde{x}) + [0.5, 2] \cdot (x - \widetilde{x})$$

demonstrating the existence of such a function s_f. Of course, the usual approach would be to compute a function s_f and from that the interval $[S]$. For a large class of functions this process can be automized as will be discussed in Chapter 3. Then, setting $R := 0.5$, we have

$$\begin{aligned} & -R \cdot f(\widetilde{x}) + \{1 - R \cdot s_f(\widetilde{x}, \widetilde{x} + [Y])\} \cdot [Y] \\ \subseteq\ & [-0.35, -0.34] + \{1 - 0.5 \cdot [0.5, 2]\} \cdot [-3.2, 0.2] \\ \subseteq\ & [-0.35, -0.34] + [0, 0.75] \cdot [-3.2, 0.2] \\ \subseteq\ & [-2.75, -0.34] \subsetneqq [Y] = [-3.2, -0.2] \end{aligned}$$

This demonstrates by Theorem 2.1 that $R \neq 0$ and $s \neq 0$ for all $s \in S$ and the existence of some $\widehat{x} \in \widetilde{x} + [-2.75, -0.34] = [-2.05, 0.36]$ with $f(\widehat{x}) = 0$. In our example, we have in fact infinitely many zeros, a point, where f is non-differentiable as well as infinitely many zeros of f' within the inclusion interval $[-2.05, 0.36]$.

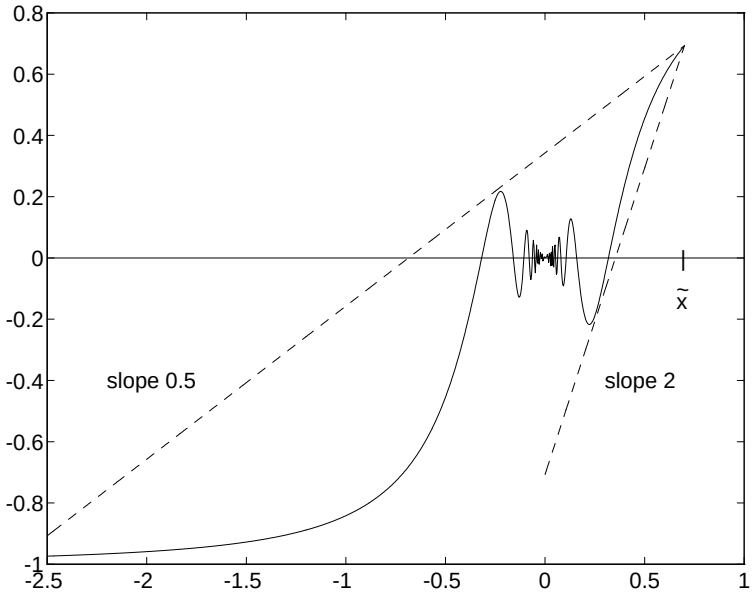


Figure 2.1 Graph of $|x| \cdot \sin 1/x$ with slopes

In one dimension, by its definition the function s_f must be continuous except in $\widetilde{x}$. This changes for $n \geq 2$. There, the matrix function $s_f(\widetilde{x}, x)$ needs only be continuous "in the direction $x - \widetilde{x}$"; otherwise discontinuities may occur. The definition of $s_f(\widetilde{x}, x)$ for $x = \widetilde{x}$ is almost arbitrary; any matrix value within $\boldsymbol{C}_{\widetilde{x}+[V]}$ does not influence the assumptions of Theorem 2.1. In fact, the set $s_f(\widetilde{x}, \widetilde{x} + [Y])$ need not even be connected.

The example above is, of course, an artificial example to demonstrate some basic observations concerning Theorem 2.1. In a practical application the diameter of $[Y]$ is usually small compared to the absolute value of $\widetilde{x}$, and this is advantageous in order to obtain

accurate inclusions. For comparing different methods consider the original Krawczyk operator [52]

$$K([X]) := \widetilde{x} - R \cdot f(\widetilde{x}) + \{I - R \cdot f'([X])\} \cdot ([X] - \widetilde{x}),$$

where f is supposed to be in C^1 and $f'([X])$ is an interval evaluation of the Jacobian of f, the latter, for example, obtained by automatic differentiation (see [69], [26]). Then the proof of Theorem 2.1 shows that $g(x) \in K([X])$ for all $x \in [X]$ provided $\widetilde{x} \in [X]$. Furthermore $K([X]) \subsetneqq X$ implies the existence of some $\widehat{x} \in [X]$ with $f(\widehat{x}) = 0$. Use of the Jacobian as slope function expands f within $[X]$ w.r.t. every $x \in [X]$, not only w.r.t. $\widetilde{x}$. This also yields the uniqueness of $\widehat{x}$ within $[X]$ (cf. also Theorem 2.3). We now compare the following two algorithms:

Given R, $\widetilde{x}$ do

I) $[Z] := \widetilde{x} - R \cdot f(\widetilde{x}); [X] := [Z]; k = 0;$
repeat
$k = k + 1;$
$[Y] := \mathrm{hull}(\widetilde{x}, [X] \circ \varepsilon);$
$[\boldsymbol{C}] := I - R \cdot f'([Y]);$
$[X] := [Z] + [\boldsymbol{C}] \cdot ([Y] - \widetilde{x});$
until $[X] \subsetneqq [Y]$ or $k = 15;$
if $[X] \subsetneqq [Y]$ then
$\exists^{1-1} \widehat{x} \in [X] : f(\widehat{x}) = 0$

II) $[Z] := -R \cdot f(\widetilde{x}); [X] := [Z]; k = 0;$
repeat
$k = k + 1;$
$[Y] := \mathrm{hull}(0, [X] \circ \varepsilon);$
$[\boldsymbol{C}] := I - R \cdot f'(\widetilde{x} + [Y]);$
$[X] := [Z] + [\boldsymbol{C}] \cdot [Y];$
until $[X] \subsetneqq [Y]$ or $k = 15;$
if $[X] \subsetneqq [Y]$ then
$\exists^{1-1} \widehat{x} \in \widetilde{x} + [X] : f(\widehat{x}) = 0$

Finally, we compare with a third algorithm which is

III) algorithm II) using an Einzelschrittverfahren.

The computing times for all three algorithms are roughly the same provided the same number of iterations is performed. For good approximations $R \approx f'(\widetilde{x})^{-1}$ and $\widetilde{x}$ with $f(\widetilde{x}) \approx 0$ all three algorithms perform similarly, in the number of iterations as well as in the accuracy of the inclusion intervals. We are interested in the effect of bad approximations R and $\widetilde{x}$, the quality of which we do not know a priori. In practice, if the problem is not too ill-conditioned, a few Newton iterations can improve the quality of an approximate solution, and therefore we can usually assume $\widetilde{x}$ to be fairly good. The quality of R, however, may be poor if it originates from inverting $f'(\overline{x})$ with a poor starting value $\overline{x}$. Moreover, improving R is expensive, and therefore one might try a verification step with the given one.

Consider the following example by Branin (cf. [1])

$$f_1 = 2\sin(2\pi\, x_1/5) \cdot \sin(2\,\pi\, x_3/5) - x_2$$
$$f_2 = 2.5 - x_3 + 0.1 \cdot x_2 \cdot \sin(2\,\pi\, x_3) - x_1$$
$$f_3 = 1 + 0.1 \cdot x_2 \cdot \sin(2\,\pi\, x_1) - x_3$$

with a solution $\widehat{x} = (1.5, 1.809\ldots, 1.0)^T$. All of the following computations are performed in single precision ($\sim$ 7 decimal digits).

		number of iterations		relative accuracy		
algorithm	δ	average	maximum	average	minimum	# failed
I)	0.001	1	1	1.6e-6	2.6e-6	0
II)	0.001	1	1	1.2e-6	1.6e-6	0
III)	0.001	1	1	1.2e-6	1.6e-6	0
I)	0.1	1	1	2.4e-5	7.4e-5	0
II)	0.1	1	1	1.6e-6	2.8e-6	0
III)	0.1	1	1	1.4e-6	2.3e-6	0
I)	0.25	1	1	5.9e-5	1.7e-4	0
II)	0.25	1	1	2.3e-6	4.5e-6	0
III)	0.25	1	1	1.8e-6	3.4e-6	0
I)	0.5	1.4	3	1.4e-4	7.2e-4	0
II)	0.5	1.6	3	4.7e-6	2.9e-5	0
III)	0.5	1.3	2	3.1e-6	1.2e-5	0
I)	0.75	2.9	13	4.1e-4	4.2e-3	0
II)	0.75	2.4	5	1.5e-5	2.9e-4	0
III)	0.75	1.9	4	7.8e-6	9.9e-5	0
I)	1.0	4.9	15	7.4e-4	7.5e-3	19
II)	1.0	3.6	11	3.3e-4	3.8e-2	11
III)	1.0	3.0	7	7.0e-5	2.7e-3	8

Table 2.1 Comparison algorithms I, II, III

For all three algorithms we performed 2 Newton steps from the starting value $(0, 0, 0)^T$ given in Branin's example. This produces an $\widetilde{x}$ with a relative accuracy 10^{-7}, which is almost working precision. Then R is computed by

$$R_{ij} = \{f'(\widehat{x})^{-1}\}_{ij} \cdot (1 + \delta \cdot \mathrm{rand}_{ij})$$

where δ is the perturbation parameter and rand_{ij} are uniformly distributed random numbers in $[-1, 1]$. In the following table we display the average and maximum number of

interval iterations necessary (the k in the algorithm), average and minimum relative accuracy w.r.t. the midpoint of the components of the computed inclusion interval, and the number of cases in which no inclusion was achieved. For every perturbation value δ for R we performed 100 test runs.

The comparison of the three algorithms depends, of course, very much on the problem to solve and on the choice of ε. The first algorithm computes an inclusion of the solution itself; therefore we have to choose a small ε to allow convergence and to keep a good relative accuracy. In our test results we used a relative inflation by 10^{-5} for the first algorithm. The second and third algorithm enclose the error w.r.t. $\widetilde{x}$; therefore a reasonable inflation is necessary. In the example we took 20 % relative inflation. In all three algorithms, we expanded the interval adding twice the smallest positive machine number, thus taking the second predecessor, successor of the left, right bound, respectively.

The table shows that for smaller δ (up to 25 %), all algorithms use 1 interval iteration. Enclosing the error rather than the solution gains little for a good approximation R but more than one figure for $\delta = 0.25$. Remember that the value of δ is a *maximum* of random perturbations for R. For larger δ the third algorithm needs the smallest number of interval iterations whereas for a maximum of 100 % perturbation the number of failures is best for the third algorithm. The number of iterations is important w.r.t. the computing time because every iteration requires the evaluation of a Jacobian and the multiplication by R. Another interesting approach to construct a starting region $[X]$ for nonlinear systems is described by Alefeld [6].

2.2. Refinement of the solution

If an inclusion is not good enough, iterative refinement using intersection is possible. For the sake of completeness we state the following theorem. However, we do not recommend extensive use of this technique. In most cases a pure floating point iteration with subsequent verification step will be more efficient, in terms of computing time as well as accuracy.

Theorem 2.2. With the assumptions of Theorem 2.1, assume $[V] \subsetneqq [Y]$. Then $\widehat{x} \in \widetilde{x} + [V]$, and if the i-th component V_i of $[V]$, $1 \le i \le n$ is replaced by

$$V_i := V_i \cap \ \{\diamond\,(Z + \boldsymbol{C}_{\widetilde{x}+[V]} \cdot [V])\}_i,$$

then $\widehat{x} \in \widetilde{x} + [V]$ still holds true for the new $[V]$. In other words, continuing with the Einzelschrittverfahren described in Theorem 2.1 together with componentwise intersection no zero of f can be lost. If, with the assumptions of Theorem 2.1 except (23)

$$V_i \cap \ \{\diamond\,(Z + \boldsymbol{C}_{\widetilde{x}+[U]} \cdot [U])\}_i = \emptyset \quad \text{for } [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T$$

for some $1 \le i \le n$, then $\widetilde{x} + [Y]$ contains no zero of f.

Proof. Following the proof of Theorem 2.1 every zero $\widehat{x} \in \widetilde{x} + [V]$ is a fixed point of g and is therefore contained in $\diamond(Z + \boldsymbol{C}_{\widetilde{x}+[V]} \cdot [V])$ which proves the first part of the theorem. Using $g(x) \in Z + \boldsymbol{C}_{\widetilde{x}+[Y]} \cdot [Y]$ for all $x \in \widetilde{x} + [Y]$ proves the second part. ■

2.3. Verification of uniqueness

We have proven existence of a zero within a given interval, but not uniqueness of this zero. The latter can be verified by the following theorem.

Theorem 2.3. With the assumptions of Theorem 2.1, assume $[V] \subsetneqq [Y]$, i.e. there exists some $\widehat{x} \in \widetilde{x} + [V]$ with $f(\widehat{x}) = 0$. For a given interval vector $[W] \supseteq [V]$, $\widetilde{x} + [W] \subseteq D$, let the function s_f satisfy

$$y \in \widetilde{x} + [W] \Rightarrow f(y) = f(x) + s_f(x, y) \cdot (y - x) \quad \text{for all } x \in \widetilde{x} + [V] \tag{25}$$

in addition to (21).

If all $S \in s_f(\widetilde{x} + [V], \widetilde{x} + [W])$ are regular, then the zero $\widehat{x}$ of f is unique in $\widetilde{x} + [W]$.

Proof. For $\widehat{y} \in \widetilde{x} + [W]$ with $f(\widehat{y}) = 0$, (25) implies

$$0 = f(\widehat{y}) = f(\widehat{x}) + s_f(\widehat{x}, \widehat{y}) \cdot (\widehat{y} - \widehat{x}) = s_f(\widehat{x}, \widehat{y}) \cdot (\widehat{y} - \widehat{x})$$

and the regularity of $s_f(\widehat{x}, \widehat{y})$ implies $\widehat{y} = \widehat{x}$. ■

The regularity of the set of matrices $s_f(\widetilde{x} + [V], \widetilde{x} + [W])$ can be verified by means of Lemma 1.1 or 1.3. In the latter case it is simpler to use Lemma 1.6, which saves a lot of computing time. Computing large inclusion intervals containing exactly one solution is important, for example, in global optimization (see [38]) or, for verified computation of all zeros of a nonlinear system within a given domain [48]. For large banded or sparse matrices, Theorem 1.8 can be used to prove regularity.

2.4. Verification of existence and uniqueness for large inclusion intervals

In a practical implementation it can be advantegeous first to verify existence in a small solution set $\widetilde{x} + [V]$ and then to verify uniqueness in a much larger one $\widetilde{x} + [W]$. This two-step approach is superior to trying to verify existence *and* uniqueness in *one* step for the larger set $\widetilde{x} + [W]$.

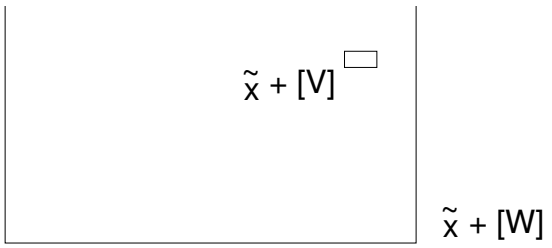


Figure 2.2 Verifying existence in $\widetilde{x} + [V]$ and uniqueness in $\widetilde{x} + [W]$

This is because we would need to expand $f(x)$ w.r.t. *every point* in the large set $\widetilde{x} + [W]$ which makes the set of matrices $s_f(\widetilde{x}+[W], \widetilde{x}+[W])$ very thick and possibly not convergent.

Given a "large" set $\widetilde{x} + [W]$ in which we want to verify existence and uniqueness of a zero $\widehat{x}$ of f we can proceed as follows:

1. Compute an approximate solution $\widetilde{x}$
2. Compute some $\widetilde{x} + [V]$ with $\widehat{x} \in \widetilde{x} + [V]$ of small size near $\widetilde{x}$ using Theorem 2.1
3. Possibly refine $\widetilde{x} + [V]$ using Theorem 2.2
4. Verify regularity of $s_f(\widetilde{x} + [V], \widetilde{x} + [W])$ using Lemma 1.6

Algorithm 2.1 Verification of existence and uniqueness within large intervals

We add some practical remarks on the above algorithm:

- The approximation $\widetilde{x}$ should be good. Very much in the sense of Wilkinson (0.3) we do as much as possible in floating point. It is better and faster. It remains the task for interval analysis to *verify* the quality of an approximation.

- The inclusion interval $\widetilde{x} + [V]$ should be of good quality. However, rather than applying 1 iteration in step 3 it is better to perform 2 iterations in step 1 (if $\widetilde{x}$ was not good enough).

- In step 4 only *regularity* of the interval matrix $[C] := s_f(\widetilde{x} + [V], \widetilde{x} + [W])$ is to be verified. The fundamental advantage of the above algorithm is that this $[C]$ is *constant*. That means, applying Lemma 1.6 and Theorem 1.5 leads to the verification existence and uniqueness of the zero of f within $\widetilde{x} + [W]$ *if and only if* $\rho(|I - R \cdot [C]|) < 1$.
  Furthermore, it is *much faster* than applying Theorem 2.1 directly to $\widetilde{x} + [W]$ because if an iteration has to be performed, the matrix $s_f(\widetilde{x} + [W], \widetilde{x} + [W])$ has to be recomputed and the multiplication by R has to be executed in every step. This means computation of a whole Jacobian or slope, and a matrix times interval matrix product. In contrast, Algorithm 2.1 needs only a real matrix times real vector multiplication in each iteration step when using Lemma 1.6.

In other words, we can expect to obtain verification of existence and uniqueness faster and within larger intervals using Algorithm 2.1. This can be of great importance in practical applications. For example, in global optimization many refinements can possibly be saved, especially in higher dimensions.

We illustrate the use of Algorithm 2.1 with a simple one-dimensional example. Let

$$f(x) := e^x - 2x - 1 \tag{26}$$

with $f(0) = 0$, $\widetilde{x} := 0.1$ and $\widetilde{x} + [Y] := [-0.1, +0.1]$. The slope in the one-dimensional case is the set of secants $s(\widetilde{x}, x) := \{f(x) - f(\widetilde{x})\}/(x - \widetilde{x})$ for $x \neq \widetilde{x}$, and it is easy to see that

$$[S] := \{\, s(\widetilde{x}, x) \mid x \in \widetilde{x} + [Y], x \neq \widetilde{x} \,\} = [s(0.1, -0.1), s(0.1, 0.1)] \subseteq [-1.0, -0.89].$$

Setting $R \approx \mathrm{mid}([S])^{-1} = -0.945^{-1}$, e.g. $R := -1$ yields

$$\begin{aligned} -R \cdot f(\widetilde{x}) + \{1 - R \cdot [-1, -0.89]\} \cdot [-0.1, +0.1] &= [-0.106, -0.083] =: [V] \\ &\subsetneq [-0.2, 0] = [Y] \end{aligned}$$

Therefore, there is a zero $\widehat{x}$ of f within $\widetilde{x} + [V] = [-0.006, +0.017]$, namely $\widehat{x} = 0$. Up to now we do not know uniqueness. Consider $\widetilde{x} + [W] := [-2, 1]$. Then the set of secants computes to

$$\begin{aligned} \{\, s(x, y) \mid x \in \widetilde{x} + [V], y \in \widetilde{x} + [W], y \neq x \,\} &= [s(-0.006, -2), s(0.017, 1)] \\ &= [-1.570, -0.269] \end{aligned}$$

which does not contain zero and therefore implies the uniqueness of $\widehat{x} = 0$ in the larger interval $[-2, 1]$. On the other hand the slope function for $\widetilde{x} + [W]$

$$\{\, s(x, y) \mid x, y \in \widetilde{x} + [W], x \neq y \,\} \subseteq [-1.8647, +0.7183]$$

contains 0 and is therefore not suitable for an inclusion. Even if we take $\widetilde{x} := \widehat{x} = 0$ and $[Y] := \widetilde{x} + [W] = [-2, 1]$ we cannot even verify existence of a zero within $[-2, 1]$, because $\{\, s(0, x) \mid x \in [-2, 1], x \neq 0 \,\} \subseteq [-1.5677, -0.2818] =: [S]$, and taking $R := \mathrm{mid}([S])^{-1} = -1.0814$ yields

$$\begin{aligned} -R \cdot f(\widetilde{x}) + \{1 - R \cdot [S]\} \cdot [W] &= \{1 + 1.0814 \cdot [-1.5677, -0.2818]\} \cdot [-2, 1] \\ &\subseteq [-1.391, +1.391] \not\subseteq [-2, 1]. \end{aligned}$$

We want to add two remarks to the previous example. First, in our computation of $s(x, y)$, we always assumed $x \neq y$. This is because we calculated s directly from the set of secants which is undefined for $x = y$. On the other hand, the assumptions of our theorems are always assumptions on an expansion of f like (25). Therefore the values $s_f(x, y)$ for $x = y$ are uninteresting, because in this case $f(x) = f(y)$. Therefore, we

could exclude these values in the assumptions of our theorems. On the other hand, in a practical application this in turn is unimportant, and for the sake of better readability we did not exclude $x = y$. The second remark concerns the computation of the function s_f, which was simple in the previous example and essentially done by hand using auxiliary information of the function f. In Chapter 3 we will discuss a method for automatic evaluation of such funtions s_f for a wide class of functions f.

In the case of functions depending on data which are afflicted with tolerances, we can use parametrized functions with parameters varying within a certain tolerance. It is straightforward to give theorems corresponding to Theorems 2.1, 2.2 and 2.3 for this case. However, in this case the solution is not a single point, but we have a whole set of solutions corresponding to parameters within the tolerances. Then we obtain an inclusion of this solution set, which is, by the principles of interval analysis, an *outer* inclusion, i.e. a set which is verified to *contain* the solution set. However, it is important to know whether this inclusion is possibly an overestimation of the true solution set or not.

2.5. Inner inclusions of the solution set

The amount of overestimation can be estimated by means of *inner* inclusions. For some set $\Sigma \in \mathbb{IPIR}^n$ we call $[X] \in \mathbb{IIR}^n$ an inner inclusion if for every component $1 \le i \le n$,

$$\inf_{\sigma\in\Sigma} \sigma_i \le \inf([X])_i \quad \text{and} \quad \sup([X])_i \le \sup_{\sigma\in\Sigma} \sigma_i$$

holds. In other words, for every component $[X]_i$ of $[X]$, there are points in Σ the i-th component of which are left of the lower bound of $[X]_i$ and right of the upper bound of $[X]_i$, respectively. Such inner bounds for the solution set of a parametrized system of nonlinear equations can be computed by means of the following theorem. Note that these are bounds for the solution set for finite perturbations of parameters within some set of parameters C.

Theorem 2.4. Let $f : D_p \times D_n \subseteq \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous w.r.t. the unknowns $x \in \mathbb{R}^n$, let $R \in \mathbb{R}^{n\times n}$, $[Y] \in \mathbb{IIR}^n$, $\widetilde{x} \in D_n$ with $\widetilde{x} + [Y] \subseteq D_n$, and for $C \in \mathbb{IPIR}^p$, $C \subseteq D_p$ let a function $s_f : D_p \times D_n \times D_n \to \mathrm{M}_{nn}(\mathbb{IR})$ be given with

$$c \in C,\ x \in \widetilde{x} + [Y] \Rightarrow f(c, x) = f(c, \widetilde{x}) + s_f(c, \widetilde{x}, x) \cdot (x - \widetilde{x}). \tag{27}$$

Define $Z := -R\cdot f(C, \widetilde{x}) \in \mathbb{IPIR}^n$, $\boldsymbol{C} : D_p \times D_n \to \mathrm{M}_{nn}(\mathbb{IR})$ with $\boldsymbol{C}(c, x) := I - R\cdot s_f(c, \widetilde{x}, x)$ and define $[V] \in \mathbb{IIR}^n$ using the following Einzelschrittverfahren:

$$\begin{aligned} 1 \le i \le n \ \ : \ \ & V_i := \{\diamond(Z + \boldsymbol{C}(C, \widetilde{x} + [U]) \cdot [U])\}_i \\ & \text{with } [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T. \end{aligned} \tag{28}$$

If

$$[V] \subsetneqq [Y], \tag{29}$$

then R and every matrix within $\boldsymbol{C}(C, \widetilde{x} + [V])$ is regular, and for every $c \in C$ there exists some $\widehat{x}_c \in \widetilde{x} + [V]$ with $f(c, \widehat{x}_c) = 0$. Define the solution set Σ of f within $\widetilde{x} + [Y]$ w.r.t. parameters $c \in C$ by

$$\Sigma := \{\, x \in \widetilde{x} + [Y] \mid \exists\, c \in C : f(c, x) = 0 \,\}. \tag{30}$$

Then $\Sigma \subseteq \widetilde{x} + [V]$, and abbreviating

$$[Z] := \diamond(Z) \quad \text{and } [\Delta] := \diamond(\boldsymbol{C}(C, \widetilde{x} + [V]) \cdot [V]),$$

the following componentwise estimations hold true

$$\begin{aligned} &\widetilde{x}_i + \inf([Z]_i) + \sup([\Delta]_i) \geq \inf_{\sigma \in \Sigma}\ \sigma_i \quad \text{and} \\ &\widetilde{x}_i + \sup([Z]_i) + \inf([\Delta]_i) \leq \sup_{\sigma \in \Sigma}\ \sigma_i. \end{aligned} \tag{31}$$

Proof. The first part of the theorem follows by applying Theorem 2.1 for every $c \in C$. Let $c \in C$ be fixed but arbitrary. Then defining $g : D_n \to \mathbb{IR}^n$ by $g(x) := x - R \cdot f(c, x)$, following the first part of the proof of Theorem 2.1, and using (28) and (29) for $1 \leq i \leq n$ we see that every fixed point of g within $\widetilde{x} + [Y]$ needs also to be in $\widetilde{x} + [V]$. But every zero of f is a fixed point of g, hence $\Sigma \subseteq \widetilde{x} + [V]$. For $c \in C$ we have for every $x \in \widetilde{x} + [Y]$

$$\begin{aligned} \widetilde{x} - R \cdot f(c, \widetilde{x}) &= \widetilde{x} - R \cdot \{f(c, x) + s_f(c, \widetilde{x}, x) \cdot (\widetilde{x} - x)\} \\ &= x - R \cdot f(c, x) - \{I - R \cdot s_f(c, \widetilde{x}, x)\}(x - \widetilde{x}). \end{aligned}$$

For this $c \in C$ there exists some $\widehat{x}_c \in \widetilde{x} + [V] \subseteq \widetilde{x} + [Y]$ with $f(c, \widehat{x}_c) = 0$ and

$$\begin{aligned} \widetilde{x} - R \cdot f(c, \widetilde{x}) &= \widehat{x}_c - \{I - R \cdot s_f(c, \widetilde{x}, \widehat{x}_c)\}(\widehat{x}_c - \widetilde{x}) \\ &\in \Sigma - \boldsymbol{C}(C, \widetilde{x} + [V]) \cdot [V] \\ &\subseteq \Sigma - [\Delta]. \end{aligned} \tag{32}$$

The left hand side of (32) is an element of $\widetilde{x} + Z$. Since $c \in C$ was choosen arbitrarily we also have

$$\widetilde{x} + Z \subseteq \Sigma - [\Delta] \quad \text{or} \quad \forall\, z \in Z\ \exists\, \sigma \in \Sigma\ \exists\, \delta \in [\Delta] : \widetilde{x} + z = \sigma - \delta. \tag{33}$$

For fixed index i between 1 and n and every $\varepsilon > 0$, there is a $c \in C$ with

$$\{-R \cdot f(c, \widetilde{x})\}_i \leq \inf([Z]_i) + \varepsilon.$$

Together with (33) this shows the existence of some $c \in C$, $\sigma \in \Sigma$ and $\delta \in [\Delta]$ with

$$\widetilde{x} + \inf([Z]_i) + \varepsilon \geq \{\widetilde{x} - R \cdot f(c, \widetilde{x})\}_i = \sigma_i - \delta_i \geq \inf_{\sigma \in \Sigma}\ \sigma_i - \sup([\Delta]_i)$$

and proves the first inequality. The second one follows similarly. ■

In a practical application the sharpness of the bounds depends on $[\Delta]$: this is exactly the difference between the inner and outer bounds. But $[\Delta]$ is the product of $I - R \cdot s_f(C, \widetilde{x} + [V])$ and $[V]$. The first factor is small for $R \approx s_f(\widetilde{c}, \widetilde{x})^{-1}$ for some $\widetilde{c} \in C$ and diameter of C not too big, and the second factor $[V]$ is the difference between $\widetilde{x}$ and the solutions $\widehat{x}_c$, and therefore also small. In other words, $[\Delta]$ is the product of small quantities for reasonable parameter tolerances. This means we can expect inner and outer bounds not too far apart.

We give a simple example. We use the nonlinear system given by Broyden [16] which we parametrize with 3 parameters:

$$\begin{aligned} p_1 \cdot \sin(xy) - y \,/\, (4\pi) - x \,/\, p_2 &= 0 \\ (1 - 1 \,/\, (4\pi)) \cdot (e^{p_3 \cdot x} - e) + ey \,/\, \pi - 2ex &= 0 \end{aligned} \tag{34}$$

with initial approximation (0.6,3) and parameter values

$$\begin{aligned} p_1 &\in 0.5 \cdot [1 - \varepsilon, 1 + \varepsilon] \\ p_2 &\in 2.0 \cdot [1 - \varepsilon, 1 + \varepsilon] \qquad\qquad \text{for } \varepsilon = 0.01 \text{ in all 3 cases.} \\ p_3 &\in 2.0 \cdot [1 - \varepsilon, 1 + \varepsilon] \end{aligned}$$

For the midpoint parameter value $(0.5, 2, 2)^T$ we have a solution $\widehat{x} = (0.5, \pi)$. For the expansion of the function f we use slopes, *not* the Jacobian. The latter gives poorer results as will be discussed in the next chapter. After a short computation, choosing $\widetilde{x} \approx \widehat{x}$, we obtain

$$[Z] \subseteq \begin{pmatrix} [-0.01819, & +0.01814] \\ [-0.03389, & +0.03358] \end{pmatrix}, \; [\Delta] \subseteq \begin{pmatrix} [-0.0052, & +0.0052] \\ [-0.0071, & +0.0072] \end{pmatrix}$$

and

$$\Sigma \subseteq \begin{pmatrix} [0.4766, & 0.5233] \\ [3.1006, & 3.1823] \end{pmatrix}.$$

Automated evaluation of the slope function s_f will be discussed in the next chapter. The problem in applying (31) is that outer bounds for $[\Delta]$ suffice, but we need *inner* bounds for $[Z]$. We could regard $[Z]$ as a good approximation for $\diamond\, \Sigma$ with error term $[\Delta]$. For linear systems the determination of inner bounds of $[Z]$ is not too difficult, as we will see in Chapter 4.

In the nonlinear case we could compute $-R \cdot f(c, \widetilde{x})$ for several random $c \in C$ and take the interval hull. This yields, especially for larger dimensions, very poor results

(an example for linear systems is given in Chapter 4). If f is differentiable w.r.t. the parameters c, a better method is to locally linearize:

$$R \cdot f(c, \widetilde{x}) \approx R \cdot f(\widetilde{c}, \widetilde{x}) + \left\{ R \cdot \frac{\partial f}{\partial c}\,(\widetilde{c}, \widetilde{x}) \right\} \cdot (c - \widetilde{c}).$$

If $C = [C]$ is a parameter interval and $\widetilde{c}$ its midpoint, then the matrix $R \cdot \frac{\partial f}{\partial c}\,(\widetilde{c}, \widetilde{x})$ is the local steepest descent direction. If we take $c \in \partial[C]$ with

$$\operatorname{sign}(c - \widetilde{c})_j = \pm \operatorname{sign}(R \cdot \frac{\partial f}{\partial c}\,(\widetilde{c}, \widetilde{x}))_{ij}, \tag{35}$$

we have the locally best choices of c for the i-th component of $[Z]$. In our example it is

$$R \cdot \frac{\partial f}{\partial c}\,(\widetilde{c}, \widetilde{x}) \approx \begin{pmatrix} -1.9 & -0.2 & -0.2 \\ -0.9 & -0.1 & 1.3 \end{pmatrix}$$

and the corresponding values for $-R \cdot f(c, \widetilde{x})$ are

$$\begin{pmatrix} -0.01818 \\ 0.01968 \end{pmatrix}, \quad \begin{pmatrix} 0.01813 \\ -0.01999 \end{pmatrix}, \quad \begin{pmatrix} -0.00965 \\ -0.03388 \end{pmatrix}, \quad \begin{pmatrix} 0.00960 \\ 0.03357 \end{pmatrix},$$

for c according to (35). Thus

$$[Z] \supseteq \begin{pmatrix} [-0.01818, & +0.01813] \\ [-0.03388, & +0.03357] \end{pmatrix}.$$

If only the elongation of few solution components is needed, the nonlinear system can be solved using the specific parameters computed by (35).

The matrix $\frac{\partial f}{\partial c}(\widetilde{c}, \widetilde{x})$ can be computed in an automated process by means of automatic differentiation [69], [26]. If f is not differentiable w.r.t. the parameters c, slopes instead of derivatives can be used as well.

Finally, applying Theorem 2.4 we obtain

$$\begin{pmatrix} [0.4870, & 0.5130] \\ [3.1149, & 3.1680] \end{pmatrix} \subseteq \diamond(\Sigma) \subseteq \begin{pmatrix} [0.4766, & 0.5233] \\ [3.1006, & 3.1823] \end{pmatrix}$$

which still gives reasonable accuracy for practical purposes. In the following figure the dashed rectangle is $[Z]$, the dotted one is the inner and the solid one is the outer inclusion for $\diamond(\Sigma)$, whereas the circles depict actual zeros of $f(c, x)$ for components c_i of the parameter c varying independently in an arithmetic progression between the bounds.

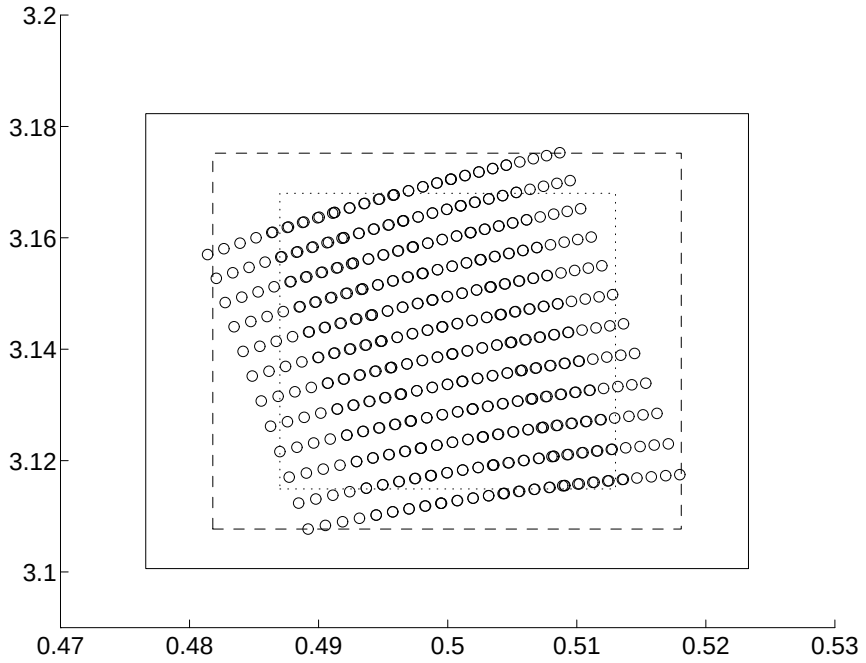


Figure 2.3 Inner and outer inclusions of the solution complex

Obviously $[Z]$ is a good approximation of $\diamond(\Sigma)$. It should be noted that with the assumptions of Theorem 2.4, we do not assure uniqueness of the zero of $f(c, x)$ for fixed parameter $c \in C$ within $\tilde{x} + [V]$. Therefore the solution complex need not be connected.

For ε larger than 0.015, the nonlinearities over the whole parameter domain become too big, and other techniques like bisection have to be used to obtain an inclusion.

Inner inclusions in the above sense were first investigated by Neumaier [65]. In comparison, the computation of inner inclusions using the methods described above is much cheaper. The above Theorem 2.4 was proved for Jacobians instead of slopes in [79].

As a larger, dense example consider

$$\begin{aligned} &\text{discretization of } u(t) + \int_0^1 H(s,t) \cdot (p \cdot u(s) + s + 1)^3 ds = 0 \\ &\text{with } p \in [0.9, 1.1] \text{ and } H(s,t) = \begin{cases} s(1-t) & \text{for } s \le t \\ t(1-s) & \text{for } s > t \end{cases} \end{aligned} \tag{36}$$

proposed by Moré and Cosnard [62]. This produces a problem with *full Jacobian matrix*. For dimension $n = 1000$ we obtained the following results. The inner and outer inclusions

for all 1000 components are of the same quality as those shown below.

$$\begin{array}{lcll} [-0.00085, -0.00081] & \subseteq & [X]_1 & \subseteq [-0.00089, -0.00077] \\ [-0.00170, -0.00163] & \subseteq & [X]_2 & \subseteq [-0.00178, -0.00155] \\ & & \dots & \\ [-0.00353, -0.00340] & \subseteq & [X]_{999} & \subseteq [-0.00364, -0.00328] \\ [-0.00176, -0.00170] & \subseteq & [X]_{1000} & \subseteq [-0.00182, -0.00164] \end{array}$$

Table 2.2. Inner and outer inclusion for (36) and dimension $n = 1000$

2.6. Sensitivity analysis with verified inclusion of the sensitivity

Computing inner inclusions for the solution set of a parametrized system of nonlinear equations yields a sensitivity analysis for finite perturbations of the input parameters. If we are interested in the sensitivity for a specific parameter value, we have to use other techniques. The difference to the previous approach is that we are looking for the sensitivity of the solution w.r.t. ε-perturbations in the limit $\varepsilon \to 0$.

For a meaningful definition of the sensitivity of a zero $\widehat{x}_{\widehat{c}}$, $f(\widehat{c}, \widehat{x}_{\widehat{c}}) = 0$ of a function f w.r.t. perturbations of a parameter $\widehat{c}$, we locally need continuous dependency of $\widehat{x}_{\widehat{c}}$ on $\widehat{c}$. Using an inclusion computed by means of Theorem 2.1 does not assure this. Consider for example $|x| + c^2$ for $\widehat{c} = 0$. Therefore we impose stronger assumptions on f. Moreover, we use a Jacobian-like function instead of s_f.

Theorem 2.5. Let $f : D_p \times D_n \subseteq \mathbb{R}^p \times \mathbb{R}^n$ be twice differentiable w.r.t. both the parameters D_p and unknowns D_n, such that for each parameter c_j at most one component function f_i is dependent on c_j. Let $\widetilde{x} \in D_n$, $R \in \mathrm{M}_{nn}(\mathbb{R})$, $[Y] \in \mathbb{IIR}^n$ such that $\widetilde{x} + [Y] \subseteq D_n$. Define

$$J(c, X) := \diamond \{ \frac{\partial f}{\partial x}(c, x) \mid x \in X \} \tag{37}$$

for $X \in \mathbb{IPIR}^n$, $X \subseteq D_n$, $c \in D_p$. For fixed parameter $\widehat{c} \in \mathrm{int}(D_p)$, let $Z := -R \cdot f(\widehat{c}, \widetilde{x})$ and let $\boldsymbol{C}([U]) := I - R \cdot J(\widehat{c}, \widetilde{x} + (0 \cup [U]))$. Define $[V] \in \mathbb{IIR}^n$ by means of the following Einzelschrittverfahren

$$1 \le i \le n : V_i := \{Z + \boldsymbol{C}([U]) \cdot [U]\}_i \quad \text{with } [U] := (V_i, \dots, V_{i-1}, Y_i, \dots, Y_n)^T.$$

Then

$$[V] \subsetneqq [Y]$$

implies the existence of a unique and simple zero $\widehat{x}_{\widehat{c}}$ of $f_{\widehat{c}}(x) = f(\widehat{c}, x)$ within $\widetilde{x} + [V]$. Let $c^* \in \mathbb{R}^p$, $c^* \ge 0$ and define

$$\begin{aligned} u &:= |R| \cdot |\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})| \cdot |c^*| \\ w &:= |I - R \cdot J(\widehat{c}, \widetilde{x} + (0 \cup [V]))| \cdot d[V]. \end{aligned}$$

Then

$$\phi := \max_i \ \frac{u_i}{(d[V] - w)_i}$$

is well defined. For small enough $\varepsilon > 0$ and for every c with $|c - \widehat{c}| \leq \varepsilon \cdot c^*$, there is a uniquely defined zero $\widehat{x}_c$ of f_c within $\widetilde{x} + [V]$, and the sensitivity vector of the zero $\widehat{x}_{\widehat{c}}$ of $f_{\widehat{c}}$ w.r.t. perturbations weighted by c^* can be defined componentwise by

$$\mathrm{Sens}(\widehat{x}_{\widehat{c}}, f, c^*)_k := \lim_{\varepsilon \to 0^+} \ \max \{ \frac{|\widehat{x}_{\widehat{c}} - \widehat{x}_c|_k}{\varepsilon} : \ |c - \widehat{c}| \leq \varepsilon \cdot c^* \}.$$

The sensitivity vector satisfies

$$\mathrm{Sens}(\widehat{x}_{\widehat{c}}, f, c^*) \in [u - \phi \cdot w, \ u + \phi \cdot w].$$

Proof. Theorem 2.4 implies the existence and uniqueness of a zero $\widehat{x}_{\widehat{c}}$ of $f_{\widehat{c}}$. This zero must be simple because of the regularity of the corresponding Jacobian which is implied by (37) and Theorem 2.4. Hence $Z + \boldsymbol{C}([V]) \subseteq [V]$. Now the proof of Theorem 2.4 in [80], which is lengthy and therefore omitted here, can be followed. ■

It should be mentioned that $\widehat{x}$ occuring in the definition of u can be replaced by $\widetilde{x} + [V]$, and the derivative w.r.t. the parameters can easily be computed using automatic differentiation in a forward or backward mode (cf. [26] or [69]).

The true sensitivity can be shown (cf. [80]) to be equal to

$$\mathrm{Sens}(\widehat{x}_{\widehat{c}}, f, c^*) = \left| \left(\frac{\partial f}{\partial x} (\widehat{c}, \widehat{x}) \right)^{-1} \right| \cdot \left| \frac{\partial f}{\partial c} (\widehat{c}, \widehat{x}) \right| \cdot c^*. \tag{38}$$

The main point of the above theorem is that the inverse Jacobian does not have to be included, but information from the inclusion process suffices to bound the difference between R and the true inverse of the Jacobian and finally to include the sensitivity. Bounds using absolute values have been investigated by Bauer [12]. Formula (38) verifies sensitivity results on linear systems by Skeel [86] and for matrix inversion and linear programming problems given by Rohn [72]. For many other standard problems in numerical analysis, (38) allows to state simple explicit formulas for the sensitivity. Using the verification scheme, inclusions of the sensitivity can also be computed (see [80]).

One main advantage of the approach we chose in Theorem 2.5 is that we are free to choose the weights c^*. A weight $c_i^* = |\widehat{c}_i|$ imposes a relative perturbation, a weight $c_i^* = 1$ an absolute perturbation and, especially, $c_i^* = 0$ imposes no perturbation at all for the parameter $\widehat{c}_i$. This is important in practical applications when $\widehat{c}_i$ is some system parameter like a system zero which need not be perturbed by construction. This is a main advantage over a norm approach.

3. Expansion of functions and slopes

In this chapter we follow two aims: to derive algorithmic principles for estimating the range of a function f over a domain X, and to expand f over X w.r.t. some point $\widetilde{x}$. For differentiable $f : \mathbb{R}^n \to \mathbb{R}$ this can be achieved by using the n-dimensional mean value theorem. For $[X] \in \mathbb{IIR}^n$,

$$\forall\, x, y \in [X] : \quad f(x) \in f(y) + [J] \cdot (y - x) \tag{39}$$

shows that $[J] := \bigcap\{\, [M] \in \mathbb{IM}_{nn}(\mathbb{R}) \mid \frac{\partial f}{\partial x}(x) \in [M] \text{ for all } x \in [X] \,\}$ allows an expansion w.r.t. *every* $y \in [X]$. The inclusion theorems given in Chapter 2 only require an expansion w.r.t. a single point $\widetilde{x}$. In turn this allows to prove existence but not uniqueness, the latter being verified with Theorem 2.3. Moreover, we do not want to restrict our functions to the class of differentiable ones, and we do not require $\widetilde{x} \in [X]$.

(39) shows that $\frac{\partial f}{\partial x}$ could serve as an expansion function. However, when applying the theorems of Chapter 2 to problems with high-nonlinearity or for the verification of large inclusion intervals, we are interested in expansion intervals of small diameter. For our example (2.6) we obtain for $[X] = [-2, 1]$

$$f'([X]) = e^{[X]} - [2] \subseteq [-1.865, 0.719],$$

which covers *all* slopes within $[X]$ w.r.t. every $x \in [X]$ rather than, e.g., the slopes $[S] = [-1.568, -0.281]$ w.r.t. the single point $\widetilde{x} = 0$. Moreover, in our example $f'([X])$ contains zero and is therefore not useful for our verification purposes. For n-dimensional functions we can shrink the diameter slightly by observing that for $[X] \in \mathbb{IIR}^n$, $\widetilde{x} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n \in C^1$,

$$\begin{aligned} &\forall\, x \in [X] : f(x) \in f(\widetilde{x}) + Z \cdot (x - \widetilde{x}) \quad \text{with} \\ &Z_{ij} := \frac{\partial f_i}{\partial x_j}\,(X_1, \ldots, X_{i-1}, \widetilde{x}_i \,\dot\cup\, X_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) \end{aligned} \tag{40}$$

holds. This was used by Hansen [28], see also Alefeld [4]. Using this also loses uniqueness of the zero. It helps, but more can be done. We are aiming for a simple method, easily and automatically executable on the computer, for computing an enclosure of the slopes of f w.r.t. a fixed point $\widetilde{x}$. More precisely, we mean the following.

Definition 3.1. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be given. We say that $s_f \in \mathbb{IPIR}^n$ *expands* f within $X \in \mathbb{IPIR}^n$, $X \subseteq D$ w.r.t. $\widetilde{x} \in D$ if

$$\forall\, x \in X : \quad f(x) \in f(\widetilde{x}) + s_f \cdot (x - \widetilde{x}). \tag{41}$$

The set s_f depends on $\widetilde{x}$ and $[X]$. We formulate the definition for power set operations, but due to the basic principle of interval operations, the isotonicity, we immediately have

$$(41) \;\Rightarrow\; f(x) \in f(\widetilde{x}) \;\diamondplus\; \diamond\,\{s_f \;\cdot\; ([X] \;\diamondtimes\; \widetilde{x})\} \quad \text{for } [X] := \diamond(X).$$

Using power set operations simplifies statements and proofs, but does not restrict the domain of applicability or the assertions when going to interval operations. In order to formulate a simple implementation scheme we need the following definition.

Definition 3.2. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $X \in \mathbb{PIR}^n$ with $X \subseteq D$ and $\widetilde{x}$ be given. We say that the triplet $(f_c, f_r, s_f) \in \mathbb{PIR} \times \mathbb{PIR} \times \mathbb{PIR}^n$ is a *slope expansion* of f w.r.t. X and $\widetilde{x}$ if

$$f(\widetilde{x}) \in f_c, \quad f(X) \subseteq f_r \quad \text{and} \quad s_f \text{ expands } f \text{ within } X \text{ w.r.t. } \widetilde{x}.$$

Let a function be given by means of a program using constants, variables, control structures, loops and so forth. To be more precise we consider a sequence of statements

$$\begin{array}{ll} 1 \le i \le n: & z_i := x_i \\ n+1 \le i \le m: & \textit{either } z_i := \text{const} \;\text{ or }\; z_i := z_{i_1} \text{ op } z_{i_2} \quad \text{with } \; i_1, i_2 < i, \end{array} \tag{42}$$

where the x_i are the values of the independent variables and z_m is the value of some function. Here, op denotes a monadic or dyadic operator to be specified in a moment. Obviously, many functions can be evaluated using a scheme (42). Next we give a theorem on how to compute a sequence of slope expansions for such functions.

Theorem 3.3. Let $X \subseteq \mathbb{PIR}^n$, $X_i := \{\, \zeta_i \mid \zeta \in X \,\} \in \mathbb{PIR}$ and $\widetilde{x} \in \mathbb{IR}^n$ be given. Then with $f : \mathbb{IR}^n \to \mathbb{IR}$:

$$\begin{array}{ll} (\widetilde{x}_i, X_i, e_i^T) & \text{is a slope expansion for} \quad f(x) \equiv x_i \;(e_i \text{ denotes the ith unit vector}) \\ (c, c, 0) & \text{is a slope expansion for} \quad f(x) \equiv c \end{array}$$

Given slope expansions (f_c, f_r, s_f) and (g_c, g_r, s_g) for functions $f : \mathbb{IR}^n \to \mathbb{IR}$ and $g : \mathbb{IR}^n \to \mathbb{IR}$, respectively w.r.t. X and $\widetilde{x}$ we have

$$\begin{array}{ll} (f_c \pm g_c,\; f_r \pm g_r,\; s_f \pm s_g) & \text{is a slope expansion for} \quad f \pm g, \text{ resp.} \\ (f_c g_c, f_r g_r, f_r s_g + g_c s_f) & \text{is a slope expansion for} \quad f \cdot g, \\ (f_c/g_c, f_r/g_r, (s_f - f_c/g_c \cdot s_g)/g_r) & \text{is a slope expansion for} \quad f/g \end{array}$$

provided the operations are well-defined.

Proof. As an example, we give the calculation for the multiplication, the other operations follow similarly. For $x \in X$ we have

$$\begin{aligned}(f \cdot g)(x) &\in f(x) \cdot [g(\widetilde{x}) + s_g \cdot (x - \widetilde{x})] \\ &\subseteq [f(\widetilde{x}) + s_f \cdot (x - \widetilde{x})] \cdot g(\widetilde{x}) + f(x) \cdot s_g \cdot (x - \widetilde{x}) \\ &\subseteq f(\widetilde{x})\ g(\widetilde{x}) + \{f(x) \cdot s_g + g(\widetilde{x}) \cdot s_f\} \cdot (x - \widetilde{x}) \\ &\subseteq f_c \cdot g_c + \{f_r \cdot s_g + g_c \cdot s_f\} \cdot (x - \widetilde{x}) \end{aligned}$$ ■

This theorem can be found in [54] and similar ideas are in [30]. In a later paper Neumaier proved an extension to transcendental functions [64]. Here he showed for example that

$$(e^{f_c}, e^{f_r}, e^{f_r} \cdot s_f) \quad \text{is a slope expansion for } \exp \circ f \tag{43}$$

provided $f_c \in f_r$. This can be proved by expanding $e^{f(x)}$ w.r.t. $e^{f(\widetilde{x})}$ for some $\zeta \in f(x) \cup f(\widetilde{x}) \subseteq f_r$

$$e^{f(x)} = e^{f(\widetilde{x})} + e^{\zeta} \cdot (f(x) - f(\widetilde{x})) \in e^{f_c} + e^{f_r} \cdot s_f \cdot (x - \widetilde{x})$$

provided $\widetilde{x} \in X$. We can give a sharper slope expansion than (43) not assuming $\widetilde{x} \in X$ by means of the following theorem.

Theorem 3.4. Let $X \in \mathbb{IPIR}^n$ and $\widetilde{x} \in \mathbb{IR}^n$ be given and let (f_c, f_r, s_f) be a slope expansion for $f : \mathbb{IR}^n \to \mathbb{IR}$ w.r.t. X and $\widetilde{x}$. For a function $g : \mathbb{IR} \to \mathbb{IR}$ let $S_g \in \mathbb{IPIR}^n$ expand g within $f(X)$ w.r.t. $f(\widetilde{x})$. Then

$$(g(f_c), g(f_r), S_g \cdot s_f) \quad \text{is a slope expansion for} \quad g \circ f \text{ w.r.t. } X \text{ and } \widetilde{x}.$$

Proof. $(g \circ f)(x) = g(f(x)) = g(f(\widetilde{x})) + S_g \cdot (f(x) - f(\widetilde{x})) \subseteq g(f_c) + S_g \cdot s_f \cdot (x - \widetilde{x})$. ■

For the set S_g we can take the set of secants within $f(X)$, i.e.

$$S_g := \{\ \frac{g(y) - g(\widetilde{y})}{y - \widetilde{y}}\ \mid \widetilde{y} := f(\widetilde{x}),\ y \in f(X),\ y \neq \widetilde{y}\ \}.$$

Defining

$$h(y) := \begin{cases} \{g(y) - g(\widetilde{y})\} \,/\, (y - \widetilde{y}) & \text{for } y \neq \widetilde{y} \\ g'(\widetilde{y}) & \text{otherwise} \end{cases}$$

for twice differentiable g, an extremum of h at some $y \neq \widetilde{y}$ requires

$$g'(y) = \frac{g(y) - g(\widetilde{y})}{y - \widetilde{y}}\ .$$

Then from the mean value theorem we know the existence of some $\xi \in \text{int}(\widetilde{y} \cup y)$ with $g'(y) = g'(\xi)$ and for twice differentiable g some $\zeta \in y \cup \xi \subseteq \widetilde{y} \cup y$ exists with $g''(\zeta) = 0$.

This proves the following theorem.

Theorem 3.5. For a twice differentiable function $g : \mathbb{R} \to \mathbb{R}$, $[X] = [\underline{X}, \overline{X}] \in \mathbb{IR}$, $\widetilde{x} \in \mathbb{R}$ and $g''(x) \neq 0$ for all $x \in \widetilde{x} \cup [X]$ the set S_g defined by

$$S_g := h(\underline{X}) \cup h(\overline{X}) \quad \text{with} \quad h(x) = \begin{cases} \dfrac{g(x) - g(\widetilde{x})}{x - \widetilde{x}} & \text{for } x \neq \widetilde{x} \\ g'(\widetilde{x}) & \text{otherwise} \end{cases}$$

expands g within $[X]$ w.r.t. $\widetilde{x}$.

For many functions this gives a simple way of computing S_g and therefore a slope expansion. For locally non-convex or non-concave functions some case distinctions are necessary.

As an example consider again example (26) with $f(x) = e^x - 2x - 1$ and $\widetilde{x} = 0, [X] = [-2, 1]$. Then

$$S := \frac{e^{-2} - 1}{-2} \cup \frac{e - 1}{1} \subseteq [0.432, 1.719]$$

expands e^x within $[X]$ w.r.t. $\widetilde{x} = 0$ and short computation yields that

$$(0, [-2.865, 5.719],\ [-1.568, -0.281]$$

is a slope expansion of f w.r.t. X and $\widetilde{x}$. This is the same result we used in Chapter 2. We want to stress that here it is obtained *automatically* using Theorems 3.3, 3.4, and 3.5. Computing smaller ranges for the slope is especially interesting in view of the inclusion Theorems 2.1, 2.3, 2.4 and following, because a necessary condition for the corresponding assumptions to hold true is the regularity of all $s \in s_f$. If we use (43) instead, we obtain in our example

$$S \subseteq e^{[X]} \cdot 1 = [e^{-2}, e^1] \subseteq [0.135, 2.719],$$

and a slope expansion

$$(0, [-2.865, 5.719],\ [-1.865, 0.719])$$

for f, with a much bigger slope interval containing zero, thus precluding verification for $[X]$. In fact, it is the same as $f'([X])$. In other words, Theorem 3.5 allows us to perform verification of existence and uniqueness according to algorithm 2.1 for *larger* inclusion sets.

The verification of uniqueness of a zero using Theorem 2.3 requires expansion of the function w.r.t. a whole set $[Y]$ rather than a point $\widetilde{x}$. This can be achieved by replacing $\widetilde{x}$ by $[Y]$. The proof uses the fact that a slope expansion is valid for every $y \in [Y]$. In our

example, we needed a slope expansion w.r.t. $[-2,1]$ and $[Y] := [-0.006, 0.017]$, the latter replacing $\widetilde{x}$. We obtain

$$S := \frac{e^{-2} - e^{[Y]}}{-2 - [Y]} \cup \frac{e^{1} - e^{[Y]}}{1 - [Y]} \subseteq [0.425, 1.755] \tag{44}$$

and

$$[-1.575, -0.245] \text{ expands } f \text{ within } [-2,1] \text{ w.r.t. every } \widetilde{x} \in [Y].$$

This proves regularity of the slope and therefore, as we have seen before, *uniqueness* of the zero of f within $[-2,1]$. The slope could be sharpened slightly by using $\inf([Y])$ and $\sup([Y])$ in (44) instead of the entire $[Y]$. Then, exactly the results as in Chapter 2 are obtained.

For Broyden's example (34) we achieve inclusions up to perturbations $\varepsilon = 0.015$ using slopes defined by Theorems 3.3 and 3.5. In contrast, using the Jacobian, we only obtained inclusions for ε not larger than 0.009. This is still true when using the improved version (40). But even for $\varepsilon = 0.009$ the results obtained by slopes are more accurate. If we denote the inner inclusion by $[S]$ and the outer inclusion by $[T]$, then $\rho_i = w([S]_i) \,/\, w([T]_i)$ is a measure for the quality of inner and outer inclusion of the ith component. In our example we obtained

$$\begin{aligned} &\rho_1 = 0.17,\ \rho_2 = 0.30 \quad && \text{using Jacobians (40)} \\ &\rho_1 = 0.62,\ \rho_2 = 0.70 \quad && \text{using slopes defined by Theorems 3.3 and 3.5.} \end{aligned}$$

We want to stress again that the process of computing a slope expansion can be fully automated by means of predefined operators implementing the rules given in the theorems of this chapter. This is very much in the same spirit as automatic differentiation. We also mention that slopes can be computed in a backward mode, achieving attractive computing times as in the case of automatic differentiation. That means computing an entire slope for a function in n variables takes about 5 times the computing time for one function evaluation, independent of the number of variables. The possibility of easy and automatic computation of slopes make them suitable for practical applications. The computation of slope expansions for functions $f : \mathbb{R}^n \to \mathbb{R}^n$ can be performed for every component function individually.

4. Dense systems of linear equations

Consider the linear system

$$Ax = b \quad \text{for} \quad A \in \mathrm{M}_{nn}(\mathbb{R}),\ b \in \mathbb{R}^n$$

with dense system matrix A. If we regard it as a zero finding problem of $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = Ax - b$ we can apply the theorems of Chapters 2 and 3. In the linear case the

slope function is constant and equal to the Jacobian, namely A itself:

$$f(x) = f(\widetilde{x}) + A \cdot (x - \widetilde{x}) \quad \text{for all } x, \widetilde{x} \in \mathbb{IR}^n. \tag{45}$$

Therefore, as an application of Theorem 2.1, we obtain the following result.

Theorem 4.1. Let $A \in \mathrm{M}_{nn}(\mathbb{IR})$, $b \in \mathbb{IR}^n$ be given, $R \in \mathrm{M}_{nn}(\mathbb{IR})$, $[Y] \in \mathbb{IIR}^n$, $\widetilde{x} \in \mathbb{IR}^n$ and define

$$Z := R \cdot (b - A\widetilde{x}) \in \mathbb{IR}^n, \; C := I - R \cdot A \in \mathrm{M}_{nn}(\mathbb{IR}).$$

Define $[V] \in \mathbb{IIR}^n$ by means of the following Einzelschrittverfahren for $1 \leq i \leq n$:

$$V_i := \{\diamond(Z + C \cdot [U])\}_i \quad \text{where} \quad [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T. \tag{46}$$

If

$$[V] \subsetneqq [Y], \tag{47}$$

then R and A are regular and the unique solution $\widehat{x} = A^{-1}b$ of $Ax = b$ satisfies $\widehat{x} \in \widetilde{x} + [V]$.

Proof. Applying Theorem 2.1 to $f(x) = Ax - b$ yields regularity of R and A and the existence of some $\widehat{x} \in \widetilde{x} + [V]$ with $f(x) = 0$. $\widehat{x}$ is unique because of the regularity of A. ■

In the case of linear systems we do not need to use the powerful Theorem 2.1 but can proceed in a more elementary way. Moreover, convexity of the inclusion set is not necessary.

Theorem 4.2. Let $A \in \mathrm{M}_{nn}(\mathbb{IR})$, $b \in \mathbb{IR}^n$ be given, $R \in \mathrm{M}_{nn}(\mathbb{IR})$, $\emptyset \neq Y \subseteq \mathbb{IR}^n$ closed and bounded, $\widetilde{x} \in \mathbb{IR}^n$ and define

$$Z := R \cdot (b - A\widetilde{x}) \in \mathbb{IR}^n, \quad C := I - RA \in \mathrm{M}_{nn}(\mathbb{IR}).$$

If

$$Y^* := Z + C \cdot Y \subseteq \mathrm{int}(Y), \tag{48}$$

then R and A are regular and the unique solution $\widehat{x} = A^{-1}b$ of $Ax = b$ satisfies $\widehat{x} \in \widetilde{x} + Y^*$.

Proof. (48) and Lemma 1.1 imply $\rho(C) < 1$, and therefore regularity of R and A. $g(x) := Z + C \cdot x$ is a contractive mapping which maps Y into itself. Hence, the Fixed Point Theorem of Banach-Weissinger [33] implies the existence of a unique fixed point $\widehat{y} \in Y$ of g which is $\widehat{x} - \widetilde{x}$. ■

4.1. Optimality of the inclusion formulas

When applying Theorem 4.1, we have a sufficient criterion for $\widetilde{x} + [V]$ to contain the solution of $Ax = b$. The information available is the approximate inverse $R \approx A^{-1}$ and the approximate solution $A\widetilde{x} \approx b$. Given R and $C = I - RA$ the iteration given in Theorem 1.5 will produce some $[X] \in \mathbb{IIR}^n$ satisfying $Z + C \cdot [X] \subsetneqq [X]$ if and only if $\rho(|C|) < 1$. Thus from a theoretical point of view the quality of $\widetilde{x}$ is not important. The only and important information for the "behaviour" of the iteration is R. Thus in order to judge the quality of Theorem 4.1 it suffices to consider $\widetilde{x} = 0$:

$$R \cdot b + \{I - RA\} \cdot [X] \subsetneqq [X]. \tag{49}$$

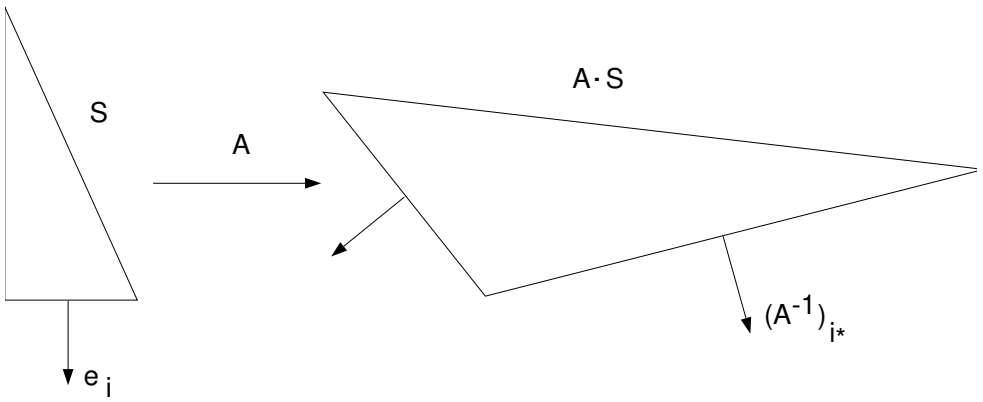
Below we will show that (49) makes "optimal" use of the available information, namely the approximate inverse R. Optimality is shown by geometrical considerations. Let *ch* denote the convex hull of a set and let

$$S := ch(\underline{x}, \underline{x} + \varepsilon_1 e_1, \ldots, \underline{x} + \varepsilon_n e_n)$$

be a standard simplex, i.e. a simplex having main edges parallel to the coordinate axes e_i. Then $A \cdot S = ch(A\underline{x}, A\underline{x} + \varepsilon_1 A_1, \ldots, A\underline{x} + \varepsilon_n A_n)$ where A_i denotes the i-th column of A. $A \cdot S$ is a general simplex. If we could show

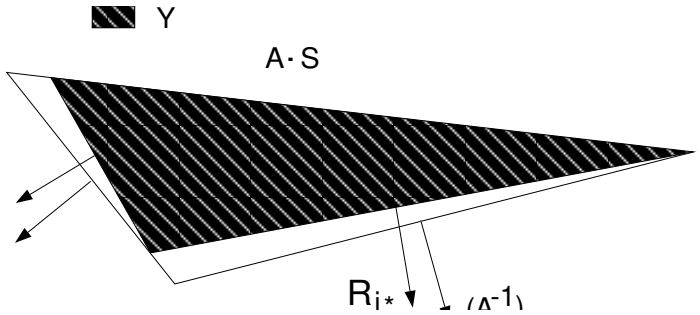
$$b \in A \cdot S, \quad \text{then} \quad b \in \{\, A \cdot x \mid x \in S \,\} \quad \text{and} \quad \exists\, \widehat{x} \in S : A\widehat{x} = b,$$

thus S would contain a solution of $Ax = b$. In order to show $b \in A \cdot S$ we need an *inner* inclusion of $A \cdot S$. Note that an *outer* inclusion, i.e. some $[X] \in \mathbb{IIR}$ with $A \cdot S \subseteq [X]$, can easily be computed.



The normal vectors of the hyperplanes bounding S are the unit vectors e_i [there is a $(n+1)$-st one, but this is not important for the following considerations]. The normal vector of a hyperplane bounding $A \cdot S$ must be normal to $A \cdot e_j$, for all $j \neq i$. So this is the i-th row $(A^{-1})_{i*}$ of A^{-1} for regular A and *approximately* the i-th row R_{i*} of R.

We need to calculate an inner estimation of $A \cdot S$. Thus from a geometrical point of view we could try to use the rows R_{i*} of R to fill $A \cdot S$ from the interior as in the following diagram.



The hyperplanes defined by the normal vectors R_{i*} are "put" in the vertices of $A \cdot S$. If the so defined "interior" Y, the shaded area, contains b then

$$b \in Y \subseteq A \cdot S \quad \text{implies} \quad \exists\, \widehat{x} \in S : \; A\widehat{x} = b.$$

This vague description of what is intended has been formulated in mathematical terms by Jansson [37]. Max of a matrix denotes the column vector of maxima of the rows.

Theorem 4.3 (Jansson). Let A, R be $n \times n$ matrices, $C := R \cdot A$, $b, \underline{x}, \varepsilon \in \mathbb{R}^n$ with $\varepsilon > 0$ and $S := ch(\{\underline{x}, \underline{x} + \varepsilon_1 e_1, \ldots, \underline{x} + \varepsilon_n e_n\})$. With

$$(t_1, \ldots, t_n)^T := C\underline{x} + \mathrm{Max}\Big\{(C - \mathrm{Diag}(C)) \cdot \mathrm{Diag}(\varepsilon)\Big\}$$

$$t_{n+1} := (\varepsilon^{-1})^T C\underline{x} + \mathrm{Min}\{(\varepsilon^{-1})^T C \cdot \mathrm{Diag}(\varepsilon)\}$$

the simplex

$$Y := \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{l} r^i y \geq t_i, \;\; i = 1, \ldots, n \\ (\varepsilon^{-1})^t R y \leq t_{n+1} \end{array} \right\} \tag{50}$$

satisfies $Y \subseteq A \cdot S$. Moreover, every simplex

$$\widetilde{Y} := \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{l} r^i y \geq \widetilde{t}_i, \;\; i = 1, \ldots, n \\ (\varepsilon^{-1})^t R y \leq \widetilde{t}_{n+1}, \end{array} \right\}$$

with $\widetilde{t}_i \in \mathbb{R}$ for $i = 1, \ldots, n+1$, $\widetilde{t}_i \neq t_i$ for at most one i with $\widetilde{Y} \subseteq A \cdot S$ is contained in Y.

The last statement in Theorem 4.3 states the *geometrical optimality*. Since $Y \subseteq A \cdot S$, the standard simplex S defined by (50) contains a solution of $Ax = b$ if $b \in Y$. This solution is also unique, as has been proved by Jansson [37].

Theorem 4.4 (Jansson). Let A, R be $n \times n$ matrices, $C := R \cdot A$, $b, \underline{x}, \varepsilon \in \mathbb{R}^n$ with $\varepsilon > 0$. If the inequalities

$$\mathrm{Rb} > C\underline{x} + \mathrm{Max}\Big\{(C - \mathrm{Diag}(C)) \cdot \mathrm{Diag}(\varepsilon)\Big\} \tag{51}$$

$$(\varepsilon^{-1})^T Rb < (\varepsilon^{-1})^T C\underline{x} + \mathrm{Min}\{(\varepsilon^{-1})^T C \cdot \mathrm{Diag}(\varepsilon)\} \tag{52}$$

are valid, then R and A are nonsingular, and the unique solution $\widehat{x}$ of $Ax = b$ is contained in the standard simplex $S = \mathrm{ch}(\{\underline{x}, \underline{x} + \varepsilon_1 e_1, \ldots, \underline{x} + \varepsilon_n e_n\})$.

The interesting fact is now that conditions (51), (52), which make *optimal* use of the given information R, can be shown to be *equivalent* to (49) with a minor technical condition. The following theorem has been given in [81].

Theorem 4.5. With the assumptions of Theorem 4.4 and R scaled such that $\mathrm{diag}(R \cdot A) = 1$,

$$\text{(51) and (52) together are equivalent to} \quad R \cdot b + \{I - R \cdot A\} \cdot S \subseteq \mathrm{int}(S).$$

Beside the geometrical optimality Theorem 4.5 gives an important information on the choice of R, namely to scale it by left multiplication of a diagonal matrix such that $\mathrm{diag}(RA) \approx I$. For practical applications floating point multiplication suffices.

4.2. Inner inclusions and sensitivity analysis

The other theorems for systems of nonlinear equations can be specialized to linear equations as well. Because of their mutual importance we formulate the theorems for estimating the sensitivity explicitly, that is for finite perturbations of the input data with inner inclusions and ε-perturbations for $\varepsilon \to 0$. Theorem 2.4 yields the following.

Theorem 4.6. Let $[A] \in \mathbb{IM}_{nn}(\mathbb{R})$, $[b] \in \mathbb{IIR}^n$ be given, $R \in \mathrm{M}_{nn}(\mathbb{R})$, $[Y] \in \mathbb{IIR}^n$, $\widetilde{x} \in \mathbb{R}^n$ and define

$$[Z] := \diamond\{R \cdot ([b] - [A] \cdot \widetilde{x})\} \in \mathbb{IIR}^n, \quad [\boldsymbol{C}] := \diamond\{I - R \cdot [A]\} \in \mathbb{IM}_{nn}(\mathbb{R}).$$

Define $[V] \in \mathbb{IIR}^n$ by means of the following Einzelschrittverfahren

$$1 \le i \le n : V_i := \{\diamond([Z] + [\boldsymbol{C}] \cdot [U])\}_i \quad \text{where } [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T.$$

If

$$[V] \subsetneqq [Y],$$

then R and every matrix $A \in [A]$ are regular and for every $A \in [A]$, $b \in [b]$ the unique solution $\widehat{x} = A^{-1}b$ of $Ax = b$ satisfies $\widehat{x} \in \widetilde{x} + [V]$. Define the solution set Σ by

$$\Sigma([A],[b]) := \{\, x \in \mathbb{IR}^n \mid \exists\, A \in [A]\ \exists\, b \in [b] : Ax = b \,\}.$$

Then with $[\Delta] := \diamond\{[\boldsymbol{C}] \cdot [V]\} \in \mathbb{IIR}^n$ the following estimations hold true for every $1 \le i \le n$:

$$\begin{aligned} \widetilde{x}_i + \inf([Z]_i) + \sup([\Delta]_i) &\ge \inf_{\sigma\in\Sigma} \sigma_i \quad \text{and} \\ \widetilde{x}_i + \sup([Z]_i) + \inf([\Delta]_i) &\le \sup_{\sigma\in\Sigma} \sigma_i. \end{aligned}$$

Even for linear systems the solution complex $\Sigma = \Sigma([A],[b])$ need not be convex. Moreover, Rohn and Poljak[68], and Rohn [73] have shown that the computation of $\diamond\Sigma$, the interval hull of the true solution set Σ, is NP-hard. Nevertheless, Theorem 4.6 gives inner and outer bounds on Σ, where the quality is determined by the width of Δ. This in turn is the product of small quantities provided the width of $[A]$ is not too big.

For the application of Theorem 4.6 we need an *inner* inclusion of $Z = R \cdot ([b] - [A] \cdot \widetilde{x})$. Fortunately, this is not too difficult. For intervals $[b]$ and $[A]$,

$$\{\, b - A\widetilde{x} \mid b \in [b], A \in [A] \,\} = [b] \ \diamondsuit\ [A] \ \diamondsuit\ \widetilde{x} \tag{53}$$

holds. In most theorems throughout this paper it is not important to distinguish between interval and power set operations, as has been explained in the introduction. Here, we need inner inclusions. (53) can be seen by expanding the r.h.s. componentwise and observing that every interval component of $[A]$ and $[b]$ occurs exactly once. That means no overestimation is introduced; power set operations and interval operations yield identical results. This changes when multiplying by R, because the hyperrectangle $[b] - [A] \cdot \widetilde{x}$ becomes a parallel epiped under the linear mapping R. For example

$$R = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},\ [v] = [b] - [A] \cdot \widetilde{x} = \begin{pmatrix} [1,2] \\ [-2,-1] \end{pmatrix} \text{ with } R \diamond [v] = \begin{pmatrix} [-1,\ 1] \\ [-4,-2] \end{pmatrix}$$

but no $v \in [v]$ exists with $R \cdot v = (-1,-4)^T$. However, the interval vector $R \ \diamond\ [v]$ is still sharp:

$$[X] \in \mathbb{IIR}^n \text{ with } R \cdot [v] = \{\, R \cdot v \mid v \in [v] \,\} \subseteq [X] \Rightarrow R \ \diamond\ [v] \subseteq [X],$$

i.e. $R \diamond [v]$ is the interval hull of $R \cdot [v]$. This can also be seen by expanding $R \cdot [v]$ componentwise, and observing that for every component $1 \le i \le n$, every interval component of $[v]$ occurs exactly once. Thus every *component* $(\ \diamond\{R \cdot ([b] - [A] \cdot \widetilde{x})\})_i$ is sharp as required by Theorem 4.6. If we go to rounded arithmetic we have to compute

$[b] - [A] \cdot \widetilde{x}$ as well as the product by R with inward and outward rounding. If we can use a precise dot product as proposed by Kulisch [55], [56], this task is simplified because we obtain the *exact* components of $[b] - [A] \cdot \widetilde{x}$, which we only have to round inward and outward.

The bounds obtained by using Theorem 4.6 are essentially sharp as long as $[\Delta]$ does not become too big. In turn, $[\Delta]$ is the product of small numbers as long as the width of $[A]$ does not become too big.

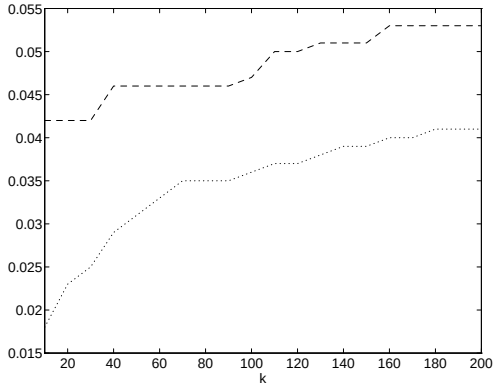
A frequently used heuristic approach to error and sensitivity analysis is to run a calculation several times with varying input data. The number of figures of the solution which agree in all calculations is taken to indicate the precision or width of the solution set. For example, Bellmann [14] writes: "Considering the many assumptions that go into the construction of mathematical models, the many uncertainties that are always present, we must view with some suspicion at any particular prediction. One way to contain confidence is to test the consequences of various changes in the basic parameters".

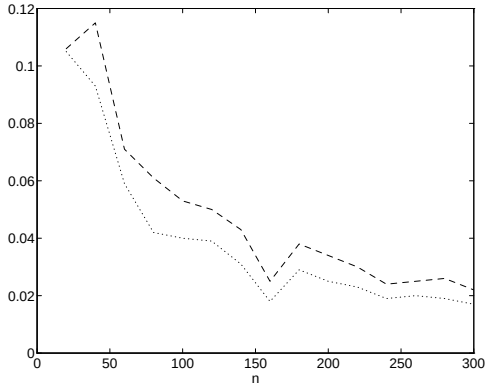
Inner bounds on $\Sigma([A], [b])$ obtained by Monte Carlo like methods may be much weaker than those computed by Theorem 4.6. Consider $A \in \mathrm{M}_{nn}(\mathbb{IR})$, $b \in \mathbb{IR}^n$ with randomly chosen components uniformly distributed within $[-1, 1]$.We set $[A] := A \cdot [1-e, 1+e]$ and $[b] := b \cdot [1-e, 1+e]$ for $e = 10^{-5}$. Then we use the following Monte Carlo approach.

$$\Sigma_{MC} := \emptyset;\ \text{for } i = 1 \text{ to } k \text{ do } \{\text{Take } A \in \partial[A],\ b \in \partial[b] \text{ randomly},\ \widehat{x} := A^{-1}b \text{ and set } \Sigma_{MC} = \diamondsuit(\Sigma_{MC} \cup \widehat{x})\}.$$

Thus we take only linear systems with A, b on the boundary of $[A]$ and $[b]$ in order to maximize Σ_{MC}; however, there are 2^{n^2+n} such A and b. (Remember that the exact computation of $\diamond\ \Sigma([A], [b])$ is NP-hard).

Σ_{MC} is an inner inclusion of $\Sigma := \diamond\ \Sigma([A], [b])$, that is $\Sigma_{MC} \subseteq \Sigma$. We may ask for the difference in width between Σ_{MC} and $\diamond\ \Sigma$. In all our examples the ratio of the width of the inner inclusion to the width of the outer inclusion computed by Theorem 4.6 was greater than 0.99. In other words we know $w(\Sigma)$ with an error less than 1 % using Theorem 4.6. Define $r \in \mathbb{IR}^n$ by $r_i := w(\Sigma_{MC})_i \ / \ w(\Sigma)_i$ and $r_{\max} := \max_i r_i$, $r_{av} := \sum_i r_i/k$. r depends on the number k of samples used to compute Σ_{MC}. In the first diagram we display $r_{\max}$ (dashed) and r_{av} (dotted) for a fixed (dense random) matrix of dimension $n = 100$ for different values of k, in the second plot we always use $k = n$ samples for every (dense random) matrix up to dimension $n = 300$. In other words n linear systems with n unknowns have been solved in the second graph to obtain Σ_{MC}, i.e. $\frac{1}{3}\, n^4$ operations where the computation of Σ requires $3n^3$ operations.





We see that for increasing n the underestimation of Σ by Σ_{MC} goes rapidly below 5 % although we used n samples (sic!) for computing Σ_{MC}.

Next we give an example for a larger linear system with full matrix. Consider the Hadamard matrix $A \in \mathbb{R}^{n\times n}$ with (cf. [25], example 3.14)

$$A_{ij} := \left(\frac{i+j}{p}\right) \text{ and } p = n+1 \quad \text{for } n = 1008, \tag{54}$$

and right hand side b such that the true solution $\widehat{x} = A^{-1}b$ satisfies $\widehat{x}_i = (-1)^{i+1}/i$. We introduce relative tolerances of 10^{-5} and define

$$[A] := A \cdot [1-e, 1+e] \text{ and } [b] := b \cdot [1-e, 1+e] \text{ with } e = 10^{-5},$$

and will include $\Sigma([A],[b]) = \{\, x \in \mathbb{R}^n \mid \exists\, A \in [A]\ \exists\, b \in [b] : Ax = b\}$. The computation is performed in *single precision* ($\sim$ 7 decimals). The following results are obtained for the inner inclusion $[X]$ and outer inclusion $[Y]$, with $[X] \subseteq \diamond\Sigma([A],[b]) \subseteq [Y]$ (see [42]).

inner and outer inclusions for some components						$\frac{w([X])}{w([Y])}$
[0.999 873,	1.000 127]	$\subseteq$	$\Sigma([A],[b])_1$	$\subseteq$	[0.999 869, 1.000 131]	0.96980
[−0.500 127,	−0.499 873]	$\subseteq$	$\Sigma([A],[b])_2$	$\subseteq$	[−0.500 131, −0.499 869]	0.96975
[0.333 206,	0.333 460]	$\subseteq$	$\Sigma([A],[b])_3$	$\subseteq$	[0.333 203, 0.333 464]	0.96978
			...			
[−0.001 121,	−0.000 867]	$\subseteq$	$\Sigma([A],[b])_{1006}$	$\subseteq$	[−0.001 125, −0.000 863]	0.96979
[0.000 866,	0.001 120]	$\subseteq$	$\Sigma([A],[b])_{1007}$	$\subseteq$	[0.000 862, 0.001 124]	0.96981
[−0.001 119,	−0.000 865]	$\subseteq$	$\Sigma([A],[b])_{1008}$	$\subseteq$	[−0.001 123, −0.000 861]	0.96977

In the last column the ratio of the width of the inner and outer inclusion is given in order to judge the quality. The worst of these ratios is achieved in component 116 with a value 0.96967. This means that we know the size of the solution complex $\Sigma([A],[b])$ up to an accuracy of about 3 %.

There are other methods for computing an inclusion of systems of linear equations [43], [29], [52], [66]. Because of their underlying principle, these methods require strong regularity for the system matrix $[A]$. Therefore Theorem 1.5 implies that the scope of applicability cannot be larger than the one of Theorem 4.7 together with an iteration with ε-inflation as demonstrated by Theorem 1.5. For example, Neumaier [66] uses $R \approx \mathrm{mid}([A])^{-1}$, $\langle R \cdot [A] \rangle \cdot \widetilde{u} \approx |R \cdot [b]| + \varepsilon$ and assumes $\alpha > 0$ to be given with

$$\langle R \cdot [A] \rangle \cdot \widetilde{u} \geq \alpha \cdot |R \cdot [b]|. \tag{55}$$

Here, $\langle \cdot \rangle$ denotes Ostrowski's comparison matrix [64]. If $[A]$ is strongly regular, then $\alpha^{-1} \cdot \widetilde{u} \cdot [-1, 1]$ is an inclusion of $\Sigma([A], [b])$. In [81] it has been shown that replacing $\geq$ by $>$ in (4.10) already implies strong regularity of $[A]$ and, moreover, $[X] := \alpha^{-1} \cdot \widetilde{u} \cdot [-1, 1]$ satisfies (55). Although having in principle the same scope of applicability, the methods differ in speed and the quality of the inclusion. The differences are marginal; it seems for large widths Neumaier's method is advantageous, whereas for moderate widths it is the other way around. For numerical results see [81].

All those methods are *by their underlying principle* not applicable for matrices $[A]$ not being strongly regular. The only method which can go across this border is based on Theorem 1.8, and will be discussed in Chapter 5.

Next we go to ε-perturbations of a linear system $Ax = b$ for $\varepsilon \to 0$.

Theorem 4.7. Let the assumption of Theorem 4.1 hold true implying $\widehat{x} := A^{-1}b \in \widetilde{x} + [V]$. Let $A^* \in \mathrm{M}_{nn}(\mathbb{IR})$, $b^* \in \mathbb{IR}^n$, $A^* \geq 0$, $b^* \geq 0$ be given and define

$$u := |R| \cdot (b^* + A^* \cdot |\widehat{x}|) \quad \text{and} \quad w := |I - RA| \cdot d([V]). \tag{56}$$

Then

$$\phi := \max_i \left\{ \frac{u_i}{(d([V]) - w)_i} \right\}$$

is well defined. The componentwise sensitivity of $\widehat{x}$ w.r.t. perturbations weighted by A^* and b^* defined by

$$\mathrm{Sens}_k(\widehat{x}, A, b, A^*, b^*) := \lim_{\varepsilon \to 0^+} \max \left\{ \frac{|\widehat{x} - \widetilde{x}|_k}{\varepsilon} \;\middle|\; \widetilde{A}\widetilde{x} = \widetilde{b} \right\}$$

for some $\widetilde{A}$, $\widetilde{b}$ with $|A - \widetilde{A}| \leq \varepsilon \cdot A^*, |b - \widetilde{b}| \leq \varepsilon \cdot b^*$ satisfies for $1 \leq k \leq n$

$$\mathrm{Sens}_k(\widehat{x}, A, b, A^*, b^*) \in u \pm \phi \cdot w. \tag{57}$$

Proof. Apply Theorem 2.5 to $f : \mathbb{IR}^{n^2+n} \times \mathbb{IR}^n \to \mathbb{IR}^n$ defined by $f((A, b), x) := Ax - b$ with $n^2 + n$ parameters A_{ij} and b_i, $1 \leq i, j \leq n$, each parameter occuring at most once in

every equation. ■

In a practical application, $\widehat{x}$ in (56) can be replaced by $\widetilde{x} + [V]$. Theorem 4.7 also confirms a result by Skeel for the exact value of the sensitivity of the solution of a linear system [86]. It is

$$\mathrm{Sens}(\widehat{x}, A, b, A^*, b^*) = |A^{-1}| \cdot (b^* + A^* \cdot |\widehat{x}|) \tag{58}$$

which is included in (57). Skeel states this result for relative perturbations $A^* = |A|$, $b^* = |b|$. The advantage of (57) and (58) is the freedom we have for the perturbations, especially specific components may be kept unaltered.

The above defined sensitivity is the *absolute* sensitivity of the solution $\widehat{x}$. The relative sensitivity, i.e. the relative change of the solution is

$$\mathrm{Sens}_{rel}(\widehat{x}, A, b, A^*, b^*) := \left(\frac{\mathrm{Sens}_i(\widehat{x}, A, b, A^*, b^*)}{|\widehat{x}|_i} \right)$$

provided $\widehat{x}_i \neq 0$. As an example, consider

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2\varepsilon & 2\varepsilon \\ 1 & 2\varepsilon & -\varepsilon \end{pmatrix}, \; b = \begin{pmatrix} 3 + 3\varepsilon \\ 6\varepsilon \\ 2\varepsilon \end{pmatrix} \quad \text{with } \widehat{x} = A^{-1}b = \begin{pmatrix} \varepsilon \\ 1 \\ 1 \end{pmatrix} \tag{59}$$

given by Fox and Kahan [27]. Then for relative perturbations $A^* = |A|$, $b^* = |b|$ we have

$$\mathrm{Sens}_{rel}(\widehat{x}, A, b, |A|, |b|) = (9.6, 4.8, 6.0)^T, \tag{60}$$

i.e. a very stable solution whereas for absolute perturbations $A^* = (1)$, $b^* = (1)$ we get

$$\mathrm{Sens}_{rel}(\widehat{x}, A, b, (1), (1)) = (1.8/\varepsilon, 0.9/\varepsilon, 1.8/\varepsilon) \tag{61}$$

If the data of the linear system is given with reasonable precision where the small components may result from the chosen units, we have a stable solution. This is no longer true if the data is only given in absolute precision. The condition number $\|A\| \cdot \|A^{-1}\| \approx 3.6/\varepsilon$ does not reflect this behaviour. Theorem 4.7 allows computation of enclosures of the sensitivities (60), (61) without computing an inclusion of A^{-1}. For example, for $\varepsilon = 2^{-30} \approx 10^{-10}$ we obtain at least 7 correct digits for the sensitivities (60), (61) when computing in double precision.

4.3. Data dependencies in the input data

When applying Theorem 4.6 to the solution of an interval linear system with matrix $[A] \in \mathbb{IM}_{nn}(\mathbb{IR})$ and right hand side $[b] \in \mathbb{IIR}^n$, i.e. computing inner and outer bounds for

$$\Sigma([A], [b]) := \{ \, x \in \mathbb{IR}^n \mid \; \exists \, A \in [A] \; \exists \, b \in [b] : Ax = b \, \} \tag{62}$$

we implicitly assumed A and b to vary componentwise *independently* within $[A]$ and $[b]$. In practical applications this need not to be the case. We may have further constraints on the matrices within $[A]$ possibly in connexion with $[b]$. A simple example is symmetric matrices, that is only $A \in [A]$ with $A = A^T$ are considered, and we define

$$\Sigma^{sym}([A],[b]) := \{\, x \in \mathbb{IR}^n \mid \exists\, A \in [A]\ \exists\, b \in [b] : A = A^T \text{ and } Ax = b \,\}. \tag{63}$$

Obviously $\Sigma^{sym}([A],[b]) \subseteq \Sigma([A],[b])$. Another example are Toeplitz matrices which belong to the larger class of *persymmetric* matrices, the latter being characterized by $A = E\ A^T\ E$ where $E = [e_n, \ldots, e_1]$ is an $n \times n$ permutation matrix. Persymmetric matrices are symmetric w.r.t. the northeast-southwest diagonal. As an example of linear systems that also have dependencies in the right hand side, we mention the Yule-Walker problem [24], which is

$$T_n(p) \cdot y = -p \quad \text{for} \quad p \in \mathbb{IR}^n$$

where $T_n(p)$ is defined by

$$T_n(p) = \begin{pmatrix} 1 & p_1 & p_2 & \ldots & p_{n-1} \\ p_1 & 1 & p_1 & \ldots & p_{n-2} \\ p_2 & p_1 & 1 & \ldots & p_{n-3} \\ & & \ldots & & \\ p_{n-1} & p_{n-2} & p_{n-3} & \ldots & 1 \end{pmatrix}. \tag{64}$$

Those problems arise in conjunction with linear prediction problems. $T_n(p)$ does not depend on p_n. We define for $[p] \in \mathbb{IIR}^n$

$$\Sigma([p]) = \{\, x \in \mathbb{IR}^n \mid\ \exists\, p \in [p] : T_n(p) \cdot x = -p \,\}. \tag{65}$$

Replacing the p_i by $[p]_i$ in (64) we have $\Sigma([p]) \subseteq \Sigma(T_n([p]), -[p])$. The general inclusion (62) may yield large overestimations compared to (63) or (65).

Computing inclusions for $\Sigma([A],[b])$ with data dependencies was first considered by Jansson [39]. He treated symmetric and skew-symmetric matrices as well as dependencies in the right hand side. In the following, we give a straightforward generalization to affine-linear dependencies of the matrix and the r.h.s. on a set of parameters $p \in \mathbb{IR}^k$. This covers all of the above-mentioned problems including symmetric, persymmetric, and Toeplitz systems and the Yule-Walker problem.

For a parameter vector $p \in \mathbb{IR}^k$ consider linear systems $A(p) \cdot x = b(p)$ where $A(p) \in \mathrm{M}_{nn}(\mathbb{IR})$ and $b(p) \in \mathbb{IR}^n$ depend on p. If p is allowed to vary within a range $[p] \in \mathbb{IIR}^n$, we may ask for *outer and inner inclusions* of the set of solutions of all $A(p) \cdot x = b(p)$, $p \in [p]$

$$\Sigma(A(p), b(p), [p]) := \{\, x \in \mathbb{IR}^n \mid\ \exists\, p \in [p] : A = A(p),\ b = b(p) \text{ and } Ax = b \,\}. \tag{66}$$

Consider $A(p)$, $b(p)$ depending linearly on p, that is

$$\begin{array}{l}\text{There are vectors } w(i,j) \in \mathbb{R}^k \text{ for } 0 \le i \le n,\ 1 \le j \le n \text{ with} \\ \{A(p)\}_{ij} = w(i,j)^T \cdot p \quad \text{and} \quad \{b(p)\}_j = w(0,j)^T \cdot p.\end{array} \tag{67}$$

Each individual component $\{A(p)\}_{ij}$ and $\{b(p)\}_j$ of $A(p)$ and $b(p)$ depends linearly on p. For example, for symmetric matrices we could use

$$\{A(p)\}_{ij} := \begin{cases} p_{ij} & \text{for } i < j \\ p_{ii} & \text{for } i = j \\ p_{ji} & \text{for } i > j \end{cases} \quad \text{and } \{b(p)\}_j := p_{0j} \tag{68}$$

or for the Yule-Walker problem

$$\{A(p)\}_{ij} := p_{|i-j|}, \quad \text{and } \{b(p)\}_j := -p_j \quad \text{with } p_0 := 1 \tag{69}$$

Now Theorem 2.4 or, with obvious modifications, Theorem 4.6 can be applied directly, even for nonlinear dependencies of A, b w.r.t. p. In order to obtain sharp inclusions, the problem is to obtain sharp bounds for $Z = -R \cdot f([p], \widetilde{x}) = R \cdot \{b([p]) - A([p]) \cdot \widetilde{x}\}$, because straightforward evaluation causes overestimation. Fortunately, linear dependencies (67) of $A(p)$ and $b(p)$ allow a sharp inner and outer estimation of Z.

Theorem 4.8. Let $A(p) \cdot x = b(p)$ with $A(p) \in \mathrm{M}_{nn}(\mathbb{R})$, $b(p) \in \mathbb{R}^n$, $p \in \mathbb{R}^k$ be a parametrized linear system, where $A(p)$, $b(p)$ are given by (67). Let $R \in \mathrm{M}_{nn}(\mathbb{R})$, $[Y] \in \mathbb{IR}^n$, $\widetilde{x} \in \mathbb{R}^n$ and define $[Z] \in \mathbb{IR}^n$, $[\boldsymbol{C}] \in \mathbb{IM}_{nn}(\mathbb{R})$ by

$$Z_i := \left(\sum_{j,\nu=1}^{n} \left\{ R_{ij} \cdot (w(0,j) - x_\nu \cdot w(j,\nu)) \right\}^T \right) \cdot [p], \quad \boldsymbol{C} := I - R \cdot A([p]). \tag{70}$$

Define $[V] \in \mathbb{IR}^n$ by means of the following Einzelschrittverfahren

$$1 \le i \le n: \ V_i := \{\diamond([Z] + [\boldsymbol{C}] \cdot [U])\}_i \quad \text{where } [U] := (V_1, \ldots, V_{i-1}, Y_i, \ldots, Y_n)^T.$$

If

$$[V] \subsetneqq [Y],$$

then R and every matrix $A(p), p \in [p]$ is regular, and for every $A = A(p)$, $b = b(p)$ with $p \in [p]$ the unique solution $\widehat{x} = A^{-1}b$ of $Ax = b$ satisfies $\widehat{x} \in \widetilde{x} + [V]$. Define the solution set Σ by

$$\Sigma := \Sigma(A(p), b(p), [p]) = \{\, x \in \mathbb{R}^n \mid \ \exists\, p \in [p] : A = A(p), b = b(p) \text{ and } Ax = b \,\}.$$

Then with $[\Delta] := \diamond([\boldsymbol{C}] \cdot [V]) \in \mathbb{IR}^n$ the following inner and outer estimations hold for every $1 \le i \le n$:

$$\begin{array}{l}\widetilde{x}_i + \inf([Z]_i) + \sup([\Delta]_i) \ge \inf\limits_{\sigma \in \Sigma} \ \sigma_i \quad \text{and} \\ \widetilde{x}_i + \sup([Z]_i) + \inf([\Delta]_i) \le \sup\limits_{\sigma \in \Sigma} \ \sigma_i.\end{array}$$

Proof. Consider $f : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^n$ with $f(p,x) := A(p) \cdot x - b(p)$. Then application of Theorem 2.4 completes the proof if we can show $\diamond(-R \cdot f([p], \widetilde{x})) = [Z]$ with $[Z]$ defined by (70). It is for $1 \le i \le n$

$$\begin{aligned}
&\Big\{ \diamond\,(-R \cdot f([p], \widetilde{x}))\Big\}_i = \Big\{ \diamond\,\{-R \cdot f(p, \widetilde{x}) \mid p \in [p]\}\Big\}_i = \\
&= \Big[\diamond\,\{R \cdot (b(p) - A(p) \cdot \widetilde{x}) \mid p \in [p]\}\Big]_i \\
&= \diamond\Big\{ \sum_{j,\nu=1}^{n} R_{ij} \cdot (w(0,j)^T \cdot p - (w(j,\nu)^T \cdot p) \cdot \widetilde{x}_\nu) \;\Big|\; p \in [p]\Big\} \\
&= \diamond\Big\{ \sum_{j,\nu=1}^{n} \{R_{ij} \cdot (w(0,j) - \widetilde{x}_\nu \cdot w(j,\nu))\}^T \cdot p \;\Big|\; p \in [p]\Big\} \\
&= \Big(\sum_{j,\nu=1}^{n} \{R_{ij} \cdot (w(0,j) - \widetilde{x}_\nu \cdot w(j,v))\}^T\Big) \cdot [p].
\end{aligned}$$

The last equality holds since every component p_i occurs at most once in the previous expression. ■

We illustrate Theorem 4.8 with our previous two examples (68) and (69). Only the determination of $[Z]$ is important. For *symmetric systems* as in (68) we have

$$Z_i := \sum_{j=1}^{n} R_{ij} \cdot [b]_j - \sum_{j=1}^{n} \sum_{\nu=j+1}^{n} (R_{ij} \cdot x_\nu + R_{i\nu} \cdot x_j) \cdot [A]_{j\nu} - \sum_{j=1}^{n} R_{ij} x_j \cdot [A]_{jj}.$$

Here we used $[b]_j$ and $[A]_{j\nu}$, $j \le \nu$ as parameters, that is only the upper triangle of $[A]$ including diagonal. The formula can be derived by computing the components of $[Z]$ following the lines of the proof of Theorem 4.8. The main point is that in every component every parameter occurs at most once. For the Yule-Walker example we obtain, after short computation,

$$\begin{aligned}
&\sum_{j,\nu=1}^{n} R_{ij}(b(p)_j - A(p)_{j\nu} \cdot x_\nu) = -\sum_{j,\nu=1}^{n} R_{ij} \cdot (p_j + p_{|j-\nu|} \cdot x_\nu) \\
&= -\{R_{i*} \cdot x + \sum_{k=1}^{n} \{R_{ik} + R_{i*} \cdot y^{(k)}\} \cdot p_k\},
\end{aligned}$$

where R_{i*} denotes the i-th row of R and

$$y_\nu^{(k)} := \begin{cases} x_{\nu+k} & \text{for } \nu \le k \\ x_{\nu-k} & \text{for } \nu > n-k \\ x_{\nu-k} + x_{\nu+k} & \text{otherwise.} \end{cases}$$

Thus we have

$$Z_i = -\{R_{i*} \cdot x + \sum_{k=1}^{n} \{R_{ik} + R_{i*} \cdot y^{(k)}\} \cdot [p]_k\}.$$

As a first example we consider a linear system with symmetry constraint. We choose the following 2×2 example given by Behnke [13] to be able to plot $\Sigma([A],[b])$ vs. $\Sigma^{\mathrm{sym}}([A],[b])$

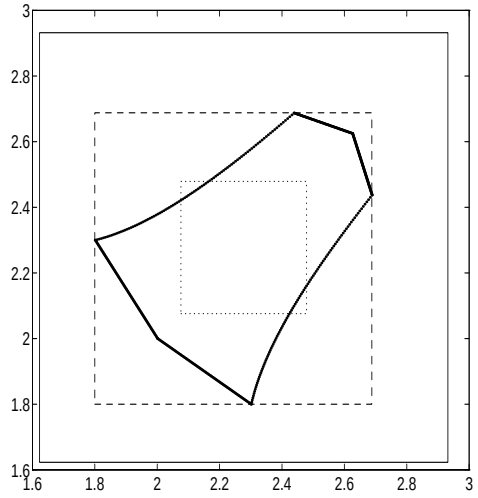
$$[A] := \begin{pmatrix} 3 & [1,2] \\ [1,2] & 3 \end{pmatrix}, \quad [b] := \begin{pmatrix} [10, & 10.5] \\ [10, & 10.5] \end{pmatrix}$$

Then the following inner and outer inclusions were computed using Theorem 4.8.

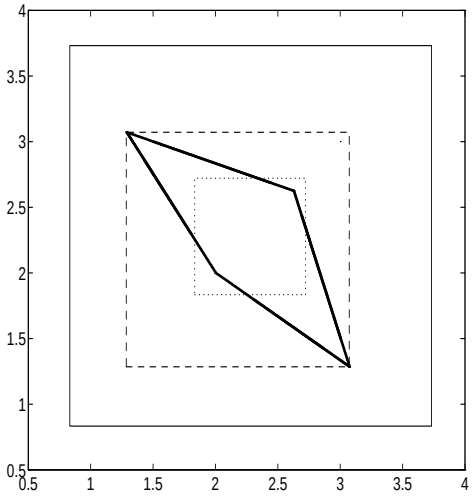
$$\begin{pmatrix} [2.076,\ 2.479] \\ [2.076,\ 2.479] \end{pmatrix} \subseteq \diamond\ \Sigma^{\mathrm{sym}} = \begin{pmatrix} [1.8100,\ 2.688] \\ [1.8100,\ 2.688] \end{pmatrix} \subseteq \begin{pmatrix} [1.623,\ 2.932] \\ [1.623,\ 2.932] \end{pmatrix}$$

and

$$\begin{pmatrix} [1.834,\ 2.722] \\ [1.834,\ 2.722] \end{pmatrix} \subseteq \diamond\Sigma = \begin{pmatrix} [1.285,\ 3.072] \\ [1.285,\ 3.072] \end{pmatrix} \subseteq \begin{pmatrix} [0.833,\ 3.723] \\ [0.833,\ 3.723] \end{pmatrix}.$$



Σ^{sym} for Behnke's example



Σ for Behnke's example

In the graph the vertices of the inner and outer inclusions can be seen. Note that Σ^{sym} is much smaller than Σ and the initial data has tolerances of 5 or 50 %, respectively. Σ_{sym} fits exactly into Σ (a different scale is used).

As a second example, consider (4.20) from Gregory/Karney [25], $a = 1$:

$$A = \begin{pmatrix} -1 & 2a & 1 & & & \\ 2a & 0 & 2a & 1 & & \\ 1 & 2a & 0 & 2a & 1 & \\ & 1 & 2a & 0 & 2a & 1 \\ & & & & \ddots & \\ & & & 1 & 2a & -1 \end{pmatrix} \quad \text{and } b := A \cdot \widehat{x} \text{ with } \widehat{x}_i = (-1)^{i+1}.$$

We set

$$[A] := A \cdot [1-e,\ 1+e],\ [b] := b \cdot [1-e,\ 1+e] \quad \text{for } e = 10^{-3},\ n = 50. \tag{71}$$

The third example is the Yule-Walker problem

$$(69) \quad \text{with } p = (100, 0, 0, 0, 0, 1)^T \cdot [1-e,\ 1+e] \quad \text{and } e = 10^{-3}. \tag{72}$$

Let

$$[X^*] \subseteq \diamond\Sigma([A],[b]) \subseteq [X] \quad \text{and} \quad [Y^*] \subseteq \diamond\Sigma(A(p),b(p),[p]) \subseteq [Y]$$

denote the inner and outer inclusion computed using Theorems 4.6 and 4.8, respectively.

We obtained the following results.

	min.	$\frac{w([X_i^*])}{w([X_i])}$	av.	min.	$\frac{w([Y_i^*])}{w([Y_i])}$	av.	min.	$\frac{w([X_i])}{w([Y_i])}$	av.
(71)	0.92		0.91	0.92		0.91	1.6		1.6
(72)	0.99		0.99	0.99		0.99	12.3		6600

The table shows that the inner and outer inclusions almost coincide, whereas the $[X]$ is larger, sometimes much larger than $[Y]$. More drastic examples can easily be constructed (see [39]).

5. Special nonlinear systems

Many standard problems in numerical analysis can be formulated as the solution of a system of nonlinear equations. For example,

$$f\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} Ax - \lambda x \\ e_k^T x - \zeta \end{pmatrix} : \mathbb{IR}^{n+1} \to \mathbb{IR}^{n+1} \quad \text{with } A \in \mathbb{IR}^{n\times n},\ 0 \neq \zeta \in \mathbb{IR},\ 1 \leq k \leq n \tag{73}$$

characterizes an eigenvector/eigenvalue pair (x,λ) of the matrix A with normalization $x_k = \zeta$. A slope function s_f can be given by

$$s_f((\widetilde{x},\widetilde{\lambda})^T,(x,\lambda)^T) := \begin{pmatrix} A - \widetilde{\lambda} I & -x \\ e_k^T & 0 \end{pmatrix} \tag{74}$$

satisfying (21) for all $x, \widetilde{x} \in \mathbb{IR}^n$, $\lambda, \widetilde{\lambda} \in \mathbb{IR}$. Following the lines of Chapter 2, we can formulate an inclusion theorem similar to Theorem 2.1. This yields the existence of an eigenvector/eigenvalue *pair* within the inclusion interval. However, for this special nonlinear system it is possible to prove much more, even when using slopes rather than a set of Jacobians (cf. [74], [75], [78]). We state a theorem for the generalized eigenvalue problem (see also [58], [60]).

Theorem 5.1. Let $T \in \{\mathbb{IR}, \mathbb{C}\}$, $A, B \in T^{n\times n}$, $R \in T^{(n+1)\times(n+1)}$, $\widetilde{x} \in T^n$, $\widetilde{\lambda}, \zeta \in T$ and define

$$G\begin{pmatrix} Y \\ M \end{pmatrix} := Z + \{I_{n+1} - R \cdot S(Y)\} \cdot \begin{pmatrix} Y \\ M \end{pmatrix}$$

$$\text{with} \quad Z := -R \cdot \begin{pmatrix} A\widetilde{x} - \widetilde{\lambda}B\widetilde{x} \\ e_k^T\widetilde{x} - \zeta \end{pmatrix} \text{ and } S(Y) := \begin{pmatrix} A - \widetilde{\lambda}B & -B(\widetilde{x}+Y) \\ e_k^T & 0 \end{pmatrix}$$

for $Y \in \mathbb{IP}T^n$, $M \in \mathbb{IP}T$ and a fixed integer k between 1 and n. If for nonempty, compact and convex $X \in \mathbb{IP}T^n$, $\Lambda \in \mathbb{IP}T$ and $\zeta \neq 0$

$$G\begin{pmatrix} X \\ \Lambda \end{pmatrix} \subsetneqq \begin{pmatrix} X \\ \Lambda \end{pmatrix},$$

then for the pencil $Ax = \lambda Bx$ the following holds true:

I) There exists one and only one eigenvector $\widehat{x}$ normalized to $e_k^T\widehat{x} = \zeta$ satisfying $\widehat{x} \in \widetilde{x} + X$.

II) There exists one and only one eigenvalue $\widehat{\lambda}$ satisfying $\widehat{\lambda} \in \widetilde{\lambda} + \Lambda$.

III) $\widehat{x}$ and $\widehat{\lambda}$ belong together: $A\widehat{x} = \widehat{\lambda}B\widehat{x}$.

The **proof** splits into four parts. First, the existence of an eigenvalue/eigenvector pair follows by Theorem 2.1. Then, second the uniqueness of the *pair* is proved. Third, for every eigenvector x of the pencil with $e_k^Tx = \zeta$ and $x \in \widetilde{x} + X$ it is proved that the corresponding eigenvalue λ satisfies $\lambda \in \widetilde{\lambda} + \Lambda$, demonstrating the uniqueness of the eigenvector. Fourth, the proof proceeds similarly for the eigenvalue. The proof can be found in [74] and [75] for the real case and the ordinary algebraic eigenvalue problem, and in [78] for the complex case and the generalized eigenvalue problem; see also [13]. ■

Note that Theorem 5.1 demonstrates the uniqueness of the pair $(\widehat{x}, \widehat{\lambda})$ and even the *individual uniqueness* of $\widehat{x}$ *and* $\widehat{\lambda}$ within $\widetilde{x} + X$ and $\widetilde{\lambda} + \Lambda$, resp. This holds although we only used slopes to expand the nonlinear function, not the Jacobian. Moreover, we did not assume $\widetilde{x} \in \widetilde{x} + X$ or $\widetilde{\lambda} \in \widetilde{\lambda} + \Lambda$. The ordinary algebraic eigenproblem is included in the above approach by setting $B = I$. Also, the nonlinear system can easily be reduced to dimension n (see [74]).

Theorem 5.1 immediately extends to data $[A], [B] \in \mathbb{I}T^{n\times n}$, $T \in \{\mathbb{R}, \mathbb{C}\}$ afflicted with tolerances. In this case inner inclusions can also be computed following the lines of Theorem 2.4. For the algebraic eigenproblem we want to derive a special technique for computing inner bounds. We describe it for real matrices; the method easily extends to the complex case.

Let $A, B \in \mathbb{R}^{n\times n}$, $A^*, B^* \in \mathbb{R}^{n\times n}$ with $A^*, B^* \geq 0$. Let λ be a simple eigenvalue of $Ax = \lambda Bx$ with right and left eigenvector x and y, respectively. Then the sensitivity of λ w.r.t. perturbations of A, B weighted by A^*, B^* is

$$|y^T| \cdot \{|A^*| + |\lambda| \cdot |B^*|\} \cdot |x| \;/\; |y^TBx|$$

for x, y subject to the normalization $x^T x = y^T y = 1$. This result follows from (38), see [80]. It has been formulated by Wilkinson [89] for $A^* = |A|$, $B^* = |B|$. The eigenvalue of the perturbed problem $\widetilde{A}\widetilde{x} - \widetilde{\lambda}\widetilde{B}\widetilde{x}$ is

$$\widetilde{\lambda} \approx \lambda + y^T \cdot \{\overline{A} - \lambda\overline{B}\} \cdot x \;/\; (y^T \overline{B} x) \quad \text{with } \overline{A} = \widetilde{A} - A,\ \overline{B} = \widetilde{B} - B.$$

Thus the largest change of λ is achieved for

$$\operatorname{sign}(\widetilde{A} - A) = \operatorname{sign}(yx^T) \text{ and } \operatorname{sign}(\widetilde{B} - B) = -\operatorname{sign}(\lambda) \cdot \operatorname{sign}(yx^T). \tag{75}$$

This can also be concluded from (38). Thus for given $[A]$, $[B]$ a method for computing *inner bounds* for an eigenvalue can be as follows [83].

1) Compute an inclusion $[\Lambda]$ of an eigenvalue λ of $Ax = \lambda Bx$, $A \in [A]$, $B \in [B]$ using Theorem 5.1.
2) Using (75) compute $\widetilde{A}_1 \in [A]$, $\widetilde{B}_1 \in [B]$ and $\widetilde{A}_2 \in [A]$, $\widetilde{B}_2 \in [B]$ for maximizing the change of λ.
3) Compute an inclusion $[\Lambda_1]$, $[\Lambda]_2$ of an eigenvalue of the corresponding pencils $\widetilde{A}_1 x = \lambda_1 \widetilde{B}_1 x$ and $\widetilde{A}_2 x = \lambda_2 \widetilde{B}_2 x$, respectively.
   If $[\Lambda]_1, [\Lambda]_2 \subseteq [\Lambda]$, then for all $\lambda \in [\Lambda]_1 \dot\cup [\Lambda]_2 \setminus ([\Lambda]_1 \cup [\Lambda]_2)$ there exists some $A \in [A]$, $B \in [B]$ such that λ is an eigenvalue of $Ax = \lambda Bx$.

Note that this procedure works correctly because of the *individual* uniqueness of the eigenvalue within the inclusion interval, as demonstrated by Theorem 5.1. Therefore λ_1 and λ_2 are perturbations of the *same* eigenvalue λ of the pencil. Furthermore, note that in step 3) two *point problems* have to be solved. Therefore in general this method works better than using Theorem 2.4. Details and examples are given in [83].

The eigenvalues of a pencil $Ax = \lambda Bx$ can be treated as eigenvalues of $B^{-1}A$ provided B is regular. However, this may yield weaker results if B is ill-conditioned. Moreover, the methods described so far also work in the degenerate case when B is singular. Consider the following example [78].

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

Then $\det(A - \lambda B) = 2\lambda - 2$ and the only eigenvector/eigenvalue pair is $x = (0,1)^T$, $\lambda = 1$. Nevertheless, sharp inclusions can be calculated using Theorem 5.1 [78].

If for Hermitian matrices only existence, but not uniqueness, of an eigenvalue within an error bound is required, the easiest and best technique seems to be to use perturbation bounds. This is again in the spirit of Wilkinson (3). For $A \in \mathbb{C}^{n\times n}$, $\widetilde{\lambda} \in \mathbb{C}$, $0 \neq \widetilde{x} \in \mathbb{C}^n$ and $A^H = A$, Theorem 4.14 [88] yields

$$\exists\, i: \quad |\lambda_i(A) - \widetilde{\lambda}| \leq \|A\widetilde{x} - \widetilde{\lambda}\widetilde{x}\|_2 \;/\; \|\widetilde{x}\|_2. \tag{76}$$

Another advantage of (76) is that multiple eigenvalues can also be included. This is not possible using Theorem 5.1. (76) is especially advantageous for including singular values of a matrix. If $A \in \mathbb{R}^{n\times n}$, then the singular values of A are eigenvalues of $\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$. Hence, for approximations $\widetilde{\sigma}$ of a singular value and $0 \neq \widetilde{u}$, $\widetilde{v}$ of the left, right singular vectors to $\widetilde{\sigma}$, (76) yields

$$\exists\, i: \quad |\sigma_i(A) - \widetilde{\sigma}| \leq \|(A^T\widetilde{u} - \widetilde{\sigma}\widetilde{v}, A\widetilde{v} - \widetilde{\sigma}\widetilde{u})^T\|_2 \;/\; \|(\widetilde{v},\widetilde{u})^T\|_2,$$

or simply

$$\exists\, i: \quad |\sigma_i(A)^2 - \widetilde{\sigma}^2| \leq \|A^TA\widetilde{v} - \widetilde{\sigma}^2\widetilde{v}\| \;/\; \|\widetilde{v}\|_2.$$

(Here, for $x, y \in \mathbb{R}^n$, $(x,y) \in \mathbb{R}^{2n}$ denotes the vector consisting of the components x_i followed by y_i).

This also allows inclusion of multiple singular values. In this case the singular values are well-conditioned whereas the singular vectors are ill-conditioned. Therefore an inclusion for

$$\begin{array}{ll} A\,v = \sigma\,u & u^T\,u = 1 \\ A^T\,u = \sigma\,v & v^T\,v = 1 \end{array}$$

using an inclusion theorem for nonlinear systems would not be possible. For inclusion of singular vectors and singular values see also [5].

6. Sparse systems of nonlinear equations

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function, $R \in \mathrm{M}_{nn}(\mathbb{R})$, $[Y] \in \mathbb{IR}^n$ and $\widetilde{x} \in D$ with $\widetilde{x} + [Y] \subseteq D$. In Theorem 2.1 we gave a necessary condition for $\widetilde{x} + [Y]$ to contain a zero of f. We assumed an expansion $s_f : D \times D \to \mathrm{M}_{nn}(\mathbb{R})$ be given with

$$\forall\, x \in \widetilde{x} + [Y]: \quad f(x) = f(\widetilde{x}) + s_f(\widetilde{x}, x)\cdot(x - \widetilde{x}).$$

We have seen in Chapter 3 that, roughly speaking, if f is given by means of a sequence of arithmetic expressions using $+, -, \cdot, /, \sqrt{\cdot}$, exp, log, trigonometric functions etc., then s_f can be evaluated in an automated way for point and interval data by means of another sequence of arithmetic expressions generated from the first one. Defining $Z := -R\cdot f(\widetilde{x}) \in \mathbb{IR}^n$, $\boldsymbol{C} : D \to \mathrm{M}_{nn}(\mathbb{R})$ with $\boldsymbol{C}_x := \boldsymbol{C}(x) = I - R \cdot s_f(\widetilde{x}, x)$ then

$$Z + \boldsymbol{C}_{\widetilde{x}+[Y]} \cdot [Y] \subsetneqq [Y] \tag{77}$$

implies the regularity of R and every $C \in \boldsymbol{C}_{\widetilde{x}+[Y]}$ and the existence of some $\widehat{x} \in \widetilde{x} + [Y]$ with $f(\widehat{x}) = 0$. For simplicity we omitted the Einzelschrittverfahren used in Theorem 2.1.

In many practical applications large and sparse system matrices occur. For the application of (77), we need an appropriate R. In the dense case, R was chosen to be an

approximate inverse of some $s_f(\widetilde{x}, x)$. In general, this inverse, however, becomes full. For larger n this would imply tremendous computational effort and large amounts of memory.

The question is whether R could be replaced by some decomposition. Let us look at (77) for some $x \in \widetilde{x} + [Y]$ and $R := (LU)^{-1}$.

$$Z + \boldsymbol{C}_x \cdot y = -R \cdot f(\widetilde{x}) + (I - R \cdot S) \cdot y = (LU)^{-1} \cdot \{-f(\widetilde{x}) + (LU - S) \cdot y\} \tag{78}$$

for $S := s_f(\widetilde{x}, x)$ and $y \in [Y]$. If we could verify that the r.h.s. of (78) is contained in $[Y]$ for every $x \in \widetilde{x} + [Y]$ by means of $\subsetneqq$, that is

$$(LU)^{-1} \cdot \Big\{ - f(\widetilde{x}) + (LU - s_f(\widetilde{x}, \widetilde{x} + [Y])) \cdot [Y] \Big\} \subsetneqq [Y], \tag{79}$$

then the assertions of Theorem 2.1 would hold true. L, U coming from an approximate LU-decomposition of some $s_f(\widetilde{x}, x)$ have the same profile as $s_f(\widetilde{x}, x)$. Thus, $LU - S$ is not expensive to compute explicitly (if it is not computed together with the decomposition itself; we come to this later). If L, U are banded or sparse then, in general, L^{-1}, U^{-1} again are full. Therefore we rewrite (79) as

$$U \setminus \Big\{ L \setminus \{ - f(\widetilde{x}) + (LU - s_f(\widetilde{x}, \widetilde{x} + [Y])) \cdot [Y] \} \Big\} \subsetneqq [Y] \tag{80}$$

where $\setminus$ denotes backward and forward substitution. This makes use of the structure of L and U. In a practical application, the values of the inner braces of (80) and $[Y]$ are small and more or less symmetric to the origin. Thus, replacing (80) by

$$\begin{array}{l} U \setminus \Big\{ L \setminus \{[-w, w]\} \Big\} \subsetneqq [Y] \quad \text{with} \\ w := | - f(\widetilde{x}) + (LU - s_f(\widetilde{x}, \widetilde{x} + [Y])) \cdot [Y]| \end{array} \tag{81}$$

as in Lemma 1.6 does reflect the practical case very well. That means, our problem reduces to solving a triangular system with interval right hand side. This can be done by means of interval backward and forward substitution, as has been noted by many authors [18], [28], [2]. However, we will see that this approach is suitable only for a special class of matrices.

In the following we will give specific solution procedures for different classes of matrices.

6.1. M-matrices

For simplicity consider first a linear system $f(x) = Ax - b$. Then (80) becomes

$$U \setminus \Big\{ L \setminus \{b - A\widetilde{x} + (LU - A) \cdot [Y]\} \Big\} \subsetneqq [Y]. \tag{82}$$

The difficult part, prone to possible overestimation, is the forward and backward substitution. This process of interval forward substitution can be described by

$$L \setminus [-w, w] = [-\langle L \rangle^{-1} \cdot w, \langle L \rangle^{-1} \cdot w] \tag{83}$$

for some $0 \leq w \in \mathbb{IR}^n$, where $\langle\cdot\rangle$ denotes Ostrowski's comparison matrix (cf. [66]). Looking at (81) in the case of M-matrices, fortunately there will be no overestimation by the process of forward and backward substitution, because L and U are M-matrices as well. Therefore $\langle L\rangle^{-1} = |L^{-1}|$, implying $L \setminus [-w, w] = [-|L^{-1}| \cdot w, |L^{-1}| \cdot w]$.

In the case of dense systems, an interval iteration is cheap compared to the cost for the elimination. For large sparse systems this changes. To avoid an iteration, denote $z := b - A\widetilde{x}$, $\Delta := LU - A$ and $y := |[Y]|$. Then (82) is true for $[Y] := [-y, y]$ if

$$\langle U\rangle \setminus \Big\{\langle L\rangle \setminus \{|z| + |\Delta| \cdot y\}\Big\} < y. \tag{84}$$

Now we compute y in such a way that (84) is satisfied. Denote

$$t_1 := \langle U\rangle \setminus \{\langle L\rangle \setminus |z|\} \in \mathbb{IR}^n \quad \text{and} \quad t_2 := \langle U\rangle \setminus \{\langle L\rangle \setminus \{|\Delta| \cdot t_1\}\} \in \mathbb{IR}^n.$$

If we can find $0 < \kappa \in \mathbb{IR}$ with

$$t_1 + (1+\kappa) \cdot t_2 < (1+\kappa) \cdot t_1,$$

then (84) is satisfied for $y := (1+\kappa) \cdot t_1$ and $A^{-1}b \in \widetilde{x} + (1+\kappa) \cdot [-t_1, t_1]$. That means

$$\kappa := \max_i \frac{(t_2)_i}{(t_1 - t_2)_i} \quad \text{provided} \quad t_1 > t_2$$

is the smallest suitable value using a continuity argument. This proves the following theorem which we formulate for nonlinear and linear systems the data of which may be afflicted with tolerances.

Theorem 6.1. Let continuous $f : D_p \times D_n \in \mathbb{IR}^p \times \mathbb{IR}^n \to \mathbb{IR}^n$ with an expansion (27) for $C \subseteq D_p$ be given, $[Y] \in \mathbb{IIR}^n$, $\widetilde{x} \in D_n$, $\widetilde{x} + [Y] \subseteq D_n$. Let L, $U \in \mathrm{M}_{nn}(\mathbb{IR})$ be regular lower, upper triangular and suppose

$$U \setminus \Big\{L \setminus \{-f(C, \widetilde{x}) + (LU - s_f(C, \widetilde{x}, \widetilde{x} + [Y])) \cdot [Y]\}\Big\} \subsetneqq [Y]. \tag{85}$$

Then every $M \in s_f(C, \widetilde{x}, \widetilde{x} + [Y])$ is regular and $\forall\ c \in C\ \exists\ \widehat{x}_c \in \widetilde{x} + [Y] : f(c, \widehat{x}_c) = 0$.

For linear $f(x) = Ax - b$, $A \in [A] \in \mathbb{IIM}_{nn}(\mathbb{IR})$, $b \in [b] \in \mathbb{IIR}^n$, define $\Delta := LU - [A] \in \mathbb{IIM}_{nn}(\mathbb{IR})$ and

$$\begin{array}{rcl} t_1 & := & \langle U\rangle \setminus \{\langle L\rangle \setminus |[b] - [A]\widetilde{x}|\} \in \mathbb{IR}^n \text{ and} \\ t_2 & := & \langle U\rangle \setminus \{\langle L\rangle \setminus \{|\Delta| \cdot t_1\}\} \in \mathbb{IR}^n. \end{array} \tag{86}$$

If $t_1 > t_2$ then every $A \in [A]$ is regular and for $\kappa := \max_i (t_2)_i/(t_1 - t_2)_i$

$$\Sigma([A], [b]) = \{\, x \in \mathbb{IR}^n \mid\ \exists\ A \in [A], b \in [b] : Ax = b \,\} \subseteq \widetilde{x} + (1+\kappa) \cdot [-t_1, t_1].$$

In a practical application, L and U arise from an (approximate) decomposition of A. This could be an LU-, but also an LDL^T- or LDM^T-decomposition with obvious

changes in (85), (86). The theorem is valid for arbitrary S or $[A]$; however, in a practical application for *general* S or $[A]$ the backward substituion in (85) or (86) will cause a tremendous overestimation, usually growing exponentially with the *dimension*. Consider

$$A = L \cdot L^T \quad \text{with} \quad L = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & 1 & 1 & 1 \\ & & & & \cdots \end{pmatrix}. \tag{87}$$

A is symmetric positive definite but not an M-matrix. Consider $[b] \in \mathbb{IIR}^n$ with $[b]_i := [-1, 1]$, $1 \leq i \leq n$. As we have seen in (83), the forward substitution with L can be expressed by $L \setminus [b] = [-\langle L \rangle^{-1} \cdot (1), \langle L \rangle^{-1} \cdot (1)]$ whereas $[-|L^{-1}| \cdot (1), |L^{-1}| \cdot (1)]$ is the true solution complex for $L^{-1} \cdot [b]$. Thus the amount of overestimation by interval backward substitution is exactly $\|\langle L \rangle^{-1}\|_\infty \,/\, \|L^{-1}\|_\infty$. It is

n	20	40	60	80	100
$\|\langle L \rangle^{-1}\|_\infty \,/\, \|L^{-1}\|_\infty$	$1.2 \cdot 10^3$	$8.8 \cdot 10^6$	$8.8 \cdot 10^{10}$	$1.0 \cdot 10^{15}$	$1.2 \cdot 10^{19}$

,

and we see the exponential behaviour of the overestimation by interval forward substitution. The condition number for $n = 100$ is $\mathrm{cond}(A) = 1.3 \cdot 10^4$. The observed overestimation depends mainly on the *dimension*, not on the condition number (see [82]).

Computing time can be saved by avoiding the explicit computation of Δ. This can be estimated a priori from the floating point decomposition.

Theorem 6.2. Let $A \in \mathrm{M}_{nn}(\mathbb{IR})$ have lower, upper bandwidth p, q, resp., $\beta = \min(p, q)$. Let the floating point LDM^T-, LDL^T-, Cholesky decomposition executed with rounding $\varepsilon < 0.01$ for the M-matrix, symmetric M-matrix, symm. pos. def. matrix A produce $(\widetilde{L}, \widetilde{D}, \widetilde{M^T})$, $(\widetilde{L}, \widetilde{D})$, $\widetilde{G}$, respectively. LDM^T and LDL^T are assumed to be executed in ordinary floating point arithmetic, while we use a precise scalar product for the Cholesky decomposition. If $\widetilde{D} \geq 0$ and the Cholesky decomposition does not break down, then for

$$\left.\begin{array}{ll} B_{kk} := 1.03 \cdot \beta \cdot A_{kk} & \\ B_{ik} := 3.08 \cdot \beta \cdot \widetilde{L}_{ik} \cdot \widetilde{D}_{kk} & \text{for } i > k \\ B_{ki} := 3.08 \cdot \beta \cdot \widetilde{D}_{kk} \cdot \widetilde{M}_{ki} & \text{for } i < k \end{array}\right\} \text{for } LDM^T$$

$$\left.\begin{array}{ll} B_{kk} := 1.03 \cdot p \cdot A_{kk} & \\ B_{ik} := 3.08 \cdot p \cdot \widetilde{L}_{ik} \cdot \widetilde{D}_{kk} & \text{for } i \neq k \end{array}\right\} \text{for } LDL^T$$

$$\left.\begin{array}{ll} B_{kk} := 3.04 \cdot \widetilde{G}_{kk}^2 & \\ B_{ik} := 2.01 \cdot \widetilde{G}_{ik} \cdot \widetilde{G}_{kk} & \text{for } i \neq k \end{array}\right\} \text{ for Cholesky}$$

$$|A - \widetilde{A}| \leq \varepsilon \cdot |B| \quad \text{for } \widetilde{A} = \widetilde{L}\widetilde{D}\widetilde{M}^T,\ \widetilde{L}\widetilde{D}\widetilde{L}^T,\ \widetilde{G}\widetilde{G}^T, \quad \text{holds, resp.}$$

For the **proof** and more details, see [80]. ■

Consider the discretization of the Poisson equation on a rectangle

$$A = \begin{pmatrix} M & -I & & \\ -I & M & -I & \\ & -I & M & -I \\ & & & \cdots \end{pmatrix} \quad \text{with} \quad M = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & & \cdots \end{pmatrix} \tag{88}$$

Then for a r.h.s. b such that the solution of $Ax = b$ is $x = (x_i)$, $x_i = 1/i$, an algorithm based on Theorems 6.1 and 6.2 in single precision ($\sim$ 7 decimals) computation delivers an inclusion $[X] \in \mathbb{IIR}^n$ with the quality given in the following table [80]. For $[A] := A \cdot [1-e, 1+e]$, $[b] := b \cdot [1-e, 1+e]$, $e = 10^{-5}$ the inclusions $[Y]$ were obtained.

n	*bandwidth*	$\max_i \frac{w([X]_i)}{1/i}$	$\max_i \frac{w([Y]_i)}{1/i}$
80 000	40	$4.7 \cdot 10^{-5}$	$7.4 \cdot 10^{-4}$
200 000	20	$2.8 \cdot 10^{-5}$	$3.2 \cdot 10^{-4}$
500 000	10	$5.9 \cdot 10^{-6}$	$1.3 \cdot 10^{-4}$
1 000 000	5	$1.3 \cdot 10^{-6}$	$5.2 \cdot 10^{-5}$

In the example above we could also apply straightforward interval Gaussian elimination. This is applicable because Alefeld proved the following theorem [3].

Theorem 6.3. Let $[A] \in \mathbb{IM}_{nn}(\mathbb{IR})$ where $B := \langle [A] \rangle \in \mathrm{M}_{nn}(\mathbb{IR})$ as defined by

$$B_{ij} := \begin{cases} |\mathrm{mid}([A]_{ii})| - \dfrac{1}{2} w([A]_{ii}) & \text{for } i = j \\ -\,|[A]_{ij}| & \text{otherwise} \end{cases}$$

is an M-matrix. Then interval Gaussian elimination does not break down even without pivoting.

For M-matrices the interval Gauß-algorithm (IGA) yields good inclusions. Consider (88) with $[L], [U] \in \mathbb{IM}_{nn}(\mathbb{IR})$ produced by IGA yielding an inclusion of the solution $[X]$ by interval forward and backward computation. All operations are interval operations in single precision equivalent to 7 decimal places. Again we considered the point system

$Ax = b$ and $[A] := A \cdot [1-e, 1+e]$, $[b] := b \cdot [1-e, 1+e]$ with $A\widehat{x} = b$, $\widehat{x}_i = 1/i$. Below we give the average and maximum error of the computed inclusion $[X]$.

n	p	e	average	$\frac{w([X]_i)}{1/i}$	maximum
100 000	5	0	$1.6 \cdot 10^{-5}$		$2.5 \cdot 10^{-5}$
100 000	10	0	$7.8 \cdot 10^{-5}$		$1.2 \cdot 10^{-4}$
100 000	20	0	$4.7 \cdot 10^{-4}$		$7.7 \cdot 10^{-4}$
100 000	5	10^{-5}	$5.8 \cdot 10^{-4}$		$9.4 \cdot 10^{-4}$
100 000	10	10^{-5}	$1.8 \cdot 10^{-3}$		$3.1 \cdot 10^{-3}$
100 000	20	10^{-5}	$6.7 \cdot 10^{-3}$		$1.1 \cdot 10^{-2}$

Compared to the previous results we see only a small overestimation due to pure rounding error effects. For the case of Poisson's equation, also an interval version of Bunemann's algorithm proposed by Schwandt is applicable (cf. [84], [85], for comparisons see also [80]).

6.2. Symmetric positive definite matrices

Next we treat the large class of s.p.d. (symm. pos. def.) systems. Again, we start with a linear system $Ax = b$. We have already seen in example (87) that interval forward or backward substitution may produce vast overestimations for s.p.d. matrices that are not M-matrices. Therefore, another method has to be used. Denote the singular values of A by $\sigma_1(A) \geq \ldots \geq \sigma_n(A)$. A is s.p.d.; therefore the eigenvalues $\lambda_i(A)$ coincide with the singular values. Then for $\widehat{x} := A^{-1}b$, $\widetilde{x} \in \mathbb{IR}^n$ we have

$$\begin{aligned} \|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 &= \|A^{-1} \cdot (b - A\widetilde{x})\|_2 \\ &\leq \|A^{-1}\|_2 \cdot \|b - A\widetilde{x}\|_2 = \sigma_n(A)^{-1} \cdot \|b - A\widetilde{x}\|_2. \end{aligned} \tag{89}$$

Thus a *lower bound* on the smallest singular value $\sigma_n(A)$ delivers bounds for the solution of the linear system. If for some $\lambda \in \mathbb{IR}$, $A - \lambda I$ is pos. def. and hence the Cholesky decomposition GG^T of $A - \lambda I$ exists, then

$$\lambda_n(A - \lambda I) = \lambda_n(A) - \lambda = \sigma_n(A) - \lambda > 0 \quad \text{and } \sigma_n(A) > \lambda.$$

The existence of a Cholesky decomposition and therefore positive definiteness of $A - \lambda I$ could be verified by means of an interval Cholesky decomposition. That means, every operation is replaced by its corresponding interval operation. However, as in (87) this does not work for general s.p.d. matrices. Applying interval Cholesky decomposition to $0.1 \cdot A$, A defined as in (87) producing $[G] \in \mathbb{IM}_{nn}(\mathbb{IR})$ with $\exists\, G \in [G] : GG^T = A$ and monitoring the diameter of the last diagonal element $[G]_{nn}$ of $[G]$ gives the following results. The computation is performed in double precision (~ 17 decimals).

n	5	10	15	20	25	30	35	40
$w([G]_{nn})$	$5 \cdot 10^{-15}$	$7 \cdot 10^{-13}$	$9 \cdot 10^{-11}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-2}$	failed

Already for $n = 40$ the algorithm fails by running into a diagonal element containing 0. The factor 0.1 has been introduced because otherwise all intermediate results would be integers. Again, this is not a problem of the condition ($\text{cond}(A) \approx 2.1 \cdot 10^3$ for $n = 40$) but of the *dimension*.

Therefore, following Wilkinson's rule (3) we do as much as possible in floating point, and perform an a posteriori error analysis. Let

$$\widetilde{G} \text{ be an approximate Cholesky factor of } A - \lambda I$$

and define $E := \widetilde{G}\widetilde{G}^T - (A - \lambda I) \in \mathrm{M}_{nn}(\mathbb{R})$. Then a perturbation theorem by Wilkinson for eigenvalues of symmetric matrices [89], pp. 101–2, shows

$$|\lambda_i(A - \lambda I + E) - \lambda_i(A - \lambda I)| \leq \|E\| \quad \text{for } 1 \leq i \leq n.$$

But $\widetilde{G}\widetilde{G}^T = A - \lambda I + E$ and therefore $A - (\lambda - \|E\|) \cdot I$ is positive semidefinite, implying

$$\sigma_n(A) \geq \lambda - \|E\|.$$

Lemma 6.4. Let $A \in \mathrm{M}_{nn}(\mathbb{R})$ be symmetric, $\lambda \in \mathbb{R}$ and $\widetilde{G} \in \mathrm{M}_{nn}(\mathbb{R})$. If for some norm

$$\tau := \|\widetilde{G}\widetilde{G}^T - (A - \lambda I)\| \quad \text{with } \lambda > \tau,$$

then A is positive definite and for the smallest singular value of A

$$\sigma_n(A) \geq \lambda - \tau > 0$$

holds. For $[A] \in \mathbb{IM}_{nn}(\mathbb{R})$ and

$$\tau := \|\widetilde{G}\widetilde{G}^T - ([A] - \lambda I)\| \quad \text{with } \lambda > \tau, \tag{90}$$

every *symmetric* $A \in [A]$ is regular and

$$\forall\, A \in [A] \text{ with } A = A^T: \quad \sigma_n(A) \geq \lambda - \tau > 0.$$

If the 2-norm is used in (90), then *every* $A \in [A]$ is regular.

Proof. The first part has been proved before by observing that $\widetilde{G}\widetilde{G}^T$ is symmetric positive semidefinite and therefore $\sigma_n(A) = \lambda_n(A) \geq \lambda - \tau > 0$. Applying this to every $A \in [A]$ also implies the second part. The matrix of eigenvectors X of $\widetilde{G}\widetilde{G}^T$ is orthogonal with $\|X\|_2 = 1$. Therefore, the Baur-Fike Theorem [24] implies regularity for *every*

$A \in [A]$ (note that for unsymmetric $A \in [A]$ the eigenvalues and singular values do not necessarily coincide). ■

Note that A is not assumed to be positive definite but this is verified a posteriori by Lemma 6.4. We formulate the following theorem for nonlinear and linear systems the data of which may be afflicted with tolerances.

Theorem 6.5. Let continuous $f : D_p \times D_n \subseteq \mathbb{R}^p \times R^n \to \mathbb{R}^n$ with an expansion (27) for $C \subseteq D_p$ be given, $0 < y \in \mathbb{R}^n$, $[Y] := [-y, y] \in \mathbb{IR}^n$, $\widetilde{x} \in D_n$, $\widetilde{x} + [Y] \subseteq D_n$. Let $S \in \mathrm{M}_{nn}(\mathbb{R})$ symmetric, $\lambda \in \mathbb{R}$, $\widetilde{G} \in \mathrm{M}_{nn}(\mathbb{R})$, for some norm let

$$\tau := \|\widetilde{G}\widetilde{G}^T - (S - \lambda I)\| \quad \text{and assume} \quad \lambda > \tau. \tag{91}$$

If

$$(\lambda - \tau)^{-1} \cdot \| - f(C, \widetilde{x}) + (S - s_f(C, \widetilde{x}, \widetilde{x} + [Y])) \cdot [Y]\|_2 < \|y\|_2, \tag{92}$$

then every matrix $M \in s_f(C, \widetilde{x}, \widetilde{x}+[Y])$ is regular and $\forall\, c \in C\ \exists\, \widehat{x}_c \in \widetilde{x}+[Y] : f(c, \widetilde{x}) = 0$. The assertions remain true when choosing $S := \mathrm{mid}(s_f(C, \widetilde{x}, \widetilde{x} + [Y]))$ and replacing (92) by

$$\|f(C, \widetilde{x})\|_2 + \|\mathrm{rad}(s_f(C, \widetilde{x}, \widetilde{x} + [Y]))\|_2 \cdot \|y\|_2 < (\lambda - \tau) \cdot \|y\|_2.$$

For linear $f(x) = Ax - b$, $A \in [A] \in \mathbb{IM}_{nn}(\mathbb{R})$, $b \in [b] \in \mathbb{IR}^n$ let $\lambda \in \mathbb{R}$, $\widetilde{G} \in \mathrm{M}_{nn}(\mathbb{R})$, for some norm let

$$\tau := \|\widetilde{G}\widetilde{G}^T - ([A] - \lambda I)\| \quad \text{and assume} \quad \lambda > \tau. \tag{93}$$

Then every symmetric $A \in [A]$ is regular with $\sigma(A) \geq \lambda - \tau > 0$ and for all symmetric $A \in [A]$ and $\forall\, b \in [b]$

$$\|A^{-1}b - \widetilde{x}\|_\infty \leq \|A^{-1}b - \widetilde{x}\|_2 \leq (\lambda - \tau)^{-1} \cdot \|[b] - [A] \cdot \widetilde{x}\|_2. \tag{94}$$

If the 2-norm is used in (93), then the assertions hold for every $A \in [A]$.

Proof. Lemma 6.4 and (92) imply $\sigma_n(S) \geq \lambda - \tau > 0$ and therefore the regularity of S. Then (92) and the symmetry of $[Y]$ imply

$$S^{-1} \cdot \{-f(C, \widetilde{x}) + (S - s_f(C, \widetilde{x}, \widetilde{x} + [Y]) \cdot [Y]\} \subseteq \mathrm{int}([Y]), \tag{95}$$

and using the same deduction as for (78) shows that the assumptions of Theorem 2.4 are satisfied and finishes the first part of the proof. The second part follows by (89) and Lemma 6.4. ■

In the linear interval case the assertions are true for symmetric $A \in [A]$, whereas in the nonlinear case only the symmetry of S is needed to be able to apply Lemma 6.4. In Lemma 6.6 we show how to bound the smallest singular value of a general matrix from below.

Setting $S := \mathrm{mid}(s_f(C, \tilde{x}, \tilde{x} + [Y]))$ is the optimal choice for a preconditioner, see [66] and the work by Rex and Rohn [71]. The regularity of S is shown a posteriori by (91). The computation of S^{-1}, which, in general, is full, is avoided by estimating the effect of S^{-1} on a vector by (92).

The solution of large linear systems has direct applications in the verified solution of ordinary and partial differential equations (cf. [67], [63]). Beside this, a verified lower bound for the smallest singular value of a symmetric or the smallest eigenvalue of an s.p.d. matrix are useful in the work of [67] and [13], see also the papers by Behnke, Goerisch and Plum in this volume. Unfortunately, we do not have the space to go into much detail on sparse systems. It will be treated in a separate, forthcoming paper.

In order to apply Theorem 6.5 we have to compute an approximate Cholesky decomposition of $S - \lambda I$ or $m([A]) - \lambda I$ and an upper bound on τ. These two steps can be performed simultaneously in one algorithm. Denote $S - \lambda I$ or $m([A]) - \lambda I$ by $B \in \mathrm{M}_{nn}(\mathbb{IR})$. Then

$$\begin{aligned} &\text{for } i \;= 1 \ldots n \quad \text{for } j = 1 \ldots i \\ &\qquad \{r = B_{ij} - \sum_{\nu=1}^{j-1} \widetilde{G}_{i\nu}\widetilde{G}_{j\nu}; \quad \text{if } i = j \quad \text{then } \widetilde{G}_{ii} = r^{1/2} \quad \text{else } \widetilde{G}_{ij} = r \;/\; \widetilde{G}_{jj}\} \end{aligned}$$

computes $\widetilde{G}$ with $\widetilde{G}\widetilde{G}^T = B$. But also in every step

$$E_{ij} = r - \widetilde{G}_{ij} \cdot \widetilde{G}_{jj} \quad \text{with } E := B - \widetilde{G}\widetilde{G}^T.$$

That means if we perform the computation of r by interval arithmetic, yielding $[r]$, use $m([r])$ and floating point arithmetic to compute $\widetilde{G}_{ii}$ and $\widetilde{G}_{ij}$, and interval arithmetic to compute E_{ij}, then both $\widetilde{G}$ and E and therefore τ are computed. For interval input data we add $\mathrm{rad}([B]_{ij}) \cdot [-1, 1]$ to E_{ij}. If the precise scalar product proposed by Kulisch [55], [56] is available, then r can be kept in the accumulator, rounded once to nearest for computing $\widetilde{G}_{ii}$ or $\widetilde{G}_{ij}$ and, after subtracting $\widetilde{G}_{ij} \cdot \widetilde{G}_{jj}$, to the smallest enclosing interval. The computational costs for some $[A]$ with bandwidth p are $\frac{1}{2}\, n \cdot p^2$ operations.

This yields the following algorithmic approach.

1) Compute an approximate Cholesky decomposition of $m([A])$, and using this obtain an approximate solution $\tilde{x}$ of $m([A]) \cdot x = m([b])$. Compute an approximation $\widetilde{\lambda}$ of the smallest eigenvalue of $m([A])$ by means of inverse power iteration; set $\lambda := 0.9 \cdot \widetilde{\lambda}$.

2) Compute an approximate Cholesky decomposition $\widetilde{G}$ of $m([A]) - \lambda I$ and an upper bound τ for $\|\widetilde{G}\widetilde{G}^T - ([A] - \lambda I)\|$ using the method described above.

3) For $\lambda > \tau$ (94) holds.

Inverse iteration in step 1) is inexpensive because only floating point forward and backward substitutions are necessary. Thus the main costs are the two *approximate* Cholesky decompositions yielding a total of

$n \cdot p^2$ floating point operations.

The estimation (94) gives a norm-wise estimate, no componentwise estimate. That is, if some components are much smaller in magnitude than others the relative accuracy of the inclusion decreases. A componentwise estimation has been presented by the author at the conference "Numerical Results with Automatic Result Verification " at Lafayette, Louisiana in February 1993 but not yet published. The principle of the estimation (94) contains no overestimation as long as the approximation λ is not too bad. This is because, as we mentioned before, $[b] - [A] \cdot \widetilde{x}$ is essentially symmetric to the origin thus containing the singular vector belonging to the smallest singular value.

There is another approach presented at the conference just mentioned [9]. However, the singleton method and others implicitly calculate an inverse of L and U. Thus the computing time n^2p grows quadratically with n compared to np^2 in our approach.

In the following examples we used a r.h.s. of $Ax = b$ s.t. $(A^{-1}b)_i = (-1)^{i+1} \ / \ i$. *iter* denotes the number of inverse power iterations. All computations were performed in double precision ($\sim$ 17 decimals). For our matrix (87) we obtained

n	cond(A)	*iter*	$\|\widehat{x} - \widetilde{x}\|_\infty \ / \ \|\widetilde{x}\|_\infty$
10 000	$1.2 \cdot 10^8$	3	$3.4 \cdot 10^{-17}$
100 000	$1.2 \cdot 10^{10}$	3	$3.4 \cdot 10^{-15}$
1 000 000	$1.2 \cdot 10^{12}$	3	$3.4 \cdot 10^{-13}$

For (61) from Gregory/Karney [25] we obtained

n	cond(A)	*iter*	$\|\widehat{x} - \widetilde{x}_\infty\| \ / \ \|\widetilde{x}\|_\infty$
1 000	$1.7 \cdot 10^{11}$	2	$3.9 \cdot 10^{-14}$
10 000	$1.6 \cdot 10^{15}$	2	$5.4 \cdot 10^{-10}$
20 000	$2.6 \cdot 10^{16}$	2	$1.8 \cdot 10^{-8}$
50 000	$1.0 \cdot 10^{18}$	2	failed

The failure in the last example is due to the large condition number, which is beyond the critical value of 10^{17} for a precision of 17 decimals. The behaviour of our method depends mainly on the condition number, *not* on the dimension.

6.3. General matrices

First, we consider again the linear case. For $A \in \mathrm{M}_{nn}(\mathbb{R})$ we might try to estimate $\sigma_n(A)$ using Lemma 6.4 for $A^T A$. However, this would limit the scope of applicability to $\mathrm{cond}(A) \lesssim 10^{l/2}$ if l decimal digits precision are used. Instead, we observe for an LDM^T-decomposition

$$LDM^T = A \quad \Rightarrow \quad \sigma_n(A) \geq \sigma_n(L) \cdot \sigma_n(D) \cdot \sigma_n(M). \tag{96}$$

The heuristic is that $\sigma_n(L)$ and $\sigma_n(M)$ are of the same order of magnitude and not too small, where $\sigma_n(D)$ is simply the minimum of $|D_{ii}|$. For other decompositions like LDL^T or LU similar considerations hold.

In practice, an approximate decomposition $\widetilde{L}\widetilde{D}\widetilde{M}^T \approx A$ is used and the error $\Delta := \widetilde{L}\widetilde{D}\widetilde{M}^T - A$ has to be considered. For $[\Delta] \in \mathbb{IM}_{nn}(\mathbb{R})$ we define $\|[\Delta]\| := \max\{\, \|\Delta\| \mid \Delta \in [\Delta]\,\}$.

Lemma 6.6. Let $[A] \in \mathbb{IM}_{nn}(\mathbb{R})$, $\lambda \in \mathbb{R}$ and $\widetilde{A} \in \mathrm{M}_{nn}(\mathbb{R})$. Define

$$[\Delta] := \widetilde{A} - [A] \in \mathbb{IM}_{nn}(\mathbb{R}),\ \tau := \|[\Delta]\|_2 = \| \, |[\Delta]| \, \|_2 \text{ and assume } \sigma_n(\widetilde{A}) > \tau. \tag{97}$$

Then every $A \in [A]$ is regular with

$$\sigma_n(A) \geq \sigma_n(\widetilde{A}) - \tau > 0. \tag{98}$$

Proof. $\sigma_n(\widetilde{A}) > \tau$ implies the regularity of $\widetilde{A}$. For fixed but arbitrary $A \in [A]$ define $\Delta := \widetilde{A} - A \in [\Delta]$. Then, perturbation theory for singular values (Corollary 8.3–2, [24]) and using

$$B \in \mathbb{R}^{n \times n} \ \Rightarrow \ \|B\|_2 = \sigma_1(B) = \lambda_1 \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \leq \lambda_1 \begin{pmatrix} 0 & |B| \\ |B^T| & 0 \end{pmatrix} = \| \, |B| \, \|$$

shows

$$\sigma_n(A) \geq \sigma_n(\widetilde{A}) - \|\Delta\|_2 \geq \sigma_n(\widetilde{A}) - \| \, |[\Delta]| \, \|_2 \geq \sigma_n(\widetilde{A}) - \tau. \qquad \blacksquare$$

For the practical application we have to observe that the 2-norm is not absolute. We may use $\|[\Delta]\|_2 \leq \{\, \|[\Delta]\|_1 \cdot \|[\Delta]\|_\infty \}^{1/2}$. If $\widetilde{A} := \widetilde{L}\widetilde{U}$, $\widetilde{A} := \widetilde{L}\widetilde{D}\widetilde{L}^T$ or $\widetilde{A} := \widetilde{L}\widetilde{D}\widetilde{M}^T$, then $\sigma_n(\widetilde{A})$ can be estimated from below using (96). $[\Delta]$ can be estimated during the elimination process or by using Theorem 6.2. This is also true for $\widetilde{A} := \widetilde{L}\widetilde{U}$. Finally we obtain the following theorem.

Theorem 6.7. Let continuous $f : D_p \times D_n \subseteq \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$ with an expansion (27) for $C \subseteq D_p$ be given, $0 < y \in \mathbb{R}^n$, $[Y] := [-y, y] \in \mathbb{IR}^n$, $\widetilde{x} \in D_n$, $\widetilde{x} + [Y] \subseteq D_n$. If for regular $\widetilde{P}, \widetilde{Q} \in \mathrm{M}_{nn}(\mathbb{R})$

$$\sigma_n(\widetilde{Q})^{-1} \cdot \sigma_n(\widetilde{P})^{-1} \cdot \| \, -f(C, \widetilde{x}) + (\widetilde{P}\widetilde{Q} - s_f(C, \widetilde{x}, \widetilde{x} + [Y])) \cdot [Y] \|_2 < \| \, [Y] \, \|_2,$$

then every matrix $M \in s_f(C, \widetilde{x}, \widetilde{x}+[Y])$ is regular and $\forall\, c \in C\ \exists\, \widehat{x}_c \in \widetilde{x}+[Y] : f(c, \widetilde{x}) = 0$.

For linear $f(x) = Ax - b$, $A \in [A]$, $b \in [b]$, let $\widetilde{A} \in \mathrm{M}_{nn}(\mathbb{IR})$ with

$$[\Delta] := \widetilde{A} - [A] \in \mathbb{IM}_{nn}(\mathbb{IR}), \quad \tau := \|[\Delta]\|_2 \quad \text{and assume } \sigma_n(\widetilde{A}) > \tau.$$

Then every $A \in [A]$ is regular with $\sigma_n(A) \geq \sigma_n(\widetilde{A}) - \tau > 0$ and

$$\forall\, A \in [A]\ \forall\, b \in [b] : \|A^{-1}b - \widehat{x}\|_\infty \leq \|A^{-1}b - \widehat{x}\|_2 \leq \{\sigma_n(\widetilde{A}) - \tau\}^{-1} \cdot \|\,[b] - [A] \cdot \widetilde{x}\|_2.$$

Proof. The first part follows using (95) like in the proof of Theorem 6.5, the second part is a consequence of (89) applied to every $A \in [A]$. ■

In the application of Theorem 6.7, $\widetilde{P}$ and $\widetilde{Q}$ are the factors of an approximate decomposition of some $S \in s_f(C, \widetilde{x}, \widetilde{x}+[Y])$ where $\sigma_n(\widetilde{P})$ is estimated by applying Lemma 6.4 to $A := \widetilde{P}\widetilde{P}^T$, similarly for $\widetilde{Q}$. The heuristic is that the condition of S is equally distributed among the factors $\widetilde{P}$ and $\widetilde{Q}$ and that the condition of $\widetilde{P}\widetilde{P}^T$ and $\widetilde{Q}\widetilde{Q}^T$ does not exceed the condition of S by too much. In the linear case $\widetilde{A} := \widetilde{P}\widetilde{Q}$, where $\widetilde{P}$ and $\widetilde{Q}$ are approximate factors of $\mathrm{mid}([A])$.

Consider example (65) from [25] for $a = 1$, the matrix being symmetric but not positive definite. Using Theorem 6.7 and an $\widetilde{L}\widetilde{D}\widetilde{L}^T$-decomposition yields the following results.

n	$\mathrm{cond}(A)$	*iter*	$\|\widehat{x} - \widetilde{x}\|_\infty \,/\, \|\widetilde{x}\|_\infty$
1 000	$6.2 \cdot 10^2$	3	$1.0 \cdot 10^{-18}$
10 000	$5.5 \cdot 10^3$	3	$7.6 \cdot 10^{-17}$
100 000	$6.3 \cdot 10^4$	3	$7.6 \cdot 10^{-14}$

The approximation of the smallest singular value of $\widetilde{L}\widetilde{L}^T$ is computed by *iter* steps of inverse power iteration. The difference $\widetilde{G}\widetilde{G}^T - (\widetilde{L}\widetilde{L}^T - \lambda I)$ can be computed together with the Cholesky decomposition for $\widetilde{L}\widetilde{L}^T - \lambda I$ as described before. For more details see [82].

We finish with some sparse examples from the Harwell test case library [21]. p, q denote the lower, upper bandwidth, where profile is the number of nonzero elements (see [82]):

Matrix	n	p	q	profile	cond	$\|\widehat{x} - \widetilde{x}\|_\infty \,/\, \|\widetilde{x}\|_\infty$
*gre*_216	216	14	36	876	$2.7 \cdot 10^2$	$1.0 \cdot 10^{-18}$
*gre*_343	343	18	49	1435	$2.5 \cdot 10^2$	$1.0 \cdot 10^{-18}$
*gre*_512	512	24	64	2192	$3.8 \cdot 10^2$	$1.0 \cdot 10^{-18}$
*west*0167	167	158	20	507	$2.8 \cdot 10^6$	$1.0 \cdot 10^{-18}$
*west*0381	381	363	153	2157	$2.0 \cdot 10^6$	$1.0 \cdot 10^{-18}$
*bcsstk*08	1074	590	590	7017	$6.1 \cdot 10^6$	$1.0 \cdot 10^{-18}$
*bcsstk*14	1806	161	161	3263	$4.3 \cdot 10^4$	$1.0 \cdot 10^{-18}$

7. Implementation issues: An interval library

Finally we will discuss computational and performance aspects of verification algorithms on the arithmetical and programming level. This is necessary, because directed roundings are used in one way or another, and because most simple operations, such as a sign test or switching rounding, are expensive for today's machines as compared to floating point operations.

Not too long ago, the paradigm was that the computing time is essentially proportional to the number of multiplications. This paradigm includes that in most numerical algorithms: divisions are rare, and addition/subtraction used to be much faster than multiplication. So we learned that Gaussian elimination needs $\frac{1}{3}n^3 + O(n^2)$ operations.

Meanwhile the computing paradigm changed dramatically. To see this we do *not* have to go to large vector or parallel machines, PC's or workstations suffice. Consider, for example, an IBM RS/6000 Model 370, a 63 MHz machine with 25 Linpack MFlops. If we look at

floating point multiplication	$x * y$
floating point addition	$x + y$
floating point comparison	$x < y$
floating point Mult & Add	$x * y + z$
switching rounding mode,	

then ultimately *each of these operations can be executed in 1 cycle* or 63 Million times per second. This is true provided the operands are in the registers; otherwise some 2 or 3 cycles are needed. The main point is that this performance *can be achieved* if the code is written in a proper way and the problem is formulated in a suitable way. Here we see high impact of implementational issues on the design of algorithms, in other words: Scientific Computing. Consider, for example, Gaussian elimination and matrix multiplication. Then for a full matrix with $n = 300$ we have on the IBM RS/6000 Model 370

Linpack LU-decomposition	0.9 sec $= \frac{1}{3} \cdot 2.7$ sec
Matrix multiplication	1.8 sec.

In other words, $3 \cdot (\frac{1}{3}n^3)$ operations need not to be equivalent to n^3 operations, it depends on the algorithm. The above numbers hold for non-blocked versions; the blocked matrix multiplication needs about 0.6 sec.

That means we have vast differences in computing times depending on whether the cache can be used effectively, on the "simplicity" of the code so it can be optimized by the computer, and much more.

For interval operations these arguments are significantly amplified by the fact that sign tests, comparisons and so forth are necessary. It is of utmost importance to the final performance of a verification algorithm that the above arguments are taken into account.

Therefore we designed and implemented a C-library BIAS [46], Basic Interval Arithmetic Subroutines, for general purpose machines and a C++ class library PROFIL [47], [49], [50] providing convenient access to the operations defined in BIAS. These libraries have been developed with emphasis on providing

- a concise interface for basic interval routines
- an interface independent of the specific interval representation
- portability
- speed

Let us first look at the basic arithmetic, vector and matrix operations for points and intervals. The directed rounding, which used to be a big problem, can be handled using IEEE-arithmetic [34], [35] and coprocessors implementing it. Today, many PC's, workstations and mainframes do support IEEE arithmetic. However, switching the rounding mode may be made dramatically faster by writing a one- or two-line assembler program rather than using the built-in routines. For example, on the IBM RS/6000 mentioned above, the

built-in library function needs	$\approx$ 45 cycles
whereas an assembler routine needs	1 cycle.

If no IEEE arithmetic is available, the rounding may be simulated through multiplication by $1-\varepsilon$, $1+\varepsilon$. This requires careful implementation near underflow, but offers thereby the advantage of portability to a wide variety of machines (cf. [45]).

Having routines for switching rounding mode, a fast implementation of the basic arithmetic routines $+, -, \cdot, /$ for reals and intervals is not too difficult. After finishing an interval operation, optionally, one may leave the rounding mode as is and not switch it back to nearest. This saves about 10 % computing time.

For vector and matrix operations things change. Consider $[Y] := r \cdot [X]$, $r \in \mathbb{R}$, $[X], [Y] \in \mathbb{IR}^n$. Then a straightforward implementation is

$$\text{for} \quad i = 1 \ldots n \quad \text{do} \quad [Y]_i := r \cdot [X]_i; \tag{99}$$

where the multiplication is a $\mathbb{R} \times \mathbb{IR}$-multiplication. However, in (99) n sign-tests on r and about $2n$ switches of the rounding mode are executed. Sign-tests and rounding

switches are expensive, therefore this is a very inefficient implementation. Consider

$$\begin{array}{lllll} \text{if } r>0 \text{ then } \{ & \text{set-rounding-down;} & \text{for } i=1\ldots n \text{ do } [\underline{Y}]_i := r\cdot[\underline{X}]_i; \\ & \text{set-rounding-up;} & \text{for } i=1\ldots n \text{ do } [\overline{Y}_i] := r\cdot[\overline{X}]_i \,\} \\ \text{else } \{ & \text{set-rounding-down;} & \text{for } i=1\ldots n \text{ do } [\underline{Y}]_i := r\cdot[\overline{X}]_i; \\ & \text{set-rounding-up;} & \text{for } i=1\ldots n \text{ do } [\overline{Y}_i] := r\cdot[\underline{X}]_i\}; \end{array} \tag{100}$$

where $[\underline{X}]_i$, $[\overline{X}]_i$ denote the lower, upper bound of $[X]_i$, respectively. If we compare (99) and (100) for $n = 100$ we get

	comparisons	rounding switches	cycles
(99) [traditional]	n	$2n$	2663
(100) [BIAS]	1	2	546

In this way simple observations do imply vast performance improvements. The same method as above is applicable to matrix-matrix multiplication. Let $R \in \mathrm{M}_{nn}(\mathbb{IR})$, $[A] \in \mathbb{IM}_{nn}(\mathbb{IR})$, then contrary to the standard implementation for $[C] := R \cdot [A]$,

$$[C]_{ij} := \sum_{k=1}^{n} R_{ik} \cdot [A]_{kj}, \tag{101}$$

the BIAS implementation is a rowwise update of $[C]$:

$$\begin{array}{l} \text{for } \quad i = 1\ldots n \text{ do} \\ \qquad [C]_{i*} := 0 \\ \qquad \text{for } j = 1\ldots n \text{ do } [C]_{i*} = [C]_{i*} + R_{ij} \cdot [A]_{j*} \end{array} \tag{102}$$

Comparing (101) and (102) for $n = 300$ we obtain

	comparisons	rounding switches	computing time
(101) [traditional]	n^3	$2n^3$	22 sec
(102) [BIAS]	n^2	$2n^2$	4 sec

Improvements like (100) and (102) for a number of vector and matrix operations for real and complex operands are implemented in BIAS (cf. [46], [49]).

For the implementation of verification algorithms it is convenient to call interval operations by means of an operator concept. Therefore, a C++ class library PROFIL has been written [47], [49], [50] implementing all real and interval operations for scalars, vectors and matrices including real and interval transcendental functions. The operator concept causes a minor loss in performance which is outweighed by the ease of notation.

The support and speed of specific operations may influence the design and speed of verification algorithms. Consider for example two ways of computing an inclusion of the

inverse of a real matrix $A \in \mathrm{M}_{nn}(\mathbb{R})$. For $R \in \mathrm{M}_{nn}(\mathbb{R})$, $[X] \in \mathbb{IM}_{nn}(\mathbb{R})$, $0 < Y \in \mathrm{M}_{nn}(\mathbb{R})$

$$R + (I - RA) \cdot [X] \subsetneq [X] \quad \Rightarrow \quad A^{-1} \in [X] \tag{103}$$

$$|R \cdot (I - AR)| + |I - RA| \cdot Y < Y \quad \Rightarrow \quad A^{-1} \in R + [-Y, Y] \tag{104}$$

holds. This is a consequence of Theorems 2.1 and 1.7. Obviously, the first formula (103) looks much simpler. In a practical implementation this changes. The actual computing times for $n = 300$ on the IBM RS/6000 for (103) are

$[C] := I - RA$	$\mathrm{M}_{nn}(\mathbb{R}) \times \mathrm{M}_{nn}(\mathbb{R}) \to \mathbb{IM}_{nn}(\mathbb{R})$	3.6 sec
$[C] \cdot [X]$	$\mathbb{IM}_{nn}(\mathbb{R}) \times \mathbb{IM}_{nn}(\mathbb{R}) \to \mathbb{IM}_{nn}(\mathbb{R})$	24.4 sec
		28.0 sec

In (104) $|I-RA|$ can be computed as a product of *real matrices* with rounding upwards and downwards, storing the absolute value, which is a real matrix again. The multiplication $|I - RA| \cdot Y$ can be performed as a *real* matrix multiplication with rounding upwards. This yields for (104):

$[\mathrm{Res}] := I - AR$	$\mathrm{M}_{nn}(\mathbb{R}) \times \mathrm{M}_{nn}(\mathbb{R}) \to \mathbb{IM}_{nn}(\mathbb{R})$	3.6 sec
$\lvert R \cdot [\mathrm{Res}]\rvert$	$\mathrm{M}_{nn}(\mathbb{R}) \times \mathbb{IM}_{nn}(\mathbb{R}) \to \mathrm{M}_{nn}(\mathbb{R})$	4.1 sec
$C := \lvert I - RA\rvert$	$\mathrm{M}_{nn}(\mathbb{R}) \times \mathrm{M}_{nn}(\mathbb{R}) \to \mathrm{M}_{nn}(\mathbb{R})$	3.6 sec
$C \cdot Y$	$\mathrm{M}_{nn}(\mathbb{R}) \times \mathrm{M}_{nn}(\mathbb{R}) \to \mathrm{M}_{nn}(\mathbb{R})$	1.8 sec
		13.1 sec

Therefore we see that (104), which at the first sight seems to be more expensive than (103), is in the actual implementation faster by a factor of 2. All these computing times were achieved with general purpose, unblocked algorithms not tuned for a specific machine. In blocked versions both computing times improve and the ratio stays approximately the same.

As a user of traditional floating point algorithms using a standard compiler with optimization, one may ask, how much performance loss (or gain) one has to expect when going to verification methods. The *comparison* is still not fair, of course, because the verification algorithm gives rigorous information and the verification of correctness. But life is sometimes not fair. Comparisons like this can be found in [51] or [40], [41].

For the solution of a dense system of linear equations a standard floating point algorithm can be found in LAPACK [10]. We used F2C [22] to transform the LAPACK-code from Fortran into C which might cause a loss in performance. We compare to a verification algorithm based on Theorem 4.1 [74] and used the unblocked general purpose BIAS/PROFIL routines which are, as all algorithms BIAS/PROFIL, not specialized to

a specific architecture. Using blocked versions gains a lot and adaptation to the specific architecture again gains performance. For an $n \times n$ linear system, we obtained the following results on the IBM RS/6000 Model 370. They demonstrate that when using BIAS/PROFIL the theoretical factor 6 is actually achieved.

n	LAPACK	Inclusion Method using PROFIL	
		point data	interval data
100	0.03 sec	0.24 sec	0.27 sec
200	0.3 sec	1.7 sec	1.9 sec
300	1.0 sec	6.4 sec	7.2 sec

Example: Solution of $Ax = b$

Recently, Corliss [19] presented test suites for comparing interval libraries in accuracy and speed. His test results for several libraries including BIAS/PROFIL can be found in [49].

The PROFIL / BIAS library [49] is constantly under development. Recently, a test matrix library, a list handling module, an automatic differentiation module, and several miscellaneous functions have been added [50]. The libraries BIAS and PROFIL and extensions are available in source code via anonymous ftp for non-commercial use, ready to use for IBM RS/6000, HP 9000/700, SUN Sparc and PC's with coprocessor. This also includes the documentation [46], [47], [50].

8. Conclusion

The presented theorems on general dense and sparse systems of equations can be specialized or extended to many standard problems in numerical analysis. Frequently, the special structure can be used to prove more general assertions under weaker assumptions (see, for example, the algebraic eigenvalue problem in Chapter 5).

For polynomials there are several interesting methods described by Böhm [15]. These include multivariate polynomials, simultaneous inclusion of all zeros and inclusion of clusters of zeros. Also, a generalization of the Theorem of Gargantini/Henrici [23] is given which *constructs* inclusion intervals rather than refines them.

Specific theorems can be given for linear, quadratic and convex programming problems [53], [75]. In the case of linear programming problems, Jansson treated the basis unstable case ([36] and his paper in this volume). This interesting work allows presentation of several solutions to the user that are optimal w.r.t. some data within the tolerances, and offers more freedom in the choice of the solution.

This paper summarizes some basic principles for computing an inclusion of the solution of dense and sparse systems of equations. There are many other methods, and many more

details which could not be treated due to limited space. We apologize to authors for not being mentioned.

We are still at the beginning, and the work is very much in progress. The fruitful combination of numerical methods and verification methods is very promising. This monograph is written in this spirit, and we hope we could pass this to the reader.

Acknowledgement. The author wants to thank the referees for the thorough reading and for very many helpful remarks.

REFERENCES

1. J.P. Abbott and R.P. Brent. Fast Local Convergence with Single and Multistep Methods for Nonlinear Equations. *Austr. Math. Soc. 19 (Series B)*, pages 173–199, 1975.
2. ACRITH High-Accuracy Arithmetic Subroutine Library, Program Description and User's Guide. IBM Publications, No. SC 33-6164-3, 1986.
3. G. Alefeld. Zur Durchführbarkeit des Gaußschen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten. In R. Albrecht and U. Kulisch, editors, *Grundlagen der Computer-Arithmetik*, volume 1. COMPUTING Supplementum, 1977.
4. G. Alefeld. Intervallanalytische Methoden bei nichtlinearen Gleichungen. In S.D. Chatterji et al., editor, *Jahrbuch Überblicke Mathematik 1979*, pages 63–78. Bibliographisches Institut, Mannheim, 1979.
5. G. Alefeld. Rigorous Error Bounds for Singular Values of a Matrix Using the Precise Scalar Product. In E. Kaucher, U. Kulisch, and Ch. Ullrich, editors, *Computerarithmetic*, pages 9–30. Teubner Stuttgart, 1987.
6. G. Alefeld. Inclusion Methods for Systems of Nonlinear Equations. In J. Herzberger, editor, *Topics in Validated Computations — Studies in Computational Mathematics*, pages 7–26, Amsterdam, 1994. North-Holland.
7. G. Alefeld and J. Herzberger. *Einführung in die Intervallrechnung.* B.I. Wissenschaftsverlag, 1974.
8. G. Alefeld and J. Herzberger. *Introduction to Interval Computations.* Academic Press, New York, 1983.
9. F.L. Alvarado. Practical Interval Matrix Computations. talk at the conference "Numerical Analysis with Automatic Result Verification", Lafayette, Louisiana, February 1993.
10. E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, and S. Hammarling. *LAPACK User's Guide.* SIAM Publications, Philadelphia, 1992.
11. H. Bauch, K.-U. Jahn, D. Oelschlägel, H. Süsse, and V. Wiebigke. *Intervallmathematik, Theorie und Anwendungen*, volume Bd. 72 of *Mathematisch-naturwissenschaftliche Bibliothek.* B.G. Teubner, Leipzig, 1987.
12. F.L. Bauer. Optimally scaled matrices. *Numerische Mathematik 5*, pages 73–87, 1963.
13. H. Behnke. *Die Bestimmung von Eigenwertschranken mit Hilfe von Variationsmethoden und Intervallarithmetik.* Dissertation, Inst. für Mathematik, TU Clausthal, 1989.
14. R. Bellmann. *Adaptive Control Processes.* Princeton University Press, 1975.
15. H. Böhm. *Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrucke mit garantierter, maximaler Genauigkeit.* Dissertation, University of Karlsruhe, 1983.
16. C.G. Broyden. A new method of solving nonlinear simultaneous equations. *Comput. J.*, 12:94–99, 1969.
17. L. Collatz. Einschließungssatz für die charakteristischen Zahlen von Matrizen. *Math. Z.*,

48:221–226, 1942.
18. D. Cordes and E. Kaucher. Self-Validating Computation for Sparse Matrix Problems. In *Computer Arithmetic: Scientific Computation and Programming Languages*. B.G. Teubner Verlag, Stuttgart, 1987.
19. G.F. Corliss. Comparing software packages for interval arithmetic. Preprint presented at SCAN'93, Vienna, 1993.
20. J.W. Demmel. The Componentwise Distance to the Nearest Singular Matrix. *SIAM J. Matrix Anal. Appl.*, 13(1):10–19, 1992.
21. I.S. Duff, A.M. Erisman, and J.K. Reid. *Direct Methods for Sparse Matrices*. Clarendon Press, Oxford, 1986.
22. S.I. Feldman, D.M. Gay, M.W. Maimone, and N.L. Schreyer. A Fortran-to-C Converter. Computing Science Technical Report 149, AT&T Bell Laboratories, Murray Hill, NJ., 1991.
23. I. Gargantini and P. Henrici. Circular Arithmetic and the Determination of Polynomial Zeros. *Numer. Math.*, 18:305–320, 1972.
24. G.H. Golub and Ch. Van Loan. *Matrix Computations*. Johns Hopkins University Press, second edition, 1989.
25. R.T. Gregory and D.L. Karney. *A Collection of Matrices for Testing Computional Algorithms*. John Wiley & Sons, New York, 1969.
26. A. Griewank. *On Automatic Differentiation*, volume 88 of *Mathematical Programming*. Kluwer Academic Publishers, Boston, 1989.
27. R. Hamming. *Introduction to Applied Numerical Analysis*. McGraw Hill, New York, 1971.
28. E.R. Hansen. On Solving Systems of Equations Using Interval Arithmetic. *Math. Comput. 22*, pages 374–384, 1968.
29. E.R. Hansen. On the Solution of Linear Algebraic Equations with Interval Coefficients. *Linear Algebra Appl. 2*, pages 153–165, 1969.
30. E.R. Hansen. A generalized interval arithmetic. In K. Nickel, editor, *Interval Mathematics*, volume 29, pages 7–18. Springer, 1975.
31. E.R. Hansen. *Global Optimization using Interval Analysis*. Marcel Dekker, New York, 1992.
32. G. Heindl, 1993. private communication.
33. H. Heuser. *Lehrbuch der Analysis*, volume Band 2. B.C. Teubner, Stuttgart, 1988.
34. *IEEE 754 Standard for Floating-Point Arithmetic*, 1986.
35. *ANSI/IEEE 854-1987, Standard for Radix-Independent Floating-Point Arithmetic*, 1987.
36. C. Jansson. *Zur linearen Optimierung mit unscharfen Daten*. Dissertation, Universität Kaiserslautern, 1985.
37. C. Jansson. A Geometric Approach for Computing A Posteriori Error Bounds for the Solution of a Linear System. *Computing*, 47:1–9, 1991.
38. C. Jansson. A Global Minimization Method: The One-Dimensional Case. Technical Report 91.2, Forschungsschwerpunkt Informations- und Kommunikationstechnik, TU Hamburg-Harburg, 1991.
39. C. Jansson. Interval Linear Systems with Symmetric Matrices, Skew-Symmetric Matrices, and Dependencies in the Right Hand Side. *Computing*, 46:265–274, 1991.
40. C. Jansson. A Global Optimization Method Using Interval Arithmetic. In L. Atanassova and J. Herzberger, editors, *Computer Arithmetic and Enclosure Methods*, IMACS, pages 259–267. Elsevier Science Publishers B.V., 1992.
41. C. Jansson and O. Knüppel. A Global Minimization Method: The Multi-dimensional case. Technical Report 92.1, Forschungsschwerpunkt Informations- und Kommunikationstechnik, TU Hamburg-Harburg, 1992.
42. C. Jansson and S.M. Rump. Algorithmen mit Ergebnisverifikation — einige Bemerkungen

zu neueren Entwicklungen. In *Jahrbuch Überblicke Mathematik 1994*, pages 47–73. Vieweg, 1994.
43. W.M. Kahan. A More Complete Interval Arithmetic. *Lecture notes for a summer course at the University of Michigan*, 1968.
44. W.M. Kahan. The Regrettable Failure of Automated Error Analysis. A Mini-Course prepared for the conference at MIT on Computers and Mathematics, 1989.
45. R.B. Kearfott, M. Dawande, K. Du, and C. Hu. INTLIB: A portable Fortran-77 elementary function library. *Interval Comput.*, 3(5):96–105, 1992.
46. O. Knüppel. BIAS — Basic Interval Arithmetic Subroutines. Technical Report 93.3, Forschungsschwerpunkt Informations- und Kommunikationstechnik, Inst. f. Informatik III, TU Hamburg-Harburg, 1993.
47. O. Knüppel. PROFIL — Programmer's Runtime Optimized Fast Interval Library. Technical Report 93.4, Forschungsschwerpunkt Informations- und Kommunikationstechnik, TUHH, 1993.
48. O. Knüppel. *Einschließungsmethoden zur Bestimmung der Nullstellen nichtlinearer Gleichungssysteme und ihre Implementierung.* PhD thesis, Technische Universität Hamburg-Harburg, 1994.
49. O. Knüppel. PROFIL / BIAS — A Fast Interval Library. *Computing*, 53:277–287, 1994.
50. O. Knüppel and T. Simenec. PROFIL/BIAS extensions. Technical Report 93.5, Forschungsschwerpunkt Informations- und Kommunikationstechnik, Technische Universität Hamburg-Harburg, 1993.
51. C.F. Korn. *Die Erweiterung von Software-Bibliotheken zur effizienten Verifikation der Approximationslösung linearer Gleichungssysteme.* PhD thesis, Universität Basel, 1993.
52. R. Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. *Computing*, 4:187–201, 1969.
53. R. Krawczyk. Fehlerabschätzung bei linearer Optimierung. In K. Nickel, editor, *Interval Mathematics*, volume 29 of *Lecture Notes in Computer Science*, pages 215–222. Springer Verlag, Berlin, 1975.
54. R. Krawczyk and A. Neumaier. Interval Slopes for Rational Functions and Associated Centered Forms. *SIAM J. Numer. Anal.*, 22(3):604–616, 1985.
55. U. Kulisch. *Grundlagen des numerischen Rechnens (Reihe Informatik 19).* Bibliographisches Institut, Mannheim, Wien, Zürich, 1976.
56. U. Kulisch and W.L. Miranker. *Computer Arithmetic in Theory and Practice.* Academic Press, New York, 1981.
57. MATLAB User's Guide. The MathWorks Inc., 1987.
58. G. Mayer. Enclosures for Eigenvalues and Eigenvectors. In L. Atanassova and J. Herzberger, editors, *Computer Arithmetic and Enclosure Methods*, IMACS, pages 49–67. Elsevier Science Publisher B.V., 1992.
59. G. Mayer. Epsilon-inflation in verification algorithms. *J. Comput. Appl. Math.*, 60:147–169, 1993.
60. G. Mayer. Taylor-Verfahren für das algebraische Eigenwertproblem. *ZAMM*, 73(718):T857 – T860, 1993.
61. R.E. Moore. *Interval Analysis.* Prentice-Hall, Englewood Cliffs, N.J., 1966.
62. J.J. Morè and M.Y. Cosnard. Numerical solution of non-linear equations. *ACM Trans. Math. Software*, 5:64–85, 1979.
63. M.R. Nakao. A Numerical Verification Method for the Existence of Weak Solutions for Nonlinear Boundary Value Problems. *Journal of Mathematical Analysis and Applications*, 164:489–507, 1992.

64. A. Neumaier. Existence of solutions of piecewise differentiable systems of equations. *Arch. Math.*, 47:443–447, 1986.
65. A. Neumaier. *Rigorous Sensitivity Analysis for Parameter-Dependent Systems of Equations.* J. Math. Anal. Appl. 144, 1989.
66. A. Neumaier. *Interval Methods for Systems of Equations.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.
67. S. Oishi. Two topics in nonlinear system analysis through fixed point theorems. *IEICE Trans. Fundamentals*, E77-A(7):1144–1153, 1994.
68. S. Poljak and J. Rohn. Checking Robust Nonsingularity Is NP-Hard. *Math. of Control, Signals, and Systems 6*, pages 1–9, 1993.
69. L.B. Rall. Automatic Differentiation: Techniques and Applications. In *Lecture Notes in Computer Science 120.* Springer Verlag, Berlin-Heidelberg-New York, 1981.
70. H. Ratschek and J. Rokne. *Computer Methods for the Range of Functions.* Halsted Press (Ellis Horwood Limited), New York (Chichester), 1984.
71. G. Rex and J. Rohn. A Note on Checking Regularity of Interval Matrices. *Linear and Multilinear Algebra 39*, pages 259–262, 1995.
72. J. Rohn. A New Condition Number for Matrices and Linear Systems. *Computing*, 41:167–169, 1989.
73. J. Rohn. Enclosing Solutions of Linear Interval Equations is NP-Hard. *Proceedings of the SCAN-93 conference Vienna*, 1993.
74. S.M. Rump. *Kleine Fehlerschranken bei Matrixproblemen.* PhD thesis, Universität Karlsruhe, 1980.
75. S.M. Rump. Solving Algebraic Problems with High Accuracy. Habilitationsschrift. In U.W. Kulisch and W.L. Miranker, editors, *A New Approach to Scientific Computation*, pages 51–120. Academic Press, New York, 1983.
76. S.M. Rump. New Results on Verified Inclusions. In W.L. Miranker and R. Toupin, editors, *Accurate Scientific Computations*, pages 31–69. Springer Lecture Notes in Computer Science 235, 1986.
77. S.M. Rump. Algebraic Computation, Numerical Computation, and Verified Inclusions. In R. Janßen, editor, *Trends in Computer Algebra*, pages 177–197. Lecture Notes in Computer Science 296, 1988.
78. S.M. Rump. Guaranteed Inclusions for the Complex Generalized Eigenproblem. *Computing*, 42:225–238, 1989.
79. S.M. Rump. Rigorous Sensitivity Analysis for Systems of Linear and Nonlinear Equations. *Math. of Comp.*, 54(10):721–736, 1990.
80. S.M. Rump. Estimation of the Sensitivity of Linear and Nonlinear Algebraic Problems. *Linear Algebra and its Applications (LAA)*, 153:1–34, 1991.
81. S.M. Rump. On the Solution of Interval Linear Systems. *Computing*, 47:337–353, 1992.
82. S.M. Rump. Validated Solution of Large Linear Systems. In R. Albrecht, G. Alefeld, and H.J. Stetter, editors, *Validation numerics: theory and applications*, volume 9 of *Computing Supplementum*, pages 191–212. Springer, 1993.
83. S.M. Rump. Zur Außen- und Inneneinschließung von Eigenwerten bei toleranzbehafteten Matrizen. *Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)*, 73(7-8):T861–T863, 1993.
84. H. Schwandt. An Interval Arithmetic Approach for the Construction of an almost Globally Convergent Method for the Solution of the Nonlinear Poisson Equation on the Unit Square. *SIAM J. Sci. Stat. Comp.*, 5(2):427–452, 1984.
85. H. Schwandt. The Interval Bunemann Algorithm for Arbitrary Block Dimension. In R.

Albrecht, G. Alefeld, and H.J. Stetter, editors, *Validation Numerics*, volume 9, pages 213–232. COMPUTING Supplementum, 1993.
86. R. Skeel. Iterative Refinement Implies Numerical Stability for Gaussian Elimination. *Math. of Comp.*, 35(151):817–832, 1980.
87. B.T. Smith, J.M. Boyle, J.J. Dongarra, B.S. Garbow, Y. Ikebe, V.C. Klema, and C.B. Moler. Matrix Eigensystem Routines — EISPACK Guide. volume 6 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, 1976.
88. G.W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.
89. J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford Univerity Press, Oxford, 1969.
90. J.H. Wilkinson. Modern Error Analysis. *SIAM Rev. 13*, pages 548–568, 1971.