# Estimation of the Sensitivity of linear and nonlinear algebraic problems[*]

by S. M. Rump, Hamburg

**Abstract**

Methods are presented for performing a rigorous sensitivity analysis for general systems of linear and nonlinear equations w.r.t. weighted perturbations in the input data. The weights offer the advantage that all or part of the input data may be perturbed e.g. relatively or absolutely. System zeroes may, depending on the application, stay zero or not.
The main purpose of the paper is to give methods for computing *rigorous bounds* on the sensitivity of each individual component of the solution on the computer. The methods presented are very effective with the additional property that, due to an automatic error control mechanism, every computed result is guaranteed to be correct. Examples are given for linear and nonlinear systems demonstrating that the computed bounds are in general very sharp. Interesting comparisons to traditional condition numbers are given.
For linear systems the solution set for finite perturbations in the coefficients is estimated. Moreover, some theoretical results for eigenvectors/values and singular values are given.

**AMS Subject Classification:** 65G10, 65G05, 65H10

**Keywords and Phrases:** Sensitivity, perturbation theory, linear systems, nonlinear systems, guaranteed bounds.

# 1 Introduction

Let $K$ denote one of the sets $\mathbb{R}$ (real numbers) or $\mathbb{C}$ (complex numbers). Vectors $v \in K^n$ and matrices $A \in K^{n\times n}$ consist of $n$ resp. $n \times n$ components. Let $T$ denote one of the sets $K$,

[*] published in LAA, 153:1–34, 1991

$VK$, or $MK$. The power set over those sets is denoted by $\mathbb{P}\mathrm{T}$, $\mathbb{P}\mathrm{VT}$, $\mathbb{P}\mathrm{MT}$, respectively. For a set of real or complex floating-point numbers $\mathbb{F} \subseteq T$ let $S$ denote one of the sets $\mathbb{F}$, $\mathrm{V}\mathbb{F}$, or $\mathrm{M}\mathbb{F}$.

If not stated otherwise operations $+, -, \cdot, /$ are throughout this paper power set operations, defined in the usual way. Sets occuring several times in an expression are treated independently, e.g.

$$Z \in \mathbb{P}\mathrm{T} : \quad Z * Z := \{\, z_1 * z_2 \mid z_1, z_2 \in Z \,\} \quad \{\, z * z \mid z \in Z \,\}$$

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

Intervals over $T$ resp. $S$ are defined in the usual way by

$$[X] \in \mathbb{I}\mathrm{T} : [X] = \{\, x \in T \mid \underline{x} \le x \le \overline{x} \,\} \quad \text{for } \underline{x}, \overline{x} \in T \quad \text{and}$$

$$[X] \in \mathbb{I}\mathrm{S} : [X] = [\underline{x}, \overline{x}] = \{\, x \in T \mid \underline{x} \le x \le \overline{x} \,\} \quad \text{for } \underline{x}, \overline{x} \in S$$

where in the case of interval vectors and matrices the induced componentwise ordering is used. Interval operations

$$\Diamond\!\!\!\!\!\ast\; : \mathbb{I}\mathrm{T} \times \mathbb{I}\mathrm{T} \to \mathbb{I}\mathrm{T} \;\text{ resp. }\; \Diamond\!\!\!\!\!\ast\; : \mathbb{I}\mathrm{S} \times \mathbb{I}\mathrm{S} \to \mathbb{I}\mathrm{S} \quad \text{for } * \in \{+, -, \cdot, /\}$$

can be defined using the rounding $\Diamond : \mathbb{P}\mathrm{T} \to \mathbb{I}\mathrm{T}$ resp. $\Diamond : \mathbb{P}\mathrm{T} \to \mathbb{I}\mathrm{S}$ and

$$A \;\Diamond\!\!\!\!\!\ast\; B := \Diamond(A * B).$$

The definition holds similar for interval vectors and matrices. There are very effective implementations for all those interval operations (cf. [3], [20], [11]).

Infimum $\inf(z)$ and supremum $\sup(z)$ of nonempty and bounded sets $Z \in \mathbb{P}\mathrm{T}$ resp. $Z \in \mathbb{P}\mathrm{S}$ are defined in the usual way, in case of vectors and matrices componentwise (that means for $A \in \mathbb{P}\mathrm{MT}$ is $\inf(A) \in MT$). The diameter $d(Z)$ and the radius $r(Z)$ of some nonempty, bounded $Z \in \mathbb{P}\mathrm{T}$ resp. $Z \in \mathbb{P}\mathrm{S}$ are defined by

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

The diameter of $A \in \mathbb{P}\mathrm{MT}$ is the matrix of diameters. $q$ denotes the $n$-dimensional version of the Hausdorff metric over $\mathbb{P}\mathrm{T}$ (cf. [3]). The definitions extend immediately to $\mathbb{I}\mathrm{T}$ resp. $\mathbb{I}\mathrm{S}$ using the canonical embedding.

# 2 Sensitivity of the Solution of a System of Nonlinear Equations

Let a parametrized nonlinear function $f : D_p \times D_n \to \mathbb{R}^n$ with suitable differentiability properties be given where $D_p \subseteq \mathbb{R}^p,\ D_n \subseteq \mathbb{R}^n$. For the parametrized nonlinear equations $f_c(x) = 0$ where $f_c : D_n \to \mathbb{R}^n$ and $f_c(x) := f(c, x)$, we are seeking a componentwise sensitivity of an individual zero $\widehat{x}$ of $f_{\widehat{c}}$ for fixed $\widehat{c} \in \operatorname{int}(D_p)$ to perturbations in $\widehat{c}$. The perturbations in $\widehat{c}$ are allowed to be weighted by some $c^* \in \mathbb{R}^p$, $c^* \geq 0$ which means that we are looking for zeros of $f_{\widetilde{c}}$ where $|\widetilde{c} - \widehat{c}| \leq \varepsilon \cdot |c^*|$ for $\varepsilon \to 0$.

Weighted perturbations bear the advantage that zero parameters may stay zero or not, depending on the application. Our aim is to give rigorous lower and upper bounds for the sensitivity which can be calculated on digital computers, including all rounding errors during the evaluation.

More precisely our general assumptions for $f$ are the following:

$$f : D_p \times D_n \to \mathbb{R}^n \text{ with } D_p \subseteq \mathbb{R}^p,\ D_n \subseteq \mathbb{R}^n \text{ and } f \in C^2(D_p \times D_n). \tag{1}$$

Slightly weaker assumptions are possible for the following; for simplicity we use (2.1). Assume $\widehat{x}$ is a simple zero of $f_{\widehat{c}}$, $\widehat{c} \in \operatorname{int}(D_p)$, and let $C_\varepsilon := \{\, \widetilde{c} \mid |\widetilde{c} - \widehat{c}| \leq \varepsilon \cdot |c^*| \,\}$, $0 < \varepsilon \in \mathbb{R}$, $0 \leq c^* \in \mathbb{R}^p$. Because $\widehat{x}$ is simple for small enough $\varepsilon$ and every $\widetilde{c} \in C_\varepsilon$ there is a uniquely determined zero $\widetilde{x} \in U_\delta(\widehat{x})$ of $f_{\widetilde{c}}$. Therefore, for small enough $\varepsilon$, the set

$$\Sigma(f, C_\varepsilon, \widehat{x}) := \{\, \widetilde{x} \in U_\delta(\widehat{x}) \mid \exists\, \widetilde{c} \in C_\varepsilon : f_{\widetilde{c}}(\widetilde{x}) = 0 \,\} \tag{2}$$

is well-defined and connected.

**Definition 2.1.** Let $f$ with (2.1) be given. Then the (absolute) sensitivity of the k$^{th}$ component, $1 \leq k \leq n$, of the simple zero $\widehat{x}$ of $f_{\widehat{c}}$, $\widehat{c} \in \operatorname{int}(D_p)$, to perturbations in $\widehat{c}$ weighted by $c^*$ is defined by

$$\operatorname{Sens}_k(\widehat{x}, f, c^*) := \lim_{\varepsilon \to 0^+} \frac{\operatorname{rad}\Big(\Sigma(f, C_\varepsilon, \widehat{x})\Big)_k}{\varepsilon}$$

Obviously an equivalent definition of the sensitivity is

$$\operatorname{Sens}_k(\widehat{x}, f, c^*) = \lim_{\varepsilon \to 0^+} \max \left\{ \frac{|\widetilde{x}_k - \widehat{x}_k|}{\varepsilon} \;\Big|\; f_{\widetilde{c}}(\widetilde{x}) = 0 \text{ for } \widetilde{c} \in \mathbb{C}_\varepsilon, \widetilde{x} \text{ is connected to } \widehat{x} \right\}.$$

The vector of sensitivities of $\widehat{x}$ is denoted by $\operatorname{Sens}(\widehat{x}, f, c^*)$. In contrast to traditional perturbation theory, where the distance between $\widetilde{c}$ and $\widehat{c}$ is frequently bounded by some norm, here we investigate the sensitivity to perturbations $\widetilde{c}$ of $\widehat{c}$ having an absolute distance to

$\widehat{c}$ bounded by the weights $c^*$. In our approach the weights may switch smoothly between a relative and an absolute distance in the $|\cdot|$-sense [taking $c^* := |\widehat{c}|$ or $c^* := (1,\ldots,1)^T$, respectively] for each individual parameter $c_i,\ 1 \le i \le p$. This offers a great flexibility for practical applications.

For calculating an inclusion of a zero $\widehat{x}$ of $f_{\widehat{c}},\ \widehat{c} \in \mathrm{int}(D_p)$ we use the following theorem (see [26]).

**Theorem 2.2.** Let $f$ with (2.1) be given, let $\widetilde{x} \in D_n,\ R \in \mathbb{R}^{n\times n}$, and $\emptyset \neq X \in \mathbb{IR}^n$ such that $\overline{x} + (0\underline{\cup}X) \subseteq D_n$. Define for $c \in \mathrm{int}(D_p)$, $Y \in \mathbb{PR}^n$ with $Y \subseteq D_n$,

$$J(c,Y) := \bigcap \left\{ M \in \mathbb{IR}^{n\times n} \;\Big|\; \frac{\partial f}{\partial x}(c,y) \in M \text{ for all } y \in Y \right\}. \tag{3}$$

If for some $\widehat{c} \in \mathrm{int}(D_p)$

$$-R\cdot f(\widehat{c},\widetilde{x}) + \Big\{ I - R\cdot J\Big(\widehat{c},\overline{x} + (0\underline{\cup}X)\Big)\Big\}\cdot X \subseteq \mathrm{int}(X), \tag{4}$$

then $R$ and every matrix $M \in J\Big(\widehat{c},\overline{x} + (0\underline{\cup}X)\Big)$ are not singular, and there is a unique and simple zero $\widehat{x}$ of $f_{\widehat{c}}$ in $\overline{x} + \mathrm{int}(X)$.

**Remark.** All operations in the above Theorem are power set operations.

It is a straightforward generalization of Theorem 2.2 to replace $\widehat{c}$ by some

$$C_\varepsilon := \{\, \widetilde{c} \mid |\widetilde{c} - \widehat{c}| \le \varepsilon \cdot |c^*| \}$$

for some $c^* \in \mathbb{R}^p$, $c^* \ge 0$. For small enough $\varepsilon$ (2.4) remains valid and we conclude that every $f_c,\ c \in C_\varepsilon$, has a unique and simple zero $\widehat{x}_c$ within $\overline{x} + \mathrm{int}(X)$. $\Big[f(C_\varepsilon,\widetilde{x})$ is defined as usual by $\{\, f(\widetilde{c},\overline{x}) \mid \widetilde{c} \in C_\varepsilon\}\Big]$.

Using Theorem 2.2 can already give upper bounds for the sensitivity (see Figure 1).

**Fig. 1**

To obtain lower bounds for this set we need to find a hyperrectangle $Y$ with the property that for every hyperplane bounding $Y$ there are points in $\Sigma(f,C_\varepsilon,\widehat{x})$ going beyond it. This is accomplished by the following theorem [28].

**Theorem 2.3.** Let $f$ satisfying (2.1) be given; let $\overline{x} \in D_n$, $R \in \mathbb{IR}^{n\times n}$. For $\varepsilon > 0$ and some $\widehat{c} \in \operatorname{int}(D_p)$ let $\emptyset \neq X_\varepsilon \in \mathbb{IIR}^n$ such that $\overline{x} + (0\overline{\cup}X) \subseteq D_n$, let

$$C_\varepsilon := \left\{ \widetilde{c} \in \mathbb{IR}^p \,\middle|\, |\widetilde{c} - \widehat{c}| \leq \varepsilon \cdot |c^*| \right\} \tag{5}$$

for some $c^* \in \mathbb{IR}^p$, $c^* \geq 0$, and $C_\varepsilon \subseteq \operatorname{int}(D_p)$, and define

$$\begin{aligned} Z_\varepsilon &:= \Diamond\Big( -R \cdot f(C_\varepsilon, \overline{x})\Big), \\ \Delta_\varepsilon &:= \Big\{ I - R \cdot J\Big(C_\varepsilon, \overline{x} + (0\underline{\cup}X_\varepsilon)\Big)\Big\} \cdot X_\varepsilon. \end{aligned} \tag{6}$$

If

$$Z_\varepsilon + \Delta_\varepsilon \subseteq \operatorname{int}(X_\varepsilon) \tag{7}$$

then

$$\overline{x} + Z_\varepsilon \overset{\vee}{+} \Delta_\varepsilon \subseteq \Diamond\Sigma(f, C_\varepsilon, \widehat{x}) \subseteq \overline{x} + Z_\varepsilon + \Delta_\varepsilon \tag{8}$$

for $\Sigma(f, C_\varepsilon, \widehat{x})$ as defined in (2.2).

The inward directed addition $\overset{\vee}{+}$ is defined by $X + Y := \Big[\operatorname{inf}(X) + \operatorname{sup}(Y),\ \operatorname{sub}(X) + \operatorname{inf}(Y)\Big]$ for $X, Y \in \mathbb{IT}$ (see [28]).

A heuristic interpretation of (2.8) is that $Z_\varepsilon$ is an "approximation" to the smallest hyperrectangle enclosing $\Sigma(f, C_\varepsilon, \widehat{x})$ wheras adding $\Delta_\varepsilon$ to the vertices of $Z_\varepsilon$ in the proper direction yields an inner and an outer estimate for

**Fig. 2.**

$\Diamond\ \Sigma(f, C_\varepsilon, \widehat{x})$, see Figure 2. In practice $\Delta_\varepsilon$ is small, which implies very sharp bounds.

For (2.8) in Theorem 2.3 it is crucial that $Z_\varepsilon$ is the precise smallest rectangle enclosing $-R\cdot f(C_\varepsilon, \overline{x})$; the latter is defined by $-R\cdot f(C_\varepsilon, \overline{x}) := \{\, x \mid X = -R\cdot f(c, \overline{x}) \text{ for some } c \in C_\varepsilon \,\}$.

This set will not be computed exactly, except in special cases, but rather will be estimated by some $Z_1, Z_2 \in \mathbb{IIR}^n$ with $Z_1 \subseteq \Diamond\Big( - R \cdot f(C_\varepsilon, \overline{x})\Big) \subseteq Z_2$, yielding

$$Z_1 \overset{\vee}{+} \Delta_\varepsilon \subseteq \Diamond\ \Sigma(f, C_\varepsilon, \widehat{x}) \subseteq Z_2 + \Delta_\varepsilon$$

However, in the limit $\varepsilon \to 0$ the size of $\Delta_\varepsilon$ can be estimated, yielding lower and upper bounds for the sensitivity of a zero $\widehat{x}$ of $f_{\widehat{c}}$.

**Theorem 2.4.** Let $f$ satisfying (2.1) be given such that each parameter $c_j$ occurs in at most one component $f_i$ of $f$, let $\overline{x} \in D_n$, $R \in \mathbb{R}^{n \times n}$, and $\emptyset \neq X \in \mathbb{IIR}^n$ such that $\overline{x} + (0 \underline{\cup} X) \subseteq D_n$. Define

$$J(c, Y) := \bigcap \left\{ M \in \mathbb{IIR}^{n \times n} \;\Big|\; \frac{\partial f}{\partial x}(c, y) \in M \quad \text{for all } y \in Y \right\} \tag{9}$$

for $c \in \operatorname{int}(D_p)$, $Y \in \mathbb{IPR}^n$ with $Y \subseteq D_n$, and let

$$-R \cdot f(\widehat{c}, \overline{x}) + \Big\{ I - R \cdot J\Big(\widehat{c}, \overline{x} + (0 \underline{\cup} X)\Big)\Big\} \cdot X \subseteq \operatorname{int}(X) \tag{10}$$

for some $\widehat{c} \in \operatorname{int}(D_p)$. Then there is a unique and simple zero $\widehat{x}$ of $f_{\widehat{c}}$ in $\overline{x} + \operatorname{int}(X)$. Let $c^* \in \mathbb{R}^p$, $c^* \geq 0$, and define

$$\begin{aligned} u &:= |R \cdot | \, \frac{\partial f}{\partial c} \, (\widehat{c}, \widehat{x}) \, | \cdot |c^*|, \\ w &:= |I - R \cdot J\Big(\widehat{c}, \overline{x} + (0 \underline{\cup} X)\Big)| \cdot d(X). \end{aligned} \tag{11}$$

Then

$$\phi := \max_i \frac{u_i}{(d(X) - w)_i} \tag{12}$$

is well defined, and the sensitivity of the zero $\widehat{x}$ of $f_{\widehat{c}}$ to perturbations weighted by $c^*$ satisfies

$$\operatorname{Sens}(\widehat{x}, f, c^*) \in u \pm \phi \cdot w \tag{13}$$

**Remark.** In practical applications an inclusion of $u$ can be calculated by using $\overline{x} + X$ instead of $\widehat{x}$.

**Proof.** According to assumption (2.1) $f$ is differentiable w.r.t. $c$, so that for small enough $\varepsilon > 0$ and $\widetilde{c} \in C_\varepsilon$

$$f(\widetilde{c}, \widehat{x}) = f(\widehat{c}, \widehat{x}) + \frac{\partial f}{\partial c} \, (\widehat{c}, \widehat{x}) \cdot (\widetilde{c} - \widehat{c}) + 0(\varepsilon^2). \tag{14}$$

By assumption, for $1 \le j \le q$ at most one function $f_i$ is depending on $c_j$. This implies that in every column of $\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})$ there is at most one nonzero element. Using the $\widetilde{c}$ and $-\widetilde{c}$ with $\operatorname{sgn}(\widetilde{c} - \widehat{c})_j = \operatorname{sgn}\left(\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})\right)_{ij}$ and observing $f(\widehat{c}, \widehat{x}) = 0$ proves

$$f(C_\varepsilon, \widehat{x}) = \pm\varepsilon \cdot |\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})| \cdot |c^*| + 0(\varepsilon^2).$$

In other words $f(C_\varepsilon, \widehat{x})$ is a full rectangle symmetric to the origin up to terms of $0(\varepsilon^2)$. Therefore

$$\begin{aligned} \diamond\Big(-R \cdot f(C_\varepsilon, \widehat{x})\Big) &= \pm\varepsilon \cdot |R| \cdot |\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})| \cdot |c^*| + \varepsilon^2 \cdot P_\varepsilon \\ &= \pm\varepsilon \cdot u + \varepsilon^2 \cdot P_\varepsilon \end{aligned} \tag{15}$$

for small enough $\varepsilon > 0$, where $P_\varepsilon \in \mathbb{IIR}^n$ is bounded for $\varepsilon \to 0$.

Using $Z := -R \cdot f(\widehat{c}, \overline{x}) \in \mathbb{R}^n$ and

$$C := \Big\{I - R \cdot J\Big(\widehat{c}, \overline{x} + (0\underline{\cup}X)\Big)\Big\} \in \mathbb{IPR}^{n\times n},$$

(2.12) becomes $Z + C \cdot X \subseteq \operatorname{int}(X)$ implying (cf. [27], Lemma 2)

$$C \cdot Y \subseteq \operatorname{int}(Y) \quad \text{for } Y := X - X \in \mathbb{IIR}^n \tag{16}$$

where $X - X = \Big\{\, x_1 - x_2 \mid x_1, x_2 \in X \,\Big\}$. $Y$ is componentwise fully symmetric w.r.t. the origin, i.e. $Y_i = -Y_i$ for $1 \le i \le n$. Let $Y_\kappa \cdot Y$ and $y := |Y| = \max\Big\{\, |x_1 - x_2| \,\Big|\, x_1, x_2 \in X \Big\} = d(X)$. For small enough $\kappa > 0$ we have $\widehat{x} + (0\underline{\cup}Y_\kappa) = \widehat{x} + Y_\kappa \subseteq \overline{x} + X \subseteq D_n$, using $\widehat{x} \in \overline{x} + \operatorname{int}(X)$. Therefore $J\Big(\widehat{c}, \widehat{x} + (0\underline{\cup}Y_\kappa)\Big) \subseteq J\Big(\widehat{c}, \overline{x} + (0\underline{\cup}X)\Big)$, and by using

$$\frac{\partial f}{\partial x}(\widetilde{c}, \widehat{x} + x) = \frac{\partial f}{\partial x}(\widehat{c}, \widehat{x} + x) + 0(\varepsilon),$$

which holds for every $\widetilde{c} \in C_\varepsilon$ and every $x \in 0\underline{\cup}Y_\kappa$ we get

$$J\Big(\widetilde{c}, \widehat{x} + (0\underline{\cup}Y_\kappa)\Big) \subseteq J\Big(\widehat{c}, \overline{x} + (0\underline{\cup}X)\Big) + \varepsilon \cdot Q_\varepsilon \tag{17}$$

for every $\widetilde{c} \in C_\varepsilon$ where $Q_\varepsilon \in \mathbb{IR}^{n\times n}$ is bounded for $\varepsilon \to 0$.

Therefore using (2.15) and (2.17),

$$\begin{aligned} &-R \cdot f(C_\varepsilon, \widehat{x} + \Big\{I - R \cdot J\Big(C_\varepsilon, \widehat{x} + (0\underline{\cup}Y_\kappa)\Big)\Big\} \cdot Y_\kappa \subseteq \\ &\pm\varepsilon u + \varepsilon^2 \cdot P_\varepsilon + \Big\{I - R \cdot J\Big(\widehat{c}, \overline{x} + (0\underline{\cup}X)\Big)\Big\} \cdot Y_\kappa + \varepsilon \cdot R \cdot Q_\varepsilon \cdot Y_\kappa \end{aligned} \tag{18}$$

for every small enough $\kappa > 0$. Using the abbreviation $v := |R| \cdot |Q_\varepsilon| \cdot y$, the right-hand side of (2.18) is surely constained in $\operatorname{int}(Y_\kappa)$ if

$$\varepsilon \cdot u + \varepsilon^2 \cdot |P_\varepsilon| + \kappa \cdot w + \varepsilon \cdot \kappa \cdot v < \kappa \cdot y \tag{19}$$

because $|Y_\kappa = \kappa \cdot |Y| = \kappa \cdot d(X)$. $v$ is bounded for $\varepsilon \to 0$. By (2.16) and $Y_i = -Y_i$, $1 \le i \le n$ we have $C \cdot Y = \pm|C| \cdot |Y| \subseteq \mathrm{int}(Y)$, i.e. $|C| \cdot y < y$ and therefore $0 \le w < y$ implying $y - w - \varepsilon v > 0$ for small enough $\varepsilon$. Define

$$\kappa = \kappa(\varepsilon) := \varepsilon \cdot \max_i \frac{\{u + \varepsilon \cdot |P_\varepsilon|\}_i}{\{y - w - \varepsilon v\}_i} \cdot (1 + \varepsilon) + \varepsilon^2. \tag{20}$$

Then $\kappa > 0$ is well defined for small enough $\varepsilon$, and (2.19) is true. Hence the l.h.s. of (2.21) is contained in $\mathrm{int}(Y_\kappa)$, and therefore the assumptions of Theorem 2.4 are satisfied for $\overline{x} := \widehat{x}$, $X_\varepsilon := Y_\kappa$, and small enough $\varepsilon > 0$. In this case $Z_\varepsilon$ and $\Delta_\varepsilon$ from (2.8), on replacing $X_\varepsilon$ by $Y_\kappa$, compute to

$$Z_\varepsilon = \pm\varepsilon \cdot u + \varepsilon^2 \cdot P_\varepsilon \quad \text{and} \quad \Delta_\varepsilon \subseteq \pm(\kappa \cdot w + \varepsilon\kappa v)$$

according to (2.15) and (2.18). The point is that $Z_\varepsilon$ and $\Delta_\varepsilon$ are symmetric w.r.t. the origin up to terms of $0(\varepsilon^2)$. The inclusion (2.8), together with the definition (2.2) of $\Sigma(f, C_\varepsilon, \widehat{x})$, gives for $1 \le i \le n$

$$\begin{aligned} \left\{u - \varepsilon \cdot |P_\varepsilon| - \frac{\kappa}{\varepsilon} \cdot (w + \varepsilon \cdot v)\right\}_i &\le \frac{\mathrm{inf}\Sigma(f, C_\varepsilon, \widehat{x})}{\varepsilon} \le \frac{\mathrm{sup}\Sigma(f, C_\varepsilon, \widehat{x})}{\varepsilon} \\ &\le \left\{u + \varepsilon \cdot |P_\varepsilon| + \frac{\kappa}{\varepsilon} \cdot (w + \varepsilon \cdot v)\right\}_i \end{aligned} \tag{21}$$

for all small enough $\varepsilon > 0$, with corresponding $\kappa$ given by (2.20). Noting the definition of $\kappa$ and $\phi$ and taking the limit $\varepsilon \to 0$ finishes the proof. ■

It should be stressed that if $f$ is given in explicit form,

$$\frac{\partial f}{\partial x}(c, x) \quad \text{and} \quad \frac{\partial f}{\partial c}(c, x)$$

can be calculated by so-called automatic differentiation. This method has been found and forgotten several times dating back to the forties, and is slowly finding its place in numerical analysis. For details and improvements the reader is referred to [9], [30], [8]. In particular, $J\big(\widehat{c}, \overline{x} + (0 \underline{\cup} X)\big)$ can be calculated and rigorously estimated using interval arithmetic and automatic differentiation. This is performed by replacing all operations with their corresponding interval operations [24].

Operations allowed in the computation of $f$ cover transcendental functions very well, because in [6], [15] algorithms have been described for computing very sharp bounds for $t(X), X \in \mathbb{IC}$, where $t$ is any trigonometric, inverse trigonometric, hyperbolic, inverse hyperbolic, exponential, or logarithmic function, and also for $X^Y, X, Y \in \mathbb{IC}$. Calculating $J\big(\widehat{c}, \overline{x} + (0 \underline{\cup} X)\big)$ using this method normally introduces little overestimation. This is because $X$ encloses the error of $\overline{x}$ and is usually very small.

Theorem 2.4 gives an estimate of the sensitivity of $\widehat{x}$ which can be rigorously calculated without knowing $\widehat{x}$ precisely. All potential errors made by replacing $\widehat{x}$ by $\overline{x}+X$ and by using $J\Big(\widehat{c}, \widetilde{x} + (0\underline{\cup}X)\Big)$ are covered by (2.13).

The exact sensitivity of the zero $\widehat{x}$ of $f_{\widehat{c}}$ to perturbations weighted by $c^*$ is readily obtained as

$$|\frac{\partial f}{\partial x}(\widehat{c}, \widehat{x})^{-1}| \cdot |\frac{\partial f}{\partial c}(\widehat{c}, \widehat{x})| \cdot |c^*|. \tag{22}$$

This confirms the results by Skeel [29] for systems of linear equations. Formula (2.22) can be proved directly; using Theorem 2.5 it can be seen by setting

$$\overline{x} := \widehat{x}, \quad R := \frac{\partial f}{\partial x}(\widehat{c}, \widehat{x})^{-1}$$

and observing that $I - R \cdot \frac{\partial f}{\partial x}(\widehat{c}, x)$ is convergent for every $x$ sufficiently close to $\widehat{x}$, thus allowing one to find a positive vector $y$ with

$$|I - R \cdot \frac{\partial f}{\partial x}(\widehat{c}, x)| \cdot y < y.$$

This is true provided $\widehat{x}$ is a simple zero of $f$

The quality of the estimate (2.13) is essentially determined by the minimum difference of the components of $d(X) - w$ [which, according to (2.16), is always positive]. This difference in turn is small if the spectral radius of $|I - R \cdot J\Big(\widehat{c}, \overline{x} + (0\underline{\cup}X)\Big)|$ is small. in practice the latter value rarely exceeds 1/2 as long as (2.10) holds. Therefore, in view of (2.12), it is likely that in practical applications (2.13) can be written as $\text{Sens}(\widehat{x}, f, c^*) \in u \cdot (1 \pm \delta)$, where $\delta \ll 1/2$. For a sensitivity information this is a satisfactory result, because in practical applications knowing the magnitude of the sensitivity is usually sufficient. This heuristic is verified by the numerical results given in Section 5.

# 3 Sensitivity of polynomial zeroes

As an application of theorem 2.5 we mention the sensitivity of a simple real zero of a polynomial $P \in \mathbb{R}[x]$. We write the problem as a parameterized nonlinear equation

$$f(c, x) : \mathbb{R}^{n+1} x \ \mathbb{R} \to \mathbb{R} \text{ with } f(c, x) := \sum_{i=0} n c_i \cdot x^i. \tag{23}$$

Let $P^*$ be a polynomial with $P^*(x) = \sum_{i=0} n p_i^* \cdot x^i$. We do not assume $P_n^* \neq 0$. Using the canomical isomorphisen we identify $P$ with its vector of coefficients and define

$$P_\varepsilon := \Big\{ \widetilde{P} \,\Big|\, |\widetilde{P} - P| \leq \varepsilon \cdot |P^*| \Big\}.$$

The sensitivity of $\widehat{x}$ w.r.t. perturbation in the coefficients of $P$ weighted by $P^*$ is then - similar to the nonlinear case — defined by

$$\mathrm{Sens}(\widehat{x}, P, P^*) = \lim_{\varepsilon \to 0} \max \left\{ \frac{|\widetilde{x} - \widehat{x}|}{\varepsilon} \,\Big|\, \widetilde{x} \text{ is connected to } P_\varepsilon \right\}.$$

**Theorem 3.1.** Let $P \in \mathrm{IR}[x]$ and $\widetilde{x}, r \in \mathrm{IR}$, $\emptyset \neq X \in \mathrm{IIR}$, $0 \in X$ be given with

$$-r \cdot P(\widetilde{x}) + \Big\{1 - r \cdot P'(\widetilde{x} + X)\Big\} \cdot X \subseteq \mathrm{int}(X). \tag{24}$$

Then there is exactly one root $\widehat{x}$ of $P$ within $\widetilde{x} + X$; $\widehat{x}$ is simple. Let $P^* \in \mathrm{IR}[x]$ be some polynomial of at most the degree of $P$ having non-negative coefficients and let $w := |1 - r \cdot P'(\widetilde{x} + X)| \cdot |X|$. Then the sensitivity of $\widehat{x}$ w.r.t. $\varepsilon$-perturbations in the coefficients of $P$ weighted by $P^*$ satisfies

$$\mathrm{Sens}(\widehat{x}, P, P^*) \in |r| \cdot P^*(|\widetilde{x}|) \cdot \left(1 \pm \frac{w}{|X| - w}\right). \tag{25}$$

**Proof.** Follows by straightforward application of theorem 2.5 to (3.1). ■

In practice $X$ will be obtained by means of an iteration process (see [26], [5]). Unless the problem is extremely ill-conditioned the term $w$ will be very small as compared to $|X|$ due to a small residue $|1 - r \cdot P'(\widetilde{x} + X)|$ where $r \approx P'(\widetilde{x})^{-1}$.

The estimation (3.3) clearly shows how its quality depends on how small is $w$ as compared to $X$. An exact value for the sensitivity is obtained by setting $\widetilde{x} := \widehat{x}$, $r := P'(\widehat{x})^{-1}$ and $X := \kappa \cdot [-1, 1]$. For small enough $\kappa$ (3.2) is satisfied yielding

$$\mathrm{Sens}(\widehat{x}, P, P^*) = |P'(\widehat{x})^{-1}| \cdot |P^*(|\widehat{x}|)$$

repeating a well-known result form perturbation theory [32].

In theorem 3.1 we used direct and independent perturbations of the coefficients $p_i$ of $P$ (weighted by $P^*$). Without going into detail we mention that according to theorem 2.5 any continuously differentiable functional relation between the coefficients of $P$ and those of $P^*$ can be handled, that means the sensitivity of $\widehat{x}$ weighted by $P^*$ in this functional relationship is estimated by (2.15). In practical applications this covers a large class of dependencies between the coefficients of $P$.

# 4 Sensitivity of linear problems for larger perturbations

In this chapter we will derive bounds for the sensitivity of the solution of a system of linear equations $Ax = b$ subject to perturbations in the matrix $A$ and the right hand side $b$ weighted by some nonnegative $A^*, b^*$. We are especially interested in the range of the solution for finite perturbations rather than in the limit for $\varepsilon \to 0$. For this purpose we give the following definition.

**Definition 4.1.** Let $A \in \mathrm{IR}^{n\times n}$, $b \in \mathrm{IR}^n$, $A$ being nonsingular and $\widehat{x} := A^{-1}b$. For nonnegative $A^* \in \mathrm{IR}^{n\times n}$, $b^* \in \mathrm{IR}^n$ and $0 \le \varepsilon \in \mathrm{IR}$ we define

$$\begin{aligned} &A_\varepsilon := \{\, \widetilde{A} \mid |\widetilde{A} - A| \le \varepsilon \cdot |A^*| \,\} \text{ and} \\ &b_\varepsilon := \{\, \widetilde{b} \mid |\widetilde{b} - b| \le \varepsilon \cdot |b^+| \,\}. \end{aligned} \tag{26}$$

Then the $\varepsilon$-elongation of the $k^{\text{th}}$ component of $\widehat{x}$, $1 \le k \le n$ w.r.t. perturbations in $A$ and $b$ weighted by $A^*$ and $b^*$ is defined by

$$\mathrm{Elon}_k^\varepsilon(A^{-1}b, A^*, b^*) := \max \left\{ \frac{|\widetilde{x}_k - \widehat{x}_k|}{\varepsilon} \,\Big|\, \widetilde{A}\widetilde{x} = \widetilde{b} \text{ with } \widetilde{A} \in A_\varepsilon,\ \widetilde{b} \in b_\varepsilon \right\}. \tag{27}$$

The vector of $\varepsilon$-elongations of $A^{-1}b$ is denoted by $\mathrm{Elon}^\varepsilon(A^{-1}b, A^*, b^*)$.

The $\varepsilon$-elongation is the true range of the solution $\widetilde{A}^{-1} \cdot \widetilde{b}$ where $\widetilde{A}$, $\widetilde{b}$ are within the range of $\varepsilon$-perturbations of $A$ and $b$ weighted by $A^*$, $b^*$. In the limit $\varepsilon \to 0$ the $\varepsilon$-elongation coincides with the traditional sensitivity of $\widehat{x}$ w.r.t. perturbations in $A$, $b$ weighted by $A^*$, $b^*$.

The $\varepsilon$-elongation is estimated by the following theorem.

**Theorem 4.2.** Let $A \in \mathrm{IR}^{n\times n}$, $b \in \mathrm{IR}^n$ and $\widetilde{x} \in \mathrm{IR}^n$, $R \in \mathrm{IR}^{n\times n}$, $\emptyset \ne X \in \mathrm{IIR}^n$ with

$$R \cdot (b - A\widetilde{x} + \{I - R \cdot A\} \cdot X \subseteq \mathrm{int}(X). \tag{28}$$

Then $A$ and $R$ are not singular and the unique solution $\widehat{x} := A^{-1}b$ of $Ax = b$ satisfies $\widehat{x} \in \widetilde{x} + \mathrm{int}(X)$.

For nonnegative $A^* \in \mathrm{IR}^{n\times n}$, $b^* \in R^n$ not both being identical zero define

$$\begin{aligned} &u := |R| \cdot (|b^*| + |A^*| \cdot |\widetilde{x}|), \\ &v := |R| \cdot |A^*| \cdot |X| \text{ and} \\ &w := |I - RA| \cdot |X|. \end{aligned} \tag{29}$$

Then both

$$\varepsilon^* := \min_i \left\{ \frac{(|X| - w)_i}{(u+v)_i} \,\Big|\, u_i + v_i \neq 0 \right\} \tag{30}$$

and for $0 \le \varepsilon < \varepsilon^*$

$$\phi_\varepsilon := \max_i \left\{ \frac{u_i}{(|X| - \varepsilon v - w)_i} \,\Big|\, u_i + v_i \neq 0 \right\} \tag{31}$$

are well-defined and it is

$$\mathrm{Elon}^\varepsilon(A^{-1}b, A^*, b^*) \in u \pm \phi_\varepsilon(\varepsilon v + w). \tag{32}$$

**Note.** in practical applications $u$ can be computed by using $\widetilde{x} \in \widetilde{x} + X$.

**Proof.** The first part of the theorem is an immediate consequence of theorem 2.1 for $f : \mathbb{R}^n \to \mathbb{R}^n,\ f(x) := Ax - b$ (see also [26]).

Consider $f : (\mathbb{R}^{n^2} \times \mathbb{R}^n) \times \mathbb{R}^n \to \mathbb{R}^n$ with $n^2 + n$ parameters $A$, $b$ and $f(A, b, x) := Ax - b$. Following the lines of the proof of theorem 2.5 for every $\varepsilon \ge 0$

$$b_\varepsilon - A_\varepsilon \cdot \widehat{x} = \left\{ \widetilde{b} - \widetilde{A}\widehat{x} \,\Big|\, \widetilde{b} \in b_\varepsilon, \widetilde{A} \in A_\varepsilon \right\} = b - A\widehat{x} \pm \varepsilon \cdot (|b^*| + |A^*| \cdot |\widehat{x}|) = \pm\varepsilon \cdot (|b^*| + |A^*| \cdot |\widehat{x}|)$$

implying

$$\begin{aligned} Z_\varepsilon &:= \Diamond f(A_\varepsilon, b_\varepsilon, \widehat{x}) = \Diamond R \cdot (b_\varepsilon - A_\varepsilon \cdot \widehat{x}) \\ &= \pm\varepsilon \cdot |R| \cdot (|b^*| + |A^*| \cdot |\widehat{x}|) = \pm\varepsilon u. \end{aligned} \tag{33}$$

Therefore $P_\varepsilon = 0$ using the notation of the proof of theorem 2.5.

The jacobian of $f$ is identical to $A$ implying $Q_\varepsilon = |A^*|$. If $A^*$ and $b^*$ are not both identical zero then the non-singularity of $R$ implies $u + v \not\equiv 0$. Hence by (4.5) $\varepsilon^*$ is well-defined. For $0 \le \varepsilon < \varepsilon^*$ follows

$$y - w_y - \varepsilon v_y \ge \sigma \cdot (|X| - w - \varepsilon(u+v) + \varepsilon v) > \sigma\varepsilon v \ge 0$$

for the components $k$ with $u_k + v_k \neq 0$ using (4.5) and using $y > w_y$ for the others. Then for every $\delta_1, \delta_2 > 0$

$$\kappa := \varepsilon \cdot \max_i \left\{ \frac{u_i}{\{y - w_y - \varepsilon v_y\}_i} \,\Big|\, u_i + v_i \neq 0 \right\} \cdot (1 + \delta_1) + \delta_2 \tag{34}$$

is well-defined and $\kappa > 0$. Then again using $w_y < y$ for the components $k$ with $u_k + v_k = 0$

$$\varepsilon u + \kappa w_y + \varepsilon\kappa v_y < \kappa(y - w_y - \varepsilon v_y) + \kappa w_y + \varepsilon\kappa v_y = \kappa y$$

which is the equivlaent to (2.22). Thus (2.24) proves

$$\mathrm{Elon}^{\varepsilon}(A^{-1}b, A^{*}, b^{*}) \in u \pm \frac{\kappa}{\varepsilon} \cdot (\varepsilon v_y + w_y)$$

for every $0 \le \varepsilon < \varepsilon^{*}$ and corresponding $\kappa$ with (4.9) for any $\delta_1, \delta_2 > o$. Taking the limit $\delta_1, \delta_2 \to 0$ and regarding $\sigma \cdot \frac{\kappa}{\varepsilon} \to \phi_\varepsilon$ finishes the proof. ■

In a practical application we set $\tilde{x} \approx A^{-1}b, R \approx A^{-1}$. $X$ is obtained by means of an iteration [26]. It can be shown that a properly defined iteration using interval operations finishes with some $X$ satisfying (4.3) if an only if $\rho(|I - RA|) < 1$ (cf. [27] and the following theorem 4.2.1).

Theorem 4.2 serves theoretical purposes to be discussed in the following. In practical applications $\varepsilon^{*}$ from (4.5) is small due to the fact that the exact solution of the linear system is, in general, not exactly representable an approximated by some $\tilde{x}$.

For a practical applications the $\varepsilon$-elongation of the solution of the linear system $Ax = b; A \in [A], b \in [b]$ w.r.t. perturbations weighted by $A^{*} \in \mathbb{R}^{n\times n}, A^{*} \ge 0$ can be estimated by directly computing estimations for $\Sigma(A_\varepsilon, b_\varepsilon)$ where $A_\varepsilon := \left\{ \tilde{A} \,\middle|\, |tildeA - A| \le \varepsilon \cdot A^{*} \right\}$, $b_\varepsilon := \left\{ \tilde{b} \,\middle|\, |\tilde{b} - b| \le \varepsilon \cdot b^{*} \right\}$. This approach has been described in [28]. It is

$$Z_\varepsilon \overset{\vee}{+} \Delta_\varepsilon \subseteq \Diamond\Sigma(A_\varepsilon, b_\varepsilon) - \tilde{x} \subseteq Z_\varepsilon + \Delta_\varepsilon \tag{35}$$

for $Z_\varepsilon := R \cdot (b_\varepsilon - A_\varepsilon \tilde{x})$, $\Delta_\varepsilon := \{I - R \cdot A_\varepsilon\} \cdot X$ where $R \in \mathbb{R}^{n\times n}$, $\tilde{x} \in \mathbb{R}^{n}, X \in \mathbb{IR}^{n}$ and $Z_\varepsilon + \Delta_\varepsilon \subseteq \mathrm{int}(X)$. Using this approach the problem is to find an appropriate $X$ satisfying $Z_\varepsilon + \Delta_\varepsilon \subseteq \mathrm{int}(X)$. Such an $X$ can be determinded by means of an iteration where in each step it is indispensable to apply a so-called $\varepsilon$-inflation introduced in [25]. We define

$$X \in \mathbb{IR}^{n} : x \circ \delta = x \pm \delta \text{ for some } 0 < \delta \in \mathbb{R}^{n}. \tag{36}$$

Here we use $\delta$ to avoid a conflict with the $\varepsilon$ used in (4.10).

For given $X^{0} \in \mathbb{IR}^{n}$ and $C_\varepsilon := I - R \cdot A_\varepsilon$ we define the iteration

$$Y^{k} := X^{k} \circ \delta;\ X^{k+1} := Z_\varepsilon + C_\varepsilon \cdot Y^{k} \text{ for } 0 \le k \in \mathbb{N}. \tag{37}$$

If for some $k \in \mathbb{N}\quad X^{k+1} \subseteq \mathrm{int}(Y^{k})$ then $\Delta_\varepsilon := C_\varepsilon \cdot Y^{k}$ satisfies (4.10). Using iteration (4.12) explicit conditions can be stated under which some $k \in \mathbb{N}$ exists with $X^{k+1} \subseteq \mathrm{int}(Y^{k})$.

**Theorem 4.2.1.** Let $Z_\varepsilon \in \mathbb{IR}^{n}, C_\varepsilon \in \mathbb{IR}^{n\times n}$ and $0 < \delta \in \mathbb{R}^{n}$ be given. Then the following is equivalent:

a) For every $X^{0} \in \mathbb{IR}^{n}$ there exist a $k \in \mathbb{N}$ such that

$$X^{k+1} \subseteq \mathrm{int}(Y^k)$$

using iteration (4.12) for $X^k, Y^k$

b) $\rho(|I - R \cdot A_\varepsilon|) < 1$.

**Note.** For $B \in \mathbb{IIR}^{n\times n}$ is $(|B|)_{ij} := \max\{\, |b_{ij}| \,\big|\, b_{ij} \in B_{ij} \,\}$.

**Proof.** $X^{k+1} \subseteq \mathrm{int}(Y^k)$ implies $\rho(|C_\varepsilon|) < 1$ as has been shown in [27]. Defining $E := \pm\delta$ we have

$$X^{k+1} \to X_\varepsilon$$

implying

$$Z_\varepsilon + E + C_\varepsilon \cdot X_\varepsilon \text{ and } Z_\varepsilon + C_\varepsilon \cdot X_\varepsilon + \rho \cdot E \subseteq \mathrm{int}(X_\varepsilon)$$

for every $0 \le \rho < 1$. We have $q(X^{k+1}, X_\varepsilon) \to 0$ and therefore

$$q(Z_\varepsilon + C_\varepsilon \cdot X^k,\ Z + C_\varepsilon \cdot X_\varepsilon) = q(C_\varepsilon \cdot X^k,\ C_\varepsilon \cdot X_\varepsilon) \le |C_\varepsilon| \cdot q(X^k, X_\varepsilon).$$

Hence there exist a $k \in N$ with $q(Z_\varepsilon + C_\varepsilon \cdot X^k,\ Z + C_\varepsilon \cdot X_\varepsilon) < \delta/2$ and $q(X^k, X_\varepsilon) < \delta/2$ implying $Z_\varepsilon + C_\varepsilon \cdot X^k \subseteq \mathrm{int}(X^k)$. ■

Theorem 4.2.1 also holds in the complex case where the absolute value for a complex interval matrix $B \in \mathbb{IIC}^n$ is defined by $|B| = |\mathrm{Re}(B)| + |\mathrm{im}(B)|$ (see [3]).

Setting $\widetilde{x} := A^{-1}b$, $R := A^{-1}$ and $X := \pm(1, \ldots, 1)^T$ satisfies the assumptions of theorem 4.2 yielding the exact sensitivity of $\widehat{x}$ w.r.t. perturbations in $A, b$ weighted by $A^*, B^*$ to be

$$\mathrm{Sens}(\widehat{x}, A, b, A^*, b^*) = |A^{-1}| \cdot (|b^+| + |A^*| \cdot |\widehat{x}|), \tag{38}$$

a result which can be found in the literature for $A^* = |A|,\ b^* = |b|$ ([4], [23], [29]). For some $\widetilde{x}, R, X$ satisfying (4.3) lower and upper bounds for this value are obtained by

$$\mathrm{Sens}(\widehat{x}, A, b, A^*, b^*) \in |R| \cdot (|b^*| + |A^*| \cdot |\widehat{x}|) \pm \phi \cdot w \tag{39}$$

where

$$\phi := \max_i \left\{ \frac{u_i}{\{|X| - w\}_i} \,\Big|\, u_i + v_i \neq 0 \right\} \text{ and } w := |I - RA| \cdot |X|.$$

In our approach we use a componentwise absolute value combined with weights as a measure for the maximum elongation resp. the sensitivity of a solution. In many practical applications this is what a user is really interested in. The componentwise absolute value avoids equilibration effects due to norm estimates. This effect may be significant when components in the solution and/or in $A, B$ show large differences in size.

Consider the example

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2\varepsilon & 2\varepsilon \\ 1 & 2\varepsilon & -\varepsilon \end{pmatrix}, \quad b = \begin{pmatrix} 3+3\varepsilon \\ 6\varepsilon \\ 2\varepsilon \end{pmatrix} \tag{40}$$

given by Hamming [10] and discussed e.g. in Deif [7]. The sensitivity of the three components of the solution $\widehat{x} = (\varepsilon, 1, 1)^T$ w.r.t. relative changes in all components of $A$ and $b$ (with weights $A^* = |A|, \ b^* = |b|$) compute approximately to

9.6, 4.6 and 6.0.

Using (4.13) the solution is very stable w.r.t. perturbations whereas $\|A\| \cdot \|A^{-1}\| \approx 0.8/\varepsilon$. For other right hand sides the problem is very sensitive such as for $b = (6, 2, 1)^T$ where we have sensitivities of approximately

2, $0.67/\varepsilon$ and $2.67/\varepsilon$.

The condition number, more explicitely the smallest singular value $\sigma_3 = 2\varepsilon$ of $A$ indicates that a singular matrix is near by $A$. Indeed

$$B = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2.4\varepsilon & 1.2\varepsilon \\ 1 & 1.2\varepsilon & 0.6\varepsilon \end{pmatrix} \tag{41}$$

is the nearest singular matrix in the $\|\cdot\|_2$ norm with a distance equal to $\sigma_3 = 2\varepsilon$. However, the *relative* distance from $B$ to $A$ is very large namely

$$\min\{\, \delta \,\big|\, |B - A| \le \delta \cdot |A| \,\} = 1.6. \tag{42}$$

In the $\|\cdot\|_2$-sense we have, roughly spoken, a distance relative to the largest element in absolute value of $A$ whereas (4.17) is the *relative* distance to $A$ taken for every individual component of $A$.

The value of this observation for practical applications depends very much on the application itself. If the data of $A$ are afflicted with an error being *absolutely* not greater than some value $\varepsilon^*$ then $B$ of (4.16) may very well lie in or near the domain of possible data; in case of a given *relative* precision for every component of $A$, $B$ is with 160 % relative distance to far away.

This leads to the questian of the distance of a matrix $A$ to the next singular matrix weighted by some nonnegative matrix $A^*$:

$$\operatorname{SingRad}(A, A^*) := \min\left\{ \varepsilon \,\Big|\, \exists \text{ singular } \widetilde{A} \text{ with } |\widetilde{A} - A| \le \varepsilon \cdot |A^*| \right\}, \tag{43}$$

which has strong connections to the term *strongly regular* for interval matrices (cf. [23]).

**Lemma 4.3.** Let $A \in \mathrm{I\!R}^{n\times n}$ be a nonsingular matrix, $A^* \in \mathrm{I\!R}^{n\times n}$ be nonnegative. Then

$$\{\rho(|A^{-1}| \cdot |A^*|)\}^{-1} \leq \mathrm{SingRad}(A, A^*) \tag{44}$$

and

$$\{\rho(|A^{-1}| \cdot |A|)\}^{-1} \leq \mathrm{SingRad}(A, |A|) \leq 1. \tag{45}$$

For lower or upper diagonal $A$ holds

$$\{\rho(|A^{-1}| \cdot |A|)\}^{-1} = \mathrm{SingRad}(A, |A|) = 1. \tag{46}$$

**Proof.** Let $\widetilde{A} := A + \delta A$ be given with $|\widetilde{A} - A| \leq \varepsilon \cdot |A^*|$ and $\varepsilon < \{\rho(|A^{-1}| \cdot |A^*|)\}^{-1}$. Then

$$\rho(I - A^{-1}\widetilde{A}) = \rho(A^{-1} \cdot \delta A) \leq \rho(|A^{-1}| \cdot |\delta A|) \leq \varepsilon \cdot \rho(|A^{-1}| \cdot |A^*|) < 1$$

implying the nonsingularity of $\widetilde{A}$ and (4.2=). To prove (4.21) we use the fact that with $A$ also $A^{-1}$ is lower resp. upper diagonal implying that $|A^{-1}| \cdot |A|$ has this property with all diagonal elements equal to 1. ■

The quality of the estimations in lemma 4.3 has been tested in several experiments. The true value of the radius of singularity was calculated through approaching it by checking the determinant for all $2^k$ possibilities ($k$ the number of nonzero components in $A^*$). In our experiments $1/\rho(|A^{-1}||A^*|)$ was a reasonable estimation delivering roughly the magnitude of the true value.

In our example (4.15) the matrix

$$\widetilde{A} \text{ with } \widetilde{A}_{ij} = A_{ij} \cdot (1 + \delta B_{ij}), \quad B = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \tag{47}$$

$$\text{and } \delta = \mathrm{SingRad}(A, |A|) \approx 0.37778$$

was found to be the nearest singular matrix to $A$ w.r.t. perturbations relative to $|A|$ itself. Lemma 4.3 gives an estimation of approximately 0.302 for $\delta$.

Theorem 4.2 gave an estimation on $\mathrm{Elon}^{\varepsilon}(\widehat{x}, A, b, A^*, b^*)$ for finite values of $\varepsilon$ based on an inclusion $\widetilde{x} + X$ for $\widehat{x} = A^{-1}b$. This estimation (4.7) can be calculated on digital computers.

Following we will give a theoretical bound for the $\mathrm{Elon}^{\varepsilon}(\widehat{x}, A, b.A^*, b^*)$ being applicable for all $\varepsilon < \rho(|A^{-1}||A^*|)^{-1}$.

**Corollary 4.4.** Let $A \in \mathrm{IR}^{n\times n}$ being nonsingular and $b \in \mathrm{IR}^n$ with $\widehat{x} := A^{-1}b$ be given. For nonnegative $A^* \in \mathrm{IR}^{n\times n}$, $b^* \in \mathrm{IR}^n$, $A^*$ and $b^*$ not both being identically zero and $\varepsilon < \varepsilon^* := \{\rho(|A^{-1}| \cdot |A^*|)\}^{-1}$ let

$$u := |A^{-1}| \cdot (|b^*| + |A^*| \cdot |\widehat{x}|) \text{ and } C := |A^{-1}| \cdot |A^*|.$$

Then

$$\mathrm{Elon}^{\varepsilon}(A^{-1}b, A^*, b^*) \in u \pm \varepsilon \cdot C \cdot (I - \varepsilon C)^{-1} \cdot u \tag{48}$$

**Proof.** By assumption $\varepsilon$ is small enough to make $I - \varepsilon C$ nonsingular. Perron-Frobenius-Theory shows that $\varepsilon C \geq 0$, and $\rho(\varepsilon C) < 1$ implies $(I - \varepsilon C)^{-1} > 0$ (see [31], Theorem 3.8) and therefore

$$Y := (I - \varepsilon C)^{-1} \cdot u > 0. \tag{49}$$

Let $A_\varepsilon := \left\{ \widetilde{A} \,\middle|\, |\widetilde{A} - A| \leq \varepsilon |A^*| \right\}$, $b_\varepsilon := \left\{ \widetilde{b} \,\middle|\, |\widetilde{b} - b| \leq \varepsilon \cdot |b^*| \right\}$ and $A_\kappa := [-\kappa y, +\kappa y]$. Then

$$A^{-1} \cdot (b_\varepsilon - A_\varepsilon \cdot \widehat{x}) = \pm \varepsilon \cdot |A^{-1}| \cdot (|b^*| + |A^*| \cdot |\widehat{x}|) \tag{50}$$

and

$$\begin{aligned}(I - A^{-1} \cdot A_\varepsilon) \cdot Y_\kappa &= \left\{ (I - A^{-1}\widetilde{A}) \cdot y_\kappa \,\middle|\, \widetilde{A} \in A_\varepsilon \, y_\kappa \in Y_\kappa \right\} \\ &= \left\{ -A^{-1} \cdot \delta A \cdot y_\kappa \,\middle|\, |\delta A| \leq \varepsilon \cdot |A^*|, \, y_\kappa \in Y_\kappa \right\} \\ &= \pm \varepsilon \cdot \kappa \cdot |A^{-1}| \cdot |A^*| \cdot y = \pm \varepsilon \kappa C y.\end{aligned} \tag{51}$$

For $\kappa > \varepsilon$ is $\kappa y - \varepsilon \kappa C y - \varepsilon u = \kappa(I - \varepsilon C) y - \varepsilon u = \kappa u - \varepsilon u > 0$. Therefore, defining $\widetilde{x} := \widehat{x}$, $R := A^{-1}$ and $X := \pm \kappa y$ (4.3) is satisfied for every $\kappa > \varepsilon$.

Applying theorem 4.2 gives

$$\mathrm{Elon}^{\varepsilon}(\widehat{x}, A, b, A^*, b^*) \in u \pm \phi_\varepsilon \cdot \varepsilon \cdot v$$

because $w = 0$ where $v = \kappa C \cdot y$ and using $y - \varepsilon C y = u$

$$\phi_\varepsilon = \max_i \left\{ \frac{u_i}{\{\kappa(y - \varepsilon C Y)\}_i} \,\middle|\, u_i + v_i \neq 0 \right\} = 1/\kappa.$$

Taking the limit $\kappa \to \varepsilon^+$ finishes the proof. ■

Using Corollary 4.4 usually gives reasonable bounds for the sensitivity of a linear system the coefficients of which perturbing within wider ranges. Consider for instance (4.15) with

$\varepsilon = 10^7$. Let $A^* := |A|$ and $b^* := |b|$, i.e. we are looking for the set of solutions of $\widetilde{A}x = \widetilde{b}$ for $\widetilde{A}, \widetilde{b}$ within relative distance $\leq \delta$ to $A, b$. As we saw before values of $\delta$ up to 0.302, i.e. ±30.2% change in every component of $A$ and $b$, are suitable. The following table displays $\mathrm{Elon}^\delta(A^{-1}b, |A|, |b|)$ for the three components of the linear system (4.15) computed after (4.23) for different values of $\delta$.

| $\delta[\%]$ | lower bound<br>upper bound | for | $\mathrm{Elon}^\delta(A^{-1}b, \lvert A\rvert, \lvert b\rvert)$ |
|---|---|---|---|
| 0.1 | 9.59<br>9.61 | 4.79<br>4.81 | 5.99<br>6.01 |
| 1 | 9.50<br>9.70 | 4.74<br>4.85 | 5.93<br>6.07 |
| 10 | 8.5<br>10.7 | 4.2<br>5.4 | 5.2<br>6.8 |
| 20 | 7.2<br>12.0 | 3.6<br>6.0 | 4.3<br>7.7 |
| 30 | 5.6<br>13.6 | 2.8<br>6.8 | 3.2<br>8.8 |

**Table 4.1.** Analysis of the linear system (4.15)

Even for a value of 30 % the bounds clearly give the magnitude of the total sensitivity. The same example with $b = (6, 2, 1)^T$ give the following results:

| $\delta[\%]$ | lower bound<br>upper bound | for | $\mathrm{Elon}^\delta(A^{-1}b, \lvert A\rvert, \lvert b\rvert)$ |
|---|---|---|---|
| 1 | 1.98<br>2.02 | 6.58<br>6.75 $\cdot 10^6$ | 2.64<br>2.69 $\cdot 10^7$ |
| 10 | 1.78<br>2.22 | 5.7<br>7.6 $\cdot 10^6$ | 2.41<br>2.92 $\cdot 10^7$ |
| 20 | 1.5<br>2.5 | 4.6<br>8.7 $\cdot 10^6$ | 2.1<br>3.3 $\cdot 10^7$ |
| 30 | 1.1<br>2.9 | 3.2<br>10.2 $\cdot 10^6$ | 1.7<br>3.7 $\cdot 10^7$ |

**Table 4.2.** Analysis of the linear system (4.15) with $b = (6, 2, 1)^T$

The figures show that even for larger perturbartions the magnitude of the sensitivity is reasonably estimated. For practical purposes the knowledge of the magnitude of the sensitivity is usually sufficient.

Theorem 4.2 applies immediately to matrix inversion by treating the linear systems $Ax = I$. The $\varepsilon$-elongation of $A^{-1}$ is weighted by some nonnegative $A^* \in \mathrm{I\!R}^{n\times n}$ defined similarly to definition 4.1 by

$$\mathrm{Elon}^{\varepsilon}_{kl}(A^{-1}, A^*) := \max\left\{ \frac{|\widetilde{A}^{-1}_{kl} - A^{-1}_{kl}|}{\varepsilon} \;\Big|\; |\widetilde{A} - A| \le \varepsilon |A^*| \right\}.$$

The matrix of $\varepsilon$-elongations is denoted by $\mathrm{Elon}^{\varepsilon}(A^{-1}, A^*)$. Then for $b := I$ and $\widetilde{x} := \mathrm{I\!R} \approx A^{-1}$ we obtain the following result.

**Corollary 4.5.** Let $A \in \mathrm{I\!R}^{n\times n}$ and $R \in \mathrm{I\!R}^{n\times n}$, $\emptyset \ne X \in \mathrm{I\!I\!R}^{n\times n}$ with

$$R \cdot (I - A \cdot R) + \{I - R \cdot A\} \cdot X \subseteq \mathrm{int}(X). \tag{52}$$

Then $A$ and $R$ are not singular and $A^{-1} \in R + \mathrm{int}(X)$. For nonnegative $A^* \in \mathrm{I\!R}^{n\times n}$ not being identical zero define

$$u := |R| \cdot |A^*| \cdot |A^{-1}|, \;\; v := |R| \cdot |A^*| \cdot |X| \;\text{ and } w := |I - RA| \cdot |X|.$$

Then both

$$\varepsilon^* := \min_{i,j} \left\{ \frac{(|X| - w)_{ij}}{(u+v)_{ij}} \;\Big|\; u_{ij} + v_{ij} \ne 0 \right\}$$

and

$$\phi_\varepsilon := \max_{i,j} \left\{ \frac{u_{ij}}{(|X| - \varepsilon v - w)_{ij}} \;\Big|\; u_{ij} + v_{ij} \ne 0 \right\}$$

are well-defined and

$$\mathrm{Elon}^{\varepsilon}(A^{-1}, A^*) \in u \pm \phi_\varepsilon \cdot (\varepsilon \cdot v + w). \tag{53}$$

$u$ can be estimated by using $A^{-1} \in R + X$. (4.28) yields rigorous estimations on the $\varepsilon$-elongation of the inverse of A weighted by $A^*$ which can be calculated on digital computers. Similar to corollary 4.4 bounds hardly being exactly computable but of theoretical interest are

$$\mathrm{Elon}^{\varepsilon}(A^{-1}, A^*) \in |A^{-1}| \cdot |A^*| \cdot |A^{-1}| \pm \varepsilon \cdot |A^{-1}||A^*| \cdot (I - \varepsilon|A^{-1}||A^*|)^{-1} \cdot |A^{-1}| \cdot |A^*| \cdot |A^{-1}|$$

which is true for all $\varepsilon < \varepsilon^* := \left\{\rho(|A^{-1}| \cdot |A^*|)\right\}^{-1}$.

# 5 Sensitivity of eigenvectors/eigenvalues and singular values/vectors

Let $A \in \mathrm{I\!R}^{n \times n}$ be a matrix with simple eigenvalue $\widehat{\lambda} \in \mathrm{I\!R}$.We restrict our attention in this chapter to real eigenvalues/eigenvectors. However, all of the following results immediately extend to the complex case.

In the following we consider two formulations of the eigenproblem as a nonlinear system $f^i : \mathrm{I\!R}^{n^2} \times (\mathrm{I\!R}^n \times \mathrm{I\!R}) \to \mathrm{I\!R}^n \times \mathrm{I\!R}$, namely

$$f^1(A, x, \lambda) := \begin{pmatrix} Ax & - & \lambda x \\ x^T x & - & 1 \end{pmatrix} \text{ and } f^2(A, x, \lambda) := \begin{pmatrix} Ax & - & \lambda x \\ e_l^T x & - & 1 \end{pmatrix} \tag{54}$$

for some $1 \leq 1 \leq n$. Both are equivalent for an eigenvector-/eigenvalue pair $(\widehat{x}, \widehat{\lambda})$ provided $\widehat{x}_1 \neq 0$. They reflect a different normalization of $\widehat{x}$ and lead to different sensitivities of the eigenvector $\widehat{x}$.

**Definition 5.1.** Let $A \in \mathrm{I\!R}^{n \times n}$ be a matrix with simple eigenvalue $\widehat{\lambda} \in \mathrm{I\!R}$ and let $A^* \in \mathrm{I\!R}^{n \times n}$ be a nonnegative matrix. Let $\varepsilon > 0$ be small enough such that all eigenvalues $\widehat{\lambda}$ being connected to $\widehat{\lambda}$ within $A_\varepsilon := \left\{ \widetilde{A} \;\middle|\; |\widetilde{A} - A| \leq \varepsilon \cdot |A^*| \right\}$ are simple. Then the sensitivity of $\widehat{\lambda}$ w.r.t. perturbations in $A$ weighted by $A^*$ is defined by

$$\mathrm{Sens}(\widehat{\lambda}, A, A^*) := \lim_{\varepsilon \to 0^+} \max \left\{ \frac{|\widetilde{\lambda} - \widehat{\lambda}|}{\varepsilon} \;\middle|\; \widetilde{\lambda} \text{ connected to } \widehat{\lambda} \text{ within } A_\varepsilon \right\}.$$

The sensitivity of the $k^{\text{th}}$ component of an eigenvector $\widehat{x}$ of $A$ with $A\widetilde{x} = \widehat{\lambda}\widehat{x}$ is defined by

$$\begin{aligned} &\mathrm{Sens}_k(\widehat{x}, \|\cdot\|_2, A, A^*) := \\ &:= \lim_{\varepsilon \to 0^*} \max \left\{ \frac{|\widetilde{x}_k - \widehat{x}_k|}{\varepsilon} \;\middle|\; \widetilde{x} \text{ connected to } \widehat{x} \text{ within } A_\varepsilon \text{ w.r.t.} f^1 \right\} \end{aligned} \tag{55}$$

resp.

$$\begin{aligned} &\mathrm{Sens}_k(\widehat{x}, e_l^T, A, A^*) := \\ &:= \lim_{\varepsilon \to 0^+} \max \left\{ \frac{|x_k - \widehat{x}_k|}{\varepsilon} \;\middle|\; \text{connected to } \widehat{x} \text{ within } A_\varepsilon \text{ w.r.t.} f^2 \right\}. \end{aligned} \tag{56}$$

where $1 \leq l \leq n$ with $e_l^T \widehat{x} \neq 0$.

In the second case of course $\mathrm{Sens}_l(\widehat{x}, e_l^T, A, A^*) = 0$. Applying theorem 2.5 to $f^1$ resp. $f^2$ yields estimations for the sensitivities of the eigenvalue $\widehat{\lambda}$ resp. the eigenvalue $\widehat{x}$ of $A$ for both kinds of normalizations.

**Theorem 5.2.** Let $A \in \mathbb{R}^{n\times n}$ and $\widetilde{x} \in \mathbb{R}^n$, $\widetilde{\lambda} \in \mathbb{R}$, $R \in \mathbb{R}^{(n+1)\times(n+1)}$, $\emptyset \neq X \in \mathbb{IIR}^n$, $\emptyset \neq \Lambda \in \mathbb{IIR}^n$ with $0 \in X$, $0 \in \Lambda$. Define

$$M_1 := \begin{pmatrix} A - (\widetilde{\lambda} + \Lambda)I & -(\widetilde{x} + X) \\ 2\widetilde{x}^T & 0 \end{pmatrix} \text{ and } M_2 := \begin{pmatrix} A - (\widetilde{\lambda} + \Lambda)I & -(\widetilde{x} + X) \\ e_l^T & 0 \end{pmatrix},$$

$M_\alpha \in \mathbb{IIR}^{(n+1)\times(n+1)}$ for $\alpha = 1, 2$ and let

$$-R \cdot \begin{pmatrix} A\widetilde{x} - \widetilde{\lambda}\widetilde{x} \\ \widetilde{x}^T\widetilde{x} - 1 \end{pmatrix} + \{I - R \cdot M_1\} \cdot \begin{pmatrix} X \\ \Lambda \end{pmatrix} \subseteq \operatorname{int} \begin{pmatrix} X \\ \Lambda \end{pmatrix} \tag{57}$$

resp.

$$-R \cdot \begin{pmatrix} A\widetilde{x} - \widetilde{\lambda}\widetilde{x} \\ e_l^T \cdot x - 1 \end{pmatrix} + \{I - R \cdot M_2\} \cdot \begin{pmatrix} X \\ \Lambda \end{pmatrix} \subseteq \operatorname{int} \begin{pmatrix} X \\ \Lambda \end{pmatrix} \tag{58}$$

be satisfied. Then $R$ and every matrix $M \in M_\alpha$, $\alpha = 1$ resp. $\alpha = 2$ are nonsingular and there exist an eigenvector/eigenvector pair $(\widehat{x}, \widehat{\lambda})$ of $A$ with $\widehat{x} \in \widetilde{x} + \operatorname{int}(X)$ and $\widehat{\lambda} \in \widetilde{\lambda} + \operatorname{int}(X)$. It is $\widehat{x}^T\widehat{x} = 1$ resp. $e_l^T\widehat{x} = 1$; $\widehat{\lambda}$ is simple.

For nonnegative $A^* \in \mathbb{R}^{n\times n}$ define

$$u := |S| \cdot |A^*| \cdot |\widehat{x}| \text{ and } w_\alpha := |I - R \cdot M_\alpha| \cdot |(X, \Lambda)^T| \tag{59}$$

where $S \in \mathbb{R}^{(n+1)\times n}$ is the matrix of first $n$ columns of $R$ and $\alpha = 1$ resp. $\alpha = 2$ depending on whether (5.4) or (5.5) holds. Then

$$\phi := \max_{1\le i\le n+1} \left\{ \frac{U - i}{\big((|X|, |\Lambda|)^T - w\big)_i} \right\} \tag{60}$$

is well-defined and for the sensitivity of the eigenvector $\widehat{x}$ w.r.t. perturbations in $A$ weighted by $A^*$ holds for $1 \le k \le n$

$$\mathrm{Sens}_k(\widehat{x}, \|\cdot\|_2, A, A^*) \in \{u \pm \phi \cdot w_1\}_k \text{ resp.}$$

$$\mathrm{Sens}_k(\widehat{x}, e_l^T, A, A^*) \in \{u \pm \phi \cdot w_2\}_k.$$

The sensitivity of the eigenvalue $\widehat{\lambda}$ satisfies

$$\mathrm{Sens}(\widehat{\lambda}, A, A^*) \in |r| \cdot |A^*| \cdot |\widehat{x} \pm \phi \cdot w_\alpha \tag{61}$$

where $r$ is the row vector of the first $n$ components of the last row of $R$, $\alpha = 1$ resp. $\alpha = 2$.

**Proof.** Follows immediately by applying theorem 2.5 to $f^1$ resp. $f^2$ defined by (5.1) and regarding (in the notation of the proof of theorem 2.5

$$Z_\varepsilon = \Diamond - R \cdot \begin{pmatrix} A_\varepsilon \widehat{x} - \widehat{\lambda}\widehat{x} \\ N \end{pmatrix} = \pm |R| \cdot \begin{pmatrix} |A^*||\widehat{x}| \\ 0 \end{pmatrix} = \pm |S| \cdot |A^*| \cdot |\widehat{x}|$$

where $N = \widehat{x}^T\widehat{x} - 1 = 0$ resp. $N = e_l^T\widehat{x} - 1 = 0$ for $\alpha = 1$ resp. $\alpha = 2$ and $A_\varepsilon := \{\, \widetilde{A} \,\big|\, |\widetilde{A} - A| \le \varepsilon \cdot |A^*| \,\}$. ■

Estimation (5.8) can be calculated on digital computers. For the computation of $u$ from (5.6) note that $\widehat{x} \in \widetilde{x} + X$. By examination of the inverse of the Jacobian of $f^1$ for $f^1(x, \lambda) = 0$ we get

$$\begin{aligned} \left\{ \frac{\partial f^1}{\partial(x,\lambda)}(x,\lambda) \right\}^{-1} &= \begin{pmatrix} A - \lambda I & -x \\ 2x^T & 0 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} B & h \\ g^T & \zeta \end{pmatrix}, B \in \mathbb{R}^{n \times n},\ g,\ h \in \mathbb{R}^n \text{ and } \zeta \in \mathbb{R}. \end{aligned} \tag{62}$$

Then $\zeta = 0$ because $\det(A - \lambda I) = 0$, $(A - \lambda I)h = 0$ and $2x^T h = 1$ implying $h = 0.5x^T$. By $g^T(A - \lambda I) = 0$ together with $-g^T \cdot x = 1$ it follows that $g$ is the left eigenvector $y$ to $\lambda$ subject to normalization $y^T \cdot x = -1$. Hence by applying theorem 5.2 together with (2.25) we get

$$\mathrm{Sens}(\lambda, A, A^*) = |y^T| \cdot |A| \cdot |x| / |y^t x| \tag{63}$$

which can also be obtained by classical perturbation theory [32]. For a symmetric matrix $A$ this means

$$\mathrm{Sens}(\lambda, A, A^*) = |x^T| \cdot |A| \cdot |x| / |x^T x| \tag{64}$$

for *unsymmetric* perturbations of $A$ weighted by $A^*$. Allowing only for symmetric perturbations weighted by $A^*$ we only have to examine the last component of $Z_\varepsilon$ again. Let $A = A^T$ and

$$A_\varepsilon^s := \left\{ \widetilde{A} \text{ symmetric} \,\middle|\, |\widetilde{A} - A| \le \varepsilon \cdot |A^*| \right\}. \tag{65}$$

Then using $f^1$ from (5.1)

$$Z_\varepsilon = \Diamond - R \cdot \begin{pmatrix} A_\varepsilon^s \widehat{x} & - & \widehat{\lambda}\widehat{x} \\ \widehat{x}^T\widehat{x} & - & 1 \end{pmatrix} = \Diamond\ R \cdot \begin{pmatrix} \Delta A \cdot \widehat{x} \\ 0 \end{pmatrix} \tag{66}$$

where $\Delta A \in \mathrm{IR}^{n\times n}$, $|\Delta A| \le \varepsilon \cdot |A^*|$ and $\Delta A$ symmetric. If $R$ is the exact inverse of the Jacobian of $f^1$ at $(\widehat{x}, \widehat{\lambda})$ then $g^T$ in (5.9) equals $y^T$ which is $-\widehat{x}^T/\widehat{x}^T\widehat{x}$ because of the symmetry of $A$ and the normalization $g^T \cdot \widehat{x} = 1$. The sensitivity w.r.t. symmetric perturbations is $\max\limits_{Z\in Z} |Z_\varepsilon|/\varepsilon$ in the limit $\varepsilon \to 0$. We have

$$\widehat{x}^T \cdot \Delta A \cdot \widehat{x} = \sum_{i,j} \widehat{x}_i \cdot \Delta A_{ij} \cdot \widehat{x}_j = \mathrm{diag}(\Delta A) \cdot \widehat{x} + 2 \cdot \sum_{i>j} \widehat{x}_i \cdot \Delta A_{ij} \cdot \widehat{x}_j. \tag{67}$$

In the sum (5.14) all dependencies are eliminated yielding $\max |\widehat{x}^T \cdot \Delta A \cdot \widehat{x}| = \varepsilon |\widehat{x}^T| \cdot |A^*| \cdot |\widehat{x}|$.

**Corollary 5.3.** The sensitivity of a simple eigenvalue $\lambda$ with corresponding eigenvector $x$ of a symmetric matrix w.r.t. unsymmetric perturbations weighted by $A^* \in \mathrm{IR}^{n\times n}$, $A^* \ge 0$ is the same as the sensitivity w.r.t. symmetric perturbations weighted by $A^*$, namely

$$|x^T| \cdot |A^*| \cdot |x|/x^T x.$$

For a simple singular value $\sigma$ of some matrix $A\sigma^2$ is an eigenvalue of $A^T A$. Let $u$ and $v$ be the left and right singular vector of $A$ with $\|u\|_2 = \|v\|_2 = 1$. Then $A^T A v = \sigma A^T u = \sigma^2 v$ and the sensitivity of the eigenvalue $\sigma^2$ of $A^T A$ is, similar to (5.14), is equal to

$$\lim_{\varepsilon\to 0} \max v^T \cdot \Delta(A^T A) \cdot v/\varepsilon$$

where $\Delta(A^T A) := \{ \widetilde{A}^T \widetilde{A} - A^T A \mid |\widetilde{A} - A| \le \varepsilon \cdot |A^*| \}$. Neglecting $0(\varepsilon^3)$ terms we have to examine $v^T(A^T \cdot \Delta A)^T A)v$ for $|\Delta A| \le \varepsilon \cdot |A^*|$. By using $Av = \sigma u$ and $A^T u = \sigma v$

$$v^T \cdot (A^T \cdot \Delta A + (\Delta A)^T A)v = \sigma u^T \Delta A v + \sigma v^T (\Delta A)^T u = 2\sigma \cdot u^T \cdot \Delta A \cdot v.$$

Here all dependencies are eliminated and the sensitivity of the eigenvalue $\sigma^2$ of $A^T A$ turns out to be $2\sigma |u^T| \cdot |A^*| \cdot |v|$ yielding the sensitivity of the singular value $\sigma$ of $A$ to be $|u^T| \cdot |A^*| \cdot |v|$.

**Corollary 5.4.** The sensitivity of a simple singular value $\sigma$ with left and right singular vector $u$ and $v$ w.r.t. perturbations in $A$ weighted by some nonnegative $A^* \in \mathrm{IR}^{n\times n}$ is

$$|u^T| \cdot |A^*| \cdot |v|.$$

# 6 Numerical example

In this section we will give some numerical examples for systems of linear and nonlinear equations. Throughout this section we will use a short notation for intervals by giving coinciding digits of the left and right bounds only once. E.g. an interval [2.718281, 2.718282] will be noted by

$$2.71828_1^2.$$

The notation bears the advantage that the sharpness of an interval is realized immediately. In many examples we display only 3 figures of the result. In this case for the example above

$$2.718$$

would be displayed. The notation indicates that *all displayed figures are correct.* Using the interpretation that an interval is an inclusion for a correct result we can define a *relative error* $\delta([A])$ of an interval $[A] = [\underline{a}, \overline{a}]$ by

$$\delta([A]) = \frac{|\overline{a} - \underline{a}|}{\min(|\underline{a}|, |\overline{a}|)} = \frac{\operatorname{diam}([A])}{q(0, [A])}$$

provided that $0 \notin [A]$. Indead the relative error for any $x \in [A]$ w.r.t. any $y \in [A]$ is bounded by $\delta([A])$. To be perfectly clear it should be stated that all given results are correct in the sense that the true result is between the diplayed left and right bounds.

In order to implement the estimations derived in the previous chapters on digital computers an appropriate floating-point arithmetic is necessary. That means an arithmetic with directed roundings to preserve the central property of interval arithmetic, the isotonicity:

$$a, b \in \mathbb{F} : a * b \in [a, a] * [b, b] \quad \text{and}$$

$$[A], [B] \in \mathbb{IF} : \forall\, a, b \in \mathbb{R} : a \in [A],\; b \in [B] \Rightarrow a * b \in [A] * [B].$$

Here $\mathbb{F} \subseteq \mathbb{R}$ denotes some set of floating-point numbers,

$$[A] = [\underline{a}, \overline{a}] := \{\, x \in \mathbb{R} \mid \underline{a} \leq x \leq \overline{a} \,\} \quad \text{for } \underline{a}, \overline{a} \in \mathbb{F}$$

are floating-point intervals ($[A] \in \mathbb{IF}$) and $* \in \{+, -, \cdot, /\}$. Operations over floating-point intervals as well as for floating-point interval vectors and matrices are well-defined and fastly executable on digital computers (cf. [3], [5], [20], [23]).

In the following we use an implementation on a Personal Computer with Coprocessor taking advantage of the IEEE 754 arithmetic. All results are produced using double precision, i.e. 54 bit in the mantissa equivalent to approximately 18 decimal figures. The programming language in use is TPX (cf. [11]), an extension of TURBO-PASCAL developed at the technical university of Hamburg allowing a general operator-concept, function- and procedure-name overloading, general result types for functions as well as dynamic array handling. TPX is transpiled into TURBO-PASCAL by means of a precompiler. The code for the examples presented below as well as the precompiler itself are freely available.

We start with an application of theorem 2.5 using an example given by Abbot and Brent [1], a discretization of

$$3y''y + y'^2 = 0, y(0) = 0, y(1) = 20. \tag{68}$$

The exact solution is $y(t) = 20 \cdot t^{3/4}$, our initial approximation is very poor namely $\widetilde{x}_i \equiv 10.0$, $1 \leq i \leq n$. We solve the discretized system $f : \mathrm{IR}^n \times \mathrm{IR}^p$ with

$$\begin{aligned} f_1 &= p_1 \cdot x_1(x_2 - 2x_1) + x_2^2/4 \\ f_i &= p_1 \cdot x_i(x_{i+1} - 2x_i + x_{i-1}) + p_2(x_{i+1} - x_{i-1})^2/4 \;\; 2 \leq i \leq n-1 \\ f_n &= p_1 \cdot x_n(20 - 2x_n + x_{n-1}) + p_2(20 - x_{n-1})^2/4 \end{aligned} \tag{69}$$

with parameters $p_1$ and $p_2$. The necessary derivatives are calculated using automatic differentiation (cf. [24]). For different values of $n$ the following table displays the computed inclusion $[x_n]$ for $x_n$ and its relative error $\delta[x_n]$. We are interested in the sensitivity of the solution w.r.t. perturbations in the coefficients 3.0 and 1.0 of the first and second term in (6.1), that means in the notation (2.15) $c^* = (3.0,\ 1.0)^T$. We display the sensitivities $\sigma_1$ and $\sigma_n$ of the first and $n^{\text{th}}$ component of the solution $\widehat{x}$ of the discretized system (6.2) and the maximum sensitivity $\sigma_{\max}$ of all components $x_i, 1 \leq i \leq n$.

Finally the maximum relative error $[\sigma]_{\max}$ of the inclusions of the sensitivities of the components $\widehat{x}_i, 1 \leq i \leq n$ is displayed.

| | $n = 20$ | $n = 50$ | $n = 100$ |
|---|---|---|---|
| $[x_n]$ | $19.27738548068119^7_3$ | $19.70448048678683^9_5$ | $19.85565911919249^9_6$ |
| $\delta[x_n]$ | $1.2E-16$ | $1.6E-16$ | $1.9E-16$ |
| $\sigma_1$ | $2.035289918869^6_4$ | $1.364512098913^6_3$ | $0.979339959727^8_2$ |
| $\sigma_n$ | $0.349629110027^6_4$ | $0.145809614137^9_7$ | $0.065358146488^4_1$ |
| $\sigma_{\max}$ | 3.57 | 3.63 | 3.79 |
| $\delta[\sigma]_{\max}$ | $3.3E-13$ | $1.8E-12$ | $3.4E-12$ |

**Table 6.1.** Discretization of (6.2)

Obviously the discretized nonlinear system is very stable under perturbations of the coefficients 3.0 and 1.0. To estimate the sensitivity of a problem usually 1 or 2 figures suffice; with our achieved accuracy of 12 figures and more for the sensitivity practical needs are more than satisfied.

We mention that perturbing the boundary conditions yields results of similar quality. The maximum sensitivity for $n = 100$ increases to 19.85.

Next we investigate some ill-conditioned linear systems $Ax = b$. Until noted otherwise we set $A^* := |A|$ and $b^* := |b|$, i.e. regard relative perturbations w.r.t. all components of the linear system. In this approach zero elements stay zero. In order to display the dependency of the sensitivity w.r.t. different right hand sides we use

$$b := A^{-1} \cdot (+1, -1, +1, \ldots)^T \text{ and } b := V_1$$

for a singular value decomposition $A = U\Sigma V^T$. The minimum $\text{Sens}_{\min}$ and maximum $\text{Sens}_{\max}$ sensitivity of the components of the solution for these two right hand sides w.r.t. perturbations in $A$ and $b$ weighted by $A^*$ and $b^*$ are displayed. Finally we display the condition numbers delivered by LINPACK condition estimator.

Our first example are Hilbert matrices. In order to keep the coefficients exactly representable we use the inverse, i.e.

$$A := H^{-1} \text{ with } H_{ij} := (i + j - 1)^{-1}.$$

The following results were obtained using theorem 2.5 for the linear case.

| | $b = A^{-1} \cdot (+1, -1, +1, \ldots)^T$ | | $b = V_1$ | | |
|---|---|---|---|---|---|
| $n$ | $\text{Sens}_{\min}$ | $\text{Sens}_{\max}$ | $\text{Sens}_{\min}$ | $\text{Sens}_{\max}$ | cond(A) |
| 5 | 2.489E5 | 5.402E5 | 2.108E5 | 4.929E7 | 4.7E5 |
| 7 | 1.951E8 | 4.372E8 | 1.797E8 | 9.166E11 | 4.7E8 |
| 9 | 1.705E11 | 3.886E11 | 1.454E11 | 7.492E15 | 4.9E11 |
| 11 | $1.5^{9}_{7}$ E14 | $3.^{70}_{58}$ E14 | $1.^{46}_{30}$ E14 | $1.3^{8}_{3}$ E16 | 5.2E14 |
| 13 | $^{2.0}_{0}$ E17 | $^{1.4}_{0}$ E19 | $^{1.18}_{0.41}$ E17 | $^{5.2}_{0}$ E17 | 2.4E17 |

**Table 6.2.** Sensitivity Hilbert matrices

For $n = 13$ in some cases only upper bounds for the sensitivity are obtained which need not be sharp. However, for a number of components still lower and upper bounds are obtained like $(0.77,\ 2.04) \cdot 10^{17}$ implying a minimum sensitivity of the linear system.

Next we display the results for Pascal matrices defined by

$$P_{ij} := \begin{pmatrix} i + j \\ i \end{pmatrix}$$

| | $b = A^{-1} \cdot (+1, -1, +1, \ldots)^T$ | | $b = V_1$ | | |
|---|---|---|---|---|---|
| $n$ | $\text{Sens}_{\min}$ | $\text{Sens}_{\max}$ | $\text{Sens}_{\min}$ | $\text{Sens}_{\max}$ | cond($A$) |
| 5 | 3.564E3 | 3.637E4 | 3.606E3 | 4.742E5 | 6.3E4 |
| 7 | 1.139E5 | 4.176E6 | 1.125E5 | 1.982E8 | 1.4E7 |
| 9 | 3.625E6 | 4.872E8 | 3.533E6 | 8.708E10 | 3.0E9 |
| 11 | 1.158E8 | 5.804E10 | 1.119E8 | 4.141E13 | 6.7E11 |
| 13 | 3.718E9 | 7.009E12 | 3.572E9 | 3.187E16 | 1.5E14 |

**Table 6.3.** Sensitivity Pascal matrices

The results for Zielke matrices defined by

$$Z_{ij} := \frac{\begin{pmatrix} n-i-1 \\ i-1 \end{pmatrix} \cdot n \cdot \begin{pmatrix} n-1 \\ n-j \end{pmatrix}}{i+j-1}$$

are even better. Supposedly this is due to the fact that the inverse is the same matrix with a chessboard-like sign distribution and is therefore exactly representable.

| | $b = A^{-1} \cdot (+1, -1, +1, \ldots)^T$ | | $b = V_1$ | | |
|---|---|---|---|---|---|
| n | $\text{Sens}_{\text{min}}$ | $\text{Sens}_{\text{max}}$ | $\text{Sens}_{\text{min}}$ | $\text{Sens}_{\text{max}}$ | cond$(A)$ |
| 7 | 1.150E6 | 8.780E8 | 1.097E8 | 2.166E8 | 1.8E9 |
| 9 | 1.360E8 | 1.447E12 | 9.632E10 | 1.889E11 | 4.9E12 |
| 10 | 1.497E9 | 6.414E13 | 2.580E12 | 3.443E12 | 2.7E14 |
| 11 | 1.658E10 | 2.533E15 | 8.966E13 | 1.752E14 | 1.5E16 |

**Table 6.4.** Sensitivity Zielke matrices

The tables verify again the well-known fact that the condition number may over- or underestimate the true sensitivity of the solution of a linear system. All results are of high quality. Moreover, as is well-known, the sensitivity may depend significantly on the right hand side. An advantage of the methods presented is that the sensitivity of each individual components of the solution can be rigorously estimated w.r.t. weighted perturbations in the full set of components or part of them. We will investigate this in a final example.

For a 50 × 50 random matrix with random right hand side (all numbers being uniformly distributed in $[0, 1]$) we obtain for the sensitivity $\text{Sens}_i$ of the $i^{\text{th}}$ component of the solution $\widehat{x} = A^{-1} \cdot b$ for $i = 9$ and $i = 19$ with $A^* := |A|$, $b^* := |b|$

$$\begin{aligned} \text{Sens}_{10} &= 248 \\ \text{Sens}_{9} &= 17900\ . \\ \text{cond}(A) &= 685 \end{aligned}$$

We have the possibility to check the sensitivity w.r.t. individual sets of coefficients in $A$ and $b$, for example w.r.t. the columns $A_i$ of $A$. We obtain for the sensitivity of the $9^{\text{th}}$ component of the solution $\widehat{x}$

$$\begin{array}{llll} \text{Sens}_9 \text{ w.r.t. } A^* = |A_{19}|, & b^* = 0 & 8600 \\ \text{Sens}_9 \text{ w.r.t. } A^* = |A_9|, & b^* = 0 & 6 \\ \text{Sens}_9 \text{ w.r.t. } A^* = 0, & b^* = |b| & 1270. \end{array}$$

For another right hand side, namely $b = U - 50$ with $A = U \Sigma V^T$, things change. We obtain for the minimum and maximum sensitivity of the solution $\widehat{x}$ w.r.t. perturbations weighted by $A^* = |A|$, $b^* = |b|$

| | | |
|---|---|---|
| | minimum | 210 |
| Sensitivity of$\widehat{x}$ | | |
| | maximum | 122851 |

as compared the condition number 685.

# 7 Conclusion

Methods have been described to compute rigorous bounds for the sensitivity of linear or nonlinear systems of equations w.r.t. weighted perturbations in the input data. Together with rigorous estimations on the solution the sensitivity information comes virtually free of cost. The calculated estimations are very sharp.

A criticism of inclusion algorithm for data afflicted with tolerances was that correct bounds for the solution set are computed and all experience showed that those bounds are very sharp, but the degree of sharpness could not be estimated (see [14]). Another criticism was that even a guaranteed and very sharp error bound may mislead a user in case of an extremely sensitive problem. The presented theorems and practical results together with those presented in [28] fill those gaps.

The sensitivity analysis offers the additional advantage that rather than a single number estimating the condition of the problem in use a whole sensitivity vector can be computed estimating variations of individual components of the solution for weighted perturbations in the input data. As is well-known traditional condition numbers do not necessarily reflect the true sensitivity of individual components of a solution.

The methods described can be implemented very effectively on digital computers. No special computer arithmetic is necessary; a state of the art arithmetic e.g. described in the IEEE 754 binary floating-point standard [12] or the arithmetic developed by Kulisch [17] suffices. Especially all kinds of arithmetics representing sets on computers are suitable; in our implementation we used a rectangular real or complex arithmetic.

# References

[1] Abbot, J.P.; Brent, R.P.: Fast Local Convergence with Single and Multistep Methods for Nonlinear Equations, Austr. Math. Soc. 19 (Series B), 173–199 (1975)

[2] ACRITH, High-Accuracy Arithmetic Subroutine Library, Program Description and User's Guide, IBM Publications, Document Number SC 33-6164-3 (1986)

[3] Alefeld, G.; Herzberger, J.: Introduction to Interval Computations, Academic Press (1983)

[4] Arioli, M.; Demmel, J.W.; Duff, I.S.: Solving Sparse Linear Systems with Backward Error, SIAM J. Matrix Anal. Appl. 10, No. 2, 165 – 190 (1989)

[5] Bauch, H.; Jahn, K.-U.; Oelschlägel, D.; Süsse, H.; Wiebigke, V.: Intervallmathematik, Theorie und Anwendungen; Mathematisch-Naturwissenschaftliche Bibliothek, Band 72, B.G. Teubner Leipzig (1987)

[6] Braune, K.D.: Hochgenaue Standardfunktionen für reelle und komplexe Punkte und Intervalle in beliebigen Gleitpunktrastern, Dissertation, Universität Karlsruhe (1987)

[7] Deif, A.: Sensitivity Analysis in Linear Systems, Springer New York (1986)

[8] Fischer, H.: Automatic Differentiation: Fast Method to Compute the Quadratic Form of Hessian Matrix and Given Vector, FACTA UNIVERSITAS (NIS), Ser. Math. Inform. 3, 51 – 59 (1988)

[9] Griewank, A.: On Automatic Differentiation, to appear

[10] Hamming, R.: Introduction to applied numerical analysis, McGraw Hill New York (1971)

[11] Husung, D.: Precompiler for Scientific Computation (TPX), Informatik III, TU Hamburg-Harburg (1989)

[12] IEEE 754 Standard for Floating-point Arithmetic (1986)

[13] Jansson, C.: A Self-Validating Method for Solving Linear Programming Problems, Computing, Suppl. 6, 33 – 46 (1988)

[14] Kahan, W.; LeBlanc, E.: Anomalies in the IBM ACRITH Package, Proceedings of the 7th Symposium on Computer Arithmetic, edited by Kai Hwang, Urbana, Illinois (1985)

[15] Krämer, W.: Inverse Standardfunktionen für reelle und komplexe Intervallargumente mit a priori Fehlerabschätzungen für beliebige Datenformate, Dissertation, Universität Karlsruhe (1987)

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

[17] Kulisch, U.: Computer Arithmetic in Theory and Practice, Academic Press (1981)

[18] Moore, R.E.: Interval Analysis. Englewood Cliffs: Prentice Hall (1966)

[19] Moore, R.E.: A Test for Existence of Solutions for Non-Linear Systems, SIAM J. Numer. Anal. 4 (1977)

[20] Moore, R.E.: Methods and Applications of Interval Analysis. Philadelphia: SIAM (1979)

[21] Neumaier, A.: Overestimation in Linear Interval Equations, SIAM J. Numer. Anal., Vol. 24, No. 1, 207–214 (1987)

[22] Neumaier, A.: Rigorous Sensitivity Analysis for Parameter-Dependent Systems of Equations (1988)

[23] Neumaier, A.: Interval Methods for Systems of Equations, Cambridge University Press, to be published

[24] Rall, L.B.: Automatic Differentiation: Techniques and Applications, Lecture Notes in Computer Science, No. 120, Springer-Verlag, Berlin-Heidelberg-New York (1981)

[25] Rump, S.M.: Kleine Fehlerschranken bei Matrixproblemen, Dr.-Dissertation, Institut für Angewandte Mathematik, Universität Karlsruhe (1980)

[26] Rump, S.M.: Solving Algebraic Problems with High Accuracy, Habilitationsschrift, in: A New Approach to Scientific Computation, Hrsg. U.W. Kulisch und W.L. Miranker, Academic Press, 51–120 (1983)

[27] Rump, S.M.: New Results on Verified Inclusions, in: Accurate Scientific Computations, eds. W.L. Miranker and R. Toupin, Springer Lecture Notes in Computer Science, 235 (1986)

[28] Rump, S.M.: Rigorous Sensitivity Analysis for Systems of Linear and Nonlinear Equations, to appear in MATH. of COMP. (1990)

[29] Skeel, R.: Iterative Refinement implies Numerical Stability for Gaussian Elimination, MATH. of COMP. 35, Nor. 151, 817 – 832 (1980)

[30] Speelpennig, B.: Compiling Fast Partial Derivatives of Functions given by Algorithms, Ph. D. Thesis, University of Illinois at Urbana-Champaign (1980)

[31] Varga, R.S.: Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs (1962)

[32] Wilkinson, J.H.: The Algebraic Eigenvalue Prroblem, Oxford University Press, Oxford (1969)