3. VL, 22.4.09, G. Bärwolff
Einführung in die Numerische Mathematik

Nachdem die Fehlerverstärkung bei Grundoperationen im einem Gleitpunkt-Zahlensystem betrachtet wurde, soll nun etwas Allgemeiner das Problem der Fehlerfortpflanzung bei Rechenalgorithmen diskutiert werden.

1.7 Stabilität - Vorwärtsanalyse - Rückwärtsanalyse

Allgemein beschreibt die Stabilität die Robustheit numerischer Verfahren gegenüber Störungen in den Eingabedaten.

Ein gegebenes Problem oder ein Algorithmus soll durch die Funktion

$$f : \mathbb{R}^n \to \mathbb{R}^m \qquad (1.21)$$

beschrieben werden, wobei eine explizite Formel für f vorliegen soll. Im Allg. bedeutet (1.21), dass ausgehend von n Eingangsdaten m Ergebnisse des Problems berechnet werden.
Der besseren Übersichtlichkeit halber betrachten wir skalarwertige Probleme, d.h. $m = 1$.

Definition 1.21
Die absolute normweise Kondition des Problems $x \longmapsto f(x)$ ist die kleinste Zahl $K_{abs} \geq 0$, so dass

$$\| f(\tilde{x}) - f(x) \| \leq K_{abs} \cdot \| \tilde{x} - x \| \quad , \tilde{x} \to x$$

Das Problem heißt schlecht gestellt, falls es keine solche Zahl gibt ($\kappa_{abs} = \infty$)
Analog ist die relative normweise Kondition die kleinste Zahl $\kappa_{rel} \geq 0$ mit

$$\frac{\| f(\tilde{x}) - f(x) \|}{\| f(x) \|} \lesssim \kappa_{rel} \frac{\| \tilde{x} - x \|}{\| x \|}$$

Bemerkung 1.12

$\lesssim$ bedeutet dabei, dass die Relation in erster Näherung gilt

Die verwendeten Normen können z.B. euklidische Normen des $\mathbb{R}^n$ bzw. $\mathbb{R}^m$ sein

κ klein bedeutet grob ein gut konditioniertes
κ groß — " — schlecht — " — Problem

Beispiele

(a) f diff'bar

$$\| f(\tilde{x}) - f(x) \| = \| f'(x)(\tilde{x} - x) + o(\|\tilde{x} - x\|) \|$$

$$\leq \underbrace{\| f'(x) \|}_{\kappa_{abs}} \| \tilde{x} - x \| + o(\|\tilde{x} - x\|)$$

$\| f'(x) \|$ ← Matrix-Norm für die Abl.-Matrix

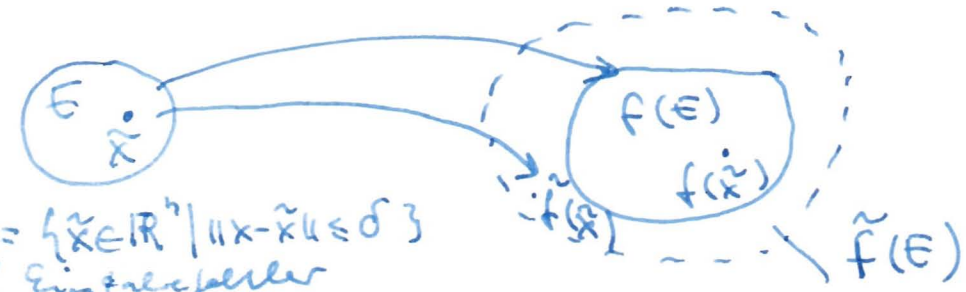
$$\rightarrow \quad \| f(\tilde{x}) - f(x) \| \lesssim \kappa_{abs} \| \tilde{x} - x \|$$

$$\rightarrow \quad \frac{\| f(\tilde{x}) - f(x) \|}{\| f(x) \|} \leq \underbrace{\frac{\| f'(x) \| \| x \|}{\| f(x) \|}}_{\kappa_{rel}} \frac{\| \tilde{x} - x \|}{\| x \|}$$

In Folgenden sollen Fehleranalysen durchgeführt werden. Dabei geht es uns um den Fehler, der durch den Übergang vom ursprünglichen Problem f zu einem numerischen Verfahren $\tilde{f}$ entsteht. Statt $f(x)$ wird eine Näherung $\tilde{f}(x)$ berechnet

1.8 Stabilitätskonzepte

Vorwärtsanalyse:



$E = \{\tilde{x} \in \mathbb{R}^n \mid \|x - \tilde{x}\| \leq \delta\}$
δ Eingabefehler

Analyse der Vergrößerung von $f(E)$ zu $\tilde{f}(E)$ fragt nach der Stabilität im Sinne der Vorwärtsanalyse. Sie beinhaltet den Einfluss des Eingabefehlers ($\tilde{x} \in E$ statt $\{x\}$) und des Verfahrensfehlers ($\tilde{f}$ anstelle von f)

Definition 1.12
Ein Verfahren heißt stabil, wenn es eine Konstante $C \in \mathbb{R}$ gibt, so dass gilt:

$$\|f(\tilde{x}) - \tilde{f}(\tilde{x})\| \leq \|x\| \, C \, eps$$

(eps Maschinengenauigkeit). C quantifiziert die Stabilität im Sinne der Vorwärtsanalyse

Rückwärtsanalyse (James Hardy Wilkinson):
Die Idee besteht darin, ein fehlerbehaftetes Resultat $\tilde{y} = \tilde{f}(\tilde{x})$ durch $\tilde{y} = f(\hat{x}) = f(\tilde{x} + \Delta\tilde{x})$

darzustellen.

Die formale Definition des Rückwärtsfehlers eines Algorithmus $\tilde{f}$ für die fehlerbehafteten (gerundeten) Eingabedaten $\tilde{x}$ mit $\|\tilde{x}\| \neq 0$ lautet:

Definition 1.13

$$\varepsilon_R(\tilde{x}) = \inf \left\{ \frac{\|\Delta\tilde{x}\|}{\|\tilde{x}\|} \;\middle|\; \hat{x} = \tilde{x} + \Delta\tilde{x} \in D_f \wedge f(\tilde{x}+\Delta\tilde{x}) = \tilde{f}(\tilde{x}) \right\}$$

Man nennt einen Algorithmus rückwärtsstabil, wenn der relative Rückwärtsfehler $\varepsilon_R(\tilde{x})$ für alle $\tilde{x} \in D_{\tilde{f}}$ kleiner als der unvermeidbare relative Eingabefehler ist, d.h.

$$\varepsilon_R(\tilde{x}) \leq eps \quad \text{f.a. } \tilde{x} \in D_{\tilde{f}} \tag{1.20}$$

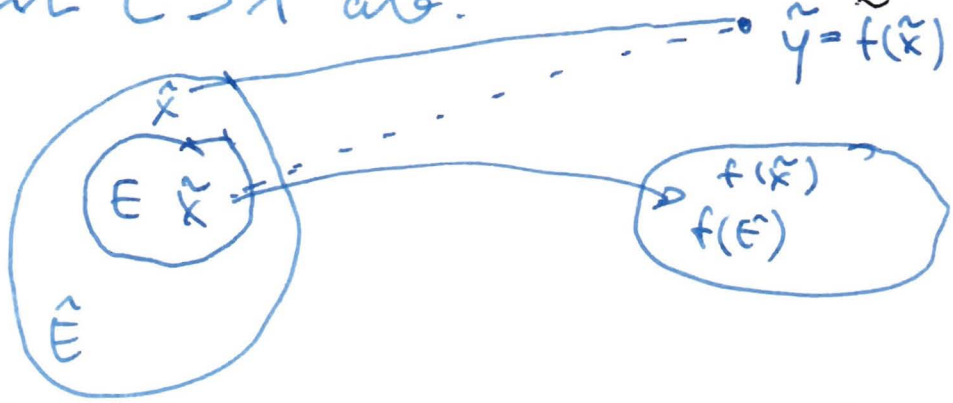
Bemerkung 1.14

Oft interessiert man sich auch nur dafür, ob der relative Rückwärtsfehler überhaupt beschränkt ist.

Außerdem schwächt man für manche Anwendungen (1.20) zu

$$\varepsilon_R(\tilde{x}) \leq C\, eps \tag{1.21}$$

mit $C > 1$ ab.



Rückwärtsanalyse untersucht das Verhältnis von

$$\hat{E} = \bigcup_{\tilde{x} \in E} \{ \hat{x} : f(\hat{x}) = \tilde{f}(\tilde{x}) \text{ und } \|\hat{x} - \tilde{x}\| \text{ minimal} \}$$

Bemerkung 1.15
Man kann zeigen, dass Rückwärtsstabilität Vorwärtsstabilität impliziert

Beispiel Addition zweier Zahlen

$$f(a,b) = a+b$$

Vorwärtsanalyse

$$\tilde{f}(\tilde{a},\tilde{b}) = \tilde{a} \mathbin{\tilde{+}} \tilde{b} = a(1+\varepsilon_1) \mathbin{\tilde{+}} b(1+\varepsilon_2)$$
$$= (a + a\varepsilon_1 + b + b\varepsilon_2)(1+\varepsilon_3)$$
$$\doteq a+b + \underbrace{a\varepsilon_1 + b\varepsilon_2 + (a+b)\varepsilon_3}_{\tau}$$
$$\min|\varepsilon_k| \le \text{eps}$$

$$\rightsquigarrow \tilde{f}(\tilde{a},\tilde{b}) = a+b+\tau = f(a,b) + \tau \qquad (1.22)$$

$$\rightsquigarrow |\tilde{f}(\tilde{a},\tilde{b}) - f(a,b)| \le \tau \le (|a|+|b|+|a+b|)\,\text{eps}$$

Rückwärtsanalyse
Ausgehend von (1.22) machen wir den Ansatz

$$\tilde{f}(\tilde{a},\tilde{b}) = \underline{a+b+\tau = \hat{a}+\hat{b}} = f(\hat{a},\hat{b})$$

Gleichung hat unendlich viele Lösungen, aber nur eine, die $\left\|\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} - \begin{pmatrix}a\\ b\end{pmatrix}\right\|$ minimal macht! (euklidische Norm)

nämlich $\hat{a} = a + \tau/2$, $\hat{b} = b + \tau/2$

$$\rightsquigarrow \varepsilon_R(a,b) = \frac{\left\|\begin{pmatrix}\tau/2\\ \tau/2\end{pmatrix}\right\|}{\left\|\begin{pmatrix}a\\ b\end{pmatrix}\right\|} = \frac{\tau}{\sqrt{2}\sqrt{a^2+b^2}} \le \frac{|a|+|b|+|a+b|}{\sqrt{2}\sqrt{a^2+b^2}}\,\text{eps}$$

wie sieht es mit der Kondition der Addition aus?

$$\frac{f(a+\Delta a, b+\Delta b) - f(a,b)}{f(a,b)} = \frac{\Delta a + \Delta b}{a+b}$$

$$\approx \frac{a}{a+b}\frac{\Delta a}{a} + \frac{b}{a+b}\frac{\Delta b}{b}$$

$$\rightsquigarrow \left|\frac{\Delta a+\Delta b}{a+b}\right| \leq \left|\frac{a}{a+b}\right| \left|\frac{\Delta a}{a}\right| + \left|\frac{b}{a+b}\right|\left|\frac{\Delta b}{b}\right|$$

$$\leq \max\left\{\left|\frac{a}{a+b}\right|, \left|\frac{b}{a+b}\right|\right\} 2\,eps$$

κ_{rel}

$f(x_1, x_2) = x_1 + x_2$

$\frac{\partial f}{\partial x_1} = 1 = \frac{\partial f}{\partial x_2}$

$\rightsquigarrow$ Addition ist für $sign(a) = -sign(b)$ und $|a| \approx |b|$ schlecht konditioniert,

d.h. heißt der Fehler, der unvermeidbar ist, kann im ungünstigsten Fall

$$\max\left\{\left|\frac{a}{a+b}\right|, \left|\frac{b}{a+b}\right|\right\} 2\,eps$$

sein

1.9. Fehlerabschätzungen für lin. Gleichungssysteme

Sei A eine reelle reguläre Matrix vom Typ $(n \times n)$ und $\vec{b}$ ein (Spalten) Vektor aus dem $\mathbb{R}^n$. Zu lösen ist

$$A\vec{x} = \vec{b} \qquad (1.23)$$

Beispiel 1.15

$$A = \begin{bmatrix} 3 & 1{,}001 \\ 6 & 1{,}997 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} 1{,}999 \\ 4{,}003 \end{bmatrix}$$

(1.23) $\rightarrow$

$$3x_1 + 1{,}001 x_2 = 1{,}999$$
$$6x_1 + 1{,}997 x_2 = 4{,}003$$

Lösung ist Schnittpunkt von 2 Geraden, und zwar $(x_1, x_2) = (1, -1)$.

Ändert man $\vec{b}$ geringfügig zu

$$\tilde{\vec{b}} = \vec{b} + \Delta\vec{b} = \begin{bmatrix} 2{,}002 \\ 4{,}000 \end{bmatrix}, \text{ d.h. } \Delta\vec{b} = \begin{bmatrix} 0{,}003 \\ -0{,}003 \end{bmatrix},$$

dann hat $A\tilde{\vec{x}} = \tilde{\vec{b}}$ die Lösung

$$(\tilde{x}_1, \tilde{x}_2) = (0{,}4004, 0{,}8),$$

d.h. eine geringfügige Veränderung der Geraden hat eine recht große Auswirkung auf den Schnittpunkt!

Grund: Beide Geraden sind fast parallel

Werkzeuge zur mathematischen Beschreibung dieses Phänomens:

Vektor- und Matrixnormen ($\vec{x} \in \mathbb{R}^n$, A $(n \times n)$-Matrix)

$$\|\vec{x}\|_1 = \sum_{k=1}^{n} |x_k| \qquad \text{Summennorm}$$

$$\|\vec{x}\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2} \qquad \text{Euklidische Norm}$$

$$\|\vec{x}\|_\infty = \max_{1 \le k \le n} \{|x_k|\} \qquad \text{Max.-Norm}$$

Durch Vektornormen induzierte Matrixnormen

$$\|A\|_G = \max_{\vec{x} \in \mathbb{R}^n, \vec{x} \neq 0} \frac{\|A\vec{x}\|_G}{\|\vec{x}\|_G} = \max_{\vec{y} \in \mathbb{R}^n, \|\vec{y}\|=1} \|A\vec{y}\|_G$$

$G \in \{1, 2, \infty\}$

Es gilt

a) $\|A\vec{x}\|_G \le \|A\|_G \|\vec{x}\|_G$ f.a. $\vec{x} \in \mathbb{R}^n$ — Verträglichkeit

b) $\|AB\|_G \le \|A\|_G \|B\|_G$ — Submultiplikativität

Zurück zum Beispiel 1.15

Es ist $\|\Delta\vec{b}\|_\infty = 0{,}003$, $\|\vec{b}\|_\infty = 4{,}003$

$$\rightsquigarrow \frac{\|\Delta\vec{b}\|_\infty}{\|\vec{b}\|_\infty} \approx 7{,}5 \cdot 10^{-4}$$

$$\|\Delta\vec{x}\|_\infty = \|\vec{x} - \tilde{\vec{x}}\|_\infty = 1{,}8, \quad \|\vec{x}\|_\infty = 1$$

$$\rightsquigarrow \frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = 1{,}8$$

$$\rightsquigarrow \frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} \Big/ \frac{\|\Delta\vec{b}\|_\infty}{\|\vec{b}\|_\infty} \approx 2400$$

Wie hängt die Fehlerverstärkung von A ab?

$$\|\vec{x} - \tilde{\vec{x}}\| = \|A^{-1}\vec{b} - A^{-1}\tilde{\vec{b}}\| = \|A^{-1}(\vec{b} - \tilde{\vec{b}})\|$$

$$\rightsquigarrow \|A\vec{x}\| = \|A^{-1}\Delta\vec{b}\|$$

$$\rightsquigarrow \frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} = \frac{\|A^{-1}\Delta\vec{b}\| \cdot \|\vec{b}\|}{\|\vec{x}\|\,\|\vec{b}\|}$$

$$= \frac{\|A\vec{x}\| \cdot \|A^{-1}\Delta\vec{b}\|}{\|\vec{x}\|\,\|\vec{b}\|} \leq \frac{\|A\|\,\|\vec{x}\| \cdot \|A^{-1}\|\,\|\Delta\vec{b}\|}{\|\vec{x}\| \cdot \|\vec{b}\|}$$

$$= \|A\| \cdot \|A^{-1}\| \frac{\|\Delta\vec{b}\|}{\|\vec{b}\|}$$

also

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \leq \|A\|\,\|A^{-1}\| \frac{\|\Delta\vec{b}\|}{\|\vec{b}\|} \qquad (1.24)$$

Definition 2.26

Die Zahl

$$\text{cond}(A) = \kappa(A) = \|A\| \cdot \|A^{-1}\|$$

heißt Konditionszahl der Matrix A und ist eine Schranke für die Fehlerverstärkung bei der Lösung von (2.2) bei gestörter rechter Seite.

Zurück zum Beispiel 2.25

$$\|A\|_\infty \cdot \|A^{-1}\|_\infty = 4798{,}2 = \text{cond}(A)$$

Noch ein paar Normen:

p-Norm, $p \in \mathbb{N}, p \geq 1$, $\vec{x} \in \mathbb{R}^n$

$$\|\vec{x}\|_p = \left(|x_1|^p + |x_2|^p + \ldots + |x_n|^p\right)^{1/p},$$

induziert $\|A\|_p$

Frobenius-Norm einer $(m \times n)$-Matrix A

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$$ also die

(Euklidische bzw. 2-Norm von A als Vektor geschrieben)