Numerische Mathematik

Vorlesung von Johann Linhart

Wintersemester 2004/05

Inhaltsverzeichnis

Kapitel 1

Einleitung

Am 2. 12. 2004 wurde ein zusätzlicher Abschnitt "Iterative Lösung von linearen Gleichungssystemen" angefügt!

Letzte Änderung: August 13, 2007

Die Numerische Mathematik beschäftigt sich mit der *Konstruktion* und *Analyse* von Algorithmen. Unter einem *Algorithmus* versteht man eine Rechenvorschrift mit folgenden charakteristischen Eigenschaften:

1. Jeder Schritt der Rechenvorschrift muss exakt und eindeutig festgelegt sein (*"Definitheit"*).

2. Der Rechenvorgang wird nach endlich vielen Schritten beendet (*"Finitheit"*).

3. Der Rechenvorgang ist auf eine ganze Problemklasse anwendbar (*"Allgemeingültigkeit"*).

Ein typisches Beispiel ist der bekannte euklidische Algorithmus zur Bestimmung des größten gemeinsamen Teilers. In der Numerischen Mathematik geht es aber vor allem um Algorithmen, die mit der Verarbeitung von reellen oder komplexen Zahlen zu tun haben und auf Methoden der Analysis und Linearen Algebra aufbauen. Ein bekanntes Beispiel ist das Newton-Verfahren zur Bestimmung einer Nullstelle einer nichtlinearen Funktion, bei dem die Funktion linear approximiert und die Lösung iterativ immer mehr verbessert wird (siehe Kapitel 8.2). Andere Algorithmen, wie Sortierverfahren, Suchverfahren und dergleichen, werden in Vorlesungen über Algorithmen und Datenstrukturen, Diskrete Mathematik, Graphentheorie etc. besprochen.

Bei der Analyse von Algorithmen unterscheiden wir zwei Hauptrichtungen, *Fehleranalyse* und *Komplexitätsanalyse*.

1.1 Fehleranalyse

Es geht dabei nicht um das Auffinden von Fehlern im Sinne von Programmierfehlern, unrichtigen mathematischen Formeln oder dergleichen, sondern um die Untersuchung von Rechenungenauigkeiten, auch *numerische Fehler* genannt: Auch bei grundsätzlich völlig richtigen Algorithmen entstehen oft fehlerhafte oder ungenaue Ergebnisse. Der Ursache nach unterscheidet man folgende Arten von Fehlern.

a) *Datenfehler:* Wenn die Eingabedaten z.B. von Messungen stammen, sind sie fast immer mit einem gewissen Fehler behaftet. Diese führen dann natürlich normalerweise auch zu Fehlern im Ergebnis.

b) *Verfahrensfehler:* Wenn der Algorithmus z.B. die Berechnung von Grenzwerten erfordert, muss nach endlich vielen Schritten abgebrochen werden. Das Ergebnis ist daher in den meisten Fällen unvermeidlich mit einem gewissen Fehler behaftet, auch wenn die Eingabedaten völlig exakt sind.

c) *Rundungsfehler:* Immer, wenn reelle Zahlen näherungsweise durch endlich viele Ziffern dargestellt werden, entstehen Rundungsfehler. Diese können sich zu sehr großen Fehlern akkumulieren, insbesondere, wenn die Anzahl der Rechenoperationen groß ist.

Als ein typisches *Beispiel* können wir die Berechnung des Flächeninhalts eines Gebietes ansehen, das durch eine einfach geschlossene Randkurve gegeben ist. Nehmen wir an, dass diese Kurve durch eine große Anzahl von Messpunkten gegeben ist, dann haben wir jedenfalls gewisse Messfehler. Die Berechnung des Flächeninhaltes kann nun z.B. dadurch erfolgen, dass wir die Kurve durch ein Polygon mit deutlich weniger Ecken approximieren. Durch diese Approximation entsteht ein Verfahrensfehler. Der Flächeninhalt des Polygons kann berechnet werden, indem man das Polygon in Dreiecke zerlegt. Dabei treten sicher gewisse Rundungsfehler auf, die besonders gravierend werden können, wenn man auch "negative" Dreiecke zulässt, da dann unter Umständen zwei fast gleich große Zahlen subtrahiert werden. Man spricht dann von "Auslöschung" (siehe Kapitel 3.2.2). Die Zulassung von "negativen" Dreiecken erleichtert allerdings die Zerlegung beträchtlich: man kann einfach einen festen Punkt (z.B. den Ursprung) mit je zwei aufeinanderfolgenden Ecken des Polygons verbinden.

1.2 Komplexitätsanalyse

Hier sind vor allem zwei Komplexitätsmaße von Bedeutung:

a) *Zeitkomplexität:* Das ist an und für sich die Rechenzeit in Abhängigkeit von der Anzahl der zu verarbeitenden Daten. Da die Rechenzeit sehr vom Rechner abhängt, verwendet man statt dessen oft die Anzahl der elementaren Rechenoperationen (wie Addition, Multiplikation und Division) bei n Eingabedaten. Von besonderem Interesse ist dabei das asymptotische Verhalten für $n \to \infty$, und zwar meist nur eine Abschätzung der Größenordnung. Zum Beispiel braucht man für die Berechnung des größten gemeinsamen Teilers von zwei natürlichen Zahlen mit höchstens n Dezimalziffern mit dem euklidischen Algorithmus $O(n)$ elementare Rechenoperationen (siehe Vorlesungen über Diskrete Mathematik oder Zahlentheorie), das heißt, die Anzahl dieser Rechenoperationen ist $\leq cn$ mit einer Konstanten c.

In der numerischen Mathematik wird die Zeitkomplexität auch in Abhängigkeit von der geforderten Genauigkeit untersucht. Die Zeitkomplexität ergibt sich dann aus der Fehleranalyse. Wenn wir z.B. wissen, dass bei einer gegen a konvergenten Folge (x_k) von reellen Zahlen der Verfahrensfehler $|x_k - a| \leq 2^{-k}$ ist, so erhalten wir daraus eine Schranke für die Anzahl der Schritte, welche nötig sind, damit dieser Fehler kleiner als eine vorgegebene Schranke ε_0 wird: $2^{-k} \leq \varepsilon_0$ ist ja gleichbedeutend mit $k \geq -\log_2 \varepsilon_0$. Die Anzahl der nötigen Schritte ist hier daher $O(|\log \varepsilon_0|)$ (vgl. z.B. Kapitel 8.3).

b) *Speicherkomplexität:* Das ist die Größe des notwendigen Speicherplatzes in Abhängigkeit von der Anzahl der Eingabedaten (und der geforderten Genauigkeit). Auch hier ist vor allem eine Abschätzung der asymptotischen Größenordnung interessant. In dieser Vorlesung wird die Speicherkomplexität allerdings nicht behandelt.

Die Fehler- und Komplexitätsanalyse eines Algorithmus dient oft dazu, einen besseren Algorithmus zu konstruieren. In der Praxis kommt es dabei nicht nur auf das aysmptotische Verhalten an, sondern auch auf den Zeit- und Platzbedarf bei den tatsächlich zu verarbeitenden Datenmengen und der geforderten Genauigkeit.

1.3 Literatur

Es gibt viele gute Lehrbücher über Numerische Mathematik. Diese Vorlesung orientiert sich aber in erster Linie an den Büchern von Hämmerlin/Hoffmann [6], Deuflhard/Hohmann [3] und Stoer [8], die sich deshalb für eine Vertiefung oder Weiterführung des Studiums besonders gut eignen.

Kapitel 2

Zahlendarstellungen

2.1 b-adische Entwicklung reeller Zahlen

Sei b eine beliebige natürliche Zahl > 1. Jede nicht-negative reelle Zahl kann bekanntlich folgenderweise durch eine (formal) beiderseits unendliche Folge von "*Ziffern*" $a_i \in \{0, \ldots, b-1\}$ dargestellt werden:

$$z \mapsto (\ldots, a_m, a_{m-1}, a_{m-2}, \ldots)$$

mit $a_m \neq 0$ und $a_i = 0$ für alle $i > m$, sodass

$$z = a_m b^m + a_{m-1} b^{m-1} + \ldots = \sum_{i=-\infty}^{\infty} a_i b^i.$$

Diese Darstellung ist eindeutig, wenn man ausschließt, dass von einem bestimmten Index an alle Ziffern $= b-1$ sind. Das heißt, wir verlangen, dass es keine ganze Zahl k gibt mit $a_i = b-1$ für alle $i \leq k$.

Die so definierte Folge $(a_i)_{i \in \mathbb{Z}}$ heißt die *b-adische Entwicklung* oder *Darstellung* der Zahl z zur *Basis* b.

Für $b = 10$ ist das äquivalent zur üblichen Dezimalbruchdarstellung. Für das Rechnen mit Computern hat die Darstellung zur Basis 2 besondere Bedeutung (*Dualdarstellung*).

2.2 Gleitkommadarstellung reeller Zahlen

2.2.1 Abschneiden

Wenn man reelle Zahlen durch ihre b-adische Entwicklung in einem Computer darstellen will, kann man natürlich nur endlich viele Ziffern verwenden. Man schneidet daher die obige Darstellung nach einer bestimmten Anzahl von Stellen ab und erhält die näherungsweise Darstellung

$$z \approx z' = a_m b^m + a_{m-1} b^{m-1} + \ldots + a_k b^k.$$

Man sagt, z' entsteht aus z durch *Abschneiden ("chopping") auf* $s = m-k+1$ Stellen.

Wenn wir hier b^{m+1} herausheben, erhalten wir

$$z' = (a_m b^{-1} + a_{m-1} b^{-2} + \ldots + a_{m-s+1} b^{-s}) b^{m+1}.$$

Das ist die sogenannte *normalisierte Gleitkommadarstellung*. Die Zahl

$$a = a_m b^{-1} + a_{m-1} b^{-2} + \ldots + a_{m-s+1} b^{-s}$$

heißt *Mantisse*, sie liegt im Intervall $[0, 1)$. Die Anzahl s der Stellen heißt auch *Mantissenlänge*.

Die Zahl $m + 1$ heißt *Exponent* dieser Darstellung. Das ist eine ganze Zahl, die in einem Computer immer auf einen bestimmten Bereich eingeschänkt ist: $e_{\min} \leq m + 1 \leq e_{\max}$. Im Folgenden ignorieren wir aber der Einfachheit halber diese Einschränkung meistens.

Die Menge aller Zahlen in normalisierter Gleitkommadarstellung (mit gegebenen $s, b, e_{\min}$ und $e_{\max}$) ist die Menge der *Maschinenzahlen*.

Zum Beispiel sieht die normalisierte Gleitkommadarstellung der Zahl π zur Basis 10 auf 3 Stellen so aus:

$$\pi \approx (3 \cdot 10^{-1} + 1 \cdot 10^{-2} + 4 \cdot 10^{-3}) \times 10^1.$$

Die Mantisse ist hier also 0.314, und der Exponent ist 1. Wir können auch schreiben

$$\pi \approx 0.314 \times 10^1.$$

2.2.2 Runden

Eine etwas genauere Darstellung erhalten wir, wenn wir die letzte Ziffer a_k durch $a_k + 1$ ersetzen, falls $a_{k-1} \geq b/2$ ist. Wenn dadurch $a_k = b$ würde, so

setzen wir $a_k = 0$ und erhöhen die vorhergehende Ziffer a_{k+1} um 1, usw.. Unter Umständen ändern sich dadurch mehrere vorhergehende Ziffern. Rundet man z.B. 0.1357996 auf 6 Stellen, so ergibt sich 0.135800.

Auf diese Weise erhalten wir eine Maschinenzahl $\tilde{z}$, welche den kleinstmöglichen Abstand von der gegebenen Zahl z hat.

Bezeichnung: $\tilde{z} = \mathrm{rd}_{s,b}(z)$ bei Rundung auf s Stellen. ($\tilde{z}$ ist eigentlich auch von $e_{\min}$ und $e_{\max}$ abhängig, das drücken wir aber nicht in der Bezeichnung aus.)

2.2.3 Relativer und absoluter Fehler

Wenn $\tilde{z}$ irgendeine Näherung der Zahl z ist, so heißt $|z - \tilde{z}|$ der *absolute Fehler* (dieser Näherung). Für $z \neq 0$ heißt

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

der *relative Fehler*. Diese zwei Arten von Fehlern spielen sowohl bei Datenfehlern als auch bei Verfahrens- und Rundungsfehlern eine Rolle.

Beispiel: Bei der obigen Näherung der Kreiszahl $\pi = 3.141\,592\ldots$, nämlich $\tilde{\pi} = 3.14$, ist der absolute Fehler gleich $0.00159\ldots$ und der relative Fehler gleich $\frac{0.00159\ldots}{3.14159\ldots} \approx 5{\times}10^{-4}$.

Manchmal drückt man den relativen Fehler auch in Prozent aus. Die Prozentzahl ist natürlich 100-mal so groß. In unserem Beispiel beträgt der relative Fehler also ungefähr 0.05 %.

2.2.4 Rundungsfehler

Wenn obiges $\tilde{z}$ durch Rundung entstanden ist, sprechen wir von *absolutem* bzw. *relativem Rundungsfehler.*

Überlegen wir uns nun, wie groß diese Rundungsfehler sein können. Sei z' die Maschinenzahl, welche durch Rundung von z auf s Stellen zur Basis b entstanden ist. Dann gilt (mit den Bezeichnungen von vorhin) offensichtlich

$$|z - \tilde{z}| \leq \frac{1}{2} b^k$$

und daher (für $z \neq 0$)

$$\left|\frac{z - \tilde{z}}{z}\right| \leq \frac{1}{2} b^{k-m} = \frac{1}{2} b^{1-s}.$$

Diese Zahl nennt man die *relative Maschinengenauigkeit*. Sie wird oft mit **eps** bezeichnet. Z.B. ergibt sich für $b = 2$ und $s = 24$ die Maschinengenauigkeit $\mathbf{eps} = \frac{1}{2}2^{1-24} = 2^{-24} \approx 5.96 \times 10^{-8}$. (Das entspricht dem, was man normalerweise unter *single precision* versteht.)

2.2.5 Gleitkommaarithmetik

Bei der Durchführung der Grundrechnungsarten mit reellen Zahlen in s-stelliger Gleitkommadarstellung wird das Ergebnis auch wieder gerundet. Anstelle der gewöhnlichen Addition, Multiplikation und Division handelt es sich dann also im Wesentlichen um folgende Verknüpfungen auf der Menge der Maschinenzahlen:

$$x \oplus y := \mathrm{rd}_{s,b}(x + y),$$
$$x \odot y := \mathrm{rd}_{s,b}(x \cdot y),$$
$$x \oslash y := \mathrm{rd}_{s,b}(x / y).$$

(Die Subtraktion kann auf die Addition und eine Vorzeichenänderung zurückgeführt werden.)

Die konkrete Realisierung dieser Verknüpfungen auf einem Rechner kann davon geringfügig abweichen, das spielt aber für unsere Überlegungen keine Rolle. Es soll hier nur darauf hingewiesen werden, dass diese Verknüpfungen nicht den üblichen Rechenregeln gehorchen. Das folgende Beispiel zeigt etwa, dass die Addition $\oplus$ nicht assoziativ ist:

Sei $s = 2$ und $b = 10$. Dann gilt

$$(0.50 \oplus 0.54) \oplus (-0.53) = 1.0 \oplus (-0.53) = 0.47,$$

$$0.50 \oplus (0.54 \oplus (-0.53)) = 0.50 \oplus 0.01 = 0.51.$$

In ähnlicher Weise kann man zeigen, dass auch die Multiplikation nicht assoziativ ist und das Distributivgesetz nicht gilt. Aus diesem Grund kann es vorkommen, dass zwei mathematisch äquivalente Ausdrücke bei der Auswertung durch einen Computer zu völlig verschiedenen Ergebnissen führen (siehe Übungsaufgaben).

Kapitel 3

Fehleranalyse

Wir diskutieren hier vor allem die Auswirkungen von Fehlern in den Eingabedaten auf das Ergebnis. Gleichzeitig werden wir auch Einblick in die Auswirkungen von Rundungsfehlern erhalten. Verfahrensfehler werden später bei den einzelnen numerischen Verfahren behandelt.

Der folgende Abschnitt über Normen dient als Vorbereitung.

3.1 Vektor- und Matrixnormen

Unter einer (Vektor-)*Norm* in einem reellen Vektorraum V versteht man bekanntlich eine Abbildung $V \to \mathbb{R} : x \mapsto \|x\|$ mit folgenden Eigenschaften (für alle $x, y \in V$ und $\lambda \in \mathbb{R}$):

$$\begin{aligned} &\|x\| \geq 0, \\ &\|x\| = 0 \Leftrightarrow x = o, \\ &\|\lambda x\| = |\lambda| \, \|x\| \,, \\ &\|x + y\| \leq \|x\| + \|y\| \,. \end{aligned}$$

Wichtige Beispiele für Normen im $\mathbb{R}^n$ sind die folgenden (wobei $x = (x_1, \ldots, x_n)$):

$$\|x\|_1 := \sum\nolimits_{i=1}^{n} |x_i| \quad (\textit{Summennorm}),$$

$$\|x\|_2 := \sqrt{\sum\nolimits_{i=1}^{n} x_i^2} \quad \textit{(euklidische Norm)},$$

$$\|x\|_\infty := \max_{1 \leq i \leq n} |x_i| \quad \textit{(Maximumsnorm)}.$$

Diese Normen heißen auch *1-Norm*, *2-Norm* bzw. ∞*-Norm*.

Zu jeder Vektornorm im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$ gibt es eine in natürlicher Weise dazugehörige *Matrixnorm*, die folgenderweise definiert ist:

$$\|A\| := \max_{\|x\|=1} \|Ax\| .$$

Hier ist A eine (reelle) $m \times n$-Matrix, und die Normstriche bedeuten auf der rechten Seite eine Vektornorm im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$.

Man kann unschwer nachprüfen, dass dadurch wirklich eine Norm auf dem Vektorraum $\mathbb{R}^{m,n}$ aller reellen $m \times n$-Matrizen definiert wird.

Die zu den oben angeführten drei Vektornormen gehörigen Matrixnormen bezeichnen wir ebenfalls mit $\|.\|_1$, $\|.\|_2$ bzw. $\|.\|_\infty$.

Beispiel 3.1 *Sei A eine orthogonale Matrix. Dann ist bekanntlich $\|Ax\|_2 = \|x\|_2$ für alle $x \in \mathbb{R}^n$, und daher $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = 1$.*

Bemerkung 3.2 *Aus der Definition von $\|A\|$ ergibt sich leicht die folgende Ungleichung (für alle $x \in \mathbb{R}^n$):*

$$\|Ax\| \leq \|A\| \, \|x\| . \tag{3.1}$$

Beweis: Für $x = o$ ist die Ungleichung trivialerweise richtig. Für $x \neq o$ sei $x' = \frac{1}{\|x\|} x$. Dann ist $\|x'\| = 1$ und daher gilt $\|Ax'\| \leq \|A\|$ nach Definition von $\|A\|$. Daher gilt $\|Ax'\| = \left\| A(\frac{1}{\|x\|} x) \right\| = \left\| \frac{1}{\|x\|} Ax \right\| = \frac{1}{\|x\|} \|Ax\| \leq \|A\|$, und daraus folgt die Behauptung. ■

Bemerkung 3.3 *Die Matrixnorm hat außerdem noch folgende schöne Eigenschaft:*

$$\|AB\| \leq \|A\| \, \|B\| \tag{3.2}$$

für je zwei Matrizen A und B, deren Produkt definiert ist.

Beweis: Sei $\|x\| = 1$. Dann ergibt sich durch zweimalige Anwendung von (3.1):

$\|ABx\| = \|A(Bx)\| \leq \|A\| \, \|Bx\| \leq \|A\| \, \|B\| \, \|x\| = \|A\| \, \|B\|$,

und daraus folgt schon die Behauptung. ■

In analoger Weise kann man für beliebige lineare Abbildungen zwischen normierten Vektorräumen eine Norm definieren. Seien V, V' solche Vektorräume und $f : V \to V'$. Dann heißt

$$\|f\| := \sup_{\|x\|=1} \|f(x)\|$$

die Norm von f, und alles Obige lässt sich ohne Weiteres übertragen. (Hier wurde sup statt max geschrieben, weil in unendlichdimensionalen Vektorräumen das Maximum unter Umständen nicht existiert.)

Die zu der 1-, 2- und ∞-Norm gehörigen Matrixnormen kann man auch auf Grund des folgenden Satzes berechnen.

Satz 3.4 *Für jede (reelle) $m \times n$-Matrix A gilt*

$$\|A\|_1 = \max_{1\le k\le n} \sum_{i=1}^{m} |a_{ik}|,$$

$$\|A\|_2 = \sqrt{\rho(A^T A)},$$

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{k=1}^{n} |a_{ik}|,$$

wobei ρ den Spektralradius bedeutet. Dieser ist folgenderweise für jede quadratische Matrix definiert:

$$\rho(M) := \max\{\ |\lambda| : \lambda \textit{ Eigenwert von } M\}.$$

$\|A\|_1$ ist also die größte "absolute Spaltensumme" und $\|A\|_\infty$ die größte "absolute Zeilensumme". Diese Normen heißen daher auch *Spaltensummennorm* bzw. *Zeilensummennorm*.

Beweis des Satzes:

Die Beweise für $\|A\|_1$ und $\|A\|_\infty$ stellen Übungsaufgaben dar.

Beweis für $\|A\|_2$:

Sei $M := A^T A$. Das ist eine reelle symmetrische $n \times n$-Matrix. Sie besitzt daher (der Vielfachheit nach) n reelle Eigenwerte $\lambda_1, \ldots, \lambda_n$, und es gibt eine orthogonale Matrix Q, sodass $D := Q^T M Q$ eine Diagonalmatrix ist, in deren Diagonalen die Eigenwerte stehen.

Wir betrachten jetzt die zu M gehörige quadratische Form:

$x^T M x = x^T A^T A x = (Ax)^T Ax = (Ax) \cdot (Ax) = \|Ax\|_2^2 \ge 0$ für alle $x \in \mathbb{R}^n$.

Wir sehen, sie ist positiv semidefinit, und daher sind alle Eigenwerte von M nicht-negativ.

Wir wollen nun $\|A\|_2^2 = \max \|Ax\|_2^2$ bestimmen, wobei das Maximum über die Menge $S^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ genommen wird (das ist die Einheitssphäre des $\mathbb{R}^n$).

Die zu Q gehörige lineare Abbildung lässt die Norm unverändert und ist bijektiv. Daraus folgt, dass sie die Sphäre S^{n-1} auf sich abbildet, das heißt $\{Qx : \|x\|_2 = 1\} = \{x : \|x\|_2 = 1\}$. Wir können daher bei der Bildung des Maximums an Stelle von x auch Qx schreiben (wobei stets $\|x\|_2 = 1$):

$\|A\|_2^2 = \max \|Ax\|_2^2 = \max x^T Mx = \max (Qx)^T M(Qx) = \max x^T Q^T MQx = \max x^T Dx = \max \sum \lambda_i x_i^2$.

Wir wollen zeigen, dass dieses Maximum gleich dem betragsgrößten Eigenwert ist. Da alle Eigenwerte ≥ 0 sind, ist dieser gleich dem größten Eigenwert (ohne Betrag), den wir mit λ_{i_0} bezeichnen.

Zunächst ist $\sum \lambda_i x_i^2 \leq \lambda_{i_0} \sum x_i^2 = \lambda_{i_0}$ für alle $x \in S^{n-1}$. Sei nun $\widetilde{x}_{i_0} = 1$ und $\widetilde{x}_i = 0$ für alle $i \neq i_0$. Dann ist $\|\widetilde{x}\|_2 = 1$ und $\sum \lambda_i \widetilde{x}_i^2 = \lambda_{i_0}$, was zu zeigen war. ■

3.2 Kondition einer Aufgabe

3.2.1 Allgemeines

Eine sehr allgemeine Klasse von Aufgaben kann man folgenderweise beschreiben. Jedem n-Tupel $x = (x_1, \ldots, x_n)$ von Eingabedaten soll ein m-Tupel $y = (y_1, \ldots, y_m)$ von Ergebnisdaten zugeordnet werden. Wenn es sich bei diesen Daten um reelle Zahlen handelt, dann geht es um eine Abbildung

$$\varphi : \mathbb{R}^n \to \mathbb{R}^m.$$

Häufig ist der Definitionsbereich nur eine Teilmenge T des $\mathbb{R}^n$, also

$$\varphi : \mathbb{R}^n \supset T \to \mathbb{R}^m : x \mapsto (\varphi_1(x), \ldots, \varphi_m(x)).$$

Gewöhnlich setzt man voraus, dass T eine offene Menge ist.

Wir interessieren uns nun dafür, wie sich Fehler der Eingabedaten auf das Ergebnis auswirken. Wenn wir annehmen, dass die Funktion φ differenzierbar ist, so können wir die Änderungen der Ergebnisdaten näherungsweise aus den Änderungen der Eingabedaten mit Hilfe der Jacobi-Matrix von φ berechnen.

Die *Jacobi-Matrix*[1] von φ ist die $m \times n$-Matrix aller partiellen Ableitungen (an einer Stelle x). Wir bezeichnen sie mit $\varphi'(x)$, also

$$\varphi'(x) = (a_{ik}) \quad \text{mit} \quad a_{ik} = \frac{\partial \varphi_i}{\partial x_k}(x).$$

Für festes $x \in \mathbb{R}^n$ ist $\varphi'(x)$ die Matrix einer linearen Abbildung, mit der die Abbildung φ in der Nähe von x approximiert werden kann. Wir schreiben das so:

$$\varphi(\tilde{x}) - \varphi(x) \doteq \varphi'(x)(\tilde{x} - x),$$

wobei der Punkt über dem Gleichheitszeichen bedeutet, dass der Unterschied zwischen linker und rechter Seite von kleinerer Größenordnung als $\|\tilde{x} - x\|$ ist. Genauer heißt das

$$\lim_{\tilde{x} \to x} \frac{\|\varphi(\tilde{x}) - \varphi(x) - \varphi'(x)(\tilde{x} - x)\|}{\|\tilde{x} - x\|} = 0.$$

(siehe Vorlesungen über Analysis).

Daraus können wir nun schließen, dass

$$\|\varphi(\tilde{x}) - \varphi(x)\| \ \dot{\leq} \ \|\varphi'(x)\| \, \|\tilde{x} - x\| \,,$$

wobei der Punkt über dem Ungleichheitszeichen natürlich bedeutet, dass die Ungleichung auch nur bis auf einen Term gilt, welcher von kleinerer Größenordnung als $\|\tilde{x} - x\|$ ist.

Denken wir uns $x = (x_1, \ldots, x_n)$ als die exakten Eingabedaten und $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ als Näherungswerte davon, so nennen wir $\Delta x := \tilde{x} - x$ den *Fehler der Eingabedaten* und $\Delta y := \varphi(\tilde{x}) - \varphi(x)$ den (dadurch verursachten) *Fehler des Ergebnisses.* Diese Fehler sind also Vektoren des $\mathbb{R}^n$ bzw. $\mathbb{R}^m$, und wir interessieren uns hier vor allem für ihre Normen. $\|\Delta x\|$ nennen wir den *absoluten Fehler der Eingabedaten*, und analog $\|\Delta y\|$ den *absoluten Fehler des Ergebnisses.*

Wir können also schreiben:

$$\|\Delta y\| \ \dot{\leq} \ \|\varphi'(x)\| \, \|\Delta x\| \,.$$

Die Zahl

$$\kappa_{\text{abs}} := \|\varphi'(x)\|$$

[1] Von Carl Gustav Jacob Jacobi (1804 - 1851, Potsdam, Berlin) stammen wichtige Arbeiten über Zahlentheorie, elliptische Funktionen (elliptische Integrale) und Differenzialgleichungen (siehe [9]).

heißt die *absolute Kondition* der betrachteten Aufgabe. Wir können somit schreiben:

$$\|\Delta y\| \stackrel{\cdot}{\leq} \kappa_{\text{abs}} \|\Delta x\| .$$

κ_{abs} gibt also eine obere Schranke für den Faktor an, um den sich der absolute Fehler der Eingabedaten vergrößern kann.

Für $x \neq o$ nennen wir $\frac{\|\Delta x\|}{\|x\|}$ den *relativen Fehler der Eingabedaten*, und analog (für $\varphi(x) \neq o$) $\frac{\|\Delta y\|}{\|y\|}$ den *relativen Fehler des Ergebnisses* (wobei natürlich $y := \varphi(x)$ ist).

Wir wollen nun den relativen Fehler des Ergebnisses durch den relativen Fehler der Eingabendaten abschätzen:

$$\frac{\|\Delta y\|}{\|y\|} \stackrel{\cdot}{\leq} \frac{1}{\|y\|} \|\varphi'(x)\| \|\Delta x\| \;=\; \frac{\|x\|}{\|y\|} \|\varphi'(x)\| \frac{\|\Delta x\|}{\|x\|}.$$

Der hier auftretende Faktor heißt die *relative Kondition* und wird mit κ_{rel} bezeichnet:

$$\kappa_{\text{rel}} = \frac{\|x\|}{\|y\|} \|\varphi'(x)\| .$$

Es gilt also

$$\frac{\|\Delta y\|}{\|y\|} \stackrel{\cdot}{\leq} \kappa_{\text{rel}} \frac{\|\Delta x\|}{\|x\|}.$$

κ_{rel} gibt eine obere Schranke für den Faktor an, um den sich der relative Fehler der Eingabedaten vergrößern kann.

Wenn κ_{abs} oder κ_{rel} wesentlich größer als 1 ist, so spricht man von einer *schlecht konditionierten Aufgabe*. Das ist natürlich ein vager Begriff. Was man unter "wesentlich größer" versteht, welche der beiden Konditionszahlen betrachtet wird, und welche Norm man verwendet, hängt von der jeweiligen Aufgabe ab.

Als einfaches *Beispiel* betrachten wir das Ziehen der Quadratwurzel. Hier ist $\varphi : \mathbb{R}^+ \to \mathbb{R} : x \mapsto \sqrt{x}$ und $\varphi'(x) = \frac{1}{2\sqrt{x}}$. Daher erhalten wir $\kappa_{\text{abs}} = \left|\frac{1}{2\sqrt{x}}\right|$ und $\kappa_{\text{rel}} = \left|\frac{x}{\sqrt{x}} \frac{1}{2\sqrt{x}}\right| = \frac{1}{2}$. Die absolute Kondition ist also für x in der Nähe von 0 schlecht, während die relative Kondition für alle $x \neq 0$ gut ist.

3.2.2 Kondition der Grundrechnungsarten

Die Konditionen der Grundrechnungsarten stellen einerseits wichtige Beispiele dar, und andererseits bilden sie die Grundlage für die Analyse von Rundungsfehlern.

Addition und Subtraktion

Hier betrachten wir die Abbildung $\varphi : \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto x_1 + x_2$.

$\frac{\partial\varphi}{\partial x_1} = \frac{\partial\varphi}{\partial x_2} = 1$, die Jacobimatrix sieht also so aus: $J := \varphi'(x_1, x_2) = \begin{pmatrix} 1 & 1 \end{pmatrix}$.

$\|J\|_1 = 1$, $\|J\|_\infty = 2$,

$$\|J\|_2^2 = \rho\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \end{pmatrix}\right) = \rho\left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = 2,$$

denn die Eigenwerte dieser Matrix sind 0 und 2. (Das charakteristische Polynom lautet $\begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda = \lambda(\lambda - 2)$.)

Wir sehen, dass die Addition bezüglich κ_{abs} als gut konditioniert angesehen werden kann, denn bei Verwendung der 1-Norm ist

$$\kappa_{\text{abs}} = 1.$$

Für die relative Kondition erhalten wir bei Verwendung der 1-Norm:

$$\kappa_{\text{rel}} = \frac{|x_1| + |x_2|}{|x_1 + x_2|}.$$

Wenn x_1 und x_2 nicht negativ sind, dann ist $\kappa_{\text{rel}} = 1$. Dasselbe ergibt sich, wenn beide Zahlen nicht positiv sind. Dagegen kann κ_{rel} für zwei Zahlen mit verschiedenem Vorzeichen beliebig groß werden, da für $x_1 \approx -x_2$ der Nenner beliebig nahe an Null herankommt. (Für $x_1 = -x_2$ ist κ_{rel} nicht definiert.) Die Addition von zwei reellen Zahlen mit verschiedenen Vorzeichen und annähernd gleichem Absolutbetrag ist daher extrem schlecht konditioniert (in Bezug auf die relative Kondition). Wir können das auch so ausdrücken: *Die Subtraktion zweier ungefähr gleich großer positiver Zahlen ist extrem schlecht konditioniert.* Dieses Phänomen ist vielfach für große numerische Fehler verantwortlich. Man nennt es auch *Auslöschung*, weil bei der Subtraktion solcher Zahlen oft mehrere Stellen (der Gleitpunktdarstellung) gewissermaßen "ausgelöscht" werden.

Beispiel: In 4-stelliger Gleitkommadarstellung (zur Basis 10) ist $\pi \approx 0.3142\times10^1$ und $\sqrt[3]{31} \approx 0.3141\times10^1$. Bilden wir die Differenz, so erhalten wir $0.1\times10^{-2} = 10\times10^{-4}$. Die Ziffern $3, 1, 4$ "verschwinden" also sozusagen. Der exakte Wert der Differenz ist $\pi - \sqrt[3]{31} = 2.12\ldots\times10^{-4}$. Bei unserem Ergebnis stimmt also nicht einmal die erste Stelle. Der absolute Fehler ist zwar nur $7.88\ldots\times10^{-4}$, der relative Fehler beträgt aber $\frac{7.88\ldots\times10^{-4}}{2.12\ldots\times10^{-4}} \approx 3.7$, das sind 370 % (!), obwohl der relative Fehler der Eingabedaten jedenfalls $\leq \frac{1}{2} \times 10^{1-4} = 5\times10^{-4} = 0.05$ % ist (siehe Abschnitt 2.2.4).

Multiplikation

Hier betrachten wir $\varphi : \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto x_1 x_2$.

$\frac{\partial \varphi}{\partial x_1} = x_2$, $\frac{\partial \varphi}{\partial x_2} = x_1$, und $J := \varphi'(x_1, x_2) = \begin{pmatrix} x_2 & x_1 \end{pmatrix}$.

Bei Verwendung der 1-Norm erhalten wir

$$\kappa_{\text{abs}} = \max(|x_2|, |x_1|).$$

Die absolute Kondition ist also umso schlechter, je größer die Faktoren sind. Das ist auch unmittelbar klar: Jeder Fehler in einem Faktor wird mit dem anderen Faktor multipliziert:

$$(x_1 + \delta)\, x_2 = x_1 x_2 + \delta x_2.$$

Wenn wir etwa im Anschluss an obiges Beispiel $10000 \cdot (\pi - \sqrt[3]{31})$ mit 4-stelliger Gleitkommaarithmetik berechnen, so erhalten wir den Wert 10 statt des genauen Werts $2.12\ldots$. Der absolute Fehler ist 10000-mal so groß wie bei $\pi - \sqrt[3]{31}$, also $7.88\ldots$.

Für die relative Kondition ergibt sich:

$$\kappa_{\text{rel}} = \frac{|x_1| + |x_2|}{|x_1 x_2|} \max(|x_2|, |x_1|).$$

Sei o.B.d.A. $|x_1| \leq |x_2|$. Dann sehen wir $\kappa_{\text{rel}} = \frac{|x_1|+|x_2|}{|x_1|} = 1 + \frac{|x_2|}{|x_1|} \geq 2$. Die relative Kondition der Multiplikation ist also auch nicht besonders gut. (Für die ∞-Norm erhalten wir dasselbe.) Der Wert 2 ergibt sich genau dann, wenn $|x_1| = |x_2|$ ist.

Sehen wir uns auch noch die 2-Norm an: Hier ist $\|J\|_2^2 = \rho\left(\begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \begin{pmatrix} x_2 & x_1 \end{pmatrix}\right) =$
$\rho\left(\begin{pmatrix} x_2^2 & x_1 x_2 \\ x_1 x_2 & x_1^2 \end{pmatrix}\right) = x_1^2 + x_2^2$,
denn $\begin{vmatrix} x_2^2 - \lambda & x_1 x_2 \\ x_1 x_2 & x_1^2 - \lambda \end{vmatrix} = \lambda\,(\lambda - x_1^2 - x_2^2)$. Es folgt

$$\kappa_{\text{rel}} = \frac{\sqrt{x_1^2 + x_2^2}}{|x_1 x_2|} \sqrt{x_1^2 + x_2^2} = \frac{x_1^2 + x_2^2}{|x_1 x_2|}.$$

Auch diese Konditionszahl ist stets ≥ 2, denn wegen $x_1^2 + x_2^2 - 2\,|x_1 x_2| = (|x_1| - |x_2|)^2 \geq 0$ ist $x_1^2 + x_2^2 \geq 2\,|x_1 x_2|$.

Interessanterweise erweist sich die Multiplikation aber als gut konditioniert (bezüglich κ_{rel}), wenn wir sie nur in Abhängigkeit von einem Faktor betrachten. Wir denken uns also einen Faktor festgehalten, sagen wir x_2, und betrachten die Abbildung $\mathbb{R} \to \mathbb{R} : x_1 \mapsto x_1 x_2$. Es kommt auf dasselbe hinaus, wenn wir $\varphi : \mathbb{R} \to \mathbb{R} : x \mapsto cx$ mit einer Konstanten c setzen. Dann erhalten wir $\kappa_{\text{rel}} = \left|\frac{x}{cx}\, c\right| = 1$. (Hier ist es egal, welche Norm wir nehmen, da im $\mathbb{R}^1$ alle betrachteten Normen übereinstimmen.)

Division

Da wir die Division durch die Multiplikation mit dem reziproken Wert des Divisors ersetzen können, untersuchen wir hier nur die Abbildung $x \mapsto 1/x$. Für diese erhalten wir

$$\begin{aligned} \kappa_{\text{abs}} &= \frac{1}{x^2}, \\ \kappa_{\text{rel}} &= \frac{x}{1/x}\,\frac{1}{x^2} = 1. \end{aligned}$$

Die absolute Kondition der Division durch x ist also besonders schlecht in der Nähe von $x = 0$. Das entspricht einer Faustregel, nach der jede Aufgabe in der Nähe einer Stelle, wo sie nicht lösbar ist, schlecht konditioniert ist. Die relative Kondition der Division ist dagegen (vielleicht zunächst etwas überraschend) überall, wo sie definiert ist, gut. (Dabei halten wir allerdings, ähnlich wie bei der Multiplikation, den Dividenden konstant).

3.2.3 Kondition eines linearen Gleichungssystems

Ein lineares Gleichungssystem ist von der Form $Ax = b$, wobei A eine reelle $m \times n$-Matrix und $b \in \mathbb{R}^n$ ist. Wir betrachten hier nur den Fall, dass $m = n$ und A invertierbar ist. Die Menge aller invertierbaren reellen $n \times n$-Matrizen bildet bezüglich der Matrizenmultiplikation eine Gruppe. Sie heißt *allgemeine lineare Gruppe* (engl. *general linear group*) und wird mit $\mathbb{GL}_n$ bezeichnet.

Unter den genannten Voraussetzungen hat $Ax = b$ eine eindeutige Lösung, nämlich $x = A^{-1}b$. Es geht also um die folgende Abbildung:

$$\varphi : \mathbb{GL}_n \times \mathbb{R}^n \to \mathbb{R}^n : (A, b) \mapsto A^{-1}b.$$

Kondition von $Ax = b$ bezüglich b

Zur Vereinfachung betrachten wir zunächst nur die Abhängigkeit von b, ähnlich wie vorhin bei der Multiplikation und Division. Das heißt, wir studieren

die Abbildung

$$f : \mathbb{R}^n \to \mathbb{R}^n : b \mapsto A^{-1}b.$$

Das ist eine lineare Abbildung, und die Jacobimatrix von f ist gleich A^{-1}. Wir erhalten daher (für $x = A^{-1}b \neq o$)

$$\begin{aligned} \kappa_{\text{abs}} &= \left\|A^{-1}\right\|, \\ \kappa_{\text{rel}} &= \frac{\|b\|}{\|A^{-1}b\|}\left\|A^{-1}\right\| = \frac{\|Ax\|}{\|x\|}\left\|A^{-1}\right\|. \end{aligned}$$

Wegen $\|Ax\| \leq \|A\|\,\|x\|$ können wir die relative Kondition folgenderweise abschätzen:

$$\kappa_{\text{rel}} \leq \|A\|\left\|A^{-1}\right\|.$$

Die hier auftretende Zahl

$$\kappa(A) := \|A\|\left\|A^{-1}\right\|$$

heißt die *Kondition* oder *Konditionszahl der Matrix* A. Sie beschreibt gewissermaßen die relative Kondition von $Ax = b$ für alle möglichen rechten Seiten b bei fester (exakter) Matrix A. Diese Zahl ist stets ≥ 1, denn $1 = \|E\| = \|AA^{-1}\| \leq \|A\|\,\|A^{-1}\|$ (siehe Ungleichung (3.2)). (E bedeutet natürlich die Einheitsmatrix.)

Beispiel:

Sehen wir uns das folgende einfache Gleichungssystem an (mit $a \geq 0$):

$$\begin{aligned} x_1 + ax_2 &= 1 \\ ax_1 + x_2 &= 1 \end{aligned}$$

Hier ist $A = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$ und $A^{-1} = \frac{1}{1-a^2}\begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix}$. Betrachten wir z.B. die 1-Norm:

$\|A\|_1 = 1 + a$, $\quad \|A^{-1}\|_1 = \frac{1}{|1-a^2|}(1+a)$, also $\kappa(A) = \frac{(1+a)^2}{|1-a^2|}$. Für $a \approx 1$ ist die Konditionszahl sehr groß, und das ist auch anschaulich plausibel, da es sich in diesem Fall um den Schnitt von zwei fast parallelen Geraden handelt. Betrachten wir z.B. den Fall $a = 0.99$ und überlegen wir uns, wie es sich auf die Lösung auswirkt, wenn wir auf der rechten Seite einmal 1.01 statt 1 schreiben. Wir erhalten folgende Lösungen (gerundet):

$$x = \frac{1}{1-0.99^2}\begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.5025 \\ 0.5025 \end{pmatrix},$$

$$\tilde{x} = \frac{1}{1-0.99^2}\begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix}\begin{pmatrix} 1.01 \\ 1 \end{pmatrix} = \begin{pmatrix} 1.0050 \\ 0.0050 \end{pmatrix}$$

also $\|x - \tilde{x}\|_1 = 1$, während $\left\|b - \tilde{b}\right\|_1 = 0.01$ ist. Für die entsprechenden relativen Fehler ergibt sich 0.995 bzw. 0.005. Die Konditionszahl ist $\frac{1.99^2}{1-0.99^2} = 199$, das ist hier gleich dem Verhältnis der relativen Fehler.

Wählen wir dagegen $b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ und $\tilde{b} = \begin{pmatrix} 1.01 \\ 0 \end{pmatrix}$, so ergibt sich $x = \begin{pmatrix} 50.25 \\ -49.75 \end{pmatrix}$, $\tilde{x} = \begin{pmatrix} 50.75 \\ -50.25 \end{pmatrix}$, also wieder $\|x - \tilde{x}\|_1 = 1$, aber der relative Fehler des Ergebnisses ist jetzt nur 0.01. Das Verhältnis ist der relativen Fehler ist hier also sehr viel kleiner als die Konditionszahl, nämlich 1.

Bemerkung 3.5 *Es scheint vielleicht naheliegend zu sein, die Determinante einer Matrix als Maß für ihre Kondition zu verwenden. Das ist jedoch nicht sinnvoll, da* $\det(\lambda A) = \lambda^n \det A$ *für* $\lambda \in \mathbb{R}$*, während* λA *für alle* $\lambda \neq 0$ *dieselbe Kondition wie* A *hat.*

Bemerkung 3.6 *Alle orthogonalen Matrizen haben bezüglich der 2-Norm die Konditionszahl* 1*.*

Die orthogonalen Matrizen haben also diesbezüglich die bestmögliche Kondition.

Beweis: Sei Q eine orthogonale $n \times n$-Matrix. Dann ist $\|Q\|_2 = 1$ (siehe Beispiel 3.1). Da $Q^{-1} = Q^T$ ebenfalls orthogonal ist, gilt auch $\|Q^{-1}\|_2 = 1$ und daher $\kappa_2(Q) = 1$. ■

Kondition von $Ax = b$ bezüglich A

Versuchen wir nun, die Kondition in Abhängigkeit von A zu studieren. Da geht es um die Abbildung

$$g : \mathbb{GL}_n \to \mathbb{R}^n : A \mapsto A^{-1}b.$$

Da wir jetzt b als konstant ansehen wollen, ist das Wesentliche die Abbildung

$$h : \mathbb{GL}_n \to \mathbb{GL}_n : A \mapsto A^{-1}.$$

Die zugehörige absolute Kondition ist analog zu 3.2.1 durch die Norm der Ableitung definiert, nur dass es sich hier nicht um eine Abbildung vom $\mathbb{R}^n$ in den $\mathbb{R}^m$ handelt. In der folgenden Einschaltung wird daher erklärt, was man unter der Ableitung in beliebigen normierten Vektorräumen versteht und wie man sie (im endlichdimensionalen Fall) berechnet.

Differenziation in normierten Vektorräumen

Definition 3.7 *Seien V,V' reelle normierte Vektorräume und T eine offene Teilmenge von V. Eine Abbildung*

$$f : T \to V'$$

heißt ***differenzierbar*** *in $x_0 \in T$, wenn es eine lineare Abbildung*

$$f'(x_0) : V \to V'$$

gibt, sodass

$$\lim_{u \to o} \frac{\|f(x_0 + u) - f(x_0) - f'(x_0)(u)\|}{\|u\|} = 0.$$

Die Abbildung $f'(x_0)$ heißt dann die ***Ableitung*** *von f an der Stelle x_0.*

Wie kann man nun $f'(x_0)$ berechnen (falls es existiert)? Dazu überlegen wir uns Folgendes.

Sei $u \in V,\ u \neq o$. (Hier ist also u ein fester Vektor im Gegensatz zu vorhin.) Dann gilt nach obiger Definition (falls f in x_0 differenzierbar ist):

$$\lim_{\lambda \to 0+} \frac{\|f(x_0 + \lambda u) - f(x_0) - f'(x_0)(\lambda u)\|}{\|\lambda u\|} = 0.$$

Dabei bedeutet $\lambda \to 0+$, dass wir uns auf positive Werte von λ beschränken.

Da $f'(x_0)$ linear ist, können wir das auch so schreiben:

$$\lim_{\lambda \to 0+} \frac{\|f(x_0 + \lambda u) - f(x_0) - \lambda f'(x_0)(u)\|}{|\lambda| \, \|u\|} = 0.$$

Da u konstant ist, können wir $\|u\|$ weglassen. Außerdem können wir bei $|\lambda|$ die Betragsstriche weglassen und die Division durch λ innerhalb der Normstriche durchführen. Das ergibt

$$\lim_{\lambda \to 0} \left\| \frac{f(x_0 + \lambda u) - f(x_0)}{\lambda} - f'(x_0)(u) \right\| = 0.$$

Das heißt aber nichts anderes als

$$f'(x_0)(u) = \lim_{\lambda \to 0} \frac{f(x_0 + \lambda u) - f(x_0)}{\lambda},$$

das ist die Ableitung der Abbildung $\mathbb{R} \to V' : \lambda \mapsto f(x_0 + \lambda u)$ an der Stelle $\lambda = 0$.

Damit haben wir prinzipiell eine Möglichkeit, die Werte der linearen Abbildung $f'(x_0)$ für beliebige Vektoren $u \neq o$ zu berechnen. (Für $\|u\| = 1$ nennt man das auch die Richtungsableitung von f in Richtung u (an der Stelle x_0)). Wenn wir eine Basis von V kennen, können wir auf diese Weise die Spaltenvektoren der entsprechenden Matrix von $f'(x_0)$ berechnen. Für $V = \mathbb{R}^m$ ergibt das mit der kanonischen Basis die bekannte Jacobimatrix.

Differenziation der Matrixinversion

Wir wenden das im letzten Abschnitt Gesagte nun auf die Abbildung

$$h : \mathbb{GL}_n \to \mathbb{GL}_n : A \mapsto A^{-1}$$

an. Das ist sinnvoll, denn $\mathbb{GL}_n$ ist eine offene Teilmenge des Vektorraums aller $n \times n$-Matrizen, und dieser kann z.B. durch die Normen $\|.\|_p$ mit $p = 1, 2$ oder ∞ normiert werden.

(Die Tatsache, dass $\mathbb{GL}_n$ offen ist, ergibt sich folgendermaßen: Eine $n \times n$-Matrix ist genau dann invertierbar, wenn ihre Determinante $\neq 0$ ist. Die Determinante ist aber eine stetige Abbildung. Also ist $\mathbb{GL}_n$ das Urbild der offenen Menge $\mathbb{R} \setminus \{0\}$ unter einer stetigen Abbildung und daher offen.)

Nach Obigem ist

$$h'(A_0)(U) = \lim_{\lambda \to 0} \frac{(A_0 + \lambda U)^{-1} - A_0^{-1}}{\lambda}.$$

Das ist die Ableitung der Funktion $\lambda \mapsto (A_0 + \lambda U)^{-1}$ an der Stelle $\lambda = 0$. Um diese zu berechnen, betrachten wir folgende Gleichung:

$$(A_0 + \lambda U)(A_0 + \lambda U)^{-1} = E.$$

Diese Gleichung ist offensichtlich richtig, soferne $(A_0 + \lambda U)^{-1}$ existiert, und das ist für genügend kleine λ wegen der Offenheit von $\mathbb{GL}_n$ der Fall.

Differenzieren wir nun obige Gleichung nach λ, so erhalten wir nach der Produktregel

$$U(A_0 + \lambda U)^{-1} + (A_0 + \lambda U) \frac{d}{d\lambda}(A_0 + \lambda U)^{-1} = 0.$$

Das ergibt für $\lambda = 0$:

$$UA_0^{-1} + A_0 \left. \frac{d}{d\lambda}(A_0 + \lambda U)^{-1} \right|_{\lambda=0} = 0,$$

also

$$h'(A_0)(U) = \left. \frac{d}{d\lambda}(A_0 + \lambda U)^{-1} \right|_{\lambda=0} = -A_0^{-1} U A_0^{-1}.$$

Abschätzung der Kondition von $Ax = b$ bezüglich A

Jetzt können wir die Kondition von $Ax = b$ bezüglich der Matrix A abschätzen. Dazu brauchen wir die Ableitung der Abbildung $g : \mathbb{GL}_n \to \mathbb{R}^n : A \mapsto A^{-1}b$. Nach Obigem ist

$$g'(A)(U) = -A^{-1}UA^{-1}b.$$

Daher ist $\|g'(A)(U)\| = \|A^{-1}UA^{-1}b\| \leq \|A^{-1}\| \, \|A^{-1}b\|$ für $\|U\| = 1$, und das ergibt mit $x = A^{-1}b$:

$$\begin{aligned} \kappa_{\text{abs}} &= \|g'(A)\| \leq \left\|A^{-1}\right\| \|x\|, \\ \kappa_{\text{rel}} &= \frac{\|A\|}{\|x\|} \|g'(A)\| \leq \|A\| \left\|A^{-1}\right\| = \kappa(A). \end{aligned}$$

Hier gilt also dieselbe Schranke für κ_{rel} wie bezüglich des Vektors b.

Beispiel: Wir nehmen $A = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ und sehen uns an, was eine Änderung von a bewirkt. Setzen wir etwa $\tilde{a} = 0.98$, so erhalten wir (gerundet)

$\tilde{x} = \frac{1}{1-0.98^2} \begin{pmatrix} 1 & -0.98 \\ -0.98 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 25.25 \\ -24.75 \end{pmatrix}$.

Mit $a = 0.99$ hatten wir vorhin $x = \begin{pmatrix} 50.25 \\ -49.75 \end{pmatrix}$, also ist $\|x - \tilde{x}\|_1 = 50$.

Andererseits ist $A - \tilde{A} = \begin{pmatrix} 0 & 0.01 \\ 0.01 & 0 \end{pmatrix}$, also $\left\|A - \tilde{A}\right\|_1 = 0.01$.

Es ergeben sich somit die relativen Fehler $\|x - \tilde{x}\|_1 / \|x\|_1 = 50/100 = 0.5$ und $\left\|A - \tilde{A}\right\|_1 / \|A\|_1 = 0.01/1.99 = 1/199$. Das Verhältnis ist hier also gleich der halben Konditionszahl.

Andere Darstellungen der Kondition einer Matrix

Die folgenden Sätze dienen einerseits zu einem besseren Verständnis der Konditionszahl, und andererseits werden wir sie später brauchen.

Satz 3.8 *Sei A eine reguläre $n \times n$-Matrix. Dann gilt*

$$\kappa(A) = \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|}.$$

Die Kondition einer Matrix kann man sich also so vorstellen: Sei E das Bild der Einheitssphäre unter der Abbildung $x \mapsto Ax$ und y_1 ein Punkt aus E mit größtmöglichem Abstand vom Ursprung, y_2 ein Punkt aus E mit kleinstmöglichem Abstand vom Ursprung. Dann ist $\kappa(A)$ gleich dem Verhältnis dieser beiden Abstände, das heißt $\frac{\|y_1\|}{\|y_2\|}$. (Wenn es sich um die 2-Norm handelt, dann ist E für $n = 2$ eine Ellipse und für $n = 3$ ein Ellipsoid.)

Beweis: Wir können die Norm von A auch so schreiben:

$$\|A\| = \max_{x \neq o} \frac{\|Ax\|}{\|x\|},$$

denn $\frac{\|Ax\|}{\|x\|} = \left\| A\left(\frac{1}{\|x\|}x\right)\right\|$ und $\left\|\frac{1}{\|x\|}x\right\| = 1$.

Daher gilt

$$\left\|A^{-1}\right\| = \max_{y \neq o} \frac{\|A^{-1}y\|}{\|y\|} = \max_{x \neq o} \frac{\|x\|}{\|Ax\|} = \frac{1}{\min_{x \neq o} \frac{\|Ax\|}{\|x\|}} = \frac{1}{\min_{\|x\|=1} \|Ax\|},$$

und daraus folgt die Behauptung. ■

Dieser Satz ermöglicht eine natürliche Definition der Kondition auch für nicht-quadratische und nicht-reguläre Matrizen. Wir setzen einfach

$$\kappa(A) := \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|}$$

und vereinbaren $\kappa(A) := \infty$, falls der Nenner gleich Null ist.

Für gewisse Matrizen kann man die Konditionszahl auch durch die Eigenwerte darstellen:

Satz 3.9 *Sei A eine symmetrische positiv definite $n \times n$-Matrix. Dann gilt (bezüglich der 2-Norm):*

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{\max_{\|x\|=1} \left(x^T Ax\right)}{\min_{\|x\|=1} \left(x^T Ax\right)},$$

wobei natürlich $\lambda_{\max}(A)$ bzw. $\lambda_{\min}(A)$ einen größten bzw. kleinsten Eigenwert von A bedeutet.

Beweis: Es gibt in diesem Fall einer Orthonormalbasis $(u_1, \ldots, u_n)$ aus Eigenvektoren von A, und die zugehörigen Eigenwerte $\lambda_1, \ldots, \lambda_n$ sind alle reell und positiv. Jedes $x \in \mathbb{R}^n$ lässt sich so darstellen:

$$x = \xi_1 u_1 + \ldots + \xi_n u_n,$$

und es gilt

$$\|x\|^2 = \xi_1^2 + \ldots + \xi_n^2.$$

Wegen $Au_i = \lambda_i u_i$ gilt dann

$$\begin{aligned} Ax &= \lambda_1\xi_1 u_1 + \ldots + \lambda_n\xi_n u_n, \\ \|Ax\|^2 &= \lambda_1^2\xi_1^2 + \ldots + \lambda_n^2\xi_n^2. \end{aligned}$$

Nehmen wir an, dass die Eigenwerte der Größe nach geordnet sind, also $0 < \lambda_1 \leq \ldots \leq \lambda_n$. Dann gilt für $\|x\| = 1$

$$\lambda_1^2 = \lambda_1^2(\xi_1^2 + \ldots + \xi_n^2) \leq \|Ax\|^2 \leq \lambda_n^2(\xi_1^2 + \ldots + \xi_n^2) = \lambda_n^2$$

und in beiden Ungleichungen kann Gleichheit eintreten (nämlich für $(\xi_1, \ldots, \xi_n) = (1, 0, \ldots, 0)$ bzw. $(0, \ldots, 0, 1)$). Daraus erkennen wir

$$\begin{aligned} \max_{\|x\|=1} \|Ax\|^2 &= \lambda_n^2, \\ \min_{\|x\|=1} \|Ax\|^2 &= \lambda_1^2, \end{aligned}$$

und mit dem vorigen Satz folgt $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.
Die zweite Gleichung folgt ganz ähnlich:

$$x^T Ax = x \cdot (Ax) = \lambda_1\xi_1^2 + \ldots + \lambda_n\xi_n^2,$$

also gilt für $\|x\| = 1$

$$\lambda_1 = \lambda_1(\xi_1^2 + \ldots + \xi_n^2) \leq x^T Ax \leq \lambda_n(\xi_1^2 + \ldots + \xi_n^2) = \lambda_n,$$

und auch hier kann wie oben Gleichheit eintreten. Es folgt

$$\begin{aligned} \max_{\|x\|=1} x^T Ax &= \lambda_n, \\ \min_{\|x\|=1} x^T Ax &= \lambda_1, \end{aligned}$$

und daraus die Behauptung. ■

3.3 Kondition eines Algorithmus

3.3.1 Unterschied zwischen der Kondition eines Algorithmus und der Kondition einer Aufgabe

Im Allgemeinen besteht die Lösung einer Aufgabe aus mehreren Schritten. Die zugehörige Abbildung φ kann dann als Zusammensetzung von Abbildungen dargestellt werden, welche den einzelnen Schritten entsprechen, sagen wir

$\varphi = \psi_n \circ \ldots \circ \psi_1$. Es kann dabei allerdings passieren, dass eine oder mehrere der Teilaufgaben eine wesentlich schlechtere Kondition als die Gesamtaufgabe haben, z.B. wenn einmal zwei annähernd gleiche Zahlen subtrahiert werden.

Wenn ein Algorithmus in dieser Weise als Zusammensetzung von Teilaufgaben beschrieben werden kann, so heißt er *gut konditioniert* oder *stabil*, wenn alle Teilaufgaben gut konditioniert sind. Andernfalls heißt er *schlecht konditioniert* oder *instabil*.

Als *Beispiel* sehen wir uns die Lösung einer quadratischen Gleichung an. Sei

$$x^2 + 2px + q = 0$$

die zu lösende Gleichung, und nehmen wir der Einfachheit halber an, dass wir uns nur für die größere der beiden Wurzeln interessieren. Wir betrachten also

$$\varphi(p, q) = -p + \sqrt{p^2 - q}.$$

Die Jacobimatrix von φ sieht so aus:

$$\varphi'(p, q) = \begin{pmatrix} -1 + \frac{p}{\sqrt{p^2-q}} & \frac{-1}{2\sqrt{p^2-q}} \end{pmatrix}.$$

Die absolute Kondition der Aufgabe ist also offensichtlich schlecht, wenn $q \approx p^2$ ist. In diesem Fall berührt die Parabel mit der Gleichung $y = x^2 + 2px + q$ fast die x-Achse. Es ist daher auch anschaulich klar, dass geringfügige Änderungen der Parabelgleichung große Auswirkungen auf die Lage der Nullstellen haben werden.

Untersuchen wir nun die relative Kondition, und zwar der Einfachheit halber getrennt bezüglich der Abhängigkeit von p und q.

$$\kappa_{\text{rel}}(p) = \left| \frac{p}{-p+\sqrt{p^2-q}} \left(-1 + \frac{p}{\sqrt{p^2-q}} \right) \right| = \left| \frac{p}{-p+\sqrt{p^2-q}} \cdot \frac{p-\sqrt{p^2-q}}{\sqrt{p^2-q}} \right| = \left| \frac{p}{\sqrt{p^2-q}} \right|,$$

$$\kappa_{\text{rel}}(q) = \left| \frac{q}{-p+\sqrt{p^2-q}} \cdot \frac{-1}{2\sqrt{p^2-q}} \right| = \left| \frac{q(p+\sqrt{p^2-q})}{2q\sqrt{p^2-q}} \right| = \left| \frac{p+\sqrt{p^2-q}}{2\sqrt{p^2-q}} \right|.$$

Für $q < 0$ ist $\sqrt{p^2 - q} > |p|$. Daher sind in diesem Fall beide Konditionszahlen < 1, die Aufgabe ist also gut konditioniert, und zwar für alle Werte von p. Trotzdem kann es zu großen Fehlern kommen, je nachdem, wie wir die Lösung berechnen.

Zunächst ist es naheliegend, die Lösung einfach durch Auswertung der Formel $x_1 = -p + \sqrt{p^2 - q}$ zu berechnen. Bei genauerer Betrachtung handelt es sich dabei um Folgendes:

1. Algorithmus zur Berechnung der größeren Nullstelle von $x^2 + 2px + q$*:*

$$\begin{aligned} s &:= p^2, \\ t &:= s - q, \\ u &:= \sqrt{t}, \\ x_1 &:= -p + u. \end{aligned}$$

(Dabei bedeutet ":=" natürlich eine Zuweisung, wie sie in praktisch allen Programmiersprachen vorgesehen ist.)

Die ersten drei Schritte dieses Algorithmus sind für $q < 0$ gut (relativ) konditioniert, da ja nur Wurzelziehen sowie Multiplikation und Addition positiver Zahlen auftreten. (Es ist ja $-q > 0$.) Beim vierten Schritt wird jedoch für $p > 0$ eine Subtraktion positiver Zahlen durchgeführt. Für $p \approx u$, das heißt für $p^2 \gg |q|$, tritt dort Auslöschung ein. Für $p < 0$ ist dagegen auch der vierte Schritt gut konditioniert.

Es gibt jedoch auch eine andere Möglichkeit, die Lösung zu berechnen. Es gilt nämlich

$$-p + \sqrt{p^2 - q} = \frac{-q}{p + \sqrt{p^2 - q}},$$

wie man leicht nachrechnen kann (Multiplikation beider Seiten mit $p+\sqrt{p^2 - q}$). Das führt zum

2. Algorithmus zur Berechnung der größeren Nullstelle von $x^2 + 2px + q$*:*

$$\begin{aligned} s &:= p^2, \\ t &:= s - q, \\ u &:= \sqrt{t}, \\ v &:= p + u, \\ x_1 &:= -q/v. \end{aligned}$$

Die ersten drei Schritte stimmen mit dem 1. Algorithmus überein. Der 4. Schritt ist hier für $p > 0$ gut konditioniert, dagegen für $p \approx -u$ schlecht.

Wir sehen also, dass man auch bei einer gut konditionierten Aufgabe bei der Konstruktion eines Algorithmus darauf achten muss, dass nicht durch einen schlecht konditionierten Schritt insgesamt eine schlechte Kondition entsteht.

3.3.2 Vorwärtsanalyse

Um die Genauigkeit eines bestimmten Ergebnisses abzuschätzen, kann man, wie schon vorhin angedeutet, der Reihe nach für jeden einzelnen Rechenschritt mit Hilfe der Konditionszahlen Fehlerschranken berechnen. (Das ist

in konkreten Fällen oft recht mühsam.) Durch Zusammensetzung erhält man dann Fehlerschranken für das Endergebnis. Diese Vorgangweise heißt *Vorwärtsanalyse*. Man spricht in diesem Zusammenhang auch von *Fehlerfortpflanzung*. Dabei wird allerdings in jedem Schritt das "schlechtest mögliche" angenommen, sodass diese Methode oft zu einer starken Überschätzung des Fehlers führt. Aus diesem Grund verzichten wir hier auf eine ausführlichere Behandlung dieser Methode.

3.3.3 Rückwärtsanalyse

Da die Eingabedaten im Allgemeinen nicht ganz genau sind, ist es oft zweckmäßig, zu überlegen, ob das berechnete Ergebnis vielleicht die exakte Lösung einer Aufgabe ist, die durch eine geringfügige Änderung der Eingabedaten entsteht. Auf diese Art ist es möglich, auch bei schlecht konditionierten Aufgaben zu einer befriedigenden Lösung zu kommen.

Ein typisches und wichtiges Beispiel für die Rückwärtsanalyse tritt bei der Lösung linearer Gleichungssysteme auf. Es erweist sich dabei als günstiger, an Stelle der Norm der Fehler die Absolutbeträge der Fehler der einzelnen Koordinaten (Komponenten) zu betrachten. Wir verwenden dazu folgende Bezeichnungsweise:

Definition 3.10 *Für $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ setzen wir*

$$|x| := (|x_1|, \ldots, |x_n|) = (|x_i|)_{i=1,\ldots,n}.$$

Analog setzen wir für eine $m \times n$-Matrix $A = (a_{ik})$

$$|A| := (|a_{ik}|)_{i=1,\ldots,m;\ k=1,\ldots,n}\,.$$

Bemerkung 3.11 *Ungleichheitszeichen zwischen Matrizen oder Vektoren verstehen wir immer komponentenweise. Zum Beispiel bedeutet*

$$|x - \tilde{x}| \leq y$$

mit $x, \tilde{x}, y \in \mathbb{R}^n$, dass

$$|x_i - \tilde{x}_i| \leq y_i$$

für alle $i \in \{1, \ldots, n\}$.

Wir gehen nun davon aus, dass wir für jedes Matrixelement a_{ik} eine (absolute) Genauigkeitsschranke $\Delta a_{ik} \geq 0$ kennen, das heißt, wir nehmen an, dass der absolute Fehler von a_{ik} nicht größer als Δa_{ik} ist. Analog setzen wir für jede Koordinate b_i von b eine Genauigkeitsschranke Δb_i voraus. Dann gilt der folgende grundlegende Satz:

Satz 3.12 *(Prager und Oettli 1964) Seien $A = (a_{ik})$ und $\Delta A = (\Delta a_{ik})$ reelle $m \times n$-Matrizen mit $\Delta a_{ik} \geq 0$, und $b = (b_i)$, $\Delta b = (\Delta b_i)$ Vektoren aus dem $\mathbb{R}^n$ mit $\Delta b_i \geq 0$. Sei $\tilde{x}$ eine Näherungslösung von $Ax = b$ und $r = A\tilde{x} - b$ das zugehörige Residuum. Wenn*

$$|r| \leq (\Delta A)\,|\tilde{x}| + \Delta b,$$

dann gibt es eine Matrix $\tilde{A}$ und einen Vektor $\tilde{b}$, sodass

$$\begin{aligned} \left|\tilde{A} - A\right| &\leq \Delta A, \\ \left|\tilde{b} - b\right| &\leq \Delta b \end{aligned}$$

und

$$\tilde{A}\tilde{x} = \tilde{b}.$$

Beweis: Wir suchen geeignete Änderungen δa_{ik} und δb_k mit $|\delta a_{ik}| \leq \Delta a_{ik}$ und $|\delta b_k| \leq \Delta b_k$, sodass mit

$$\begin{aligned} \tilde{a}_{ik} &:= a_{ik} + \delta a_{ik} \\ \tilde{b}_i &:= b_i + \delta b_i \end{aligned}$$

das Gleichungssystem $\tilde{A}\tilde{x} = \tilde{b}$ erfüllt ist.

Sei $t := (\Delta A)\,|\tilde{x}| + \Delta b$, das heißt

$$t_i = \sum_k (\Delta a_{ik})\,|\tilde{x}_k| + \Delta b_i.$$

Dann gilt nach Voraussetzung $|r_i| \leq t_i$ (für alle $i \in \{1, \ldots, n\}$).

Wir zeigen nun, dass man die δa_{ik} und δb_k folgenderweise wählen kann:

$$\delta a_{ik} := \begin{cases} -(\text{sign } \tilde{x}_k)\frac{r_i}{t_i}\,\Delta a_{ik} & \text{für } t_i \neq 0, \\ 0 & \text{für } t_i = 0, \end{cases}$$

$$\delta b_i := \begin{cases} \frac{r_i}{t_i}\,\Delta b_i & \text{für } t_i \neq 0, \\ 0 & \text{für } t_i = 0. \end{cases}$$

Wegen $|r_i| \leq t_i$ gilt offensichtlich $|\delta a_{ik}| \leq \Delta a_{ik}$ und $|\delta b_k| \leq \Delta b_k$. Wir haben also nur $\tilde{A}\tilde{x} = \tilde{b}$ nachzuprüfen, das heißt $(A + \delta A)\tilde{x} = b + \delta b$. Das ist gleichbedeutend mit $A\tilde{x} - b = \delta b - (\delta A)\tilde{x}$, das heißt $r = \delta b - (\delta A)\tilde{x}$. Das ist aber tatsächlich der Fall:

Für $t_i \neq 0$ gilt nämlich

$$\delta b_i - \sum_k (\delta a_{ik})\tilde{x}_k = \frac{r_i}{t_i}\left(\Delta b_i + \sum_k (\Delta a_{ik})\,|\tilde{x}_k|\right) = \frac{r_i}{t_i}\,t_i = r_i,$$

und für $t_i = 0$ ist $\delta b_i - \sum_k (\delta a_{ik})\tilde{x}_k = r_i$ trivialerweise richtig, da dann auch $r_i = 0$ sein muss. ■

Beispiel: Wir betrachten wieder $A = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, und nehmen an, dass die Elemente der Matrix bis auf einen absoluten Fehler von 0.005 genau sind, während die Koordinaten von b einen absoluten Fehler bis zu 0.01 enthalten können. (Hier stimmen die absoluten Fehler im Wesentlichen mit den relativen Fehlern überein.)

Angenommen, wir haben auf irgendeine Weise folgende Näherungslösung berechnet: $\tilde{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Um festzustellen, ob diese Lösung akzeptabel ist, berechnen wir zunächst das Residuum

$$r = A\tilde{x} - b = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.01 \\ 0.00 \end{pmatrix}$$

und vergleichen dieses mit der Schranke des Satzes:

$$(\Delta A)\,|\tilde{x}| + \Delta b = \begin{pmatrix} 0.005 & 0.005 \\ 0.005 & 0.005 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0.01 \\ 0.01 \end{pmatrix} = \begin{pmatrix} 0.015 \\ 0.015 \end{pmatrix}.$$

Die Lösung ist also ohne Weiteres akzeptabel, obwohl $\tilde{x}$ von der exakten Lösung in jeder Koordinate um ca. 0.5 abweicht (vgl. das erste Beispiel in Abschnitt 3.2.3).

Obwohl sich der Satz von Prager und Oettli auf die absoluten Fehler bezieht, kann man daraus auch eine Aussage über die relativen Fehler herleiten. Sei ε eine Schranke für die relativen Fehler der Komponenten der Matrix A und des Vektors b. Das heißt

$$\begin{aligned} |\delta a_{ik}| &\leq \varepsilon\,|a_{ik}|\,, \\ |\delta b_i| &\leq \varepsilon\,|b_i|\,. \end{aligned}$$

Dann können wir $\Delta a_{ik} := \varepsilon\,|a_{ik}|$ und $\Delta b_i := \varepsilon\,|b_i|$ setzen und erhalten:

Folgerung 3.13 *Sei $\varepsilon > 0$. Wenn $|r| \leq \varepsilon(|A|\,|\tilde{x}| + |b|)$ ist, dann gibt es eine Matrix $\tilde{A}$ und einen Vektor $\tilde{b}$, deren Komponenten bezüglich A bzw. b einen relativen Fehler $\leq \varepsilon$ aufweisen, sodass $\tilde{A}\tilde{x} = \tilde{b}$.*

Wir können diese Aussage auch so interpretieren:

Folgerung 3.14 *$\tilde{x}$ ist eine akzeptable Lösung von $Ax = b$, wenn wir bei A und b relative Fehler zulassen, welche*

$$\leq \max_i \frac{|r_i|}{(|A|\,|\tilde{x}| + |b|)_i}$$

sind.

Beweis: Sei $\varepsilon = \max_i \frac{|r_i|}{(|A||\tilde{x}|+|b|)_i}$. Dann ist $\frac{|r_i|}{(|A||\tilde{x}|+|b|)_i} \leq \varepsilon$ und daher $|r_i| \leq \varepsilon(|A|\,|\tilde{x}| + |b|)_i$ für alle i, das heißt $|r| \leq \varepsilon(|A|\,|\tilde{x}| + |b|)$. ■

In obigem Beispiel ist die Lösung also auch akzeptabel, wenn wir bei den Eingabedaten, also auch bei b, nur relative Fehler der Größe

$\frac{0.01}{1.99} \approx 0.005$ zulassen, denn hier ist

$$|A|\,|\tilde{x}| + |b| = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1.99 \\ 2.0 \end{pmatrix}.$$

Kapitel 4

Lineare Gleichungssysteme

4.1 Das Eliminationsverfahren

Wir betrachten hier nur Gleichungssysteme der Form $Ax = b$ mit einer regulären (= invertierbaren) $n \times n$-Matrix A und erinnern uns an das bekannte Gauß'sche Eliminationsverfahren[1]: Im ersten Schritt subtrahieren wir das (a_{i1}/a_{11})-fache der ersten Zeile von der i-ten Zeile, für i von 2 bis n. Dadurch entsteht aus A eine Matrix, die wir mit $A^{(1)}$ bezeichnen. Für ihre Elemente gilt:

$$a_{ik}^{(1)} = \begin{cases} a_{ik} & \text{für } i = 1, \\ 0 & \text{für } i \geq 2 \text{ und } k = 1, \\ a_{ik} - (a_{i1}/a_{11})a_{1k} & \text{für } i \geq 2 \text{ und } k \geq 2. \end{cases}$$

Dabei müssen wir natürlich $a_{11} \neq 0$ voraussetzen. Der Vektor b wird ebenso verändert:

$$b_i^{(1)} = \begin{cases} b_i & \text{für } i = 1, \\ b_i - (a_{i1}/a_{11})b_1 & \text{für } i \geq 2. \end{cases}$$

Im nächsten Schritt subtrahieren wir in der Matrix $A^{(1)}$ das $(a_{i2}^{(1)}/a_{22}^{(1)})$-fache der 2. Zeile von der i-ten Zeile, für i von 3 bis n, und erhalten damit die Matrix $A^{(2)}$, usw.. Dabei bleibt die Lösungsmenge des Gleichungssystems unverändert, und wir erhalten schließlich eine Matrix $A^{(n-1)}$, bei der alle Elemente unterhalb der Hauptdiagonalen gleich Null sind. So eine Matrix nennt man eine *rechte (obere) Dreiecksmatrix* und bezeichnet sie oft mit $R = (r_{ik})$. Die rechte Seite des Gleichungssystems muss natürlich auch immer entsprechend verändert werden.

[1] Carl Friedrich Gauß (1777 - 1855, Braunschweig, Göttingen) ist einer der bedeutendsten Mathematiker, der in verschiedensten Richtungen bahnbrechend war (z.B. Zahlentheorie und Differentialgeometrie), siehe [9].

Gleichungssysteme $Rx = b$ mit einer rechten Dreiecksmatrix R sind besonders einfach zu lösen. Sei

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n &= b_1 \\ r_{22}x_2 + \ldots + r_{2n}x_n &= b_2 \\ \ddots \quad \vdots \quad &\vdots \\ r_{nn}x_n &= b_n \end{aligned}$$

so ein Gleichungssystem. Aus der letzten Gleichung erhält man $x_n = b_n/r_{nn}$. Setzt man das in die vorletzte Gleichung ein, kann man leicht x_{n-1} berechnen, usw.. Man nennt das auch *Rückwärtseinsetzen*.

4.1.1 Pivotisierung

Das beschriebene Verfahren funktioniert allerdings nur dann, wenn im s-ten Schritt die Zahl $a_{ss}^{(s-1)}$, durch die wir dividieren, ungleich Null ist. Diese Zahl nennt man das *Pivotelement*. (Das Wort *pivot* bedeutet im Englischen und Französischen "Drehpunkt" oder "Angelpunkt".) Wenn das einmal nicht der Fall ist, so gibt es sicher einen Index $i > s$, sodass $a_{is}^{(s-1)} \neq 0$, denn sonst wäre die Matrix $A^{(s-1)}$ und damit auch A singulär. Durch Vertauschung der Zeilen mit den Nummern i und s erhält man daher ein Pivotelement $\neq 0$.

Bei Berücksichtigung von Rundungsfehlern ist klar, dass es nicht genügt, wenn sich die Pivotelemente nur geringfügig von Null unterscheiden. Da sie durch Subtraktion entstehen, sind sie unter Umständen mit einem großen relativen Fehler behaftet, besonders wenn sie sehr kleinen Betrag haben. Aus diesem Grunde ist es zweckmäßig, in jedem Schritt durch (eventuelle) Zeilenvertauschungen ein Pivotelement mit möglichst großem Betrag zu erzeugen (*Zeilenpivotisierung*). Man wählt also im s-ten Schritt den Index $i_0 \geq s$ so, dass $|a_{i_0 s}| \geq |a_{is}|$ für alle $i \in \{s, \ldots, n\}$ und vertauscht dann die i_0-te Zeile mit der s-ten (falls $i_0 \neq s$).

Noch günstiger, aber aufwändiger ist es, wenn man auch Spaltenvertauschungen verwendet. Man sucht also Indizes i_0 und $k_0 \geq s$, sodass $a_{i_0 k_0}$ größtmöglichen Betrag hat, und vertauscht dann nicht nur die i_0-te Zeile mit der s-ten Zeile, sondern auch die k_0-te Spalte mit der s-ten Spalte (*vollständige Pivotisierung*). Dabei ist allerdings zu berücksichtigen, dass auch die Koordinaten der Lösung entsprechend vertauscht werden müssen.

4.1.2 Zeitkomplexität

Als Maß für die Zeitkomplexität untersucht man bei diesem und ähnlichen Verfahren oft nur die Anzahl der "Punktoperationen", das sind Multiplikationen und Divisionen. Im 1. Schritt brauchen wir $n-1$ Divisionen und $n(n-1)$ Multiplikationen, im 2. Schritt $n-2$ Divisionen und $(n-1)(n-2)$ Multiplikationen, usw.. Die Anzahl der Punktoperationen bis zur Dreiecksform beträgt daher

$$\sum_{i=1}^{n}(i+1)(i-1) = \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \ \sim \ \frac{1}{3}n^3,$$

wobei die Welle bedeutet, dass asymptotische Gleichheit gilt (das heißt, der Quotient zwischen linker und rechter Seite geht für $n \to \infty$ gegen Eins). Beim Rückwärtseinsetzen benötigen wir für die i-te Zeile $n-i$ Multiplikationen und eine Division, also insgesamt

$$\sum_{i=1}^{n}(n-i+1) = \frac{1}{2}n^2 + \frac{1}{2}n \ \sim \ \frac{1}{2}n^2$$

Punktoperationen. Die Gesamtanzahl der Punktoperationen beim Gauß'schen Eliminationsverfahren ist daher

$$\sim \ \frac{1}{3}n^3.$$

Unter Umständen muss man auch den Aufwand für die Pivotisierung in Betracht ziehen. Dabei spielt vor allem die Anzahl der notwendigen Vergleiche eine Rolle. Diese beträgt bei Zeilenpivotisierung $\sum_{i=1}^{n-1}(n-i+1) = \frac{1}{2}n^2 + \frac{1}{2}n - 1 \sim \frac{1}{2}n^2$ und bei vollständiger Pivotisierung $\sum_{i=1}^{n-1}(n-i+1)^2 = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n \sim \frac{1}{3}n^3$.

4.1.3 Lineare Gleichungssysteme mit mehreren rechten Seiten

Es kommt häufig vor, dass man mehrere lineare Gleichungssysteme zu lösen hat, die alle dieselbe Koeffizientenmatrix A haben und sich nur in den rechten Seiten unterscheiden. Diese kann man simultan lösen, indem man die verschiedenen Spaltenvektoren $b_1, \ldots, b_r$ zu einer Matrix $B = (b_1, \ldots, b_r)$ zusammenfasst. Man löst dann eigentlich eine Matrixgleichung der Form

$$AX = B.$$

Das Eliminationsverfahren funktioniert genau so wie vorhin, nur dass man eben mehrere rechte Seiten umzuformen hat. Der Aufwand für die Umformung der Matrix auf Dreiecksform ist genau so groß wie im einfachen Fall. Für den Gesamtaufwand erhält man asymptotisch

$$\frac{1}{3}n^3 + rn^2.$$

Speziell für $B = E$ handelt es sich um die Aufgabe der *Matrixinversion*. Wir sehen also, dass diese asymptotisch

$$\frac{4}{3}n^3$$

Punktoperationen erfordert.

4.2 Lineare Ausgleichsrechnung

4.2.1 Problemstellung

Es geht hier um die Lösung von "überbestimmten" linearen Gleichungssystemen, das sind Gleichungssysteme der Form $Ax = b$ mit $A \in \mathbb{R}^{m,n}$ und $b \in \mathbb{R}^n$, wobei $m > n$ ist. Solche Gleichungssysteme haben im Allgemeinen keine Lösung im strengen Sinn. Man versucht daher, x so zu bestimmen, dass $\|Ax - b\|$ möglichst klein ist, kurz: $\|Ax - b\| = \min$. Dabei kommen natürlich verschiedene Normen in Frage, meistens verwendet man jedoch die euklidische Norm. Das kann man dann auch so interpretieren, dass man versucht, die Summe der Quadrate

$$\sum_{i=1}^{m}(Ax - b)_i^2 = \sum_{i=1}^{m}(a_{i1}x_1 + \ldots + a_{in}x_n - b_i)^2 = \|Ax - b\|_2^2$$

zu minimieren. Diese "Methode der kleinsten Quadrate" geht auf Gauß zurück.

Im Folgenden bezeichnen wir mit $\|.\|$ stets die euklidische Norm.

4.2.2 Äquilibrierung

Die Lösungsmenge eines linearen Gleichungssystems ändert sich nicht, wenn man eine Gleichung mit einem von Null verschiedenen Faktor multipliziert. Die Lösung des Ausgleichsproblems $\|Ax - b\| = \min$ ändert sich auf diese

Weise jedoch schon, da sich ja der entsprechende Summand $(Ax-b)_i^2$ verändert. Man kann durch einen solchen Faktor einer Gleichung mehr oder weniger Gewicht verleihen. Es ist also bei linearen Ausgleichsproblemen immer darauf zu achten, dass man die Gleichungen eventuell mit "richtigen" Faktoren multipliziert. In manchen Fällen ist es sinnvoll, wenn man die Gleichungen so normiert, dass jede Zeile von A die (euklidische) Norm 1 bekommt, in anderen Fällen haben die Differenzen $(Ax-b)_i$ eine unmittelbare Bedeutung, so dass eine Normierung nicht sinnvoll wäre. Beispiele dazu werden am Ende des nächsten Abschnitts diskutiert.

4.2.3 Normalgleichungen

Wir suchen also zu gegebener Matrix $A \in \mathbb{R}^{m,n}$ und gegebenem Vektor $b \in \mathbb{R}^n$ einen Punkt $x \in \mathbb{R}^m$, sodass $\|Ax-b\|$ möglichst klein ist. Das heißt, wir suchen einen Punkt aus dem Bildraum

$$\operatorname{bild} A := \{Ax : x \in \mathbb{R}^n\}$$

mit kleinstmöglichen Abstand von b. Das ist aber nichts anderes wie die orthogonale Projektion von b auf bild A.

Satz 4.1 *Zu $A \in \mathbb{R}^{m,n}$ und $b \in \mathbb{R}^n$ gibt es genau einen Punkt des $\mathbb{R}^m$ mit minimalem Abstand von* bild A*, und dieser kann so berechnet werden: Sei x' eine Lösung des "Normalgleichungssystems"*

$$A^T Ax = A^T b.$$

Dann ist Ax' der gesuchte Punkt, und somit x' eine Lösung des Ausgleichsproblems $\|Ax-b\| = \min$.

Beweis: Sei x' eine Lösung des Normalgleichungssystems, das heißt

$$A^T(Ax'-b) = o.$$

Das ist gleichbedeutend mit $x \cdot A^T(Ax'-b) = 0$ für alle $x \in \mathbb{R}^n$.

Das können wir auch so schreiben: $(Ax)\cdot(Ax'-b) = 0$ für alle $x \in \mathbb{R}^n$, also $b-Ax'$ orthogonal zu bild A.

Sei nun y ein von Ax' verschiedener Punkt aus bild A. Dann gilt nach dem "Satz von Pythagoras":

$$\|b - y\|^2 = \|b - Ax'\|^2 + \|Ax' - y\|^2 ,$$

und es folgt $\|b - y\| > \|b - Ax'\|$. Daher ist Ax' der eindeutige Punkt aus bild A mit minimalem Abstand von b. ■

Im Allgemeinen ist die Lösung des Normalgleichungssystems nicht eindeutig. Es gilt aber:

Bemerkung 4.2 *Sei $A \in \mathbb{R}^{m,n}$ mit $m \geq n$. Wenn A den höchstmöglichen Rang n hat, dann ist $A^T A$ regulär, und das Ausgleichsproblem $\|Ax - b\| =$ min hat daher eine eindeutige Lösung.*

Das ergibt sich unmittelbar aus folgendem Satz, da $A^T A$ eine $n \times n$-Matrix ist.

Satz 4.3 *Sei $A \in \mathbb{R}^{m,n}$ mit beliebigen $m, n \in \mathbb{N}$. Dann gilt*

$$\operatorname{rang} A = \operatorname{rang}(A^T A).$$

Beweis:

Nach der Dimensionsformel für lineare Abbildungen gilt

$$n = \operatorname{rang} A + \dim(\operatorname{kern} A)$$

und

$$n = \operatorname{rang} A^T A + \dim(\operatorname{kern} A^T A).$$

Also ist

$$\operatorname{rang} A = \operatorname{rang} A^T A + \dim(\operatorname{kern} A^T A) - \dim(\operatorname{kern} A).$$

Es genügt daher zu zeigen, dass $\operatorname{kern} A = \operatorname{kern} A^T A$ ist.

Sei $x \in \operatorname{kern} A$. Dann ist $Ax = o$ und daher auch $A^T Ax = o$. Es folgt $x \in \operatorname{kern} A^T A$.

Sei umgekehrt $x \in \operatorname{kern} A^T A$. Dann ist $A^T Ax = o$ und daher $x^T A^T Ax = (Ax) \cdot (Ax) = 0$. Es folgt $Ax = o$ und somit $x \in \operatorname{kern} A$. ■

Kondition der Normalgleichungen

Die Lösung eines linearen Ausgleichsproblems mit den Normalgleichungen ist zwar recht bequem, aber insoferne problematisch, als die Normalgleichungen häufig schlecht konditioniert sind. Das hängt mit folgendem Satz zusammen.

Satz 4.4 *Sei A eine $m \times n$-Matrix mit $m \geq n$ und* $\operatorname{rang} A = n$. *Dann gilt bezüglich der* 2*-Norm*

$$\kappa(A^T A) = \kappa(A)^2.$$

Beweis: Wir benützen hier die Definition der Kondition für beliebige Matrizen (siehe Bemerkung nach Satz 3.8) und den Satz 3.9 (die Matrix $A^T A$ ist ja symmetrisch und positiv definit, siehe Beweis von Satz **??**):

$$\kappa(A)^2 = \frac{\max \|Ax\|^2}{\min \|Ax\|^2} = \frac{\max(Ax) \cdot (Ax)}{\min(Ax) \cdot (Ax)} = \frac{\max \left(x \cdot A^T Ax\right)}{\min \left(x \cdot A^T Ax\right)} = \kappa(A^T A),$$

wobei Maximum und Minimum jeweils über alle x mit $\|x\| = 1$ genommen werden. ■

Eine genauere Analyse (vgl. [3]) zeigt, dass die Methode der Normalgleichungen insbesondere für kleine Residuen nicht empfehlenswert ist. Daher wird in Abschnitt 4.2.4 eine alternative Methode diskutiert.

Beispiele:

1. Ausgleichung im Vermessungswesen Zur Bestimmung der Koordinaten eines Punktes kann man so vorgehen, dass man mehrere (d.h. drei oder mehr) Geraden vermisst, die theoretisch durch diesen Punkt hindurchgehen. Infolge der unvermeidlichen Mess- und Rundungsfehler gehen diese Geraden jedoch nicht genau durch einen Punkt. Man nimmt dann denjenigen Punkt, für den die Summe der Quadrate der Abstände von den Geraden minimal ist. Um das zu erreichen, betrachtet man die Hesse'sche Normalform der Geradengleichungen ("Äquilibrierung", vgl. Abschnitt 4.2.2). Das sind Gleichungen der Form $a_1 x_1 + a_2 x_2 = b$ mit $a_1^2 + a_2^2 = 1$. Diese haben nämlich die Eigenschaft, dass für jeden Punkt (x_1, x_2) die Differenz zwischen linker und rechter Seite bis auf das Vorzeichen gleich dem Abstand dieses Punktes von der Geraden ist.

2. Kurvenanpassung Hier geht es darum, zu m durch Messung erhaltenen Zahlenpaaren $(x_1, y_1), \ldots, (x_m, y_m)$ eine Funktion f zu finden, die sich als Linearkombination von gewissen einfachen Funktionen (z.B. $\sin, \cos$ oder $1, x, x^2$) darstellen lässt und möglichst gut den Messwerten entspricht, das heißt, dass der Graph der Funktion f möglichst durch die Punkte (x_i, y_i) durchgeht.

Seien $g_1, \ldots, g_n$ die Funktionen, welche eine Basis des Vektorraums der in Betracht gezogenen Funktionen bilden. Dann suchen wir reelle Koeffizienten $\alpha_1, \ldots, \alpha_n$ so, dass die Summe der Quadrate der Abweichungen von den Messwerten minimal wird, das heißt

$$\sum_{i=1}^{m} \left(\sum_{k=1}^{n} \alpha_k g_k(x_i) - y_i \right)^2 = \min .$$

Auch hier handelt es sich um die Lösung eines überbestimmten linearen Gleichungssystems, falls $m > n$ ist:

$$\begin{aligned} g_1(x_1)\alpha_1 + \ldots + g_n(x_1)\alpha_n &= y_1 \\ &\vdots \\ g_1(x_m)\alpha_1 + \ldots + g_n(x_m)\alpha_n &= y_m \end{aligned}$$

Dabei sind $\alpha_1, \ldots, \alpha_n$ die Unbekannten. (Eine Äquilibrierung wie vorhin wäre hier nicht sinnvoll.)

Wählt man als Basisfunktionen 1 und x, dann sieht die Matrix so aus:

$$A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} .$$

Es handelt sich hier um die Berechnung einer *Ausgleichsgeraden,* und wir erhalten für die Matrix des Normalgleichungssystems

$$A^T A = \begin{pmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} .$$

Der Vektor der rechten Seite ist

$$A^T y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} .$$

Sehen wir uns z.B. die Kondition (bezüglich der 1-Norm) für den Fall $x_i = i$ an. Hier ist

$$A^TA = \begin{pmatrix} m & \frac{1}{2}m(m+1) \\ \frac{1}{2}m(m+1) & \frac{1}{6}m(m+1)(2m+1) \end{pmatrix},$$

$$\left\|A^TA\right\|_1 = \tfrac{1}{2}m(m+1) + \tfrac{1}{6}m(m+1)(2m+1) = \tfrac{2}{3}m + m^2 + \tfrac{1}{3}m^3 \sim \tfrac{1}{3}m^3,$$

$$\left(A^TA\right)^{-1} = \tfrac{2}{m(m^2-1)} \begin{pmatrix} 1+3m+2m^2 & -3-3m \\ -3-3m & 6 \end{pmatrix},$$

$$\left\|\left(A^TA\right)^{-1}\right\|_1 = \tfrac{2}{m(m^2-1)}(4+6m+2m^2) \sim 4m^{-1},$$

also $\kappa\left(A^TA\right) \sim \frac{4}{3}m^2$. Die Kondition ist somit ziemlich schlecht.

4.2.4 QR-Zerlegung

Allgemeines

Die im Folgenden dargestellte Methode eignet sich sowohl zur Lösung von gewöhnlichen als auch von überbestimmten linearen Gleichungssystemen. Die Kondition ist deutlich besser als bei den Normalgleichungen, und bezüglich der Rundungsfehler verhält sich diese Methode im Allgemeinen besser als das Gauß'sche Eliminationsverfahren.

Der Grundgedanke besteht darin, das betrachtete Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m,n}$, $m \geq n$, durch eine *orthogonale Transformation* so umzuformen, dass eine (rechte) Dreiecksmatrix entsteht. Genauer heißt das: Wir suchen eine orthogonale Matrix $Q \in \mathbb{R}^{m,m}$, sodass

$$QA = \begin{pmatrix} R \\ O \end{pmatrix}$$

mit einer rechten Dreiecksmatrix $R \in \mathbb{R}^{n,n}$ und der Nullmatrix $O \in \mathbb{R}^{m-n,n}$.

Die Methode beruht vor allem darauf, dass die 2-Norm jedes Vektors bei Multiplikation mit einer orthogonalen Matrix unverändert bleibt. In diesem Zusammenhang sei auch daran erinnert, dass die Konditionszahl orthogonaler Matrizen gleich Eins ist (siehe Bemerkung 3.6).

Nehmen wir einmal an, wir hätten eine solche Matrix Q gefunden. Dann multiplizieren wir auch die rechte Seite b des Gleichungssystems mit Q und bezeichnen die ersten n Koordinaten von Qb mit c und die restlichen mit d:

$$Qb = \begin{pmatrix} c \\ d \end{pmatrix}$$

mit $c \in \mathbb{R}^n$, $d \in \mathbb{R}^{m-n}$. Dann gilt für beliebiges $x \in \mathbb{R}^n$ bezüglich der 2-Norm:

$$\|Ax-b\|^2 = \|Q(Ax-b)\|^2 = \|QAx-Qb\|^2 = \left\| \begin{pmatrix} Rx-c \\ -d \end{pmatrix} \right\|^2 =$$

$$= \|Rx-c\|^2 + \|d\|^2 \geq \|d\|^2$$

mit Gleichheit genau dann, wenn $Rx = c$. Das heißt, wenn wir das Gleichungssystem $Rx = c$ lösen, erhalten wir eine Lösung des Ausgleichsproblems $\|Ax-b\| = \min$. Da dann A in der Form

$$A = Q^T \begin{pmatrix} R \\ O \end{pmatrix}$$

dargestellt werden kann, nennt man eine derartige Methode auch QR-Methode. Die Matrix $\begin{pmatrix} R \\ O \end{pmatrix}$ nennt man dann auch eine rechte (obere) Dreiecksmatrix.

Es gibt verschiedene Methoden, zu einer gegebenen Matrix A so eine orthogonale Matrix Q zu finden. Zur Konstruktion von Q bieten sich in natürlicher Weise Drehungen und Spiegelungen an. Drehungen werden bei den sogenannten *Givens-Rotationen* verwendet (Givens[2] 1953). Hier wird aber nur eine Spiegelungsmethode besprochen, die auf Householder[3] (1958) zurückgeht.

Householder-Spiegelungen

Wir überlegen uns zunächst, wie man die Spiegelung an einer durch den Ursprung gehenden Hyperebene mit Normalvektor $v \neq o$ im $\mathbb{R}^m$ beschreiben kann. Wir fassen dabei, wie üblich, v als einspaltige Matrix auf. Dann ist $v^T v = v \cdot v = \|v\|^2 \in \mathbb{R}$ und

$$v\, v^T = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \begin{pmatrix} v_1 & v_2 & \cdots & v_m \end{pmatrix} = \begin{pmatrix} v_1^2 & v_1v_2 & \cdots & v_1v_m \\ v_2v_1 & v_2^2 & \cdots & v_2v_m \\ \vdots & \vdots & & \vdots \\ v_mv_1 & v_mv_2 & \cdots & v_m^2 \end{pmatrix} \in \mathbb{R}^{m,m}.$$

(Wir bezeichnen hier mit $\|.\|$ immer die 2-Norm.)

Hilfssatz 4.5 *Die Matrix*

$$Q := E - \frac{2}{v^T v}(v\, v^T)$$

[2] James Wallace **Givens** (1910 - 1993, USA) entwickelte diese Methode zur Berechnung von Eigenwerten symmetrischer Matrizen (siehe Notices AMS 41 (1994), 29 - 33).

[3] Alston **Householder** (1904 - 1993, USA) ist vor allem durch seine Arbeiten über Matrizen in der numerischen Mathematik bekannt (siehe [9]). Die von ihm eingeführten Spiegelungen eignen sich ebenfalls zur Berechnung von Eigenwerten symmetrischer Matrizen.

beschreibt die Spiegelung an der Hyperebene H_v durch o mit Normalvektor v.

Beweis: Es ist nur zu zeigen, dass $Qv = -v$ und $Qx = x$ für alle $x \in H_v$.
a) $Qv = v - \frac{2}{v^T v}(v\ v^T)v = v - \frac{2}{v^T v}v(v^T v) = v - 2v = -v$.
b) Sei $x \in H_v$. Dann ist $v \cdot x = v^T x = 0$ und daher $Qx = x - \frac{2}{v^T v}(v\ v^T)x = x - \frac{2}{v^T v}v(v^T x) = x$. ■

Wir versuchen nun in einem ersten Schritt, eine Spiegelung Q_1 zu finden, die die erste Spalte a_1 von A auf ein Vielfaches des ersten kanonischen Basisvektors e_1 abbildet, das heißt

$$Q_1 a_1 = \lambda_1 e_1 = \begin{pmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

mit einer reellen Zahl λ_1. Da Q_1 die Norm unverändert lässt, muss jedenfalls $|\lambda_1| = \|a_1\|$ sein, das heißt für λ_1 kommt nur $\pm\|a_1\|$ in Frage. Wenn bei einer Spiegelung an einer Hyperebene H (durch o) der Punkt a_1 auf den Punkt $\lambda_1 e_1$ abgebildet wird, dann ist zumindest anschaulich klar, dass der Verbindungsvektor dieser beiden Punkte senkrecht zu H stehen muss, das heißt $a_1 - \lambda_1 e_1$ muss ein Normalvektor von H sein. Das wird im Folgenden noch genauer ausgeführt.

Hilfssatz 4.6 *Seien $a, b \in \mathbb{R}^m$, $\|a\| = \|b\| \neq 0$. Wenn $a \neq b$ ist, dann bildet die Spiegelung an der Hyperebene durch o mit Normalvektor $a - b$ den Punkt a auf den Punkt b ab.*

Beweis: Sei $Q = E - \frac{2}{v^T v}(v\ v^T)$ mit $v = a - b \neq o$.
Betrachen wir den Mittelpunkt p von a und b, also

$$p := \frac{1}{2}(a + b),$$

und berechnen wir das Q-Bild von p:
Zunächst ist $Qp = p - \frac{2}{v^T v}(v\ v^T)p = p - \frac{2}{v^T v}v(v^T p)$.
Nun ist aber $v^T p = p \cdot v = \frac{1}{2}(a + b) \cdot (a - b) = \frac{1}{2}(\|a\|^2 - \|b\|^2) = 0$, also $Qp = p$.
Zur Berechnung von Qa stellen wir nun a durch p und v dar: Aus $a + b = 2p$ und $a - b = v$ ergibt sich durch Addition $2a = 2p + v$, also $a = p + \frac{1}{2}v$. Wir erhalten dann wegen $Qv = -v$:

$Qa = Qp + \frac{1}{2}Qv = p - \frac{1}{2}v = \frac{1}{2}(a+b) - \frac{1}{2}(a-b) = b$,w.z.z.w. ■

Wir verwenden diesen Hilfssatz jetzt mit $a = a_1$, $b = \lambda_1 e_1$ und $\lambda_1 = \pm \|a_1\|$. Es ist noch zu überlegen, welche der beiden Möglichkeiten für λ_1 wir nehmen sollen. Wenn λ_1 dasselbe Vorzeichen wie a_{11} hat, dann besteht die Gefahr einer Auslöschung bei der Berechnung der ersten Koordinate von $a_1 - \lambda_1 e_1$. Wir wählen daher das Vorzeichen von λ_1 so, dass

$$\operatorname{sign} \lambda_1 = -\operatorname{sign} a_{11},$$

falls $a_{11} \neq 0$. Wenn $a_{11} = 0$, können wir willkürlich etwa $\operatorname{sign} \lambda_1 = +1$ wählen. Auf diese Weise erreichen wir auch, dass $a_1 - \lambda_1 e_1$ sicher $\neq o$ ist, solange nur $a_1 \neq o$.

Mit diesen Vereinbarungen setzen wir also

$$Q_1 := E - \frac{2}{v_1^T v_1}(v_1\ v_1^T) \quad \text{mit } v_1 := a_1 - \lambda_1 e_1$$

und erhalten

$$Q_1 A = \left(\begin{array}{c|ccc} \lambda_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|c} \lambda_1 & * \\ \hline o & A_2 \end{array}\right),$$

wobei die Sterne bedeuten, dass dort irgendwelche Zahlen stehen, die uns nicht weiter interessieren. o bezeichnet hier den Nullvektor des $\mathbb{R}^{m-1}$ (als Spalte aufgefasst), und A_2 die Matrix, die durch Weglassung der ersten Zeile und ersten Spalte von $Q_1 A$ entsteht.

In derselben Weise können wir nun eine Spiegelung Q_2' finden, sodass

$$Q_2' A_2 = \left(\begin{array}{c|c} \lambda_2 & * \\ \hline o & A_3 \end{array}\right).$$

Setzen wir dann

$$Q_2 := \left(\begin{array}{c|c} 1 & o^T \\ \hline o & Q_2' \end{array}\right),$$

so gilt

$$Q_2 Q_1 A = \left(\begin{array}{c|c} 1 & o^T \\ \hline o & Q_2' \end{array}\right)\left(\begin{array}{c|c} \lambda_1 & * \\ \hline o & A_2 \end{array}\right) = \left(\begin{array}{cc|ccc} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{array}\right).$$

Wenn wir auf diese Weise fortfahren, erhalten wir orthogonale Matrizen $Q_1, \ldots, Q_n$, sodass

$$Q_n \cdots Q_1 A = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \lambda_n \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} =: R \in \mathbb{R}^{m,n}.$$

Mit $Q := Q_n \cdots Q_1$ gilt also $QA = R$, und das ist gleichbedeutend mit $A = Q^T R$, denn das Produkt orthogonaler Matrizen ist wieder orthogonal, und für solche Matrizen ist die Inverse gleich der Transponierten. Q^T ist natürlich auch orthogonal, und wir haben somit folgenden Satz bewiesen.

Satz 4.7 *Zu jeder Matrix $A \in \mathbb{R}^{m,n}$ mit $m \geq n$ gibt es eine orthogonale Matrix $Q \in \mathbb{R}^{m,m}$ und eine rechte Dreiecksmatrix $R \in \mathbb{R}^{m,n}$, sodass*

$$A = QR.$$

Das Paar (Q, R) heißt dann eine QR-Zerlegung von A.

Bemerkung 4.8 *Die QR-Zerlegung ist im Allgemeinen nicht eindeutig. Man kann jedoch zeigen, dass sie für reguläre quadratische Matrizen eindeutig bestimmt ist, wenn man die Vorzeichen der Diagonalelemente von R vorschreibt.*

Wenn man also eine QR-Zerlegung von A kennt, dann kann man das Ausgleichsproblem $\|Ax - b\| = \min$ folgendermaßen lösen (vgl. Abschnitt "Allgemeines"):

1. Berechne $Q^T b$ und bezeichne die ersten n Koordinaten dieses Vektors mit c.

2. Bezeichne die aus den ersten n Zeilen von R bestehende Matrix mit R'.

3. Löse das Gleichungssystem $R'x = c$ durch Rückwärtseinsetzen.

Bemerkung 4.9 *Man braucht also eigentlich nur die ersten n Zeilen von Q^T bzw. die ersten n Spalten von Q.*

Bemerkung 4.10 *Sei Q' die $m \times n$-Matrix, die aus den ersten n Spalten von Q besteht, und R' die Dreiecksmatrix, die aus den ersten n Zeilen von R besteht. Dann gilt*

$$A = Q'R'.$$

Kapitel 5

Interpolation

5.1 Problemstellung

Nehmen wir an, dass wir von einer (reellen) Funktion f nur die Funktionswerte an n Stellen $x_1, \ldots, x_n$ kennen:

$$f(x_i) = y_i \quad \text{für alle } i \in \{1, \ldots, n\}.$$

Wie kann man auf Grund dieser Daten näherungsweise Funktionswerte an beliebigen Stellen x berechnen? Dieses Problem tritt z.B. auf, wenn die Funktionswerte durch Messungen ermittelt wurden oder wenn die direkte Berechnung der Funktionswerte sehr aufwändig ist. Darüber hinaus spielt die Interpolation bei der numerischen Differenziation und Integration und auch in der Computergraphik eine wichtige Rolle.

Die Stellen x_i heißen *Stützstellen*. Die im Folgenden besprochenen Interpolationsmethoden eignen sich vor allem für die Bestimmung des Funktionswerts an Stellen x, die zwischen der kleinsten und größten Stützstelle liegen. Wenn x außerhalb dieses Intervalls liegt, spricht man auch von *Extrapolation*.

Der Grundgedanke der Interpolation besteht darin, eine möglichst einfache Funktion zu finden, die an den Stützstellen genau die gegebenen Funktionswerte hat, und von der zu hoffen ist, dass sie die betrachtete Funktion gut approximiert.

Wir werden im Folgenden als "möglichst einfache" Funktion eine Polynomfunktion mit möglichst niedrigem Grad nehmen. Es gibt aber auch andere sinnvolle Möglichkeiten, z.B. trigonometrische Polynome oder Funktionen, die sich stückweise durch Polynome darstellen lassen (sogenannte *Splines*, siehe insbesondere Vorlesungen oder Bücher über Computergeometrie).

Bei Polynomen über beliebigen Körpern wird oft zwischen Polynomen und Polynomfunktionen unterschieden, weil es passieren kann, dass zwei verschiedene Polynome (d.h. Polynome mit verschiedenen Koeffizienten) dieselbe Funktion liefern. Über dem Körper der reellen Zahlen ist so eine Unterscheidung aber nicht notwendig, da in diesem Fall Polynome und Polynomfunktionen einander umkehrbar eindeutig entsprechen. Das heißt, wenn zwei reelle Polynome in allen Funktionwerten übereinstimmen, dann müssen sie auch dieselben Koeffizienten haben.

5.2 Existenz und Eindeutigkeit des Interpolationspolynoms

Die Interpolation durch Polynomfunktionen beruht auf dem folgenden grundlegenden Satz.

Satz 5.1 *Seien $x_1, \ldots, x_n$ (paarweise) verschiedene reelle Zahlen und $y_1, \ldots, y_n$ beliebige reelle Zahlen. Dann gibt es genau ein Polynom p mit Grad $\leq n-1$, sodass*

$$p(x_i) = y_i \quad \textit{für alle } i \in \{1, \ldots, n\}.$$

Beweis: Das gesuchte Polynom hat die Form

$$p(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}.$$

Wir suchen also Zahlen $a_0, \ldots, a_{n-1}$, sodass

$$a_0 + a_1 x_i + \ldots + a_{n-1} x_i^{n-1} = y_i$$

für alle $i \in \{1, \ldots, n\}$. Das ist ein lineares Gleichungssystem in den Unbekannten a_i mit der Matrix

$$\begin{pmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{pmatrix}.$$

Die Determinante dieser Matrix ist nach Vandermonde[1] benannt und hat bekanntlich den Wert

$$\prod_{1 \leq i < k \leq n} (x_k - x_i).$$

[1] In den Arbeiten von Alexandre-Théophile Vandermonde (1735 - 1796, Paris) kommt diese Determinante merkwürdigerweise gar nicht vor (siehe [9]).

Sie ist also jedenfalls ungleich Null, und daher hat das Gleichungssystem genau eine Lösung. ■

Das gemäß diesem Satz eindeutig bestimmte Polynom heißt das *Interpolationspolynom* zu den Punkten $(x_1, y_1), \ldots, (x_n, y_n)$. (Diese Punkte nennen wir auch *Stützpunkte*.)

5.3 Fehlerabschätzung

Die Differenz zwischen der interpolierten Funktion f und dem Interpolationspolynom p kann zwischen den Stützstellen im Prinzip beliebig groß sein. Wenn man jedoch eine Schranke für die n-te Ableitung von f kennt bzw. annimmt, so ist eine Fehlerabschätzung möglich. Es geht dabei nur um den *Verfahrensfehler*, die Rundungsfehler lassen wir hier unberücksichtigt.

Zur Abschätzung des Fehlers verwendet man das folgende *Stützstellenpolynom:*

$$\omega(x) := (x - x_1)(x - x_2) \cdots (x - x_n).$$

Wir nehmen im Folgenden immer an, dass f stetig und auf einem Intervall I definiert ist. Mit $\|f\|$ bezeichnen wir dann die Maximums-Norm von f, das heißt

$$\|f\| := \max_{x \in I} |f(x)| .$$

Satz 5.2 *Seien $x_1, \ldots, x_n$ (paarweise) verschiedene Punkte in dem Intervall $I = [a, b]$, f sei eine in I n-mal stetig differenzierbare Funktion und p das Interpolationspolynom zu den Punkten $(x_i, f(x_i))$. Dann gibt es zu jedem Punkt $x \in I$ einen Punkt $\xi \in I$, sodass*

$$f(x) - p(x) = \frac{1}{n!}\ f^{(n)}(\xi)\ \omega(x). \tag{5.1}$$

Daher gilt

$$|f(x) - p(x)| \leq \frac{1}{n!}\ \left\|f^{(n)}\right\|\ \ |\omega(x)| \tag{5.2}$$

und

$$\|f - p\| \leq \frac{1}{n!}\ \left\|f^{(n)}\right\|\ \ \|\omega\| . \tag{5.3}$$

Beweis: Wir können natürlich annehmen, dass x mit keiner der Stützstellen übereinstimmt, da sonst in (5.1) auf beiden Seiten Null steht. Wir denken uns nun x festgehalten und betrachten die Funktion

$$g(z) := f(z) - p(z) - (f(x) - p(x))\frac{\omega(z)}{\omega(x)}.$$

$g(x_k) = 0$ für alle $k \in \{1, \ldots, n\}$, und $g(x) = 0$. Die Funktion g hat also mindestens $n + 1$ verschiedene Nullstellen in I. Nach dem Satz von Rolle liegt zwischen je zwei solchen Nullstellen eine Nullstelle der Ableitung g'. Daher hat g' mindestens n Nullstellen in I. Analog fortfahrend sehen wir, dass $g^{(n)}$ mindestens eine Nullstelle in I hat. Sei ξ eine solche Nullstelle, also $g^{(n)}(\xi) = 0$.

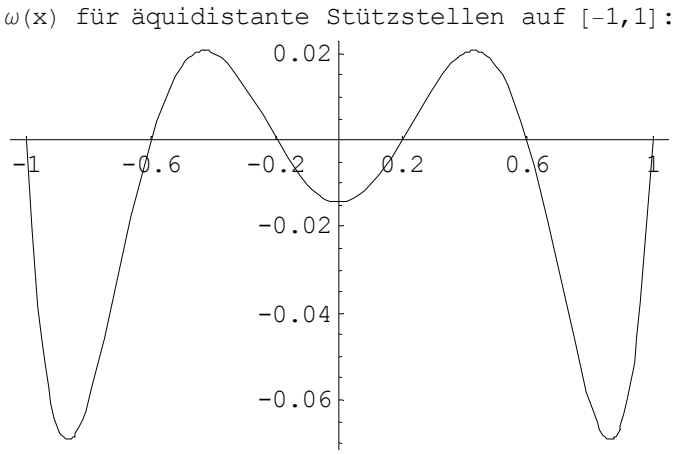
$$g^{(n)}(z) = f^{(n)}(z) - p^{(n)}(z) - (f(x) - p(x))\frac{\omega^{(n)}(z)}{\omega(x)} = f^{(n)}(z) - (f(x) - p(x))\frac{n!}{\omega(x)},$$

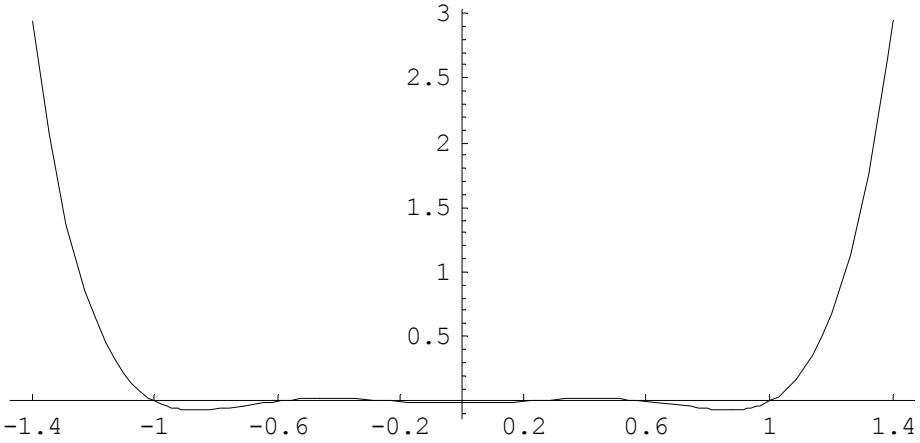
denn $\omega^{(n)}(z) = n!$, da $\omega(z)$ ein Polynom der Form $z^n + c_{n-1}z^{n-1} + \ldots + c_0$ ist, und $p^{(n)}(z) = 0$, da der Grad von p kleiner als n ist. Es folgt:

$$0 = g^{(n)}(\xi) = f^{(n)}(\xi) - (f(x) - p(x))\frac{n!}{\omega(x)},$$

und das ergibt die Beziehung (5.1). ■

Der Interpolationsfehler hängt also sehr von $\omega(x)$ ab. Betrachten wir z.B. äquidistante Stützstellen auf einem Intervall $I = [a, b]$, sodass a und b mit der ersten bzw. letzten Stützstelle übereinstimmen. In diesem Fall ist $|\omega(x)|$ am kleinsten in der Nähe der mittleren Stützstellen, dagegen sehr groß in der Nähe des Randes von I, und noch größer außerhalb von I. Die folgenden zwei Graphiken zeigen $\omega(x)$ für $n = 6$ und $I = [-1, 1]$, einmal für $-1 \leq x \leq 1$ und das zweite Mal für $-1.4 \leq x \leq 1.4$.

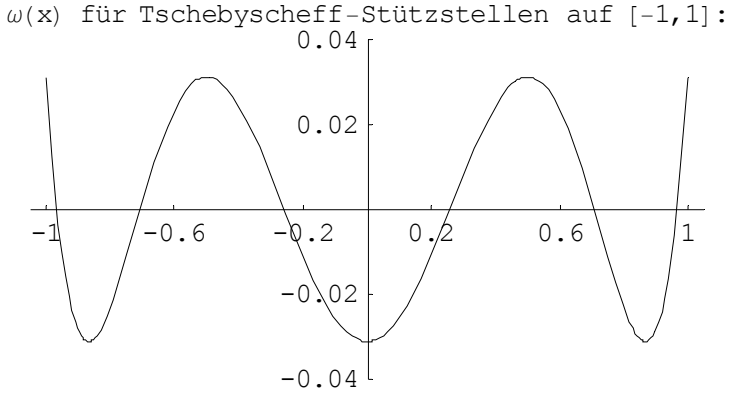




Bei größeren Werten von n eignen sich also äquidistante Stützstellen am ehesten, wenn nur im mittleren Bereich interpoliert werden soll. Will man über das ganze Intervall eine möglichst gleichmäßige Approximation, so verteilt man oft die Stützstellen so, dass sie am Rande enger liegen. In gewissem Sinne optimal ist für das Intervall $[-1, 1]$ die folgende Wahl:

$$x_k = \cos \frac{(k - \frac{1}{2})\pi}{n}.$$

(Das sind die Nullstellen der sogenannten Tschebyscheff[2]-Polynome, siehe z.B. [3].) Bei dieser Wahl haben die lokalen Extremwerte von $\omega(x)$ alle denselben Betrag. In der Graphik ist wieder der Fall $n = 6$ dargestellt.



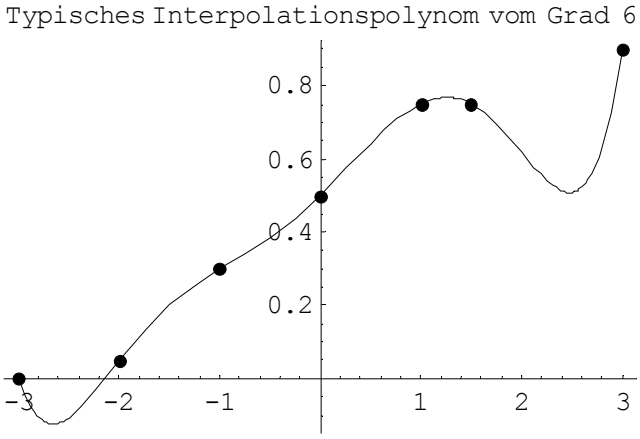
Für diese Stützstellen gilt $\|\omega\| = \frac{1}{2^{n-1}}$, und das ist das Minimum von $\|\omega\|$ über alle n-Tupel von Stützstellen im Intervall $[-1, 1]$ (siehe [3]). Wir bekommen

[2]Pafnuty Lwowitsch Tschebyscheff (= Chebyshev) (1821 - 1894, Russland) ist durch seine Beiträge zur Approximations- und Wahrscheinlichkeitstheorie bekannt (siehe [9]). Unter "Tschebyscheff-Approximation" versteht man die Approximation von Funktionen bezüglich der Supremumsnorm.

damit folgende Fehlerabschätzung:

$$\|f - p\| \leq \frac{1}{n! 2^{n-1}} \left\| f^{(n)} \right\|.$$

Generell ist zu sagen, dass Interpolationspolynome, vor allem bei höherem Grad, sehr stark zur "Welligkeit" neigen, wie etwa bei folgendem Beispiel.



Durch Erhöhung des Grades kann im Allgemeinen die Güte der Interpolation nur in einem kleinen Bereich verbessert werden. Lässt man die Anzahl der Stützstellen gegen unendlich gehen, erhält man eine Folge von Interpolationspolynomen, die möglicherweise nicht gegen die interpolierte Funktion konvergiert. Das hängt damit zusammen, dass $\left\| f^{(n)} \right\|$ unter Umständen sehr schnell gegen unendlich gehen kann.

5.4 Berechnung des Interpolationspolynoms

Im Prinzip kann das Interpolationspolynom durch Lösung eines linearen Gleichungssystems bestimmt werden (siehe Beweis von Satz 5.1). Es gibt aber auch direktere Methoden, die sich für verschiedene Zwecke als günstiger erweisen.

5.4.1 Die Lagrange'sche Form des Interpolationspolynoms

Zur Berechnung des Interpolationspolynoms zu den Punkten $(x_1, y_1), \ldots, (x_n, y_n)$ kann man die folgenderweise definierten *Lagrange-Polynome*[3] verwenden (für $k \in \{1, \ldots, n\}$):

$$L_k(x) := \prod_{\substack{i=1 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}.$$

Das sind offensichtlich Polynome vom Grad $n-1$ mit folgender Eigenschaft:

$$L_k(x_j) = \delta_{jk},$$

denn für $j = k$ sind hier alle Faktoren gleich Eins, und für $j \neq k$ ist der j-te Faktor gleich Null. Setzen wir nun

$$p(x) := \sum_{k=1}^{n} y_k \; L_k(x), \tag{5.4}$$

so ist das ein Polynom vom Grad $\leq n-1$, und

$$p(x_j) = \sum_{k=1}^{n} y_k \delta_{jk} = y_j,$$

also ist $p(x)$ das gesuchte Polynom. Diese Form des Interpolationspolynoms ist besonders elegant und vor allem für theoretische Zwecke nützlich (z.B. für die Herleitung der Newton-Cotes-Formeln in Kapitel 7.1). Die Lagrange-Polynome $L_k(x)$ nennt man hier auch *Lagrange-Koeffizienten*, da sie ja als Koeffizienten der y_k aufgefasst werden können.

1. Beispiel: Im einfachsten Fall $n = 2$ erhalten wir eine Formel für *lineare Interpolation*:

$$p(x) = y_1 \frac{x - x_2}{x_1 - x_2} + y_2 \frac{x - x_1}{x_2 - x_1}.$$

Das können wir auch so schreiben:

$$p(x) = \frac{(x_2 - x)y_1 + (x - x_1)y_2}{x_2 - x_1}. \tag{5.5}$$

Diese Funktion stellt natürlich eine Gerade durch die beiden gegebenen Punkte dar.

[3] Joseph-Louis Lagrange (1736 - 1813, Turin, Paris) war ein berühmter Mathematiker und Physiker (siehe [9]).

2. Beispiel: Sei $f(x) = \int_0^x e^{\sin t} dt$. Angenommen, wir haben die folgenden Funktionswerte irgendwie berechnet oder aus einer Tabelle entnommen:

x_i	0.6	0.7	0.8
$y_i = f(x_i)$	0.81355	0.99671	1.19441

Dann erhalten wir z.B. für das erste Lagrange-Polynom:

$L_1(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} = \frac{(x-0.7)(x-0.8)}{(-0.1)(-0.2)} = 50\,(x-0.7)\,(x-0.8) =$

$= 50x^2 - 75x + 28.$

Analog ergibt sich $L_2(x) = -100x^2 + 140x - 48$ und $L_3(x) = 50x^2 - 65x + 21$. Daraus erhalten wir das Interpolationspolynom

$p(x) = 0.81355 L_1(x) + 0.99671 L_2(x) + 1.19441 L_3(x) =$

$= 0.727\,x^2 + 0.8865x + 0.01993.$

Damit können wir z.B. für $f(0.66)$ den Näherungswert $p(0.66) = 0.921\,701\,2$ berechnen. Der exakte Wert $f(0.66) = 0.921\,697\,88\ldots$ stimmt hier (auf 5 Stellen gerundet) sogar mit dem Näherungswert überein.

Wenn wir dagegen versuchen zu extrapolieren, so sehen wir z.B. für $x = 0.9$ schon eine merkliche Differenz:

$p(0.9) = 1.406\,65, \quad f(0.9) = 1.406\,35\ldots .$

Zum Vergleich berechnen wir auch die Fehlerabschätzung für $x = 0.66$ gemäß (5.2). Dazu brauchen wir zunächst eine Abschätzung von $\left\|f^{(3)}\right\|$ auf dem Intervall $[0.6,\ 0.8]$:

$f'(x) = e^{\sin x}, \quad f''(x) = e^{\sin x}\cos x,$

$f^{(3)}(x) = e^{\sin x}\,(-\sin x + \cos^2 x)$

Zur Abschätzung von $\left|f^{(3)}(x)\right|$ können wir z.B. die Extremwerte von $g(x) := -\sin x + \cos^2 x$ auf $[0.6,\ 0.8]$ bestimmen.

$\frac{dg}{dx} = -(1 + 2\sin x)\cos x = 0$ würde $\cos x = 0$ oder $\sin x = -\frac{1}{2}$ bedeuten, und das kann in unserem Intervall nicht eintreten.

Wir brauchen daher nur die Funktionswerte am Rand zu berechnen:

$g(0.6) = 0.116\ldots, \quad g(0.8) = -0.231\ldots$

Daher ist $\left\|f^{(3)}\right\| \le e^{\sin 0.8} \cdot 0.232 = 0.475\ldots$

Jetzt brauchen wir noch $\omega(0.66) = (0.66 - 0.6)(0.66 - 0.7)(0.66 - 0.8) = 0.000\,336$ und erhalten

$|f(0.66) - p(0.66)| \le \frac{1}{3!}\left\|f^{(3)}\right\|\omega(0.66) \le \frac{1}{6} \cdot 0.476 \cdot 0.000336 = 26.656 \times 10^{-6}$

Der tatsächliche Fehler beträgt ungefähr 3.3×10^{-6}, ist also deutlich kleiner.

5.4.2 Das Neville-Schema

Die Lagrange'sche Form des Interpolationspolynoms ist für die praktische Berechnung nicht besonders günstig. Das im Folgenden erklärte Schema ist dazu besser geeignet, insbesondere, wenn man nicht an dem Interpolationspolynom als Ganzem interessiert ist, sondern nur einzelne Funktionswerte berechnen will (vgl. insbesondere das Kapitel 5.5 über Extrapolation). Es eignet sich auch besonders gut, wenn die Anzahl der Stützstellen nicht von vornherein festgelegt ist, sondern sukzessive Stützstellen hinzugefügt werden. Dieses Schema stammt von Neville[4]. Ein ähnliches Verfahren wurde von Aitken[5] untersucht, daher spricht man auch vom *Aitken-Neville-Schema*.

Satz 5.3 *Für $0 \leq k < i$ sei T_{ik} das Interpolationspolynom zu den Punkten $(x_{i-k}, y_{i-k}), \ldots, (x_i, y_i)$. Diese Polynome kann man nach folgenden Formeln rekursiv berechnen:*

$$\begin{aligned} T_{i0}(x) &= y_i, \\ T_{ik}(x) &= \frac{(x_i - x)T_{i-1,k-1}(x) + (x - x_{i-k})T_{i,k-1}(x)}{x_i - x_{i-k}} \quad \textit{für } 0 < k < i. \end{aligned}$$

Bemerkungen:

1. Durch Vergleich mit Formel (5.5) erkennen wir, dass $T_{ik}(x)$ durch lineare Interpolation zwischen $T_{i-1,k-1}(x)$ und $T_{i,k-1}(x)$ entsteht (bezüglich der beiden Stützstellen x_{i-k} und x_i).

2. Die Polynome T_{ik} bzw. für festes x die Zahlen $T_{ik}(x)$ können nach folgendem Dreiecksschema der Reihe nach berechnet werden (eine Zeile nach der anderen oder eine Spalte nach der anderen):

$$\begin{array}{llllll} T_{10}, & & & & & \\ T_{20}, & T_{21}, & & & & \\ T_{30}, & T_{31}, & T_{32}, & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ T_{n0}, & T_{n1}, & T_{n2}, & \ldots, & T_{n,n-1}. & \end{array}$$

Zur Berechnung eines T_{ik} mit $k > 0$ verwendet man jeweils die beiden Eintragungen $T_{i-1,k-1}$ und $T_{i,k-1}$, welche schräg links oberhalb bzw. links von

[4] Eric Harold Neville (1889 - 1961, England) interessierte sich vor allem für Differenzialgeometrie und elliptische Funktionen (siehe J. London Math. Soc. 37 (1962), 479-482).

[5] Alexander Craig Aitken (1895 - 1967, Neuseeland, Schottland) lieferte wertvolle Beiträge zur Statistik, Numerik und Algebra, insbesondere Matrizen und Determinanten (siehe [9]).

T_{ik} stehen. $T_{n,n-1}$ ist dann schließlich das Interpolationspolynom zu allen gegebenen Punkten $(x_1, y_1), \ldots, (x_n, y_n)$.

Beweis des Satzes: Die Gleichung $T_{i0}(x) = y_i$ ist natürlich trivial. Für die zweite Gleichung stellen wir zunächst fest, dass die rechte Seite ein Polynom vom Grad $\leq k$ ist. Wir haben also nur nachzuprüfen, dass dieses Polynom durch die Punkte $(x_{i-k}, y_{i-k}), \ldots, (x_i, y_i)$ geht. Die Polynome $T_{i-1,k-1}$ und $T_{i,k-1}$ haben die Stützstellen x_j mit $i-k+1 \leq j \leq i-1$ gemeinsam. Ersetzen wir x durch x_j mit so einem Index j, so erhalten wir

$$\frac{(x_i - x_j)T_{i-1,k-1}(x_j) + (x_j - x_{i-k})T_{i,k-1}(x_j)}{x_i - x_{i-k}} = \frac{(x_i - x_j)y_j + (x_j - x_{i-k})y_j}{x_i - x_{i-k}} = y_j.$$

Für $j = i - k$ fällt der zweite Summand im Zähler weg, und wir erhalten

$$\frac{(x_i - x_{i-k})T_{i-1,k-1}(x_{i-k})}{x_i - x_{i-k}} = T_{i-1,k-1}(x_{i-k}) = y_{i-k}.$$

Für $j = i$ fällt der erste Summand im Zähler weg, und wir erhalten

$$\frac{(x_i - x_{i-k})T_{i,k-1}(x_i)}{x_i - x_{i-k}} = T_{i,k-1}(x_i) = y_i.$$

Damit ist der Satz bewiesen. ■

Beispiel: Sei $C(x) = \int_x^\infty \frac{\cos t}{t} dt$. (Das Negative dieser Funktion ist unter dem Namen *Integral-Kosinus* bekannt.)

Angenommen, wir kennen folgende (gerundeten) Funktionswerte:

x	5.0	5.2	5.5	5.6
y	0.19003	0.17525	0.14205	0.12867

und wollen $C(5.3)$ näherungsweise durch Interpolation bestimmen.

Das Neville-Schema sieht dann folgenderweise aus, wenn wir auf der linken Seite eine Spalte mit den x-Werten hinzufügen:

5.0	0.19003			
5.2	0.17525	T_{21}		
5.5	0.14205	T_{31}	T_{32}	
5.6	0.12867	T_{41}	T_{42}	T_{43}

Wir berechnen nun zuerst (mit $x = 5.3$)

$T_{21}(x) = \frac{(5.2-x)\cdot 0.19003 + (x-5.0)\cdot 0.17525}{5.2-5.0} = \frac{-0.1\times 0.19003 + 0.3\times 0.17525}{0.2} = 0.16786,$

dann in analoger Weise $T_{31}(x) = 0.16418$, und daraus

$T_{32}(x) = \frac{(5.5-x)T_{21}(x)+(x-5.0)T_{31}(x)}{5.5-5.0} = \frac{0.2\times 0.16786+0.3\times 0.16418}{0.5} = 0.16565.$

Analog erhält man $T_{42}(x) = 0.16534$ und schließlich $T_{43}(x) = 0.16550$, wobei immer auf 5 Stellen gerundet wurde. Der exakte Wert ist $C(5.3) = 0.165505958\ldots$.

Nach Durchführung der Berechnung sieht die Tabelle so aus:

5.0	0.19003			
5.2	0.17525	0.16786		
5.5	0.14205	0.16418	0.16565	
5.6	0.12867	0.16881	0.16534	0.16550

Hier bekommt man den Eindruck, dass die Diagonalelemente gut gegen den gesuchten Wert konvergieren. Das ist oft der Fall, muss aber nicht so sein.

Es ist übrigens nicht notwendig, dass die Werte x_i der Größe nach geordnet sind. Manchmal ist es zweckmäßiger, sie nach wachsendem Abstand von der Stelle x zu ordnen, damit man leichter beurteilen kann, ob man noch weitere Stützstellen hinzunehmen soll.

5.5 Extrapolation

Wie schon im Abschnitt 5.3 erklärt wurde, eignet sich das Interpolationspolynom im Allgemeinen nur schlecht zur Annäherung von Funktionswerten an Stellen außerhalb des durch die Stützstellen bestimmten Intervalls. Es gibt jedoch einen wichtigen Fall, wo sich Extrapolation bewährt, nämlich wenn die Stützstellen x_i in gewissem Sinne den Anfang einer Folge bilden, die schnell gegen eine Stelle x konvergiert, und der Funktionswert $f(x)$ an dieser Stelle gesucht ist.

5.5.1 Extrapolation für $x = 0$

Es bedeutet natürlich keine wesentliche Einschränkung der Allgemeinheit, wenn wir $x = 0$ annehmen. Als Stützstellen nehmen wir

$$x_i = \frac{h}{2^{i-1}} = 2h \cdot 2^{-i}$$

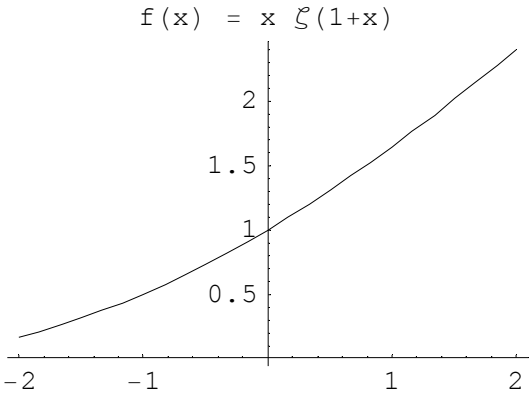
mit einem festen $h > 0$, also $(x_i) = (h, \frac{h}{2}, \frac{h}{4}, \frac{h}{8}, \ldots)$. Die Rekursionsformel des Neville-Schemas vereinfacht sich dann zu

$$\begin{aligned} T_{i0} &= f(x_i), \\ T_{ik} &= \frac{2^{-i}T_{i-1,k-1} - 2^{k-i}T_{i,k-1}}{2^{-i} - 2^{k-i}} = \frac{2^k T_{i,k-1} - T_{i-1,k-1}}{2^k - 1}, \end{aligned}$$

wobei das Argument (0) weggelassen wurde.

Beispiel:

Sei $f(x) = x \cdot \zeta(1+x)$, wobei ζ die *Riemann'sche Zetafunktion*[6] bezeichnet, welche so definiert ist: $\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$. Für $s = 1$ ist diese Reihe divergent, und daher ist $f(0)$ nicht definiert. Der folgende Plot von $f(x)$ lässt aber vermuten, dass man f durch die Zusatzdefinition $f(0) := 1$ im Punkt 0 stetig ergänzen kann.



Wir können nun versuchen, $f(0)$ durch Extrapolation der Werte $f(\frac{1}{2^{i-1}})$ für $i = 1, \ldots, 4$ näherungsweise zu bestimmen. Wir erhalten folgendes Neville-Schema (auf 6 Nachkommastellen gerundet):

1.000	1.644 934			
0.500	1.306 188	0.967 441		
0.250	1.148 778	0.991 368	0.999 344	
0.125	1.073 280	0.997 782	0.999 920	1.000 003

Wir erhalten hier also ein zufriedenstellendes Resultat.

[6] Georg Friedrich Bernhard Riemann (1826 - 1866, Deutschland, Italien) ist einer der bedeutendsten Mathematiker des 19. Jahrhunderst (siehe [9]).

Richardson[7]-Extrapolation

Wenn f eine *gerade Funktion* ist, das heißt $f(-x) = f(x)$, dann kommen bei der Taylorentwicklung nur gerade Potenzen vor, d.h. $f(x) = c_0 + c_2x^2 + c_4x^4 + \ldots$. In diesem Fall ist es günstiger, auch zur Interpolation ein Polynom zu verwenden, bei dem nur gerade Potenzen auftreten, d.h. ein Polynom in der Variablen $t = x^2$. Man setzt dann also $t_i = x_i^2 = \frac{h^2}{4^{i-1}}$ und erhält folgende Rekursionsformel

$$T_{ik} = \frac{4^k T_{i,k-1} - T_{i-1,k-1}}{4^k - 1}. \tag{5.6}$$

Diese Variante des Extrapolationsverfahrens wird insbesondere bei der numerischen Differenziation (Kapitel 6.3) und Integration (Kapitel 7.3, "Romberg-Integration") verwendet.

5.5.2 Summation einer Reihe mittels Extrapolation

Man kann das beschriebene Verfahren unter Umständen auch zur Berechnung der Summe einer (unendlichen) Reihe verwenden. Sei $\sum_{i=1}^{\infty} a_i$ eine solche Reihe. Dann setzen wir

$$f(\frac{1}{n}) := \sum_{i=1}^{n} a_i,$$

das ist also die n-te Partialsumme. Obwohl f nur für die Zahlen der Form $1/n$ definiert ist, können wir versuchen, durch Extrapolation einen brauchbaren Wert für $f(0)$ zu finden, und das entspricht natürlich der unendlichen Summe $\sum_{i=1}^{\infty} a_i$.

Beispiel:

Wir betrachten wieder die Riemann'sche Zetafunktion und versuchen, $\zeta(2) = \sum_{k=1}^{\infty} \frac{1}{k^2}$ näherungsweise zu berechnen. Hier ist $f(1) = 1,\ f(1/2) = 1 + \frac{1}{4} = 1.25,\ f(1/4) = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} = 1.423\,61\ldots$ usw., und es ergibt sich folgenes Neville-Schema:

1.000	1			
0.500	1.25	1.5		
0.250	1.423 611	1.597222	1.62963	
0.125	1.527 422	1.631233	1.64257	1.644419

[7]Lewis Fry Richardson (1881 - 1953, England, Schottland) hat unter anderem auch versucht, die Ursache von Kriegen mathematisch zu untersuchen (siehe [9]).

Der exakte Wert ist $\zeta(2) = \frac{1}{6}\pi^2 = 1.644\,934\ldots$. Wir erhalten also einen relativ guten Wert, obwohl wir nur die ersten 8 Summanden berücksichtigt haben. (Die Summe der ersten 8 Summanden ist $f(1/8) = 1.527\,422\ldots$.)

Kapitel 6

Numerische Differenziation

6.1 Motivation

Wenn eine Funktion in "geschlossener" Form gegeben ist, das heißt als Zusammensetzung von endlich vielen "elementaren" Funktionen, kann man im Prinzip den Differenzialquotienten immer exakt berechnen. Trotzdem ist die numerische Differenziation von nicht geringer Bedeutung, vor allem aus folgenden Gründen.

- Die exakte Differenziation kann zu sehr komplizierten Formeln führen, bei deren Auswertung erhebliche Rundungsfehler (insbesondere durch Auslöschung) auftreten können.

- Wenn eine Funktion durch Grenzwerte definiert ist (z.B. unendliche Reihen oder Integrale), kann die exakte Differenziation problematisch sein. So ist etwa die gliedweise Differenziation einer Reihe nur unter bestimmten Voraussetzungen erlaubt (vgl. Vorlesungen oder Bücher über Analysis).

- Wenn die Funktionswerte nur einzeln, etwa durch Messung, bestimmt werden können, ist exakte Differenziation unmöglich.

- Schließlich spielt die näherungsweise Differenziation bei der numerischen Lösung von Differenzialgleichungen eine wesentliche Rolle.

6.2 Differenziation des Interpolationspolynoms

Da Polynome sehr leicht zu differenzieren sind, verwendet man zur näherungsweisen Differenziation vorzugsweise ein Interpolationspolynom, das durch einige Funktionswerte in der Nähe der zu untersuchenden Stelle definiert ist.

6.2.1 Abschätzung des Verfahrensfehlers

Nach Satz 5.2 gilt für das Interpolationspolynom p zu einer Funktion f

$$f(x) = p(x) + \frac{1}{n!}\ f^{(n)}(\xi(x))\ \omega(x),$$

mit der Stützstellenfunktion $\omega(x) = \prod_{i=1}^{n}(x - x_i)$. Wir haben hier $\xi(x)$ an Stelle von ξ geschrieben, um deutlich zu machen, dass ξ von x abhängt. Nehmen wir einmal an, dass die Funktion $\xi(x)$ differenzierbar ist, obwohl das überhaupt nicht klar ist. Dann ergibt sich

$$f'(x) = p'(x) + \frac{1}{n!}\ \left(f^{(n+1)}(\xi(x))\ \xi'(x)\ \omega(x) + f^{(n)}(\xi(x))\ \omega'(x)\right).$$

Wählt man für x eine der Stützstellen x_k, so fällt der erste Summand in dem großen Klammerausdruck weg, da ja $\omega(x_k) = 0$ ist, und wir erhalten

$$f'(x_k) = p'(x_k) + \frac{1}{n!}\ f^{(n)}(\xi(x_k))\ \omega'(x_k).$$

Man kann nun (mit einiger Mühe) zeigen, dass diese Formel trotz der problematischen Annahme über die Differenzierbarkeit von $\xi(x)$ stimmt (siehe z.B. [7]). Das heißt, es gilt folgender Satz, der analog zu Satz 5.2 zur Fehlerabschätzung verwendet werden kann:

Satz 6.1 *Sei p das Interpolationspolynom der n-mal stetig differenzierbaren Funktion f zu den Stützstellen $x_1, \ldots, x_n$ im Intervall I. Dann gibt es zu jedem $k \in \{1, \ldots, n\}$ ein $\xi_k \in I$, sodass*

$$f'(x_k) = p'(x_k) + \frac{1}{n!}\ f^{(n)}(\xi_k)\ \omega'(x_k),$$

wobei ω das zugehörige Stützstellenpolynom ist.

Bemerkung 6.2 $\omega'(x_k) = \prod_{i=1, i\neq k}^{n}(x_k - x_i)$.

Beweis:
Nach der Produktregel gilt $\omega'(x) = \sum_{j=1}^{n} \prod_{i=1, i\neq j}^{n} (x - x_i)$ und daher $\omega'(x_k) = \prod_{i=1, i\neq k}^{n} (x_k - x_i)$, da für $x = x_k$ alle anderen Summanden wegfallen. ■

Zweipunktformel

Der einfachste (sinnvolle) Fall ist $n = 2$. Hier handelt es sich um lineare Interpolation. Mit $h := x_2 - x_1$ sieht das Interpolationspolynom so aus (siehe (5.5)):

$$p(x) = \frac{1}{h}((x_2 - x)y_1 + (x - x_1)y_2),$$

also

$$p'(x) = \frac{1}{h}(y_2 - y_1),$$

das ist natürlich der Anstieg der die beiden Punkte verbindenden "Sehne".
$\omega'(x_1) = x_1 - x_2 = -h$, somit erhalten wir

$$f'(x_1) = \frac{1}{h}(y_2 - y_1) - \frac{h}{2}\ f^{(2)}(\xi).$$

Im Allgemeinen ist es aber günstiger, eine größere, ungerade Anzahl von Stützstellen zu verwenden (z.B. 3, 5 oder 7). Für äquidistante Stützstellen ergeben sich auch dann noch relativ einfache Formeln:

Dreipunktformel

Für drei Stützstellen $x_1,\ x_2 = x_1 + h,\ x_3 = x_1 + 2h$ sehen die Lagrange-Koeffizienten folgendermaßen aus:
$L_1(x) = \frac{x-x_2}{x_1-x_2} \cdot \frac{x-x_3}{x_1-x_3} = \frac{1}{2h^2}(x - x_2)(x - x_3)$,
$L_1'(x) = \frac{1}{2h^2}\left((x - x_3) + (x - x_2)\right)$, also $L_1'(x_1) = \frac{1}{2h^2}(-3h) = -\frac{3}{2h}$.
Analog ergibt sich $L_2'(x_1) = \frac{2}{h}$ und $L_3'(x_1) = -\frac{1}{2h}$.
Wegen $\omega'(x_1) = (x_1 - x_2)(x_1 - x_3) = 2h^2$ erhalten wir damit die folgende *Dreipunktformel:*

$$f'(x_1) = \frac{1}{2h}\left(-3\ f(x_1) + 4\ f(x_2) - f(x_3)\right) + \frac{h^2}{3}\ f^{(3)}(\xi) \qquad (6.1)$$

für ein $\xi \in [x_1,\ x_1 + 2h]$. Der Fehler ist hier also nur mehr ein $O(h^2)$ im Gegensatz zum $O(h)$ bei der Zweipunktformel.

Für die Ableitung an der mittleren Stützstelle x_2 ergibt sich $L_1'(x_2) = -\frac{1}{2h}$, $L_2'(x_2) = 0$, $L_3'(x_2) = \frac{1}{2h}$, $\omega'(x_2) = -h^2$ und somit

$$f'(x_2) = \frac{1}{2h}\,(f(x_3) - f(x_1)) - \frac{h^2}{6}\; f^{(3)}(\xi).$$

Diese Formel wird etwas schöner, wenn wir die Bezeichnung folgendermaßen ändern:

$$x_0 := x_2, \quad x_1 = x_0 - h, \quad x_3 = x_0 + h.$$

Wir erhalten dann die *zentrale Dreipunktformel*

$$f'(x_0) = \frac{1}{2h}(f(x_0 + h) - f(x_0 - h)) - \frac{h^2}{6}\; f^{(3)}(\xi), \tag{6.2}$$

die in vielen Fällen günstiger als (6.1) ist. Dagegen kann (6.1) auch verwendet werden, wenn f in x_1 nur rechtsseitig differenzierbar ist.

Fünfpunktformel

In analoger Weise kann man die folgende *zentrale Fünfpunktformel* herleiten:

$$f'(x_0) = \frac{1}{12h}(f(x_0-2h) - 8\; f(x_0-h) + 8\; f(x_0+h) - f(x_0+2h)) + \frac{h^4}{30}\; f^{(5)}(\xi). \tag{6.3}$$

Merkwürdigerweise kommt in den letzten beiden Formeln der mittlere Punkt $(x_0, f(x_0))$ gar nicht vor!

6.2.2 Fehleranalyse

Die Restglieder bei obigen Formeln ermöglichen eine Abschätzung des Verfahrensfehlers, falls eine Schranke für $\left|f^{(n)}(x)\right|$ bekannt ist. Bei der numerischen Differenziation ist es aber besonders wichtig, auch den Einfluss von Daten- und Rundungsfehlern zu berücksichtigen.

Die allgemeine Gestalt einer n-Punkt-Formel für numerische Differenziation lautet (bei äquidistanten Stützstellen)

$$f'(x) \approx \frac{1}{h}\sum_{i=1}^{n} c_i\; f(x_i)$$

mit gewissen Konstanten c_i. Im Allgemeinen können aber die Funktionswerte $f(x_i)$ nur näherungsweise bestimmt werden (entweder auf Grund der unvermeidlichen Messfehler oder wegen der notwendigen Rundungen). Angenommen, $\tilde{f}(x_i)$ ist eine solche Näherung von $f(x_i)$. Der dadurch entstehende

absolute Fehler bei Anwendung obiger Formel ist dann

$$\begin{aligned}\left|\frac{1}{h}\sum_{i=1}^{n} c_i\ \tilde{f}(x_i) - \frac{1}{h}\sum_{i=1}^{n} c_i\ f(x_i)\right| &\leq \frac{1}{h}\sum_{i=1}^{n}|c_i|\ \left|\tilde{f}(x_i) - f(x_i)\right| \\ &\leq \frac{1}{h}\left(\sum_{i=1}^{n}|c_i|\right)\ \max_i\left|\tilde{f}(x_i) - f(x_i)\right|.\end{aligned}$$

Diese Schranke wird umso größer, je kleiner die Schrittweite h gewählt wird! Bezeichnen wir den maximalen Datenfehler mit $|\Delta f|$, so erhalten wir für den Gesamtfehler $|\Delta f'|$ (ohne Rundungsfehler) bei der numerischen Differenziation mit einer n-Punkt-Formel also eine Abschätzung der Form

$$|\Delta f'| \leq C_1\ \frac{1}{h}\ |\Delta f|\ +\ C_2\ h^{n-1}\ \left\|f^{(n)}\right\|$$

mit gewissen Konstanten C_1 und C_2.

Wenn man $|\Delta f|$ und $\left\|f^{(n)}\right\|$ kennt, kann man auf Grund dieser Formel h so bestimmen, dass die Schranke für den Gesamtfehler minimal wird: Sei $a = C_1\,|\Delta f|$ und $b = C_2\,\left\|f^{(n)}\right\|$. Dann geht es um das Minimum der Funktion

$$s(h) := \frac{a}{h} + bh^{n-1}.$$

Berechnen wir die erste Ableitung und setzen sie gleich Null,

$$s'(h) = -\frac{a}{h^2} + (n-1)bh^{n-2} = 0,$$

so erhalten wir für die Stelle des Minimums

$$h = \sqrt[n]{\frac{a}{(n-1)b}}.$$

Es handelt sich tatsächlich um ein Minimum, denn

$$s''(h) = \frac{2a}{h^3} + (n-1)(n-2)bh^{n-3} > 0$$

für $n \geq 2$.

Wie schon oben erwähnt, wurde bei dieser Analyse aber noch nicht berücksichtigt, dass bei der Auswertung der Differenziationsformeln beträchtliche Rundungsfehler durch Auslöschung auftreten können, da normalerweise ungefähr gleich große Zahlen voneinander subtrahiert werden (siehe z.B. die Formeln (6.2) und (6.3)).

6.2.3 Zweite Ableitung

Um zu geeigneten Formeln für die numerischen Berechnung der zweiten Ableitung zu kommen, ist es am einfachsten, wenn man vom Taylor'schen[1] Satz ausgeht. Nach diesem gilt bekanntlich (falls f 4-mal differenzierbar ist):

$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2}\ f''(x_0)h^2 + \frac{1}{6}\ f^{(3)}(x_0)h^3 + \frac{1}{24}\ f^{(4)}(\xi_1)h^4$

für ein $\xi_1 \in (x_0,\ x_0+h)$, und

$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{1}{2}\ f''(x_0)h^2 - \frac{1}{6}\ f^{(3)}(x_0)h^3 + \frac{1}{24}\ f^{(4)}(\xi_2)h^4$

für ein $\xi_2 \in (x_0-h,\ x_0)$.

Durch Addition folgt

$$f(x_0+h) + f(x_0-h) = 2f(x_0) + f''(x_0)h^2 + \frac{1}{24}\left(f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right)h^4.$$

Wenn $f^{(4)}$ stetig ist, dann gibt es nach dem Zwischenwertsatz ein ξ aus dem Intervall $[x_0-h,\ x_0+h]$, sodass

$$f^{(4)}(\xi) = \frac{1}{2}\left(f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right),$$

und es folgt

$$f''(x_0) = \frac{1}{h^2}\left(f(x_0+h) - 2f(x_0) + f(x_0-h)\right) - \frac{h^4}{12}f^{(4)}(\xi). \tag{6.4}$$

Das kann man auch so schreiben:

$$f''(x_0) = \frac{1}{h}\left(\frac{f(x_0+h)-f(x_0)}{h} - \frac{f(x_0)-f(x_0-h)}{h}\right) - \frac{h^4}{12}f^{(4)}(\xi).$$

Das ist also eine "Dreipunktformel" für die zweite Ableitung.

6.3 Differenziation durch Extrapolation

Wie wir in Abschnitt 6.2.2 gesehen haben, kann man bei der numerischen Differenziation die Schrittweite h nicht beliebig klein wählen. Da man sich aber in Wirklichkeit für den Wert bei $h = 0$ interessiert, ist es naheliegend, Extrapolation anzuwenden. Betrachten wir z.B. die zentrale Dreipunktformel

[1] Brook Taylor (1685 - 1731, England) ist nur einer von mehreren berühmten Mathematikern, die die heute so genannten Taylor-Reihen mehr oder weniger unabhängig voneinander entdeckten, z.B. Gregory, Newton, Leibnitz, Bernoulli, de Moivre (siehe [9]).

(6.2) und schreiben x statt h. Dann geht es um den Limes der folgenden Funktion für $x \to 0$:

$$D(x) := \frac{1}{2x}\,(f(x_0 + x) - f(x_0 - x)).$$

Diese Funktion ist gerade, d.h. $D(-x) = D(x)$, und daher können wir die diesbezügliche Extrapolationsformel (5.6) verwenden. Wir haben also (für festes h)

$$T_{i0} = D\left(\frac{h}{2^{i-1}}\right) = \frac{2^{i-2}}{h}\left(f\left(x_0 + \frac{h}{2^{i-1}}\right) - f\left(x_0 - \frac{h}{2^{i-1}}\right)\right),$$

und

$$T_{ik} = \frac{4^k T_{i,k-1} - T_{i-1,k-1}}{4^k - 1} \text{ für } 0 < k < i.$$

Speziell für $i = 2$ und $k = 1$ ergibt das

$T_{21} = \frac{1}{3}(4T_{20} - T_{10}) = \frac{1}{6h}(8f(x_0 + \frac{h}{2}) - 8f(x_0 - \frac{h}{2}) - f(x_0 + h) + f(x_0 - h))$.

Das ist dasselbe wie die zentrale Fünfpunktformel, nur mit $h/2$ an Stelle von h, allerdings ohne Fehlerglied.

Im nächsten Schritt erhalten wir

$T_{31} = \frac{1}{3}(4T_{30} - T_{20}) = \frac{1}{3h}\left(8f\left(x_0 + \frac{h}{4}\right) - 8f\left(x_0 - \frac{h}{4}\right) - f(x_0 + \frac{h}{2}) + f(x_0 - \frac{h}{2})\right)$

und daraus

$T_{32} = \frac{1}{15}(16T_{31} - T_{21}) =$

$= \frac{1}{90h}(-f(x_0 - h) + 40f(x_0 - \frac{h}{2}) - 256f\left(x_0 - \frac{h}{4}\right) +$

$\quad +256f(x_0 + \frac{h}{4}) - 40f(x_0 + \frac{h}{2}) + f(x_0 + h))$.

Auch die letzte Formel könnte man durch Differenziation des Interpolationspolynoms durch die entsprechenden 7 Punkte erhalten, die hier allerdings nicht äquidistant sind.

Bemerkung: Im Hinblick auf mögliche Rundungsfehler ist es im Allgemeinen nicht so günstig, die näherungsweise Ableitung durch Einsetzen in eine der obigen Formel zu ermitteln. Besser ist es, die T_{ik} auf Grund der Rekursionsformel zu berechnen.

Kapitel 7

Numerische Integration

Es geht hier um die näherungsweise Berechnung von bestimmten Integralen, die man auch "*numerische Quadratur*" nennt. Im Gegensatz zur Differenziation ist es auch bei Zusammensetzungen von elementaren Funktionen im Allgemeinen nicht möglich, das Integral wieder durch elementare Funktionen darzustellen. Daher ist in vielen Fällen die numerische Integration die einzige Möglichkeit, den (numerischen) Wert eines bestimmten Integrals zu berechnen. Im übrigen gelten die für die Bedeutung der numerischen Differenziation angeführten Gründe auch hier.

Der Grundgedanke ist hier derselbe wie bei der numerischen Differenziation: Wir integrieren an Stelle der gegebenen Funktion ein Interpolationspolynom.

7.1 Newton-Cotes Formeln

Die Newton[1]-Cotes[2] Formeln beruhen auf der Integration eines Interpolationspolynoms mit äquidistanten Stützstellen.

[1]Sir Isaac Newton (1643 - 1727, England) gilt als Begründer der klassischen theoretischen Physik und entwickelte die Differenzial- und Integralrechnung gleichzeitig mit Leibniz, aber unabhängig von ihm (siehe [9]).

[2]Roger Cotes (1682 - 1716, England) hat sich intensiv mit den Arbeiten von Newton auseinandergesetzt (siehe [9]).

7.1.1 Geschlossene Newton-Cotes Formeln

Sei $[a, b]$ das Integrationsintervall. Dann betrachten wir zu gegebenem $n \in \mathbb{N}$ die Stützstellen

$$x_k = a + kh \quad \text{für } k = 0, 1, \ldots, n$$

mit der Schrittweite

$$h = \frac{b-a}{n}.$$

Die Anzahl der Stützstellen ist hier also $n+1$, da wir die Nummerierung mit 0 beginnen.

Sei nun p das Interpolationspolynom der zu integrierenden Funktion f zu diesen Stützstellen, und

$$y_k := f(x_k) = p(x_k).$$

$p(x) = \sum_{k=0}^{n} y_k L_k(x)$ mit den Lagrange-Koeffizienten $L_k(x)$ gemäß (5.4). Daher ist

$$\int_a^b p(x)\ dx = \sum_{k=0}^{n} y_k \int_a^b L_k(x)\ dx = \sum_{k=0}^{n} A_k y_k$$

mit

$$A_k := \int_a^b L_k(x)\ dx.$$

Damit haben wir das Problem im Prinzip schon gelöst. Das Interessante ist aber, dass man durch eine geeignete Transformation erreichen kann, dass die Integrale $\int_a^b L_k(x)\ dx$ nicht jedesmal neu berechnet werden müssen. Wir setzen

$$t := \frac{x-a}{h}, \quad \text{das heißt } x = a + th$$

Dadurch wird das Intervall $[a, b]$ auf $[0, n]$ transformiert, und wegen $x_i = a + ih$ gilt

$$L_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i} = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{th - ih}{kh - ih} = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{t-i}{k-i}$$

und daher

$$A_k = h\ \alpha_k^{(n)}$$

mit

$$\alpha_k^{(n)} = \int_0^n \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{t-i}{k-i} \right)\ dt.$$

Diese Koeffizienten hängen nicht mehr von a, b, und h ab. Es gibt Tabellen für alle in der Praxis vorkommenden Werte von n.

Wir erhalten somit

$$\int_a^b f(x)\ dx \approx h \sum \alpha_k^{(n)} f(x_k).$$

Das nennt man die *geschlossene Newton-Cotes-Formel vom Grad n.*

Ähnlich wie bei der numerischen Differenziation kann man auch hier den Verfahrensfehler durch Zurückführung auf den Interpolationsfehler abschätzen. Sei $E_n(f)$ der Fehler, der sich bei Anwendung der n-ten Newton-Cotes Formel ergibt, das heißt

$$E_n(f) := \int_a^b f(x)\ dx - h \sum \alpha_k^{(n)} f(x_k)$$

mit $h = (b-a)/n$ und $x_k = a + kh$. Dann gilt:

Satz 7.1 *Wenn f auf $[a,b]$ $(n+2)$-mal stetig differenzierbar ist, dann gibt es ein $\xi \in [a,b]$, sodass für gerades n*

$$E_n(f) = \frac{1}{(n+2)!} \left(\int_0^n t\ \omega_n(t)\ dt \right) h^{n+3} f^{(n+2)}(\xi), \tag{7.1}$$

und für ungerades n

$$E_n(f) = \frac{1}{(n+1)!} \left(\int_0^n \omega_n(t)\ dt \right) h^{n+2} f^{(n+1)}(\xi),$$

mit dem (transformierten) Stützstellenpolynom

$$\omega_n(t) := \prod_{i=0}^{n} (t-i).$$

Der *Beweis* dieses Satzes ist relativ lang und wird hier nicht ausgeführt (vgl. z.B. [7]). ■

Bemerkung: Man sieht hier, dass der Exponent von h bei ungerader Anzahl n genauso groß ist wie bei der vorhergehenden geraden Anzahl. Im Allgemeinen ist es daher eher ungünstig, n ungerade zu wählen (abgesehen vom Fall $n = 1$).

Die geschlossene Newton-Cotes-Formel vom Grad n ist auf Grund der Herleitung exakt, wenn f mit dem Interpolationspolynom (vom Grad n) übereinstimmt. Interessanterweise ist sie aber auf Grund des obigen Satzes für gerades n sogar für alle Polynome vom Grad $\leq n+1$ exakt, welche durch die Stützpunkte hindurchgehen, denn wenn f ein solches Polynom ist, dann ist $f^{(n+2)}(\xi) = 0$ für alle ξ.

Für kleine Werte von n kann man die Koeffizienten $\alpha_k^{(n)}$ und den Fehlerterm leicht berechnen:

$n = 1$: **(Sehnentrapezregel)**

$\alpha_0^{(1)} = \int_0^1 \frac{t-1}{0-1}\, dt = -\int_0^1 (t-1)\, dt = \frac{1}{2}$,
$\alpha_1^{(1)} = \int_0^1 \frac{t-0}{1-0}\, dt = \int_0^1 t\, dt = \frac{1}{2}$,
$\int_0^1 \omega_1(t)\, dt = \int_0^1 (t-0)(t-1)\, dt = \int_0^1 (t^2-t)\, dt = -\frac{1}{6}$,
also (mit $h = b-a$)
$E_1(f) = \frac{1}{2}\left(-\frac{1}{6}\right) h^3 f''(\xi) = -\frac{h^3}{12} f''(\xi)$
und somit

$$\int_a^b f(x)\, dx = \frac{h}{2}(f(a)+f(b)) - \frac{h^3}{12}\, f''(\xi)$$

für ein $\xi \in [a,b]$. Diese einfache, aber wichtige Näherungsformel heißt Sehnentrapezregel, da sie (nach Weglassung des Fehlerterms) die Fläche des Trapezes angibt, das entsteht, wenn man die Punkte $(a, f(a))$, $(b, f(b))$ durch eine geradlinige Strecke ("Sehne") verbindet.

$n = 2$: **(Simpson'sche Regel[3])**

$\alpha_0^{(2)} = \int_0^2 \frac{(t-1)(t-2)}{(0-1)(0-2)}\, dt = \frac{1}{2}\int_0^2 (t^2-3t+2)\, dt = \frac{1}{3}$,
und analog $\alpha_1^{(2)} = \frac{4}{3}$, $\alpha_2^{(2)} = \frac{1}{3}$.
$\int_0^2 t\ \omega_2(t)\, dt = \int_0^2 t(t-0)(t-1)(t-2)\, dt = \int_0^2 (t^4-3t^3+2t^2)\ dt = -\frac{4}{15}$,
also (mit $h = (b-a)/2$)
$E_2(f) = \frac{1}{4!}\left(-\frac{4}{15}\right) h^5 f^{(4)}(\xi) = -\frac{h^5}{90}\ f^{(4)}(\xi)$.
Das ergibt

$$\int_a^b f(x)\, dx = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) - \frac{h^5}{90}\ f^{(4)}(\xi).$$

[3] nach Thomas Simpson (1710 - 1761, England), siehe [9].

$n = 3$: **(3/8-Regel)**

Hier ergibt sich in analoger Weise (mit $h = (b-a)/3$ und $x_k = a + kh$):

$$\int_a^b f(x)\ dx = \frac{3h}{8}(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)) - \frac{3h^5}{80} f^{(4)}(\xi).$$

$n = 4$: **(Milne-Regel[4])**

$$\int_a^b f(x)\,dx = \frac{2h}{45}\,(7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)) - \frac{8h^7}{945} f^{(6)}(\xi).$$

Für einige weitere Werte von n findet man die expliziten Newton-Cotes Formeln z.B. in [10].

Bemerkung: Wenn man die Fehlerterme dieser Formeln vergleichen will, muss man beachten, dass h von n abhängt. So sieht es z.B. auf den ersten Blick so aus, als ob der Betrag des Fehlerterms bei $n = 3$ größer als bei $n = 2$ wäre, da $\frac{3}{80} > \frac{1}{90}$ ist. Setzt man aber die richtigen Werte für h ein, so sieht man

$$\frac{3(b-a)^5}{3^5 \cdot 80} = \frac{(b-a)^5}{6480} < \frac{(b-a)^5}{2^5 \cdot 90} = \frac{(b-a)^5}{2880}.$$

Abgesehen davon ist aber der Fall $n = 3$ doch eher ungünstig, wie schon oben erklärt wurde.

7.1.2 Offene Newton-Cotes Formeln

Diese Formeln werden unter anderem angewendet, wenn die Funktionswerte an den Intervallendpunkten schwierig oder nur ungenau berechnet werden können, weil sie z.B. nur als Limes definiert sind.

Hier setzen wir

$$x_k = a + (k+1)h \text{ für } k = 0, \ldots, n$$

mit

$$h = \frac{b-a}{n+2}.$$

Analog zu den geschlossenen Formeln ergibt sich durch Transformation auf das Intervall $[-1, n+1]$

$$\int_a^b f(x)\ dx = h \sum_{k=0}^{n} \tilde{\alpha}_k^{(n)} f(x_k)$$

[4] Edward Arthur Milne (1896 - 1950, England, Irland) war in erster Linie an Astronomie interessiert (z.B. Sternatmosphären), siehe [9].

mit

$$\tilde{\alpha}_k^{(n)} = \int_{-1}^{n+1} \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{t-i}{k-i} \right) \, dt.$$

Auch die Fehlerabschätzung unterscheidet sich nur durch die Integrationsgrenzen von den geschlossenen Formeln.

Satz 7.2 *Wenn f auf $[a,b]$ $(n+2)$-mal stetig differenzierbar ist, dann gibt es ein $\xi \in [a,b]$, sodass für gerades n*

$$E_n(f) = \frac{1}{(n+2)!} \left(\int_{-1}^{n+1} t \; \omega_n(t) \; dt \right) h^{n+3} f^{(n+2)}(\xi),$$

und für ungerades n

$$E_n(f) = \frac{1}{(n+1)!} \left(\int_{-1}^{n+1} \omega_n(t) \; dt \right) h^{n+2} f^{(n+1)}(\xi).$$

Auch hier ergeben sich für kleine n relativ einfache Formeln. Insbesondere ist hier auch der Fall $n = 0$ sinnvoll. (Dabei ist zu beachten, dass ein leeres Produkt den Wert 1 hat.)

$n = 0$: **(Mittelpunktsregel)**

$\tilde{\alpha}_0^{(0)} = \int_{-1}^{1} 1 \; dt = 2,$
$\int_{-1}^{1} t \; \omega_0(t) \; dt = \int_{-1}^{1} t^2 \; dt = \frac{2}{3},$
also gilt mit $h = (b-a)/2$

$$\int_a^b f(x) \; dx = 2h \; f\left(\frac{a+b}{2}\right) + \frac{h^3}{3} f''(\xi)$$

für ein $\xi \in [a,b]$. Dieser Näherungswert entspricht dem Flächeninhalt eines Rechtecks, dessen Höhe gleich dem Funktionswert im Mittelpunkt des Intervalls ist. Man kann ihn aber auch als Flächeninhalt eines Trapezes auffassen, das oben durch die Tangente im Punkt $(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right))$ an den Funktionsgraphen begrenzt wird. Daher wird diese Regel auch *Tangententrapezregel* genannt.

$n = 1$:

$\tilde{\alpha}_0^{(1)} = \int_{-1}^{2} \frac{t-1}{0-1}\, dt = \int_{-1}^{2} (1-t)\, dt = \frac{3}{2}$, und ebenso $\tilde{\alpha}_1^{(1)} = \frac{3}{2}$,
$\int_{-1}^{2} \omega_1(t)\, dt = \int_{-1}^{2} t(t-1)\, dt = \frac{3}{2}$, also

$$\int_a^b f(x)\, dx = \frac{3h}{2}\left(f(x_0) + f(x_1)\right) + \frac{3h^3}{4} f''(\xi)$$

mit $h = (b-a)/3$, $x_0 = a + h$, $x_1 = a + 2h$.

Der hier auftretende Faktor des Fehlerterms ist kleiner als bei $n = 0$, denn $\frac{3(b-a)^3}{3^3 \cdot 4} = \frac{(b-a)^3}{36} < \frac{(b-a)^3}{2^3 \cdot 3} = \frac{(b-a)^3}{24}$.

$n = 2$:

$$\int_a^b f(x)\, dx = \frac{4h}{3}(2f(x_0) - f(x_1) + 2f(x_2)) + \frac{14h^5}{45} f^{(4)}(\xi)$$

mit $h = (b-a)/4$ und $x_k = a + (k+1)h$.

7.1.3 Rundungsfehler bei den Newton-Cotes Formeln

Bei diesen Formeln treten ab einem gewissen n negative Koeffizienten auf, und zwar bei den geschlossenen Formeln ab $n = 8$ (siehe z.B. [10]) und bei den offenen bereits ab $n = 2$ (siehe oben). Dadurch ist es leicht möglich, dass durch Auslöschung große Rundungsfehler entstehen. Abgesehen davon konvergieren die Näherungswerte für $n \to \infty$ im Allgemeinen nicht gegen den exakten Wert des Integrals (vgl. Bemerkung am Ende von Kapitel 5.3).

Aus diesen Gründen werden diese Formeln nur selten für größere Werte von n verwendet. Statt dessen geht man meist zu zusammengesetzten Formeln über.

7.2 Zusammengesetzte Formeln

Wir teilen das Intervall $[a, b]$ in N gleich große Teile mit Länge

$$H = \frac{b-a}{N}$$

und setzen

$$z_j := a + jH \quad \text{für } j = 0, \ldots, N.$$

Auf jedem Teilintervall $[z_{j-1}, z_j]$ wenden wir dann eine bestimmte Integrationsformel an, die einen Näherungswert

$$I_n^{(j)}(f) \approx \int_{z_{j-1}}^{z_j} f(x)\ dx$$

liefert. Dabei bedeutet n wie vorhin, dass die Formel auf $n+1$ Stützstellen beruht.

Im Folgenden wird die Verwendung der geschlossenen Newton-Cotes Formeln für diesen Zweck näher ausgeführt. In diesem Fall ist

$$I_n^{(j)}(f) = h \sum_{k=0}^{n} \alpha_k^{(n)} f(z_{j-1} + kh)$$

mit

$$h = \frac{H}{n} = \frac{b-a}{nN}.$$

Auf jedem Teilintervall $[z_{j-1}, z_j]$ treten also dieselben Koeffizienten $\alpha_k^{(n)}$ auf (mit $k \in \{0, \ldots, n\}$). Insgesamt erhalten wir den Näherungswert

$$I_N(f) = \sum_{j=1}^{N} I_n^{(j)} \approx \int_a^b f(x)\ dx.$$

Wir wollen nun den Verfahrensfehler bei dieser Methode abschätzen.

$$E_N(f) := \int_a^b f(x)\ dx - I_N(f) = \sum_{j=1}^{N} E_n^{(j)}(f),$$

wobei die $E_n^{(j)}(f)$ gemäß Satz 7.1 bestimmt sind. Dabei tritt für jedes Teilintervall $[z_{j-1}, z_j]$ ein Zwischenwert ξ_j auf. Dies lässt sich vereinfachen. Betrachten wir z.B. den Fall, dass n gerade ist. Dann ist

$$E_N(f) = \frac{1}{(n+2)!} \left(\int_0^n t\ \omega_n(t)\ dt \right) h^{n+3} \sum_{j=1}^{N} f^{(n+2)}(\xi_j). \tag{7.2}$$

Das $\frac{1}{N}$-fache der hier auftretenden Summe ist ein arithmetisches Mittel von Werten der Funktion $f^{(n+2)}$. Wenn diese Funktion stetig ist, gibt es daher eine Stelle $\xi \in [a, b]$, sodass dieses arithmetisches Mittel gleich $f^{(n+2)}(\xi)$ ist. Damit erhalten wir folgenden Satz:

Satz 7.3 *Bei der oben beschriebenen zusammengesetzten Integration einer $(n+2)$-mal stetig differenzierbaren Funktion f gilt für gerades n*

$$E_N(f) = \frac{b-a}{(n+2)!\ n} \left(\int_0^n t\ \omega_n(t)\ dt \right) h^{n+2} f^{(n+2)}(\xi) \tag{7.3}$$

für ein $\xi \in [a,b]$. Für ungerades n gilt entsprechend

$$E_N(f) = \frac{b-a}{(n+1)!\ n} \left(\int_0^n \omega_n(t)\ dt \right) h^{n+1} f^{(n+1)}(\xi).$$

Dabei ist $h = \frac{b-a}{nN}$.

Beweis für gerades n:
$h^{n+3} \sum_{j=1}^N f^{(n+2)}(\xi_j) = h^{n+2} \cdot \frac{b-a}{n} \cdot \frac{1}{N} \sum_{j=1}^N f^{(n+2)}(\xi_j) = h^{n+2} \cdot \frac{b-a}{n} \cdot f^{(n+2)}(\xi)$,
wie vorhin erklärt wurde. Setzt man das in (7.2) ein, so erhält man (7.3).

Bemerkung 7.4 *Lassen wir bei konstantem Grad n die Anzahl N der Teilintervalle gegen unendlich gehen, so konvergiert h gegen 0 und daher auch $E_N(f)$ (da ja $f^{(n+2)}$ bzw. $f^{(n+1)}$ wegen der vorausgesetzten Stetigkeit beschränkt ist). Im Gegensatz zu den einfachen Newton-Cotes Formeln konvergieren hier also die Näherungswerte gegen den exakten Wert des Integrals, und der Fehler ist von der Ordnung $O(h^{n+2}) = O\left(\frac{1}{N^{n+2}}\right)$.*

Bemerkung 7.5 *Der Exponent von h ist um 1 kleiner als bei den einfachen Newton-Cotes Formeln. Dieser kleine Nachteil wird durch den in der vorigen Bemerkung genannten Vorteil bei weitem aufgewogen.*

Am häufigsten werden wohl die zwei folgenden zusammengesetzten Formeln verwendet:

$n = 1$: **(Summierte Sehnentrapezregel)**

$$I_N(f) = \frac{h}{2} \sum_{j=1}^N (f(x_{j-1}) + f(x_j)) = \frac{h}{2} \left(f(x_0) + f(x_N) + 2 \sum_{j=1}^{N-1} f(x_j) \right)$$

mit $h = H = (b-a)/N$, $x_j = z_j = a + jh$ und dem Fehler

$$E_N(f) = -\frac{b-a}{12}\ h^2\ f''(\xi),$$

denn $\int_0^1 \omega_1(t)\ dt = -\frac{1}{6}$ (siehe einfache Sehnentrapezregel).

$n = 2$: **(Summierte Simpson'sche Regel)**

Mit $h = \frac{b-a}{2N}$ und $x_k := a + kh$ ist hier $z_j = x_{2k}$, und es gilt

$$\begin{aligned} I_N(f) &= \frac{h}{3} \sum_{j=1}^{N} \left(f(z_{j-1}) + 4f\left(\frac{z_{j-1} + z_j}{2}\right) + f(z_j) \right) \\ &= \frac{h}{3} \left(f(x_0) + f(x_{2N}) + 2 \sum_{k=1}^{N-1} f(x_{2k}) + 4 \sum_{k=1}^{N} f(x_{2k-1}) \right) \end{aligned}$$

mit dem Fehler

$$E_N(f) = -\frac{b-a}{180}\ h^4\ f^{(4)}(\xi),$$

denn $\int_0^2 t\ \omega_2(t)\ dt = -\frac{4}{15}$ (siehe einfache Simpson'sche Regel).

7.3 Romberg-Integration

Sei $T(h)$ der Wert, der sich bei Anwendung der summierten Sehnentrapezregel mit Schrittweite h ergibt, also

$$T(h) := \frac{h}{2} \left(f(x_0) + f(x_N) + 2 \sum_{j=1}^{N-1} f(x_j) \right)$$

mit $N = \frac{b-a}{h}$ und $x_j = a + jh$. Dann gilt

$$\lim_{h \to 0} T(h) = \int_a^b f(x)\ dx.$$

(Das ist richtig, obwohl $T(h)$ nur für diejenigen h definiert ist, für welche $(b-a)/h$ ganzzahlig ist.)

Der Grundgedanke der Romberg-Integration besteht nun darin, diesen Limes mit Hilfe der in Kapitel 5.5.1 beschriebenen Richardson-Extrapolation zu berechnen. Dabei ist die Tatsache von Bedeutung, dass in der Taylor-Entwicklung von $T(h)$ nur gerade Potenzen vorkommen. Das ergibt sich aus dem folgenden Satz, der auch an sich interessant ist und eine Möglichkeit zur Fehlerabschätzung liefert. Einen Beweis findet man z.B. in [8].

Satz 7.6 (Euler[5]-Maclaurin[6]-Entwicklung) *Sei f auf $[a,b]$ $2m$-mal stetig differenzierbar. Dann gibt es ein $\xi \in [a,b]$, sodass*

$$T(h) = \int_a^b f(x)\ dx + c_2 h^2 + c_4 h^4 + \ldots + c_{2m-2} h^{2m-2} + R(\xi)\ h^{2m},$$

mit

$$c_{2k} = \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right)$$

und

$$R(\xi) = \frac{B_{2m}}{(2m)!}\ f^{(2m)}(\xi)\ (b-a).$$

Bemerkung 7.7 *Mit B_{2k} werden die Bernoulli'schen Zahlen[7] bezeichnet, die durch folgende "erzeugende Funktion" definiert werden können:*

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n,$$

das heißt, B_n ist definitionsgemäß gleich der n-ten Ableitung der Funktion $\frac{x}{e^x-1}$ an der Stelle 0 (wo sie allerdings nur als Grenzwert definiert ist). Einige Werte sind (vgl. z.B.[10]):

$$\begin{aligned} B_0 &= 1, \quad B_1 = -\frac{1}{2}, \quad B_{2k+1} = 0 \quad \textit{für alle } k \in \mathbb{N}, \\ B_2 &= \frac{1}{6}, \quad B_4 = -\frac{1}{30}, \quad B_6 = \frac{1}{42}. \end{aligned}$$

Die folgende asymptotische Formel ergibt für größere k eine sehr gute Näherung:

$$B_{2k} \sim (-1)^{k+1}\ 2\ \frac{(2k)!}{(2\pi)^{2k}}$$

(siehe z.B. [1]).

[5] Leonhard Euler (1707 - 1783, Basel, Sankt Petersburg) ist einer der berühmtesten Mathematiker und Physiker (siehe [9]).

[6] Colin Maclaurin (1698 - 1746, Schottland) hat die Differenzial- und Integralrechnung von Newton weiterentwickelt und beschäftigte sich insbesondere auch mit Taylor-Reihen (siehe [9]).

[7] Bernoulli ist der Name einer Reihe hervorragender Mathematiker. Hier handelt es sich um Jacob Bernoulli (1654 - 1705, Basel). Er trug gemeinsam mit seinem Bruder Johann entscheidend zur Anwendung der Infinitesimalrechnung auf Geometrie und Mechanik bei, förderte die Wahrscheinlichkeitsrechnung und lieferte wichtige Beiträge zur Theorie der Differenzialgleichungen (siehe [9]).

Aus der letzten Formel erkennen wir, dass die Koeffizienten $\frac{B_{2k}}{(2k)!}$ exponentiell gegen Null gehen. Trotzdem konvergiert die Euler-MacLaurin-Entwicklung im Allgemeinen nicht gegen den gewünschten Wert. So sind z.B. für eine periodische Funktion f (mit Periode $b-a$) alle Koeffizienten c_{2k} gleich Null, sodass das Restglied immer gleich dem Integrationsfehler ist.

Bemerkung 7.8 *Bei der Romberg-Integration kommt es nur darauf an, dass für festes m (und $h \to 0$) gilt:*

$$T(h) = \int_a^b f(x)\ dx + c_2 h^2 + c_4 h^4 + \ldots + c_{2m-2} h^{2m-2} + O(h^{2m}).$$

Für die Berechnung des Integrals können wir daher ganz analog zur numerischen Differenziation folgendes Rekursionsverfahren verwenden:

$$T_{i0} = T(h_i) \quad \text{mit } h_i = (b-a)/2^{i-1} \ \text{ für } i = 1, 2, 3, \ldots$$

und

$$T_{ik} = \frac{4^k T_{i,k-1} - T_{i-1,k-1}}{4^k - 1} \ \text{für } 0 < k < i.$$

Bemerkung 7.9 *Man kann beweisen, dass die "Diagonalelemente" $T_{i,i-1}$ für jede Riemann-integrierbare Funktion gegen das Integral konvergieren (R. Bulirsch 1964, siehe [6]). Im Allgemeinen konvergieren sie sogar ziemlich rasch, es kann aber passieren, dass sie nicht schneller als die Trapezsummen T_{i0} konvergieren.*

Es ist vielleicht interessant, zu bemerken, dass in der zweiten Spalte des Rombergschemas (d.h. für $k = 1$) gerade die summierte Simpson-Formel aufscheint. Mit $N = 2^{i-2}$, also $2N = 2^{i-1}$, und $x_j = a + j\frac{b-a}{2^{i-1}}$ gilt nämlich

$$\begin{aligned} T(h_i) &= \frac{h_i}{2}\left(f(x_0) + f(x_{2N}) + 2\sum_{j=1}^{2N-1} f(x_j)\right), \\ T(h_{i-1}) &= h_i\left(f(x_0) + f(x_{2N}) + 2\sum_{j=1}^{N-1} f(x_{2j})\right), \end{aligned}$$

und daher

$$\begin{aligned} T_{i1} &= \frac{1}{3}(4T(h_i) - T(h_{i-1})) \\ &= \frac{h_i}{3}\left(f(x_0) + f(x_{2N}) - 2\sum_{j=1}^{N-1} f(x_{2j}) + 4\sum_{j=1}^{2N-1} f(x_j)\right), \end{aligned}$$

und das stimmt mit der Simpson'schen Regel überein. (Man beachte, dass die Summanden mit geradem Index in beiden Summen vorkommen und daher insgesamt mit dem Faktor +2 auftreten.)

Man kann daher das Romberg-Schema auch mit der Simpson'schen Regel beginnen, dann erhält man gleich die Spalte mit $k = 1$ und kann im nächsten Schritt schon die Spalte mit $k = 2$ berechnen.

Beispiel: Wir wollen das Integral

$$\int_0^2 e^{-x^2} dx$$

näherungsweise berechnen. Mit der einfachen Simpson'schen.Regel erhalten wir

$T_{21} = \frac{1}{3}(1 + 4e^{-1} + e^{-4}) \approx 0.829\,944.$

Als nächstes verwenden wir die summierte Simpson-Formel mit $N = 2$:

$T_{31} = \frac{1}{6}(1 + 4e^{-0.25} + 2e^{-1} + 4e^{-2.25} + e^{-4}) \approx 0.881\,812.$

Daraus ergibt sich der Extrapolationswert $T_{32} = (16\,T_{31} - T_{21})/15 \approx 0.885\,270.$

Setzen wir nun das Romberg-Verfahren fort, so erhalten wir

$T_{41} \approx 0.882\,066, \quad T_{42} = (16\,T_{41} - T_{31})/15 \approx 0.882\,082,$

$T_{43} = (64\,T_{42} - T_{32})/63 \approx 0.882\,032.$

Der genaue Wert ist $0.882\,081\,390\ldots$, und dem kommt T_{42} am nächsten. Hier sieht man also, dass nicht unbedingt die Diagonalelemente die besten Näherungen ergeben. Wenn man bereits nach der 3. Zeile abbricht, so ist T_{31} und nicht T_{32} der beste Näherungswert.

Trotz der an diesem Beispiel gezeigten Problematik wird man aber normalerweise die Diagonalfolge nehmen (vgl. obige Bemerkung 7.9).

Kapitel 8

Iterative Lösung von Gleichungen

8.1 Das Kontraktionsprinzip

8.1.1 Allgemeines

Es geht hier um Verfahren zur Lösung von (normalerweise nicht-linearen) Gleichungen. Eine solche Gleichung hat im Allgemeinen die Form

$$f(x) = 0.$$

Man kann sie aber auf verschiedene Weise auf die Form

$$\varphi(x) = x$$

bringen, und für solche Gleichungen gibt es unter bestimmten Voraussetzungen sehr effiziente Lösungsverfahren. Die Lösungen heißen dann *Fixpunkte* von φ, die Form $\varphi(x) = x$ daher auch *Fixpunktgleichung.*

Beispiel: Die Gleichung

$$x^3 + 4x^2 - 10 = 0$$

kann z.B. auf folgende Arten in eine Fixpunktgleichung übergeführt werden:

a) Addition von x auf beiden Seiten:

$$x^3 + 4x^2 + x - 10 = x.$$

b) Wenn man sich nur für positive Lösungen interessiert, kann man x^2 isolieren und dann die Wurzel ziehen:

$$x = \frac{1}{2}\sqrt{10 - x^3}.$$

c) Die folgende Umformung schließt ebenfalls eventuelle negative Lösungen aus:

$$x = \sqrt{\frac{10}{x+4}}.$$

Sie entsteht, indem man zuerst x^2 aus den beiden ersten Summanden heraushebt.

Eine sehr allgemeine Methode der Überführung von $f(x) = 0$ in eine Fixpunktgleichung besteht darin, eine Funktion h zu wählen, die in dem betrachteten Bereich keine Nullstellen besitzt, und dann

$$\varphi(x) := x - h(x)\ f(x)$$

zu setzen. Die Fixpunkte von φ sind dann genau die Nullstellen von f.

In vielen Fällen erhält man die Lösung einer Fixpunktgleichung durch folgendes einfache *Iterationsverfahren (Picard-Iteration)*[1]. Man wählt einen Startwert x_0 und berechnet dann sukzessive $x_1, x_2, \ldots$ nach der Vorschrift

$$x_{n+1} := \varphi(x_n).$$

Wenn diese Folge konvergiert und φ stetig ist, so ist der Limes ein Fixpunkt:

$$\xi = \lim x_n \ \Rightarrow\ \varphi(\xi) = \lim \varphi(x_n) = \lim x_{n+1} = \xi.$$

Versuchen wir das etwa bei den obigen Beispielen b) und c), so erhalten wir mit $x_0 = 1$ (gerundet) die Folgen

(1, 1.5, 1.28695, 1.40254, 1.34546, 1.37517, 1.36009, ...)

bzw.

(1, 1.41421, 1.35904, 1.36602, 1.36513, 1.36524, 1.36523, ...).

Die zweite Folge konvergiert offensichtlich viel schneller gegen die Lösung $x = 1.365230\,0134\ldots$.

Bei Beispiel a) erhalten wir dagegen keine konvergente Folge:

(1, −4, −14, −1984, −7 793 788 874, ...).

Es kommt also unter Umständen sehr darauf an, auf welche Art man die Überführung in eine Fixpunktgleichung durchführt.

Die Bedeutung der Picard-Iteration liegt unter anderem darin, dass sie nicht nur für reelle Funktionen, sonder sehr viel allgemeiner anwendbar ist, nämlich in beliebigen vollständigen metrischen Räumen.

[1] Charles Emile Picard (1856 - 1941, Paris) benutzte diese Iterationsmethode, um die Existenz der Lösung von gewöhnlichen Differentialgleichungen zu zeigen (vgl. [9]).

Zunächst sei daran erinnert, was man hier unter "vollständig" versteht: Ein metrischer Raum heißt *vollständig*, wenn in ihm jede Cauchy-Folge konvergiert. Eine Folge (x_n) in einem metrischen Raum (M, d) heißt dabei *Cauchy-Folge*[2], wenn es zu jedem $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$ gibt, sodass gilt:

$$n, m \geq n_\varepsilon \;\Rightarrow\; d(x_n, x_m) < \varepsilon.$$

Bekanntlich ist der $\mathbb{R}^n$ mit der euklidischen Metrik vollständig, aber auch z.B. mit den Metriken, die durch die 1-Norm bzw. ∞-Norm induziert werden:

$$\begin{aligned} d_1(x, y) &:= \|x - y\|_1 , \\ d_\infty(x, y) &:= \|x - y\|_\infty . \end{aligned}$$

Jede abgeschlossene Teilmenge eines vollständigen metrischen Raumes ist, für sich betrachtet, ebenfalls ein vollständiger metrischer Raum.

Definition 8.1 *Sei (M, d) ein metrischer Raum und*

$$\varphi : M \to M$$

eine Abbildung von M in sich. Wenn es eine Zahl $L > 0$ gibt, sodass

$$d(\varphi(x), \varphi(y)) \leq L\, d(x, y) \quad \textit{für alle } x, y \in M,$$

so sagt man, φ erfüllt eine ***Lipschitz-Bedingung***[3] *mit der* ***Lipschitz-Konstanten*** *L. Wenn $L < 1$ ist, heißt φ eine (strikte)* ***Kontraktion*** *(des metrischen Raumes (M, d)).*

1. Bemerkung: Jede Funktion, die eine Lipschitz-Bedingung erfüllt, ist gleichmäßig stetig.

2. Bemerkung: Aus $d(\varphi(x), \varphi(y)) < d(x, y)$ für alle $x, y \in M$ folgt noch nicht, dass φ eine Kontraktion ist (siehe Übungsaufgaben).

Der folgende Satz ist grundlegend für alle Anwendungen der Picard-Iteration:

[2] Augustin Louis Cauchy (1789 - 1857, Paris) war einer der vielseitigsten Mathematiker (siehe [9]). Sein Name ist z.B. auch durch die Cauchy'sche Integralformel der Funktionentheorie bekannt.

[3] Von Rudolf Lipschitz (1832 - 1903, Königsberg, Bonn) stammen bedeutende Arbeiten über (algebraische) Zahlentheorie und (gewöhnliche und partielle) Differentialgleichungen, siehe [9].

Satz 8.2 (Fixpunktsatz von Banach[4]) *Sei φ eine Kontraktion des vollständigen metrischen Raumes (M, d) mit der Lipschitzkonstanten $L < 1$. Dann gibt es genau einen Fixpunkt ξ von φ, und für jedes $x_0 \in M$ konvergiert die durch*

$$x_{n+1} = \varphi(x_n)$$

definierte Folge gegen ξ. Es gelten dabei folgende Fehlerabschätzungen:

$$\begin{aligned} &1)\ d(x_n, \xi) \ \leq\ \frac{L^n}{1-L}\, d(x_0, x_1) \qquad \textit{(a-priori-Fehlerschranke),} \\ &2)\ d(x_n, \xi) \ \leq\ \frac{L}{1-L}\, d(x_{n-1}, x_n) \qquad \textit{(a-posteriori-Fehlerschranke).} \end{aligned}$$

Beweis: Zunächst ist leicht einzusehen, dass es höchstens einen Fixpunkt gibt: Wären nämlich ξ und η zwei verschiedene Fixpunkte von φ, so wäre

$$0 < d(\xi, \eta) = d(\varphi(\xi), \varphi(\eta)) \leq L\ d(\xi, \eta),$$

also $1 \leq L$, im Widerspruch zur vorausgesetzten Kontraktionseigenschaft von φ.
Da

$$d(x_s, x_{s+1}) = d(\varphi(x_{s-1}), \varphi(x_s)) \leq L\ d(x_{s-1}, x_s) \quad \text{für alle } s \in \mathbb{N},$$

sieht man mit vollständiger Induktion, dass

$$d(x_{n+k}, x_{n+k+1}) \leq L^k\ d(x_n, x_{n+1}) \quad \text{für alle } n, k \in \mathbb{N}.$$

Daraus folgt nun für $n < m$ mit der Dreiecksungleichung

$$\begin{aligned} d(x_n, x_m) \ &\leq\ d(x_n, , x_{n+1}) + d(x_{n+1}, x_{n+2}) + d(x_{n+2}, x_{n+3}) + \ldots + d(x_{m-1}, x_m) \\ &\leq\ (1 + L + L^2 + \ldots + L^{m-1-n})\ d(x_n, x_{n+1}) \\ &\leq\ \frac{1}{1-L}\, d(x_n, x_{n+1}) \\ &\leq\ \frac{1}{1-L}\, L^n\ d(x_0, x_1). \end{aligned}$$

Wegen $0 < L < 1$ geht L^n gegen Null, und daher folgt, dass (x_n) eine Cauchy-Folge ist und somit gegen ein $\xi \in M$ konvergiert. Auf Grund der Stetigkeit der Abbildung d erhalten wir mit $m \to \infty$

$$d(x_n, \xi) \leq \frac{L^n}{1-L} d(x_0, x_1)$$

[4]Der polnische Mathematiker Stefan Banach (1892 - 1945, Krakau, Lemberg [Lvov]) gilt als Begründer der modernen Funktionalanalysis. Dort spielen die vollständigen normierten Vektorräume, welche heute nach ihm benannt werden, eine zentrale Rolle (siehe [9]).

und

$$d(x_n, \xi) \leq \frac{1}{1-L}\ d(x_n, x_{n+1}) \leq \frac{1}{1-L}\ L\ d(x_{n-1}, x_n).$$

■

Bemerkung: Auf Grund der a-priori-Fehlerschranke kann man eine Schranke für die Anzahl der Iterationen angeben, die man für die Erreichung einer bestimmten Genauigkeit braucht:

$\frac{1}{1-L} L^n d(x_0, x_1) < \varepsilon \ \Leftrightarrow\ L^n < \frac{\varepsilon(1-L)}{d(x_0,x_1)} \ \Leftrightarrow\ n \log L < \log \varepsilon + \log \frac{1-L}{d(x_0,x_1)}$.

Wegen $\log L < 0$ folgt daraus:

$$\text{Für } n > \frac{-\log \varepsilon + \log \frac{d(x_0,x_1)}{1-L}}{|\log L|} \text{ ist sicher } d(x_n, \xi) < \varepsilon.$$

Wir können das auch so ausdrücken: Für die Anzahl n_ε der Iterationsschritte, die nötig sind, um die (absolute) Genauigkeit ε zu erreichen, gilt

$$n_\varepsilon \leq \left\lceil \frac{1}{|\log L|} \left(-\log \varepsilon + \log \frac{d(x_0, x_1)}{1-L} \right) \right\rceil . \tag{8.1}$$

8.1.2 Anwendung des Kontraktionsprinzips im $\mathbb{R}^s$

(Wir schreiben hier $\mathbb{R}^s$ statt dem üblichen $\mathbb{R}^n$, da wir n als Index der Iterationsfolge verwenden.)

Wir benötigen die folgende Version des Mittelwertsatzes der Differenzialrechnung für Abbildungen von $\mathbb{R}$ in den $\mathbb{R}^s$.

Satz 8.3 *Sei $[a, b]$ ein kompaktes Intervall und $g : [a, b] \to \mathbb{R}^s$ eine differenzierbare Abbildung. Wenn es eine Zahl K gibt, sodass*

$$\|g'(t)\| \leq K \quad \textit{für alle } t \in [a, b],$$

dann gilt

$$\|g(b) - g(a)\| \leq K \cdot (b - a).$$

(Dabei kann für $\|.\|$ irgendeine Norm des $\mathbb{R}^s$ genommen werden, also z.B. die 1-Norm, die 2-Norm oder die ∞-Norm.)

Beweis: Sei $g_i(t)$ die i-te Koordinate von $g(t)$. Nach dem Hauptsatz der Differenzial- und Integralrechnung gilt für jedes $i \in \{1, \ldots, s\}$

$$\int_a^b g_i'(t)\ dt = g_i(b) - g_i(a).$$

Diese s Gleichungen können wir zu einer Vektorgleichung zusammenfassen:

$$\int_a^b g'(t)\ dt = g(b) - g(a).$$

Nun gilt aber

$$\left\| \int_a^b g'(t)\ dt \right\| \leq \int_a^b \|g'(t)\|\ dt,$$

da für die entsprechenden Riemann'schen Summen die folgende Ungleichung gilt:

$$\left\| \sum_j g'(\xi_j)(t_{j+1} - t_j) \right\| \leq \sum_j \left\| g'(\xi_j) \right\| (t_{j+1} - t_j).$$

(Das ergibt sich aus der verallgemeinerten Dreiecksungleichung $\|\sum a_j\| \leq \sum \|a_j\|$.)
Wir erhalten daher

$$\|g(b) - g(a)\| = \left\| \int_a^b g'(t)\ dt \right\| \leq \int_a^b \|g'(t)\|\ dt \leq \int_a^b K\ dt = K \cdot (b - a).$$

■

Damit können wir nun den folgenden Mittelwertsatz für Abbildungen vom $\mathbb{R}^m$ in den $\mathbb{R}^s$ beweisen:

Satz 8.4 *Sei $\varphi : \mathbb{R}^m \supset D \to \mathbb{R}^s$ eine stetig differenzierbare Abbildung. Wenn die Verbindungsstrecke $[x, y]$ von x und y in D enthalten ist, dann gilt*

$$\|\varphi(x) - \varphi(y)\| \leq \left(\max_{z \in [x,y]} \|\varphi'(z)\| \right) \|x - y\|.$$

Beweis: Sei

$$g(t) := \varphi(x + t(y - x)) \quad \text{für } t \in [0, 1].$$

Dann ist nach der "Kettenregel"

$$g'(t) = \varphi'(x + t(y - x))\ (y - x)$$

und daher (nach (3.1))

$$\|g'(t)\| \leq \|\varphi'(x + t(y - x))\|\ \ \|y - x\|.$$

Nach dem vorigen Satz gilt somit

$$\|g(1) - g(0)\| \leq \max_{t \in [0,1]} \|\varphi'(x + t(y - x))\|\ \ \|y - x\|\ (1 - 0).$$

Da $g(0) = \varphi(x)$ und $g(1) = \varphi(y)$, folgt daraus die Behauptung. ■

Wenn eine Teilmenge des $\mathbb{R}^s$ mit je zwei Punkten auch deren Verbindungsstrecke enthält, nennt man sie *konvex*.

Damit können wir jetzt den zentralen Satz dieses Teilkapitels formulieren und beweisen:

Satz 8.5 *Sei D eine abgeschlossene konvexe Teilmenge des $\mathbb{R}^s$, und φ sei eine stetig differenzierbare Abbildung $D \to D$. Wenn es ein $L < 1$ gibt, sodass*

$$\|\varphi'(z)\| \leq L \quad \textit{für alle } z \in D,$$

dann ist φ eine (strikte) Kontraktion auf D (mit der Lipschitzkonstanten L).

Beweis: Seien $x, y \in D$. Nach dem vorigen Satz folgt aus $\|\varphi'(z)\| \leq L$ sofort $\|\varphi(x) - \varphi(y)\| \leq L\,\|x - y\|$. ■

1. Beispiel (mit $s = 1$):

Wir wollen die Gleichung

$$\sin x = 1 - x$$

lösen. Die einfachste zugehörige Fixpunktgleichung ist

$$1 - \sin x = x.$$

Wir brauchen nun ein Intervall, das durch die Abbildung

$$\varphi(x) := 1 - \sin x$$

in sich abgebildet wird. Dazu bieten sich etwa $[0, 1]$ oder $[0, \pi/2]$ an. Allerdings ist φ auf diesen Intervallen nicht (strikt) kontrahierend, da $\varphi'(0) = -1$ ist. Wir nehmen daher z.B. $I = [0.1,\ 1]$. Dieses Intervall wird auch durch φ in sich abgebildet, denn $\varphi(0.1) = 1 - \sin 0.1 = 0.90\ldots,\ \varphi(1) = 0.15\ldots$, und φ ist auf I monoton fallend (da $\varphi'(x) = -\cos x < 0$). Auf dem Intervall I ist nun

$$|\varphi'(x)| = |\cos x| \leq |\cos 0.1| = 0.995\ldots \ =: L.$$

Die Iterationsfolge konvergiert daher für jeden Startwert $x_0 \in I$. Zum Beispiel erhält man für $x_0 = 0.5$ (gerundet):

$x_1 = 0.52057, x_5 = 0.51653,\ x_{10} = 0.50817, x_{15} = 0.51239,\ x_{20} = 0.51026,\ x_{25} = 0.51133$.

Der genaue Wert des Fixpunkts ist $\xi = 0.51097\ldots$. Man sieht also, dass die Konvergenz ziemlich langsam erfolgt.

Aus der a-priori-Fehlerschranke erhalten wir z.B. für $\varepsilon = 0.0001$ die folgende Abschätzung für die Anzahl n_ε der nötigen Iterationsschritte (siehe (8.1)):

$$n_\varepsilon \leq \left\lceil \frac{1}{|\log L|} \left(-\log \varepsilon + \log \frac{d(x_0, x_1)}{1 - L} \right) \right\rceil = 2122.$$

Der tatsächliche Fehler wird hier allerdings sehr überschätzt. In Wirklichkeit braucht man nur 35 Iterationen, um den Fehler kleiner als 0.0001 zu machen. Das liegt daran, dass in der Nähe der Lösung für L der wesentlich bessere Wert $\varphi'(0.5) = \cos 0.5 = 0.877\ldots$ genommen werden könnte. Wenn wir z.B. vom Intervall $[0.5,\ 0.521]$ ausgehen (das auch durch φ in sich abgebildet wird), dann erhalten wir $n_\varepsilon \leq 57$.

2. Beispiel (mit $s = 2$):

Sei

$$\varphi(x, y) = 0.9\ e^{-y}\ (\cos x, \sin x).$$

φ bildet offensichtlich das Einheitsquadrat $[0, 1]^2$ in sich ab. Um festzustellen, ob es sich um eine Kontraktion handelt, berechnen wir

$$\varphi'(x, y) = \begin{pmatrix} \frac{\partial \varphi_1}{\partial x} & \frac{\partial \varphi_1}{\partial y} \\ \frac{\partial \varphi_2}{\partial x} & \frac{\partial \varphi_2}{\partial y} \end{pmatrix} (x, y) = 0.9\ e^{-y} \begin{pmatrix} -\sin x & -\cos x \\ \cos x & -\sin x \end{pmatrix}.$$

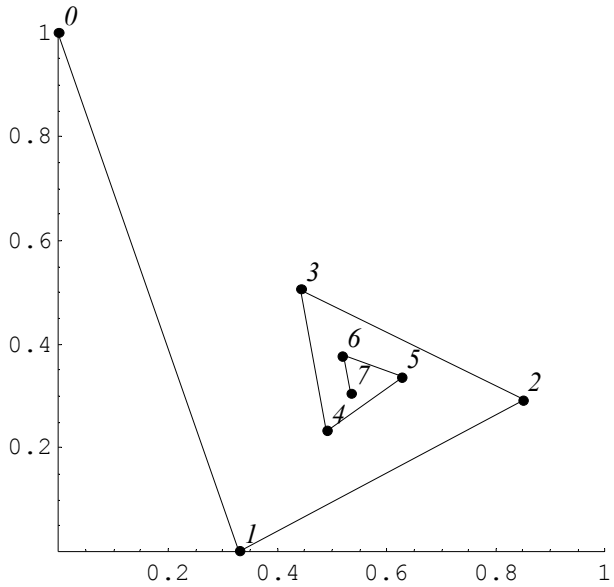
$\|\varphi'(x, y)\|_1 = \|\varphi'(x, y)\|_\infty = 0.9\ e^{-y}(\sin x + \cos x)$. Das ergibt z.B. für $x = \pi/4$ und $y = 0$ den Wert $1.272\ldots$. Bezüglich der 1- und ∞-Norm liegt also keine Kontraktion vor. Sehen wir uns nun die 2-Norm an, die hier leicht zu berechnen ist.

Sei $A := \begin{pmatrix} -\sin x & -\cos x \\ \cos x & -\sin x \end{pmatrix}$. Das ist eine orthogonale Matrix, daher ist $\|A\|_2 = 1$ (siehe Beispiel 3.1). Wir sehen also:

$$\|\varphi'(x, y)\|_2 = 0.9\ e^{-y} \cdot 1 \leq 0.9 =: L$$

für alle $y \geq 0$. Die folgende Abbildung zeigt die ersten 8 Punkte der Iterationsfolge für den Startpunkt $(0, 1)$, wobei aufeinanderfolgende Punkte durch

eine Strecke verbunden wurden.



Es ergibt sich z.B. (gerundet) $x_9 = (0.54053, 0.34643),\ x_{10} = (0.54575, 0.32753)$. Mit diesen beiden Werten können wir die folgende a-posteriori-Fehlerabschätzung gewinnen:

$d(x_{10}, \xi) \leq \frac{0.9}{0.1}\ d(x_9, x_{10}) = 9 \left\| x_9 - x_{10} \right\|_2 = 9 \cdot 0.019609 = 0.17648.$

Die exakte Lösung ist $\xi = (0.54887\ldots, 0.33566\ldots)$, somit ist der tatsächliche Fehler $d(x_{10}, \xi) = 0.00871\ldots$ wesentlich kleiner.

Aus der a-priori-Fehlerabschätzung ergibt sich z.B. folgende Schranke für n_ε mit $\varepsilon = 0.0001$:

$$n_\varepsilon \leq \left\lceil \frac{1}{|\log 0.9|} \left(-\log \varepsilon + \log \frac{d(x_0, x_1)}{0.1} \right) \right\rceil = 110.$$

In Wirklichkeit braucht man allerdings nur 21 Iterationen, um diese Genauigkeit zu erreichen.

8.1.3 Konvergenzordnung

Bei der Picard-Iteration einer Kontraktion wird der (absolute) Fehler, also der Abstand zum Fixpunkt, bei jedem Iterationsschritt mindestens um einen bestimmten Faktor kleiner:

$$d(x_{n+1}, \xi) = d(\varphi(x_n), \varphi(\xi)) \leq L\ d(x_n, \xi).$$

Der Fehler ist hier also in jedem Schritt durch eine feste lineare Funktion des vorhergehenden Fehlers beschränkt. Es gibt aber auch Verfahren, bei denen der Fehler jeweils durch das Quadrat oder eine andere Potenz des vorhergehenden Fehlers abgeschätzt werden kann. Zur Angabe der Konvergenzgeschwindigkeit verwendet man daher folgenden Begriff.

Definition 8.6 *Sei $p \geq 1$ und (x_n) eine gegen ξ konvergente Folge in einem metrischen Raum (M, d). Man sagt, diese Folge konvergiert mit der Ordnung $p \geq 1$, wenn es eine Konstante $K > 0$ und ein $n_0 \in \mathbb{N}$ gibt, sodass*

$$d(x_{n+1}, \xi) \leq K \cdot (d(x_n, \xi))^p \quad \text{für alle } n \geq n_0,$$

*wobei für $p = 1$ zusätzlich $K < 1$ verlangt wird. Im Falle $p = 1$ sprechen wir auch von **linearer**, für $p = 2$ von **quadratischer** Konvergenz.*

Bemerkung 8.7 *Manchmal verwendet man den folgenden, etwas modifizierten Begriff: Sei (x_n) eine Folge in M und (a_n) eine Folge positiver reeller Zahlen, die in obigem Sinne mit der Ordnung p gegen Null konvergiert. Wenn es eine Konstante $c \in \mathbb{R}$ gibt, sodass*

$$d(x_n, \xi) \leq c \cdot a_n \quad \text{für alle } n \geq n_0,$$

dann sagt man, dass (x_n) mit der Ordnung p gegen ξ konvergiert, auch wenn die Bedingung von Definition 8.6 nicht erfüllt ist.

Hier ist ein Beispiel einer Folge, die nur im Sinne dieser Variante linear gegen Null konvergiert:

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4^2}, \frac{1}{4^2}, \frac{1}{4^3}, \frac{1}{4^3}, \dots\right).$$

(Man kann hier $a_n = 2^{-n}$ und $c = n_0 = 1$ wählen.)

Der folgende Satz gibt nun eine hinreichende Bedingung dafür an, dass eine Iterationsfolge quadratisch konvergiert.

Satz 8.8 *Sei D eine abgeschlossene Teilmenge des $\mathbb{R}^s$ und $\varphi : D \to D$ eine Kontraktion mit einem Fixpunkt ξ im Inneren von D, die in einer Umgebung $U \subset D$ von ξ zweimal stetig differenzierbar ist. Wenn $\varphi'(\xi) = O$ ist, so konvergiert die durch $x_{n+1} = \varphi(x_n)$ definierte Iterationsfolge für jeden Startpunkt $x_0 \in D$ quadratisch gegen ξ (bezüglich der Maximumsnorm).*

Beweis: Wir können annehmen, dass U eine Kugel bezüglich der betrachteten Norm ist, d.h.

$$U = \{x \in \mathbb{R}^s : \|x - \xi\|_\infty \leq \delta\} \text{ für ein } \delta > 0.$$

(Jede Umgebung von ξ enthält nämlich so eine "Kugel", die eigentlich ein s-dimensionaler Würfel ist.)

Nach dem Taylor'schen Satz gilt für $x = (x_1, \ldots, x_s) \in U$ wegen $\varphi'(\xi) = O$: Zu jedem $i \in \{1, \ldots, s\}$ gibt es einen Punkt $\zeta(i) \in [\xi, x]$, sodass

$$\varphi_i(x) - \xi_i = \varphi_i(x) - \varphi_i(\xi) = \frac{1}{2} \sum_{j,k=1}^{s} \frac{\partial^2 \varphi_i}{\partial x_j \partial x_k}(\zeta(i)) \; (x_j - \xi_j)(x_k - \xi_k).$$

Aus der Stetigkeit der zweiten Ableitungen folgt die Existenz einer Konstanten K mit

$$\left| \frac{\partial^2 \varphi_i}{\partial x_j \partial x_k}(z) \right| \leq K \quad \text{für alle } z \in U \text{ und alle } i, j, k \in \{1, \ldots, s\}.$$

Da hier $x_1, x_2, \ldots$ die Koordinaten von x sind, bezeichnen wir die Iterationsfolge mit

$$x^{(0)}, x^{(1)}, x^{(2)}, \ldots \; .$$

Da diese Folge auf jeden Fall gegen ξ konvergiert, gibt es ein n_0, sodass $x^{(n)} \in U$ für alle $n \geq n_0$. Aus der obigen Taylor-Formel folgt daher mit $x^{(n)}$ an Stelle von x:

$$\left\| x^{(n+1)} - \xi \right\|_\infty \leq \frac{s^2}{2} K \left\| x^{(n)} - \xi \right\|_\infty^2 .$$

(Dabei haben wir benützt, dass U konvex ist: Aus $x^{(n)} \in U$ und $\xi \in U$ folgt $[\xi, x^{(n)}] \subset U$ und daher $\zeta(i) \in U$.) ■

Bemerkung: Auf die Voraussetzung, dass φ eine Kontraktion sein soll, kann man in gewissem Sinne verzichten. Wenn φ' stetig ist und $\varphi'(\xi) = O$, gibt es nämlich für $0 < L < 1$ auf jeden Fall eine Kugel-Umgebung U von ξ, wo $\|\varphi'(x)\| \leq L$. Für alle $x \in U$ ist dann nach Satz 8.5

$$\|\varphi(x) - \xi\| = \|\varphi(x) - \varphi(\xi)\| \leq L \|x - \xi\| < \|x - \xi\| .$$

φ bildet also U auf sich ab. Für alle $x, y \in U$ gilt aber wiederum nach Satz 8.5

$$\|\varphi(x) - \varphi(y)\| \leq L \|x - y\| ,$$

und daher ist φ eine Kontraktion auf U. Man muss dann nur sicher sein, dass der Startpunkt in U liegt. Das ist allerdings in der Praxis oft schwer überprüfbar.

Im Eindimensionalen diskutieren wir eine verschärfte Version dieses Satzes für den Fall von mehrfachen Nullstellen. Dazu:

Definition 8.9 *Sei f eine p-mal differenzierbare Funktion $\mathbb{R} \supset D \to \mathbb{R}$. Ein $\xi \in D$ heißt* ***p-fache Nullstelle*** *oder eine* ***Nullstelle der Ordnung p*** *von f, wenn es eine p-mal differenzierbare Funktion g mit $g(\xi) \neq 0$ gibt, sodass*

$$f(x) = (x - \xi)^p g(x) \quad \textit{für alle } x \in D.$$

Bemerkung 8.10 *Sei f eine p-mal differenzierbare Funktion $\mathbb{R} \supset D \to \mathbb{R}$. Wenn ξ eine p-fache Nullstelle von f ist, dann gilt*

$$f(\xi) = f'(\xi) = \ldots = f^{(p-1)}(\xi) = 0 \quad \textit{und} \quad f^{(p)}(\xi) \neq 0.$$

Beweis: Sei ξ eine p-fache Nullstelle von f. Dann ist jedenfalls $f(\xi) = 0$, und für $p > 1$ gilt:

$$f'(x) = p(x - \xi)^{p-1} g(x) + (x - \xi)^p g'(x) = (x - \xi)^{p-1}(p\ g(x) + (x - \xi)g'(x)).$$

Daher ist $f'(\xi) = 0$, und ξ ist eine $(p-1)$-fache Nullstelle von f'. Mit Induktion folgt die Behauptung. ■

Jetzt kommt die angekündigte verschärfte Version von Satz 8.8:

Satz 8.11 *Sei φ eine $(p+1)$-mal stetig differenzierbare Kontraktion des Intervalls $[a, b]$ mit Fixpunkt ξ, und $p > 1$. Wenn φ' in ξ eine $(p-1)$-fache Nullstelle hat, so konvergiert die durch $x_{n+1} = \varphi(x_n)$ definierte Iteration für jeden Startwert $x_0 \in [a, b]$ mit der Ordnung p gegen ξ.*

Beweis: Auf Grund der vorhergehenden Bemerkung (mit φ' an Stelle von f und $p - 1$ an Stelle von p) haben wir $\varphi^{(i)}(\xi) = 0$ für $i = 1, \ldots, p - 1$ und $\varphi^{(p)}(\xi) \neq 0$. Nach dem Taylor'schen Satz gilt somit

$$x_{n+1} = \varphi(x_n) = \varphi(\xi) + \frac{1}{p!}(x_n - \xi)^p \varphi^{(p)}(\xi) + o(|x_n - \xi|^p)$$

und daher

$$\lim_{n\to\infty} \frac{x_{n+1} - \xi}{(x_n - \xi)^p} = \frac{1}{p!}\ \varphi^{(p)}(\xi) =: c.$$

Da die rechte Seite eine von n unabhängige Konstante $\neq 0$ ist, folgt z.B.

$$|x_{n+1} - \xi| \leq 2\,|c|\,|x_n - \xi|^p$$

für genügend große n, und damit ist der Satz bewiesen. ■

8.2 Das Newton-Verfahren

Wir betrachten im Folgenden zunächst wieder Gleichungen der Form $f(x) = 0$ mit einer gewöhnlichen reellen Funktion (das heißt $f : \mathbb{R} \supset D \to \mathbb{R}$), die genügend oft stetig differenzierbar ist.

Bei der Erklärung des Kontraktionsprinzips haben wir gesehen, dass die Lösung einer Gleichung der Form $f(x) = 0$ äquivalent ist zur Bestimmung der Fixpunkte von

$$\varphi(x) := x - h(x)\ f(x)$$

mit einer beliebigen Funktion h, die nur in dem betrachteten Bereich keine Nullstellen haben darf. Auf Grund von Satz 8.8 versuchen wir nun, h so zu wählen, dass $\varphi'(\xi) = 0$ ist für die (bzw. eine) Lösung ξ. Wegen $f(\xi) = 0$ lautet diese Bedingung

$$\varphi'(\xi) = 1 - h(\xi)f'(\xi) = 0,$$

das heißt

$$h(\xi) = \frac{1}{f'(\xi)}.$$

Um das zu erreichen, setzen wir

$$h(x) := \frac{1}{f'(x)}$$

für alle x aus dem Definitionsbereich von f. Wir müssen dabei allerdings eventuell den Definitionsbereich so weit einschränken, dass er keine Nullstellen von f' enthält. Das ist genau dann möglich, wenn $f'(\xi) \neq 0$ ist.

Bei geeignetem Startwert x_0 konvergiert also die durch

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)}$$

definierte Folge *quadratisch* gegen eine Nullstelle ξ von f, falls $f'(\xi) \neq 0$ ist. Diese Methode zur Lösung von Gleichungen heißt *Newton-Verfahren*. Anschaulich wird dabei in jedem Schritt die Funktion näherungsweise durch ihre Tangente im Punkt $(x_n,\ f(x_n))$ ersetzt. x_{n+1} ist dann die Nullstelle dieser Näherung, das heißt der Schnittpunkt der Tangente mit der x-Achse.

Das Hauptproblem besteht natürlich in der Wahl eines geeigneten Startwertes.

Das Newton-Verfahren lässt sich ohne weiteres auf Funktionen von mehreren Variablen übertragen, das heißt auf die Lösung von Gleichungssystemen. Das ergibt folgenden Satz.

Satz 8.12 *Sei $f : \mathbb{R}^s \supset D \to \mathbb{R}^s$ eine dreimal stetig differenzierbare Funktion. Wenn die Gleichung $f(x) = o$ eine Lösung ξ im Inneren von D hat, sodass die Jacobi-Matrix $f'(\xi)$ regulär ist, dann gibt es eine Umgebung U von ξ, sodass für jedes $x_0 \in U$ die durch*

$$x_{n+1} := x_n - (f'(x_n))^{-1} f(x_n)$$

definierte Folge quadratisch gegen ξ konvergiert.

Beweis: Wenn $\det(f'(\xi)) \neq 0$ ist, gibt es wegen der Stetigkeit von f' und det eine Umgebung U_1 von ξ, sodass $\det(f'(x)) \neq 0$ für alle $x \in U_1$.

Sei

$$\varphi(x) := x - (f'(x))^{-1} f(x).$$

φ ist für alle $x \in U_1$ wohldefiniert und differenzierbar. Wegen $f(\xi) = o$ ist

$$\varphi'(\xi) = E - (f'(\xi))^{-1} f'(\xi) = O.$$

(Hier haben wir benützt, dass die bekannte Produktregel der Differenziation auch für das Produkt einer Matrix mit einem Vektor gilt.)

Da φ' stetig ist, gibt es zu $0 < L < 1$ eine Kugel-Umgebung $U_L \subset U_1$ von ξ, sodass

$$\|\varphi'(x)\|_\infty < L \quad \text{für alle } x \in U_L.$$

U_L wird durch φ auf sich abgebildet, also ist φ nach Satz 8.5 eine Kontraktion auf U_L. Da f als dreimal stetig differenzierbar vorausgesetzt wurde, ist φ mindestens zweimal stetig differenzierbar. Es sind also alle Voraussetzungen von Satz 8.8 erfüllt, und daher konvergiert die angegebene Iterationsfolge quadratisch gegen ξ. ■

1. Beispiel: Wir betrachten wieder die Gleichung $\sin x = 1 - x$ und schreiben sie jetzt in der Form

$$x + \sin x - 1 = 0.$$

Wir nehmen also $f(x) := x + \sin x - 1$ und berechnen mit dem Newton-Verfahren eine Nullstelle.

$f'(x) = 1 + \cos x$, die Iterationsvorschrift lautet also

$$x_{n+1} = x_n - \frac{x_n + \sin x_n - 1}{1 + \cos x_n}.$$

Mit dem Startwert $x_0 = 0.5$ erhalten wir die Folge

$$(0.5, \quad 0.510\,957\,953, \quad 0.510\,973\,429, \quad 0.510\,973\,429, \ \ldots),$$

wo also bereits x_2 auf 9 Stellen genau ist.

Wie sehr es auf die Wahl des Startwertes ankommt, sieht man hier z.B., wenn man $x_0 = 3$ wählt. Man erhält (mit der üblichen ungefähr 15-stelligen Standardgenauigkeit) die Folge

$(3.,\ -210.951,\ 1783.06,\ 301.852,\ 148.737,\ -128.001,\ 294.923,\ \ldots),$

die auch nach 100 Iterationsschritten noch keinerlei Konvergenzverhalten zeigt. Der Grund dafür liegt darin, dass der Startwert hier in der Nähe einer Nullstelle der Ableitung liegt und daher die entsprechende Tangente fast waagrecht ist: $f'(\pi) = 0, \quad f'(3) \approx 0.01$.

2. Beispiel (mit $s = 2$): Wir nehmen wieder dasselbe Beispiel wie im Abschnitt 8.1.2. Es geht also um die Nullstellen der Abbildung

$$f : \mathbb{R}^2 \to \mathbb{R}^2 : (x, y) \mapsto (x, y) - 0.9\ e^{-y}(\cos x, \sin x).$$

Mit

$$f'(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.9\ e^{-y} \begin{pmatrix} -\sin x & -\cos x \\ \cos x & -\sin x \end{pmatrix}$$

und dem Startwert $(0, 1)$ erhalten wir die Iterationsfolge

$$\begin{array}{ll} (0., & 1.), \\ (0.596765, & 0.197584), \\ (0.549291, & 0.330794), \\ (0.548870, & 0.335661), \\ (0.548872, & 0.335666), \\ \ldots & \end{array}$$

die also offensichtlich viel schneller konvergiert.

1. Bemerkung: Das Newton-Verfahren würde, so wie es angegeben ist, im höherdimensionalen Fall in jedem Schritt eine Matrixinversion erfordern (bzw. die allgemeine Berechnung der inversen Jacobi-Matrix). Da man jedoch nur das Produkt dieser Matrix mit einem bestimmten Vektor braucht, ist es zumindest für $s > 2$ einfacher, das entsprechende lineare Gleichungssystem zu lösen. Man bestimmt also den Vektor z so, dass

$$f'(x_n)\ z = f(x_n)$$

und setzt dann

$$x_{n+1} := x_n - z.$$

2. Bemerkung: Die Berechnung der Matrix $f'(x)$ ist unter Umständen sehr aufwändig. Wenn das der Fall ist, nimmt man oft statt der partiellen Ableitungen von f einfache Näherungswerte, wie sie in Kapitel 6 besprochen wurden. Im Eindimensionalen kann man auch die Sekantenmethode wählen (siehe Kapitel 8.3.2).

8.3 Spezielle eindimensionale Iterationsverfahren

8.3.1 Intervallhalbierung

Hier handelt es sich um ein sehr einfaches, aber recht sicheres Verfahren, das häufig angewendet wird. Die Konvergenz ist allerdings nur linear mit dem Faktor $K = \frac{1}{2}$.

Satz 8.13 *Sei f eine stetige Funktion auf dem Intervall $[a, b]$ mit*

$$f(a) < 0, \quad f(b) > 0.$$

Dann konvergiert die folgenderweise definierte Folge gegen eine Nullstelle ξ von f:

$$x_1 := \frac{a+b}{2}, \qquad x_{k+1} := \begin{cases} x_k + \frac{b-a}{2^{k+1}} & \textit{falls } f(x_k) < 0, \\ x_k - \frac{b-a}{2^{k+1}} & \textit{falls } f(x_k) \geq 0. \end{cases}$$

Es gilt folgende Fehlerabschätzung:

$$|x_k - \xi| \leq \frac{b-a}{2^k}.$$

Bemerkung: Falls $f(a) > 0$ und $f(b) < 0$ ist, können wir dieses Verfahren auf $-f$ an Stelle von f anwenden. Es kommt also nur darauf an, dass $f(a)$ und $f(b)$ verschiedenes Vorzeichen haben.

Beweis des Satzes: Wir betrachten eine Folge von ineinander geschachtelten Intervallen $[a_k, b_k]$ und bezeichnen mit m_k die Mittelpunkte dieser Intervalle, also $m_k := (a_k + b_k)/2$.

$[a_1, b_1] := [a, b]$,

$$[a_{k+1}, b_{k+1}] := \begin{cases} [m_k, b_k] & \text{falls } f(m_k) < 0, \\ [a_k, m_k] & \text{falls } f(m_k) \geq 0. \end{cases}$$

Für jedes Intervall $[a_k, b_k]$ gilt

$$f(a_k) < 0 \quad \text{und} \quad f(b_k) \geq 0,$$

wie man unmittelbar durch Induktion sieht. Da f stetig ist, enthält also jedes dieser Intervalle mindestens eine Nullstelle von f.

Die Folge (a_k) ist monoton wachsend und nach oben durch b beschränkt, und daher konvergiert sie. Die Folge (b_k) ist monoton fallend und nach unten durch a beschränkt, daher konvergiert sie ebenfalls. Wegen

$$|b_k - a_k| = (b-a)/2^{k-1} \to 0 \quad \text{für } k \to \infty$$

stimmen die beiden Grenzwerte überein. Sei

$$\xi := \lim a_k = \lim b_k = \lim m_k.$$

Nun ist (wieder auf Grund der Stetigkeit von f) einerseits

$$f(\xi) = \lim f(a_k) \leq 0,$$

andererseits

$$f(\xi) = \lim f(b_k) \geq 0,$$

und daher $f(\xi) = 0$.

Für jedes k ist $a_k \leq \xi \leq b_k$ und daher

$$|m_k - \xi| \leq (b_k - a_k)/2 = (b-a)/2^k.$$

Es ist nun noch zu zeigen, dass die Folge (m_k) mit der im Satz definierten Iterationsfolge (x_k) übereinstimmt. Für $f(m_k) < 0$ ist $a_{k+1} = m_k$, $b_{k+1} = b_k$ und daher

$$m_{k+1} = \frac{a_{k+1} + b_{k+1}}{2} = \frac{m_k + b_k}{2} = m_k + \frac{b_k - a_k}{4} = m_k + \frac{b-a}{2^{k+1}}.$$

Für $f(m_k) \geq 0$ gilt analog

$$m_{k+1} = m_k - \frac{b-a}{2^{k+1}}.$$

Die beiden Folgen (m_k) und (x_k) erfüllen also dieselbe Rekursionsbedingung. Da $m_1 = (a+b)/2 = x_1$, stimmen sie überein.

Bemerkung: Wenn einmal $f(x_k) = 0$ ist, wird man natürlich die Iteration abbrechen. Im Allgemeinen iteriert man so lange, bis $(b-a)/2^k$ kleiner als die gegebene Toleranzgrenze für den absoluten Fehler ist. Beginnt man z.B. mit $a = 0$ und $b = 1$, so braucht man 21 Iterationen, um 6-stellige Genauigkeit zu erreichen (d.h. einen absoluten Fehler $< 0.5 \times 10^{-6}$). Natürlich kann man auch den relativen Fehler als Abbruchkriterium verwenden, falls ξ nicht zu nahe bei Null liegt.

8.3.2 Die Sekantenmethode

Ersetzt man im eindimensionalen Newton-Verfahren die Ableitung $f'(x_n)$ durch den Differenzenquotienten, so erhält man die Iterationsvorschrift

$$x_{n+1} := x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

Geometrisch bedeutet das, dass für die Berechnung von x_{n+1} die Funktion näherungsweise durch die Verbindungsgerade der Punkte $(x_{n-1}, f(x_{n-1}))$ und $(x_n, f(x_n))$ ersetzt wird. Man braucht hier natürlich zwei (verschiedene) Startwerte x_0 und x_1. Die Iteration ist nicht von der Form $x_{n+1} = \varphi(x_n)$, sondern $x_{n+1} = \varphi(x_n, x_{n-1})$.

Diese und ähnliche Interationsverfahren werden auch als *regula falsi* bezeichnet. Sie werden hauptsächlich verwendet, wenn die Berechnung der Ableitung zu aufwändig wäre. Es ist allerdings zu beachten, dass es bei der Bildung der Differenzen insbesondere in der Nähe des Grenzwerts sehr leicht zu größeren Fehlern kommen kann, da dann ungefähr gleich große Zahlen voneinander subtrahiert werden (Auslöschung!).

Die Konvergenz dieses Verfahrens ist zwar etwas langsamer als beim Newton-Verfahren, aber schneller als bei der Intervallhalbierung. Genauer gilt Folgendes:

Satz 8.14 *Sei ξ eine Nullstelle der zweimal stetig differenzierbaren Funktion f mit $f'(\xi) \neq 0$. Dann gibt es eine Umgebung U von ξ, sodass für beliebige Startwerte $x_0 \neq x_1$ aus U das Sekantenverfahren mit der Ordnung $(1 + \sqrt{5})/2 \approx 1.618$ gegen ξ konvergiert.*

Beweis: siehe z.B. [6] oder [8]. ■

8.3.3 Das Newton-Verfahren bei mehrfachen Nullstellen

Im Kapitel 8.2 über das Newton-Verfahren haben wir vorausgesetzt, dass $f'(\xi) \neq 0$ ist. Man kann zeigen, dass das Newton-Verfahren auch für $f'(\xi) = 0$ konvergiert, allerdings nur linear. Es gibt aber eine Modifikation dieses Verfahrens, die auch in diesem Fall quadratisch konvergiert. Man betrachtet dazu an Stelle von f die Funktion

$$F(x) := \frac{f(x)}{f'(x)},$$

die natürlich zunächst einmal nur für $x \neq \xi$ definiert ist.

Sei ξ eine p-fache Nullstelle von f, also $f(x) = (x-\xi)^p g(x)$ mit $g(\xi) \neq 0$ (siehe Definition 8.9). Dann ist

$$F(x) = \frac{(x-\xi)^p g(x)}{p(x-\xi)^{p-1} g(x) + (x-\xi)^p g'(x)} = \frac{(x-\xi) g(x)}{p\ g(x) + (x-\xi) g'(x)},$$

das heißt

$$F(x) = (x-\xi) G(x) \quad \text{(zunächst nur für } x \neq \xi\text{)}$$

mit

$$G(x) := \frac{g(x)}{p\ g(x) + (x-\xi) g'(x)},$$

also

$$G(\xi) = \frac{1}{p} \neq 0.$$

Mit der Zusatzdefinition $F(\xi) := 0$ gilt $F(x) = (x-\xi)G(x)$ auch für $x = \xi$. Wir sehen, dass ξ dann eine einfache Nullstelle von F ist, und F ist in einer Umgebung von ξ stetig differenzierbar. Wenn wir annehmen, dass G (und damit auch F) dreimal stetig differenzierbar ist, dann liefert die Iteration

$$x_{n+1} := x_n - \frac{F(x_n)}{F'(x_n)}$$

eine quadratisch gegen ξ konvergente Folge. Auf Grund der Definition von F gilt

$$\frac{F(x)}{F'(x)} = \frac{\frac{f(x)}{f'(x)}}{\frac{f'(x)^2 - f(x)f''(x)}{(f'(x))^2}} = \frac{f(x)f'(x)}{f'(x)^2 - f(x)f''(x)},$$

und daher können wir die Iteration auch so schreiben:

$$x_{n+1} := x_n - \frac{f(x_n)f'(x_n)}{f'(x_n)^2 - f(x_n)f''(x_n)}.$$

Bei der praktischen Anwendung dieser Formel ist allerdings darauf zu achten, dass es im Nenner durch die Differenzbildung leicht zu Auslöschung kommen kann.

Bemerkung: Dieses Iterationsverfahren funktioniert auch bei einfachen Nullstellen, d.h. für $p = 1$. Man muss also nicht unbedingt wissen, ob es sich um eine mehrfache Nullstelle handelt.

Beispiel: Sei

$$f(x) := 1 + \cos x.$$

Diese Funktion hat bei $\xi = \pi$ eine doppelte Nullstelle. Es ist $f(\pi) = f'(\pi) = 0$ und $f''(\pi) = 1$.

Das gewöhnliche Newtonverfahren ergibt mit dem Startwert $x_0 = 3$ die folgenden Iterationswerte (gerundet):

n	x_n	$\pi - x_n$
0	3.0000	0.1416
1	3.0709	0.0707
2	3.1063	0.0353
3	3.1239	0.0177
4	3.1328	0.0088
5	3.1372	0.0044
6	3.1394	0.0022
7	3.1405	0.0011

Man sieht hier deutlich, dass der Abstand von der Nullstelle in jedem Schritt ungefähr halbiert wird. Das bedeutet lineare Konvergenz.

Iteriert man dagegen $F(x) = \frac{1+\cos x}{-\sin x}$, so erhält man bereits nach zwei Schritten einen sehr genauen Wert:

n	x_n	$\pi - x_n$
0	3.000 000 000 000	0.141 592 653 590
1	3.141 120 008 060	0.000 472 645 530
2	3.141 592 653 572	0.000 000 000 018

8.4 Nullstellen von Polynomen

8.4.1 Allgemeines

Die Nullstellen eines Polynoms mit Grad ≤ 4 kann man theoretisch exakt berechnen (siehe Vorlesungen über Algebra). Diese Formeln werden in der Praxis allerdings nur für Grad ≤ 2 verwendet, da sie sonst numerisch sehr schlecht konditioniert sind. Für Polynome mit Grad > 4 ist es im Allgemeinen nicht möglich, die Lösung exakt zu berechnen.

Die Bestimmung der Wurzeln algebraischer Gleichungen höheren Grades ist grundsätzlich sehr problematisch, da sich die Wurzeln bei geringfügigen Änderungen der Koeffizienten unter Umständen sehr stark ändern. Das heißt, es handelt sich um eine schlecht konditionierte *Aufgabe*. Aus diesem Grund werden Verfahren zur Nullstellenbestimmung von Polynomen meist nur für

relativ kleine Grade angewendet, etwa bis zum Grad 8. Insbesondere für die Bestimmung der Eigenwerte von Matrizen gibt es wesentlich bessere Verfahren (siehe z.B. [6], [3]).

Ein bekanntes Beispiel für die schlechte Kondition des hier behandelten Problems stammt von Wilkinson[5] (1959). Er betrachtete das Polynom vom Grade 20 mit den Nullstellen $1, 2, 3, \ldots, 20$:

$$p(x) = \prod_{k=1}^{20}(x-k) = x^{20} - 210x^{19} + - \ldots + 20!.$$

Ersetzt man den Koeffizienten von x^{19}durch

$$-210 - 2^{-23} = -210.\,000\,000\,119\ldots,$$

so erhält man ein Polynom, das nur mehr 10 reelle Nullstellen hat. Z.B. liegen zwei konjugiert komplexe Nullstellen in der Nähe von $16.73 \pm 2.81\ i$, also ziemlich weit weg von der reellen Zahlenachse.

Grundsätzlich kann man jede algebraische Gleichung auf die Bestimmung der Eigenwerte einer Matrix zurückführen, wie der folgende Satz zeigt.

Satz 8.15 *Sei $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_{n-1}x^{n-1} + x^n$. Dann gibt es eine Matrix, deren charakteristisches Polynom (bis auf das Vorzeichen) mit $p(x)$ übereinstimmt und deren Eigenwerte daher gleich den Nullstellen von p sind, nämlich*

$$A = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & \cdots & 0 & -a_1 \\ 0 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & \ddots & \ddots & 0 & -a_{n-2} \\ 0 & \cdots & \cdots & 0 & 1 & -a_{n-1} \end{pmatrix}.$$

Beweis: Wir berechnen die Determinante von $A - \lambda E$, indem wir nach der letzten Spalte entwickeln:

$(-1)^{n+1}\det(A - \lambda E) =$

[5] James Hardy Wilkinson (1919 - 1986, England) hat bedeutende Beiträge zur numerischen Mathematik (insbesondere Eigenwertberechnung) geliefert (siehe [9]).

$$= -a_0 + a_1 \begin{vmatrix} -\lambda & 0 & & & & \\ & 1 & -\lambda & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & -\lambda \\ & & & & & 1 \end{vmatrix} - a_2 \begin{vmatrix} -\lambda & 0 & & & & \\ 1 & -\lambda & 0 & & & \\ & 0 & 1 & -\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -\lambda \\ & & & & 0 & 1 \end{vmatrix} +$$

$$\ldots$$

$$\ldots + (-1)^{n-1}(-a_{n-1} - \lambda) \begin{vmatrix} -\lambda & & & & & \\ 1 & -\lambda & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & -\lambda & \\ & & & & 1 & -\lambda \end{vmatrix} =$$

$$= -a_0 - a_1\lambda - a_2\lambda^2 - \ldots - a_{n-2}\lambda^{n-2} - (a_{n-1} + \lambda)\lambda^{n-1} = -p(\lambda). \blacksquare$$

Aus diesem Satz ergeben sich insbesondere Abschätzungen für die Lage der Nullstellen, die für die geeignete Wahl eines Startwertes nützlich sein können. Wir benötigen dazu folgende einfache Tatsache:

Lemma 8.16 *Für jeden Eigenwert λ einer (quadratischen) Matrix A gilt $|\lambda| \leq \|A\|$ (bezüglich jeder durch eine Vektornorm induzierte Matrixnorm).*

Beweis: Sei u ein normierter Eigenvektor von A zum Eigenwert λ, also $Au = \lambda u$ und $\|u\| = 1$. Dann gilt

$$|\lambda| = |\lambda| \, \|u\| = \|\lambda u\| = \|Au\| \leq \|A\| \, \|u\| = \|A\| \, .$$

$\blacksquare$

Satz 8.17 *Seien $\xi_1, \ldots, \xi_n$ die (komplexen) Nullstellen des Polynoms $p(x) = \sum_{k=0}^{n} a_k x^k$ mit $a_k \in \mathbb{C}$, $a_n = 1$. Dann gilt für alle $j \in \{1, \ldots, n\}$:*

$$\begin{aligned} &a) \quad \left|\xi_j\right| \;\leq\; \max\left\{1, \; \sum_{k=0}^{n-1} |a_k|\right\}, \\ &b) \quad \left|\xi_j\right| \;\leq\; \max\left\{|a_0| \, , \; 1 + |a_1| \, , \ldots , \; 1 + |a_{n-1}|\right\}. \end{aligned}$$

Bemerkung: Die Voraussetzung $a_n = 1$ bedeutet keine Einschränkung, denn das können wir (für $a_n \neq 0$) immer erreichen, indem wir $p(x)$ durch a_n dividieren.

Beweis des Satzes: Die zu $p(x)$ gemäß Satz 8.15 gehörige Matrix A hat die Eigenwerte ξ_j. Nach obigem Hilfssatz ist $|\xi_j| \leq \|A\|$ für alle j. Speziell für die 1- und ∞-Norm ergeben sich daraus schon die angegebenen Abschätzungen. ■

Es gibt viele numerische Verfahren zur Lösung von algebraischen Gleichungen. Im Folgenden werden einige der wichtigsten besprochen. Weitere interessante Verfahren sind z.B. in [5] beschrieben.

8.4.2 Das Horner-Schema

Für die numerische Berechnung der Nullstellen eines Polynoms muss jedenfalls das Polynom an mehreren Stellen ausgewertet werden, eventuell auch die erste Ableitung des Polynoms. Zu diesem Zweck verwendet man meist das im Folgenden beschriebene sogenannte *Horner-Schema*[6], das sich besonders gut für die Anwendung des Newton-Verfahrens eignet. Es beruht auf einer einfachen Umformung:

$$\sum_{k=0}^{n} a_k x_0^k = (\ldots((a_n x_0 + a_{n-1})x_0 + a_{n-2})x_0 + \ldots + a_1)x_0 + a_0.$$

Zur direkten Auswertung der linken Seite sind $2n-1$ Multiplikationen erforderlich, für die rechte Seite dagegen nur n. Außerdem erweist sich die rechte Seite als weniger anfällig gegenüber Rundungsfehlern (Auslöschung).

Die rechte Seite kann nach der Rekursionsformel

$$\begin{aligned} &b_n := a_n, \\ &b_k := b_{k+1}x_0 + a_k \quad \text{für } k = n-1, n-2, \ldots, 1, 0 \end{aligned} \tag{8.2}$$

ausgewertet werden. b_0 ist dann gleich dem gesuchten Wert $p(x_0)$.

Angenommen, eine Nullstelle ξ_1 von p wurde bereits berechnet. Um weitere Nullstellen zu finden, dividiert man $p(x)$ durch $x - \xi_1$, denn das ergibt bekanntlich ein Polynom mit denselben Nullstellen wie p, nur ohne ξ_1:

$$q(x) := \frac{p(x)}{x - \xi_1}.$$

Wenn ξ_1 eine s-fache Nullstelle von p mit $s > 1$ ist, dann ist ξ_1 eine $(s-1)$-fache Nullstelle von q.

[6] nach William George Horner (1786 - 1837, England).

Man berechnet nun eine Nullstelle ξ_2 von q und dividiert dann durch $x - \xi_2$, usw.. Bei diesem Prozess des *Abspaltens* (auch *Deflation* genannt) ist allerdings darauf zu achten, dass sich unter Umständen Rundungsfehler akkumulieren, insbesondere bei nahe beieinander liegenden Nullstellen. Um diesem Problem zu begegnen, kann man z.B. die berechneten Näherungswerte für die Nullstellen von q als Startwerte für ein neuerliches Newtonverfahren nehmen, wobei man aber das ursprüngliche Polynom p verwendet (*"Nachiteration"*).

Wenn man eine komplexe Nullstelle $\xi \in \mathbb{C} \setminus \mathbb{R}$ eines reellen Polynoms gefunden hat, muss auch die konjugiert komplexe Zahl $\bar{\xi}$ eine Nullstelle sein, sodass man gleich durch $(x - \xi)(x - \bar{\xi})$ dividieren kann.

Das Horner-Schema eignet sich nun interessanterweise nicht nur zur Berechnung von Funktionswerten $p(x)$, sondern auch zur Durchführung der Division durch $x - \xi$ und zur Berechnung von $p'(x)$:

Satz 8.18 *Sei* $p(x) = \sum_{k=0}^{n} a_k x^k$ *und* $x_0 \in \mathbb{C}$ *beliebig. Wenn* $b_n, b_{n-1}, \ldots, b_1, b_0$ *gemäß (8.2) bestimmt sind, dann gilt*

$$p(x) = (x - x_0)q(x) + b_0$$

mit

$$q(x) := b_n x^{n-1} + b_{n-1} x^{n-2} + \ldots + b_2 x + b_1.$$

Insbesondere folgt

$$p(x_0) = b_0$$

und

$$p'(x_0) = q(x_0).$$

1. Bemerkung: Wenn $x_0 = \xi_1$ eine Nullstelle von p ist, so ist $b_0 = 0$ und daher $p(x) = (x - \xi_1)q(x)$, das heißt $q(x) = \frac{p(x)}{x - \xi_1}$.

2. Bemerkung: Zur Berechnung von $p'(x_0) = q(x_0)$ kann man wieder das Horner-Schema verwenden. Dabei ist darauf zu achten, dass die Nummerierung der b_k um 1 verschoben ist: b_k ist nicht der Koeffizient von x^k, sondern von x^{k-1}.

Beweis des Satzes: Auf Grund der Rekursionsformel ist

$$a_k = b_k - b_{k+1} x_0 \quad \text{für } k = n - 1, \ldots, 0.$$

Daher folgt

$q(x)(x - x_0) + b_0 =$

$= b_n x^n + (b_{n-1} - b_n x_0)x^{n-1} + \ldots + (b_1 - b_2 x_0)x + (b_0 - b_1 x_0) =$
$= a_n x^n + a_{n-1}x^{n-1} + \ldots a_1 x + a_0 = p(x).$

$p'(x) = q'(x)(x - x_0) + q(x)$, also $p'(x_0) = q(x_0)$.■

3. Bemerkung: Bei manueller Rechnung ist es zweckmäßig, die Zahlen folgendermaßen anzuordnen:

x_0	a_n	a_{n-1}	a_{n-2}	$\ldots$	a_1	a_0
		$b_n x_0$	$b_{n-1}x_0$	$\ldots$	$b_2 x_0$	$b_1 x_0$
	b_n	b_{n-1}	b_{n-2}	$\ldots$	b_1	$b_0 = p(x_0)$

Will man auch $p'(x_0)$ berechnen, so wendet man das Horner-Schema nochmals an, mit den b_k $(k \neq 0)$ an Stelle der a_k, dadurch entstehen zwei zusätzliche Zeilen (vgl. untenstehendes Beispiel).

4. Bemerkung: Bei der Anwendung des Newton-Verfahrens auf Polynome mit reellen Koeffizienten ist darauf zu achten, dass dieses bei einem reellen Startwert eine reelle Folge liefert und daher keine (echt) komplexen Nullstellen finden kann. Zur Auffindung von komplexen Nullstellen kann man entweder mit einem nicht-reellen Startwert beginnen oder ein anderes Verfahren verwenden, z.B. die Methode von Muller (siehe Abschnitt 8.4.3).

Beispiel: Sei $p(x) = 2x^4 - 3x^2 + 3x - 4$. Division durch 2 ergibt $x^4 - 1.5x^2 + 1.5x - 2$. Die Schranken für die Nullstellen gemäß Satz 8.17 lauten:

a) $\left|\xi_j\right| \leq \max\{1,\ 1.5 + 1.5 + 2\} = 5,$

b) $\left|\xi_j\right| \leq \max\{2,\ 2.5,\ 2.5\} = 2.5.$

Es scheint also sinnvoll zu sein, etwa $x_0 = 2$ als Startwert zu nehmen. Das Horner-Schema sieht dann so aus:

2	2	0	−3	3	−4
		4	8	10	26
	2	4	5	13	**22**
		4	16	42	
	2	8	21	**55**	

Wir erhalten also $p(2) = 22$ und $p'(2) = 55$. Der erste Schritt des Newtonverfahrens lautet somit

$$x_1 = 2 - \frac{p(2)}{p'(2)} = 2 - \frac{22}{55} = 2 - \frac{2}{5} = 1.6.$$

Die nächsten Schritte verlaufen analog und liefern (gerundet)

$x_2 = 1.36203, \quad x_3 = 1.26871, \quad x_4 = 1.25515, \quad x_5 = 1.25488, \quad x_6 = 1.25488.$

x_6 stimmt also schon auf 6 Stellen mit x_5 überein. Wir können daher x_6 als ausreichende Näherung für die erste Nullstelle ξ_1 ansehen und dividieren nun $p(x)$ durch $x - x_6$:

1.25488	2	0	−3	3	−4
		2.50976	3.14945	0.18754	3.99998
	2	2.50976	0.14945	3.18754	−0.00002

Die letzte Zahl (−0.00002) ist (abgesehen von den Rundungsfehlern) gleich $p(x_6)$, sie ermöglicht also eine gewisse Genauigkeitskontrolle. Eine Nachiteration scheint hier nicht nötig zu sein.

Auf das Polynom mit den Koeffizienten der letzten Zeile (ohne die letzte Zahl) kann man nun das ganze Verfahren von neuem anwenden. In dem konkreten Beispiel kann man wieder mit $x_0 = 2$ beginnen, günstiger ist allerdings z.B. $x_0 = -2$. Man findet dann $\xi_2 \approx -1.73895$. Nach Abspaltung dieser Nullstelle bleibt nur mehr eine quadratische Gleichung, die man unmittelbar lösen kann. (Sie hat zwei konjugiert komplexe Wurzeln $\approx 0.242 \pm 0.926i$.)

8.4.3 Die Methode von Muller

Diese von David E. Muller 1956 publizierte Methode kann grundsätzlich auf die Nullstellenbestimmung beliebiger reeller oder komplexer Funktionen (in einer Variablen) angewendet werden, eignet sich aber besonders für die Lösung algebraischer Gleichungen. Der Grundgedanke ist ähnlich wie bei der Sekantenmethode, nur dass statt einer Geraden durch zwei Punkte eine Parabel durch drei Punkte genommen wird. Dadurch erreicht man, dass auch bei reellen Startwerten nicht-reelle Wurzeln gefunden werden können.

Wir wollen also eine Gleichung der Form $f(x) = 0$ lösen und benötigen dazu drei Startwerte x_0, x_1, x_2. Sei $y_i := f(x_i)$ für $i = 0, 1, 2$. Es ist zweckmäßig, hier das Interpolationspolynom in der Form

$$p(x) = a(x - x_2)^2 + b(x - x_2) + c$$

anzusetzen. Die unbekannten Koeffizienten a, b, c ergeben sich dann als Lösung des folgenden linearen Gleichungssystems:

$$\begin{aligned} a(x_0 - x_2)^2 + b(x_0 - x_2) + c &= y_0 \\ a(x_1 - x_2)^2 + b(x_1 - x_2) + c &= y_1 \\ c &= y_2 \end{aligned}$$

Es handelt sich also eigentlich nur um ein Gleichungssystem in den zwei Unbekannten a und b:

$$\begin{aligned} a(x_0 - x_2)^2 + b(x_0 - x_2) &= y_0 - y_2 \\ a(x_1 - x_2)^2 + b(x_1 - x_2) &= y_1 - y_2 \end{aligned}$$

Die Lösung kann man z.B. mit der Cramer'schen Regel leicht berechnen: Die Deteminante des Systems ist

$$\begin{aligned} d &:= (x_0 - x_2)^2(x_1 - x_2) - (x_1 - x_2)^2(x_0 - x_2) \\ &= (x_0 - x_2)(x_1 - x_2)((x_0 - x_2) - (x_1 - x_2)) \\ &= (x_0 - x_2)(x_1 - x_2)(x_0 - x_1), \end{aligned}$$

und damit ist

$$a = \tfrac{1}{d}\left((x_1 - x_2)(y_0 - y_2) - (x_0 - x_2)(y_1 - y_2)\right),$$

$$b = \tfrac{1}{d}\left((x_0 - x_2)^2\,(y_1 - y_2) - (x_1 - x_2)^2\,(y_0 - y_2)\right).$$

Um zu einer Näherung x_3 für eine Nullstelle zu kommen, bestimmt man nun die Nullstellen des quadratischen Polynoms $p(x)$. Man verwendet dazu normalerweise nicht die übliche Lösungsformel, sondern die folgende (vgl. Kapitel 3.3.1):

$$x_3 - x_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Das Vorzeichen der Wurzel wird dabei natürlich so gewählt, dass x_3 möglichst nahe bei x_2 liegt, das heißt, dass der Nenner möglichst großen Betrag hat. Dadurch erreich man auch, dass bei der Berechnung des Nenners möglichst keine Auslöschung auftritt. Wir erhalten somit:

$$x_3 := \begin{cases} x_2 - \frac{2c}{b+\sqrt{b^2-4ac}} & \text{falls } \left|b + \sqrt{b^2 - 4ac}\right| > \left|b - \sqrt{b^2 - 4ac}\right|, \\ x_2 - \frac{2c}{b-\sqrt{b^2-4ac}} & \text{sonst.} \end{cases}$$

Dieses Verfahren versagt natürlich, wenn $y_0 = y_1 = y_2$ ist, da dann das Interpolationspolynom $p(x)$ eine Gerade ist, welche die x-Achse nicht schneidet. Man kann allerdings zeigen, dass die Methode von Muller abgesehen von diesem Sonderfall (zumindest theoretisch) für jedes beliebige Tripel von Startwerten konvergiert, und zwar mit Konvergenzordnung ≈ 1.84 (vgl. [8]). Deflation und Nachiteration können wie beim Newtonverfahren erfolgen.

Bemerkung: Im Allgemeinen muss man hier mit komplexen Zahlen rechnen, auch wenn a, b, c reell sind, denn $b^2 - 4ac$ kann ja negativ sein.

Beispiel: Wir wollen eine Nullstelle von $x^4 - x^3 - 2x + 1$ berechnen und versuchen es mit den Startwerten $x_0 = 0$, $x_1 = 1$, $x_2 = 2$.

Wir erhalten $y_0 = 1$, $y_1 = -1$, $c = y_2 = 5$, $d = -2$, und weiter $a = 4$, $b = 10$.

Das ergibt $b^2 - 4ac = 20$,

$\left|b + \sqrt{b^2 - 4ac}\right| = 10 + \sqrt{20} > \left|b - \sqrt{b^2 - 4ac}\right| = 10 - \sqrt{20}$.

Somit: $x_3 = x_2 - \frac{2c}{b+\sqrt{b^2-4ac}} = 2 - \frac{10}{10+\sqrt{20}} = 1.309\ldots$.

Mit x_1, x_2, x_3 an Stelle von x_0, x_1, x_2 ergibt sich analog $x_4 = 1.513\ldots$.

So fortfahrend erhalten wir $x_5 = 1.554\ldots$, $x_6 = 1.559\ldots$. (Die exakte Lösung ist $1.558\,979\ldots$.)

8.4.4 Die Methode von Bairstow

Diese Methode eignet sich besonders zum Auffinden von Paaren konjugiert komplexer Nullstellen eines Polynoms p mit reellen Koeffizienten. Es geht dabei um das Abspalten eines quadratischen Faktors

$$(x - \xi)(x - \bar{\xi}) = x^2 - ux - v$$

mit

$$u = \xi + \bar{\xi} = 2\,\mathrm{Re}\,\xi \quad \text{und} \quad v = -\xi\bar{\xi} = -\left|\xi\right|^2.$$

Wir suchen also reelle Zahlen u und v, sodass das gegebene Polynom $p(x)$ durch $x^2 - ux - v$ teilbar ist. Wenn wir solche Zahlen gefunden haben, dann ist es leicht, daraus ξ (und $\bar{\xi}$) zu berechnen: Das sind ja einfach die Lösungen der quadratischen Gleichung $x^2 - ux - v = 0$.

Wir betrachten zunächst ganz allgemein die Division (mit Rest) von $p(x)$ durch so ein quadratisches Polynom: Es gibt ein Polynom q mit $\mathrm{grad}\,q = \mathrm{grad}\,p - 2$ und reelle Zahlen b und c, sodass

$$p(x) = q(x)(x^2 - ux - v) + bx + c. \tag{8.3}$$

Sowohl q als auch b und c sind dabei eindeutig bestimmt und hängen natürlich von u und v ab. Wir können daher (für festes p) b und c als Funktionen von u und v auffassen und schreiben

$$\begin{aligned} b &= b(u, v), \\ c &= c(u, v). \end{aligned}$$

$b(u, v)$ und $c(u, v)$ können zu gegebenen u und v ähnlich wie beim Horner-Schema (8.2) rekursiv berechnet werden (Details siehe weiter unten).

u und v sollen nun so bestimmt werden, dass $b = c = 0$ ist. Es geht also um die Lösung des (nichtlinearen) Gleichungssystems

$$\begin{aligned} b(u,v) &= 0 \\ c(u,v) &= 0 \end{aligned}$$

Dazu kann man z.B. das zweidimensionale Newtonverfahren verwenden. Auf diese Weise findet man übrigens auch Paare von reellen Lösungen.

Details der Methode von Bairstow

Zur Berechnung von $b = b(u,v)$ und $c = c(u,v)$ schreiben wir zunächst die Gleichung (8.3) ausführlich an (mit $p(x) = \sum_{k=0}^{n} a_k x^k$ und $q(x) = \sum_{k=0}^{n-2} b_k x^k$):

$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 =$

$= (b_{n-2}x^{n-2} + b_{n-3}x^{n-3} + \ldots + b_1 x + b_0)(x^2 - ux - v) + bx + c.$

Jetzt vergleichen wir der Reihe nach die Koeffizienten von $x^n, x^{n-1}, \ldots$:

$$\begin{array}{ll} x^n: & a_n = b_{n-2}, \\ x^{n-1}: & a_{n-1} = b_{n-3} - b_{n-2}u, \\ x^k: & a_k = b_{k-2} - b_{k-1}u - b_k v \quad \text{für } 2 \le k \le n-2, \ . \\ x^1: & a_1 = -b_0 u - b_1 v + b, \\ x^0: & a_0 = -b_0 v + c. \end{array}$$

Daraus erhalten wir

$$\begin{aligned} b_{n-2} &= a_n, \\ b_{n-3} &= a_{n-1} + b_{n-2}u, \\ b_{k-2} &= a_k + b_{k-1}u + b_k v \quad \text{für } k = n-2, n-3, \ldots, 2, \\ b &= a_1 + b_0 u + b_1 v, \\ c &= a_0 + b_0 v. \end{aligned}$$

Für die Anwendung des Newtonverfahrens brauchen wir die partiellen Ableitungen von b und c nach u und v. Sei

$$U_k := \frac{\partial b_k}{\partial u}.$$

Dann gilt auf Grund obiger Rekursion:

$$\begin{aligned} U_{n-2} &= 0, \\ U_{n-3} &= b_{n-2}, \\ U_{k-2} &= U_{k-1}u + b_{k-1} + U_k v \quad \text{für } k = n-2, n-3, \ldots, 2, \end{aligned}$$

und somit

$$\begin{aligned}\frac{\partial b}{\partial u} &= U_0 u + b_0 + U_1 v,\\ \frac{\partial c}{\partial u} &= U_0 v.\end{aligned}$$

Für

$$V_k := \frac{\partial b_k}{\partial v}$$

ergibt sich

$$\begin{aligned}&V_{n-2} = 0,\\ &V_{n-3} = 0,\\ &V_{k-2} = V_{k-1}u + V_k v + b_k \quad \text{für } k = n-2, n-3, \ldots, 2,\\ &\tfrac{\partial b}{\partial v} = V_0 u + V_1 v + b_1,\\ &\tfrac{\partial c}{\partial v} = V_0 v + b_0.\end{aligned}$$

Bei genauerer Betrachtung dieser Rekursionsformeln sehen wir, dass

$$V_{k-2} = U_{k-1}$$

ist (für $k = n-2, n-3, \ldots, 2$, wenn man zusätzlich $U_{n-1} = 0$ setzt). Der Induktionsschluss sieht so aus: Angenommen, $V_{k-2} = U_{k-1}$ für alle $k > k_0$. Dann gilt

$$U_{k_0-1} = U_{k_0}u + b_{k_0} + U_{k_0+1}v = V_{k_0-1}u + b_{k_0} + V_{k_0}v = V_{k_0-2}.$$

Die Rekursion für V_{k-2} braucht daher gar nicht eigens durchgeführt zu werden, und es gilt:

$$\begin{aligned}\frac{\partial b}{\partial v} &= U_1 u + U_2 v + b_1,\\ \frac{\partial c}{\partial v} &= U_1 v + b_0.\end{aligned}$$

Jetzt steht also der Anwendung des Newtonverfahrens nichts mehr im Wege.

Im Allgemeinen ist dieses Verfahren nicht sehr empfindlich bezüglich der Wahl des Startpunktes. Als Abbruchbedingung wird empfohlen:

$|b| + |c| \leq \varepsilon(|a_1| + |a_0|)$, wobei ε die vorgegebene relative Genauigkeit ist.

Zur Berechnung von weiteren Nullstellen müssten wir das gegebene Polynom durch $x^2 - ux - v$ dividieren. Das ist aber bereits geschehen: Die b_j sind ja die Koeffizienten des Quotienten, wenn $b = c = 0$ ist.

Beispiel: Wir betrachten wieder das Polynom $x^4 - x^3 - 2x + 1$ und versuchen einen quadratischen Faktor abzuspalten. Hier ist $n = 4$, und die Rekursionsformeln liefern

$b_2 = 1$, $b_1 = -1 + u$, $b_0 = (u-1)u + v$,

$U_2 = 0$, $U_1 = b_2 = 1$, $U_0 = u + b_1 = 2u - 1$,

also

$\frac{\partial b}{\partial u} = U_0 u + b_0 + U_1 v = (2u-1)u + (u-1)u + 2v$,

$\frac{\partial c}{\partial u} = U_0 v = (2u-1)\, v$,

$\frac{\partial b}{\partial v} = U_1 u + U_2 v + b_1 = 2u - 1$,

$\frac{\partial c}{\partial v} = U_1 v + b_0 = 2v + (u-1)u$.

Nehmen wir z.B. als Startwerte $u_0 = v_0 = -1$, so erhalten wir $b_1 = -2$, $b_0 = 1$, somit

$b = a_1 + b_0 u_0 + b_1 v_1 = -1$,

$c = a_0 + b_0 v_0 = 0$,

und folgende Jacobimatrix:

$$J = \begin{pmatrix} \frac{\partial b}{\partial u} & \frac{\partial b}{\partial v} \\ \frac{\partial c}{\partial u} & \frac{\partial c}{\partial v} \end{pmatrix} = \begin{pmatrix} 3 & -3 \\ 3 & 0 \end{pmatrix}, \text{ also } J^{-1} = \begin{pmatrix} 3 & -3 \\ 3 & 0 \end{pmatrix}^{-1} = \tfrac{1}{3} \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

Der erste Schritt des Newtonverfahrens sieht daher so aus:

$$\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} - J^{-1} \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix} - \tfrac{1}{3} \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \tfrac{1}{3} \begin{pmatrix} -3 \\ -4 \end{pmatrix}.$$

Nach vier Schritten erhalten wir mit etwa 6-stelliger Genauigkeit

$u = -1.031193$, $v = -1.358381$

und daraus

$\xi_{1,2} = \frac{u}{2} \pm \sqrt{\frac{u^2}{4} + v} \approx -0.515\,596 \pm 1.045\,247\, i.$

Wenn wir als Startwerte z.B. $u_0 = v_0 = 1$ verwenden, erhalten wir die folgenden beiden reellen Wurzeln, von denen wir eine schon im vorigen Abschnitt berechnet haben:

$x_1 \approx 0.472213$, $x_2 \approx 1.558980$.

8.5 Iterative Lösung von linearen Gleichungssystemen

8.5.1 Allgemeines

Iterationsverfahren werden auch zur Lösung von linearen Gleichungssystemen verwendet, insbesondere wenn sehr große Matrizen auftreten, da dann die direkten Lösungsmethoden, wie zum Beispiel das in Kapitel 4.1 besprochene Eliminationsverfahren, oft wesentlich aufwändiger sind und (wegen der auftretenden Rundungsfehler) auch weniger genaue Resultate liefern.

Auch bei linearen Gleichungssystemen gibt es natürlich verschiedene Möglichkeiten, ein gegebenes Gleichungssystem $Ax = b$ (mit quadratischer Matrix A) auf eine Fixpunktgleichung umzuformen, die wir hier in der Form

$$x = Tx + c$$

(mit einer anderen quadratischen Matrix T) annehmen. Wir betrachten dann also die Abbildung

$$\varphi(x) := Tx + c.$$

Wie wir beim Kontraktionsprinzip (Satz 8.5) gesehen haben, besitzt φ einen eindeutigen Fixpunkt, falls es ein $L < 1$ gibt, sodass $\|\varphi'(x)\| \leq L$ für alle x. Hier ist aber einfach $\varphi'(x) = T$ für alle x, sodass wir erhalten:

Satz 8.19 *Sei $c \in \mathbb{R}^s$ und T eine $s \times s$-Matrix mit $\|T\| < 1$. Dann besitzt das lineare Gleichungssystem $x = Tx + c$ eine eindeutig bestimmte Lösung ξ, und die durch*

$$x^{(n+1)} = Tx^{(n)} + c$$

definierte Iterationsfolge konvergiert für jeden Startwert $x^{(0)}$ linear gegen ξ.

(Wir verwenden hier obere Indizes zur Unterscheidung von den Indizes, welche die Koordinaten bezeichnen. Es ist also $x^{(n)} = (x_1^{(n)}, \ldots, x_s^{(n)})$.)

Als Norm wird hier meist wieder die Zeilen- oder Spaltensummennorm oder die Spektralnorm verwendet. Die Fehlerabschätzung kann genau so wie bei der allgemeinen Picard-Iteration erfolgen (siehe Satz 8.2). Insbesondere konvergiert das Verfahren umso schneller, je kleiner $\|T\|$ ist.

Die im Folgenden beschriebenen Iterationsmethoden unterscheiden sich nur in der Wahl der Matrix T und des zugehörigen Vektors c.

8.5.2 Das Jacobi-Verfahren (Gesamtschrittverfahren)

Bei dieser Methode wird einfach die j-te Zeile des Gleichungssystems nach x_j aufgelöst, für j von 1 bis s. Die j-te Zeile lautet zunächst

$$a_{j1}x_1 + \ldots + a_{jj}x_j + \ldots + a_{js}x_s = b_j.$$

Auflösung nach x_j ergibt

$$x_j = \frac{1}{a_{jj}} \left(\sum_{\substack{k=1 \\ k \neq j}}^{s} (-a_{jk}) x_k + b_j \right).$$

Die entsprechende Matrix $T = (t_{jk})$ (siehe voriger Abschnitt) hat daher folgende Elemente:

$$t_{jk} = \begin{cases} -a_{jk}/a_{jj} & \text{für } j \neq k, \\ 0 & \text{für } j = k. \end{cases}$$

Der Vektor c hat hier die Koordinaten

$$c_j = b_j/a_{jj}.$$

Natürlich muss hier $a_{jj} \neq 0$ sein. Das lässt sich aber durch Zeilen- oder Spaltenvertauschung immer erreichen, wenn A regulär ist.

Für manche Untersuchungen ist es zweckmäßig, das in folgender Form zu schreiben:

$$\begin{aligned} T &= -D^{-1}(L + R), \\ c &= D^{-1}b, \end{aligned}$$

wobei die Matrizen D, L und R folgenderweise definiert sind:

$$D = \begin{pmatrix} a_{11} & & & O \\ & a_{22} & & \\ & & \ddots & \\ O & & & a_{ss} \end{pmatrix},$$

$$L = \begin{pmatrix} 0 & & & O \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{s1} & \cdots & a_{s,s-1} & 0 \end{pmatrix},$$

$$R = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1s} \\ & 0 & \ddots & \vdots \\ & & \ddots & a_{s-1,s} \\ O & & & 0 \end{pmatrix}.$$

D ist also eine Diagonalmatrix, L eine linke (untere) und R eine rechte (obere) Dreiecksmatrix. Die Fixpunktgleichung hat nun die Form

$$x = -D^{-1}(L+R)x + D^{-1}b.$$

Die Iterationsvorschrift gemäß Satz 8.19 lautet also für das Jacobiverfahren:

$$x_j^{(n+1)} = \frac{1}{a_{jj}} \left(\sum_{\substack{k=1 \\ k \neq j}}^{s} (-a_{jk}) x_k^{(n)} + b_j \right) \quad \text{für } j = 1, \ldots, s. \tag{8.4}$$

Wenn man die Zeilensummennorm verwendet, so ergibt sich aus Satz 8.19 das folgende hinreichende Kriterium für die Konvergenz dieses Iterationsverfahrens:

Satz 8.20 *Die durch $x^{(n+1)} = Tx^{(n)} + c$ definierte Jacobi-Iteration (mit T und c wie oben) konvergiert gegen eine Lösung des linearen Gleichungssystems $Ax = b$, wenn*

$$\sum_{\substack{k=1 \\ k \neq j}}^{s} \left| \frac{a_{jk}}{a_{jj}} \right| < 1 \quad \textit{für alle } j \in \{1, \ldots, s\}.$$

.

Bemerkung: Diese Bedingung kann man auch so schreiben:

$$\sum_{\substack{k=1 \\ k \neq j}}^{s} |a_{jk}| < |a_{jj}| \quad \text{für alle } j \in \{1, \ldots, s\}.$$

Matrizen, die diese Bedingung erfüllen, heißen deshalb auch *diagonaldominant.*

8.5.3 Das Gauß-Seidel-Verfahren (Einzelschrittverfahren)

Die Iteration (8.4) kann unter Umständen folgenderweise verbessert werden. Wenn man die Koordinaten von $x^{(n+1)}$ der Reihe nach berechnet, so kennt man für $j > 1$ bereits die ersten $j-1$ Koordinaten von $x^{(n+1)}$. Es ist daher naheliegend, diese bei der Berechnung von $x_j^{(n+1)}$ an Stelle der ersten $j-1$ Koordinaten von $x^{(n)}$ zu verwenden. Das führt zu folgender Iterationsvorschrift:

$$x_j^{(n+1)} = \frac{1}{a_{jj}} \left(\sum_{k=1}^{j-1} (-a_{jk}) x_k^{(n+1)} + \sum_{k=j+1}^{s} (-a_{jk}) x_k^{(n)} + b_j \right) \quad \text{für } j = 1, \ldots, s. \tag{8.5}$$

Im Gegensatz zum Jacobiverfahren können hier nicht alle Koordinaten von $x^{(n+1)}$ gleichzeitig berechnet werden (Parallelverarbeitung), sondern nur hintereinander. Das erklärt die Bezeichnungen "Gesamtschritt-" und "Einzelschritt-Verfahren". Um zu einer übersichtlichen Matrizenschreibweise dieses Iterationsverfahrens zu kommen, schreiben wir es zunächst in folgender Form:

$$\begin{array}{lcr} a_{11}x_1^{(n+1)} & = & -a_{12}x_2^{(n)} - \ldots\ldots\ldots - a_{1s}x_s^{(n)} + b_1 \\ a_{21}x_1^{(n+1)} + a_{22}x_2^{(n+1)} & = & -a_{23}x_3^{(n)} - \ldots - a_{2s}x_s^{(n)} + b_2 \\ \vdots & \vdots & \vdots \\ a_{s1}x_1^{(n+1)} + \ldots\ldots\ldots + a_{ss}x_s^{(n+1)} & = & b_s \end{array}$$

Mit den Bezeichnungen des vorigen Abschnitts heißt das

$$(D + L)x^{(n+1)} = -Rx^{(n)} + b,$$

also

$$x^{(n+1)} = -(D + L)^{-1}Rx^{(n)} + (D + L)^{-1}b.$$

Hier ist somit

$$\begin{array}{rcl} T & = & -(D + L)^{-1}R, \\ c & = & (D + L)^{-1}b. \end{array}$$

Diese Darstellung ist allerdings für das praktische Rechnen ungeeignet. Man kann sie aber z.B. benützen, um zu zeigen, dass das Analogon von Satz 8.20 auch für die Gauß-Seidel-Iteration gilt:

Satz 8.21 *Wenn A diagonal dominant ist, konvergiert die Gauß-Seidel-Iteration gegen eine Lösung des Gleichungssystems $Ax = b$.*

Beweis: siehe z.B. [6]. ■

Beispiel:

$$\text{Sei } A = \begin{pmatrix} 10 & -1 & 2 & 0 \\ -1 & 11 & -1 & 3 \\ 2 & -1 & 10 & -1 \\ 0 & 3 & -1 & 8 \end{pmatrix}, \; b = \begin{pmatrix} 6 \\ 25 \\ -11 \\ 15 \end{pmatrix}.$$

A ist offensichtlich diagonaldominant. Die Iterationsvorschrift des Gauß-Seidel-Verfahrens lautet hier:

$$\begin{array}{lcrcrcrcrcr}
x_1^{(n+1)} & = & & 0.1 & x_2^{(n)} & - & 0.2 & x_3^{(n)} & & & + & 0.6 \\
x_2^{(n+1)} & = & & \frac{1}{11} & x_1^{(n+1)} & + & \frac{1}{11} & x_3^{(n)} & - & \frac{3}{11} & x_4^{(n)} & + & \frac{25}{11} \\
x_3^{(n+1)} & = & - & 0.2 & x_1^{(n+1)} & + & 0.1 & x_2^{(n+1)} & + & 0.1 & x_4^{(n)} & - & 1.1 \\
x_4^{(n+1)} & = & & & & - & 0.375 & x_2^{(n+1)} & + & 0.125 & x_3^{(n+1)} & + & 1.875
\end{array}$$

Nimmt man als Startwert den Nullvektor, so erhält man nach 5 Iterationen bereits die sehr gute Näherung $(1.00009, 2.00002, -1.00003, 0.999988)$, die sich von der exakten Lösung $(1, 2, -1, 1)$ um weniger als 10^{-4} unterscheidet. Mit dem Jacobi-Verfahren braucht man in diesem Fall 12 Iterationsschritte für dieselbe Genauigkeit. Es gibt jedoch Gleichungssysteme, wo das Jacobi-Verfahren schneller als das Gauß-Seidel-Verfahren konvergiert.

Die Anzahl der Multiplikationen und Divisionen beträgt beim Einzelschrittverfahren s^2 pro Iterationsschritt. Die entsprechende Anzahl beim Gauß'schen Eliminationsverfahren ist $\sim \frac{1}{3}s^3$ (siehe 4.1.2). Wenn also die Anzahl der Iterationsschritte kleiner als $\frac{s}{3}$ ist, dann ist das Iterationsverfahren der direkten Methode überlegen. Das kommt insbesondere bei großen Werten von s oft vor.

8.5.4 Relaxationsverfahren

Wenn das Jacobi- oder Gauß-Seidel-Verfahren zu langsam oder gar nicht konvergiert, erreicht man oft auf folgende Weise ein besseres Konvergenzverhalten: Man wählt für $x^{(n+1)}$ einen geeigneten Punkt auf der Verbindungsgeraden von $x^{(n)}$ und $Tx^{(n)} + c$:

$$x^{(n+1)} = (1 - \omega)x^{(n)} + \omega(Tx^{(n)} + c).$$

ω heißt der *Relaxationsfaktor*. Für $\omega = 1$ ist das nichts Neues. $0 < \omega < 1$ bedeutet eine Verkleinerung der Schritte (*"Unterrelaxation"*). Dadurch kann

unter Umständen aus einer nicht konvergenten Iterationsfolge eine konvergente werden. $\omega > 1$ bewirkt eine Vergrößerung der Schritte (*"Überrelaxation"*) und daher unter Umständen eine Erhöhung der Konvergenzgeschwindigkeit.

Der optimale Relaxationsfaktor ω_0 hängt sehr von der Matrix T ab (insbesondere von ihrem Spektralradius). Für spezielle Klassen von Matrizen gibt es Formeln zur Berechnung von ω_0. Z.B. gilt für das Gauß-Seidel-Verfahren

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \rho(T)}},$$

falls A positiv definit und tridiagonal ist. (*"Tridiagonal"* heißt, dass nur die Hauptdiagonale und die beiden benachbarten Nebendiagonalen Elemente ungleich Null enthalten.) Die Durchführung der Berechnung von ω_0 ist aber unter Umständen aufwändiger als die Lösung des betrachteten Gleichungssystems (vgl. z.B. [4] oder [2]).

Literaturverzeichnis

[1] Abramowitz, M.; Stegun, I. A.(ed.): Handbook of mathematical functions with formulas, graphs, and mathematical tables. 10th printing, National Bureau of Standards. Wiley-Interscience, New York etc. 1972.

[2] Burden, Richard L.; Faires, J.Douglas: Numerical analysis, 5th ed. London: ITP International Thomson Publishing (1993).

[3] Deuflhard, P.; Hohmann, A.: Numerische Mathematik I. Eine algorithmisch orientierte Einführung. 3. Aufl., Walter de Gruyter, Berlin, New York 2002.

[4] Finck von Finckenstein, Karl Graf: Einführung in die numerische Mathematik. München: Carl Hanser Verlag. (1977).

[5] Gander, Walter: Computermathematik. Basel - Boston - Stuttgart: Birkhäuser Verlag. 257 S. (1985).

[6] Hämmerlin, G.; Hoffmann, K.-H.: Numerische Mathematik. 4. Aufl., Springer-Verlag, Berlin 1994.

[7] Isaacson, E.; Keller, H.B.: Analyse numerischer Verfahren. Verlag Harri Deutsch, Zürich, Frankfurt 1973.

[8] Stoer, Josef: Numerische Mathematik I, 8. Aufl., Springer-Verlag, Berlin 1999.

[9] The MacTutor History of Mathematics archive, University of St. Andrews, Scotland, http://www-history.mcs.st-and.ac.uk/~history/

[10] Weisstein, Eric W.: MathWorld, A Wolfram Web Resource, http://mathworld.wolfram.com

Stichwortverzeichnis