Technische Universität München
Institut für Informatik
Prof. Dr. Thomas Huckle
Michael Rippl
Fabio Gratl

SoSe 2018

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 $cond(f, x)$ für reellwertige Funktionen f ist definiert als:

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

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

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

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

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

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

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

Insgesamt ergibt sich für die interessanten Fälle:

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

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) an der gegebenen Stelle m.

c) 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

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

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) 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 große 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 rd_G immer ein relativer Fehler $\leq \varepsilon_{Ma}$ entsteht:

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

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

$$\mathrm{rd}_G(x) = (1 + \varepsilon)x, \tag{2}$$

$$-\varepsilon_{Ma} \leq \varepsilon \leq \varepsilon_{Ma} \tag{3}$$

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 op_M wird einen neuen relativen Fehler ε_i erzeugt:
$$(a \,\mathrm{op}_M\, b) = \mathrm{rd}_M(a \,\mathrm{op}\, b) = (a \,\mathrm{op}\, b) \cdot (1 + \varepsilon_i) \text{ mit } |\varepsilon_i| \leq \varepsilon_{Ma}.$$
- Terme höheren Ordnung werden vernachlässigt: $\varepsilon_i \cdot \varepsilon_j \doteq 0 \quad \forall i, j.$

Untersuchen Sie die Stabilität von den Funktionen

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

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.
$$(a \,\mathrm{op}_M\, b) = (a \,\mathrm{op}\, b) \cdot (1 + \varepsilon_i) \text{ mit } |\varepsilon_i| \leq \varepsilon_M$$
- $\varepsilon_i \cdot \varepsilon_j \doteq 0$

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:

$$\left|\frac{\mathrm{rd}(f_1)(x) - f_1(x)}{f_1(x)}\right| = \left|\frac{f_1(x) + f_1(x)\cdot\varepsilon_1 - f_1(x)}{f_1(x)}\right| = |\varepsilon_1|$$

$$\left|\frac{\mathrm{rd}(f_2)(x) - f_2(x)}{f_2(x)}\right| = \left|\frac{f_2(x) + f_2(x)\cdot(\varepsilon_1+\varepsilon_2) - f_2(x)}{f_2(x)}\right| = |\varepsilon_1+\varepsilon_2|$$

$$\left|\frac{\mathrm{rd}(f_3)(x) - f_3(x)}{f_3(x)}\right| = \left|\frac{f_3(x)\cdot\varepsilon_3 + 3e^x(\varepsilon_1+\varepsilon_2)}{f_3(x)}\right| = \left|\varepsilon_3 + (\varepsilon_1+\varepsilon_2)\frac{3e^x}{3e^x-3}\right|$$

$$= \left|\varepsilon_3 + (\varepsilon_1+\varepsilon_2)\frac{e^x}{e^x-1}\right| \to \infty \text{ für } x \to 0.$$

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: Ermittlung von π nach Archimedes

Viele mathematische und naturwissenschaftliche Probleme können nicht oder nur schwer durch Angabe einer direkten Lösungsformel gelöst werden, stattdessen ist aber oftmals eine schrittweise Annäherung an die exakte Lösung möglich. Solche sogenannten iterativen Approximationen lassen sich meist algorithmisch beschreiben und in ein Computer-Programm umsetzen. Dabei ist allerdings große Sorgfalt geboten, wie die folgende Aufgabe zeigt.

Ein klassisches Beispiel für die iterative Approximation ist die Bestimmung der Kreiszahl π. Eines der ersten bekannten Verfahren geht auf Archimedes von Syrakus (um 250 v.Chr.) zurück. Die Formel Kreisumfang gleich zweimal Kreisradius mal π war Archimedes bereits bekannt. Durch immer genauere Approximation des Umfangs eines Kreises mit Radius eins mit Hilfe von in den Kreis einbeschriebenen Polygonen konnte er somit π approximieren. Für ihn war das eine mühselige Arbeit mit Tinte und Papyrus, wir können das heute dank Computer im Bruchteil einer Sekunde erledigen.

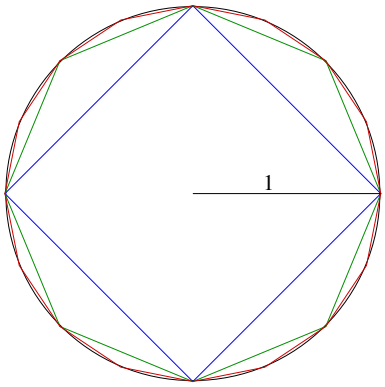


Abbildung 1: Einheitskreis und einbeschriebene Polygone

Die von Archimedes entwickelte Formel (siehe letzte Zusatzaufgabe) sah möglicherweise folgendermaßen aus:

$$s_1 = 2, \qquad s_{n+1} = \sqrt{2 - \sqrt{4 - s_n^2}}, \tag{4}$$

mit s_n der Seitenlänge des 2^n-Ecks. Der gesamte Umfang des 2^n-Ecks ergibt sich somit zu

$$U_n = 2^n \cdot s_n \approx 2\pi$$

und damit also

$$\pi_n = U_n/2 = 2^{n-1} \cdot s_n.$$

Zum Beispiel $\pi_1 = 2,\ \pi_2 = 2\sqrt{2}$ usw.

Aufgabe Mit Formel (4) für die Seitenlänge der Polygone lassen sich nun schrittweise Näherungen für π berechnen. Berechnen Sie den Fehler $\|\pi_n - \pi\|$ bis $n = 30$. Was beobachten Sie? Beheben Sie das Problem!

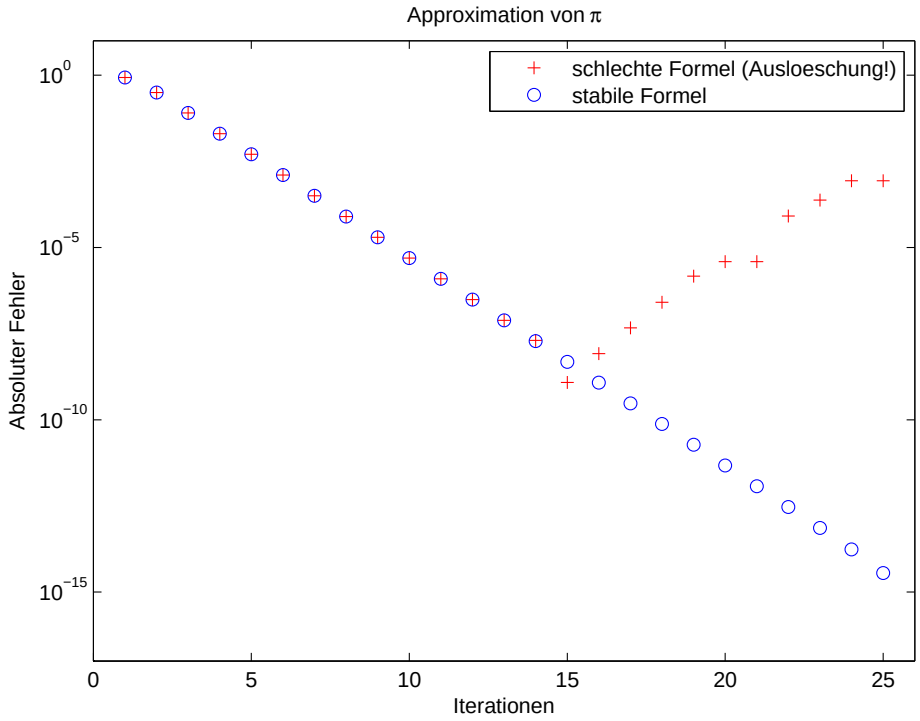
Lösung: Vermeintlich höhere Genauigkeit durch mehr Rekursionen verbessert das Ergebnis nicht sondern zerstört den gesamten Wert!
Problem: Auslöschung!

Algebraische Umformung zur Vermeidung der Auslöschung:

$$\begin{aligned} s_{n+1} &= \sqrt{2 - \sqrt{4 - s_n^2}} = \sqrt{2 - \sqrt{4 - s_n^2}} \cdot \frac{\sqrt{2 + \sqrt{4 - s_n^2}}}{\sqrt{2 + \sqrt{4 - s_n^2}}} \\ &= \frac{|s_n|}{\sqrt{2 + \sqrt{4 - s_n^2}}} \end{aligned}$$

Vergleich der absoluten Fehler der beiden Formeln in Bezug auf echte Lösung π in semilogarithmischer Skala (erstellt mit matlab-Programm archimedes.m aus www):

Falls Sie in **float**-Genauigkeit, statt **double**-Genauigkeit rechnen, tritt das Problem früher auf!



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:

i) $x \to -1$

ii) $x \to 0$

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

Lösung:

a) Kondition

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

i)

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

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

schlecht konditioniert.

ii)

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

$$= \lim_{x\to 0}\left|\frac{1}{x+1}\right| = 1$$

gut konditioniert.

b) Stabilität

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

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

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

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) $x \to 0$

$$\lim_{x\to 0}\left|\frac{(1+\varepsilon_2)\ln(1+\varepsilon_1)}{\ln(x+1)} + \varepsilon_2\right| = \infty$$

instabil.