2.4.2 Rundungsfehleranalyse

Ein Algorithmus zur Lösung des Problems

$$y = \varphi(x)$$

lässt sich in elementare Operationen (Operationen, die am Rechner zur Verfügung stehen und eine relative Genauigkeit von eps besitzen) zerlegen:

$$x = x^{(0)} \mapsto x^{(1)} \mapsto x^{(2)} \mapsto \ldots \mapsto x^{(r)} \mapsto x^{(r+1)} = y.$$

Die Abbildung, die das Zwischenergebnis $x^{(s)}$ auf y abbildet, wird mit $\psi^{(s)}$ bezeichnet und heißt Restabbildung, wobei $s = 1, 2, \ldots, r$.

Beispiel: Ein möglicher Algorithmus zur Berechnung von

$$y = \frac{1}{2h}[f(x+h) - f(x-h)]$$

ist durch folgende Einzelschritte gegeben:

Algorithmus:

0. $x^{(0)} = x$
1. $x^{(1)} = f(x^{(0)} - h)$
2. $x^{(2)} = f(x^{(0)} + h)$
3. $x^{(3)} = x^{(2)} - x^{(1)}$
4. $x^{(4)} = x^{(3)}/(2h)$

Die dazugehörigen Restabbildungen lauten:

$$\begin{aligned} \psi^{(1)}(x^{(1)}) &= [f(x+h) - x^{(1)}]/(2h), \\ \psi^{(2)}(x^{(2)}) &= [x^{(2)} - f(x-h)]/(2h), \\ \psi^{(3)}(x^{(3)}) &= x^{(3)}/(2h). \end{aligned}$$

Sowohl bei der Dateneingabe als auch bei der Ausführung einer elementaren Operation entstehen Rundungsfehler. Dadurch erhält man anstelle von y nur eine Nährung $\bar{y}$.

Nach der Fehlerformel (2.6) verfälschen Datenfehler $\Delta x^{(0)}$ das Endergebnis näherungsweise um den Beitrag $\varphi'(x)\,\Delta x^{(0)}$.

Bei der Berechnung des Zwischenergebnisses $x^{(s)}$ für $s = 1, 2, \ldots, r$ entsteht ein neuer absoluter Fehler $\Delta x^{(s)}$, der laut (2.6) zu einer Verfälschung des Endergebnisses um näherungsweise $(\psi^{(s)})'(x^{(s)})\,\Delta x^{(s)}$ führt, wenn man annimmt, dass alle anderen Operationen exakt ausgeführt werden.

Schließlich entsteht noch ein weiterer Fehler $\Delta x^{(r+1)}$ bei der letzten elementaren Operation.

In erster Ordnung ist es gerechtfertigt, den Gesamteinfluss der einzelnen Rundungsfehler durch die Addition der oben beschriebenen Einzeleffekte zu erfassen. Somit erhält man für den gesamten Rundungsfehler Δy^R näherungsweise:

$$\Delta y^R = \bar{y} - y \approx \varphi'(x)\,\Delta x^{(0)} + \sum_{s=1}^{r} (\psi^{(s)})'(x^{(s)})\,\Delta x^{(s)} + \Delta x^{(r+1)}.$$

Der Gesamtrundungsfehler setzt sich aus zwei Anteilen zusammen, dem unvermeidbaren Fehler

$$\Delta y^0 = \varphi'(x)\,\Delta x^{(0)},$$

der sich aus der Datenfehlerfortpflanzung ergibt und der unabhängig vom gewählten Algorithmus ist, und dem restlichen Rundungsfehler

$$\Delta y^r = \sum_{s=1}^{r} (\psi^{(s)})'(x^{(s)})\,\Delta x^{(s)} + \Delta x^{(r+1)},$$

der vom gewählten Algorithmus abhängt.

Ein Algorithmus heißt **numerisch stabil**, wenn der restliche Rundungsfehler Δy^r den unvermeidbaren Fehler Δy^0 nicht dominiert.

Die Fehler $\Delta x^{(s)}$ für $s = 0, 1, \ldots, r, r+1$ lassen sich mit Hilfe der Maschinengenauigkeit abschätzen, siehe die Abschnitte über die Größe des Rundungsfehlers bei der Eingabe und bei elementaren Operationen:

$$|\Delta x^{(s)}| \leq \text{eps}\,|x^{(s)}| \quad \text{für } s = 0, 1, \ldots, r+1.$$

Daraus ergeben sich folgende Abschätzungen

$$|\Delta y^0| \leq \text{eps}\,|\varphi'(x)| \cdot |x| \quad \text{und} \quad |\Delta y^r| \leq \text{eps} \left[\sum_{s=1}^{r} |(\psi^{(s)})'(x^{(s)})|\,|x^{(s)}| + |y| \right].$$

Zur Beurteilung der numerischen Stabilität eines Algorithmus vergleicht man im Allgemeinen diese oberen Schranken für den unvermeidbaren Fehler bzw. den restlichen Rundungsfehler.

Beispiel: Für den obigen Algorithmus zur Berechnung von

$$y_h = \frac{1}{2h}\,[f(x+h) - f(x-h)]$$

erhält man im Fall $h \ll x$ für den unvermeidbaren Fehler

$$|\Delta y_h^0| \quad \lesssim \quad \text{eps}\,|f''(x)||x|$$

und für den restlichen Rundungsfehler

$$\begin{aligned} |\Delta y_h^r| &\lesssim \text{eps}\left(\frac{1}{2h}|f(x)| + \frac{1}{2h}|f(x)| + |f'(x)| + |f'(x)|\right) \\ &= \text{eps}\left(\frac{1}{h}|f(x)| + 2\,|f'(x)|\right) \approx \frac{\text{eps}}{h}\,|f(x)|. \end{aligned}$$

Falls $f(x)$ von Null verschieden ist und h hinreichend klein ist, ist der Algorithmus numerisch instabil. In diesem Fall erhält man für den gesamten Rundungsgfehler $\Delta y_h^R = \bar{y}_h - y_h$:

$$|\Delta y_h^R| \leq |\Delta y^0| + |\Delta y^r| \lesssim \text{eps}\,|f''(x)||x| + \frac{\text{eps}}{h}\,|f(x)| \approx \frac{\text{eps}}{h}\,|f(x)|.$$

Auch ohne detaillierte Fehleranalyse sieht man, dass es im letzten Schritt des Algorithmus zur Auslöschung kommt. Die Subtraktion selbst ist harmlos. Aber die an sich kleinen Rundungsfehler, die in den ersten 4 Schritten des Algorithmus entstehen, werden stark verstärkt.

Sieht man von eventuell vorhandenen zusätzlichen Datenfehlern ab, setzt sich der Gesamtfehler $\Delta y^G = \bar{y}_h - y$, der bei der Verwendung des zentralen Differenzenquotienten zur Approximation der ersten Ableitung $f'(x)$ entsteht, aus dem Verfahrensfehler $\Delta y^V = y_h - y$ und dem Rundungsfehler $\Delta y^R = \bar{y}_h - y_h$ zusammen:

$$\Delta y^G = \bar{y}_h - y = \bar{y}_h - y_h + y_h - y = \Delta y^R + \Delta y^V$$

mit

$$\Delta y^V \approx \frac{h^2}{6} f'''(x).$$

und

$$|\Delta y^R| \lesssim \frac{\text{eps}}{h}\,|f(x)|$$

zusammen. Also

$$|\Delta y^G| \lesssim \frac{h^2}{6}\,|f'''(x)| + \frac{\text{eps}}{h}\,|f(x)|.$$

Der Verfahrensfehler wird umso kleiner, je kleiner die Schrittweite h ist, der Rundungsfehler steigt hingegen mit kleiner werdender Schrittweite (Auslöschung). Die Abbildung 2.1 zeigt dieses gegenläufige Verhalten am Beispiel

$$f(x) = \sin x.$$

Für den Differenzenquotienten gilt hier

$$\frac{1}{2h}[f(x+h) - f(x-h)] = \cos x\,\frac{\sin h}{h}.$$

Der letzte Ausdruck ermöglicht für dieses Beispiel die Vermeidung der Auslöschung und damit eine (fast) rundungsfehlerfreie Auswertung des Differenzenquotienten. Wie Abbildung 2.1 zeigt, fällt der Verfahrensfehler proportional zu h^2, während der Rundungsfehler proportional zu $1/h$ steigt.

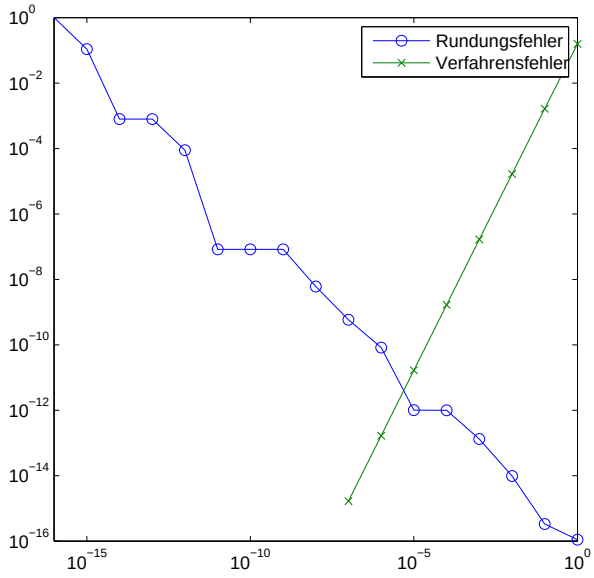

Abbildung 2.1: Numerische Differentiation: Verfahren- und Rundungsfehler

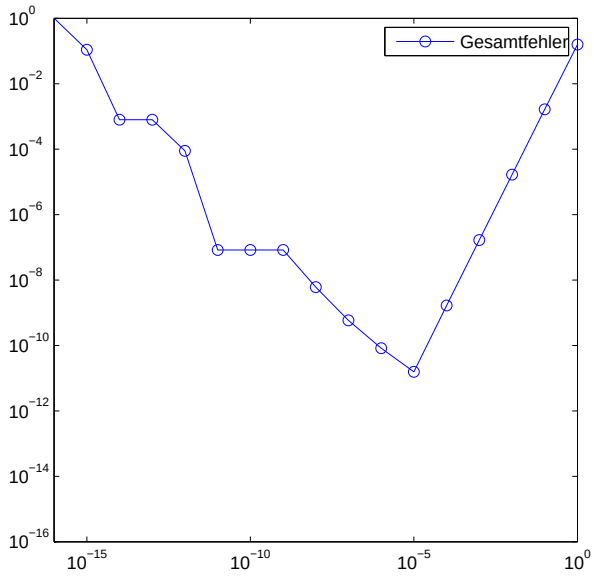

Abbildung 2.2: Numerische Differentiation: Gesamtfehler

Die optimale Schrittweite ist von der Größenordnung $\text{eps}^{1/3}$, siehe auch Abbildung 2.2.

Oft muss bei der Auswertung von f von einem wesentlich größeren relativen Fehler $\varepsilon_f \gg \text{eps}$ ausgegangen werden. Dann erhält man völlig analog die Abschätzung

$$|\Delta y^G| \lesssim \frac{h^2}{6} |f'''(x)| + \frac{\varepsilon_f}{h} |f(x)|.$$

Die optimale Schrittweite ist diesmal von der Größenordnung ${\varepsilon_f}^{1/3}$.

Kapitel 3

Direkte Verfahren zur Lösung linearer Gleichungssysteme

Lineare Gleichungssysteme entstehen z.B. bei der **Diskretisierung** von linearen partiellen Differentialgleichungen, die bei der Beschreibung zahlreicher physikalisch-technischer Probleme auftreten. Aber auch die Lösung nichtlinearer Probleme wird häufig auf die Lösung einer Folge von linearen Gleichungssystemen zurückgeführt (**Linearisierung**).

3.1 Ein Beispiel

In der Einleitung wurde das Problem der Berechnung der Temperaturverteilung in einem Stab diskutiert, das im stationären Fall auf ein Randwertproblem der folgender Art führt:

Gesucht ist eine Funktion u mit

$$\begin{aligned} -u''(x) &= f(x), \qquad x \in (0,1) \\ u(0) &= u(1) = 0, \end{aligned}$$

wobei f eine vorgegebene Funktion ist.

Durch Diskretisierung (Finite Differenzen Methode) entstand folgendes lineare Gleichungssystem:

$$-\frac{1}{h^2}\left(u_{i-1} - 2u_i + u_{i+1}\right) = f(x_i) \quad \text{für } i = 1, 2, \ldots, N$$

mit den Randbedingungen

$$u_0 = u_{N+1} = 0$$

für die Unbekannten u_i, $i = 1, 2, \ldots, N$, die als Näherungen für die exakte Lösung $u(x_i)$ in den Punkten x_i interpretiert werden können.

Man erhält also ein lineares Gleichungssystem

$$K_h \underline{u}_h = \underline{f}_h$$

mit der Matrix

$$K_h = (K_{ij})_{i,j=1,2,\ldots,N} = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

und den Vektoren

$$\underline{u}_h = (u_i)_{i=1,2,\ldots,N}, \quad \underline{f}_h = (f_i)_{i=1,2,\ldots,N} \text{ mit } f_i = f(x_i).$$

3.2 Datenstabilität

Wir betrachten nun folgende allgemeine Problemstellung: Gegeben ist eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ und ein Vektor $b = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n$. Gesucht ist ein Vektor $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ mit

$$Ax = b.$$

Im Sinne der allgemeinen Diskussion in Kapitel 2 ist das Problem, aus gegebenen Daten A und b die gesuchte Lösung x auszurechnen, rein formal durch

$$x = A^{-1}b = \varphi(A, b)$$

gegeben. Verfälschte Daten $\bar{b} = b+\Delta b$ und $\bar{A} = A+\Delta A$ führen auf ein verfälschtes Ergebnis $\bar{x} = x + \Delta x$

Die in Kapitel 2 vorgestellte Methode der Beurteilung der Datenstabilität würde es erforderlich machen, für jede der n Komponenten der Lösung und jede der $n^2 + n$ Komponenten der Daten eine Konditionszahl zu berechnen und abzuschätzen, eine völlig unübersichtliche Situation.

Besser ist es, mit Hilfe von Normen die Größe eines Vektors oder einer Matrix auf jeweils nur eine Zahl zu komprimieren.

So lässt sich statt der n komponentenweise gebildeten relativen Fehler $\varepsilon_{x_j} = (\bar{x}_j - x_j)/x_j$ die Abweichung im Ergebnis auch durch eine einzige Zahl (relativ gut) messen: $\varepsilon_x = \|\bar{x} - x\|_2/\|x\|_2$. Dabei bezeichnet $\|x\|_2$ die Euklidische Länge eines Vektors x:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

Analog misst man den relativen Fehler der rechten Seite b: $\varepsilon_b = \|\bar{b} - b\|_2/\|b\|_2$.

Anstelle der Euklidischen Norm kann auch eine andere Vektornorm verwendet werden.

Beispiel: Will man vor allem den komponentenweisen maximalen Fehler erfassen, bietet sich die Maximumnorm an:

$$\|x\|_\infty = \max_{i=1,2,\ldots,n} |x_i|.$$

Beispiel: Lassen sich die Komponenten eines Vektors z.B. als einzelne Massendefekte interpretieren, so ist man gelegentlich am Gesamtmassendefekt interessiert. Das führt auf die Verwendung der Betragssummennorm:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

Beispiel: Alle oben genannten Normen sind Spezialfälle der l_p-Norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

für alle $p \in [1, \infty)$. Der Fall $p = \infty$ ist als Grenzfall $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$ zu verstehen.

Beispiel: Die Euklidische Norm eines Vektors lässt sich mit Hilfe des Euklidischen Skalarprodukts darstellen:

$$\|x\|_2 = \sqrt{(x,x)_2} \quad \text{mit} \quad (x,y)_2 = \sum_i x_i y_i.$$

Für jede symmetrische und positiv definite Matrix A lässt sich durch

$$(x,y)_A = (Ax,y)_2 = \sum_{i,j} a_{ij} x_i y_j$$

ein Skalarprodukt und damit eine Norm definieren:

$$\|x\|_A = \sqrt{(x,x)_A}.$$

Alle diese Vektornormen in $\mathbb{R}^n$ erfüllen die drei für eine Norm charakteristischen Eigenschaften:

1. Definitheit: $\|v\| \geq 0$ und $\|v\| = 0$ nur, wenn $v = 0$,
2. Homogenität: $\|\lambda v\| = |\lambda| \|v\|$ für alle reellen Zahlen λ,
3. Dreiecksungleichung: $\|v + w\| \leq \|v\| + \|w\|$.

Auch die Größe von Matrizen lässt sich durch Normen (Matrixnormen) messen. Wenn Matrix- und Vektornormen gemeinsam verwendet werden, fordert man gewisse Verträglichkeitsbedingungen, vor allem soll folgende Eigenschaft gelten:

$$\|Ax\| \leq \|A\| \, \|x\| \quad \text{für alle } x \in \mathbb{R}^n.$$

Man sieht sofort, dass diese Eigenschaft für folgende Definition von $\|A\|$ erfüllt ist:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Man nennt diese Norm die der entsprechenden Vektornorm zugeordnete Matrixnorm.

Tatsächlich stellt sich heraus, dass durch diese Definition eine Matrixnorm entsteht, dass also die für eine Norm charakteristischen Eigenschaften (Definitheit, Homogenität und Dreiecksungleichung) erfüllt sind.

Darüber hinaus gelten zusätzlich noch die folgenden Rechenregeln für jede Vektornorm und die zugeordnete Matrixnorm:

1. Die Matrixnorm ist passend zur Vektornorm, d.h.:

$$\|Ax\| \leq \|A\| \, \|x\| \quad \text{für alle } A \in \mathbb{R}^{n \times n} \text{ und alle } x \in \mathbb{R}^n.$$

2. Die Matrixnorm ist submultiplikativ, d.h.:

$$\|AB\| \leq \|A\| \, \|B\| \quad \text{für alle } A, B \in \mathbb{R}^{n \times n}.$$

3. Für die Einheitsmatrix gilt: $\|I\| = 1$.

Beispiele:

1. Die der Euklidischen Norm $\|x\|_2$ zugeordnete Matrixnorm lässt sich folgendermaßen darstellen:
$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)},$$
wobei $\lambda_{\max}(B)$ den größten Eigenwert einer Matrix B bezeichnet. Sie heißt Spektralnorm.

2. Die der Maximumsnorm $\|x\|_\infty$ zugeordnete Matrixnorm lässt sich folgendermaßen darstellen:
$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$
Sie heißt Zeilenbetragssummennorm.

3. Die der Betragssummennorm $\|x\|_1$ zugeordnete Matrixnorm lässt sich folgendermaßen darstellen:
$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$
Sie heißt Spaltenbetragssummennorm.

Beispiel: Die Frobenius–Norm ist eine Matrixnorm, gegeben durch:

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}.$$

Sie entspricht der Euklidischen Norm, wenn man die Matrix $A \in \mathbb{R}^{n\times n}$ als Vektor in $\mathbb{R}^{n^2}$ interpretiert. Offensichtlich gilt:

$$\|I\|_F = \sqrt{n},$$

sie kann also nicht eine einer Vektornorm zugeordnete Matrixnorm sein. Sie ist aber trotzdem eine submultiplikative und zur Euklidischen Norm passende Matrixnorm und wesentlich einfacher berechenbar als die Spektralnorm.

Mit Hilfe einer dieser Matrixnormen lässt sich der relative Fehler in der Matrix A durch die Größe $\varepsilon_A = \|\bar{A} - A\|/\|A\|$ messen.

Mit diesen Vorbereitungen lässt sich nun die Datenstabilität eines linearen Gleichungssystems leicht untersuchen. Zuerst wird der Spezialfall $\bar{A} = A$ betrachtet:

Satz 3.1. *Seien $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix, $b \in \mathbb{R}^n$, $\Delta b \in \mathbb{R}^n$ und $\bar{b} = b + \Delta b$. $x \in \mathbb{R}^n$ erfülle das Gleichungssystem $Ax = b$, $\bar{x} = x + \Delta x$ erfülle das Gleichungssystem $A\bar{x} = \bar{b}$. Dann gilt für jede Vektornorm und eine dazu passende Matrixnorm:*

$$\varepsilon_x \leq \kappa(A)\, \varepsilon_b$$

mit $\varepsilon_x = \|\Delta x\|/\|x\|$, $\varepsilon_b = \|\Delta b\|/\|b\|$ und $\kappa(A) = \|A\|\, \|A^{-1}\|$.

Beweis. Durch Subtraktion von $x = A^{-1}b$ und $\bar{x} = x + \Delta x = A^{-1}\,\bar{b} = A^{-1}\,(b + \Delta b)$ erhält man $\Delta x = A^{-1}\,\Delta b$. Also

$$\|\Delta x\| \leq \|A^{-1}\|\, \|\Delta b\| = \|A^{-1}\|\, \|b\|\, \frac{\|\Delta b\|}{\|b\|}.$$

Mit $\|b\| = \|A\,x\| \leq \|A\|\, \|x\|$ folgt

$$\|\Delta x\| \leq \|A^{-1}\|\, \|A\|\, \|x\|\, \frac{\|\Delta b\|}{\|b\|},$$

woraus nach Division mit $\|x\|$ die Behauptung folgt. □

Die Zahl $\kappa(A)$ heißt die Konditionszahl der Matrix A. Sie ist für die Größe der Auswirkung von Datenfehlern in b auf die Lösung x verantwortlich.

Berücksichtigt man auch Störungen in A, so erhält man:

Satz 3.2. *Für jede Vektornorm und jede dazu passende und submultiplikative Matrixnorm gelten folgende Aussagen:*

1. *Falls $A \in \mathbb{R}^{n\times n}$ regulär ist und $\Delta A \in \mathbb{R}^{n\times n}$ die Abschätzung $\|\Delta A\| < 1/\|A^{-1}\|$ erfüllt, dann ist auch $\bar{A} = A + \Delta A$ regulär.*

2. *Seien zusätzlich $b \in \mathbb{R}^n$, $\Delta b \in \mathbb{R}^n$ und $\bar{b} = b + \Delta b$. $x \in \mathbb{R}^n$ erfülle das Gleichungssystem $Ax = b$, $\bar{x} = x + \Delta x$ erfülle das Gleichungssystem $\bar{A}\bar{x} = \bar{b}$. Dann gilt:*

$$\varepsilon_x \leq \frac{\kappa(A)}{1 - \kappa(A)\, \varepsilon_A}\, (\varepsilon_A + \varepsilon_b)$$

mit $\varepsilon_A = \|\Delta A\|/\|A\|$.