Technische Universität München
Institut für Informatik
Prof. Dr. Hans-Joachim Bungartz
Dipl.-Inf. Christoph Riesinger
Dipl.-Math. Jürgen Bräckle

WS 2013/2014

Numerisches Programmieren, Übungen

Musterlösung 2. Übungsblatt: Kondition, Stabilität und Gauß-Elimination

1) Kondition und Stabilität

a) i) $cond(f_1, x) = \left| \frac{x}{a \cdot x} \cdot a \right| = 1$

$\Rightarrow$ keine Zunahme des relativen Fehlers durch Multiplikation $\Rightarrow$ gut konditioniert

ii) $cond(f_2, x) = \left| \frac{x}{a - x} \cdot (-1) \right| = \left| \frac{x}{a - x} \right|$

$\Rightarrow cond(f_2, x) \gg 1$ für $x \approx a$ und $a \neq 0$ $\Rightarrow$ schlecht konditioniert für $x \approx a$

iii) $cond(f_3, x) = \left| \frac{x}{3e^x - 3} \cdot 3e^x \right| = \left| \frac{x \cdot e^x}{e^x - 1} \right|$

$$\begin{aligned} \lim_{x \to 0} cond(f_3, x) &= \lim_{x \to 0} \left| \frac{x \cdot e^x}{e^x - 1} \right| = \left| \lim_{x \to 0} \frac{x \cdot e^x}{e^x - 1} \right| \overset{\text{l'Hosp.}}{=} \left| \lim_{x \to 0} \frac{e^x + x \cdot e^x}{e^x} \right| \\ &= \left| \lim_{x \to 0} (1 + x) \right| = 1 \end{aligned}$$

Insgesamt ergibt sich für die interessanten Fälle:

$$cond(f_3, x) \to \begin{cases} \infty, & ; \text{für } x \to \infty \\ 1, & ; \text{für } x \to 0 \\ 0, & ; \text{für } x \to -\infty \end{cases}$$

b) Mit Rundungsfehlern $\varepsilon_1,\ \varepsilon_2,\ \varepsilon_3$ im Bereich der Maschinengenauigkeit ε_M, d.h. $|\varepsilon_i| \leq \varepsilon_M$, und ohne Eingabefehler stellt sich die gerundete f-Auswertung wie folgt dar (Term mit $\varepsilon_i \cdot \varepsilon_j \doteq 0$, d.h. vernachlässigbar klein):

$$\begin{aligned} rd(f_3(x)) &= (3e^x(1 + \varepsilon_1)(1 + \varepsilon_2) - 3)(1 + \varepsilon_3) \\ &\doteq (3e^x(1 + \varepsilon_1 + \varepsilon_2) - 3)(1 + \varepsilon_3) \\ &= (3e^x - 3)(1 + \varepsilon_3) + 3e^x(\varepsilon_1 + \varepsilon_2)(1 + \varepsilon_3) \\ &\doteq f_3(x)(1 + \varepsilon_3) + 3e^x(\varepsilon_1 + \varepsilon_2). \end{aligned}$$

Damit ergibt sich für den die Stabilität beschreibenden relativen Fehler:

$$\begin{aligned}\left|\frac{rd(f_3(x)) - f_3(x)}{f_3(x)}\right| &= \left|\frac{f_3(x)\cdot\varepsilon_3 + 3e^x(\varepsilon_1+\varepsilon_2)}{f_3(x)}\right| = \left|\varepsilon_3 + (\varepsilon_1+\varepsilon_2)\frac{3e^x}{3e^x-3}\right| \\ &\overset{\Delta\text{-Ugl.}}{\leq} |\varepsilon_3| + \left|(\varepsilon_1+\varepsilon_2)\frac{e^x}{e^x-1}\right| \leq \varepsilon_M + 2\varepsilon_M\left|\frac{e^x}{e^x-1}\right| \\ &= \varepsilon_M\left(1 + 2\cdot\left|\frac{e^x}{e^x-1}\right|\right) \to \infty \text{ für } x \to 0.\end{aligned}$$

Der Algorithmus zur Auswertung von $f_3(x)$ ist also stabil für alle Werte von x außer der Null. **Achtung:** Für zunehmendes x haben wir bereits in Teilaufgabe a,iii) festgestellt, dass auch die Konditionszahl von $f_3(x)$ steigt. Also ist zwar die Auswertung von $f_3(x)$ für große x per Definition stabil, diese Aussage ist jedoch in Anbetracht der schlechten Kondition wertlos!
Für $x \approx 0$ haben wir den Fall eines instabilen Algorithmus trotz guter Kondition!

2) Beispiel für schlechte Kondition: Schnittpunkt zweier Geraden

a) Durch gleichsetzen der beiden Geraden

$$\begin{aligned}g_1 = g_2 &\Leftrightarrow x = mx + 1 \\ &\Leftrightarrow x = \frac{1}{1-m} =: f(m)\end{aligned}$$

erhalten wir den x-Wert des Schnittpunktes.

b) Mit der Ableitung

$$f'(m) = (-1)\cdot(1-m)^{-2}\cdot(-1) = \frac{1}{(1-m)^2}$$

gilt für die Konditionszahl

$$cond(f,m) = \left|\frac{m(1-m)^{-2}}{(1-m)^{-1}}\right| = \left|\frac{m}{1-m}\right|.$$

Für $m \approx 1$ ($\hat{=}$ Geraden fast parallel) folgt somit $cond \to \infty$, d.h. es liegt ein schlecht konditioniertes Problem vor. Das wollen wir uns jetzt an einem genauen Zahlenbeispiel verdeutlichen. Für die gegebene Stelle $m = 1.005$ haben wir die Kondition

$$cond(f, 1.005) = \left|\frac{1+\frac{1}{200}}{\frac{1}{200}}\right| = \underline{201}.$$

Damit erwarten wir, dass mit $m = 1.005$ ein relativer Eingabefehler um ungefähr das 200-fache verstärkt wird.

c) Um wieviel wird der Eingabefehler nun wirklich verstärkt? Dazu berechnen wir den tatsächlichen Schnittpunkt, den fehlerhaften Schnittpunkt, und die relativen Fehler in

Ein- und Ausgabe.

$$
\begin{aligned}
x = f(m) &= \frac{1}{1-m} = \frac{1}{1-1.005} = -200 \\
\tilde{x} = f(\tilde{m}) &= \frac{1}{1-\tilde{m}} = \frac{1}{1-1.01} = -100 \\
\text{relativer Eingabefehler} &: \left|\frac{m-\tilde{m}}{m}\right| = \left|\frac{\frac{1}{200}}{\frac{201}{200}}\right| = \frac{1}{201} \\
\text{relativer Ausgabefehler} &: \left|\frac{f(m)-f(\tilde{m})}{f(m)}\right| = \left|\frac{-200+100}{-200}\right| = \frac{1}{2}
\end{aligned}
$$

Somit ist der verstärkende Faktor

$$
\frac{\text{relativer Ausgabefehler}}{\text{relativer Eingabefehler}} = \frac{\frac{1}{2}}{\frac{1}{201}} = \underline{100.5}
$$

3) Ableitungsapproximation

a) Laut der Taylorformel gilt

$$
f'(x_0) \;=\; \frac{1}{h}\left(f(x_0+h) - f(x_0) - \frac{1}{2!}f''(z)\cdot h^2\right), \quad z \in (x_0; x_0+h).
$$

Damit erhalten wir für den absoluten Fehler err_{abs}:

$$
\begin{aligned}
err_{abs} \quad &= \quad \left|f'(x_0) - rd\big(D_f(x_0,h)\big)\right| \;=\; \left|\frac{f(x_0+h)-f(x_0)}{h} - \frac{1}{2!}f''(z)\cdot h \;-\; rd\big(D_f(x_0,h)\big)\right| \\
&= \quad \left|-\frac{1}{2!}f''(z)\cdot h \;+\; D_f(x_0,h) - rd\big(D_f(x_0,h)\big)\right| \\
&\overset{\Delta\text{-Ugl.}}{\leq} \quad \underbrace{\left|\frac{1}{2!}f''(z)\cdot h\right|}_{=\text{ analytischer Fehler}} \;+\; \underbrace{\left|D_f(x_0,h) - rd\big(D_f(x_0,h)\big)\right|}_{=\text{ numerischer Fehler}}.
\end{aligned}
$$

Jetzt betrachten wir den Rundungsfehler in der Eingabe $(\varepsilon_0, \varepsilon_1)$ und der Berechnungsvorschrift $(\varepsilon_2, \varepsilon_3)$ für den Differenzenquotienten. Dabei gehen wir vereinfacht davon aus, dass bei $x_0 + h$ kein zusätzlicher Fehler auftritt.

$$
\begin{aligned}
rd\big(D_f(x_0,h)\big) \;&=\; \frac{\left(f(x_0+h)(1+\varepsilon_1) - f(x_0)(1+\varepsilon_0)\right)(1+\varepsilon_2)}{h}(1+\varepsilon_3) \\
&\doteq\; \frac{1+\varepsilon_2+\varepsilon_3}{h}\left(f(x_0+h) + \varepsilon_1 f(x_0+h) - f(x_0) - \varepsilon_0 f(x_0)\right) \\
&=\; \frac{1+\varepsilon_2+\varepsilon_3}{h}\left(f(x_0+h) - f(x_0)\right) \;+\; \frac{1+\varepsilon_2+\varepsilon_3}{h}\left(\varepsilon_1 f(x_0+h) - \varepsilon_0 f(x_0)\right) \\
&\doteq\; \frac{1+\varepsilon_2+\varepsilon_3}{h}\left(f(x_0+h) - f(x_0)\right) \;+\; \frac{1}{h}\left(\varepsilon_1 f(x_0+h) - \varepsilon_0 f(x_0)\right). \qquad (1)
\end{aligned}
$$

Die Terme höherer Ordnung $(\varepsilon_i \cdot \varepsilon_k)$ können wir dabei vernachlässigen $(\doteq)$.

Nach dem Mittelwertsatz gibt es ein $y \in (x_0; x_0+h)$, so dass $f'(y) = \frac{f(x_0+h)-f(x_0)}{x_0+h-x_0}$. Damit gilt also $D_f(x_0,h) = f'(y)$ und mit Gleichung (1):

$$\begin{aligned} rd\big(D_f(x_0,h)\big) &\doteq D_f(x_0,h) + (\varepsilon_2+\varepsilon_3)f'(y) \; + \; \frac{1}{h}\left(h\cdot\varepsilon_1 f'(y) + (\varepsilon_1-\varepsilon_0)f(x_0)\right) \\ &= D_f(x_0,h) + (\varepsilon_1+\varepsilon_2+\varepsilon_3)f'(y) \; + \; \frac{1}{h}(\varepsilon_1-\varepsilon_0)f(x_0). \end{aligned}$$

Somit erhalten wir damit für den numerischen Fehler (mit $|\varepsilon_0|, |\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \leq \varepsilon_{Ma},\ h > 0$):

$$\begin{aligned} \big|D_f(x_0,h) - rd\big(D_f(x_0,h)\big)\big| &\doteq \left|(\varepsilon_1+\varepsilon_2+\varepsilon_3)f'(y) \; + \; \frac{1}{h}(\varepsilon_1-\varepsilon_0)f(x_0)\right| \\ &\overset{\Delta\text{-Ugl.}}{\leq} 3\;\varepsilon_{Ma}|f'(y)| + \frac{2\;\varepsilon_{Ma}}{h}|f(x_0)| \\ &\leq \varepsilon_{Ma}\cdot \underbrace{3 \max_{y\in[x_0;x_0+h]} |f'(y)|}_{=:\;c_1} + \frac{\varepsilon_{Ma}}{h}\cdot \underbrace{2|f(x_0)|}_{=:\;c_2} \\ &= \varepsilon_{Ma}\cdot c_1 + \frac{\varepsilon_{Ma}}{h}\cdot c_2 \overset{h\to 0}{\to} \infty. \end{aligned}$$

Insgesamt ergibt sich also für den absoluten Fehler:

$$err_{abs} \leq \left|\frac{1}{2!}f''(z)\cdot h\right| + \varepsilon_{Ma}\left(c_1 + \frac{c_2}{h}\right) \leq h\cdot \underbrace{\frac{1}{2}\max_{z\in[x_0;x_0+h]}|f''(z)|}_{=:\;c_3} + \varepsilon_{Ma}\left(c_1+\frac{c_2}{h}\right) \quad (2)$$

Hierbei wird angenommen, dass c_1 und c_3 unabhängig von h sind (z.B. falls $f'(x)$ und $f''(x)$ beschränkt) . Wichtige Beobachtung: Der analytische Fehler verhält sich proportional zu h, der numerische proportional zu $1/h$ (vgl. Bild auf Angabenblatt bzw. Abb. 1).

b) Die Abschätzung (2) aus a) hängt von h ab. Um einen möglichst kleinen absoluten Fehler err_{abs} zu erhalten, müssen wir ein 1-D-Minimierungsproblem in h lösen:

$$\begin{aligned} \min_h \left\{h\cdot c_3 + \varepsilon_{Ma}\left(c_1+\frac{c_2}{h}\right)\right\} \;\Rightarrow\; 0 &\overset{!}{=} \frac{d}{dh}\left(h\cdot c_3 + \varepsilon_{Ma}\left(c_1+\frac{c_2}{h}\right)\right) \\ &= c_3 - \varepsilon_{Ma}\frac{c_2}{h^2} \end{aligned}$$

Damit erhalten wir als optimales h : $h_{opt} = \sqrt{\frac{\varepsilon_{Ma}\cdot c_2}{c_3}}$

In Abbildung 1 ist der absolute Fehler err_{abs} über der Schrittweite h aufgetragen für die Approximation der Ableitung von $sin(x)$ an der Stelle $x = 1$ mit Hilfe von zwei Differenzenquotienten. Der Graph besitzt doppelt logarithmische Skalen und wurde mit dem Matlab-Programm deriv_approx_sin.m erstellt, das auch auf der Webseite zu den Übungen zu finden ist.

$$\begin{aligned} \text{forward difference quotient} &:= \frac{f(x_0+h)-f(x_0)}{h} \\ \text{central difference quotient} &:= \frac{f(x_0+h/2)-f(x_0-h/2)}{h} \end{aligned}$$

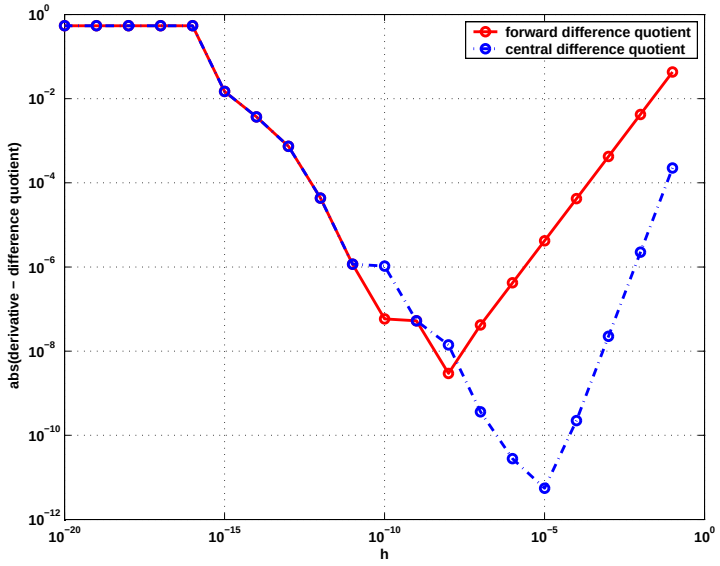

Abbildung 1: Darstellung der optimalen Schrittweite h in 10er-Potenzen für die Approximation von $\frac{d}{dx}(sin(x))|_{x=1}$ durch Differenzenquotienten.

4) Gauß-Elimination

a) $\left(\begin{array}{cc|c} 1 & 2 & 1 \\ 3 & 4 & 1 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & -2 & -2 \end{array}\right)$

Rückwärtssubstitution: $x_2 = \frac{-2}{-2} = 1, \quad x_1 = 1 - 2 \cdot x_2 = -1$

$\Rightarrow x = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

Die Lösung ist eindeutig. Sie entspricht einem Punkt im $\mathbb{R}^2$.

b) $\left(\begin{array}{cc|c} 1 & 2 & 1 \\ 3 & 6 & 3 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 0 & 0 \end{array}\right)$

Rückwärtssubstitution: $x_2 = \lambda, \quad x_1 = 1 - 2 \cdot x_2 = 1 - 2 \cdot \lambda$

$\Rightarrow x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda \cdot \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

Es gibt unendlich viele Lösungen. Sie entsprechen einer Geraden im $\mathbb{R}^2$.

c) $\left(\begin{array}{ccc|c} 0 & 4 & 8 & 0 \\ 1 & -5 & 2 & 0 \\ 2 & 1 & -7 & 1 \end{array}\right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -5 & 2 & 0 \\ 0 & 4 & 8 & 0 \\ 2 & 1 & -7 & 1 \end{array}\right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -5 & 2 & 0 \\ 0 & 4 & 8 & 0 \\ 0 & 11 & -11 & 1 \end{array}\right)$

$\rightsquigarrow \left(\begin{array}{ccc|c} 1 & -5 & 2 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 11 & -11 & 1 \end{array}\right) \rightsquigarrow \left(\begin{array}{ccc|c} 1 & -5 & 2 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & -33 & 1 \end{array}\right)$

Rückwärtssubstitution: $x_3 = -\frac{1}{33}, \quad x_2 = -2 \cdot x_3 = \frac{2}{33}, \quad x_1 = 5 \cdot x_2 - 2 \cdot x_3 = \frac{12}{33}$

$$\Rightarrow x = \frac{1}{33}\begin{pmatrix} 12 \\ 2 \\ -1 \end{pmatrix}$$

Die Lösung ist eindeutig. Sie entspricht einem Punkt im $\mathbb{R}^3$.

Wiederholung: Gleitpunktzahlen, Rundung und Kondition

a)

$$\begin{aligned} rd(f(x)) &= \frac{(e^x(1+\varepsilon_1) - e^{-x}(1+\varepsilon_2))(1+\varepsilon_3)}{3}(1+\varepsilon_4) \\ &= \frac{e^x(1+\varepsilon_1)(1+\varepsilon_3)(1+\varepsilon_4) - e^{-x}(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4)}{3} \\ &\doteq \frac{e^x(1+\varepsilon_1+\varepsilon_3+\varepsilon_4) - e^{-x}(1+\varepsilon_2+\varepsilon_3+\varepsilon_4)}{3} \end{aligned}$$

b)

$$\begin{aligned} \left|\frac{rd(f(x)) - f(x)}{f(x)}\right| &= \left|\frac{\frac{e^x-e^{-x}}{3} - \frac{e^x(1+\varepsilon_1+\varepsilon_3+\varepsilon_4)-e^{-x}(1+\varepsilon_2+\varepsilon_3+\varepsilon_4)}{3}}{\frac{e^x-e^{-x}}{3}}\right| \\ &= \left|\frac{e^x(\varepsilon_1+\varepsilon_3+\varepsilon_4) - e^{-x}(\varepsilon_2+\varepsilon_3+\varepsilon_4)}{e^x - e^{-x}}\right| \\ &\leq \frac{|e^x| \cdot 3\varepsilon_M + |e^{-x}| \cdot 3\varepsilon_M}{|e^x - e^{-x}|} = 3\varepsilon_M \cdot \frac{e^x + e^{-x}}{|e^x - e^{-x}|} \end{aligned}$$

mit $|\varepsilon_i| \leq \varepsilon_M$ (Maschinengenauigkeit)

- Für $x \to \infty$ gilt:
$$\lim_{x\to\infty}\left(3\varepsilon_M \cdot \frac{e^x + e^{-x}}{|e^x - e^{-x}|}\right) = 3\varepsilon_M$$
- Für $x \to -\infty$ gilt wegen der Symmetrie ebenfalls
$$\lim_{x\to-\infty}\left(3\varepsilon_M \cdot \frac{e^x + e^{-x}}{|e^x - e^{-x}|}\right) = 3\varepsilon_M$$
- Wegen
$$3\varepsilon_M \cdot \frac{e^x + e^{-x}}{|e^x - e^{-x}|} \xrightarrow[x\to 0]{} \infty$$
ist die Auswertung für $x \approx 0$ instabil, d.h. die Auswertung ist nicht für alle $x \in \mathbb{R}$ stabil.

c)

$$cond(f,x) = \left|\frac{x \cdot \frac{e^x+e^{-x}}{3}}{\frac{e^x-e^{-x}}{3}}\right| = \left|x \cdot \frac{e^x + e^{-x}}{e^x - e^{-x}}\right| = \left|x \cdot \frac{e^{2x}+1}{e^{2x}-1}\right|$$

An der Stelle $x = 0$ gilt:

$$\lim_{x\to 0} cond(f,x) = \lim_{x\to 0}\left|x \cdot \frac{e^{2x}+1}{e^{2x}-1}\right| \overset{\text{l'H.}}{=} \lim_{x\to 0}\left|\frac{x \cdot 2e^{2x} + 1 \cdot (e^{2x}+1)}{2e^{2x}}\right| = 1$$

$\Rightarrow$ Für $x = 0$ ist das Problem somit also gut konditioniert!