

Numerische Verfahren und Grundlagen der Analysis

Rasa Steuding

Hochschule RheinMain Wiesbaden

Wintersemester 2011/12

Fehlerbetrachtung

Definitionen

Die Lösung einer Aufgabe der praktischen Mathematik lässt sich oft auf die Bestimmung einer oder mehrerer reeller Zahlen zurückführen. Dabei vereinbart man:

- Eine reelle Zahl **numerisch berechnen** heißt, ein Paar reeller Zahlen $(\tilde{r}, \varepsilon)$ anzugeben mit $|r - \tilde{r}| \leq \varepsilon$.
- Der **Näherungswert** $\tilde{r}$ heißt auf mindestens k Dezimalstellen genau, wenn für die **Fehlerschranke** ε gilt

$$\varepsilon \leq \frac{1}{2} \cdot 10^{-k}.$$

- Die reelle Zahl $|r - \tilde{r}|$ heißt **absoluter Fehler** von $\tilde{r}$.
- Für $r \neq 0$ heißt

$$\left| \frac{r - \tilde{r}}{r} \right|$$

relativer Fehler von $\tilde{r}$.

In der Regel gestattet es ein Verfahren nicht, die gesuchte Größe in endlich vielen Schritten exakt zu gewinnen. Es entsteht ein **Verfahrensfehler**. Häufig kann der Verfahrensfehler durch Erhöhung der Anzahl der Rechenoperationen beliebig klein gehalten werden.

- **Abbruchfehler:**

$$\text{Zu } e = \sum_{n=0}^{\infty} \frac{1}{n!} \text{ sei } \tilde{e} = \sum_{n=0}^n \frac{1}{n!}.$$

Hier gilt

$$|e - \tilde{e}| \leq \frac{\max_{x \in [0,1]} \exp^{(n+1)}(x)}{(n+1)!} (1-0)^{n+1} \leq \frac{3}{(n+1)!}.$$

- **Diskretisierungsfehler:** Z_u

$$I = \int_2^4 \frac{10}{5x-1} = \log \frac{361}{81}$$

sei

$$\tilde{I} = T_n = \frac{4-2}{n} \left(\frac{1}{2}y_0 + \sum_{j=1}^{n-1} y_j + \frac{1}{2}y_n \right)$$

mit $y_j = \frac{10}{5x_j-1}$ und $x_j = 2 + \frac{4-2}{n}j$, $0 \leq j \leq n$. Hier gilt

$$|I - \tilde{I}| \leq \frac{(4-2)^3}{n^2} \max_{x \in [2,4]} \left| \left(\frac{10}{5x-1} \right)'' \right| = \frac{8}{n^2} \max_{x \in [2,4]} \left| \frac{500}{(5x-1)^3} \right| \leq \frac{4000}{729n^2}.$$

Fehlerquellen. III

Nach Vorgabe einer **Basis** $g \in \mathbb{N}$, $g \neq 1$, kann jede reelle Zahl $r > 0$ in der Form

$$r = \sum_{k=-\infty}^n a_k g^k, \quad a_k \in \mathbb{N}_0, \quad 0 \leq a_k \leq g - 1$$

mit **Ziffern** a_k dargestellt werden. Die Zahl r wird also durch eine *unendliche* Folge (a_k) angegeben. Auf Rechenmaschinen kann man aber nur mit endlichen Folgen arbeiten, und muss sich daher mit einem rationalem Näherungswert, etwa

$$\tilde{r} = \sum_{k=-t}^n a_k g^k,$$

begnügen. Den so entstehenden, durch das Rechengesetz bedingten Fehler bezeichnet man als **Rundungsfehler**.

Die Darstellung des rationalen Näherungswertes $\tilde{r}$ von r in einer Maschine kann auf verschiedene Weise normiert werden:

Festkomma-Darstellung

Jede Zahl wird nach Notierung ihres Vorzeichens durch $s + t$ Ziffern, nämlich s Ziffern vor dem Komma und t Ziffern danach, ersetzt.

Beispiel: Für $g = 10$, $s = 4$ und $t = 6$ gilt

$$30.742 \rightarrow 0030.742000$$

$$0.0437 \rightarrow 0000.043700$$

Bei der *Addition* und *Subtraktion* entsteht kein Fehler, solange der Bereich der darstellbaren Zahlen nicht überschritten wird.

Bei der *Multiplikation* entsteht zunächst eine Zahl mit $2t$ Stellen hinter dem Komma. Beim *Runden* wird der Betrag des Ergebnisses in der Regel um $\frac{1}{2}g^{-t}$ erhöht; dann wird hinter den ersten t Nachkommastellen abgeschnitten. Dabei entsteht ein *absoluter Fehler* $\varepsilon \leq \frac{1}{2}g^{-t}$. Darüber hinaus kann durch zu viele Stellen vor dem Komma eine

Bereichsüberschreitung auftreten.

Analog kann man die gleichen Ergebnisse für die *Division* erhalten.

Gleitkomma-Darstellung. I

Jede reelle Zahl $r \neq 0$ wird durch ein Produkt

$$\tilde{r} := gl(r) := a \cdot g^b \text{ mit } \frac{1}{g} \leq |a| < 1$$

ersetzt. Die **Mantisse** a bzw. der **Exponent** b wird durch eine feste Anzahl t' bzw. s' von Ziffern und je ein Vorzeichen dargestellt.

Beispiel: $30.742 \rightarrow 0.30742 \cdot 10^2$, $0.0437 \rightarrow 0.437 \cdot 10^{-1}$.

Für den dabei auftretenden Rundungsfehler gilt jedenfalls

$$|gl(r) - r| \leq \frac{1}{2} \cdot g^{b-t'} \text{ und } \frac{|gl(r) - r|}{|r|} \leq \frac{1}{2} \cdot g^{1-t'} := \varepsilon_M.$$

Daher ist

$$gl(r) = r(1 + \varepsilon) \text{ mit } |\varepsilon| \leq \varepsilon_M.$$

Die Zahl ε_M heißt **Maschinengenauigkeit**.

Gleitkomma-Darstellung. II

Es seien $r_1 = a_1 \cdot g^{b_1}$, $r_2 = a_2 \cdot g^{b_2}$ in Gleitkomma-Darstellung gegeben mit $b_2 \leq b_1$, $r_1 \neq r_2$. Bei der *Addition* und *Subtraktion* entsteht zunächst

$$a_3^* \cdot g^{b_3^*} := r_1 \pm r_2 = (a_1 \pm a_2 \cdot g^{b_2-b_1}) \cdot g^{b_1} \text{ mit } \frac{1}{q} \leq |a_3^*| < 1$$

und weiter durch Rundung der Mantisse a_3^* auf t' Ziffern

$$gl(r_1 \pm r_2) = a_3 \cdot g^{b_3},$$

wobei b_3 so festgelegt wird, dass $\frac{1}{q} \leq |a_3| < 1$ ist. Daraus folgt $b_1 - t' \leq b_3 \leq b_1 + 1$. Für den dabei entstehenden *absoluten Fehler* gilt:

$$|r_1 \pm r_2 - gl(r_1 \pm r_2)| \leq \frac{1}{2} \cdot g^{b_3-t'}.$$

Wegen $g^{b_3-1} \leq |r_1 \pm r_2| \leq g^{b_3}$ folgt für den *relativen Fehler* ρ

$$\rho \leq \frac{1}{2} \cdot g^{1-t'}.$$

Gleitkomma-Darstellung. III

Bei der *Multiplikation* bekommt man zunächst

$$r_1 \cdot r_2 = a_1 \cdot a_2 \cdot g^{b_1+b_2}.$$

Dieses Ergebnis wird dann normiert und gerundet zu

$$gl(r_1 \cdot r_2) = a_4 \cdot g^{b_4};$$

der Exponent b_4 bleibt dabei in der Grenzen

$$b_1 + b_2 - 1 \leq b_4 \leq b_1 + b_2.$$

Dadurch entsteht ein *relativer Fehler* ρ mit

$$\rho \leq \frac{1}{2} \cdot g^{1-t'}.$$

Analoge Überlegungen für die *Division* führen zum selben Ergebnis.

In allen Fällen kann im Ergebnis ein Exponent b entstehen, der nicht durch nur s' Ziffern darstellbar ist. Bei negativem b wird dann das Ergebnis durch Null ersetzt, bei positivem b scheitert die Rechnung (*Bereichsüberschreitung*).

Beispiele für $g = 10$ und $t' = 5$

- Für $r_1 = 0.13412 \cdot 10^{-5}$ und $r_2 = -0.13410 \cdot 10^{-5}$ gilt

$$gl(r_1 + r_2) = gl(0.00002 \cdot 10^{-5}) = 0.2 \cdot 10^{-9} = r_1 + r_2.$$

- Für $r_1 = 0.86391 \cdot 10^4$ und $r_2 = 0.11411 \cdot 10^{-6}$ gilt

$$gl(r_1 \pm r_2) = gl((0.86391 \pm 0.11411 \cdot 10^{-10}) \cdot 10^4) = r_1.$$

- Für $r_1 = 0.37218 \cdot 10^4$ und $r_2 = 0.71422 \cdot 10^{-1}$ gilt

$$r_1 = 0.3721800000 \cdot 10^4$$

$$r_2 = 0.0000071422 \cdot 10^4$$

$$r_1 + r_2 = 0.3721871422 \cdot 10^4$$

und $gl(r_1 + r_2) = 0.37219 \cdot 10^4$. Der relative Fehler ist

$$\frac{|r_1 + r_2 - gl(r_1 + r_2)|}{|r_1 + r_2|} = \frac{0.0000028578 \cdot 10^4}{0.3721871422 \cdot 10^4} \approx 0.77 \cdot 10^{-5}.$$

Fehlerfortpflanzung. I

Sind $\tilde{r}_1$ und $\tilde{r}_2$ Näherungen an r_1 und r_2 mit $|r_1 - \tilde{r}_1| \leq \varepsilon_1$ und $|r_2 - \tilde{r}_2| \leq \varepsilon_2$.

Bei der Addition und Subtraktion addieren sich die Schranken für den absoluten Fehler, denn es gilt

$$|(r_1 \pm r_2) - (\tilde{r}_1 \pm \tilde{r}_2)| \leq |r_1 - \tilde{r}_1| + |r_2 - \tilde{r}_2| \leq \varepsilon_1 + \varepsilon_2.$$

Relative Fehler können dagegen erheblich verstärkt werden. Zum Beispiel: Für $r_1 = 0.13412 \cdot 10^{-5}$ und $r_2 = -0.13410 \cdot 10^{-5}$ gilt

$$gl(r_1 + r_2) = gl(0.00002 \cdot 10^{-5}) = 0.2 \cdot 10^{-9} = r_1 + r_2.$$

Besitzt etwa r_1 einen relativen Fehler ρ (> 0), so wird auf das Ergebnis $r_1 + r_2$ der relative Fehler

$$\frac{\rho \cdot r_1}{r_1 + r_2} > \rho \cdot 10^3$$

übertragen. Dieser bei der Subtraktion nahezu gleich großer Zahlen immer auftretende Effekt heißt **Auslöschung**.

Fehlerfortpflanzung. II

Bei der Multiplikation und Division addieren sich die Schranken für den relativen Fehler, denn es gilt

$$\begin{aligned} |r_1 \cdot r_2 - \tilde{r}_1 \cdot \tilde{r}_2| &\leq |r_1 \cdot r_2 - r_1 \cdot \tilde{r}_2| + |r_1 \cdot \tilde{r}_2 - \tilde{r}_1 \cdot \tilde{r}_2| \\ &\leq |r_1| \cdot \varepsilon_2 + |\tilde{r}_2| \cdot \varepsilon_1 \leq |r_1 \cdot r_2| \left(\frac{\varepsilon_1}{|r_1|} + \frac{\varepsilon_2}{|r_2|} \right) + \varepsilon_1 \cdot \varepsilon_2 \end{aligned}$$

sowie

$$\begin{aligned} |r_1 \cdot \tilde{r}_2 - \tilde{r}_1 \cdot r_2| &\leq |r_1 \cdot \tilde{r}_2 - \tilde{r}_1 \cdot \tilde{r}_2| + |\tilde{r}_1 \cdot \tilde{r}_2 - \tilde{r}_1 \cdot r_2| \\ &\leq \varepsilon_1 \cdot |\tilde{r}_2| + \varepsilon_2 \cdot |r_1| + \varepsilon_1 \cdot \varepsilon_2, \end{aligned}$$

und damit

$$\left| \frac{r_1}{r_2} - \frac{\tilde{r}_1}{\tilde{r}_2} \right| \leq \left| \frac{r_1}{r_2} \right| \left(\frac{\varepsilon_1}{|r_1|} + \frac{\varepsilon_2}{|r_2|} \right) \left(1 + \frac{\varepsilon_2}{|\tilde{r}_2|} \right).$$

Die Produkte $\varepsilon_1 \varepsilon_2$ und ε_2^2 können in der Regel vernachlässigt werden.

Absoluter Fehler bei Funktionsauswertung

Ist $f : [a, b] \rightarrow \mathbb{R}$ differenzierbar und $r, \tilde{r} \in [a, b]$ mit $|r - \tilde{r}| \leq \varepsilon$, so gilt nach dem Satz von Taylor:

$$f(r) - f(\tilde{r}) = f'(\xi) \cdot (r - \tilde{r})$$

für ein geeignetes $\xi \in [a, b]$. Daraus folgt

$$|f(r) - f(\tilde{r})| \leq M \cdot |r - \tilde{r}|$$

mit $M := \max_{x \in [a, b]} f'(x)$.

Beispiel: Sei $f(x) = \sqrt{x}$. Für $r = 0.01$ und die Näherung $\tilde{r} = 0.011$ erhält man einen absoluten Fehler $|r - \tilde{r}| = 0.001$. Für den absoluten Fehler bei Funktionsauswertung gilt

$$|f(r) - f(\tilde{r})| \leq M \cdot |r - \tilde{r}| \text{ mit } M = \max_{x \in [r, \tilde{r}]} |f'(x)| = \max_{x \in [0.01, 0.011]} \frac{1}{2\sqrt{x}} = 5.$$

Relativer Fehler bei Funktionsauswertung. I

Aus

$$|f(r) - f(\tilde{r})| \leq M \cdot |r - \tilde{r}|$$

mit $M := \max_{x \in [a,b]} |f'(x)|$ folgt

$$\frac{|f(r) - f(\tilde{r})|}{|f(r)|} \leq \frac{M \cdot |r - \tilde{r}|}{|f(r)|} = \frac{M \cdot |r|}{|f(r)|} \cdot \frac{|r - \tilde{r}|}{|r|}, \quad (1)$$

wobei

$$\frac{M \cdot |r|}{|f(r)|} \leq \frac{M \cdot \max_{x \in [a,b]} |x|}{\min_{x \in [a,b]} |f(x)|}.$$

Relativer Fehler bei Funktionsauswertung. II

Beispiel: Sei $f(x) = \sqrt{x}$. Für $r = 0.01$ und die Näherung $\tilde{r} = 0.011$ erhält man einen absoluten Fehler $|r - \tilde{r}| = 0.001$. Für den relativen Fehler bei Funktionsauswertung gilt

$$\frac{|f(r) - f(\tilde{r})|}{|f(r)|} \leq K \cdot \frac{|r - \tilde{r}|}{|r|}$$

mit

$$\begin{aligned} K &\leq \frac{\max_{x \in [r, \tilde{r}]} |f'(x)| \max_{x \in [r, \tilde{r}]} |x|}{\min_{x \in [r, \tilde{r}]} |f(x)|} \\ &= \frac{\max_{x \in [0.01, 0.011]} \frac{1}{2\sqrt{x}} \max_{x \in [0.01, 0.011]} x}{\min_{x \in [0.01, 0.011]} \sqrt{x}} \\ &= \frac{5 \cdot 0.011}{0.1} = 0.55. \end{aligned}$$

Kondition. I

Der Verstärkungsfaktor zwischen den relativen Rundungsfehlern in (1) ist für die numerische Mathematik von zentraler Bedeutung.

Die **Kondition** der Funktion $y = f(x)$

$$\kappa_f(x) := \left| \frac{x \cdot f'(x)}{f(x)} \right|$$

ist eine Maßzahl für die Sensitivität des Resultats y in Bezug auf die Eingabe x .

Wir sprechen von einer bezüglich x **gut konditionierten** Aufgabenstellung $f(x)$, wenn kleine relative Eingabefehler ε_x bei exakter Arithmetik, d.h. ohne Einfluss von Rechenfehlern, zu vergleichsweise kleinen relativen Fehlern ε_y im Resultat y führen. In diesem Fall haben die relativen Rundungsfehler ε_x und ε_y die gleiche Größenordnung. Anderenfalls liegt **schlechte Kondition** von f an der Stelle x vor.

Beispiele:

- Die Kondition der Funktion $f(x) = \exp(x)$ lautet

$$\kappa_{\exp} = |x|$$

und f ist für betragsmäßig nicht zu große Werte von x gut konditioniert.

- Die Umkehrfunktion $f(x) = \log(x)$ hat die Kondition

$$\kappa_{\log} = |1/\log(x)|$$

und ist für $x \approx 1$ schlecht konditioniert.

Eine (numerisch) **stabile** Berechnungsmethode vergrößert nicht die relativen Eingabefehler einer gut konditionierten Aufgabenstellung $y = f(x)$. Eine Berechnungsmethode, die trotz kleiner Kondition κ_f , große relative Fehler im Ergebnis produziert, heißt (numerisch) **instabil**.

Beispiel: Der Ausdruck $f(x) := 1 - \sqrt{1 - x^2}$ soll für kleine Werte $x \approx 0$ numerisch berechnet werden. Diese Aufgabenstellung ist wegen

$$\kappa_f = \left| \frac{x^2}{(1 - \sqrt{1 - x^2})\sqrt{1 - x^2}} \right| \rightarrow 2 \text{ mit } x \rightarrow 0$$

gut konditioniert. Die numerische Auswertung des Ausdrucks führt allerdings zu Ergebnissen mit extrem großen relativen Rundungsfehlern. Eine numerisch stabile Berechnungsmethode erhalten wir, wenn wir den Ausdruck durch Erweiterung von Zähler und Nenner

$$1 - \sqrt{1 - x^2} = \frac{(1 - \sqrt{1 - x^2})(1 + \sqrt{1 - x^2})}{(1 + \sqrt{1 - x^2})} = \frac{x^2}{1 + \sqrt{1 - x^2}}$$

umformen und die Auslöschung so umgehen.