FEHLERRECHNUNG

Prof. Dr. Jörg Hertling, Techn. Univ. Wien

I. EINFÜHRUNG

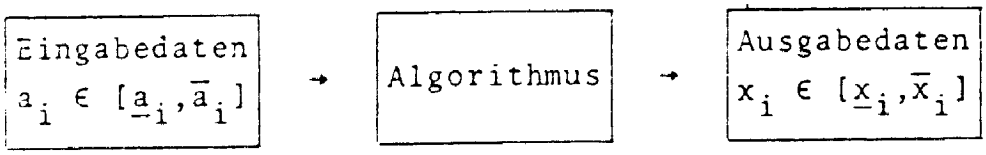
Die zunehmende Verwendung von elektronischen Rechnern birgt die Gefahr, numerischen Ergebnissen kritiklos zu vertrauen. Offensichtlich ist aber ein numerisches Ergebnis ohne eine Vorstellung über dessen Genauigkeit sinnlos.

Die Fehlerquellen können in drei Klassen aufgeteilt werden

A) Datenfehler.
B) Rundungsfehler.
C) Verfahrensfehler.

Im allgemeinen wird ein numerisches Ergebnis von all diesen Fehlerquellen beeinflußt.

ad A) In realistischen Problemen haben Eingabedaten a_i im allgemeinen auf Grund ihrer Herkunft (etwa durch Messung) nur eine beschränkte (absolute oder relative) Genauigkeit, man spricht vom *"Datenfehler"*. Tatsächlich wäre es vernünftiger, davon auszugehen, daß die Eingabedaten a_i jeweils in einem Intervall $[\underline{a}_i, \overline{a}_i]$ liegen. Naturgemäß kann man dann nach der Rechnung auch von den Ausgabedaten x_i nur erwarten, daß sie in einem Intervall $[\underline{x}_i, \overline{x}_i]$ liegen. Diesen Effekt bezeichnet man als *"Datenfehlereffekt"*.



Obwohl es Maschinen gibt, die tatsächlich in jedem Schritt mit Intervallen rechnen, ist dies i.a. zu aufwendig.

Grundsätzlich muß man aber zwei Klassen von Problemen unterscheiden:

1) Probleme, bei denen eine geringe Änderung der Eingabedaten auch eine geringe Änderung des Ergebnisses bewirkt, diese Probleme nennt man "*gut konditioniert*" oder "*stabil*".

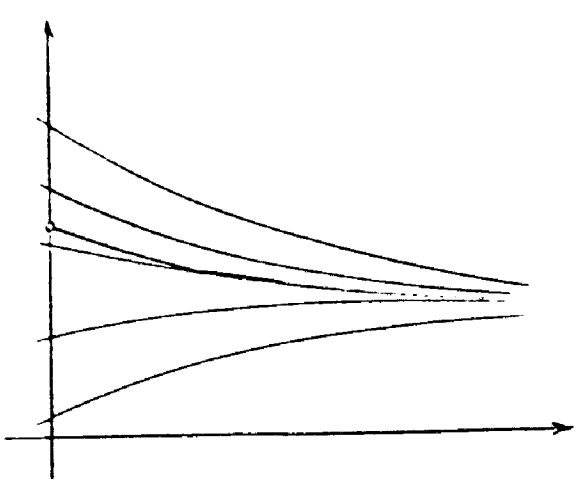
2) Probleme, bei denen eine geringe Änderung von Eingabedaten sehr große Änderungen des Ergebnisses bewirken, diese Probleme nennt man "*schlecht konditioniert*" oder "*instabil*".

Allgemein nennt man die Empfindlichkeit eines Ergebnisses bezüglich der Änderung eines Datenelements "*Kondition*".

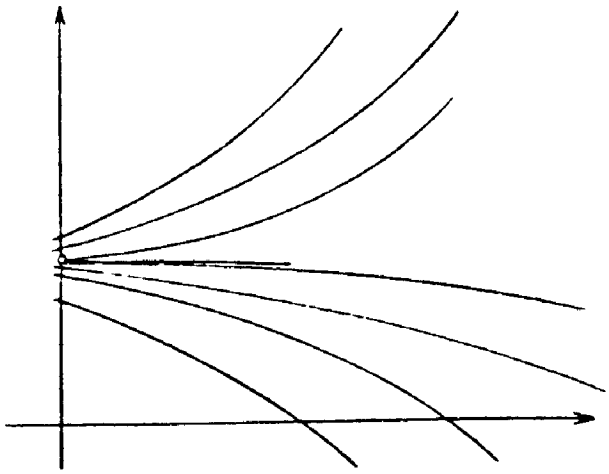
Dafür sollen nun einige Beispiele gegeben werden.

Bsp.: Stellt man ein Gleichungssystem von zwei Gleichungen in zwei Unbekannten graphisch dar, so ist das Problem schlecht konditioniert, wenn die beiden Geraden einen "schleifenden" Schnitt aufweisen; eine leichte Änderung des Anstiegs einer der beiden Geraden kann dann nämlich eine sehr große Verschiebung des Schnittpunktes (d.h. der Lösung) bewirken. Ist der Schnitt nicht schleifend, so ist das Problem gut konditioniert.

Bsp.: Löst man ein Anfangswertproblem für eine gewöhnliche Differentialgleichung erster Ordnung, numerisch (etwa mit dem Eulerverfahren), so ist dieses Problem bei einem "zusammenlaufenden" Richtungsfeld gut konditioniert; bei einem "auseinanderlaufenden" Richtungsfeld ist das Problem schlecht konditioniert, weil man dann schon nach wenigen Schritten des Verfahrens auf eine andere Lösungskurve gerät als jene, die dem Anfangswert entspricht.



gut konditioniert



schlecht konditioniert

Bsp.: Die Ritzsche Achsenkonstruktion ist verschieden konditioniert, je nachdem, ob die konjugierten Durchmesser schon nahezu orthogonal sind, oder nicht.

ad B) Der Algorithmus kann durchgeführt werden auf

a) Analogrechnern:
Dabei werden mathematische Größen und Algorithmen durch "kontinuierliche" physikalische Größen (z.B. Längen oder Spannungen) und physikalische Vorgänge simuliert. Beispiele dafür sind etwa der Rechenschieber und das Planimeter. Auf Analogrechnern ist die Genauigkeit durch die physikalische Meßgenauigkeit begrenzt.

b) Digitalrechnern:
Dabei werden mathematische Größen und Algorithmen durch "diskrete" physikalische Zustände und Verknüpfungen simuliert, die den arithmetischen Verknüpfungen entsprechen. Beispiele dafür sind der Abacus und die üblichen elektronischen Taschenrechner. Auf Digitalrechnern ist die Genauigkeit durch die Zahldarstellung begrnzt. Es treten daher *Rundungsfehler* auf. Natürlich können sich bei schlecht konditionierten Problemen auch die Rundungsfehler verheerend auswirken.

Unsere weiteren Betrachtungen gelten ausschließlich Digitalrechnern.

ad C) Selbst bei rundungsfehlerfreier Rechnung liefern viele Algorithmen nach endlich vielen Schritten die "Lösung" zwar beliebig genau, aber gewöhnlich nicht exakt, man spricht vom *Abbrechfehler*. Es sind keine infinitesimalen Operationen und Grenzprozesse möglich. Differentialquotienten müssen i.a. durch Differenzenquotienten ersetzt werden. Den hierbei auftretenden Fehler bezeichnet man als *Diskretisierungsfehler*. Fehler treten auf, wenn man, wie etwa bei der Interpolation oder Approximation eine Funktion durch eine "Ersatzfunktion" aus einer bestimmten Funktionenklasse ersetzt. Ähnliches gilt für die numerische Integration (etwa bei der Trapez- oder Simpsonregel). Solche Fehler können oft durch sogenannte "Glattheitsvoraussetzungen" abgeschätzt werden.

Naturgemäß kann über Verfahrensfehler wenig Allgemeines gesagt werden, sie müssen bei jedem Algorithmus für sich untersucht werden.

Unsere weiteren Betrachtungen gelten daher den Rundungs- und Verfahrensfehlern. Dazu definieren wir

Def.:

absoluter Fehler := berechneter Wert - idealer Wert

relativer Fehler := $\frac{\text{absoluter Fehler}}{\text{idealer Wert}}$

Entscheidend ist natürlich stets der relative Fehler.

II. ZAHLDARSTELLUNG AUF DIGITALRECHNERN, RUNDUNGSFEHLERANALYSE

Meist wird auf Digitalrechnern eine Zahldarstellung im Dualsystem verwendet:

$$x = \pm\left(\alpha_n 2^n + \alpha_{n-1} 2^{n-1} + \ldots + \alpha_0 2^0 + \alpha_{-1} 2^{-1} + \alpha_{-2} 2^{-2} + \ldots\right)$$

$$\alpha_i = 0 \text{ oder } 1 \ .$$

Die Anzahl der Dualstellen (bzw. Dezimalstellen) ist festgelegt durch die sogenannte *Wortlänge*.

Bei der sogenannten *Festpunktdarstellung* (oder *Festkommadarstellung*) ist eine Anzahl von Stellen vor dem Komma fest reserviert und eine Anzahl von Stellen nach dem Komma. Solche Maschinen finden bei kommerziellen Rechnungen Verwendung und sind für wissenschaftliche Rechnungen ungeeignet.

Bei der sogenannten *Gleitpunktdarstellung* (oder *Gleitkommadarstellung*) wird die Lage des Kommas durch einen Exponenten angegeben, d.h.

$$x = a \cdot 2^b \quad (\text{bzw. } x = a \cdot 10^b)$$

mit $|a| < 1$, b ganz. a heißt *Mantisse*, b heißt *Exponent* und die Voraussetzung $|a| < 1$ bedeutet, daß der Exponent so zu wählen ist, daß die Mantisse erst nach dem Komma Ziffern aufweist, die ungleich Null sind.

Auf Rechenanlagen ist nun eine feste Anzahl t von Stellen für die Mantisse reserviert und eine feste Anzahl e von Stellen für den Exponenten. Zusätzlich sind noch zwei Stellen für die Vorzeichen von Mantisse und Exponent reserviert. Damit ist aber die Darstellung einer Zahl noch nicht eindeutig, denn man kann die Mantisse ja nach dem Komma noch mit einer "beliebigen" Anzahl von Nullen beginnen lassen und diese "Verschie-

bung" der Mantisse durch den Exponenten kompensieren. Man wählt daher die sogenannte *Normalisierte Gleitpunktdarstellung*, d.h. der Exponent wird so gewählt, daß die erste Ziffer der Mantisse von Null verschieden ist. Das bedeutet im Dualsystem $|a| \geq 2^{-1}$, im Dezimalsystem $|a| \geq 10^{-1}$ und allgemein $|a| \geq p^{-1}$, wobei p die Basis des Zahlsystems ist. Eine Zahl in binärer normalisierter Gleitkommadarstellung hat somit folgendes Aussehen:

$$x = \pm\left(\alpha_{-1}2^{-1} + \ldots + \alpha_{-t}2^{-t}\right) \cdot 2^{\pm(\beta_0 2^0 + \ldots + \beta_{e-1}2^{e-1})} ,$$

$$\alpha_{-1} = 1 \qquad \alpha_{-i} = 0 \text{ oder } 1 \text{ für } i = 2,3,\ldots,t$$
$$\beta_i = 0 \text{ oder } 1 \text{ für } i = 0,1,\ldots,e-1$$

Offensichtlich kann die Zahl 0 nicht in normalisierter Gleitkommadarstellung dargestellt werden und besitzt daher eine Sonderstellung.

Aus alldem ergeben sich zunächst folgende Folgerungen:

1) Zur Durchführung eines Algorithmus stehen nur endlich viele Zahlen einer bestimmten Bauart zur Verfügung, sogenannte *"Maschinenzahlen"*.

2) Es gibt keine beliebig großen und beliebig kleinen Zahlen und es gibt keine beliebig benachbarten Zahlen.

3) Die Körperaxiome sind verletzt. Summe, Differenz, Produkt und Quotient zweier Maschinenzahlen sind i.a. keine Maschinenzahl; Assoziativgesetze und das Distributivgesetz sind verletzt.

Will man eine Zahl, die nicht im Bereich der Maschinenzahlen liegt, durch eine Maschinenzahl darstellen, muß somit ab der Mantissenstelle $\alpha_{-(t+1)}\ldots$ gerundet oder abgeschnitten werden.

Der zur Verfügung stehende Bereich von *Maschinenzahlen* ist somit durch die Basis p, die Länge der Mantisse t und die Länge des Exponenten e eindeutig festgelegt und wird mit $M(p,t,e)$ bezeichnet. Ist zusätzlich noch festgelegt, ob nach der letzten Stelle der Mantisse abgeschnitten oder gerundet werden soll, so spricht man von einer *Maschinenarithmetik*. Eine Maschinenarithmetik mit runden bezeichnen wir mit $[M(p,t,e),\tilde{rd}]$, eine Maschinenarithmetik mit abschneiden bezeichnen wir mit $[M(p,t,e),\hat{rd}]$.

Berücksichtigt man den Exponenten nicht, so kann der absolute Fehler beim Runden der Mantisse somit abgeschätzt werden durch

$$|rd(a) - a| \leq \frac{1}{2} p^{-t} .$$

Der relative Fehler beim Runden auf Zahlen in <u>normalisierter</u> Gleitkommadarstellung kann (unabhängig vom Exponenten, der sich wegkürzt!) somit abgeschätzt werden durch

$$\left|\frac{rd(x) - x}{x}\right| \leq \frac{1}{2} p^{-t+1} =: eps .$$

"eps" nennt man die *Maschinengenauigkeit*, und sie kann auch als die kleinste positive Maschinenzahl g definiert werden, für die gilt

$$rd(1 + g) > 1 .$$

Wird somit eine Zahl x auf eine Maschinenzahl gerundet, so gilt

$$rd(x) = x(1 + \varepsilon) \quad \text{mit} \quad |\varepsilon| \leq eps ,$$

wobei ε der realtive Fehler ist, der beim Runden entsteht.

Bsp.:

Für $[M(2,27,7),\tilde{rd}]$ ist $eps = \frac{1}{2} \cdot 2^{-26} \approx 7{,}45 \cdot 10^{-9}$

Für $[M(2,48,10),\tilde{rd}]$ ist $eps = \frac{1}{2} \cdot 2^{-47} \approx 3{,}55 \cdot 10^{-15}$

Offensichtlich ist es sinnlos, auf Maschinen mit solchen Maschinengenauigkeiten kleinere relative Fehler zu verlangen. Zwei Beispiele sollen nun binäre Maschinenarithmetiken erläutern.

Bsp.: Welche Zahlen sind in der (unrealistisch einfachen) Maschinenarithmetik $[M(2,2,1),\tilde{rd}]$ exakt darstellbar?

Als Mantissse kommen in normalisierter Gleitkommadarstellung nur ±0.10 und ±0.11 in Frage, als Exponent nur ±1 oder 0. Symmetrisch zum Nullpunkt sind daher 12 Zahlen exakt darstellbar:

$$\pm 0{,}25, \quad \pm 0.375, \quad \pm 0.5, \quad \pm 0.75, \quad \pm 1, \quad \pm 1.5 \; .$$

Zusätzlich muß noch die Zahl 0 dargestellt werden.

Bsp.: Man untersuche die Maschinenarithmetik $[M(2,t,e),\tilde{rd}]$.

Die größte Maschinenzahl ist

$$(2^{-1} + \ldots + 2^{-t}) \cdot 2^{2^0+\ldots+2^{e-1}} = (1 - 2^{-t})\; 2^{2^e-1}$$

Die zweitgrößte Maschinenzahl ist

$$(2^{-1} + \ldots + 2^{-t+1}) \cdot 2^{2^0+\ldots+2^{e-1}} = (1 - 2^{-t+1})\; 2^{2^e-1}$$

Die kleinste positive Maschinenzahl ist

$$2^{-1} \cdot 2^{-(2^0+\ldots+2^{e-1})} = 2^{-2^e}$$

Die zweitkleinste positive Maschinenzahl ist

$$(2^{-1} + 2^{-t})\ 2^{-(2^0+\ldots+2^{e-1})} = (2^{-1} + 2^{-t})\ 2^{1-2^e}.$$

Die halbe Differenz der beiden größten Maschinenzahlen ergibt (im Bereich der Maschinenzahlen) den größten absoluten Fehler beim Runden, nämlich 2^{2^e-2-t}. Die halbe Differenz der beiden kleinsten Maschinenzahlen ergibt (im Bereich der Maschinenzahlen) den kleinsten absoluten Fehler beim Runden, nämlich 2^{-t-2^e}. Während also der absolute Fehler enorm variiert, gilt für den relativen Fehler (im Bereich der Maschinenzahlen), wie schon besprochen, die obere Schranke eps = 2^{-t}. Schließlich läßt sich auch die Anzahl der Maschinenzahlen (beiderlei Vorzeichens und ohne Null) leicht berechnen; sie ist gegeben durch $2^t(2^{e+1} - 1)$.

Soll eine Zahl berechnet werden, die betragsgrößer als die größte Maschinenzahl ist, so tritt ein sogenannter *Exponentenüberlauf* auf, der i.a. als Fehler gemeldet wird. Soll eine Zahl berechnet werden, die betragskleiner als die betragskleinste Maschinenzahl ist, tritt also ein *Exponentenunterlauf* auf, so wird die Zahl meist gleich Null gesetzt. Der relative Fehler kann dann größer als eps sein, da ja für die Zahl Null keine normalisierte Gleitkommadarstellung existiert!

Für arithmetische Verknüpfungen bzw. Funktionsauswertungen gilt also, wenn man sie als Gleitpunktoperationen betrachtet folgendes:

$$\left.\begin{aligned}
rd(x + y) &=: x +^* y = (x + y)(1 + \varepsilon_1)\\
rd(x - y) &=: x -^* y = (x - y)(1 + \varepsilon_2)\\
rd(x \cdot y) &=: x \cdot^* y = (x \cdot y)(1 + \varepsilon_3)\\
rd(x / y) &=: x /^* y = (x / y)(1 + \varepsilon_4)\\
rd(\sqrt{x}) &=: \sqrt{x}^* = \sqrt{x}(1 + \varepsilon_5)\\
rd(\sin x) &=: \sin x^* = (\sin x)(1 + \varepsilon_6)
\end{aligned}\right\} |\varepsilon_i| \leq \text{eps}$$

u.s.w.

Werden zwei Zahlen mit verschiedenen Exponenten in einer Maschinenarithmetik addiert, so muß der Exponent der kleineren Zahl dem Exponenten der größeren Zahl angeglichen werden.

Bsp.: $[M[10,6,3],\tilde{rd}]$:

$$\begin{array}{r} 0.999727 \cdot 10^{2} \\ + \; 0.854621 \cdot 10^{-1} \\ \hline \end{array} \xrightarrow{\text{Angleichung d. Exponenten}} \begin{array}{r} 0.999727 \cdot 10^{2} \\ + \; 0.000855 \cdot 10^{2} \\ \hline 1.000582 \cdot 10^{2} \end{array}$$

$$\xrightarrow{\text{Normalisieren d. Ergebnisses}} 0.100058 \cdot 10^{3}$$

Man beachte, daß es bei der Subtraktion von nahezu gleich großen Zahlen zu einer Verringerung der gültigen Stellen der Mantisse kommt, man nennt diesen Effekt *Auslöschung*.

Bsp.: $[M[10,7,2],\tilde{rd}]$:

$$\begin{array}{r} 0.3141592 \cdot 10^{4} \\ - \; 0.3141333 \cdot 10^{4} \\ \hline 0.0000259 \cdot 10^{4} \end{array} \xrightarrow{\text{normalisieren}} \underbrace{0.2590000}_{\text{ungültige Stellen}}$$

Man beachte auch, daß die Analyse der Rundungsfehler von der Reihenfolge abhängt, in der die arithmetischen Operationen durchgeführt werden. Diese Reihenfolge kann durch Klammern, durch einen geordneten Baum oder durch Angabe des Algorithmus spezifiziert werden.

Bsp.: $\dfrac{a^2 - \sin b}{c}$

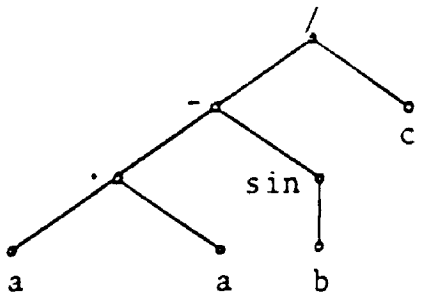
Angabe der Reihenfolge der Operationen durch Klammern:

$$((a \cdot a) - \sin b)/c$$

bzw. in Maschinenarithmetik

((a ·* a) -* sin b*) /* c .

Angabe der Reihenfolge der Operationen durch einen geordneten Baum:



Angabe der Reihenfolge der Operationen durch Angabe des Algorithmus

$$\eta_1 = a \cdot a$$
$$\eta_2 = \sin b$$
$$\eta_3 = \eta_1 - \eta_2$$
$$\eta_4 = \eta_3 / c$$

bzw. in Maschinenarithmetik

$$\tilde{\eta}_1 = a \cdot a(1 + \varepsilon_1)$$
$$\tilde{\eta}_2 = \sin b(1 + \varepsilon_2)$$
$$\tilde{\eta}_3 = (\tilde{\eta}_1 - \tilde{\eta}_2)(1 + \varepsilon_3)$$
$$\tilde{\eta}_4 = (\tilde{\eta}_3 / c)(1 + \varepsilon_4)$$

Entwickelt man in erster Ordnung in den ε_i, so muß das Ergebnis folgende Gestalt haben:

exakte Formel · (1 + rel. Fehler)

wobei der relative Fehler linear in den ε_i sein muß.

Rundungsfehleranalyse

$$\tilde{\eta}_4 = ((a \cdot a(1+\varepsilon_1) - \sin b(1+\varepsilon_2))(1+\varepsilon_3)/c)(1+\varepsilon_4)$$

$$= (a^2 - \sin b)\left(1 + \frac{a^2}{a^2-\sin b}\varepsilon_1 - \frac{\sin b}{a^2-\sin b}\varepsilon_2\right)(1+\varepsilon_3)(1+\varepsilon_4)/c$$

$$\approx \frac{a^2-\sin b}{c}\left(1 + \frac{a^2}{a^2-\sin b}\varepsilon_1 - \frac{\sin b}{a^2-\sin b}\varepsilon_2 + \varepsilon_3 + \varepsilon_4\right)$$

$$|\varepsilon_i| \leq \text{eps}$$

Setzen wir etwa die Arithmetik $[\mathbb{M}(10,6,3), \tilde{rd}]$ voraus, so ist

$$\text{eps} = \frac{1}{2}\,10^{-5} .$$

Der relative Fehler kann somit abgeschätzt werden durch

$$\left|\frac{a^2}{a^2-\sin b}\right|\text{eps} + \left|\frac{\sin b}{a^2-\sin b}\right|\text{eps} + 2\text{eps} .$$

Kritisch ist somit der Fall, wo

$$a^2 \approx \sin b .$$

Man beachte, daß bei einer Funktionsauswertung der entstehende Rundungsfehler unabhängig von der Funktion ist.

Bsp.: Die Zahlen

$a = 0{,}1732051$

$b = 0{,}3141592 \cdot 10^4$

$c = -0{,}3141333 \cdot 10^4$

sollen in der Arithmetik $[\mathbb{M}(10,7,2), \tilde{rd}]$ addiert werden und zwar nach folgenden Algorithmen:

α) $(a +^* b) +^* c$

β) $a +^* (b +^* c)$

α) $0,3141592 \cdot 10^4$
$0,0000173 \cdot 10^4$
$0,3141765 \cdot 10^4$
$-0,3141533 \cdot 10^4$
$0,0000432 \cdot 10^4$

Das normalisierte Ergebnis ist somit
$(a +^* b) +^* c = 0,4320000.$

β) $0,3141592 \cdot 10^4$
$-0,3141533 \cdot 10^4$
$0,0000259 \cdot 10^4$

$0,1732051$
$0,2590000$
$0,4322051$

Das normalisierte Ergebnis ist somit
$a +^* (b +^* c) = 0,4322051.$

Offensichtlich ist das Assoziativitätsgesetz verletzt, das Ergebnis im Fall β) ist wesentlich genauer als im Fall α), wie sich auch durch Rundungsfehleranalyse zeigen läßt.

Rundungsfehleranalyse für α):

$$(a +^* b) +^* c = ((a + b)(1 + \varepsilon_1) + c)(1 + \varepsilon_2)$$

$$\approx (a + b + c)\left(1 + \frac{a+b}{a+b+c}\,\varepsilon_1 + \varepsilon_2\right)$$

$$\frac{a+b}{a+b+c} \approx \frac{0,31 \cdot 10^4}{0,43} \approx 0,72 \cdot 10^4$$

Für diese Maschinenarithmetik ist eps = $\frac{1}{2}\,10^{-6}$ und somit kann der relative Fehler abgeschätzt werden durch

$$0,72 \cdot 10^4 \cdot \frac{1}{2} \cdot 10^{-6} + \frac{1}{2} \cdot 10^{-6} \approx 0,36 \cdot 10^{-2}$$

und der absolute Fehler durch

$$0{,}36 \cdot 10^{-2} \cdot 0{,}43 \approx 0{,}0015 \ .$$

Daraus folgt, daß nicht einmal die dritte Stelle nach dem Komma garantiert werden kann.

Rundungsfehleranalyse für β):

$$(a +^* (b +^* c)) = (a + (b + c)(1 + \varepsilon_1))(1 + \varepsilon_2)$$

$$\approx (a + b + c)\left(1 + \frac{b+c}{a+b+c}\,\varepsilon_1 + \varepsilon_2\right)$$

$$\frac{b+c}{a+b+c} \approx \frac{0{,}26}{0{,}43} \approx 0{,}6 \ .$$

Somit kann man den relativen Fehler abschätzen durch

$$0{,}6 \cdot \frac{1}{2}\,10^{-6} + \frac{1}{2}\,10^{-6} = 0{,}8 \cdot 10^{-6}$$

und den absoluten Fehler durch

$$0{,}8 \cdot 10^{-6} \cdot 0{,}43 \approx 0{,}34 \cdot 10^{-6} \ .$$

Eine Abweichung vom "exakten" Resultat kann also höchstens in der letzten Stelle des Ergebnisses erwartet werden.

Bsp.: Nun soll $a^2 - b^2$ berechnet werden und zwar nach folgenden Algorithmen

α) $(a \cdot^* a) -^* (b \cdot^* b)$

β) $(a +^* b) \cdot^* (a -^* b)$

Rundungsfehleranalyse für α):

$$((a \cdot a)(1 + \varepsilon_1) - (b \cdot b)(1 + \varepsilon_2))(1 + \varepsilon_3)$$

$$\approx (a^2 - b^2)\left(1 + \frac{a^2}{a^2-b^2}\,\varepsilon_1 - \frac{b^2}{a^2-b^2}\,\varepsilon_2 + \varepsilon_3\right)$$

Der Algorithmus ist somit "gefährlich", falls $a^2 \approx b^2$ (Auslöschung!). Der gesamte absolute Rundungsfehlereinfluß kann abgeschätzt werden durch

$$|a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3| \leq (a^2 + b^2 + |a^2 - b^2|)\text{eps} .$$

Rundungsfehleranalyse für β):

$$((a + b)(1 + \varepsilon_1) \cdot (a - b)(1 + \varepsilon_2))(1 + \varepsilon_3)$$

$$\approx (a^2 - b^2)(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

Der gesamte absolute Rundungsfehlereinfluß kann somit abgeschätzt werden durch

$$|(a^2 - b^2)(\varepsilon_1 + \varepsilon_1 + \varepsilon_1)| \leq 3\text{eps} \cdot |a^2 - b^2| .$$

Vergleicht man nun die beidem Rundungsfehler, so erkennt man, daß für $\frac{1}{3} \leq \left|\frac{a}{b}\right|^2 \leq 3$ Algorithmus β) "stabiler" ist, andernfalls aber Algorithmus α).

Bsp.: Nun soll die Wurzel $y = -p + \sqrt{p^2+q}$ der quadratischen Gleichung $y^2 + 2py - q = 0$ berechnet werden und zwar nach folgendem Algorithmus:

$$\eta_1 = p \cdot p$$

$$\eta_2 = \eta_1 + q$$

$$\eta_3 = \sqrt{\eta_2}$$

$$\eta_4 = -p + \eta_3$$

bzw.

$$\tilde{\eta}_1 = p \cdot p(1 + \varepsilon_1)$$

$$\tilde{\eta}_2 = (\tilde{\eta}_1 + q)(1 + \varepsilon_2)$$

$$\tilde{\eta}_3 = \sqrt{\tilde{\eta}_2}\ (1 + \varepsilon_3)$$

$$\tilde{\eta}_4 = (-p + \tilde{\eta}_3)(1 + \varepsilon_4)$$

$$\tilde{\eta}_4 = (-p + \sqrt{(p^2(1+\varepsilon_1)+q)(1+\varepsilon_2)}(1+\varepsilon_3))(1+\varepsilon_4)$$

Entwickelt man in erster Ordnung in ε_i, so erhält man

$$\tilde{\eta}_4 \approx y\left(1 - \frac{p}{y}\varepsilon_4 + \frac{\sqrt{p^2+q}}{y}\left(\frac{1}{2}\frac{p^2}{p^2+q}\varepsilon_1 + \frac{1}{2}\varepsilon_2 + \varepsilon_3 + \varepsilon_4\right)\right)$$

Gefährlich ist offensichtlich jener Fall, wo $p > 0$ und $q \approx 0$ ist, weil dann die Koeffizienten der ε_i sehr groß werden können. Um den Algorithmus für diesen Fall "stabil" zu gestalten, kann man so vorgehen, daß man die zweite Wurzel der quadratischen Gleichung $y_2 = -p - \sqrt{p^2+q}$ berechnet und dann aus der Relation $y_1 \cdot y_2 = -q$ das gewünschte y_1.

Bsp.: Man berechne $d = \sqrt{a^2 + b^2 - 2ab\cos\gamma}$ nach folgendem Algorithmus:

$$\tilde{\eta}_1 = a \cdot a(1+\varepsilon_1) \qquad \tilde{\eta}_6 = \tilde{\eta}_4 \cdot \tilde{\eta}_5(1+\varepsilon_6)$$

$$\tilde{\eta}_2 = b \cdot b(1+\varepsilon_2) \qquad \tilde{\eta}_7 = (\tilde{\eta}_1 + \tilde{\eta}_2)(1+\varepsilon_7)$$

$$\tilde{\eta}_3 = a \cdot b(1+\varepsilon_3) \qquad \tilde{\eta}_8 = (\tilde{\eta}_7 - \tilde{\eta}_6)(1+\varepsilon_8)$$

$$\tilde{\eta}_4 = 2 \cdot \tilde{\eta}_3(1+\varepsilon_4) \qquad \tilde{\eta}_9 = \sqrt{\tilde{\eta}_8}\,(1+\varepsilon_9)$$

$$\tilde{\eta}_5 = \cos\gamma\,(1+\varepsilon_5)$$

Rundungsfehleranalyse

$$\tilde{\eta}_9 = [((a^2(1+\varepsilon_1) + b^2(1+\varepsilon_2))(1+\varepsilon_7) -$$

$$- 2ab(1+\varepsilon_3)(1+\varepsilon_4)\cos\gamma\,(1+\varepsilon_5)(1+\varepsilon_6))(1+\varepsilon_8)]^{\frac{1}{2}}(1+\varepsilon_9)$$

$$\approx d\left(1 + \frac{a^2}{2d^2}(\varepsilon_1+\varepsilon_7+\varepsilon_8) + \right.$$

$$\left. + \frac{b^2}{2d^2}(\varepsilon_2+\varepsilon_7+\varepsilon_8) - \frac{ab\cos\gamma}{d^2}(\varepsilon_3+\varepsilon_4+\varepsilon_5+\varepsilon_6+\varepsilon_8) + \varepsilon_9\right)$$

Obwohl unsere Darstellung der Maschinenzahlen die wesentlichen Charakteristika trifft, weicht aber doch die Realisierung auf Maschinen gelegentlich etwas davon ab.

III. FORTPFLANZUNG DER DATENFEHLER

Während bei der Analyse der Rundungsfehler bei jeder arithmetischen Operation und bei Funktionsauswertungen, die durchgeführt werden, ein relativer Fehler zu berücksichtigen ist, wird nun die Auswirkung der relativen Fehler sämtlicher (oder einiger) Eingangsdaten untersucht.

Es seien ε_x und ε_y die relativen Fehler von x und y $(x,y \neq 0)$, d.h.

$$\varepsilon_x = \frac{\tilde{x} - x}{x} , \quad \varepsilon_y = \frac{\tilde{y} - y}{y}$$

oder

$$\tilde{x} = x(1 + \varepsilon_x) , \quad \tilde{y} = y(1 + \varepsilon_y) .$$

Wir untersuchen, wie sich diese relativen Fehler auf den relativen Fehler des Ergebnisses auswirken, wenn x und y durch arithmetische Operationen verknüpft werden, oder wenn Funktionen dieser Eingangsdaten gebildet werden.

Es gelten die folgenden

Fehlerfortpflanzungsformeln 1. Ordnung:

1. $x \cdot y$: $\quad \varepsilon_{xy} \approx \varepsilon_x + \varepsilon_y$

2. x / y : $\quad \varepsilon_{x/y} \approx \varepsilon_x - \varepsilon_y$

3. $x \pm y$: $\quad \varepsilon_{x\pm y} = \frac{x}{x \pm y} \varepsilon_x \pm \frac{y}{x \pm y} \varepsilon_y$ $\quad$ falls $\quad (x \pm y) \neq 0$

4. x^n : $\quad \varepsilon_{x^n} \approx n\varepsilon_x$

5. $\sqrt[n]{x}$: $\quad \varepsilon_{\sqrt[n]{x}} \approx \frac{1}{n} \varepsilon_x$

6. e^x : $\varepsilon_{e^x} \approx x\,\varepsilon_x$

7. $\ln x$: $\varepsilon_{\ln x} \approx \frac{1}{\ln x}\,\varepsilon_x$

8. $\sin x$: $\varepsilon_{\sin x} \approx (x \operatorname{ctg} x)\,\varepsilon_x$

9. $\cos x$: $\varepsilon_{\cos x} \approx (-x \operatorname{tg} x)\,\varepsilon_x$

10. $\operatorname{tg} x$: $\varepsilon_{\operatorname{tg} x} \approx \frac{2x}{\sin 2x}\,\varepsilon_x$

Beweis von 2:

$$\frac{\tilde{x}}{\tilde{y}} = \frac{x(1+\varepsilon_x)}{y(1+\varepsilon_y)} = \frac{x}{y}(1+\varepsilon_x)(1-\varepsilon_y+\varepsilon_y^2-\ldots)$$

$$= \frac{x}{y}\,(1+\underbrace{\varepsilon_x-\varepsilon_y}+\underbrace{\ldots}),$$

$\varepsilon_x - \varepsilon_y$: Näherung für den relativen Fehler des Ergebnisses;

$\ldots$: Glieder höherer Ordnung (i.a. sehr klein)

Beweis von 5:

$$\sqrt[n]{\tilde{x}} = \sqrt[n]{x(1+\varepsilon_x)} = \sqrt[n]{x}\,\sqrt[n]{(1+\varepsilon_x)}$$

$$= \sqrt[n]{x}\left(1+\frac{1}{n}\,\varepsilon_x+\underbrace{\frac{1}{n}\left(\frac{1}{n}-1\right)\frac{1}{2}\,\varepsilon_x^2+\ldots}\right)$$

Glieder höherer Ordnung (vernachlässigt);

Beweis von 6:

$$e^{\tilde{x}} = e^{x(1+\varepsilon_x)} = e^x\,e^{x\varepsilon_x} = e^x\left(1+\frac{x\varepsilon_x}{1!}+\frac{(x\varepsilon_x)^2}{2!}+\ldots\right) \approx e^x(1+x\varepsilon_x)$$

ad 3: Die Addition ist gefährlich, wenn $x \approx -y$. Die Faktoren $\frac{x}{x\pm y}$, $\frac{y}{x\pm y}$ heißen Verstärkerfaktoren oder *Konditionszahlen*.

Allgemein gilt (nach der Taylorformel):

Fehlerfortpflanzungsformel 1. Ordnung in einer Variablen:

$$f(\tilde{x}) = f(x(1+\varepsilon_x)) \approx f(x) + x\varepsilon_x f'(x)$$

$$= f(x)\left(1 + \underbrace{x\frac{f'(x)}{f(x)}\varepsilon_x}_{\varepsilon_{f(x)}}\right) = f(x)(1+\varepsilon_{f(x)})$$

Fehlerfortpflanzungsformel 1. Ordnung in zwei Variablen:

$$f(\tilde{x},\tilde{y}) = f(x(1+\varepsilon_x),y(1+\varepsilon_y)) \approx f(x,y) + x\varepsilon_x\frac{\partial f}{\partial x} + y\varepsilon_y\frac{\partial f}{\partial y}$$

$$= f(x,y)\left(1 + \underbrace{\frac{x}{f(x,y)}\frac{\partial f}{\partial x}\varepsilon_x + \frac{y}{f(x,y)}\frac{\partial f}{\partial y}\varepsilon_y}_{\varepsilon_{f(x,y)}}\right)$$

$$= f(x,y)(1+\varepsilon_{f(x,y)})$$

Nun soll für jene Beispiele, die wir einer Rundungsfehleranalyse unterzogen haben, auch eine Datenfehleranalyse durchgeführt werden.

Bsp.: $\frac{a^2 - \sin b}{c}$

Datenfehleranalyse

$$\frac{a^2(1+\varepsilon_a)^2 - \sin b(1+(b\,\mathrm{ctg}\,b)\varepsilon_b)}{c(1+\varepsilon_c)} \approx$$

$$\approx \frac{a^2 - \sin b}{c}\left(1 + \frac{2a^2}{a^2 - \sin b}\varepsilon_a - \frac{b\cos b}{a^2 - \sin b}\varepsilon_b\right)\frac{1}{1+\varepsilon_c}$$

$$\approx \frac{a^2 - \sin b}{c}\left(1 + \frac{2a^2}{a^2 - \sin b}\varepsilon_a - \frac{b\cos b}{a^2 - \sin b}\varepsilon_b - \varepsilon_c\right)$$

Dasselbe Ergebnis erhält man durch direkte Anwendung der analogen Fehlerfortpflanzungsformel 1. Ordnung in drei Variablen. Kritisch ist wieder der Fall $a^2 \approx \sin b$.

Man beachte, daß die Datenfehleranalyse unabhängig von der Reihenfolge der Auswertung ist.

Bsp.: $a + b + c$

Datenfehleranalyse

$$a(1 + \varepsilon_a) + b(1 + \varepsilon_b) + c(1 + \varepsilon_c) =$$

$$= (a + b + c)\left(1 + \frac{a}{a + b + c}\,\varepsilon_a + \frac{b}{a + b + c}\,\varepsilon_b + \frac{c}{a + b + c}\,\varepsilon_c\right)$$

Bsp.: $a^2 - b^2$

Datenfehleranalyse

$$a^2(1 + \varepsilon_a)^2 - b^2(1 + \varepsilon_b)^2 \approx (a^2 - b^2)\left(1 + \frac{2a^2}{a^2 - b^2}\,\varepsilon_a - \frac{2b^2}{a^2 - b^2}\,\varepsilon_b\right)$$

Bsp.: $y = -p + \sqrt{p^2 + q}$

Datenfehleranalyse

$$-p(1 + \varepsilon_p) + \sqrt{p^2(1 + \varepsilon_p)^2 + q(1 + \varepsilon_q)} \approx$$

$$\approx (-p + \sqrt{p^2 + q})\left(1 - \frac{p}{\sqrt{p^2 + q}}\,\varepsilon_p + \frac{p + \sqrt{p^2 + q}}{2\sqrt{p^2 + q}}\,\varepsilon_q\right)$$

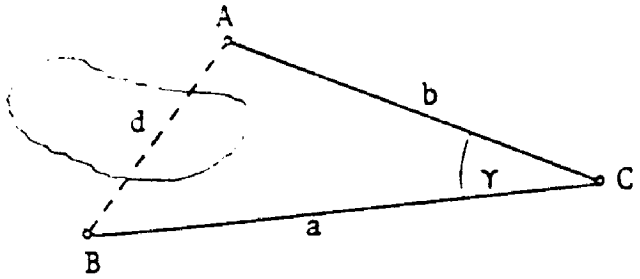
Bsp.: $d = \sqrt{a^2 + b^2 - 2ab\cos\gamma}$

Datenfehleranalyse

$$\sqrt{a^2(1+\varepsilon_a)^2 + b^2(1+\varepsilon_b)^2 - 2a(1+\varepsilon_a)b(1+\varepsilon_b)\cos\gamma(1-\gamma\varepsilon_\gamma\operatorname{tg}\gamma)} \approx$$

$$\approx d\left(1 + \frac{a^2 - ab\cos\gamma}{d^2}\,\varepsilon_a + \frac{b^2 - ab\cos\gamma}{d^2}\,\varepsilon_b + \frac{ab\,\gamma\sin\gamma}{d^2}\,\varepsilon_\gamma\right)$$

Diese Aufgabe kann etwa in folgenden Text eingekleidet werden: Der Abstand d zwischen den Punkten A und B kann wegen eines Hindernisses nicht gemessen werden. Statt dessen werden die Distanzen a und b zu einem Punkt C und der Winkel γ gemessen.



Es gilt dann $d = \sqrt{a^2 + b^2 - 2ab\cos\gamma}$.

a) Man bestimme die absoluten Konditionszahlen K_a, K_b und K_γ für die Bestimmung von d bezüglich der absoluten Fehler in a,b und γ. Man zeige, daß $|K_a|$ und $|K_b|$ stets kleiner gleich 1 sind.

b) Bei einer Messung ergab sich

$a = 74{,}83 \pm 0{,}02$ m $b = 53{,}19 \pm 0{,}02$ m

$\gamma = 35°12' \pm 3'$.

Wie stark können sich diese Meßungenauigkeiten auf den berechneten Abstand d auswirken?

Nach der obigen Datenfehleranalyse gilt

$$\underbrace{\tilde{d} - d}_{\Delta d} = \underbrace{\frac{a - b\cos\gamma}{d}}_{K_a}\underbrace{(a\varepsilon_a)}_{\Delta a} + \underbrace{\frac{b - a\cos\gamma}{d}}_{K_b}\underbrace{(b\varepsilon_b)}_{\Delta b} + \underbrace{\frac{ab\sin\gamma}{d}}_{K_\gamma}\underbrace{(\gamma\varepsilon_\gamma)}_{\Delta_\gamma}$$

$$|K_a| = \frac{|a - b\cos\gamma|}{\sqrt{a^2+b^2-2ab\cos\gamma}} = \sqrt{\frac{a^2+b^2\cos^2\gamma-2ab\cos\gamma}{a^2+b^2-2ab\cos\gamma}} \leq 1$$

$$|\Delta d| \leq 1 \cdot |\Delta a| + 1 \cdot |\Delta b| + \frac{a_{max} b_{max} \sin\gamma_{max}}{\sqrt{a^2_{min} + b^2_{min} - 2a_{max} b_{max} \cos\gamma_{min}}} \cdot |\Delta\gamma|$$

Durch einsetzen erhält man $|\Delta d| \leq 0{,}09$ m.

IV. SCHLUSSFOLGERUNGEN

Wir haben in unseren vorgehenden Betrachtungen die Analyse der Datenfehler und der Rundungsfehler streng getrennt. Dies sollte aus didaktischen Erwägungen stets getan werden, weil doch die Behandlung dieser beiden Fehlertypen durchaus verschieden ist. Natürlich treten bei der Auswertung jeder Formel i.a. sowohl Daten- wie auch Rundungsfehler auf, aber es ist leicht einzusehen, daß sich der relative Gesamtfehler als Summe des relativen Datenfehlers und des relativen Rundungsfehlers darstellen läßt.

Vorbedingung ist natürlich die Fähigkeit, eine "gestörte" Formel in erster Ordnung zu entwickeln. Mit der im III. Abschnitt angegebenen Formeltabelle ist dies in der höheren Schule durchaus möglich und man braucht dann weder auf die Formel von Taylor noch auf die Reihenentwicklungen der elementaren transzendenten Funktionen zurückzugreifen.

Jedenfalls rücken Analysen dieser Art die Mathematik in die Nähe des "realistischen Problems", sie schärfen den kritischen Verstand und es kommt ihnen ein bedeutender Bildungswert zu.

Weiterführende Literatur:

Stetter, H.J.: Numerik für Informatiker:
Computergerechte numerische Verfahren
R. Oldenbourg Verlag, München 1976.

Stoer, J.: Einführung in die Numerische Mathematik I
Springer Verlag, Berlin, 1972.