3. $\vdots$

Jetzt: Eliminiere 1. und 2. wie folgt

$$u_2 + 2u_1 = \underbrace{h^2 f_1 + \alpha}_{\text{bekannt}}$$

Nun: $Au = b$ mit $A \in \mathbb{R}^{n,n}$, $b \in \mathbb{R}^n$, $u \in \mathbb{R}^n$ und A hat die Gestalt

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix} =: \text{tridiag}(-1, 2, -1)$$

Analog reduzieren wir die Randwerte im 2d-System. Man erhält dann eine Blocktridiagonalmatrix $\begin{bmatrix} A & B & & & \\ B & A & B & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & B \\ & & & B & A \end{bmatrix} \in \mathbb{R}^{n^2,n^2}$ mit $A, B \in \mathbb{R}^{n,n}$

1.3.5 Konvergenzbetrachtung

ÜA: Diese diskrete 2. Ableitung approximiert die exakte 2. Ableitung mit $\mathcal{O}(h^2)$ falls $u \in C^4(\mathbb{R})$. Mann kann dann zeigen, dass

$$\max_k |u''(x_k) - u_k''| \leq Ch^2$$

mit einem von u unabhängigen C.

2 Rundungsfehler und numerische Stabilität

2.1 Grenzen der Genauigkeit

Wir haben uns in Kapitel I darauf verlassen, dass $\lim\limits_{h \to 0} \frac{u(x+h)-u(x)}{h} = u'(x)$, falls $u \in C^1(\mathbb{R})$ auch auf dem Computer gilt. Wir rechnen das numerisch nach. Dazu definieren wir

$$\begin{aligned} g^{(1)}(x,h) &= \frac{1}{h}\left(u(x+h) - u(x)\right) \\ g^{(2)}(x,h) &= \frac{1}{2h}\left(u(x+h) - u(x-h)\right) \end{aligned}$$

Vorwärtsdifferentienquotient bzw Mitteldifferenzenquotient Sei x fest gewählt. Wir stellen den Wert

$$E^{(i)}(h) := |g^{(i)}(x,h) - u'(x)|$$

als Funktion von h dar. Wir erwarten $E^{(i)}(h) = \mathcal{O}(h^\kappa)$ für ein $\kappa \in \mathbb{N}$. Daraus folgt: $\log(E^{(i)}(h)) = C + \kappa \cdot \log(h)$. Im doppelt logarithmischen Plot erwarten wir eine Gerade mit Steigung κ

2.2 Zahldarstellung

2.2.1 Zahlsysteme

Dezimalbasis: Jede reelle Zahl x hat zur Basis 10 die Darstellung

$$x = x_M \cdot 10^M + x_{M-1} \cdot 10^{M-1} + \ldots + x_0 \cdot 10^0 + x_{-1} \cdot 10^{-1} + \ldots$$

mit Faktoren $x_l \in \{0, \ldots, 9\}$. Die Darstellung ist nicht notwendig endlich und nicht eindeutig ($0.\overline{9} = 1.0$).

Dualbasis: Verwende 2 statt 10.

$$x = x_M \cdot 2^M + x_{M-1} \cdot 2^{M-1} + \ldots + x_0 \cdot 2^0 + x_{-1} \cdot 2^{-1} + \ldots$$

Hexadezimal: zur Bais 16, Speicheradressen: $0, \ldots, 9, A, \ldots, F$

Beispiele:

$$\begin{aligned}
9_{10} &= 8 + 1 = 2^3 + 2^0 = 1001_2 \\
9.25_{10} &= 1001.01_2 \\
0.000\overline{1100}_2 &= \sum_{k=1}^{\infty} 2^{-4k} + 2^{-4k-1} = \sum_{k=1}^{\infty} \left(\frac{1}{16}\right)^k + \frac{1}{2}\left(\frac{1}{16}\right)^k \\
&= \frac{3}{2}\left(\frac{1}{1-\frac{1}{16}} - 1\right) = \frac{1}{10}
\end{aligned}$$

Bemerkung: $\frac{1}{10}$ hat im Dezimalsystem eine endliche, im Dualsystem eine unendliche Darstellung. Jedoch gilt: $\frac{1}{2} = 5 \cdot 10^{-1}$. Daher hat jede endliche Darstellung im Dualsystem eine endliche im Dezimalsystem.

2.2.2 Maschinenzahlen

Ein Rechner kennt nur endlich viele Zahlen. Man definiert eine Abbildung $\mathrm{rd} : \mathbb{R} \to \mathbb{F}$ (Menge der Maschinenzahlen) durch *Bestapproximation* oder *Abschneiden.*. Im Dezimalsystem lautet die allgemeine Darstellung einer Maschinenzahl $y \in \mathbb{F}(10, L, E_{min}, E_{max})$:

$$y = \pm 0, \underbrace{* \cdots\cdots *}_{\substack{\text{Mantisse,} \\ L \text{ Ziffern}}} \cdot 10^e$$

mit $e \in \{E_{min}, \dots, E_{max}\} \subset \mathbb{Z}$

Die *Maschinengenauigkeit* ε hat nach Definition die Eigenschaft

$$\varepsilon := \inf\{x > 0 : \ \mathrm{rd}(1-x) < 1$$

und es gilt: $\left|\frac{x-\mathrm{rd}(x)}{x}\right| \le \varepsilon$ für $x \in [\min \mathbb{F}, \max \mathbb{F}] \setminus \{0\}$

In C oder FORTRAN

float, real*4 $\quad \varepsilon \approx 10^{-8}$
double, real*8 $\quad \varepsilon \approx 10^{-16}$

Den arithmetischen Operationen $+, -, \cdot, /$ entsprechen Operationen in der Rechnerarithmetik $\tilde{+}, \tilde{-}, \tilde{\cdot}, \tilde{/}$ und es gilt für $\circ \in \{+, -, \cdot, /\}$

$$\mathrm{rd}(x) \,\tilde{\circ}\, \mathrm{rd}(y) = x \circ y(1 + \varepsilon_{xy}) \text{ mit } |\varepsilon_{xy}| \le \varepsilon$$

Leider gleten für das Zahlensystem $\mathbb{F}$ viele der üblichen Regeln (z.B. Assoziativgesetz) ($\rightarrow$ ÜA)

2.2.3 Rundungsfehleranalyse

Differenzenquotient: Wir halten in 1.1 die Differenzenquotienten $g^{(1)}(x,h)$ und $g^{(2)}(x,h)$ definiert.

$$\begin{aligned} g^{(1)}(x,h) &= \frac{1}{h}\left(f(x+h)(1+\varepsilon_1) - f(x)(1+\varepsilon_2)\right) \cdot (1+\varepsilon_0) \\ &= \left(\frac{f(xh) - f(x)}{h} + \frac{\varepsilon_1}{h} f(x+h) - \frac{\varepsilon_2}{h} f(x)\right)(1+\varepsilon_3) \end{aligned}$$

Dann ist $|g^{(1)}(x,h) - f'(x)| = \mathcal{O}(h) + \mathcal{O}\left(\frac{\varepsilon}{h}\right)$

Die Abschätzung ist optimal, wenn beide Summanden vergleichbar sind: $h \approx \frac{\varepsilon}{h} \Rightarrow h^2 \approx \varepsilon \Rightarrow h \approx \sqrt{\varepsilon}$. Der optimale Fehler ist dann $\mathcal{O}(\sqrt{\varepsilon})$. Analog für $g^{(2)} : h \approx \sqrt[3]{\varepsilon}$ und den Fehler $\sqrt[3]{\varepsilon}^2$

Skalarprodukt: Sei $S \equiv S(y) := [1, \dots, 1] \cdot y = \sum_{k=1}^{n} y_k$ für $y \in \mathbb{R}^n$.

Nun wollen wir $y \in \mathbb{F}^n$ annehmen und die Summe $\tilde{S}$ in Rechnerarithmetik bestimmen.

Algorithmus

$\tilde{S} := y_1$
for $k = 2 : n$
$\tilde{S} = \tilde{S} \tilde{+} y_k$
end

Beispiel: $n = 3$

$$\tilde{S} = ((y_1 + y_2)(1 + \varepsilon) + y_3)(1 + \varepsilon_2) = (y_1 + y_2)(1 + \varepsilon_1)(1 + \varepsilon_2) + y_3(1 + \varepsilon_2)$$

Induktion:

$$\tilde{S} = (y_1 + y_2) \prod_{i=1}^{n-1} (1 + \varepsilon_i) + \sum_{k=3}^{n} y_k \prod_{i=k-1}^{n-1} (1 + \varepsilon_i)$$

mit $|\varepsilon_i| \leq \varepsilon$ für $i = 1, \ldots, n$

Lemma 1. *Seien $\varepsilon_i, \varepsilon$ wie oben, $\sigma_i \in \{\pm 1\}$ $(i = 1, \ldots, n)$*
Ist $n\varepsilon < 1$, so gilt

$$\prod_{i=1}^{n} (1 + \varepsilon_i)^{\sigma_i} = 1 + \vartheta_n$$

mit $\vartheta_n \in \mathbb{R}$, $|\vartheta_n| \leq \dfrac{n\varepsilon}{1 - n\varepsilon} =: \gamma_n$

Bemerkung: $n \approx 10^6$ in einfacher und $n \approx 10^{15}$ in doppelter Genauigkeit.

Beweis. *Mit Induktion ÜA* $\square$

Theorem 1. *Für die Summation von n Zahlen in Rechnerarithmetik gilt die Abschätzung*

$$|\tilde{S} - S| \leq |y_1 + y_2|\gamma_{n-1} + \sum_{k=2}^{n} |y_k|\gamma_{n-k+1}$$

sowie

$$\frac{|\tilde{S} - S|}{|S|} \leq \gamma_{n-1} \left| \frac{\sum\limits_{k=1}^{n} |y_k|}{\sum\limits_{k=1}^{n} y_k} \right| = \gamma_{n-1} \frac{S(|y|)}{|S(y)|}$$

wobei $|y|$ hier komponentenweise zu verstehen ist.

Beachte: $\gamma_{n-1} \approx n\varepsilon$, *falls* $n\varepsilon \ll 1$

Beweis. *Direkt aus der Darstellung von $\tilde{S}$ und dem Lemma folgt die erste Abschätzung. Die γ_k wachsen monoton mit k, d.h. wir können $|\tilde{S} - S| \leq \gamma_{n-1}(|y_1| + |y_2|) + \gamma_{n-1} \sum\limits_{k=3}^{n} |y_k|$ abschätzen.* $\square$

Bemerkungen

- $\gamma_{n-1} \approx n\varepsilon$
- Erst die betraglich kleinen Zahlen addieren
- Schlecht ist der Fall $|S(y)| \ll S(|y|)$, Dies gilt z.B. für Differenzenquotienten

2.3 Konditionen von Abbildungen

Erinnerung: Vektornorm, zugeordnete Operatornorm, verträgliche Operatornorm $\rightarrow$ Ergänzungsblatt

Seien gegeben: Normierte lineare Vektorräume X, Y sowie $f : X \rightarrow Y$ stetige Abbildung.

2.3.1 Norm- und komponentenweise Kondition

Definition. Normweise absolute Kondition *ist die kleinste Zahl* κ_{abs} *mit*

$$\|f(\tilde{x}) - f(x)\|_Y \leq \kappa_{\text{abs}} \|\tilde{x} - x\|_X + o(\|\tilde{x} - x\|)_X \quad (\tilde{x} \rightarrow x)$$

Normweise relative Kondition *ist die kleinste Zahl* κ_{rel} *mit*

$$\frac{\|f(\tilde{x}) - f(x)\|_Y}{\|f(x)\|_Y} \leq \kappa_{\text{rel}} \frac{\|\tilde{x} - x\|}{\|x\|_X} + o(\|\tilde{x} - x\|_X) \quad (\tilde{x} \rightarrow x)$$

für $x \neq 0, f(x) \neq 0$
Komponentenweise relative Kondition *ist die kleinste Zahl* $\kappa_r el$ *mit*

$$\left\| \frac{f(\tilde{x}) - f(x)}{f(x)} \right\|_Y \leq \kappa_{\text{rel}} \left\| \frac{\tilde{x} - x}{x} \right\|_X + o(\|\tilde{x} - x\|_X) \quad (\tilde{x} \rightarrow x)$$

Je nach Größenordnung von $\kappa \in \{\kappa_{\text{rel}}, \kappa_{\text{abs}}\}$ *nennt man eine Abbildung von* f *in* x gut *(*$\kappa \approx 1$*) oder* schlecht *(*$\kappa \gg 1$*)* konditioniert

Ist f *differenzierbare* Abbildung, so setzen wir

$$\begin{aligned}
\kappa_{\text{abs}} &:= |||f'(x)||| \\
\kappa_{\text{rel}} &:= \frac{|||f'(x)||| \cdot \|x\|_X}{\|f(x)\|_Y} \quad \text{(normweise)} \\
\kappa_{\text{rel}} &:= \left\| \frac{|f'(x)| \cdot |x|}{|f(x)|} \right\|_Y \quad \text{(komponentenweise)}
\end{aligned}$$

Letzteres mit komponentenweiser Definition von $|\cdot|$ und Division. $|||\cdot|||$ Operatornorm zu $\|\cdot\|_X, \|\cdot\|_Y$

2.3.2 Beispiele

- Addition: $f : \mathbb{R}^2 \rightarrow \mathbb{R},\ [x_1, x_2] \mapsto x_1 + x_2,\ \|x\| := |x_1| + |x_2| =: |x|_1$.
  Es gilt: $f'(x) = [1, 1]$. Also folgt:

$$\begin{aligned}
\kappa_{\text{abs}} &= \max_y \frac{|[1,1] \cdot y|}{|y|_1} \leq \frac{|y_1| + |y_2|}{|y|_1} = 1 \\
\kappa_{\text{rel}} &= \frac{1 \cdot |x|_1}{|\underbrace{x_1 + x_2}_{=f(x)}|} = \frac{|x_1| + |x_2|}{|x_1 + x_2|} \quad \text{(normweise und komponentenweise)}
\end{aligned}$$

Die Addition zweier Zahlen ist „schlecht konditioniert“ falls $x_1 \approx x_2$ (*Stellenauslöschung*). Sie ist„gut konditioniert“ falls $|x_1| + |x_2| = |x_1 + x_2| \Rightarrow \kappa_{\text{rel}} = 1$.

- Multiplikation zweier Zahlen $f : \mathbb{R}^2 \to \mathbb{R},\ [x_1, x_2] \mapsto x_1 \cdot x_2,\ |\cdot|_1$.
  Es gilt: $f'(x) = [x_2, x_1]$

$$\begin{aligned}\kappa_{\text{abs}} &= \max_y \frac{|f'(x)\cdot y|}{|y|_1} = \frac{|x_2 y_1 + x_1 y_2|}{|y_1| + |y_2|} \leq \max\{|x_1|, |x_2|\} \\ \kappa_{\text{rel}} &= \left|\frac{|f'(x)| \cdot |x|}{|f(x)|}\right| = \frac{|[x_2, x_1] \cdot [x_1, x_2]|}{|x_1 \cdot x_2|} = \frac{2 \cdot |x_1 x_2|}{|x_1 x_2|} = 2\end{aligned}$$

- Lösen eines linearen Gleichungssystem:
  Gegeben: A invertierbar in $\mathbb{R}^{n,n},\ b \in \mathbb{R}^n$
  Finde $u \in \mathbb{R}^n$ sodass gilt $Au = b$

  1. Störung der rechten Seite b: $f(b) := u = A^{-1}b$
     Wir betrachten die normweise Kondition: $f'(b) = A^{-1}$

$$\Rightarrow \kappa_{\text{abs}} = |||A^{-1}|||$$

  $\|\cdot\|$ gewählte Vektornorm, $|||\cdot|||$ zugeordnete Operatornorm

$$\begin{aligned}\kappa_{\text{rel}} &= \frac{|||A^{-1}||| \cdot \|b\|}{\|A^{-1}b\|} = \frac{|||A^{-1}||| \cdot \|AA^{-1}b\|}{\|A^{-1}b\|} \leq \frac{|||A^{-1}||| \cdot |||A||| \cdot \|A^{-1}b\|}{\|A^{-1}b\|} \\ &= |||A^{-1}||| \cdot |||A||| =: \text{cond}_{|||\cdot|||}(A) \text{ (Kondition von } A)\end{aligned}$$

  2. Einfluss der Störung von A:
     Betrachte nun u als Funktion von A: $f : \mathbb{R}^{n,n} \to \mathbb{R}^n,\ f(A) = u = A^{-1}b$
     Es gilt:

$$f'(A)E = -A^{-1}EA^{-1}b = -A^{-1}Eu$$

  Daraus folgt:

$$\begin{aligned}|||f'(A)||| &= \sup_E \frac{\|f'(A)E\|}{|||E|||} = \sup_E \frac{\|A^{-1}Eu\|}{|||E|||} \\ &\leq \sup_E \frac{|||A^{-1}||| \cdot |||E||| \cdot \|u\|}{|||E|||} = |||A^{-1}||| \cdot \|u\| \\ \Rightarrow \kappa_{\text{rel}} &\leq \frac{|||A^{-1}||| \cdot \|u\| \cdot |||A|||}{\|u\|} = \text{cond}_{|||\cdot|||}(A)\end{aligned}$$

2.4 Stabilität numerischer Algorithmen

Die Kondition von f in x beschreibt den unvermeidlichen Fehler der Rechenvorschrift $x \mapsto f(x)$.
Es sei $\tilde{f}(x)$ die Vorschrift zur Berechnung von $f(x)$ wir rechnen damit, dass selbst bei exakter Arithmetik auf $\mathbb{F}$ der relative Fehler $\kappa_f(x)\varepsilon$ auftritt.

2.4.1 Vorwärtsanalyse

Definition. *Der* Stabilitätsindikator *des Algorithmus $\tilde{f}(x)$ zur Berechnung von $f(x)$ ist die kleinste Zahl σ, so dass gilt*

$$\frac{\|\tilde{f}(\tilde{x})\|_Y}{\|f(\tilde{x})\|_Y} \leq \sigma \underbrace{\kappa_f(\tilde{x})}_{\substack{\kappa_{\text{rel}} \\ \text{normw.}}} \varepsilon + o(\varepsilon) \quad (\varepsilon \to 0)$$

für alle $\tilde{x}$ mit $\|\tilde{x} - x\|_X \leq \varepsilon \cdot \|x\|_X$
Der Algorithmus $\tilde{f}$ ist stabil im Sinne der Vorwärtsanalyse, *falls σ kleiner gleich der Anzahl der elementaren Rechenoperationen ist.*

Beispiel: Die Summation:

$\tilde{S}_1 := y_1$
for $i = 2 : n$ $\tilde{S}_i = \tilde{S}_{i-1} \oplus y$

Es gilt:

$$\frac{|\tilde{S}(y) - S(y)|}{|S(y)|} \leq \gamma_{n-1}\varepsilon \cdot \frac{S(|y|)}{|S(y)|} = (n-1)\varepsilon\kappa_S + o(\varepsilon), \text{ falls } n\varepsilon \ll 1$$

Also $\sigma < n - 1$, d.h. die Summation ist vorwärtsstabil.

2.4.2 Rückwärtsanalyse

Definition. *Der* Stabilitätsindikator *der Rückwärtsanalyse des Algorithmus $x \mapsto \tilde{f}(x)$, $x \in E$ ist die kleinstmögliche Zahl ϱ, so dass für alle $\tilde{x} \in E$ mit $\|\tilde{x} - x\|_X \leq \varepsilon\|x\|_X$ ein $\hat{x} \in E$ existiert mit $\tilde{f}(\tilde{x}) = f(\hat{x})$, so dass*

$$\frac{\|\hat{x} - \tilde{x}\|_X}{\|\tilde{x}\|_X} \leq \varrho\varepsilon + o(\varepsilon) \quad (\varepsilon \to 0)$$

Der Algorithmus $\tilde{f}$ heißt stabil *im Sinne der Rückwärtsanalyse, falls ϱ kleiner gleich der Anzahl der elementaren Rechenoperationen*

Lemma 2. *(*Rückwärtsstabil $\Rightarrow$ Vorwärtsstabil*)*

$$\sigma \leq \varrho$$

Beweis. *Sei $\tilde{x} \in E$ mit $\|x - \tilde{x}\|_X \leq \varepsilon \cdot \|x\|_X$. Dann gilt*

$$\begin{aligned} \frac{\|\tilde{f}(\tilde{x}) - f(\tilde{x})\|_Y}{\|f(\tilde{x})\|_Y} &\overset{\text{Vor.}}{=} \frac{\|f(\hat{x}) - f(\tilde{x})\|_Y}{\|f(\tilde{x})\|_Y} \\ &\overset{\text{Def } \kappa_f}{\leq} \kappa_f(\hat{x})\frac{\|\hat{x} - \tilde{x}\|_X}{\|\tilde{x}\|_X} + o(\varepsilon) \\ &\overset{\text{Vor.}}{\leq} \varrho\varepsilon \cdot \kappa_f(\tilde{x}) + o(\varepsilon) \end{aligned}$$

$\Rightarrow \sigma \leq \varrho$ *nach Def. von σ* $\square$