4. VL, 27.4.2009, G. Bärwolff

Einführung in die Numerische Mathematik

Theorem 1.17

Für die Berechnung von speziellen induzierten Matrixnormen gilt (A $(m \times n)$-Matrix, reell)

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}| \quad \text{Spaltensummennorm} \qquad (1.25)$$

$$\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}| \quad \text{Zeilensummennorm} \qquad (1.26)$$

$$\|A\|_2 = \sqrt{\lambda_{max}} \quad \text{mit } \lambda_{max} \text{ als größten EW von } A^T A \qquad (1.27)$$

Zum Beweis

(1.25) und (1.26) sollen als Übung nachgewiesen werden.

Für den Nachweis von (1.27) überlegt man, dass $A^T A$ ähnlich einer Diagonalmatrix mit den EW von $A^T A$ als Diagonalelementen ist. Die EW sind nichtnegativ und aus der Def. von $\|A\|_2$ folgt dann schließlich (1.27).

Definition 1.28

Unter dem Absolutbetrag einer Matrix $A \in \mathbb{C}^{m \times n}$ versteht man die Matrix

$$|A| = B \quad \text{mit } b_{ij} = |a_{ij}|,$$

also die Matrix mit den Absolutbeträgen ihrer Elemente. Gilt für $A, B \in \mathbb{R}^{m \times n}$ und $a_{ij} \le b_{ij}$

$A \le B$

Bemerkung 1.19

Für die "Beträge" von Matrizen gelten die Beziehungen

i) $|A+B| \leq |A| + |B|$

ii) $|A \cdot B| \leq |A|\,|B|$

iii) $A \leq B$, $C \geq 0$, $D \geq 0 \longrightarrow CAD \leq CBD$

iv) $\|A\|_p \leq \||A|\|_p$, $p \geq 1$, $p \in \mathbb{N}$

v) $\|A\| = \||A|\|$ für $\|\cdot\|_1, \|\cdot\|_\infty, \|\cdot\|_F$

$|A| \leq |B| \longrightarrow \|A\| \leq \|B\|$ für diese Normen

Nachweis von i) - v) sind großteils trivial

Wir betrachten zunächst zur Fehlerfortpflanzung bei der Lösung linearer Gleichungssysteme. Seien A und $\vec{b}$ durch $\Delta A = \varepsilon F$ bzw. $\Delta b = \varepsilon \vec{f}$ gestört, also statt $A\vec{x} = \vec{b}$ wird

$$(A + \varepsilon F)\vec{x}(\varepsilon) = \vec{b} + \varepsilon \vec{f} \qquad (1.28)$$

betrachtet. $F, \vec{f}, \varepsilon$ sind Störungen und für $\varepsilon = 0$ ergibt sich das eigentliche Problem. A sei regulär. Der Satz über implizite Funktionen liefert dann die Existenz und Differenzierbarkeit der Abb $\varepsilon \longrightarrow \vec{x}(\varepsilon)$ und $\vec{x} = \vec{x}(0)$,

$$\dot{\vec{x}}(0) = A^{-1}(\vec{f} - F\vec{x})$$

für die Taylorreihe folgt

$$\vec{x}(\varepsilon) = \vec{x} + \varepsilon \dot{\vec{x}}(0) + \Theta(\varepsilon^2),$$

also $\vec{x}(\varepsilon) - \vec{x} = \varepsilon \dot{\vec{x}}(0) + \Theta(\varepsilon^2) = \varepsilon A^{-1}(\vec{f} - F\vec{x}) + \Theta(\varepsilon^2)$.

Für den relativen Fehler gilt dann

$$\frac{\|\vec{x}(\varepsilon) - \vec{x}\|}{\|\vec{x}\|} \leq \varepsilon \frac{\|A^{-1}(\vec{f} - F\vec{x})\|}{\|\vec{x}\|} + \Theta(\varepsilon^2),$$

und im Fall einer Submultiplikativen (konsistenten) Matrixnorm von A

$$\frac{\|\vec{x}(\varepsilon) - \vec{x}\|}{\|\vec{x}\|} \leq \varepsilon \|A^{-1}\| \left\{ \frac{\|\vec{f}\|}{\|\vec{x}\|} + \|F\| \right\} + \Theta(\varepsilon^2) \qquad (1.29)$$

(1.29) sagt noch nichts über die Auswirkungen der relativen Fehler

$$\frac{\|\Delta A\|}{\|A\|} \quad \text{bzw.} \quad \frac{\|\Delta \vec{b}\|}{\|\vec{b}\|}$$

aus, was aber von Interesse ist.

Mit $\vec{b} = A\vec{x}$ folgt $\|\vec{b}\| \leq \|A\| \|\vec{x}\|$, also

$$\frac{1}{\|\vec{x}\|} \leq \frac{\|A\|}{\|\vec{b}\|}$$

und mit $\|F\| = \frac{\|F\|}{\|A\|} \|A\|$ folgt aus (1.29)

$$\frac{\|\vec{x}(\varepsilon) - \vec{x}\|}{\|\vec{x}\|} \leq \|A\| \|A^{-1}\| \left\{ \varepsilon \frac{\|F\|}{\|A\|} + \varepsilon \frac{\|\vec{f}\|}{\|\vec{b}\|} \right\} + \Theta(\varepsilon^2)$$

$$= \kappa(A) \left\{ \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \vec{b}\|}{\|\vec{b}\|} \right\} + \Theta(\varepsilon^2)$$

$$\lesssim \kappa(A) \left\{ \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \vec{b}\|}{\|\vec{b}\|} \right\} \qquad (1.30)$$

(2.30) gilt für kleine ε. Im Folgenden soll eine genauere Analyse vorgenommen werden (4)

Hilfssatz 2.20

$F \in \mathbb{R}^{n\times n}$ und $\|\cdot\|$ submultiplikativ (konsistent)
Für $\|F\| < 1$ ist $E - F$ nichtsingulär,
und es gilt

$$(E-F)^{-1} = \sum_{k=0}^{\infty} F^k$$

sowie

$$\|(E-F)^{-1}\| \leq \frac{1}{1-\|F\|}$$

Beweis in Analogie zur geom. Reihe,
bzw. Verallgemeinerung derselben ⟶

Hilfssatz 2.21

Es gelte $A\vec{x} = \vec{b}$ mit regulärem $A \in \mathbb{R}^{n\times n}$
und $\vec{b} \in \mathbb{R}^n$, $\vec{b} \neq 0$. $F = \Delta A$ und $\vec{f} = \Delta\vec{b}$ seien
Störungen mit relativem Fehler $\leq \delta$, d.h.

$$\frac{\|\Delta A\|}{\|A\|} \leq \delta \quad , \quad \frac{\|\Delta\vec{b}\|}{\|\vec{b}\|} \leq \delta$$

Falls $\delta\,\kappa(A) = r < 1$ gilt,
ist $A + \Delta A$ regulär und für die Lösung $\tilde{\vec{x}}$ von
$(A+\Delta A)\tilde{\vec{x}} = \vec{b} + \Delta\vec{b}$ gilt

$$\frac{\|\tilde{\vec{x}}\|}{\|\vec{x}\|} \leq \frac{1+r}{1-r}$$

Beweis Ergänzung

$$\|A^{-1}F\| \le \|A^{-1}\|\,\|F\| \le \delta \underbrace{\|A^{-1}\|\,\|A\|}_{\kappa(A)} = r < 1$$

$\rightsquigarrow$ $A+F = A(E+A^{-1}F) =: A(E-\tilde{F})$ mit $\|\tilde{F}\| < 1$

$(E-\tilde{F})^{-1}$ ex. nach H.S. 1.20 $\rightsquigarrow$

$(A+F)^{-1} = (E-\tilde{F})^{-1}A^{-1}$ ex.

Nach Def. von $\tilde{\vec{x}}$ folgt

$$(A+F)\tilde{\vec{x}} = \vec{b} + \vec{f} = A\vec{x} + \vec{f} \quad | \cdot A^{-1}$$

$$\rightsquigarrow (E + A^{-1}F)\tilde{\vec{x}} = \vec{x} + A^{-1}\vec{f}$$

$$\rightsquigarrow \|\tilde{\vec{x}}\| \le \|\underbrace{(E+A^{-1}F)^{-1}}_{(E-\tilde{F})^{-1}}\|\,\|\vec{x} + A^{-1}\vec{f}\|$$

$$\|(E-\tilde{F})^{-1}\| \le \frac{1}{1-\|\tilde{F}\|} \le \frac{1}{1-r}, \quad \|\tilde{F}\| \le r$$

$$\rightsquigarrow \|\tilde{\vec{x}}\| \le \frac{1}{1-r}\left(\|x\| + \|A^{-1}\|\,\underbrace{\|\vec{f}\|}_{\le \delta\|\vec{b}\| = \delta\|A\vec{x}\| \le \delta\|A\|\,\|x\|}\right)$$

$$\le \frac{1}{1-r}\left(\|\vec{x}\| + \underbrace{\delta\|A\|\,\|A^{-1}\|}_{r}\,\|\vec{x}\|\right) = \frac{1+r}{1-r}\|\vec{x}\|$$

Zur Abschätzung des rel. Fehlers betrachten wir

$$\left.\begin{aligned} A\tilde{\vec{x}} + F\tilde{\vec{x}} &= \vec{b} + f \\ A\vec{x} &= \vec{b} \end{aligned}\right\} \Rightarrow A(\tilde{\vec{x}} - \vec{x}) = \vec{f} - F\tilde{\vec{x}}$$

$$\Rightarrow \tilde{\vec{x}} - \vec{x} = A^{-1}\vec{f} - A^{-1}F\tilde{\vec{x}} \qquad (1.31)$$

Theorem 2.22

Unter den Voraussetzungen des H.S. 2.21 gilt

$$\frac{\|\vec{x}-\tilde{\vec{x}}\|}{\|\vec{x}\|} \leq \frac{2r}{1-r}, \quad r = \delta\kappa(A) \qquad (1.32)$$

Beweis

Es ergibt sich aus (1.31)

$$\|\tilde{\vec{x}}-\vec{x}\| \leq \|A^{-1}\vec{f}\| + \|A^{-1}F\tilde{\vec{x}}\|$$

$$\leq \|A^{-1}\| \underbrace{\|\vec{f}\|}_{\leq \delta\|\vec{b}\| \leq \delta\|A\|\|\vec{x}\|} + \underbrace{\|A^{-1}F\|}_{\leq r} \underbrace{\|\tilde{\vec{x}}\|}_{\leq \frac{1+r}{1-r}\|\vec{x}\|}$$

$$\leq \underbrace{\delta\kappa(A)}_{r}\|\vec{x}\| + r\frac{1+r}{1-r}\|\vec{x}\|$$

$$= r\left(1+\frac{1+r}{1-r}\right)\|\vec{x}\| = \frac{r\cdot 2}{1-r}\|\vec{x}\| \quad \checkmark$$

In den bisherigen Abschätzungen wurde der Fehler F der Matrix A in der Norm $\|F\|$ betrachtet und der Fehler von $\vec{b}$ ebenso in der Vektornorm $\|\vec{f}\|$, d.h.

$$\|F\| \leq \delta\|A\| \text{ und } \|\vec{f}\| \leq \delta\|\vec{b}\|$$

wurden vorausgesetzt. Betrachtet man nur die Komponenten, d.h. fordert man

$$|f_{ij}| \leq \delta|a_{ij}| \quad \text{für } i,j = 1,2,\ldots,n$$

dann lassen sich die Abschätzungen des H.S. 1.21 und des Theorems 1.22 verbessern.

Es gelten

Hilfssatz 1.23

Sei $A \in \mathbb{R}^{n\times n}$ regulär und $\vec{b} \in \mathbb{R}^n$, $\vec{b} \neq 0$. Es gelte $A\vec{x} = \vec{b}$ und $(A+\Delta A)\tilde{\vec{x}} = \vec{b} + \Delta\vec{b}$ und die Abschätzungen

$$|\Delta A| \leq \delta |A| \quad , \quad |\Delta \vec{b}| \leq \delta |\vec{b}| .$$

Gilt die Bedingung

$$\delta \, \| \, |A^{-1}| \, |A| \, \|_M = r < 1$$

erfüllt, so ist $A+\Delta A$ regulär und es gilt

$$\frac{\|\tilde{\vec{x}}\|}{\|x\|} \leq \frac{1+r}{1-r}$$

($\| \ \|_M$ ist eine der Matrixnormen $\| \ \|_\infty$, $\| \ \|_1$, $\| \ \|_F$ und $\| \ \|$ die damit verträgliche Vektornorm).

Theorem 1.24

Unter den obigen Voraussetzungen gilt

$$\frac{\|\tilde{\vec{x}} - \vec{x}\|}{\|\vec{x}\|} \leq \frac{2r}{1-r} \qquad (1.33)$$

(für die obigen Normen)

Bemerkung 1.25

Die Aussage der Theoreme 1.22 und 1.24 findet man auch oft in der Form

$$\frac{\|\tilde{\vec{x}} - \vec{x}\|}{\|\vec{x}\|} \leq \frac{K(A)}{1 - \|A^{-1}\| \, \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta \vec{b}\|}{\|\vec{b}\|} \right)$$

Beispiel

(i) $A\vec{x} = \vec{b} \iff \begin{bmatrix} 1 & 0 \\ 0 & 10^{-6} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 10^{-6} \end{bmatrix}$, $A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{6} \end{bmatrix}$

$\|A\|_\infty = 1$, $\|A^{-1}\|_\infty = 10^6$, $\kappa_\infty(A) = 10^6$

$\Delta A = 0$, $\Delta\vec{b} = \begin{bmatrix} 10^{-7} \\ 0 \end{bmatrix} \Rightarrow \|\Delta\vec{b}\|_\infty = 10^{-7}$, $\|\vec{b}\|_\infty = 1$

$\rightsquigarrow$ $\dfrac{\|\Delta\vec{b}\|_\infty}{\|\vec{b}\|_\infty} \leq 10^{-7} = \delta$

$r = \delta\, \kappa_\infty(A) = 10^{-7} \cdot 10^6 = 0{,}1$

Abschätzung (1.32) ergibt

$$\frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} \leq \frac{2 \cdot 0{,}1}{0{,}9} = \frac{2}{9}$$

aber $\|\Delta\vec{x}\|_\infty = |\Delta x_1| = 10^{-7}$

und $\dfrac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = 10^{-7}$

$\rightsquigarrow$ Abschätzung viel zu pessimistisch

(ii) Anwendung des Theorems 1.24

$|A| = A$, $|A^{-1}| = A^{-1}$, $|A|\,|A^{-1}| = E$

$\rightsquigarrow$ $\| |A| \cdot |A^{-1}| \|_\infty = 1 = \kappa_\infty(A)$

$\rightsquigarrow$ $\delta \cdot 1 = r < 1$

$\Delta A = 0$, $\Delta\vec{b} = \begin{bmatrix} 10^{-7} \\ 0 \end{bmatrix}$

Abschätzung (1.32) bzw. (1.33) ergibt

$$\frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} \leq \frac{2 \cdot 10^{-7}}{1 - 10^{-7}} \approx 2 \cdot 10^{-7},$$

also wird der tatsächliche rel. Fehler $\dfrac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = 10^{-7}$ wesentlich realistischer abgeschätzt.

Schlussfolgerungen für die Rechnung mit Gleitpunktzahlensystemen, wobei exakte Rechnung vorausgesetzt wird

$$A \longrightarrow rd(A) = A + \Delta$$
$$\vec{b} \longrightarrow rd(\vec{b}) = \vec{b} + \vec{\delta}$$

Voraussetzung $\|\Delta\|_\infty \leq eps \|A\|_\infty$, $\|\vec{\delta}\|_\infty \leq eps \|\vec{b}\|_\infty$

$$\rightsquigarrow \quad (A + \Delta)\tilde{\vec{x}} = \vec{b} + \vec{\delta},$$

nach Theorem 1.22 folgt mit $\delta = eps$

$$\frac{\|\tilde{\vec{x}} - \vec{x}\|_\infty}{\|\vec{x}\|_\infty} \leq \frac{2\, eps\, \kappa_\infty(A)}{1 - eps\, \kappa_\infty(A)}$$

Angenommen $eps\, \kappa_\infty(A) \leq 1/2$, dann gilt bei exakter Rechnung

$$\frac{\|\tilde{\vec{x}} - \vec{x}\|_\infty}{\|\vec{x}\|_\infty} \leq \frac{2 \cdot eps\, \kappa_\infty(A)}{1/2} = 4\, eps\, \kappa_\infty(A)$$

2. Lösung linearer Gleichungssysteme (durch sukzessive Lösung von Dreiecksystemen)

LR-Zerlegung

- Zu lösen ist $A\vec{x} = \vec{b}$

Dazu soll A als Produkt einer unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R geschrieben werden, d.h. man hat

$$A\vec{x} = \vec{b} \iff L\underbrace{R\vec{x}}_{\vec{y}} = \vec{b}$$

und löse zuerst

$$L\vec{y} = \vec{b}$$

und danach

$$R\vec{x} = \vec{y}$$

Realisierung mit dem Gaußschen Eliminationsverf.

Grundprinzip:

Rangerhaltende Manipulationen der Matrix $[A|\vec{b}]$ durch linearkombinationen von Zeilen

$$L_{ij}(\lambda) = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & \lambda & \ddots & \\ & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i \\ \\ \end{matrix}$$

$$\underbrace{\qquad\qquad}_{j}$$

L_{ij} entsteht aus der Einheitsmatrix vom Typ $n \times n$, wobei an der Position $(i;j)$ die Zahl λ statt einer 0 eingefügt wurde

Multiplikation $L_{ij}(\lambda)\,A$ bewirkt die Addition des λ-fachen der j-ten Zeile von A zur i-ten Zeile von A

d.h. durch geeignete Wahl von λ erzeugt man in

$$\tilde{A} = L_{ij}(\lambda) A$$

an der Position (ij) z.B. auch eine Null (A vom Typ $n \times n$).

$L_{ij}(\lambda)$ hat den Rang n und die Determinante 1

$\rightarrow rg(\tilde{A}) = rg(A)$

Durch mehrfache Multiplikation mit

$$L_{jk}, \quad j = k+1, \dots, n$$

erhält man bei geeigneter Wahl der λ unterhalb von $\tilde{a}_{kk}$ Null-Einträge

__Nun etwas präziser__

$$A = \begin{bmatrix} a_{11} & \dots & a_{1k} & \dots & a_{1n} \\ & \ddots & & & \\ & & a_{kk} & \dots & \\ & & \vdots & & \\ & & a_{nk} & \dots & \end{bmatrix} \longrightarrow \text{Ziel} \begin{bmatrix} a_{11} & \dots & & a_{1n} \\ & \tilde{a}_{kk} & \dots & \tilde{a}_{kn} \\ & 0 & \tilde{a}_{k+1,k+1} & \dots \\ & \vdots & & \\ & 0 & & \end{bmatrix}$$

(Spalte k von A: $\vec{a}_k$)

__Vor.__ $a_{kk} \neq 0$

__Setzen__ $\vec{t} = \vec{t}^{(k)} = \vec{t}^{(k)}(\vec{a}_k) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ t_{k+1,k} \\ \vdots \\ t_{nk} \end{bmatrix}$ } k Komponenten

mit

$$t_{ik} = \begin{cases} 0 & i = 1, \dots, k \\ \dfrac{a_{ik}}{a_{kk}} & i = k+1, \dots, n \end{cases}$$

$\vec{e}_k$ sei der k-te Standardeinheitsvektor

Definition 2.1

$$M_k := \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -t_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & -t_{n,k} & & 1 \end{bmatrix}$$

(k-te Spalte)

Frobenius-Matrix
„Gauß-Transform."

Man überlegt sich, dass M_k das Produkt der oben definierten Matrizen $L_{jk}(-t_j)$, $j = k+1, \dots, n$ ist.

Eigenschaften von M_k

$$M_k = E - \vec{t}^{(k)} \vec{e}_k^T ,$$

$$M_k \vec{a}_k = \begin{bmatrix} a_{k1} \\ \vdots \\ a_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix} ,$$

$a_{k1}, \dots, a_{kk}$ bleiben bei der Multiplikation von M_k unverändert

$$\det(M_k) = 1$$

$$M_k^{-1} = E + \vec{t}^{(k)} \vec{e}_k^T , \text{ da}$$

$$M_k^{-1} M_k = E - \vec{t}^{(k)} \underbrace{\vec{e}_k^T \vec{t}^{(k)}}_{=0} \vec{e}_k = E$$

Wenn alles gut geht,
d.h. wenn jeweils $\tilde{a}_{k-1,k-1} \neq 0$ ist,

Dann erhält man nach der Multiplikation von A mit den Frobenius-Matrizen $M_1, \dots, M_{n-1}$, also

$$M_{n-1} M_{n-2} \cdot \dots \cdot M_1 A = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix} := R$$

eine obere Dreiecksmatrix R, außerdem hat die Matrix

$$M_{n-1} M_{n-2} \cdot \dots \cdot M_1$$

die inverse Matrix

$$L = M_1^{-1} M_2^{-1} \cdot \dots \cdot M_{n-1}^{-1} = \begin{bmatrix} 1 & & & 0 \\ t_{21} & 1 & & \\ t_{31} & t_{32} & \ddots & \\ \vdots & \vdots & & \ddots \\ t_{n1} & t_{n2} & & t_{n,n-1} \; 1 \end{bmatrix}$$

so dass schließlich mit

$$A = LR$$

eine LR-Zerlegung vorliegt.