Technische Universität München
Institut für Informatik
Prof. Dr. Hans-Joachim Bungartz
Hayden Liu Weng
Sebastian Wolf
Michael Obersteiner

WiSe 2021/22

# Numerisches Programmieren, Übungen

Musterlösung 2. Übungsblatt: Kondition und Stabilität

## 1) Kondition

Anhand der »Kondition« beschreibt man die Abhängigkeit der Ausgabedaten von einer Störung der Eingabedaten eines Problems. Sie ist eher ein qualitativer Begriff - man redet von »guter« oder »schlechter« Kondition eines Problems.

Die »Konditionszahl« stellt ein qualitatives Maß dafür dar und entspricht dem asymptotisch ungünstigsten Faktor um den Störungen der Eingabe in der Ausgabe verstärkt werden.

Die (relative) Konditionszahl $\operatorname{cond}(f, x)$ für reellwertige Funktionen $f$ ist definiert als:

$$\operatorname{cond}(f, x) = \left| \frac{x \cdot f'(x)}{f(x)} \right|.$$

Berechnen Sie die relative Konditionszahl der folgenden Funktionen in Abhängigkeit von $x$:

i) $f_1(x) = a \cdot x$, ii) $f_2(x) = \frac{a-x}{b}$, iii) $f_3(x) = 3e^x - 3$.

Interpretieren Sie jeweils das Ergebnis!

Wie lautet die Konditionszahl von $f_3$ an der Stelle $x = 0$ (Grenzwertbetrachtung!)?

**Lösung:**

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

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

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

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

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

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

Insgesamt ergibt sich für die interessanten Fälle:

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

## 2) Beispiel für schlechte Kondition: Schnittpunkt zweier Geraden

Gegeben seien zwei Geraden $g_1$ und $g_2$ mit

$$\begin{aligned} g_1 &: y = x \\ g_2 &: y = mx + 1, \end{aligned}$$

deren Schnittpunkt berechnet werden soll. Der tatsächliche Eingabe-Parameter $m = 1.005$ wird dabei zu $\tilde{m} = 1.01$ aufgerundet. Wir wollen nun den dadurch entstandenen Fehler im x-Wert des Schnittpunktes untersuchen.

a) Berechnen Sie den x-Wert des Schnittpunktes für ein allgemeines $m$, und stellen Sie diese Beziehung als Funktion $x = f(m)$ dar.

b) Berechnen Sie die Konditionszahl des Problems aus a). Wann ist das Problem gut konditioniert? Wann nicht?

c) Werten Sie die Konditionszahl aus b) an der gegebenen Stelle $m$ aus.

d) Wie sieht die tatsächliche Verstärkung des relativen Eingabefehlers aus?

**Lösung:**

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. Die Funktion $f$ beschreibt nun den Zusammenhang zwischen der Eingabe $m$ und der Ausgabe $x$.

b) Mit der Ableitung

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

gilt für die Konditionszahl

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

Für $m \approx 1$ ($\hat{=}$ Geraden fast parallel) folgt somit cond $\to \infty$, d.h. es liegt ein schlecht konditioniertes Problem vor.

c) Für die gegebene Stelle $m = 1.005$ haben wir die Kondition

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

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

d) 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) Stabilität

Betrachten wir kurz nochmals die Menge $G$ vom Übungsblatt 1 (1 Bit Vorzeichen, 5 Bits Exponent, 2+1 Bits Mantisse). Es gilt $1, 3, 8 \in G$, aber $1 + 8 = 9 \notin G$ und $3 \cdot 3 = 9 \notin G$. Es kommt oft vor, dass die Ausgabe einer (auch sehr einfachen) Operation op nicht in der Menge der Maschinenzahlen ist und dadurch eine Rundung benötigt:

$$\mathrm{rd_G}(1+8) = \mathrm{rd_G}(9) = \mathrm{rd_G}(2^3 \cdot 1,0\mathbf{0}|100_2) = 2^3 \cdot 1,00_2 = 8.$$

Dadurch muss potenziell nach jeder Zwischenausgabe gerundet werden.

Ein **Verfahren** heißt **stabil**, wenn Rundungen von Zwischenergebnissen nicht zu einer großen Abweichung der Endergebnisse führen; ansonsten heißt es **instabil**. Um das Stabilitätsverhalten eines Verfahrens zu analysieren, verwenden wir **Epsilontik**.

Im Übungsblatt 1 haben wir gesehen, dass bei $\mathrm{rd_G}$ immer ein relativer Fehler $\leq \varepsilon_{\mathrm{Ma}}$ entsteht:

$$\left|\frac{\mathrm{rd_G}(x) - x}{x}\right| \leq \varepsilon_{\mathrm{Ma}}. \tag{1}$$

Das heißt, man kann immer ein $\varepsilon \in \mathbb{R}$, finden, damit

$$\mathrm{rd_G}(x) = (1+\varepsilon)x, \tag{2}$$
$$-\varepsilon_{\mathrm{Ma}} \leq \varepsilon \leq \varepsilon_{\mathrm{Ma}} \tag{3}$$

gelten. Diese Eigenschaft verwenden wir bei der Epsilontik. Wir untersuchen den relativen Endergebnisfehler $\left|\frac{\mathrm{rd}(f)(x) - f(x)}{f(x)}\right|$ anhand folgender Regeln:

- Bei der Ausführung jeder Maschinenoperation $\mathrm{op_M}$ wird einen neuen relativen Fehler $\varepsilon_i$ erzeugt:

$$(a \,\mathrm{op_M}\, b) = \mathrm{rd_M}(a \,\mathrm{op}\, b) = (a \,\mathrm{op}\, b) \cdot (1 + \varepsilon_i) \text{ mit } |\varepsilon_i| \leq \varepsilon_{\mathrm{Ma}}.$$

- Terme höheren Ordnung werden vernachlässigt: $\varepsilon_i \cdot \varepsilon_j \doteq 0 \quad \forall i, j.$

Untersuchen Sie die Stabilität von den Funktionen

i) $f_1(x) = a \cdot x,$ ii) $f_2(x) = \frac{a-x}{b},$ iii) $f_3(x) = 3e^x - 3.$

mit Hilfe der Epsilontik. Hier nehmen wir an, dass die Eingabe $x$ schon eine Gleitkommazahl ist und keine Rundung benötigt ($\mathrm{rd}(x) = x$). Die Auswertung von $e^x$ erzeuge auch nur einen relativen Fehler $\leq \varepsilon_{Ma}$.

**Lösung:** Mit der Epsilontik lässt sich die Fehlerverstärkung innerhalb eines Berechnungsverfahrens analysieren. Bei der Betrachtung des relativen Ausgabefehlers $\left|\frac{\mathrm{rd}(f)(x) - f(x)}{f(x)}\right|$ gelten dabei folgende Regeln:

- Jede Operation erzeugt einen relativen Fehler $\leq \varepsilon_M$, d.h.

$$(a \,\mathrm{op_M}\, b) = (a \,\mathrm{op}\, b) \cdot (1 + \varepsilon_i) \text{ mit } |\varepsilon_i| \leq \varepsilon_{\mathrm{M}}$$

- $\varepsilon_i \cdot \varepsilon_j \doteq 0$

Mit Rundungsfehlern $\varepsilon_1,\ \varepsilon_2,\ \varepsilon_3$ stellen sich die gerundeten $f$-Auswertungen wie folgt dar:

$$\begin{aligned}
\mathrm{rd}(f_1)(x) &= a \cdot x \cdot (1+\varepsilon_1) \\
&= \boxed{f_1(x) + f_1(x)\cdot\varepsilon_1} \\
\mathrm{rd}(f_2)(x) &= \frac{(a-x)\cdot(1+\varepsilon_1)}{b}\cdot(1+\varepsilon_2) \doteq \frac{(a-x)}{b}\cdot(1+\varepsilon_1+\varepsilon_2) \\
&= \boxed{f_2(x) + f_2(x)\cdot(\varepsilon_1+\varepsilon_2)} \\
\mathrm{rd}(f_3)(x) &= (3e^x(1+\varepsilon_1)(1+\varepsilon_2)-3)(1+\varepsilon_3) \\
&\doteq (3e^x(1+\varepsilon_1+\varepsilon_2)-3)(1+\varepsilon_3) \\
&= (3e^x-3)(1+\varepsilon_3) + 3e^x(\varepsilon_1+\varepsilon_2)(1+\varepsilon_3) \\
&\doteq \boxed{f_3(x)(1+\varepsilon_3) + 3e^x(\varepsilon_1+\varepsilon_2)}.
\end{aligned}$$

Damit ergibt sich für den die Stabilität beschreibenden relativen Fehler:

$$\begin{aligned}
\left|\frac{\mathrm{rd}(f_1)(x)-f_1(x)}{f_1(x)}\right| &= \left|\frac{f_1(x)+f_1(x)\cdot\varepsilon_1 - f_1(x)}{f_1(x)}\right| = |\varepsilon_1| \\
\left|\frac{\mathrm{rd}(f_2)(x)-f_2(x)}{f_2(x)}\right| &= \left|\frac{f_2(x)+f_2(x)\cdot(\varepsilon_1+\varepsilon_2) - f_2(x)}{f_2(x)}\right| = |\varepsilon_1+\varepsilon_2| \\
\left|\frac{\mathrm{rd}(f_3)(x)-f_3(x)}{f_3(x)}\right| &= \left|\frac{f_3(x)\cdot\varepsilon_3 + 3e^x(\varepsilon_1+\varepsilon_2)}{f_3(x)}\right| \ = \ \left|\varepsilon_3 + (\varepsilon_1+\varepsilon_2)\frac{3e^x}{3e^x-3}\right| \\
&\overset{\Delta\text{-Ugl.}}{\leq} |\varepsilon_3| + \left|(\varepsilon_1+\varepsilon_2)\frac{e^x}{e^x-1}\right| \leq \varepsilon_M + 2\varepsilon_M\left|\frac{e^x}{e^x-1}\right| \\
&= \varepsilon_M\left(1 + 2\cdot\left|\frac{e^x}{e^x-1}\right|\right) \to \infty \text{ für } x \to 0.
\end{aligned}$$

Die Auswertungen der Funktionen $f_1$ und $f_2$ sind somit offensichtlich stabil. Der Algorithmus zur Auswertung von $f_3(x)$ ist instabil für alle Werte $x \approx 0$, ansonsten stabil. **Achtung:** Für zunehmendes $x$ haben wir bereits in Teilaufgabe 1 a) 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!

## 4) Zusatzaufgabe: Klausuraufgabe SoSe 2014

Gegeben ist die Funktion $f(x) = \ln(x+1)$, definitert für $x > -1$. Untersuchen Sie die Konditionszahl und die Stabilität von $f$ für:

a) $x \to -1$

b) $x \to 0$

Die Auswertung von $\ln(x)$ erzeuge auch nur einen relativen Fehler $\leq \varepsilon_{\text{Ma}}$. Die Eingabe $x$ ist schon gerundet ($\text{rd}(x) = x$).

**Lösung:**

a) Kondition

$$f'(x) = \frac{1}{x+1}, \quad \text{cond}_f(x) = \left|\frac{\frac{x}{x+1}}{\ln(x+1)}\right|$$

i)

$$\text{cond}_f(-1) = \lim_{x\to-1}\left|\frac{\frac{x}{x+1}}{\ln(x+1)}\right| = \frac{-\infty}{-\infty}(\text{l'Hospital}) = \lim_{x\to-1}\left|\frac{\frac{1}{(x+1)^2}}{\frac{1}{x+1}}\right|$$

$$= \lim_{x\to-1}\left|\frac{1}{x+1}\right| = \infty$$

schlecht konditioniert.

ii)

$$\text{cond}_f(0) = \lim_{x\to0}\left|\frac{\frac{x}{x+1}}{\ln(x+1)}\right| = \frac{0}{0}(\text{l'Hospital}) = \lim_{x\to0}\left|\frac{\frac{1}{(x+1)^2}}{\frac{1}{x+1}}\right|$$

$$= \lim_{x\to0}\left|\frac{1}{x+1}\right| = 1$$

gut konditioniert.

b) Stabilität

$$\text{rd}(f)(x) = (1+\varepsilon_2)\ln\left((1+\varepsilon_1)(x+1)\right)$$

$$= (1+\varepsilon_2)\ln(1+\varepsilon_1) + (1+\varepsilon_2)\ln(x+1)$$

$$\left|\frac{\text{rd}(f)(x) - f(x)}{f(x)}\right| = \left|\frac{(1+\varepsilon_2)\ln(1+\varepsilon_1)}{\ln(x+1)} + \varepsilon_2\right|$$

i) $x \to -1$

$$\lim_{x\to-1}\left|\frac{(1+\varepsilon_2)\ln(1+\varepsilon_1)}{\ln(x+1)} + \varepsilon_2\right| = |\varepsilon_2| \leq \varepsilon_{Ma}$$

stabil, aber laut a) schlecht konditioniert.

ii) $x \to 0$

$$\lim_{x\to0}\left|\frac{(1+\varepsilon_2)\ln(1+\varepsilon_1)}{\ln(x+1)} + \varepsilon_2\right| = \infty$$

instabil.