Einführung in die Numerische Mathematik
(Numerik 0)

Rolf Rannacher

Institut für Angewandte Mathematik
Universität Heidelberg

Vorlesungsskriptum SS 2005

27. April 2006

Adresse des Autors:

Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 293/294
D-69120 Heidelberg, Deutschland

rannacher@iwr.uni-heidelberg.de
http://numerik.uni-hd.de/

Inhaltsverzeichnis

Literaturverzeichnis

[1] J. Stoer, und R. Bulirsch: *Einführung in die Numerische Mathematik*, Teil I (1. Auflage 1971, J. Stoer), Teil II (1. Auflage 1993, J. Stoer und R. Bulirsch); Springer (Neuauflagen).

[2] G. Hämmerlin und K.-H. Hoffmann: *Numerische Mathematik*; Springer 1989.

[3] P. Deuflhard und A. Hohmann: *Numerische Mathematik 1*; Teil I (1. Auflage 1991), W. de Gruyter (3. Auflage, 2002).

[4] A. Quarteroni, R. Sacco und F. Saleri: *Numerische Mathematik 1/2*; Springer (Übersetzung 2002).

[5] R. Schaback, und H. Werner: *Praktische Mathematik* I/II; Springer Hochschultext, 1. Auflage 1970 (Neuauflagen).

[6] F. Stummel, und K. Hainer: *Praktische Mathematik*; B.G. Teubner, 1. Auflage 1978 (Neuauflage).

[7] L. Collatz und W. Wetterling: *Optimierungsaufgaben*; Springer 1966.

0 Einleitung

Aufgabenstellung der "numerischen" Mathematik ist die Entwicklung von Methoden, mit denen die Lösungen mathematischer Problemstellungen effektiv berechnet bzw. möglichst mit Fehlerangabe angenähert werden können. Bis in die 50-er Jahre unseres Jahrhunderts zeichneten sich erfolgreiche praktische Mathematiker durch ein besonderes Geschick aus, mit großen Formel- und Datenmengen umzugehen. Seit dem Aufkommen der immer leistungsfähigeren elektronischen Rechenanlagen haben sich die Gewichte verschoben. Die "praktische" Mathematik wurde zur "numerischen" Mathematik, d.h. der Theorie der auf Digitalrechnern realisierbaren numerischen Algorithmen. Eins der Hauptanwendungsgebiete numerischer Methoden ist in der Simulation komplexer Naturvorgänge auf Rechenanlagen. Man möchte teure Experimente wie z.B. Windkanalversuche bei der Flugzeugkonstruktion oder Festigkeitstests bei Betonkonstruktionen durch beliebig oft und schnell wiederholbare Modellrechnungen ersetzen. Die dabei verwendeten numerischen Verfahren sind dabei aus einer Reihe von einfachen Bausteinen zusammengesetzt (z.B. Integralberechnungen, Lösung linearer Gleichungssysteme, Berechnung von Nullstellen etc.). Diese einführende Vorlesung befaßt sich vor allem mit diesen elementaren Bausteinen, deren Ursprünge meist noch in der Vor-Computer-Zeit liegen.

Zur numerischen Lösung eines Problems der Praxis gehört unbedingt auch eine Information über den dabei gemachten Fehler, um das Resultat richtig einschätzen zu können. Der Gesamtfehler setzt sich zusammen aus den "Modellfehlern":

- Idealisierungsfehler: Zur Beschreibung eines physikalischen Sachverhalts wird ein mathematisches Modell gebildet. Bei der mathematischen Formulierung müssen Vereinfachungen (z.B. Linearisierungen) vorgenommen werden.
- Datenfehler: Die Daten eines mathematischen Modells (z.B. Koeffizienten einer Differentialgleichung) sind aufgrund ungenauer Kenntnis von Materialeigenschaften notwendig mit Fehlern behaftet.

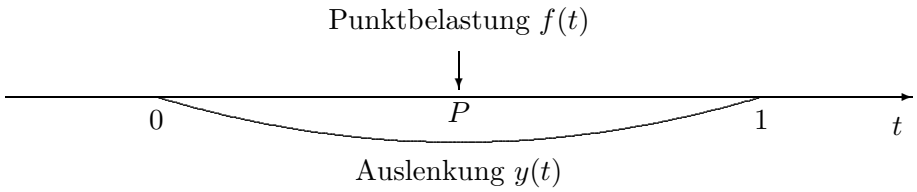
und den "numerischen" Fehlern:

- Diskretisierungsfehler: Kontinuierliche Prozesse werden durch endliche ersetzt (z.B. Approximation des Riemannschen Integrals durch Riemannsche Summen).
- Abbruchfehler: Unendliche Algorithmen werden nach endlich vielen Schritten abgebrochen (z.B. die Iteration $x_{n+1} = \frac{1}{2}(x_n + a/x_n) \to \sqrt{a}\ \ (n \to \infty)$).
- Rundungsfehler: Auf der Rechenanlage müssen alle Rechnungen auf einem endlichen Zahlbereich durchgeführt werden (z.B. $1/3 \sim 0,3\ldots3$).

Das Zusammenwirken all dieser Fehlereinflüsse soll anhand des folgenden, einfachen Beispiels demonstriert werden:

Ein Stahlseil der Länge $L = 1$ sei an den Spitzen zweier Masten befestigt, so daß es unter Einwirkung der Schwerkraft (fast) straff gespannt erscheint. Gefragt ist nun nach

der Auslenkung des Seils aus dieser Ruhelage, wenn ein Trapezkünstler in seiner Mitte steht.



Das physikalische Modell besteht in der (wohl begründeten) Annahme, daß sich die tatsächliche Auslenkung als Graph einer Funktion $y(t)$ beschreiben läßt, für welche die sog. "potentielle Gesamtenergie" (c eine Materialkonstante, $f(t)$ Belastungsdichte)

$$E(y) = \frac{c}{2} \int_0^1 \frac{y'(t)^2}{\sqrt{1+y'(t)^2}}\, dt - \int_0^1 f(t)y(t)\, dt$$

einen minimalen Wert annimmt. Dies sehen wir als das "exakte" mathematische Modell des angegebenen physikalischen Sachverhalts an. Zur Vereinfachung des Problems wird nun angenommen, daß die Belastung $f(t)$ so klein ist, daß nur kleine Auslenkungsgradienten auftreten, d.h.: $|y'(t)| \ll 1$. In diesem Fall kann das Funktional $E(y)$ vereinfacht werden zu

$$\tilde{E}(y) = \frac{c}{2} \int_0^1 y'(t)^2\, dt - \int_0^1 f(t)\, y(t)\, dt\,.$$

Dies ist nun das eigentliche "mathematische" Problem, mit dem der "praktische" Mathematiker konfrontiert ist. Der bis hierher entstandene Modellfehler ist im Augenblick nicht Gegenstand unseres Interesses. Als notwendige (und hinreichende) Bedingung für die Minimalitätseigenschaft der Funktion $y(t)$ erhält man durch Variation

$$\frac{d}{d\alpha}\, \tilde{E}(y+\alpha\varphi)\,|_{\alpha=0} = 0 \quad \forall\text{"zulässigen" } \varphi = \varphi(t)$$

die folgende (lineare) Differentialgleichung mit Randbedingungen:

$$-cy''(t) = f(t)\,, \quad t \in (0,1)\,, \qquad y(0) = y(1) = 0\,.$$

Zur Lösung des Problems wird nun eine Diskretisierung vorgenommen:

$$t_i \equiv ih\,,\ i = 0, \ldots, N+1\,, \quad f_i \equiv f(t_i)\,, \quad h = 1/(N+1)\,,$$

$$y''(t_i) \sim \frac{1}{h^2}\,\{y(t_{i+1}) - 2y(t_i) + y(t_{i-1})\}\,,$$

die auf ein lineares Gleichungssystem

$$\frac{c}{h^2}\begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & 2 & -1 \\ 0 & & & -1 & 2 \end{bmatrix}\begin{bmatrix} \eta_1 \\ \\ \vdots \\ \\ \eta_N \end{bmatrix} = \begin{bmatrix} f_1 \\ \\ \vdots \\ \\ f_N \end{bmatrix},$$

bzw. abgekürzt $A\eta = b$, für den Vektor $\eta = (\eta_1, \ldots, \eta_N)^T$ der Näherungswerte zu $y(t_i)$ führt. Dieses Gleichungssystem besitzt wegen $(\det(A) \neq 0)$ eine eindeutige Lösung. Diese kann aufgrund der Identität

$$D\eta = D\eta - A\eta + b\,, \quad D \equiv \frac{c}{h^2}\begin{bmatrix} 2 & & 0 \\ & \ddots & \\ 0 & & 2 \end{bmatrix},$$

ausgehend von einem Startvektor $\eta^{(0)} \in \mathbb{R}^N$ durch die Iteration

$$\eta^{(n+1)} = \eta^{(n)} - D^{-1}[A\eta^{(n)} - b]$$

angenähert werden. Man kann zeigen, daß $\eta^{(n)} \to \eta\ (n \to \infty)$ konvergiert. In der Praxis muß die Iteration aber nach endlich vielen Schritten mit einem Näherungswert $\eta^{(k)}$ abgebrochen werden. Tatsächlich wird aber auf dem Rechner statt $\eta^{(k)}$ eine Näherung $\tilde{\eta}^{(k)}$ geliefert, da alle arithmetischen Operationen auf einem endlichen Zahlbereich durchgeführt werden. Von diesem Ergebnis

$$\tilde{\eta}^{(k)} = (\tilde{\eta}_1^{(k)}, \ldots, \tilde{\eta}_n^{(k)})^T$$

sollen nun Rückschlüsse auf die zu erwartende Auslenkung des Stahlseils gewonnen werden. Der dabei aufgetretene Fehler setzt sich zusammen aus

- Diskretisierungsfehler: $\max_{i=1,\ldots,N} |\eta_i - y(t_i)|\,.$
- Abbruchfehler: $\max_{i=1,\ldots,N} |\eta_i^{(k)} - \eta_i|\,.$
- Rundungsfehler: $\max_{i=1,\ldots,N} |\tilde{\eta}_i^{(k)} - \eta_i^{(k)}|\,.$

1 Fehleranalyse

1.1 Zahldarstellung und Rundungsfehler

Bei der Verarbeitung numerischer Algorithmen auf dem "Computer" treten zwangsläufig Fehler auf, die durch die Endlichkeit des Bereiches der auf einer solchen Maschine darstellbaren Zahlen bedingt sind. Zur Approximation von reellen Zahlen und der elementaren arithmetischen Operationen zwischen ihnen werden sog. "Maschinenzahlen" und "Maschinenoperationen" verwendet, welche auf dem Computer realisierbar sind.

Eine "(normalisierte) Gleitkommazahl" zur Basis $b \in \mathbb{N}$, $b \geq 2$, ist eine Zahl $x \in \mathbb{R}$ dargestellt in der Form

$$x = \pm\, m \cdot b^{\pm e} \tag{1.1.1}$$

mit der "Mantisse" $m = m_1 b^{-1} + \ldots + m_r b^{-r} + \ldots \in \mathbb{R}$, und dem "Exponenten" $e = e_{s-1} b^{s-1} + \ldots + e_0 b^0 \in \mathbb{N} \cup \{0\}$, wobei $m_i, e_i \in \{0, \ldots, b-1\}$. Für $x \neq 0$ ist diese Darstellung durch die Normierungsvorschrift $m_1 \neq 0$ eindeutig bestimmt. Für $x = 0$ setzt man $m = 0$.

Bemerkung 1.1: Die Verwendung der Gleitkommadarstellung im numerischen Rechnen ist wesentlich, um Zahlen sehr unterschiedlicher Größe verarbeiten zu können; z.B. Ruhemasse Elektron $M_0 = 9.11 \cdot 10^{-28}\,\mathrm{g}$, Lichtgeschwindigkeit $c = 2.998 \cdot 10^{10}\,\mathrm{cm/sec}$.

Auf dem Rechner stehen für die Darstellung von reellen Zahlen nur endlich viele Stellen zur Verfügung:

r Ziffern + 1 Vorzeichen für die Mantisse

s Ziffern + 1 Vorzeichen für den Exponenten.

Die Speicherung einer solchen Zahl

$$x = \pm\,[m_1 b^{-1} + \ldots + m_r b^{-r}] \cdot b^{\pm[e_{s-1} b^{s-1} + \ldots + e_0 b^0]}$$

erfolgt dann in der Form $x : (\pm)\,[m_1 \ldots m_r](\pm)[e_{s-1} \ldots e_0]$. Aus technischen Gründen verwenden moderne Rechner eine Zahldarstellung mit den Basen $b = 2$ (Dualsystem) oder $b = 16$ (Sedezimalsystem) oder Mischungen davon. Die in der obigen Form auf einem Rechner dargestellten (rationalen) Zahlen werden "Maschinenzahlen" genannt; sie bilden das sog. "numerische Gleitkommagitter" $A = A(b, r, s)$. Da A endlich ist, gibt es eine größte/kleinste darstellbare Zahl:

$$x_{\max\,/\,\min} = \pm\,(b-1)\{b^{-1} + \ldots + b^{-r}\} \cdot b^{(b-1)\{b^{s-1} + \ldots + b^0\}} = \pm\,(1 - b^{-r}) \cdot b^{b^s - 1}$$

sowie eine kleinste positive/größte negative darstellbare Zahl:

$$x_{\mathrm{posmin/negmax}} = \pm\, b^{-1} \cdot b^{-(b-1)\{b^{s-1} + \ldots + b^0\}} = \pm\, b^{-b^s}\,.$$

Beispiel 1.1: Beim sog. "IEEE-Format" (üblich auf UNIX-Workstations) werden zur Darstellung von doppelt genauen Zahlen (REAL*8 in FORTRAN) 64 Bits (=8 Bytes) verwendet:

$$x = \pm m \cdot 2^{c-1022} \,.$$

Dabei stehen 1 Bit für das Vorzeichen, 52 Bits für die Mantisse $m = 2^{-1} + m_2 2^{-2} + \ldots + m_{53} 2^{-53}$ (die erste Mantissenstelle ist aus Normierungsgründen stets 1) und 11 Bits für die sog. "Charakteristik" $c = c_0 2^0 + \ldots + c_{10} 2^{10} \in [1, 2046]$ zur Verfügung, wobei $m_i \,,\, c_i \in \{0, 1\}$ Dualzahlen sind. Durch die vorzeichenfreie Darstellung des Exponenten in der Form $e = c - 1022$ wird der Zahlbereich um eine Zweierpotenz erweitert. Für REAL*8-Zahlen gilt somit:

$$\begin{aligned} x_{\max} &\sim 2^{1024} \sim 1.8 \cdot 10^{308} \quad , \qquad x_{\min} \sim -2^{1024} \sim -1.8 \cdot 10^{308} \quad , \\ x_{\text{posmin}} &= 2^{-1022} \sim 2.2 \cdot 10^{-308} \quad , \qquad x_{\text{negmax}} = -2^{-1022} \sim -2.2 \cdot 10^{-308} \quad . \end{aligned}$$

Die ausgenommenen Werte $c = 0$ und $c = 2047$ der Charakteristik werden zur Darstellung der Null ($m_2 = \ldots = m_{53} = 0$, $c_0 = \ldots = c_{10} = 0$) sowie einer Sondergröße "nan" (not a number) verwendet.

Die Ausgangsdaten $x \in \mathbb{R}$ einer numerischen Aufgabe und die Zwischenergebnisse einer Rechnung müssen durch Maschinenzahlen dargestellt werden. Für Zahlen innerhalb des "zulässigen" Bereiches

$$D := [x_{\min}, x_{\text{negmax}}] \cup \{0\} \cup [x_{\text{posmin}}, x_{\max}]$$

wird eine "Rundungsoperation" $\mathrm{rd} : D \to A$ verwendet, an die man die natürliche Forderung stellt

$$|x - \mathrm{rd}(x)| = \min_{y \in A} |x - y| \qquad \forall x \in D \,. \tag{1.1.2}$$

Dies ist beim IEEE-Format z.B. realisiert durch "natürliche" Rundung:

$$\mathrm{rd}(x) = \mathrm{sign}(x) \cdot \begin{cases} 0.m_1 \cdots m_{53} \cdot 2^e \,, & \text{für } m_{54} = 0 \\ (0.m_1 \cdots m_{53} + 2^{-53}) \cdot 2^e \,, & \text{für } m_{54} = 1 \,. \end{cases}$$

Andere manchmal vorkommende Rundungsarten, welche (1.1.2) nicht erfüllen, werden im folgenden nicht betrachtet. Für Zahlen außerhalb des zulässigen Bereiches D (z.B. als Resultat einer Division durch Null) wird von einigen Maschinen Exponentenüberlauf ("overflow" oder "underflow") registriert und die Verarbeitung abgebrochen, während im IEEE-Format in diesem Fall mit der unbestimmten Variable "nan" weitergearbeitet wird.

Der mit der Rundung verbundene sog. "absolute Rundungsfehler"

$$|x - \mathrm{rd}(x)| \le \frac{1}{2} b^{-r} b^e \tag{1.1.3}$$

hängt jeweils noch vom Exponenten e von x ab. Dagegen ist der sog. "relative Rundungsfehler"

$$\left|\frac{x-\mathrm{rd}(x)}{x}\right| \le \frac{1}{2}\,\frac{b^{-r}b^e}{|m|b^e} \le \frac{1}{2}\,b^{-r+1} \tag{1.1.4}$$

für $x \in D$, $x \neq 0$, beschränkt durch die sog. "Maschinengenauigkeit" $\mathrm{eps} := \frac{1}{2}\,b^{-r+1}$. Für $x \in D$ ist dann offenbar

$$\mathrm{rd}(x) = x(1+\varepsilon) \quad \text{mit} \quad |\varepsilon| \le \mathrm{eps}\,. \tag{1.1.5}$$

Bei Anwendung des IEEE-Formats ist der maximale relative Rundungsfehler

$$\mathrm{eps}_{\mathrm{REAL*8}} \le \frac{1}{2}2^{-52} \sim 10^{-16}\,.$$

Die arithmetischen Grundoperationen $* \in \{+,-,\cdot,/\}$ werden auf der Rechenanlage durch entsprechende "Maschinenoperationen" $\circledast \in \{\oplus,\ominus,\odot,\oslash\}$ ersetzt, welche Maschinenzahlen wieder in Maschinenzahlen überführen. Dies ist meist für $x,y \in A$ im Falle $x * y \in D$ gemäß

$$x \circledast y = \mathrm{rd}(x * y) = (x * y)(1+\varepsilon)\,, \quad |\varepsilon| \le \mathrm{eps}\,, \tag{1.1.6}$$

realisiert. Dazu werden die Operationen maschinenintern (meist unter Verwendung einer erhöhten Stellenzahl für die Mantisse) ausgeführt, in normalisierte Form gebracht und dann gerundet. Im Fall $x * y \notin D$ erscheint meist eine Fehlermeldung. Bei dem Gebrauch von "IF-Abfragen" in Programmen ist zu berücksichtigen, daß die Maschinenoperationen $\oplus$ und $\odot$ dem Assoziativgesetz und dem Distributivgesetz nur näherungsweise genügen; i. Allg. ist für $x,y,z \in A$:

$$(x \oplus y) \oplus z \;\neq\; x \oplus (y \oplus z)\,, \quad (x \oplus y) \odot z \;\neq\; (x \odot z) \oplus (y \odot z)\,.$$

Insbesondere gilt i. Allg. für Zahlen $x,y \in A$:

$$x \oplus y = x, \quad \text{für} \quad |y| \le \frac{|x|}{b}\mathrm{eps}. \tag{1.1.7}$$

Hieraus läßt sich für einen konkreten Rechner die Größe der Maschinengenauigkeit eps experimentell ermitteln.

1.2 Konditionierung numerischer Aufgaben

Eine numerische Aufgabe (z.B. Bestimmung einer Nullstelle, Lösung eines linearen Gleichungssystems etc.) wird als "gut konditioniert" bezeichnet, wenn eine kleine Störung der Eingangsdaten auch nur eine kleine Änderung der Ergebnisse zur Folge hat.

Beispiel 1.2: Als Beispiel betrachten wir das lineare Gleichungssystem

$$\begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix}$$

mit der (eindeutig bestimmten) Lösung $(x, y)^T = (2, -2)^T$. Störung der rechten Seite zu $(0.86419999, 0.14400001)^T$ erzeugt die "Näherungslösung" $(\tilde{x}, \tilde{y})^T = (0.9911, -0.4870)^T$ Diese numerische Aufgabe ist offenbar sehr schlecht konditioniert.

Zur Präzisierung des Begriffes "Konditionierung" müssen wir zunächst den der "numerischen Aufgabe" definieren. Wir wollen hier unter einer numerischen Aufgabe die Berechnung endlich vieler Größen $y_i\,(i = 1, \ldots, n)$ aus gewissen Größen $x_j\,(j = 1, \ldots, m)$ mittels einer funktionalen Vorschrift $y_i = f_i(x_1, \ldots, x_m)$ verstehen. Der Einfachheit halber betrachten wir hier nur den Fall, daß die y_i, x_j reelle (oder komplexe) Zahlen sind, und verwenden zur Abkürzung die vektorielle Schreibweise $y = f(x)$ mit

$$x = (x_1, \ldots, x_m)^T, \quad y = (y_1, \ldots, y_n)^T, \quad f = (f_1, \ldots, f_n)^T.$$

Als Beispiel kann die Berechnung eines Vektors $x \in \mathbb{R}^n$ als Lösung eines linearen Gleichungssystems $Ax = b$ dienen, wobei $f(x) = A^{-1}b$.

Definition 1.1: *Bei Verwendung fehlerhafter Eingangsdaten $x_j + \Delta x_j$ (z.B. aufgrund des Rundungsfehlers) ergeben sich fehlerhafte Resultate $y_i + \Delta y_i$. Wir bezeichnen $|\Delta y_i|$ als den "absoluten" Fehler und $|\Delta y_i / y_i|$ (für $y_i \neq 0$) als den "relativen" Fehler.*

Große absolute Fehler können offenbar, "relativ" gesehen, klein sein und umgekehrt; z.B. mag ein Fehler von ± 100 km beim Messen der Entfernung Erde-Mond als "klein" angesehen werden, während derselbe Fehler bezogen auf die Entfernung Heidelberg-Paris sicherlich als "groß" anzusehen ist.

Wir haben gesehen, daß der relative Rundungsfehler durch die Maschinengenauigkeit eps beschränkt ist. Hier wird uns auch hauptsächlich nur dieser interessieren. Im Folgenden betreiben wir eine sog. "differentielle" Fehleranalyse, die sich auf die Betrachtung des Einflusses relativ kleiner Datenfehler $|\Delta x_j| \ll |x_j|$ beschränkt. Sind die Funktionen $f_i = f_i(x_1, \ldots, x_m)$ stetig partiell differenzierbar nach den Argumenten x_j, so gilt nach dem Taylorschen Satz

$$\Delta y_i \;=\; f_i(x + \Delta x) - f_i(x) \;=\; \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x)\, \Delta x_j + R_i^f(x; \Delta x), \quad i = 1, \ldots, m, \qquad (1.2.8)$$

mit einem Restglied $R_i^f(x;\Delta x)$, welches schneller als $|\Delta x| = \max_{j=1,\dots,m}|\Delta x_j|$ gegen Null geht; wir schreiben dies abgekürzt als $R_i^f(x;\Delta x) = o(|\Delta x|)$. Der Einfachheit halber nehmen wir an, daß sogar $R_i^f(x;\Delta x) = O(|\Delta x|^2)$ gilt, was im Falle der zweimaligen Differenzierbarkeit der Funktion f gesichert ist.

Definition 1.2: *Wir verwenden hier und im Folgenden die sog. "Landauschen*[1] *Symbole" $O(\cdot)$ und $o(\cdot)$ zur quantitativen Beschreibung von Grenzprozessen. Für Funktionen $g(t)$ und $h(t)$ der Variablen $t \in \mathbb{R}_+$ bedeutet die Schreibweise*

$$g(t) = O(h(t)) \quad (t \to 0)\,,$$

daß für kleine $t \in (0, t_0]$ mit einer Konstanten $c \geq 0$ gilt

$$|\,g(t)\,| \leq c\,|\,h(t)\,|\,.$$

Entsprechend bedeutet $g(t) = o(h(t))$ für $t \to 0$, daß für kleine $t \in (0, t_0]$ mit einer Funktion $c(t) \to 0\ (t \to 0)$ gilt

$$|\,g(t)\,| \leq c(t)\,|\,h(t)\,|\,.$$

Analoge Schreibweisen verwendet man für Grenzübergänge $t \to \infty$.

Beispiel 1.3: Für eine zweimal stetige differenzierbare Funktion $g(t)$ folgt aus

$$g(t+\Delta t) = g(t) + \Delta t\, g'(t) + \frac{\Delta t^2}{2}\, g''(\tau)\,, \quad \tau \in (t, t+\Delta t)\,,$$

für den sog. "vorwärts genommenen Differenzenquotienten" die Beziehung

$$\frac{1}{\Delta t}\,\{g(t+\Delta t) - g(t)\} = g'(t) + O(\Delta t)\,.$$

Die obige Formel (1.2.8) besagt, daß der Fehler Δy_i "in erster Näherung", d.h. bis auf eine Größe der Ordnung $O(|\,\Delta x|^2)$, gleich dem ersten Summanden auf der rechten Seite ist; in Symbolen:

$$\Delta y_i \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x)\,\Delta x_j\,. \tag{1.2.9}$$

Für den komponentenweisen relativen Fehler gilt dann

[1] *Edmund Georg Hermann Landau (1877-1938): Deutscher Mathematiker; seit 1909 Professor in Göttingen (Nachfolger von Minkowski); 1934 wegen seiner jüdischen Abstammung zwangsweise pensioniert; fundamentale Beiträge zur analytischen Zahlentheorie, insbesondere zur Primzahlverteilung, und zur komplexen Funktionentheorie.*

$$\frac{\Delta y_i}{y_i} \doteq \sum_{j=1}^{m} \frac{\partial f_i}{\partial x_j}(x) \frac{\Delta x_j}{y_i} = \sum_{j=1}^{m} \underbrace{\frac{\partial f_i}{\partial x_j}(x) \frac{x_j}{f_i(x)}}_{=: k_{ij}(x)} \frac{\Delta x_j}{x_j}. \tag{1.2.10}$$

Dabei verhält sich der vernachlässigte Term wie

$$\Big| \frac{R_i^f(x;\Delta x)}{y_i} \Big| = O\Big(\frac{|\Delta x|^2}{|y_i|}\Big).$$

Unter der Voraussetzung, daß $|\Delta x| = o(|y_i|)$ kann er gegen den führenden $O(|\Delta x|)$-Term vernachlässigt werden. Wir werden im Folgenden stets annehmen, daß sich die auszuwertende Größe und die betrachteten Datenstörungen verhalten wie

$$|\Delta x| = o(|y_i|), \quad i = 1, \ldots, n.$$

Andernfalls darf der Restgliedterm nicht vernachlässigt werden.

Definition 1.3: *Die Größen $k_{ij}(x)$ heißen "(relative) Konditionszahlen" der Funktion f im Punkt x. Sie sind ein Maß dafür, wie sich kleine relative Fehler in den Ausgangsdaten im Ergebnis auswirken. Man nennt die Aufgabe, $y = f(x)$ aus x zu berechnen, "schlecht konditioniert", wenn ein $|k_{ij}(x)| \gg 1$ ist; andernfalls "gut konditioniert" oder auch "gutartig". Im Fall $|k_{ij}(x)| < 1$ spricht man von "Fehlerdämpfung" und im Fall $|k_{ij}(x)| > 1$ von "Fehlerverstärkung". Fehler durch Störungen in der Funktion $f(\cdot)$ werden hier nicht betrachtet; dies bleibt später speziellen Situationen vorbehalten (z.B.: lineare Gleichungssysteme).*

Die Konditionierung einer numerischen Aufgabe ist eng verknüpft mit der Frage nach berechenbaren Fehlerschranken für zugehörige Näherungslösungen. Beim Problem der direkten Funktionsauswertung $y = f(x)$ ist die relative Fehlerempfindlichkeit beschrieben durch die Relation (1.2.10). Interessanter ist das umgekehrte Problem der Gleichungslösung $x = f^{-1}(y)$, bei dem zu gegebenem y die Lösung der Gleichung $f(x) = y$ gesucht ist. Hierbei wird o.B.d.A. Gleichheit der Dimensionen $n = m$ angenommen. Als Beispiele können wieder die Lösung eines linearen Gleichungssystems oder die Bestimmung der Wurzeln einer quadratischen Gleichung dienen. In diesem Fall liegt es nahe, eine Fehlerabschätzung für irgendeine Näherungslösung $\tilde{x}$ auf dem Weg der "Probe" zu erzielen. Zu diesem Zweck bildet man den sog. "Defekt" $d(\tilde{x}) = y - f(\tilde{x})$. Die Frage ist nun, ob für einen kleinen Defekt auch der tatsächliche Fehler $\Delta x = \tilde{x} - x$ klein ist. Die differentielle Fehleranalyse liefert hierzu

$$\frac{\Delta x_i}{x_i} \doteq \sum_{j=1}^{n} k_{ij}^{(-1)} \frac{\Delta y_j}{y_j}, \quad k_{ij}^{(-1)} := \frac{\partial f_i^{-1}}{\partial y_j}(y) \frac{y_j}{x_i},$$

mit den Konditionszahlen $k_{ij}^{(-1)}$ der inversen Abbildung $f^{-1}(\cdot)$. Wir fassen die Konditionszahlen zu Matrizen $K = (k_{ij})_{i,j=1}^n$ und $K^{(-1)} = (k_{ij}^{(-1)})_{i,j=1}^n$ zusammen. Für deren Produkt gilt dann

$$(K^{(-1)}K)_{ij} = \sum_{k=1}^{n} k_{ik}^{(-1)} k_{kj} = \sum_{k=1}^{n} \frac{\partial f_i^{-1}}{\partial x_k} \frac{x_k}{y_i} \frac{\partial f_k}{\partial x_j} \frac{y_j}{x_k} = \frac{y_j}{y_i} \sum_{k=1}^{n} \frac{\partial f_i^{-1}}{\partial x_k} \frac{\partial f_k}{\partial x_j} = \delta_{ij} \,,$$

wobei das sog. "Kronecker[2]-Symbol" δ_{ij} für die Alternative $\delta_{ij} = 1$, für $i = j$, bzw. $\delta_{ij} = 0$, für $i \neq j$, steht. Die Matrix $K^{(-1)}$ ist also gerade die Inverse von K.

1.2.1 Arithmetische Grundoperationen

Im folgenden diskutieren wir die Konditionierung, d. h. die Anfälligkeit gegenüber kleiner Störungen der Eingangsdaten, einiger einfacher Grundaufgaben.

1) Die Addition $y = f(x_1, x_2) = x_1 + x_2$ zweier Zahlen $x_1, x_2 \in \mathbb{R}$, $x_1, x_2 \neq 0$, mit

$$k_1 = \frac{\partial f}{\partial x_1} \frac{x_1}{f} = 1 \cdot \frac{x_1}{x_1 + x_2} = \frac{1}{1 + x_2/x_1}$$
$$k_2 = \frac{\partial f}{\partial x_2} \frac{x_2}{f} = 1 \cdot \frac{x_2}{x_1 + x_2} = \frac{1}{1 + x_1/x_2}$$

ist "schlecht" konditioniert für $x_1/x_2 \sim -1$. Bei der Addition ähnlich großer Zahlen mit unterschiedlichem Vorzeichen kann bei der Fortpflanzung von kleinen Störungen in den Eingangsgrößen sog. "Auslöschung" wesentlicher Stellen auftreten.

Definition 1.4 (Auslöschung): *Unter "Auslöschung" versteht man den Verlust an wesentlichen Dezimalstellen bei der Subtraktion von Zahlen gleichen Vorzeichens. Dies ist gefährlich im Fall, daß eine oder beide der Zahlen keine Maschinenzahlen sind und vor Ausführung der Operation gerundet werden. Bei der Subtraktion von Maschinenzahlen ist Auslöschung natürlich unschädlich.*

Beispiel 1.4: *Dezimale Gleitpunktrechnung mit $r = 4$ und $s = 1$*

$$\begin{array}{rcl@{\quad}c@{\quad}rcl} x_1 & = & 0.11258762 \cdot 10^2 & \rightarrow & \mathrm{rd}(x_1) & = & 0.1126 \cdot 10^2 \\ x_2 & = & 0.11244891 \cdot 10^2 & \rightarrow & \mathrm{rd}(x_2) & = & 0.1124 \cdot 10^2 \\ x_1 + x_2 & = & 0.\underline{2250}3653 \cdot 10^2 & , & \mathrm{rd}(x_1) + \mathrm{rd}(x_2) & = & 0.\underline{2250} \cdot 10^2 \\ x_1 - x_2 & = & 0.13871 \cdot 10^{-1} & , & \mathrm{rd}(x_1) - \mathrm{rd}(x_2) & = & 0.2000 \cdot 10^{-1} \end{array}$$

Im zweiten Fall gilt $k_1 \sim k_2 \sim 810$, d.h. fast 1000-fache Fehlerverstärkung.

[2]Leopold Kronecker (1823-1891): deutscher Mathematiker; wirkte in Berlin als "Privatgelehrter"; betrieb die Arithmetisierung der Mathematik; wichtiger Vertreter des "Konstruktivismus", welcher die generelle Verwendung des Widerspruchsbeweises und des "aktual Unendlichen" in Form z.B. der allgemeinen reellen Zahlen ablehnt.

2) Die Multiplikation $y = f(x_1, x_2) = x_1 \cdot x_2$ mit

$$k_1 = \frac{\partial f}{\partial x_1}\,\frac{x_1}{f} = x_2\,\frac{x_1}{x_1 \cdot x_2} = 1\,, \qquad k_2 = \ldots = 1\,,$$

ist generell "gut" konditioniert. Dasselbe gilt auch für die Division (Übungsaufgabe).

1.2.2 Lösung quadratischer Gleichungen

Für Zahlen $p, q \in \mathbb{R}$ wird die folgende quadratische Gleichung betrachtet:

$$y^2 - py + q = 0 \quad \text{(o.b.d.A. } q \neq 0\text{)}.$$

Für die Wurzeln y_1 und y_2 gilt $y_{1,2} = y_{1,2}(p,q) = p/2 \pm \sqrt{p^2/4 - q}$ sowie $p = y_1 + y_2\,, q = y_1 \cdot y_2$ (Vietascher[3] Wurzelsatz). Damit erhält man:

$$\left.\begin{aligned} \frac{\partial y_1}{\partial p} + \frac{\partial y_2}{\partial p} &= 1 \\ \frac{\partial y_1}{\partial p}\,y_2 + y_1\,\frac{\partial y_2}{\partial p} &= 0 \end{aligned}\right\} \Rightarrow \qquad \frac{\partial y_2}{\partial p} = \frac{y_2}{y_2 - y_1}\,, \quad \frac{\partial y_1}{\partial p} = \frac{y_1}{y_2 - y_1}$$

$$\left.\begin{aligned} \frac{\partial y_1}{\partial q} + \frac{\partial y_2}{\partial q} &= 0 \\ \frac{\partial y_1}{\partial q}\,y_2 + y_1\,\frac{\partial y_2}{\partial q} &= 1 \end{aligned}\right\} \Rightarrow \qquad \frac{\partial y_1}{\partial q} = \frac{1}{y_1 - y_2} = -\frac{\partial y_2}{\partial q}$$

$$k_{11} = \frac{\partial y_1}{\partial p}\,\frac{p}{y_1} = \frac{y_1}{y_2 - y_1}\,\frac{p}{y_1} = \frac{y_1 + y_2}{y_1 - y_2} = \frac{1 + y_2/y_1}{1 - y_2/y_1}$$

$$k_{12} = \frac{\partial y_1}{\partial q}\,\frac{q}{y_1} = \frac{1}{y_1 - y_2}\,\frac{q}{y_1} = \frac{y_2}{y_1 - y_2} = \frac{1}{1 - y_2/y_1}$$

Für k_{21} und k_{22} gilt Analoges. Die Berechnung von y_1, y_2 ist schlecht konditioniert für $y_1/y_2 \sim 1$, d.h. wenn die Wurzeln relativ dicht beeinander liegen.

Beispiel 1.5: $p = 4\,, \quad q = 3.999\,, \quad y_{1,2} = 2 \pm 0.01\,,$

$$k_{12} = \frac{1}{1 - y_1/y_2} = 99.5 \quad \Rightarrow \quad \textit{fast 100-fache Fehlerverstärkung.}$$

[3] Francois Viete, lat. Franciscus Vieta (1540-1603): Französischer Mathematiker; Arbeiten über algebraische Gleichungen und sphärische Trigonometrie; gab trigonometrische Tafeln heraus und führte die systematische Buchstabenrechnung ein.

1.3 Stabilität numerischer Algorithmen

Gegeben sei wieder eine numerische Aufgabe der Art $y = f(x)$ mit einer Abbildung $f : \mathbb{R}^m \to \mathbb{R}^n$. Unter einem "Verfahren" (oder "Algorithmus") zur gegebenenfalls näherungsweisen Berechnung von y aus x verstehen wir eine endliche (oder auch abzählbar unendliche) Folge von "elementaren" Abbildungen $\varphi^{(k)}$, die durch sukzessive Anwendung einen Näherungwert $\tilde{y}$ zu y liefern:

$$x = x^{(0)} \to \varphi^{(1)}\,(x^{(0)}) = x^{(1)} \to \ldots \to \varphi^{(k+1)}\,(x^{(k)}) = x^{(k+1)} \to \ldots \tilde{y}\,.$$

Im einfachsten Fall sind die $\varphi^{(k)}$ arithmetische Grundoperationen.

Definition 1.5: *Bei der Durchführung des Algorithmus auf einer Rechenanlage treten in jedem Schritt Fehler auf (z.B. Rundungsfehler, Auswertungsfehler von transzendenten Funktionen, etc.), die sich bis zum Ende der Rechnung akkumulieren können. Der Algorithmus wird "stabil" (oder auch "gutartig") genannt, wenn die im Verlaufe der Ausführung akkumulierten Fehler den durch die Konditionierung der Aufgabe $y = f(x)$ bedingten unvermeidbaren Problemfehler nicht übersteigen.*

Eine der Hauptaufgaben der numerischen Mathematik ist es, für die in den Anwendungen auftretenden Aufgaben stabile Lösungsalgorithmen zu finden. Wir diskutieren im folgenden einige elementare Beispiele.

1.3.1 Lösung quadratischer Gleichungen

Wir betrachten die Auflösung einer quadratischen Gleichung der Form

$$y^2 - py + q = 0 \quad (0 \neq q < p^2/4)\,, \quad y_{1,2} = f(p,q) = p/2 \pm \sqrt{p^2/4 - q}\,.$$

Für $|y_1/y_2| \gg 1$, d.h. für $q \ll p^2/4$, ist die Aufgabe gut konditioniert. Der Algorithmus zur Berechnung der Wurzeln könnte wie folgt aussehen:

$$u = p^2/4\,, \quad v = u - q\,, \quad w = \sqrt{v} \quad (\geq 0)\,.$$

Im Fall $p < 0$ wird zur Vermeidung von Auslöschung zunächst $\tilde{y}_2 = p/2 - w$ berechnet mit der akzeptablen Fehlerfortpflanzung

$$\left|\frac{\Delta y_2}{y_2}\right| \dot{\leq} \underbrace{\left|\frac{1}{1 - 2w/p}\right|}_{\approx 1} \left|\frac{\Delta p}{p}\right| + \underbrace{\left|\frac{1}{1 - p(2w)}\right|}_{\approx 1} \left|\frac{\Delta w}{w}\right|.$$

Die zweite Wurzel könnte dann auf folgenden Wegen bestimmt werden:

Variante A	Variante B
$\tilde{y}_1 = p/2 + w$	$\tilde{y}_1 = q/y_2 \quad (\text{wegen } q = y_1 y_2)$

Für $q \ll p^2/4$ ist $w \approx -p/2$, d.h. bei Variante A tritt zwangsläufig Auslöschung ein. Die Rundungsfehler in p und w übertragen sich auf y_1 wie

$$\left|\frac{\Delta y_1}{y_1}\right| \dot{\leq} \underbrace{\left|\frac{1}{1+2w/p}\right|}_{\gg 1} \underbrace{\left|\frac{\Delta p}{p}\right|}_{\leq \text{eps}} + \underbrace{\left|\frac{1}{1+p/2w}\right|}_{\gg 1} \underbrace{\left|\frac{\Delta w}{w}\right|}_{\approx \text{eps}}.$$

Dieser Algorithmus ist offenbar im vorliegenden Fall $q \ll p^2/4$ sehr instabil. Bei Variante B gilt dagegen

$$\left|\frac{\Delta y_1}{y_1}\right| \dot{\leq} \underbrace{\left|\frac{\Delta q}{q}\right|}_{\leq \text{eps}} + \underbrace{\left|\frac{\Delta y_2}{y_2}\right|}_{\approx \text{eps}}.$$

d.h. dieser Algorithmus ist stabil.

Regel 1.3.1: *Bei der Lösung quadratischer Gleichungen sollten nicht beide Wurzeln aus der Lösungsformel berechnet werden.*

Beispiel 1.6: $p = -4\,, \quad q = 0.01$ (vierstellige Rechnung)

$$\left.\begin{array}{rcl} u & = & 4 \\ v & = & 3.99 \\ w & = & \underline{1.997}4984\ldots \\ \tilde{y}_2 & = & \underline{-3.997}4984\ldots \end{array}\right\} \quad \tilde{y}_1 = \left\{\begin{array}{rcl} \text{exakt} & : & -0.0025015\ldots \\ \text{A} & : & -0.0030 \quad (\text{rel. Fehler} \quad 0.2) \\ \text{B} & : & -0.0025 \end{array}\right.$$

1.3.2 Auswertung arithmetischer Ausdrücke

Im Folgenden bedienen wir uns einer sog. "Forwärtsrundungsfehleranalyse", bei welcher die Akkumulation des Rundungsfehlers ausgehend vom Startwert abgeschätzt wird. Wir beginnen mit der Auswertung eines einfachen arithmetischen Ausdrucks der Art

$$y = f(x_1, x_2) = x_1^2 - x_2^2 = (x_1 + x_2) \cdot (x_1 - x_2).$$

Der Problemfehler durch Rundung der Ausgangsdaten verhält sich wie

$$\left|\frac{\Delta y}{y}\right| \dot{\leq} \sum_{j=1}^{2} \left|\frac{\partial f}{\partial x_j}\,\frac{x_j}{f}\right| \left|\frac{\Delta x_j}{x_j}\right| \leq \left|2x_1\frac{x_1}{x_1^2 - x_2^2}\right| + \left|2x_2\frac{x_2}{x_1^2 - x_2^2}\right| = 2\left|\frac{(x_1/x_2)^2 + 1}{(x_1/x_2)^2 - 1}\right| \text{ eps}\,.$$

Für $|x_1/x_2| \approx 1$ liegt also schlechte Konditionierung vor. Zur algorithmischen Auswertung dieses Ausdrucks gibt es zwei Alternativen, wobei die Ausgangsdaten $x_1, x_2 \in A$ als *Maschinenzahlen* gegeben seien.

Algorithmus A:	Algorithmus B:
$u = x_1 \odot x_1$	$u = x_1 \oplus x_2$
$v = x_2 \odot x_2$	$v = x_1 \ominus x_2$
$\tilde{y} = u \ominus v$	$\tilde{y} = u \odot v$

Zur Rundungsfehleranalyse beachten wir, daß für die Maschinenoperationen $\circledast$ auf der Menge A der Maschinenzahlen gilt:

$$a \circledast b = \mathrm{rd}(a * b) = (a * b)(1 + \varepsilon) \quad \text{mit} \quad |\varepsilon| \leq \mathrm{eps}\,.$$

Unter Verwendung dieser Beziehung erhalten wir für den ersten Algorithmus:

$$\begin{aligned} \text{(A)} \qquad & u = x_1^2\,(1+\varepsilon_1)\,, \quad v = x_2^2\,(1+\varepsilon_2) \\ & \tilde{y} = \left[x_1^2\,(1+\varepsilon_1) - x_2^2\,(1+\varepsilon_2)\right](1+\varepsilon_3) \\ & \quad = \underbrace{x_1^2 - x_2^2}_{=\,y} + x_1^2\varepsilon_1 - x_2^2\varepsilon_2 + \underbrace{(x_1^2 - x_2^2)}_{=\,y}\,\varepsilon_3 + O(\mathrm{eps}^2) \end{aligned}$$

$$\left|\frac{\Delta y}{y}\right| \dot{\leq} \mathrm{eps}\,\frac{x_1^2 + x_2^2 + |x_1^2 - x_2^2|}{|x_1^2 - x_2^2|} \;=\; \mathrm{eps}\,\left\{1 + \left|\frac{(x_1/x_2)^2 + 1}{(x_1/x_2)^2 - 1}\right|\right\}\,.$$

Der Rundungsfehlereinfluß wird groß für $|x_1/x_2| \sim 1$, übersteigt aber nicht den Problemfehleranteil, d.h.: Der Algorithmus A ist nach unserer Definition stabil.

Für den zweiten Algorithmus gilt:

$$\begin{aligned} \text{(B)} \qquad & u = (x_1 + x_2)\,(1+\varepsilon_1)\,, \quad v = (x_1 - x_2)\,(1+\varepsilon_2) \\ & \tilde{y} = (x_1 + x_2)\,(1+\varepsilon_1)\,(x_1 - x_2)\,(1+\varepsilon_2)\,(1+\varepsilon_3) \\ & \quad = \underbrace{x_1^2 - x_2^2}_{=\,y} + \underbrace{(x_1^2 - x_2^2)}_{=\,y}\,(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) + O(\mathrm{eps}^2) \end{aligned}$$

$$\left|\frac{\Delta y}{y}\right| \dot{\leq} |\varepsilon_1 + \varepsilon_2 + \varepsilon_3| \leq 3\,\mathrm{eps}\,.$$

Algorithmus B ist offenbar i. Allg. stabiler als Algorithmus A. Es sei nochmals betont, daß hierbei von bereits gerundeten Ausfgangsdaten in A ausgegangen wird. An diesem Beispiel kann man bereits eine einfache Regel ablesen, die allgemein gültigen Charakter hat. Die Auswirkung dieser Regel wird andand des nächsten Beispiels noch klarer werden.

Regel 1.3.2: *Bei der Durchführung einer numerischen Rechnung sollte man die numerisch schlechter konditionierten Operationen möglichst frühzeitig ansetzen.*

1.3.3 Auswertung von Polynomen

$$y = p(x) = a_0 + a_1 x + \ldots + a_n x^n\,.$$

Als Modellfall betrachten wir zunächst das Polynom

$$p(x) = a_1 x + a_2 x^2 = x(a_1 + a_2 x)\,.$$

Zu seiner Auswertung in einem Punkt ξ bietet sich der Algorithmus

$$\text{A)} \quad u = \xi \odot \xi\,, \quad v = a_2 \odot u\,, \quad w = a_1 \odot \xi\,, \quad \tilde{y} = v \oplus w\,,$$

und, bei Berücksichtigung der obigen Faustregel, der Algorithmus

$$\text{B)} \quad u = a_2 \odot \xi\,, \quad v = a_1 \oplus u\,, \quad \tilde{y} = \xi \odot v\,,$$

an. Bei Algorithmus B spart man offensichtlich eine arithmetische Operation. Die zugehörige Rundungsfehleranalyse sieht wie folgt aus:

$$\begin{aligned}
\text{(A)} \qquad & u = \xi^2(1+\varepsilon_1), \quad v = a_2\xi^2(1+\varepsilon_1)(1+\varepsilon_2), \quad w = a_1\xi(1+\varepsilon_3) \\
& \tilde{y} = [a_2\xi^2(1+\varepsilon_1)(1+\varepsilon_2) + a_1\xi(1+\varepsilon_3)](1+\varepsilon_4) \\
& \quad = a_2\xi^2 + a_1\xi + (a_2\xi^2 + a_1\xi)\varepsilon_4 + a_2\xi^2(\varepsilon_1+\varepsilon_2) + a_1\xi\varepsilon_3 + O(\mathrm{eps}^2) \\
& \quad = y + y\varepsilon_4 + a_2\xi^2(\varepsilon_1+\varepsilon_2) + a_1\xi\varepsilon_3 + O(\mathrm{eps}^2)
\end{aligned}$$

$$\left|\frac{\Delta y}{y}\right| \dot{\leq} \varepsilon_4 + \frac{a_1\xi\varepsilon_3 + a_2\xi^2(\varepsilon_1+\varepsilon_2)}{a_1\xi + a_2\xi^2} = \varepsilon_4 + \varepsilon_3 + \frac{\xi}{a_1/a_2 + \xi}\,(\varepsilon_1 + \varepsilon_2 - \varepsilon_3)$$

$$\begin{aligned}
\text{(B)} \qquad & u = a_2\xi(1+\varepsilon_1), \quad v = [a_1 + a_2\xi(1+\varepsilon_1)](1+\varepsilon_2) \\
& \tilde{y} = \xi[a_1 + a_2\xi(1+\varepsilon_1)](1+\varepsilon_2)(1+\varepsilon_3) \\
& \quad = y + a_1\xi(\varepsilon_2+\varepsilon_3) + a_2\xi^2(\varepsilon_1+\varepsilon_2+\varepsilon_3) + O(\mathrm{eps}^2)
\end{aligned}$$

$$\left|\frac{\Delta y}{y}\right| \doteq \varepsilon_2 + \varepsilon_3 + \frac{\xi}{a_1/a_2 + \xi}\,\varepsilon_1\,.$$

Für $\xi \sim -a_1/a_2$ (d.h. wenn ξ nahe bei einer Nullstelle von $p(x)$ liegt) ist Algorithmus B offensichtlich etwas stabiler als Algorithmus A.

Dieses Resultat legt zur Auswertung des allgemeinen Polynoms n-ter Ordnung, ausgehend von der Darstellung

$$p(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + xa_n)\ldots))\,,$$

den folgenden Algorithmus nahe:

Definition 1.6: *Das sog. "Horner[4]-Schema"*

$$b_n = a_n, \quad k = n-1, \ldots, 0: \quad b_k = a_k + \xi b_{k+1}, \tag{1.3.11}$$

liefert den Funktionswert $p(\xi) = b_0$ *des Polynoms* $p(x)$*.*

[4] *William George Horner (1786-1837): Irischer Mathematiker; Betreiber verschiedener Schulen; bekannt durch das "Horner-Schema" (1830) zur Auswertung algebraischer Gleichungen; dessen Prinzip war aber bereits vorher anderen Autoren bekannt (früheste Quelle ist Zhu Shijie im China des 13. Jahrh.).*

Zur Auswertung eines Polynoms in der allgemeineren Darstellung

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)\ldots(x - x_{n-1})$$

mit gewissen Zahlen x_i, $i = 0, \ldots, x_{n-1}$, wird das Horner Schema wie folgt modifizert:

$$b_n = a_n, \quad k = n-1, \ldots, 0: \quad b_k = a_k + (\xi - x_k)b_{k+1}, \quad p(\xi) = b_0. \tag{1.3.12}$$

Regel 1.3.3: *Die Auswertung von Polynomen sollte mit Hilfe des Horner-Schemas erfolgen.*

Bei der Auswertung von Polynomen mit Hilfe des Horner Schemas spielt auch die Einsparung von arithmetischen Operationen eine Rolle. Dies wird hier und in vielen ähnlich gelagerten Fällen besonders wichtig, wenn die algorithmische Komponente sehr häufig wiederholt werden soll. Den Rechenaufwand zählt man dabei üblicherweise in Form von *arithmetische Operationen* (abgekürzt "a.Op."). Eine a.Op. setzt sich zusammen aus jeweils einer Addition und einer Multiplikation. Die Unterschiedliche Betrachtung von Addition und Multiplikation geschah bisher aus technischen Gründen, da auf den älteren Prozessoren eine Multiplikation deutlich mehr Zeit in Anspruch nahm als eine Addition. Inzwischen hat sich dieser Unterschied aber nivelliert, so daß Addition und Multiplikation als praktisch gleich schnell im Verhältnis zur etwas langsameren Division angesetzt werden müßten.

Beispiel 1.7: Ausführungszeiten von jeweils 10^9 a. Op.:

1) ("Antiker") Prozessor 68040 von Motorola (Macintosh Quadra 700):

Addition 1220 Sek., Multiplikation 1320 Sek., Division 2540 Sek.

2. Älterer 32 Bit-Prozessor Atlon 1.4 GHz (Standard PC):

Addition $4,5$ Sek., Multiplikation $4,5$ Sek., Division 6 Sek.

3. Aktueller 64 Bit-Prozessor Atlon 3500+ (High-end PC):

Addition $2,5$ Sek., Multiplikation $2,5$ Sek., Division $4,5$ Sek.

Bei diesen Leistungsmessungen spielt die verwendete Optimierungsstufe des Compilers eine nicht unwesentliche Rolle. Diese Zahlen sind erreichbar sowohl in C/C++ als auch in MATLAB.

1.4 Übungsaufgaben

Aufgabe 1.4.1: Man schreibe die folgenden Ausdrücke in der Form $f(h) = O(h^p)$, für $h \searrow 0$ mit möglichst großem $p \in \mathbb{N}$, bzw. $g(n) = O(n^q)$ für $n \nearrow \infty$ mit möglichst kleinem $q \in \mathbb{N}$:

$$a) \quad f(h) = 4(h^2+h)^2 - 4h^4, \qquad b) \quad g(n) = 4(n^2+n)^2 - 4n^4,$$
$$c) \quad f(h) = \frac{e^h - e^{-h}}{2h} - 1, \qquad d) \quad g(n) = \sup_{x>0} \frac{1-e^{-nx}}{1-e^{-x}}.$$

e) Wie läßt sich das asymptotische Verhalten von $f(h) = 1/\ln(h)$ beschreiben?

Aufgabe 1.4.2: Man untersuche die Konditionierung der folgenden Rechenoperationen:

$$a) \quad f(x_1, x_2) = \frac{x_1}{x_2} \quad (x_2 \neq 0), \qquad b) \quad f(x_1, x_2) = x_1^{x_2} \quad (x_1 > 0).$$

Sind die einfachen Operationen $f(x) = 1/x$ und $f(x) = \sqrt{x}$ gut konditioniert?

Aufgabe 1.4.3: Die Ausdrücke

$$a(x) = \frac{1-x}{1+2x} - \frac{1-2x}{1+x}, \qquad b(x) = \frac{3x^2}{(1+2x)(1+x)}$$

stellen für $x > 0$ dieselbe Funktion $f(x)$ dar.

a) Wie sieht es mit der Konditionierung der jeweiligen numerischen Aufgaben, $f(x)$ für $0 < |x| \ll 1$ aus diesen Darstellungen zu berechnen?

b) Wie würde man bei der praktischen Auswertung von $f(x)$ für $0 < |x| \ll 1$ zur Gewährleistung guter numerischer Stabilität vorgehen?

Aufgabe 1.4.4: Man gebe einen Weg zur experimentellen Bestimmung der Maschinengenauigkeit

$$\text{eps} := \max_{x \in D, x \neq 0} \left| \frac{\text{rd}(x) - x}{x} \right|$$

an. Dabei kann verwendet werden, daß für die Maschinenoperationen $\circledast$ gilt:

$$x \circledast y = (x * y)(1+\varepsilon), \quad x, y \in A, \ |\varepsilon| \leq \text{eps}.$$

Aufgabe 1.4.5: (Praktische Aufgabe):

a) Man bestimme mit einem Testprogramm die Maschinengenauigkeit des benutzten Rechners (in der jeweils verwendeten Programmiersprache).

b) Man schreibe ein (Matlab-)Programm zur Berechnung der Exponentialfunktion e^x mit Hilfe ihrer Taylor-Summen

$$T_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Man plotte für $n \in [0, 20]$ den relativen Fehler für die Argumente $x \in \{10, 1, -1, -10\}$. Man erkläre die schlechten Ergebnisse für negative Argumente und gebe eine Modifikation an, mit deren Hilfe negative wie positive Argumente gleich gut behandelt werden können.

Aufgabe 1.4.6: Man schreibe die folgenden Ausdrücke in der Form $f(h) = O(h^m)$ bzw. $f(h) = o(h^m)$ für $h \in \mathbb{R}_+$, $h \to 0$, mit einem möglichst großem $m \in \mathbb{N}$:

$$a) \quad f(h) = \frac{\sin(1+h) - 2\sin(1) + \sin(1-h)}{h^2} + \sin(1);$$
$$b) \quad f(h) = \frac{h}{\ln(h)}.$$

Aufgabe 1.4.7: Wie groß ist in erster Näherung der relative Fehler bei der Bestimmung der Molmenge m eines idealen Gases (mit Gaskonstante $\gamma = 0,082$) aus der Formel

$$m(P, V, T) = \frac{PV}{\gamma T},$$

wenn die Temperatur T mit $200 \pm 0,5$ Grad, der Druck P mit $2 \pm 0,01$ atm und das Volumen V mit $10 \pm 0,2$ l bestimmt wurden. Welche Messung muß verfeinert werden, um den Fehler unter 1% zu drücken?

Aufgabe 1.4.8: In vielen Fällen kann die Konvergenzordnung eines Grenzprozesses

$$a(h) \to a \ \ (h \to 0), \quad a(h) - a = O(h^\alpha),$$

nur experimentell bestimmt werden. Dazu werden bei bekanntem Limes a für zwei Werte h und $h/2$ die Fehler $a(h) - a$ und $a(h/2) - a$ berechnet und dann die Ordnung α über den formalen Ansatz $a(h) - a = ch^\alpha$ aus der folgenden Formel ermittelt:

$$\alpha = \frac{1}{\log(2)} \log\left(\left| \frac{a(h) - a}{a(h/2) - a} \right| \right).$$

a) Man rekapituliere die Rechtfertigung dieser Formel und überlege, wie man vorgehen könnte, wenn kein exakter Limes a bekannt ist.

b) Man bestimme die inhärenten Konvergenzordnungen für die folgenden von Funktionen $a(h)$ und $b(h)$ abgegriffenen Werte:

h	$a(h)$	$b(h)$
2^{-1}	7.188270827204928	8.89271737217539
2^{-2}	7.095485351135761	8.971800326329658
2^{-3}	7.047858597600531	8.992881146463981
2^{-4}	7.023726226390662	8.998220339291473
2^{-5}	7.011579000356371	8.999559782988968
2^{-5}	7.005485409034109	8.999895247704067
Limes	$a(0) = 7.0$	$b(0) =?$

Aufgabe 1.4.9: Man betrachte die Funktion

$$f(x) = \frac{1 - \cos(x)}{x}.$$

a) Für welche x ist die Auswertung von $f(x)$ gut bzw. schlecht konditioniert?

b) Man gebe für $|x| \ll 1$ einen stabilen Algorithmus zur Berechnung von $f(x)$ an. Dabei sei angenommen, daß $\cos(x)$ mit Maschinengenauigkeit berechnet wird. (Hinweis: Die Darstellung von f kann mit Hilfe der Rechenregeln für trigonometrische Funktionen umgeformt werden.)

Aufgabe 1.4.10: (Praktische Aufgabe): Man berechne Näherungswerte

$$\sum_{k=0}^{n} \frac{x^k}{k!} \approx e^x$$

für $x = -5,5$ mit $n = 1, 2, \ldots, 30$, auf die folgenden drei Arten:

1) mit der obigen Formel;

2) mit der Umformung $e^{-5,5} = 1/e^{5,5}$ und der obigen Formel;

3) mit der Umformung $e^{-5,5} = (e^{-0,5})^{11}$ und der obigen Formel.

Der exakte Wert ist $e^{-5,5} = 0,0040867714\ldots$. Wie sind die beobachteten Effekte zu interpretieren? Dies ist ein Beispiel dafür, daß scheinbar kleine Modifikationen in numerischen Algorithmen gravierende Konsequenzen für die Approximationsgenauigkeit haben können. Welche Ergebnisse ergeben sich, wenn die Auswertung der Taylor-Polynome mit Hilfe des Horner-Schemas erfolgt?

Aufgabe 1.4.11: Sei $A \in \mathbb{R}^{n\times n}$ beliebig gegeben. Man gebe einen Algorithmus an zur Auswertung des Matrixpolynoms

$$p(A) = \sum_{i=0}^{m} a_i A^i$$

mit Koeffizienten $a_i \in \mathbb{R}$, der möglichst wenig Speicherplatz und arithmetische Operationen (1 a.Op. = 1 Mult. + 1 Add.) benötigt.

Aufgabe 1.4.12: Es seien die Nullstellen eines Polynoms $p(x) = \sum_{i=0}^{m} a_i x^i$ zu bestimmen. Man zeige, daß für eine Näherung $\tilde{z}$ zu einer einfachen Nullstelle $z \neq 0$ in erster Näherung die folgende Abschätzung gilt:

$$\Big|\frac{\tilde{z}-z}{z}\Big| \stackrel{.}{\leq} \Big|\frac{p(\tilde{z})}{p'(z)z}\Big|.$$

Dies motiviert die Genauigkeitskontrolle bei der Berechnung von Nullstellen von Polynomen in der Praktischen Aufgabe 5. (Hinweis: Die Aufgabe ist leichter als sie aussieht; Taylor-Entwicklung.)

Aufgabe 1.4.13: Die Funktion $f(x) = x + 1$ stelle eine physikalische Größe dar, von der Werte $\tilde{f}(x_i) \approx f(x_i)$ an äquidistant verteilten Punkten

$$x_i = ih, \quad 0 \leq i \leq n := 10^3, \quad h = 10^{-3},$$

mit einem maximalen Fehler von $0,1\%$ gemessen werden. Man zeige, daß bei der Approximation der Ableitungswerte $f'(x_i)$ mit dem zentralen Differenzenquotienten

$$f'(x_i) \approx \frac{\tilde{f}(x_{i+1}) - \tilde{f}(x_{i-1})}{2h}, \quad i = 1, \ldots, n-1$$

aus diesen Werten ein relativer Fehler von 100% auftreten kann. Dies zeigt die Fragwürdigkeit der Approximation von Ableitungen durch Differenzenquotienten. (Hinweis: Man konstruiere spezielle Störungen.)

Aufgabe 1.4.14: Gegeben seien $n+1$ paarweise verschiedene Punkte $x_i \in \mathbb{R}^1, i = 0, 1, \ldots, n$, und die zugehörigen $n+1$ sog. Lagrangeschen Polynome

$$L_i^{(n)}(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 0, \ldots, n.$$

Man zeige, daß die Polynome $\{L_i^{(n)}, i = 0, \ldots, n\}$, eine Basis des Polynomraums P_n (Vektorraum aller Polynome vom Grad kleiner oder gleich n) bilden, und daß die folgenden

Beziehungen gelten:

$$i) \qquad \sum_{i=0}^{n} L_i^{(n)}(x) = 1, \quad x \in \mathbb{R}^1,$$

$$ii) \qquad \sum_{i=0}^{n} x_i^k L_i^{(n)}(0) = 0, \quad k = 1, \ldots, n,$$

$$iii) \qquad \sum_{i=0}^{n} x_i^{n+1} L_i^{(n)}(0) = (-1)^n \prod_{i=0}^{n} x_i.$$

(Hinweis: Man verwende die Eindeutigkeit des Lagrangeschen Interpolationspolynoms und die Darstellung des Fehlers bei der Lagrange-Interpolation.)

Aufgabe 1.4.15: (Praktische Aufgabe): Man schreibe ein Programm zur Berechnung der reellen Lösungen der quadratischen Gleichung

$$p(x) = ax^2 + bx + c = 0,$$

zu gegebenen $a, b, c \in \mathbb{R}$. Es sollen alle möglichen Fälle der Degenerierung (z. B.: $a = 0$) berücksichtigt und der Einfluß des Rundungsfehlers minimiert werden. Man erprobe das Programm anhand der folgenden Fälle:

a:	0	0	0	0	2	2	2	4	1	1	1	−1	−1	−4	−1	−4	−1	−1	$2,5 \cdot 10^9$
b:	0	0	1	2	0	0	0	2	2	2	2	0	0	0	2	8	2	2	-10^5
c:	0	1	0	1	0	1	−4	0	−3	1	2	0	−1	2	0	12	−1	−5	1

Die berechneten Lösungen sollen akzeptiert werden, wenn das heuristische Kriterium (s. Aufgabe 2)

$$\left| \frac{p(\tilde{z})}{p'(\tilde{z})\tilde{z}} \right| < 10^{-12}$$

erfüllt ist.

2 Interpolation und Approximation

Ein Grundproblem der numerischen Praxis ist die Darstellung und Auswertung von Funktionen. Dabei ergeben sich folgende Aufgabenstellungen:

(i) Eine Funktion $f(x)$ ist nur in einer diskreten Menge von Argumenten $x_0, \ldots, x_n$ bekannt und soll mit dieser Information rekonstruiert werden (z.B. zur graphischen Darstellung oder zur Auswertung an Zwischenstellen).

(ii) Eine analytisch gegebene Funktion $f(x)$ soll auf der Rechenanlage so dargestellt werden, daß jederzeit Funktionswerte zu beliebigem Argument x leicht berechnet werden können (z.B. trigonometrische Funktionen).

In beiden Fällen hat man ein System mit unendlich vielen Freiheitsgraden, nämlich die funktionale Abhängigkeit $y = f(x)$, durch einen endlichen Datensatz zu simulieren. Hierzu bedient man sich gewisser Klassen P von einfach strukturierten Funktionen; z.B.:

Polynome:	$p(x) = a_0 + a_1 x + \ldots + a_n x^n,$
rationale Funktionen:	$r(x) = \dfrac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_m x^m},$
trigonometrische Polynome:	$t(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n} \{a_k \cos(kx) + b_k \sin(kx)\}.$
Exponentialsummen:	$e(x) = \sum_{k=1}^{n} a_k \exp(b_k x).$

Definition 2.1: *Geschieht die Zuordnung eines Elementes $g \in P$ zur Funktion f durch Fixieren von Funktionswerten*

$$g(x_i) = y_i := f(x_i), \qquad i = 0, \ldots, n,$$

so spricht man von "Interpolation". Ist $g \in P$ als in einem gewissen Sinne "beste" Darstellung von f zu bestimmen, z.B.:

$$\max_{a \le x \le b} |f(x) - g(x)| \quad \textit{minimal für} \quad g \in P,$$

oder

$$\Big(\int_a^b |f(x) - g(x)|^2 \, dx \Big)^{1/2} \quad \textit{minimal für} \quad g \in P,$$

so spricht man allgemein von "Approximation". Die jeweilige Wahl der Konstruktion von $g \in P$ hängt von der zu erfüllenden Aufgabe ab. Offenbar ist die Interpolation eine spezielle Art der Approximation mit

$$\max_{i=0,\ldots,n} |f(x_i) - g(x_i)| \quad \textit{minimal für} \quad g \in P.$$

2.1 Polynominterpolation

Wir bezeichnen mit P_n den Vektorraum der Polynome vom Grad kleiner oder gleich n:

$$P_n := \{p(x) = a_0 + a_1 x + \ldots + a_n x^n \,|\, a_i \in \mathbb{R},\ i = 0, \ldots, n\}\,.$$

Definition 2.2: *Die sog. "Langrangesche*[1] *Interpolationsaufgabe" besteht darin, zu $n+1$ paarweise verschiedenen Stützstellen (auch "Knoten" genannt) $x_0, \ldots, x_n \in \mathbb{R}$ und gegebenen Knotenwerten $y_0, \ldots, y_n \in \mathbb{R}$ ein Polynom $p \in P_n$ zu bestimmen mit der Eigenschaft*

$$p(x_i) = y_i, \quad i = 0, \ldots, n. \tag{2.1.1}$$

Satz 2.1 (Langrange-Interpolation): *Die Langrangesche Interpolationsaufgabe ist eindeutig lösbar.*

Beweis: Wir zeigen zunächst die Eindeutigkeit. Sind $p_1, p_2 \in P_n$ zwei Lösungen, so gilt für $p := p_1 - p_2 \in P_n$: $p(x_i) = 0, i = 0, \ldots, n$, d. h.: p hat $n+1$ Nullstellen, und ist folglich identisch Null (folgt mit Hilfe des Satzes von Rolle[2]). Zur Existenz betrachten wir die Gleichungen $p(x_i) = y_i, i = 0, \ldots, n$. Dies kann man interpretieren als ein lineares Gleichungssystem mit $n+1$ Gleichungen für die $n+1$ unbekannten Koeffizienten $a_0, \ldots, a_n$ von $p \in P_n$. Wegen der Eindeutigkeit von p muß dieses System dann notwendig eine Lösung haben. Q.E.D.

Zur Konstruktion des Interpolationspolynoms $p \in P_n$ verwenden wir die sog. "Lagrangeschen Basispolynome"

$$L_i^{(n)}(x) := \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \in P_n\,, \qquad i = 0, \ldots, n.$$

Daß der Satz von Polynomen $\{L_i^{(n)}, i = 0, \ldots, n\}$ tatsächlich eine Basis von P_n ist, wird als Übungsaufgabe gestellt. Offenbar ist

$$L_i^{(n)}(x_k) = \left\{ \begin{array}{ll} 1\,, & \text{falls } i = k \\ 0\,, & \text{falls } i \neq k \end{array} \right\} =: \delta_{ik} \quad \text{(Kronecker-Symbol)}.$$

Definition 2.3: *Das Polynom*

$$p := \sum_{i=0}^{n} y_i L_i^{(n)} \in P_n \tag{2.1.2}$$

[1] Joseph Louis de Lagrange (1736-1813): französischer Mathematiker; 1766-87 Direktor der mathem. Klasse der Berliner Akademie, dann Prof. in Paris; bahnbrechende Arbeiten zur Variationsrechnung, zur komplexen Funktionentheorie sowie zur theor. Mechanik und Himmelsmechanik.

[2] Michel Rolle (1652-1719): französischer Mathematiker und Autodidakt; wirkte in Paris und leistete Beiträge zur Analysis, Algebra und Geometrie; der nach ihm benannten Satz wurde 1691 publiziert.

hat dann die gewünschten Eigenschaften $p(x_j) = y_j$ *. Es wird die "Lagrangesche Darstellung" des (Lagrangeschen) Interpolationspolynoms zu den Stützpunkten* $(x_0, y_0), \ldots, (x_n, y_n)$ *genannt (abgekürzt "Lagrangesches Interpolationspolynom").*

Das Lagrangesche Darstellung des Interpolationspolynoms hat den Nachteil, daß sich die verwendeten Basisfunktionen von P_n bei Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) völlig ändern. Dies wird vermieden bei Verwendung der sog. "Newtonschen[3] Basispolynome"

$$N_0(x) := 1\,; \qquad i = 1, \ldots, n: \quad N_i(x) := \prod_{j=0}^{i-1} (x - x_j).$$

Aus diesen läßt sich das Interpolationspolynom systematisch aufbauen. Für den Ansatz

$$p(x) = \sum_{i=0}^{n} a_i N_i(x)$$

findet man durch sukzessive Auswertung in $x_0, \ldots, x_n$ das gestaffelte Gleichungssystem

$$\begin{aligned}
y_0 &= p(x_0) = a_0 \\
y_1 &= p(x_1) = a_0 + a_1(x_1 - x_0) \\
&\vdots \\
y_n &= p(x_n) = a_0 + a_1(x_n - x_0) + \ldots + a_n(x_n - x_0)\ldots(x_n - x_{n-1}),
\end{aligned}$$

woraus sich die Koeffizienten a_i rekursiv berechnen lassen. Die Hinzunahme eines weiteren Punktes (x_{n+1}, y_{n+1}) ist nun leicht durch Fortsetzung dieses Prozesses mit der Basisfunktion N_{n+1} zu bewerkstelligen. In der Praxis bestimmt man die a_i jedoch auf eine andere, numerisch stabilere Weise, die im folgenden beschrieben wird.

Satz 2.2 (Newtonsche Darstellung): *Das Lagrangesche Interpolationspolynom zu den Punkten* $(x_i, y_i), i = 0, \ldots, n$*, läßt sich bzgl. der Newtonschen Polynombasis schreiben in der Form*

$$p(x) = \sum_{i=0}^{n} y[x_0, \ldots, x_i] N_i(x). \tag{2.1.3}$$

Dabei bezeichnen $y[x_0, \ldots, x_i]$ *die zu den Punkten* (x_i, y_i) *gehörenden sog. "dividierten Differenzen", welche rekursiv definiert sind durch*

$$\begin{aligned}
i = 0, \ldots, n: \qquad & y[x_i] := y_i \\
k = 1, \ldots, n - i: \qquad & y[x_i, \ldots, x_{i+k}] := \frac{y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}.
\end{aligned}$$

[3] Isaac Newton (1643-1727): englischer Physiker und Mathematiker; Professor an der Universität Cambridge; entwickelte u.a. die Grundlagen der klassischen Mechanik und der Differentialrechnung.

Beweis: Es bezeichne $p_{i,i+k} \in P_k$ das Polynom, welches die Punkte $(x_i, y_i), \ldots, (x_{i+k}, y_{i+k})$, interpoliert. Speziell ist also $p_{0,n} = p$ das gesuchte Interpolationspolynom. Wir zeigen

$$p_{i,i+k}(x) = y[x_i] + y[x_i, x_{i+1}](x - x_i) + \ldots + y[x_i, \ldots, x_{i+k}](x - x_i)\ldots(x - x_{i+k-1}),$$

was offensichtlich die Aussage des Satzes als Spezialfall beinhaltet. Der Beweis wird durch Induktion bzgl. der Indexdifferenz $k = (i+k) - i$ geführt. Für $k = 0$ ist $p_{i,i} = y_i = y[x_i]$, $i = 0, \ldots, n$. Sei die Behauptung richtig für $k - 1 \geq 0$. Konstruktionsgemäß gilt nun

$$p_{i,i+k}(x) = p_{i,i+k-1}(x) + a(x - x_i)\ldots(x - x_{i+k-1}).$$

Zu zeigen ist also, daß $a = y[x_i, \ldots, x_{i+k}]$. Offenbar ist a der Koeffizient von x^k in $p_{i,i+k}(x)$. Nach Induktionsannahme ist weiter

$$\begin{aligned} p_{i,i+k-1}(x) &= \ldots + y[x_i, \ldots x_{i+k-1}]x^{k-1}, \\ p_{i+1,i+k}(x) &= \ldots + y[x_{i+1}, \ldots, x_{i+k}]x^{k-1}, \end{aligned}$$

wobei "..." für Polynomanteile vom Grad kleiner oder gleich $k-2$ steht. Wie man leicht verifiziert, interpoliert das durch

$$\begin{aligned} q(x) &= \frac{(x - x_i)p_{i+1,i+k}(x) - (x - x_{i+k})p_{i,i+k-1}(x)}{x_{i+k} - x_i} \\ &= p_{i,i+k-1}(x) + (x - x_i)\frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i} \end{aligned}$$

gegebene Polynom $q \in P_k$ die $k+1$ Stützpunkte (x_j, y_j), $j = i, \ldots, i+k$. Wegen der Eindeutigkeit des Interpolationspolynoms ist dann notwendig $q \equiv p_{i,i+k}$. Der führende Koeffizient in $p_{i,i+k}(x)$ ist demnach

$$a = \frac{y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} = y[x_i, \ldots, x_{i+k}],$$

was den Beweis vervollständigt. Q.E.D.

Korollar 2.1: *Die Aussage von Satz 2.2 impliziert eine wichtige Invarianzeigenschaft der dividierten Differenzen. Der führende Koeffizient $y[x_0, \ldots, x_n]$ des Lagrangeschen Interpolationspolynoms ist gleichzeitig der Koeffizient des Monoms x^n in seiner Standarddarstellung. Da dieser unabhängig von der Reihenfolge in der Anordnung der Punkte $x_0, \ldots, x_n$ ist, gilt dasselbe folglich auch für die dividierte Differenz, d. h.: Es gilt*

$$y[\tilde{x}_0, \ldots, \tilde{x}_n] = y[x_0, \ldots, x_n] \tag{2.1.4}$$

für jede beliebige Permutation $\tilde{x}_0, \ldots, \tilde{x}_n$ dieser Punkte.

Die im Beweis von Satz 2.2 verwendete Beziehung zwischen den Polynomen $p_{i,i+k}$ kann direkt zur rekursiven Berechnung des Interpolationspolynoms $p = p_{0,n}$ verwendet werden.

Definition 2.4: *Das durch dir Rekursion*

$$\begin{aligned} i = 0,\ldots,n: &\quad p_{i,i}(x) = y_i, \\ k = 0,\ldots,n-i: &\quad p_{i,i+k}(x) = p_{i,i+k-1}(x) + (x - x_i)\frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i}, \end{aligned} \tag{2.1.5}$$

erzeugte Polynom $p_{0,n}$ *ist die sog. "Nevillesche*[4] *Darstellung" des Interpolationspolynoms zu den Stützpunkten* $(x_0, y_0), \ldots, (x_n, y_n)$ *.*

Bei der praktischen Berechnung der Nevilleschen darstellung geht man nach folgendem Schema vor:

x_0	y_0	$p_{0,1}(x)$	$p_{0,2}(x)$	$p_{0,3}(x)$	$\ldots$	$p_{0,n-1}(x)$	$p_{0,n}(x)$
x_1	y_1	$p_{1,2}(x)$	$p_{1,3}(x)$	$p_{1,4}(x)$	$\ldots$	$p_{1,n}(x)$	
x_2	y_2	$p_{2,3}(x)$	$p_{2,4}(x)$	$p_{2,5}(x)$	$\ldots$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$			
x_{n-1}	y_{n-1}	$p_{n-1,n}(x)$					
x_n	y_n						

Auch hier ist die Hinzunahme eines weiteren Stützpunktes (x_{n+1}, y_{n+1}) problemlos. Die Nevillesche Darstellung des Interpolationspolynoms bietet eine sehr effiziente und numerisch stabile Möglichkeit zur Berechnung einzelner Funktionswerte $p(\xi)(\xi \neq x_i)$ ohne vorherige Bestimmung der Koeffizienten in der Newtonschen Darstellung. Dazu setzt man im obigen Neville-Schema einfach $x = \xi$ und verwendet zur Berechnung von $p_{i,k} := p_{i,k}(\xi)$ (aus Stabilitätsgründen) die Rekursionsformeln

$$\begin{aligned} i = 0,\cdots,n: &\quad p_{i,i} = y_i \\ k = 1,\cdots,n-i: &\quad p_{i,i+k} = p_{i,i+k-1} + (\xi - x_i)\frac{p_{i+1,i+k} - p_{i,i+k-1}}{x_{i+k} - x_i} \\ &\qquad = p_{i,i+k-1} + \frac{p_{i,i+k-1} - p_{i+1,i+k}}{(\xi - x_{i+k})/(\xi - x_i) - 1}. \end{aligned} \tag{2.1.6}$$

2.1.1 Auswertung von Polynomen

Ist ein Polynom $p \in P_n$ in der Form $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ gegeben, so werden einzelne Werte $p(\xi)$ mit Hilfe des sog. "Horner-Schemas" berechnet (siehe Kapitel 1):

$$b_n = a_n; \qquad k = n-1,\ldots,0: \quad b_k \equiv a_k + \xi b_{k+1}; \qquad p(\xi) = b_0.$$

[4] *Eric Harold Neville (1889-1961): Englischer Mathematiker; Professor an der Universität in Reading, England (1919-1954); Beiträge zur numerischen Mathematik, u. a. zur praktischen Polynominterpolation.*

Zu $p_n := p \in P_n$ wird durch

$$p_{n-1}(x) := b_1 + b_2 x + \ldots + b_n x^{n-1}$$

ein Polynom $p_{n-1} \in P_{n-1}$ definiert. Wegen $a_k = b_k - \xi b_{k+1}$ gilt offenbar

$$p(x) = (x-\xi)p_{n-1}(x) + r_0, \qquad r_0 = p(\xi) = b_0,$$

d. h.: Das Horner-Schema leistet unter anderem die Abspaltung des Linearfaktors $x - \xi$ vom Polynom $p(x)$ (Euklidischer Algorithmus). Weiter ist dann

$$\frac{p(x) - p(\xi)}{x - \xi} = p_{n-1}(x), \qquad x \neq \xi,$$

d. h.: Für $x \to \xi$ folgt die Beziehung $p'(\xi) = p_{n-1}(\xi)$. Zur Berechnung von $p'(\xi)$ wird das Horner-Schema auf das Polynom p_{n-1} angewendet. Dies liefert Koeffizienten c_k, $k = 2, \ldots, n$, sowie ein Polynom $p_{n-2} \in P_{n-2}$ mit der Eigenschaft

$$p_{n-1}(x) = (x-\xi)p_{n-2}(x) + r_1, \qquad r_1 = p_{n-1}(\xi) = c_1.$$

Durch fortgesetzte Abspaltung des Linearfaktors $x-\xi$ erhält man so eine (endliche) Folge von Polynomen $p_n, p_{n-1}, p_{n-2}, \ldots, p_0$ mit der Eigenschaft

$$p_{n-j}(x) = (x-\xi)p_{n-j-1}(x) + r_j, \quad j = 0, \ldots, n-1; \qquad p_0 = r_n,$$

und damit die Darstellung

$$p_n(x) = r_0 + r_1(x-\xi) + \ldots + r_n(x-\xi)^n. \tag{2.1.7}$$

Vergleicht man dies mit der Taylor-Entwicklung von p an der Stelle ξ, so findet man

$$r_j = \frac{1}{j!}p^{(j)}(\xi), \qquad j = 0, \ldots, n. \tag{2.1.8}$$

Die Koeffizienten des Polynoms p_{n-j} seien mit $a_k^{(j)}$ bezeichnet:

$$p_{n-j}(x) = a_j^{(j)} + a_{j+1}^{(j)} x + \ldots + a_n^{(j)} x^{n-j}, \qquad j = 0, \ldots, n.$$

Sie werden berechnet durch die rekursive Vorschrift

$$\begin{aligned} j = 0, \ldots, n: &\qquad a_n^{(j+1)} = a_n^{(j)}, \\ k = n-1, \ldots, j: &\quad a_k^{(j+1)} = a_k^{(j)} + \xi a_{k+1}^{(j+1)}, \end{aligned}$$

und es gilt $p^{(j)}(\xi) = j! a_j^{(j+1)}$, $j = 0, \ldots, n$.

Die Koeffizienten $a_k^{(j)}$ bilden das sog. "vollständige Horner-Schema" zur Auswertung des Polynoms p im Punkt ξ.

Das Horner-Schema kann leicht zur Auswertung eines Polynoms

$$p(x) = a_0 + a_1(x - x_0) + \ldots + a_n(x - x_0)\ldots(x - x_{n-1})$$

in allgemeiner Newtonscher Darstellung modifiziert werden:

$$b_n = a_n; \qquad k = n-1, \ldots, 0: \quad b_k \equiv a_k + (\xi - x_k)b_{k+1}; \qquad p(\xi) = b_0.$$

Ist man an den Ableitung $p^{(j)}(\xi)$ eines in Newtonscher Form gegebenen Polynoms interessiert, so ist dies weitgehend äquivalent mit dem Problem, zu einer gegebenen Darstellung

$$p(x) = a_0 + a_1(x - x_0) + \ldots + a_n(x - x_0)\ldots(x - x_{n-1})$$

die Koeffizienten b_k, $k = 0, \ldots, n$, in der Darstellung

$$p(x) = b_0 + b_1(x - y_0) + \ldots + b_n(x - y_0)\ldots(x - y_{n-1})$$

bzgl. anderer (paarweise verschiedener) Punkte y_k, $k = 0, \ldots, n-1$, zu bestimmen. Dies wird durch das folgende "verallgemeinerte Horner-Schema" geleistet:

$$\begin{aligned}
&k = 0, \ldots, n: && a_k^{(0)};\\
&j = 0, \ldots, n-1: && a_n^{(j+1)} = a_n^{(j)};\\
&k = n-1, \ldots, j: && a_k^{(j+1)} = a_k^{(j)} + (y_j - x_{k-j})a_{k+1}^{(j+1)}, \quad b_j = a_j^{(j+1)}.
\end{aligned}$$

2.1.2 Interpolation von Funktionen

Wir betrachten nun den Fall, daß die Knotenwerte y_i durch eine Funktion f auf einem die Stützpunkte x_i enthaltenden Intervall $[a, b]$ gegeben sind:

$$y_i = f(x_i), \quad x_i \in [a, b], \quad i = 0, \ldots, n.$$

Zunächst stellt sich die Frage, wie gut das zugehörige Interpolationspolynom $p \in P_n$ die Funktion f auf $[a, b]$ approximiert. Im Folgenden bezeichne $\overline{(x_0, \ldots, x_n)}$ das kleinste Intervall, das alle in den Klammern eingeschlossenen Punkte enthält. Ferner bezeichnet $C[a, b]$ den Vektorraum der über $[a, b]$ *stetigen* Funktionen und analog $C^k[a, b]$ den Vektorraum der über $[a, b]$ *k-mal stetig differenzierbaren* Funktionen.

Satz 2.3 (Interpolationsfehler 1): *Sei $f \in C^{n+1}[a, b]$. Dann gibt es zu jedem $x \in [a, b]$ ein $\xi_x \in \overline{(x_0, \ldots, x_n, x)}$, so daß gilt:*

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j). \tag{2.1.9}$$

Beweis: Für $x \in \{x_0, \ldots, x_n\}$ ist alles klar. Sei also $x \in [a,b] \backslash \{x_0, \ldots, x_n\}$. Wir setzen

$$l(t) := \prod_{j=0}^{n}(t - x_j), \qquad c(x) := \frac{f(x) - p(x)}{l(x)}.$$

Die Funktion $F(t) = f(t) - p(t) - c(x)l(t)$ besitzt dann mindestens die $n+2$ Nullstellen $x_0, \ldots, x_n, x$ in $[a,b]$. Durch wiederholte Anwendung des Satzes von Rolle erschließt man, daß dann die Ableitung $F^{(n+1)}$ eine Nullstelle $\xi_x \in \overline{(x_0, \ldots, x_n, x)}$ hat. Mit

$$0 = F^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - p^{(n+1)}(\xi_x) - c(x)l^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - c(x)(n+1)\,!$$

folgt die Behauptung. Q.E.D.

Satz 2.4 (Interpolationsfehler 2): *Sei $f \in C^{n+1}[a,b]$. Dann gestattet der Fehler bei der Polynominterpolation für $x \in [a,b] \backslash \{x_0, \cdots, x_n\}$ die Darstellung*

$$f(x) - p(x) = f[x_0, \ldots, x_n, x] \prod_{j=0}^{n}(x - x_j), \tag{2.1.10}$$

mit der Notation $f[x_i, \ldots, x_{i+k}] := y[x_i, \ldots, x_{i+k}]$, und es ist

$$f[x_0, \ldots, x_n, x] = \int_0^1 \int_0^{t_1} \ldots \int_0^{t_{n-1}} \int_0^{t_n} f^{(n+1)}\left(x_0 + t_1(x_1 - x_0) + \ldots \right.$$
$$\left. + t_n(x_n - x_{n-1}) + t(x - x_n)\right)\, dt dt_n \ldots dt_2 dt_1.$$

Beweis: Der Beweis wird durch Induktion nach der Anzahl der Stützstellen (in der Reihung $x_0, x_1, x_2, \ldots$) geführt. Für $n = 0$ gilt trivialerweise

$$f(x) - p_0(x) = f(x) - f(x_0) = \begin{cases} f[x_0, x](x - x_0), \\ (x - x_0) \int_0^1 f'(x_0 + t(x - x_0))\, dt. \end{cases}$$

Sei die Behauptung nun richtig für $n - 1 \geq 0$. Dann ist

$$\begin{aligned} f(x) - p_n(x) &= f(x) - \sum_{i=0}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1}(x - x_j) \\ &= f(x) - p_{n-1}(x) - f[x_0, \ldots, x_n] \prod_{j=0}^{n-1}(x - x_j) \\ &= f[x_0, \ldots, x_{n-1}, x] \prod_{j=0}^{n-1}(x - x_j) - f[x_0, \ldots, x_n] \prod_{j=0}^{n-1}(x - x_j) \\ &= \frac{f[x_0, \ldots, x_{n-1}, x] - f[x_0, \ldots, x_n]}{x - x_n} \prod_{j=0}^{n}(x - x_j), \end{aligned}$$

und somit, wegen $f[x_0,\ldots,x_{n-1},x] = f[x,x_0,\ldots,x_{n-1}]$, unter Ausnutzung der Definition der dividierten Differenzen:

$$f(x) - p_n(x) = f[x_0,\ldots,x_n,x]\prod_{j=0}^{n}(x-x_j)\,.$$

Ferner ist nach Induktionsvoraussetzung

$$\begin{aligned}
&f[x_0,\ldots,x_{n-1},x] - f[x_0,\ldots,x_n]\\
&= \int_0^1\int_0^{t_1}\ldots\int_0^{t_{n-1}} \big\{f^{(n)}(x_0 + t_1(x_1-x_0) + \ldots + t_n(x-x_{n-1}))\\
&-f^{(n)}(x_0 + t_1(x_1-x_0) + \ldots + t_n(x_n-x_{n-1}))\big\}\; dt_n\ldots dt_1\\
&= \int_0^1\int_0^{t_1}\ldots\int_0^{t_n} \underbrace{\frac{d}{dt}f^{(n)}(x_0 + \ldots + t_n(x_n-x_{n-1}) + t(x-x_n))}_{f^{(n+1)}(\ldots)(x-x_n)}\; dtdt_n\ldots dt_1
\end{aligned}$$

und folglich, nach Definition der dividierten Differenzen,

$$f[x_0,\ldots,x_n,x] = \int_0^1\int_0^{t_1}\ldots\int_0^{t_{n-1}}\int_0^{t_n} f^{(n+1)}(\ldots)\,dtdt_n\ldots dt_1.$$

Dies vervollständigt den Beweis. Q.E.D.

Die obige Integraldarstellung der dividierten Differenzen $f[x_0,\ldots,x_n]$ gestattet ihre stetige Fortsetzung für den Fall, daß einige der Stützstellen zusammenfallen:

$$f[x_0,\ldots,x_r,x_r,\ldots,x_n] := \lim_{\varepsilon\to 0} f[x_0,\ldots,x_r,x_r+\varepsilon,\ldots,x_n].$$

Im Extremfall $x_0 = \ldots = x_n$ wird

$$f[x_0,\ldots,x_n] = \int_0^1\int_0^{t_1}\ldots\int_0^{t_{n-1}} f^{(n)}(x_0)\,dt_n\ldots dt_2dt_1 = \frac{1}{n!}f^{(n)}(x_0)\,,$$

und das Newtonsche Interpolationspolynom geht über in das Taylor-Polynom n-ten Grades von f in x_0:

$$p_n(x) = \sum_{i=0}^{n} f[x_0,\ldots,x_i]\prod_{j=0}^{i-1}(x-x_j) = \sum_{i=0}^{n}\frac{1}{i!}f^{(i)}(x_0)(x-x_0)^i\,.$$

Ferner sehen wir, daß sich die dividierten Differenzen einer hinreichend oft differenzierbaren Funktion durch Zwischenwerte der entsprechenden Ableitungen ausdrücken lassen:

$$f[x_0,\ldots,x_n,x] = \frac{1}{(n+1)!}f^{(n+1)}(\xi_x), \tag{2.1.11}$$

mit einer Zwischenstelle $\xi_x \in \overline{(x_0,\ldots,x_n,x)}$.

Wir wollen nun allgemein den Fehler bei der Lagrangeschen Interpolation diskutieren:

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j). \tag{2.1.12}$$

Für großes n wird $1/n!$ sehr klein, und das Produkt wird klein, wenn die Stützstellen immer dichter zusammenrücken. Sind die Ableitungen von f beschränkt auf $[a,b]$, so gilt also sicher

$$\max_{a \le x \le b} |f(x) - p(x)| \to 0 \quad (n \to \infty).$$

Meist haben die Ableitungen der zu interpolierenden Funktionen jedoch ein zu starkes Wachstum für $n \to \infty$, z.B.:

$$f(x) = (1+x^2)^{-1}, \qquad |f^{(n)}(x)| \approx 2^n n!\, O(|x|^{-2-n}),$$

so daß gleichmäßige Konvergenz des Approximationsprozesses nicht mehr zu erwarten ist.

Beispiel 2.1: Für die Funktion $f(x) = |x|,\ x \in [-1,1]$, ergibt die Lagrange-Interpolation in den Stützstellen $x_i = -1 + ih,\ i = 0, \ldots, 2m$, mit $h = 1/m$ und $x \notin \{x_i, i = 0, \ldots, 2m\}$ das Verhalten

$$p_m(x) \not\to f(x) \quad (m \to \infty).$$

Dieser Effekt ist nicht auf nicht differenzierbare Funktionen beschränkt, wie das obige Beispiel $f(x) = (1+x^2)^{-1},\ x \in [-5,5]$ zeigt (s. Übungsaufgabe).

Bemerkung 2.1: Der Weierstraßsche Approximationssatz besagt, daß jede Funktion $f \in C[a,b]$ beliebig gut gleichmäßig auf $[a,b]$ durch Polynome approximiert werden kann. Die Vermutung, daß dies mit Lagrangeschen Interpolationspolynomen geschehen kann, ist jedoch i. Allg. falsch.

Ein weiterer Defekt der Lagrange-Interpolation besteht in ihrer großen Fehlerempfindlichkeit. Fehlerhafte Daten $y_i + \Delta y_i$ wirken sich auf die Gestalt des Polynoms nicht nur lokal bei der Stützstelle x_i aus, sondern verändern den Verlauf auch relativ dramatisch über dem ganzen Intervall.

Beispiel 2.2: $x_i = -1 + ih,\ i = 0, \ldots, 2m,\ h = 1/m;\ y_i = 0$ für $i \neq m,\ y_m = \varepsilon;$

$$p(x) = \varepsilon \prod_{\substack{j=0 \\ j \neq m}}^{2m} \frac{x - x_j}{x_j} \qquad \text{(Lagrange-Polynom zum Aufpunkt } x_m = 0)$$

Dies liegt daran, daß auch für "gutartige" Funktionen der Term $|\prod_{j=0}^{n}(x - x_j)|$ für $x \notin \overline{(x_0, \ldots, x_n)}$ sehr schnell anwächst. Die Verwendung der Polynominterpolation zur weitreichenden "Extrapolation" ist also nur bedingt zu empfehlen. Hierfür erweist sich die Interpolation mit *rationalen* Funktionen als wesentlich geeigneter.

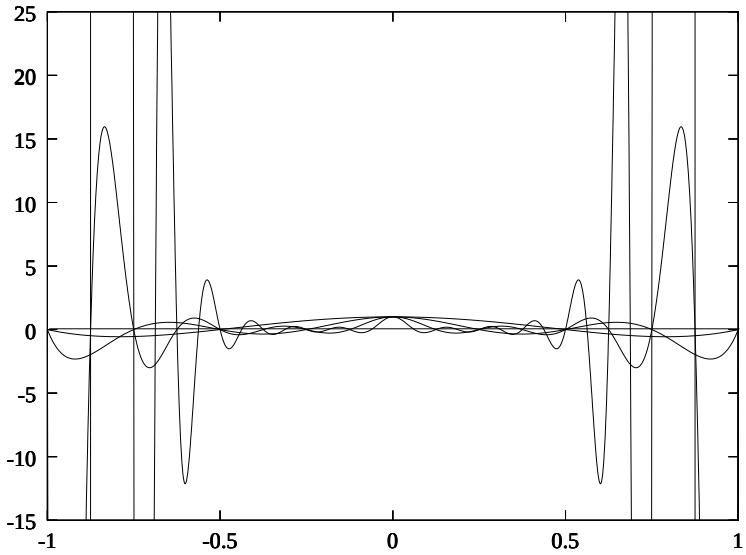


Abbildung 2.1: *Störungsverlauf bei wachsendem Polynomgrad*

2.1.3 Hermite-Interpolation

Die Aufgabenstellung der Lagrange-Interpolation läßt sich verallgemeinern zur sog. "Hermite[5]-Interpolation":.

Definition 2.5: *Die Hermitesche Interpolationsaufgabe lautet wie folgt:*

Gegeben: $x_i, \quad i = 0, \ldots, m \quad$ *(paarweise verschieden)*

$y_i^{(k)}, \quad i = 0, \ldots, m \quad k = 0, \ldots, \mu_i \quad (\mu_i \geq 0)\,.$

Gesucht: $p \in P_n, \quad n = m + \sum_{i=0}^{m} \mu_i : \quad p^{(k)}(x_i) = y_i^{(k)}\,.$

Die Punkte x_i *werden als* $(\mu_i + 1)$*-fache Stützstellen bezeichnet.*

Analog zu Satz 2.1 beweist man:

Satz 2.5 (Hermite-Interpolation): *Die Hermitesche Interpolationsaufgabe besitzt eine eindeutige Lösung.*

Sind die Knotenwerte $y_i^{(k)} = f^{(k)}(x_i)$ durch eine Funktion f gegeben, so gilt:

[5]Charles Hermite (1822-1901): französischer Mathematiker; Prof. an der École Polytechnique und der Sorbonne in Paris; Beiträge zur Zahlentheorie und zur Theorie elliptischer Funktionen.

Satz 2.6 (Interpolationsfehler): *Ist* $f \in C^{n+1}[a,b]$ *, so gibt es zu jedem* $x \in [a,b]$ *ein* $\xi_x \in (x_0,\ldots,x_m,x)$ *, so daß für die Lösung* $p \in P_n$ *der Hermiteschen Interpolationsaufgabe gilt:*

$$\begin{aligned} f(x) - p(x) &= f[x_0,\ldots,x_0,\ldots,x_m,\ldots,x_m,x] \prod_{i=0}^{m} (x-x_i)^{\mu_i+1} \\ &= \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{m} (x-x_i)^{\mu_i+1}. \end{aligned} \tag{2.1.13}$$

Beweis: Analog zu Satz 2.3 bzw. Satz 2.4. Q.E.D.

Aus Stetigkeitsgründen besitzt das Hermitesche Interpolationspolynom einer Funktion $f \in C^{n+1}[a,b]$ die Darstellung

$$\begin{aligned} p(x) = \sum_{i=0}^{m} \sum_{r=1}^{\mu_i+1} f[& \underbrace{x_0,\ldots,x_0}_{(\mu_0+1)\text{-mal}}, \ldots, \underbrace{x_{i-1},\ldots,x_{i-1}}_{(\mu_{i-1}+1)\text{-mal}}, \underbrace{x_i,\ldots,x_i}_{r\text{-mal}}] \times \\ & \times \prod_{j=0}^{i-1} (x-x_j)^{\mu_j+1} (x-x_i)^{r-1}. \end{aligned}$$

Die dividierten Differenzen $f[\ldots]$ sind durch ein zur Lagrangeschen Interpolation analoges Rekursionsschema bestimmt, wobei immer dann, wenn Differenzen wegen des Zusammenfallens von Stützpunkten nicht nach der Definitionsformel gebildet werden können, sinngemäß die Stützwerte *höherer Ordnung* einzusetzen sind; z.B.:

$$\begin{aligned} y[x_i,x_i] &= y_i^{(1)}, \qquad y[x_i,x_i,x_{i+1}] = \frac{y[x_i,x_{i+1}] - y_i^{(1)}}{x_{i+1}-x_i}, \\ y[x_i,x_i,x_i] &= \frac{1}{2} y_i^{(2)}, \qquad y[x_i,x_i,x_i,x_{i+1}] = \frac{y[x_i,x_i,x_{i+1}] - \frac{1}{2} y_i^{(2)}}{x_{i+1}-x_i}. \end{aligned}$$

Eine weitere Verallgemeinerung der Lagrange- und Hermite-Interpolation ist die sog. "Hermite-Birkhoff[6]-Interpolation", bei der Funktions- bzw. Ableitungswerte in verschiedenen Punkten beliebig gemischt vorgegeben werden, z.B.:

$$p \in P_3 : \qquad p(0) = 1\,, \quad p'(1) = 2\,, \quad p(2) = 0\,, \quad p'(3) = 3.$$

Die Frage nach der Lösbarkeit der allgemeinen Hermite-Birkhoffschen Interpolationsaufgabe ist noch nicht vollständig geklärt. Es können im Gegensatz zur Lagrangeschen und zur Hermiteschen Aufgabe alle Fälle von "eindeutig lösbar" über "unendlich mehrdeutig lösbar" bis "unlösbar" auftreten (Übungsaufgabe).

[6]George David Birkhoff (1884-1944): US-Amerikanischer Mathematiker; Professor an der Harvard University, Boston, USA; Beiträge zu sehr verschiedenen Gebieten der Mathematik: dynamische Systeme (Ergoden-Theorem), Gravitationstheorie und Mathematik der Ästhetik.

2.2 Extrapolation zum Limes

Eine wichtige Anwendung der Polynominterpolation ist die sog. "Richardsonsche[7] Extrapolation zum Limes". Ein numerischer Prozeß liefere für jeden Wert eines Parameters $h \in \mathbb{R}_+$ $(h \to 0)$ einen Wert $a(h)$. Gesucht ist die nicht direkt berechenbare Größe

$$a(0) = \lim_{h \to 0} a(h). \tag{2.2.14}$$

Zur Annäherung von $a(0)$ berechnet man $a(h_i)$ für gewisse Werte h_i, $i = 0, \ldots, n$, und nimmt den Wert $p_n(0)$ des zugehörigen Interpolationspolynoms zu $(h_i, a(h_i))$ als Schätzung für $a(0)$.

Beispiel 2.3: Wir betrachten die numerische Realisierung der l'Hospitalschen[8] Regel. Zur Berechnung von

$$a(0) := \lim_{x \to +0} \frac{\cos(x) - 1}{\sin(x)} \quad (= 0)$$

setzen wir

$$a(x) := \frac{(\cos(x) - 1)}{\sin(x)}$$

und interpolieren $a(x)$ an einigen Stützstellen h_i nahe bei 0:

$$\begin{aligned} h_0 &= \tfrac{1}{8}, & a(h_0) &= -6.258151 \cdot 10^{-2}, \\ h_1 &= \tfrac{1}{16}, & a(h_1) &= -3.126018 \cdot 10^{-2}, \\ h_2 &= \tfrac{1}{32}, & a(h_2) &= -1.562627 \cdot 10^{-2}. \end{aligned}$$

Das interpolierende Polynom ist in Lagrangescher Darstellung:

$$p_2(x) = a(h_0) \frac{(x - \frac{1}{16})(x - \frac{1}{32})}{(\frac{1}{8} - \frac{1}{16})(\frac{1}{8} - \frac{1}{32})} + a(h_1) \frac{(x - \frac{1}{8})(x - \frac{1}{32})}{(\frac{1}{16} - \frac{1}{8})(\frac{1}{16} - \frac{1}{32})} + a(h_2) \frac{(x - \frac{1}{8})(x - \frac{1}{16})}{(\frac{1}{32} - \frac{1}{8})(\frac{1}{32} - \frac{1}{16})},$$

und wir erhalten

$$a(0) \sim p_2(0) = -1.02 \cdot 10^{-5}.$$

Numerisch günstiger wäre die Berechnung von $p_2(0)$ mit Hilfe des Nevilleschen Algorithmus nach dem Schema

$$p_{i,i+k}(0) = p_{i,i+k-1}(0) + \frac{p_{i,i+k-1}(0) - p_{i+1,i+k}(0)}{x_{i+k}/x_i - 1}, \quad k = 1, 2.$$

[7] Lewis Fry Richardson (1881-1953): englischer Mathematiker und Physiker; wirkte an verschiedenen Institutionen in England und Schottland; typischer "angewandter Mathematiker"; leistete Pionierbeiträge zur Modellierung und Numerik in der Wettervorhersage.

[8] Guilliome F. A. Marquis de L'Hospital (1661-1704): Französischer Mathematiker; Schüler von Johann Bernoulli; veröffentlichte 1696 das erste Lehrbuch der Differentialrechnung, welches auch die nach ihm benannte Regel enthält.

i	x_i	$p_{i,i}(0) = a(h_i)$	$p_{i,i+1}(0)$	$p_{i,i+2}(0)$
0	$x_0 = \frac{1}{8}$	$-6.258151 \cdot 10^{-2}$	$6.115 \cdot 10^{-5}$	$\boxed{-1.02 \cdot 10^{-5}}$
1	$x_1 = \frac{1}{16}$	$-3.126018 \cdot 10^{-2}$	$7.64 \cdot 10^{-6}$	
2	$x_2 = \frac{1}{32}$	$-1.562627 \cdot 10^{-2}$		

Beispiel 2.4: Wir betrachten die numerische Differentiation. Für C^1-Funktionen f ist

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x),$$

und für $f \in C^2$ folgt durch Taylor-Entwicklung

$$\frac{f(x+h) - f(x)}{h} = f'(x) + \frac{h}{2} f''(\zeta_x)\,, \quad \zeta_x \in \overline{(x, x+h)}.$$

Im Falle $f \in C^3$ erhält man eine bessere Näherung zu $f'(x)$ durch den zentralen Differenzenquotienten

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{h^2}{6} f'''(\zeta_x)\,, \quad \zeta_x \in \overline{(x+h, x-h)}.$$

Ist f analytisch, so gilt sogar

$$a(h) := \frac{f(x+h) - f(x-h)}{2h} = f'(x) + \sum_{i=1}^{\infty} \frac{f^{(2i+1)}(x)}{(2i)!}\, h^{2i},$$

d.h.: $a(h)$ ist eine "gerade" Funktion in h. Zur Extrapolation von $a(h)$ verwendet man daher zweckmäßigerweise auch gerade Polynome, d. h. Polynome in h^2. Als Beispiel betrachten wir $f(x) = \sin(x)$ mit $f'(0) = \cos(0) = 1$ und

$$a(h) \equiv \frac{\sin(h) - \sin(-h)}{2h} = \frac{\sin(h)}{h}.$$

Auswertung von $a(h)$ in

$$h_0 = \tfrac{1}{8}\,, \quad h_1 = \tfrac{1}{16}\,, \quad h_2 = \tfrac{1}{32} \quad (\text{"heimliche" Knoten: } -\tfrac{1}{8}\,, -\tfrac{1}{16}\,, -\tfrac{1}{32}\,)$$
$$a(h_0) = 0.\underline{99}73979\,, \quad a(h_1) = 0.\underline{999}3491\,, \quad a(h_2) = 0.\underline{999}8372\,,$$

ergibt dann

$$p_2(h) = a(h_0)\, \frac{(h^2 - \frac{1}{16^2})(h^2 - \frac{1}{32^2})}{(\frac{1}{8^2} - \frac{1}{16^2})(\frac{1}{8^2} - \frac{1}{32^2})} + \ldots\,, \quad p_2(0) = 0.\underline{99999992}6\,.$$

Der folgende Satz liefert die theoretische Grundlage der Richardson-Extrapolation.

Satz 2.7 (Extrapolationsfehler): *Für die Funktion* $a(h)$, $h \in \mathbb{R}_+$, *sei bekannt, daß eine asymptotische Entwicklung der Form*

$$a(h) = a_0 + \sum_{j=1}^{n} a_j h^{jq} + a_{n+1}(h) h^{(n+1)q} \tag{2.2.15}$$

gilt, mit einem $q > 0$, *und gewissen Koeffizienten* a_j *und* $a_{n+1}(h) = a_{n+1} + o(1)$ *für* $h \to 0$. *Sei* $(h_k)_{k=0,1,2,\ldots}$ *eine monoton fallende Folge positiver Zahlen mit der Eigenschaft*

$$0 < \frac{h_{k+1}}{h_k} \leq \rho < 1\,. \tag{2.2.16}$$

Für das Interpolationspolynom $p_n^{(k)} \in P_n$ *(in* h^q*) durch* $(h_k^q, a(h_k)), \ldots, (h_{k+n}^q, a(h_{k+n}))$ *gilt dann:*

$$a(0) - p_n^{(k)}(0) = O(h_k^{(n+1)q}) \quad (k \to \infty). \tag{2.2.17}$$

Beweis: Wir setzen zur Abkürzung $z = h^q$ und $z_k = h_k^q$. Das Interpolationspolynom zu den Stützpunkten $(z_{k+i}; a(h_{k+i}))$, $i = 0, \ldots, n$, ist in Lagrangescher Darstellung:

$$p_n(z) = \sum_{i=0}^{n} a(h_{k+i}) L_{k+i}^{(n)}(z)\,, \quad L_{k+i}^{(n)}(z) = \prod_{\substack{l=0 \\ l \neq i}}^{n} \frac{z - z_{k+l}}{z_{k+i} - z_{k+l}}\,.$$

Aus der Fehlerdarstellung

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\zeta_x) \prod_{i=0}^{n} (x - x_{k+i})\,, \quad \zeta_x \in [0, h_0]\,,$$

für $f \equiv 1$ sowie $f(x) = x^r$ liest man ab, daß (Übungsaufgabe)

$$\sum_{i=0}^{n} z_{k+i}^r L_{k+i}^{(n)}(0) = \begin{cases} 1 & , \text{ für } r = 0\,, \\ 0 & , \text{ für } r = 1, \ldots, n\,, \\ (-1)^n \prod_{i=0}^{n} z_{k+i} & , \text{ für } r = n+1\,. \end{cases}$$

Damit erschließen wir:

$$\begin{aligned} p_n(0) &= \sum_{i=0}^{n} \left\{ a_0 + \sum_{j=1}^{n} a_j z_{k+i}^j + a_{n+1}(h_{k+i}) z_{k+i}^{n+1} \right\} L_{k+i}^{(n)}(0) \\ &= a_0 \sum_{i=0}^{n} L_{k+i}^{(n)}(0) + \sum_{j=1}^{n} a_j \left\{ \sum_{i=0}^{n} z_{k+i}^j L_{k+i}^{(n)}(0) \right\} + \\ &\quad a_{n+1} \sum_{i=0}^{n} z_{k+i}^{n+1} L_{k+i}^{(n)}(0) + \sum_{i=0}^{n} o(1) z_{k+i}^{n+1} L_{k+i}^{(n)}(0), \end{aligned}$$

$$p_n(0) = a_0 + a_{n+1}(-1)^n \prod_{i=0}^{n} h_{k+i}^q + o(h_k^{(n+1)q}).$$

Hier wurde (2.2.16) benutzt, um die von h_k unabhängige Abschätzung

$$\left|L_{k+i}^{(n)}(0)\right| = \prod_{l=0,l\neq i}^{n} \left|\frac{z_{k+l}}{z_{k+i} - z_{k+l}}\right| = \prod_{l=0,l\neq i}^{n} \left|\frac{1}{z_{k+i}/z_{k+l} - 1}\right| \leq \gamma(n,\rho) \tag{2.2.18}$$

zu garantieren. Wegen

$$\prod_{i=0}^{n} h_{k+i}^q = O(h_k^{(n+1)q})$$

ergibt sich (2.2.17) gleichmäßig für alle k. Q.E.D.

Üblicherweise wird ein Extrapolationsprozeß anhand des folgenden "Extrapolationstableaus" zur Berechnung der Werte $p_{i,i+k} = p_{i,i+k}(0)$ durchgeführt. Es sei daran erinnert, daß $p_{i,i+k}(h)$ das Polynom (in h^q) ist, welches die Punkte $(h_i^q, a(h_i)), \ldots, (h_{i+k}^q, a(h_{i+k}))$ interpoliert. Die Funktionswerte $p_{i,i+k}$ erhält man nach dem Neville-Algorithmus:

$$p_{i,i+k} = p_{i,i+k-1} + \frac{p_{i,i+k-1} - p_{i+1,i+k}}{x_{i+k}/x_i - 1}.$$

Der Konvention folgend setzen wir $a_{ik} \equiv p_{i-k,i}$:

	$a_{i0} = a(h_i)$						
h_0	a_{00}				Extrapolationstableau		
h_1	a_{10}	$\to$	a_{11}				
h_2	a_{20}	$\to$	a_{21}	$\to$	a_{22}		
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\ddots$	
h_i	a_{i0}		a_{i1}		a_{i2}	$\cdots$	a_{ii}

Die Tableaueinträge werden sukzessive nach der folgenden Rekursionsformel berechnet:

$$\begin{aligned} i = 0,1,2,\ldots: &\quad a_{i,0} = a(h_i)\,, \\ i = 1,2,3,\ldots,\ k = 1,2,3,\ldots,i: &\quad a_{ik} = a_{i,k-1} + \frac{a_{i,k-1} - a_{i-1,k-1}}{(h_{i-k}/h_i)^q - 1}. \end{aligned} \tag{2.2.19}$$

Nach Satz 2.7 gilt dann für festes k:

$$a(0) - a_{ik} = O(h_{i-k}^{(k+1)q}) \quad (i \to \infty), \tag{2.2.20}$$

vorausgesetzt die Schrittweitenfolge genügt der Bedingung (2.2.16). Gebräuchliche Folgen sind gegeben durch $h_i \equiv h_0/n_i$ mit

i) $n_i = 2^i$, *ii)* $n_i = 2,4,6,8,12,16,\ldots$, *iii)* $n_i = 1,2,3,\ldots$ (unzulässig).

2.2.1 Fehlerkontrolle

Zur praktischen Durchführung der Extrapolation gehört ein Kriterium, wann der Extrapolationsprozeß abzubrechen ist. Sei eine zu erzielende Fehlertoleranz TOL vorgegeben. Der Fehlerdarstellung (siehe den Beweis von Satz 2.7)

$$a_{ik} = a(0) + a_{k+1}(-1)^k \prod_{j=0}^{k} h_{i-k+j}^q + o(h_{i-k}^{(k+1)q})$$

entnehmen wir, daß für festes k und genügend großes i der Fehler $a_{ik} - a(0)$ monoton gegen Null konvergiert (falls $a_{k+1} \neq 0$). Mit den neu definierten Größen

$$b_{ik} \equiv 2a_{i+1,k} - a_{ik}$$

gilt dann im Falle $q \geq 1$:

$$\begin{aligned} b_{ik} - a(0) &= 2\{a_{i+1,k} - a(0)\} - \{a_{ik} - a(0)\} \\ &= 2a_{k+1}(-1)^k \prod_{j=0}^{k} h_{i+1-k+j}^q + o(h_{i+1-k}^{(k+1)q}) - a_{k+1}(-1)^k \prod_{j=0}^{k} h_{i-k+j}^q + o(h_{i-k}^{(k+1)q}) \\ &= \prod_{j=0}^{k} h_{i-k+j}^q \left\{-a_{k+1}(-1)^k + 2\,(\frac{h_{i+1}}{h_{i-k}})^q\, a_{k+1}(-1)^k\right\} + o(h_{i-k}^{(k+1)q}). \end{aligned}$$

Wegen $h_{i+1}^q / h_{i-k}^q \ll 1$ gilt also in erster Näherung

$$b_{ik} - a(0) \doteq -(-1)^k\, a_{k+1} \prod_{j=0}^{k} h_{i-k+j}^q,$$

und folglich

$$a_{ik} - a(0) \doteq -\,(b_{ik} - a(0))\,, \tag{2.2.21}$$

für festes k und genügend großes i. Wegen der monotonen Konvergenz $a_{ik} - a(0) \to 0$ $(i \to \infty)$ gilt also asymptotisch entweder $a_{ik} \leq a(0) \leq b_{ik}$ oder $a_{ik} \geq a(0) \geq b_{ik}$, und beide Seiten konvergieren monoton gegen $a(0)$ für $i \to \infty$. Dieses Verhalten der Folgen $(a_{ik})_{i\in\mathbb{N}}$ und $(b_{ik})_{i\in\mathbb{N}}$ (für festes k) kann zur Konstruktion eines Abbruchkriteriums herangezogen werden:

$$|a_{ik} - b_{ik}| < \text{TOL} \quad \Rightarrow \quad \text{STOP}\,. \tag{2.2.22}$$

Bemerkung 2.2: Für die Praxis lohnt es sich, den Extrapolationsprozeß vollständig durchzuführen, d.h. die Diagonalelemente a_{ii} des Tableaus zu berechnen. In der Tat kann man zeigen, daß die "Diagonalfolge" $(a_{ii})_{i\in\mathbb{N}}$ schneller gegen $a(0)$ konvergiert als jede der Spaltenfolgen $(a_{ik})_{i\in\mathbb{N}}$, $k \geq 0$.

2.3 Spline-Interpolation

Lagrangesche Interpolationspolynome eignen sich nicht besonders gut zur Approximation von (nicht glatten) Funktionen, da sie bei Vermehrung der Stützstellenzahl dazu neigen, zwischen den Stützstellen immer größere Werte anzunehmen. Dies ist eine Folge ihrer "Steifheit" bedingt durch die Forderung von C^∞-Übergängen in den Knoten. Zur Reduzierung dieser Steifheit setzt man die interpolierende Funktion φ nur als "stückweise polynomial" bzgl. der Zerlegung $a = x_0 < x_1 < \ldots < x_n = b$ an. In den Knoten x_i werden dann geeignete Differenzierbarkeitseigenschaften (z.B. $\varphi \in C^2[a,b]$) gefordert.

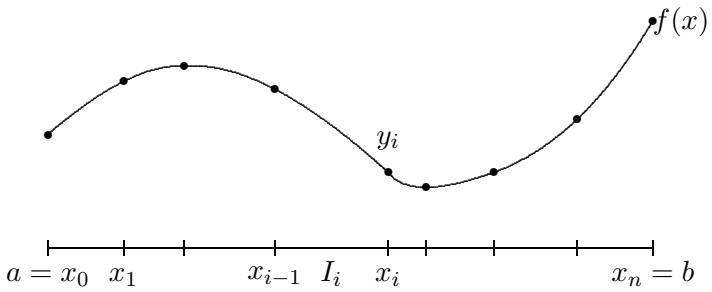


Abbildung 2.2: *Spline-Approximation*

Die Länge des Teilintervalls $I_i = [x_{i-1}, x_i]$ ist $h_i = x_i - x_{i-1}$, und die Feinheit der ganzen Intervallunterteilung wird durch die Größe $h := \max_{i=1,\ldots,n} h_i$ beschrieben. Auf einer solchen Intervallzerlegung werden Vektorräume von stückweise polynomialen Funktionen betrachtet

$$S_h^{(k,r)}[a,b] = \left\{ p \in C^r[a,b],\, p_{|I_i} \in P_k(I_i) \right\}$$

für $k, r \in \{0, 1, 2, \ldots\}$. Zu einem Satz von Stützwerten in Punkten aus dem Intervall $[a,b]$, die etwa wieder von einer zu interpolierenden Funktion f genommen werden, wird dann eine "Interpolierende" $p \in S_h^{(k,r)}[a,b]$ mit Hilfe von geeigneten Interpolationsbedingungen bestimmt. Wir betrachten im folgenden einige einfache Beispiele.

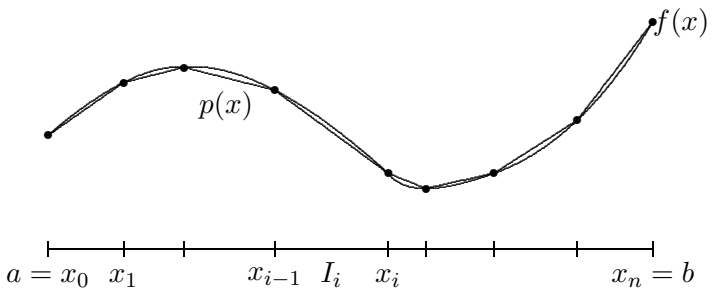


Abbildung 2.3: *Stückweise lineare Interpolation*

Beispiel 2.5: Die stückweise lineare Lagrange-Interpolation (Fall $k=1, r=0$) approximiert eine gegebene Funktion f auf $[a,b]$ durch einen Polygonzug in den Stützstellen x_i, $i=0,\ldots,n$:

$$p \in S_h^{(1,0)}[a,b] = \left\{ p \in C[a,b],\, p_{|I_i} \text{ linear} \right\}, \quad p(x_i) = f(x_i), \quad i = 0,\ldots,n.$$

Anwendung der Fehlerabschätzung für die Lagrange-Interpolation separat auf jedem der Teilintervalle I_i ergibt die globale Fehlerabschätzung

$$\max_{x\in[a,b]} |f(x) - p(x)| \;\le\; \tfrac{1}{2} h^2 \max_{x\in[a,b]} |f''(x)| .$$

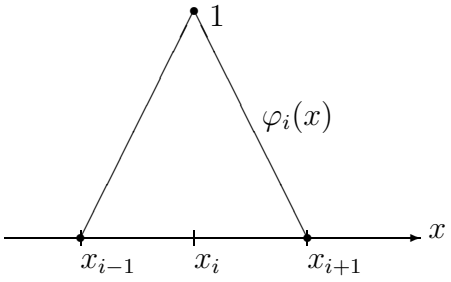


Abbildung 2.4: *"Lineare" Knotenbasisfunktion*

Für die Konstruktion der Interpolierenden verwendet man die sog. Knotenbasis von $S_h^{(1,0)}[a,b]$ bestehend aus den "Dachfunktionen" $\varphi_i \in S_h^{(1,0)}[a,b]$, $i=0,\ldots,n$, die durch die Bedingung

$$\varphi_i(x_j) = \delta_{ij}$$

eindeutig bestimmt sind. Daß dies tatsächlich eine Basis ergibt, ist durch elementare Argumente einzusehen und sei als Übungsaufgabe gestellt. Die Interpolierende p von f läßt sich dann in der Form

$$p(x) = \sum_{i=0}^{n} f(x_i)\varphi_i(x).$$

darstellen. Diese Konstruktion ist analog zur Lagrangeschen Darstellung des Lagrange-Interpolationspolynoms. Da die Dachfunktionen φ_i nur in den direkt an den jeweiligen Aufpunkt x_i angrenzenden Teilintervallen von Null verschieden sind, nennt man diese Basis "lokal". Höhere globale Glattheit als Stetigkeit ist sinnvoll mit stückweise linearen Interpolierenden offensichtlich nicht erzielbar. Dazu benötigt man Polynomansätze höherer Ordnung.

Beispiel 2.6: Wir betrachten noch die stückweise Interpolation mit kubischen Polynomen ($k=3$, $r=0$ oder $r=1$): $S_h^{(3,0)}[a,b]$ und $S_h^{(3,1)}[a,b]$. Zur Erzielung globaler Stetigkeit

$(r=0)$ verwendet man als Interpolationsbedingungen

$$p(x_i) = f(x_i)\,, \quad p(x_{ij}) = f(x_{ij}),$$

mit jeweils zwei zusätzlichen Interpolationspunkten $x_{ij} \in I_i$, $i=1,\ldots,n$, $j=1,2$. Durch stückweise kubische Lagrange-Interpolation ist dadurch eindeutig eine global stetige Interpolierende $p \in S_h^{(3,0)}[a,b]$ festgelegt. Analog erhält man durch stückweise kubische Hermite-Interpolation aus den Interpolationsbedingungen

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i), \quad i = 0,\ldots,n,$$

eine global stetig differenzierbare Interpolierende $p \in S_h^{(3,1)}[a,b]$. In beiden Fällen folgt aus den jeweiligen Fehlerabschätzungen die globale Fehlerabschätzung

$$\max_{x\in[a,b]} |f(x)-p(x)| \le \tfrac{1}{4!}h^4 \max_{x\in[a,b]} |f^{(4)}(x)|.$$

Analog zum Fall $k=1$ lassen sich auch hier wieder "lokale" Knotenbasen der Ansatzräume $S_h^{(3,0)}$ und $S_h^{(3,1)}$ angeben. Derartige rein lokale Interpolationsprozesse haben Anwendungen etwa bei der numerischen Lösung von gewöhnlichen und partiellen Differentialgleichungen.

Die Forderung nach höherer globaler Glattheit, etwa Interpolation in $S_h^{(3,2)}$, führt auf die sog. "Spline-Interpolation", welche von großer praktischer Bedeutung z.B. bei der glatten Darstellung von Flächen in der Computer-Graphik ist. Der Begriff "Spline" stammt aus dem Englischen und bedeutet soviel wie "biegsames Kurvenlineal" ("Biegestab"):

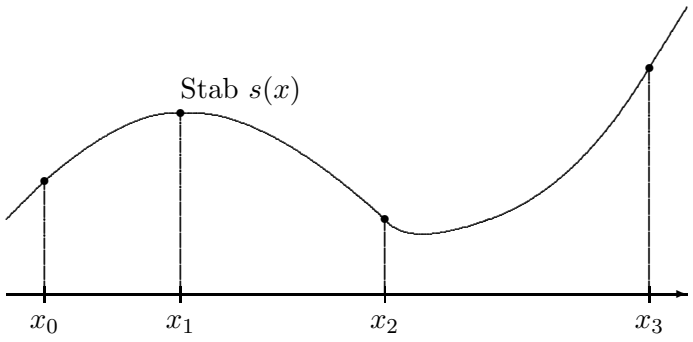


Abbildung 2.5: *Spline-Funktion*

Nach einem Grundgesetz der Mechanik (Prinzip vom Minimum der potentiellen Gesamtenergie) stellt sich der Biegestab so ein, daß die Gesamtkrümmung minimiert wird, d. h. in linearisierter Näherung:

$$\int_{x_0}^{x_n} s''(x)^2\,dx = \min\,! \tag{2.3.23}$$

bzgl. aller möglichen interpolierenden (hinreichend glatten) Funktionen. Außerhalb des Interpolationsintervalls $[x_0, x_n]$ kann $s(x)$ als linear angenommen werden, d. h.

$$s''(x_0) = s''(x_n) = 0. \tag{2.3.24}$$

Diese sog. "natürlichen" Randbedingungen stellen sich also automatisch ein, wenn der Stab nicht willkürlich zu einem anderen Verhalten gezwungen wird ("erzwungene" Randbedingungen: z.B. $s'(x_0) = s'(x_n) = 0$). Wir betrachten hier der Einfachheit halber nur die am häufigsten verwendeten "kubischen" Spline-Funktionen; sie ergeben sich auf natürliche Weise aus dem obigen Biegestabmodell aufgrund der Forderung (2.3.23).

Definition 2.6: *Eine Funktion $s_n : [a, b] \to \mathbb{R}$ heißt "kubischer Spline" bzgl. einer Zerlegung $a = x_0 < x_1 < \ldots < x_n = b$, wenn gilt:*

(i) $s_n \in C^2[a, b]$;

(ii) $s_{n|[x_{i-1},x_i]} \in P_3\,, \quad i = 1, \ldots, n\,.$

Gilt zusätzlich

(iii) $s_n''(a) = s_n''(b) = 0\,,$

so wird der kubische Spline "natürlich" genannt.

Wir fragen nun nach der Existenz des interpolierenden kubischen Splines zu vorgegebenen Knotenwerten

$$s_n(x_i) = y_i\,, \quad i = 0, \ldots, n.$$

Satz 2.8 (Spline-Interpolation): *Der interpolierende kubische Spline existiert und ist eindeutig bestimmt durch zusätzliche Vorgabe von $s_n''(a)$ und $s_n''(b)$.*

Beweis: Jeder kubische Spline (bzgl. der Zerlegung $a = x_0 < \ldots < x_n = b$) hat die Form $s(x)_{|[x_{i-1},x_i]} = p_i(x)\,, \quad i = 1, \ldots, n\,,$ mit Polynomen $p_i \in P_3$. Die jeweiligen 4 Koeffizienten dieser n Polynome ergeben $4n$ freie Parameter. Zu ihrer Bestimmung stehen folgende lineare Beziehungen zur Verfügung:

$s(x_i) = y_i\,, \quad i = 0, \ldots, n$	:	$2n$	Gleichungen
$s' \in C[a, b]$	:	$n - 1$	"
$s'' \in C[a, b]$	:	$n - 1$	"
Zusatzbedingungen	:	2	"
$\sum$	:	$4n$	"

Zum Nachweis der Existenz einer Lösung dieses (quadratischen) Gleichungssystems genügt es wieder zu verifizieren, daß das zugehörige homogene System nur die triviale Lösung hat. Dazu führen wir die folgende Funktionenmenge ein:

$$N \equiv \{w \in C^2[a,b] \mid w(x_i) = 0,\ i = 0, \ldots, n\}.$$

Seien nun $s_n^{(1)}, s_n^{(2)}$ zwei interpolierende Splines. Für die Differenz $s \equiv s_n^{(1)} - s_n^{(2)} \in N$ gilt dann mit beliebigem $w \in N$:

$$\begin{aligned}\int_a^b s''(x)w''(x)\,dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s''(x)w''(x)\,dx \\ &= \sum_{i=0}^{n-1} \left\{ s''w' \Big|_{x_i}^{x_{i+1}} - s^{(iii)}w\Big|_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} s^{(iv)}(x)w(x)\,dx \right\} \\ &= \sum_{i=0}^{n-1} s''w' \Big|_{x_i}^{x_{i+1}} = s''(b)w'(b) - s''(a)w'(a) = 0.\end{aligned}$$

Speziell für $w \equiv s \in N$ ist also

$$\int_a^b |s''(x)|^2\,dx = 0,$$

d. h.: s ist linear. Wegen $s(a) = s(b) = 0$ folgt $s \equiv 0$. Q.E.D.

Die obige Orthogonalitätsbeziehung hat die interessante Konsequenz, daß sich der interpolierende Spline durch eine besonders geringe Oszillation auszeichnet.

Satz 2.9: *Für den interpolierenden, natürlichen, kubischen Spline s_n gilt*

$$\int_a^b |s_n''(x)|^2\,dx \le \int_a^b |f''(x)|^2\,dx \tag{2.3.25}$$

bzgl. aller anderen Funktionen $f \in C^2[a,b]$ mit $f(x_i) = y_i,\ i = 0, \ldots, n$.

Beweis: Sei N wieder definiert wie im Beweis von Satz 2.8. Jede interpolierende Funktion $f \in C^2[a,b]$ kann in der Form $f = s_n + w$ mit einem $w \in N$ geschrieben werden. Wir haben (siehe oben)

$$\int_a^b s_n''(x)w''(x)\,dx = 0 \qquad \forall\, w \in N.$$

Wegen der Identität

$$\int_a^b |f''(x)|^2\,dx = \int_a^b |s_n''(x) + w''(x)|^2\,dx$$
$$= \int_a^b |s_n''(x)|^2\,dx + 2\underbrace{\int_a^b s_n''(x)w''(x)\,dx}_{=0} + \underbrace{\int_a^b |w''(x)|^2\,dx}_{\geq 0}$$

folgt die Behauptung. Q.E.D.

Die Aussage von Satz 2.9 läßt sich dahingehend umkehren, daß jede interpolierende Funktion $s \in C^2[a,b]$, $s(x_i) = y_i$, mit der Eigenschaft (2.3.25) notwendig ein natürlicher kubischer Spline ist.

Zur expliziten Berechnung des interpolierenden Splines s_n schreiben wir seine Bestandteile $s_{n|[x_{i-1},x_i]} = p_i \in P_3$ in der Form

$$p_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3, \quad i = 1, \ldots, n,$$

und bestimmen die $4n$ Koeffizienten $a_0^{(i)}, \ldots, a_3^{(i)}$. Dies wird im folgenden für den interpolierenden "natürlichen" Spline durchgeführt:

– Die Interpolationsbedingung $p_i(x_i) = y_i,\ p_i(x_{i-1}) = y_{i-1}$ impliziert:

(1) $\quad a_0^{(i)} = y_i\,, \quad i = 1, \ldots, n\,,$

und mit $h_i = x_i - x_{i-1}$:

(2) $\quad y_{i-1} - y_i = -a_1^{(i)}h_i + a_2^{(i)}h_i^2 - a_3^{(i)}h_i^3\,, \quad i = 1, \ldots, n\,.$

– Die Randbedingung $p_1''(x_0) = p_n''(x_n) = 0$ impliziert:

(3) $\quad a_2^{(1)} - 3a_3^{(1)}h_1 = 0\,, \quad a_2^{(n)} = 0\,.$

– Die Stetigkeit der 1. Ableitung $p_i'(x_i) = p_{i+1}'(x_i)$ impliziert:

(4) $\quad a_1^{(i)} = a_1^{(i+1)} - 2a_2^{(i+1)}h_{i+1} + 3a_3^{(i+1)}h_{i+1}^2\,, \quad i = 1, \ldots, n-1\,.$

– Die Stetigkeit der 2. Ableitung $p_i''(x_i) = p_{i+1}''(x_i)$ impliziert:

(5) $\quad a_2^{(i)} = a_2^{(i+1)} - 3a_3^{(i+1)}h_{i+1}\,, \quad i = 1, \ldots, n-1\,.$

Damit haben wir $4n$ Gleichungen (1)-(5) für die $a_k^{(i)}$ gefunden. Zunächst werden nun die $a_1^{(i)}$ und $a_3^{(i)}$ durch die $a_2^{(i)}$ ausgedrückt. Zur Vereinfachung wird $a_2^{(0)} := 0$ und $a_2^{(n+1)} := 0$ gesetzt.

Die Gleichungen (3) und (5) ergeben $(i \to i-1)$:

$$(6) \qquad a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i}\,, \quad i = 1, \ldots, n\,.$$

(2) und (6) ergeben:

$$(7) \qquad \begin{aligned} a_1^{(i)} &= \frac{y_i - y_{i-1}}{h_i} + a_2^{(i)} h_i - a_3^{(i)} h_i^2 \\ &= \frac{y_i - y_{i-1}}{h_i} + h_i \{a_2^{(i)} - \frac{a_2^{(i)} - a_2^{(i-1)}}{3}\} \\ &= \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\{2a_2^{(i)} + a_2^{(i-1)}\}\,, \quad i = 1, \ldots, n\,. \end{aligned}$$

(4), (7) und (6) ergeben:

$$\begin{aligned} &\underbrace{\frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\{2a_2^{(i)} + a_2^{(i-1)}\}}_{= a_1^{(i)}} = \\ &= \underbrace{\frac{y_{i+1} - y_i}{h_{i+1}} + \frac{h_{i+1}}{3}\{2a_2^{(i+1)} + a_2^{(i)}\}}_{= a_1^{(i+1)}} - 2a_2^{(i+1)} h_{i+1} + 3a_3^{(i+1)} h_{i+1}^2 \\ &= \frac{y_{i+1} - y_i}{h_{i+1}} + \frac{h_{i+1}}{3}\{2a_2^{(i+1)} + a_2^{(i)}\} - h_{i+1}\{a_2^{(i+1)} + a_2^{(i)}\}, \quad i = 1, \ldots, n-1\,. \end{aligned}$$

Dies wird für $i = 1, \ldots, n-1$ umgeschrieben zu

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1}) a_2^{(i)} + h_{i+1} a_2^{(i+1)} = 3\left\{\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right\}.$$

Damit haben wir für den $(n-1)$-Vektor $(a_2^{(1)}, \ldots, a_2^{(n-1)})^T$ ein $(n-1) \times (n-1)$-Gleichungssystem aufgestellt; die Koeffizientenmatrix

$$A = \begin{bmatrix} 2(h_1 + h_2) & h_2 & & & 0 \\ h_2 & 2(h_2 + h_3) & \ddots & & \\ & h_3 & \ldots & & \\ & & \ddots & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & & & h_{n-1} & 2(h_{n-1} + h_n) \end{bmatrix}$$

ist symmetrisch, d.h.: $a_{ij} = a_{ji}$, und strikt diagonaldominant.

$$\sum_{\substack{j=1 \\ j \neq i}}^{n-1} |a_{ij}| < |a_{ii}|. \tag{2.3.26}$$

Hieraus folgt nach einem bekannten Satz der Linearen Algebra ebenfalls die Regularität von A. Durch Lösung dieses Gleichungssystems erhält man zunächst die Koeffizienten $a_2^{(1)}, \ldots, a_2^{(n-1)}$; $a_2^{(n)} = 0$ aufgrund der Randbedingungen. Einsetzen der Werte in (7) und (6) liefert dann die anderen Koeffizienten $a_1^{(1)}, \ldots, a_1^{(n)}$ sowie $a_3^{(1)}, \ldots, a_3^{(n)}$. Zur Lösung des tridiagonalen (symmetrischen) Gleichungssystems kann eine spezielle Variante des Gaußschen Eliminationsverfahrens verwendet werden.

Interpolierende Spline-Funktionen besitzen wesentlich bessere Approximationseigenschaften für

$$h := \max_{i=0,\ldots,n-1} |x_{i+1} - x_i| \to 0$$

als die Lagrangeschen Interpolationspolynome. Allgemein konvergiert

$$\max_{a \le x \le b} |f(x) - s_n(x)| \to 0 \quad (h \to 0) \tag{2.3.27}$$

sogar für absolutstetiges f mit $\int_a^b |f'(x)|^2 \, dx < \infty$.

Satz 2.10 (Approximationsfehler): *Sei $f \in C^4[a,b]$. Erfüllt der interpolierende kubische Spline $s_n''(a) = f''(a)$ und $s_n''(b) = f''(b)$, so gilt*

$$\max_{a \le x \le b} |f(x) - s_n(x)| \le \tfrac{1}{2} h^4 \max_{a \le x \le b} |f^{(4)}(x)|. \tag{2.3.28}$$

Beweis: Siehe Werner/Schaback II. Q.E.D.

Neben den guten Approximationseigenschaften weisen Splinefunktionen auch eine wesentlich bessere Stabilität gegenüber kleinen Störungen in den Interpolationsdaten y_i auf als Lagrange-Polynome; lokale Störungen klingen nach rechts und links schnell ab.

2.4 Trigonometrische Interpolation

In vielen Anwendungsbereichen treten "periodische" Funktionen, d. h. Funktionen mit der Eigenschaft $f(x+\omega) = f(x),\ x \in \mathbb{R}$, auf, mit der sog. "Periode" $\omega > 0$. Hier bietet sich die Interpolation mit "trigonometrischen Summen" an:

$$t_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{m} \Big\{ a_k \cos\Big(\frac{kx\,2\pi}{\omega}\Big) + b_k \sin\Big(\frac{kx\,2\pi}{\omega}\Big)\Big\}, \quad n := 2m,$$

welche ebenfalls ω-periodisch sind. O.B.d.A. können wir im folgenden $\omega = 2\pi$ annehmen. Das Interpolationsintervall ist dann $[0, 2\pi]$, und die Stützstellen werden äquidistant gewählt zu

$$x_k = k\frac{2\pi}{n+1}, \quad k = 0, \ldots, n.$$

Beim Arbeiten mit trigonometrischen Summen erweist sich die "komplexe" Schreibweise als vorteilhaft.

Satz 2.11 (Trigonometrische Interpolation): *Zu gegebenen Zahlen $y_0, \ldots, y_n \in \mathbb{C}$ gibt es genau eine Funktion der Gestalt ($i = \sqrt{-1}$)*

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}, \tag{2.4.29}$$

welche den Interpolationsbedingungen $t_n^*(x_j) = y_j$ $(j = 0, \ldots, n)$ genügt. Die Koeffizienten sind bestimmt durch

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}, \quad k = 0, \ldots, n\,. \tag{2.4.30}$$

Beweis: Mit den Abkürzungen

$$w = e^{ix}, \quad w_k = e^{ix_k} = e^{ik2\pi/(n+1)}, \quad k = 0, \pm 1, \cdots \pm n,$$

wird

$$t_n^*(x) = p_n(w) = \sum_{k=0}^{n} c_k w^k, \quad p_n(\cdot) \in P_n.$$

Offenbar gilt $w_k^{n+1} = e^{ik2\pi} = 1$, d. h.: Die w_k sind sog. $(n+1)$-te "Einheitswurzeln". Ferner ist

$$w_k^j = e^{j(ik2\pi)/(n+1)} = e^{k(ij2\pi)/(n+1)} = w_j^k.$$

Die trigonometrische Interpolationsbedingung

$$t_n^*(x_k) = y_k, \quad k = 0, \ldots, n,$$

ist damit äquivalent zur polynomialen Interpolationsbedingung (im Komplexen)

$$p_n(w_k) = y_k\,, \quad k = 0, \ldots, n,$$

Diese wiederum ist nach Satz 2.1 durch ein eindeutig bestimmtes Polynom $p_n \in P_n$ erfüllbar. Zur Berechnung der zugehörigen Koeffizienten c_k schreiben wir

$$\sum_{j=0}^{n} y_j w_k^{-j} = \sum_{j=0}^{n} t_n^*(w_j) w_k^{-j} = \sum_{j=0}^{n} \Big(\sum_{l=0}^{n} c_l w_j^l \Big) w_k^{-j} = \sum_{l=0}^{n} c_l \Big(\sum_{j=0}^{n} w_j^{l-k} \Big)\,.$$

Die w_k sind Wurzeln von $w^{n+1} - 1 = (w-1)(w^n + w^{n-1} + \ldots + 1) = 0$. Wegen $w_k \neq 1$ für $k = \pm 1, \ldots, \pm n$, muß also $\sum_{j=0}^{n} w_k^j = 0$ sein. Dies ergibt

$$\sum_{j=0}^{n} w_j^{l-k} = \sum_{j=0}^{n} w_{l-k}^j = \begin{cases} n+1, & l = k, \\ 0, & l \neq k. \end{cases}$$

Also ist

$$\sum_{j=0}^{n} y_j w_k^{-j} = c_k (n+1),$$

bzw.

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}.$$

Dies vervollständigt den Beweis. Q.E.D.

Für Satz 2.11 ist es unwesentlich, ob n gerade oder ungerade ist. Im folgenden müssen diese beiden Fälle aber unterschieden werden. Wir wollen zeigen, wie mit Hilfe der Aussagen von Satz 2.11 die gestellte trigonometrische Interpolationsaufgabe gelöst werden kann. Dies führt zur sog. "diskreten Fourier[9]-Analysis".

Satz 2.12 (Diskrete Fourier-Analyse): *Für $n \in \mathbb{N}_0$ gibt es zu gegebenen reellen Zahlen $y_0, \ldots, y_n$ genau ein trigonometrisches Polynom der Form*

$$t_n(x) = \frac{1}{2} a_0 + \sum_{k=1}^{m} \{ a_k \cos(kx) + b_k \sin(kx) \} + \frac{\theta}{2}\, a_{m+1} \cos((m+1)x)\,,$$

mit $t_n(x_j) = y_j\,,\ j = 0, \ldots, n\,,$ wobei

$$\begin{aligned} &\theta = 0\,, \quad m = \tfrac{1}{2} n\,, && \textit{falls } n \textit{ gerade}, \\ &\theta = 1\,, \quad m = \tfrac{1}{2}(n-1)\,, && \textit{falls } n \textit{ ungerade}. \end{aligned}$$

[9] Jean-Baptiste Baron de Fourier (1768-1830): französischer Mathematiker und Physiker; Mitglied der Pariser Akademie lehrte an der École Polytechniqe; begleitete Napoleon auf seinem Feldzug nach Ägypten; zählt zu den bedeutendsten Mathematikern des 19. Jahrhunderts; fand bei seinen Arbeiten zur Theorie der Wärmeleitung die Darstellbarkeit periodischer Funktionen durch triginometrische Reihen.

Die Koeffizienten sind bestimmt durch

$$a_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \cos(jx_k), \quad b_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \sin(jx_k).$$

Beweis: (i) Sei t_n^* das Interpolationspolynom nach Satz 2.11:

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}, \quad c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}, \quad k = 0, \ldots, n.$$

Wegen der 2π-Periodizität von e^{-ix} gilt

$$e^{-ijx_{n+1-k}} = e^{-ij(n+1-k)2\pi/(n+1)} = e^{-ij2\pi+ijx_k} = e^{ijx_k},$$

und folglich für $k = 1, \ldots, m$:

$$c_{n+1-k} = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_{n-1-k}} = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{ijx_k} =: c_{-k}.$$

Mit dieser Bezeichnung setzen wir

$$a_k := c_k + c_{-k}, \quad b_k := i\,(c_k - c_{-k}), \quad k = 1, \ldots, m,$$

und $a_{m+1} := 2c_{m+1}$ im Falle $n = 2m+1$ ungerade, und definieren das trigonometrische Polynom

$$t_n(x) := \frac{1}{2}a_0 + \sum_{k=1}^{m} \big\{a_k \cos(kx) + b_k \sin(kx)\big\} + \frac{\theta}{2} a_{m+1} \cos((m+1)x).$$

(ii) Wir wollen zeigen, daß $t_n^*(x_j) = t_n(x_j) = y_j$, $j = 0, \ldots, n$, ist. Zunächst gilt

$$\begin{aligned} t_n(x_j) &= c_0 + \sum_{k=1}^{m} \big\{(c_k + c_{-k}) \cos(kx_j) + i\,(c_k - c_{-k}) \sin(kx_j)\big\} + \\ &\quad + \theta c_{m+1} \cos((m+1)x_j) \\ &= c_0 + \sum_{k=1}^{m} c_k \big\{\cos(kx_j) + i \sin(kx_j)\big\} + \sum_{k=1}^{m} c_{-k} \{\cos(kx_j) - \\ &\quad - i \sin(kx_j)\} + \theta\, c_{m+1} \{\cos((m+1)x_j) + i \sin((m+1)x_j)\big\}. \end{aligned}$$

Da $\sin((m+1)x_j) = 0$ (wegen $(m+1)x_j = j\pi$ für $n = 2m+1$ ungerade), folgt bei Beachtung von $e^{iz} = \cos(z) + i \sin(z)$

$$t_n(x_j) = c_0 + \sum_{k=1}^{m} c_k e^{ikx_j} + \sum_{k=1}^{m} c_{-k} e^{-ikx_j} + \theta c_{m+1} e^{i(m+1)x_j}.$$

Mit Hilfe von $c_{-k} = c_{n+1-k}$ und $e^{-ikx_j} = e^{ikj2\pi/(n+1)} = e^{i(n+1-k)j2\pi/(n+1)} = e^{i(n+1-k)x_j}$ ergibt sich wie gewünscht

$$t_n(x_j) = \sum_{k=0}^{n} c_k e^{ikx_j} = y_j.$$

Das trigonometrische Polynom $t_h(\cdot)$ erfüllt also die Interpolationsbedingungen.

(iii) Als nächstes betrachten wir die Koeffizienten a_k und b_k. Unter Verwendung der Beziehungen

$$\sin(z) = \frac{1}{2i}\,(e^{iz} - e^{-iz})\,, \quad \cos(z) = \frac{1}{2}\,(e^{iz} + e^{-iz}),$$

gilt

$$a_k = c_k + c_{-k} = \frac{1}{n+1}\sum_{j=0}^{n} y_j\,\{e^{-ijx_k} + e^{ijx_k}\} = \frac{2}{n+1}\sum_{j=0}^{n} y_j\,\cos(jx_k),$$

$$b_k = i\,(c_k - c_{-k}) = \frac{1}{n+1}\sum_{j=0}^{n} y_j i\,\{e^{-ijx_k} - e^{ijx_k}\} = \frac{2}{n+1}\sum_{j=0}^{n} y_j\,\sin(jx_k).$$

Dies zeigt die gewünschte Darstellung der Koeffizienten und auch, daß sie reell sind.

(iv) Es bleibt, die Eindeutigkeit des interpolierenden trigonometrischen Polynoms zu zeigen. Die $n+1$ Bedingungen $t_n(x_j) = y_j,\ j = 0, \ldots, n$, lassen sich als lineares Gleichungssystem für die $n+1$ unbekannten Koeffizienten a_k, b_k auffassen. Da dieses System nach dem eben Gezeigten für alle rechten Seiten $y_0, \ldots, y_n$ lösbar ist, sind die Lösungen auch eindeutig. Q.E.D.

Bei der trigonometrischen Interpolation von (stetigen) Funktionen $f : \mathbb{R} \to \mathbb{R}$ sind in Satz 2.12 die Werte y_j durch $f(x_j)$ zu ersetzen. Ist f allerdings nicht 2π-periodisch, wird es zunächst einmal zu einer 2π-periodischen Funktion $\tilde{f}$ gemacht (s. Abb. 2.6):

$$\tilde{f}(x) := \begin{cases} f(x), & x \in (0, 2\pi), \\ \frac{1}{2}\,\{f(0) + f(2\pi)\}, & x = 0, \\ 2\pi\text{-periodisch auf } \mathbb{R} \text{ fortgesetzt.} \end{cases}$$

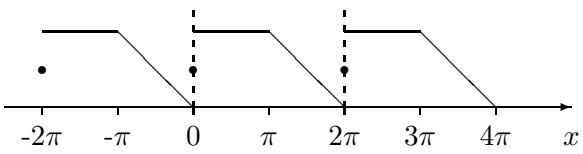


Abbildung 2.6: *Periodische Fortsetzung*

Besitzt f im Intervall $[0, 2\pi]$ eine Unstetigkeitsstelle ζ, so definiert man noch

$$\tilde{f}(\zeta) := \tfrac{1}{2}\,\{f(\zeta_+) + f(\zeta_-)\}.$$

Dies ist dadurch motiviert, daß bekanntlich die Fourier-Reihe einer stückweise stetigen Funktion in den Unstetigkeitsstellen gerade gegen diesen Mittelwert konvergiert.

In Anwendungen tritt häufig der Fall auf, daß eine Funktion f nur auf dem Intervall $[0,\pi]$ gegeben ist. Ihre 2π-periodische Fortsetzung auf ganz $\mathbb{R}$ würde dann i. Allg. bei $x=\pi$ eine Unstetigkeit in f oder in f' besitzen, welche die Approximationsgüte des trigonometrischen Interpolationspolynoms auf $[0,2\pi]$ reduziert. Durch geeignete Wahl der Fortsetzung kann dieser Effekt häufig gemildert werden. Ist f auf $[0,\pi]$ gegeben mit $f(0)=f(\pi)=0$, so wird eine 2π-periodische Fortsetzung von f erklärt durch

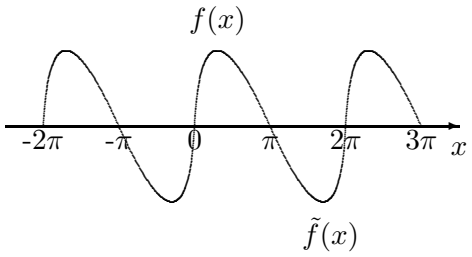
$$\tilde{f}(x) := \begin{cases} f(x), & x\in[0,\pi], \\ -f(2\pi-x), & x\in[\pi,2\pi], \\ 2\pi\text{-periodisch auf } \mathbb{R} \text{ fortgesetzt.} \end{cases}$$

Offenbar ist dann $\tilde{f}$ in $x=\pi$ stetig differenzierbar (falls f in $x=\pi$ einseitig differenzierbar war):

$$\lim_{h\to+0}\frac{\tilde{f}(\pi+h)-\tilde{f}(\pi)}{h} = \lim_{h\to+0}\frac{-f(\pi-h)+f(\pi)}{h} = f'(\pi)$$

$$\lim_{h\to+0}\frac{\tilde{f}(\pi)-\tilde{f}(\pi-h)}{h} = \lim_{h\to+0}\frac{f(\pi)+f(\pi-h)}{h} = f'(\pi).$$

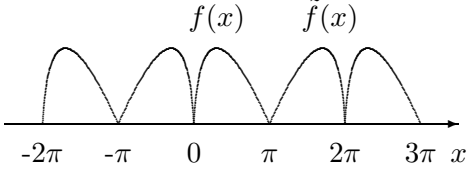
Beispiel 2.7: a) Beispiel einer "ungeraden" periodischen Fortsetzung:



$$\int_0^{2\pi}\tilde{f}(x)\,dx = 0$$

Abbildung 2.7: *Ungerade periodische Fortsetzung*

b) Beispiel einer "geraden" periodischen Fortsetzung:



$$\int_0^{2\pi}\tilde{f}(x)\,dx = 2\int_0^{\pi} f(x)\,dx$$

Abbildung 2.8: *Gerade periodische Fortsetzung*

Je nachdem ob $f:[0,\pi]\to\mathbb{R}$ als "ungerade" oder als "gerade" Funktion 2π-periodisch fortgesetzt wird, ergeben sich bei der trigonometrischen Interpolation von f die folgenden Sonderfälle: (Beachte $f(0)=f(\pi)=0$.)

a) Gerade Fortsetzung:

$$x_k = k\,\frac{2\pi}{n+1}, \quad k = 0, \ldots, n = 2m+1$$

$$y_k = \begin{cases} f(x_k)\,, & k = 0, \ldots, m\,, \\ f(2\pi - x_k)\,, & k = m+1, \ldots, n \end{cases} \quad \Rightarrow \quad y_k = y_{n+1-k},\ k = 1, \ldots, n\,.$$

Für $k = 1, \ldots, m$ folgt:

$$\begin{aligned} c_k &= \frac{1}{n+1}\sum_{j=0}^{n} y_j e^{-ijx_k} = \frac{1}{n+1}\sum_{j=0}^{n} y_{n+1-j} e^{-ijx_k} \\ &= \frac{1}{n+1}\sum_{j=0}^{n} y_j e^{-i(n+1-j)x_k} = c_{n+1-k} = c_{-k}, \\ b_k &= i(c_k - c_{-k}) = 0\,, \\ a_k &= 2c_k = \frac{2}{n+1}\sum_{j=0}^{m} y_j e^{-ijx_k} + \frac{2}{n+1}\sum_{j=m+1}^{n} y_j e^{-ijx_k} \\ &= \frac{2}{n+1}\sum_{j=0}^{m} y_j e^{-ijx_k} + \frac{2}{n+1}\sum_{j=1}^{m} \underbrace{y_{n+1-j}}_{=\,y_j} e^{-i(n+1-j)x_k} \\ &= \frac{2}{n+1}\sum_{j=0}^{m} y_j e^{-ijx_k} + \frac{2}{n+1}\sum_{j=1}^{m} y_j e^{ijx_k} \quad (n = 2m+1) \\ &= \frac{4}{n+1}\sum_{j=1}^{m} y_j \underbrace{\tfrac{1}{2}\left(e^{ijx_k} + e^{-ijx_k}\right)}_{=\,\cos(jx_k)} \quad (y_0 = y_{m+1} = 0) \\ &= \frac{2}{m+1}\sum_{j=1}^{m} y_j \cos(jx_k). \end{aligned}$$

Die gerade fortgesetzte Funktion läßt sich also durch eine "Cosinus-Summe" interpolieren:

$$s(x) = \frac{1}{2}a_0 + \sum_{k=1}^{m+1} a_k\,\cos(kx)\,, \quad a_k = \frac{2}{m+1}\sum_{j=1}^{m} \tilde{f}\,(x_j)\,\cos(jx_k).$$

b) Ungerade Fortsetzung:

$$x_k = k\,\frac{2\pi}{n+1}, \quad k = 0, \ldots, n = 2m+1,$$

$$y_k = \begin{cases} f\,(x_k)\,, & k = 0, \ldots, m \\ -f(2\pi - x_k)\,, & k = m+1, \ldots, n \end{cases} \quad \Rightarrow \quad y_k = -y_{n+1-k}\,, \quad k = 1, \ldots, n$$

Dies ergibt analog wie im Fall (a):

$$c_k = -c_{-k} \;\Rightarrow\; a_k = 0\,, \quad b_k = 2ic_k = \ldots.$$

Die ungerade fortgesetzte Funktion läßt sich also durch eine "Sinus-Summe" interpolieren:

$$s\,(x) = \sum_{k=1}^{m} b_k\, \sin(kx)\,, \quad b_k = \frac{2}{m+1} \sum_{j=1}^{m} \tilde{f}\,(x_j)\, \sin(jx_k).$$

2.4.1 "Schnelle" Fourier-Transformation ("FFT")

Wir diskutieen nun noch die effiziente Berechnung des trigonometrischen Interpolationspolynoms.

Definition 2.7: *Die in Satz 2.12 behandelte Aufgabenstellung wird "diskrete Fourier-Analyse" genannt: Den* $n+1$ *Werten* $y_j = f(x_j)$, $(j = 0, \ldots, n)$ *einer Funktion* $f : [0, 2\pi] \to \mathbb{R}$ *werden (eindeutig) die* $n+1$ *Koeffizienten* a_k, b_k *des trigonometrischen Interpolationspolynoms zugeordnet*

$$t_n\,(x) = \frac{1}{2}a_0 + \sum_{k=1}^{m} \{a_k\, \cos(kx) + b_k\, \sin(kx)\} + \frac{\theta}{2}\, a_{m+1}\, \cos((m+1)\,x).$$

Die Abbildung $\{y_j\} \to \{a_k, b_k\}$ *heißt "diskrete Fourier-Transformation"; sie ist offenbar umkehrbar.*

Interpretiert man $\cos(kx)\,, \sin(kx)$ als Grundschwingungsformen eines 2π-periodischen Prozesses $y = f\,(x)\,$, so bedeutet eine (diskrete) Fourier-Analyse die Bestimmung der jeweiligen Anteile a_k, b_k dieser Grundschwingungen am Prozeß. Zur Berechnung dieser Koeffizienten a_k, b_k bzw. zur Bestimmung des Polynoms $t_n(\cdot)$ nach den Formeln

$$a_k = c_k + c_{-k}\,, \quad b_k = i\,(c_k - c_{-k})$$

sind $n+1$ Summen der Form

$$c_k := \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{ijk2\pi/(n+1)} = \sum_{j=0}^{n} \bar{y}_j w^{jk}\,, \quad k = 0, \ldots, n,$$

zu berechnen mit den Abkürzungen (Man beachte, daß $w^{n+1} = 1$ ist.)

$$\bar{y}_j := \tfrac{1}{n+1} y_j\,, \quad w := e^{-i2\pi/(n+1)}.$$

Das Horner-Schema erfordert dazu jeweils n (komplexe) Operationen (1 komplexe Multiplikation und 1 komplexe Addition), d. h.: Insgesamt sind $n^2 + n$ Operationen erforderlich. Für große n und bei mehrfacher Ausführung der Fourier-Analyse bedeutet dies einen beträchtlichen numerischen Aufwand. Im Jahre 1965 gaben Cooley und

Tukey[10] einen Algorithmus an, der diese Aufgabe wesentlich effizienter löst, die sog. "Schnelle Fourier-Transformation" ("Fast Fourier Transform" oder auch kurz "FFT"). Wir erläutern diese für den Spezialfall $n+1 = 2^p$, $p \in \mathbb{N}$:

Idee der FFT: Durch Aufspaltung der Summe

$$\begin{aligned} c_k &= \bar{y}_0 w^{0k} + \bar{y}_1 w^{1k} + \bar{y}_2 w^{2k} + \bar{y}_3 w^{3k} + \ldots + \bar{y}_{2^p-2} w^{(2^p-2)k} + \bar{y}_{2^p-1} w^{(2^p-1)k} \\ &= \underline{\{\ \bar{y}_0 (w^2)^{0k}} + \underline{\bar{y}_1 (w^2)^{0k} w^k\}} + \underline{\{\ \bar{y}_2 (w^2)^{1k}} + \underline{\bar{y}_3 (w^2)^{1k} w^k\}} + \cdots \\ &\quad \cdots + \underline{\{\ \bar{y}_{2^p-2} (w^2)^{(2^{p-1}-1)}} + \underline{\bar{y}_{2^p-1} (w^2)^{(2^{p-1}-1)} w^k\}} \end{aligned}$$

bzgl. gerader ("—") und ungerader ("$\cdots$") Indizes erhält man

$$c_k = \sum_{j=0}^{2^{p-1}-1} \bar{y}_{2j} (w^2)^{jk} + \sum_{j=0}^{2^{p-1}-1} \bar{y}_{2j+1} (w^2)^{jk} w^k.$$

Sei $k_1 \in \{0, 1, \ldots, 2^{p-1}-1\}$ der ganzzahlige Rest bei Division von k durch 2^{p-1}:

$$k \equiv k_1 \pmod{2^{p-1}}.$$

Wegen $w^{2^p} = 1$ gilt mit einem geeigneten $\lambda \in \mathbb{N}$

$$(w^2)^{jk} = (w^2)^{\lambda 2^p j} (w^2)^{jk_1} = (w^{2^p})^{\lambda j} (w^2)^{jk_1} = (w^2)^{jk_1}.$$

Folglich ist

$$c_k = \sum_{j=0}^{2^{p-1}-1} \bar{y}_{2j} (w^2)^{jk_1} + w^k \sum_{j=0}^{p-1-1} \bar{y}_{2j+1} (w^2)^{jk_1}$$

bzw.

$$c_k = \tilde{c}_{k_1} + w^k \bar{c}_{k_1}, \quad k = 0, \ldots, 2^p - 1, k \equiv k_1 \pmod{2^{p-1}},$$

mit den Teilsummen

$$\tilde{c}_{k_1} := \sum_{j=0}^{2^{p-1}-1} \bar{y}_{2j} (w^2)^{jk_1}, \quad \bar{c}_{k_1} := \sum_{j=0}^{2^{p-1}-1} \bar{y}_{2j+1} (w^2)^{jk_1}, \quad k_1 = 0, \ldots, 2^{p-1}-1.$$

Zur Berechnung der $n+1 = 2^p$ Größen c_k genügt es also, die $2 \cdot 2^{p-1}$ Größen $\tilde{c}_{k_1}$, $\bar{c}_{k_1}$ zu bestimmen, d. h.: Die Fourier-Analyse mit $n+1 = 2^p$ Termen wird ersetzt durch zwei Fourier-Analysen mit jeweils 2^{p-1} Termen. Anwendung derselben Aufspaltung auf $\tilde{c}_{k_1}$ und $\bar{c}_{k_1}$ ergibt vier Fourier-Analysen mit jeweils 2^{p-2} Termen usw. Am Schluß verbleiben als Startpunkt des Algorithmus 2^p Fourier-Analysen mit jeweils einem Term. Diese "tri-

[10] John Wilder Tukey (1915-2000): US-Amerikanischer Mathematiker; arbeitete seit 1945 an der Princten University, seit 1956 als Direktor der Statistics research Group; bekannt vor allem durch die gemeinsam mit J.W. Cooley (IBM Cooperation) entwickelte FFT: An algorithm for the machine calculation of complex Fourier series, Math. Comput. 19, 297-301 (1965); wichtige Beiträge auch zur praktischen Statistik.

vialen" Fourier-Analysen ordnen den einzelnen Stützpunkten x_k, $(k = 1, \cdots, n)$ gerade die Werte y_k zu, welche die jeweiligen Koeffizienten der trigonometrischen Approximation 0-ter Ordnung durch konstante Funktionen darstellen.

Satz 2.13 (Schnelle Fourier-Transformation): *Im Fall* $n+1 = 2^p$ *löst die "Schnelle Fourier-Transformation" das Problem der Berechnung der* $n+1$ *Fourier-Koeffizienten* $\{c_0, \ldots, c_n\}$, *mit* $2\,(n+1)\log_2(n+1)$ *(komplexen) Operationen.*

Beweis: Sei r_p die Anzahl der (komplexen) Operationen zur Berechnung aller Koeffizienten $\{c_0, \ldots c_n\}$ im Falle $n+1 = 2^p$. Sind die Größen $\tilde{c}_{k_1}$, $\bar{c}_{k_1}$ $(k_1 = 0, \ldots, 2^{p-1}-1)$ bekannt, so erfordert die Berechnung der Potenzen w^k $(k = 1, \ldots, n)$ und der Koeffizienten $c_k = \tilde{c}_{k_1} + w^k \bar{c}_{k_1}$ $(k = 0, \ldots, n)$ offenbar höchstens $(n-1) + (n+1) = 2n \leq 2 \cdot 2^p$ Operationen. (Eine große Ersparnis wäre durch vorausgehende Berechnung und Speicherung der w^k zu erzielen!) Also wird

$$r_p \leq 2 \cdot r_{p-1} + 2 \cdot 2^p\,, \quad p = 1, 2, 3, \ldots.$$

Wir wollen zeigen, daß $r_p \leq 2p \cdot 2^p$ gilt. Ausgehend von $r_0 = 0$ folgt durch Induktion, daß

$$r_p \leq 2 \cdot r_{p-1} + 2 \cdot 2^p = 2 \cdot \left(2(p-1) \cdot 2^{p-1}\right) \leq 2p \cdot 2^p = 2\,\log_2(n+1)\,(n+1),$$

was den Beweis vervollständigt. Q.E.D.

Beispiel 2.8: $n = 2^7 - 1 = 127$

$$n^2 + n = 16.256\,, \quad 2\,(n+1)\underbrace{\log_2\,(n+1)}_{\sim 7} = 1.792.$$

Implementierung: Bei der konkreten Implementierung der FFT geht man genau entgegengesetzt zu ihrer obigen Herleitung vor. Zunächst werden die 2^p ein-elementigen Fourier-Analysen durchgeführt, welche die gegebenen Stützwerte y_j, $(j = 1, \ldots, 2^p - 1)$ verwenden. Danach werden nur noch die rekursiven Formeln

$$c_k = \tilde{c}_{k_1} + w^k \bar{c}_{k_1}\,, \quad k = 0, \ldots, 2^p - 1\,, \;\; k \equiv k_1 \,(\mathrm{mod}\, 2^{p-1}),$$

abgearbeitet. Ein FORTRAN-Unterprogramm zur Durchführung der FFT im Spezialfall $n+1 = 2^p$ lautet etwa wie folgt (in Anlehnung an das in der Arbeit von Cooley, Lewis und Welch, IEEE Transactions, E-12, 1 (March 1965) angegebene Programm):

FORTAN-Programm FFT (Fall $n+1=2^p$):

```
1.  SUBROUTINE  FFT(C,P)
2.  COMPLEX C(1),U,W,T
3.  C  Setze M:=N+1, C:=Y/(N+1)
4.  M=2**P
5.  M2=M/2
6.  M1=M-1
7.  J=1
8.  C  Umsortieren der C(I)
9.  DO 3 I=1,M1
10. IF(I.GE.J) GO TO 1
11. T=C(J)
12. C(J)=C(I)
13. C(I)=T
14. 1  K=M2
15. 2  IF(K.GE.J) GO TO 3
16. J=J-K
17. K=K/2
18. GO TO 2
19. 3  J=J+K
20. DO 5 L=1,P
21. LE=2**L
22. LE1=LE/2
23. U=(1.,0.)
24. ANG=3.14159265358979/LE1
25. W=CMPLX(COS(ANG),SIN(ANG))
26. DO 5 J=1,LE1
27. DO 4 I=J,M,LE
28. IP=I+LE1
29. T=C(IP)*U
30. C(IP)=C(I)-T
31. 4  C(I)=C(I)+T
32. 5  U=U*W
33. RETURN
```

2.5 Gauß-Approximation

Wir fassen im folgenden die Menge $C[a,b]$ der über einem Intervall $[a,b]$ stetigen reell- bzw. komplex-wertigen Funktionen als einen (unendlich dimensionalen) Vektorraum über dem Zahlkörper $\mathbb{K} = \mathbb{R}$ bzw. $\mathbb{K} = \mathbb{C}$ auf.

Bemerkung 2.3: *Die Aussagen dieses Abschnitts gelten sinngemäß auch für den Vektorraum $R[a,b]$ der über dem Intervall $[a,b]$ Riemann-integrierbaren Funktionen oder sogar für den Vektorraum $L^2(a,b)$ der über (a,b) im Lebesgueschen Sinne quadrat-integrablen Funktionen.*

Gegeben sei eine Funktion $f \in C[a,b]$ sowie ein endlich dimensionaler Teilraum $S \subset C[a,b]$, dessen Elemente zur Approximation von f dienen sollen, z.B.: $S = P_n$, Raum der Polynome vom Grad $\leq n$. Im Gegensatz zur Interpolation verwendet die "Gauß-Approximation" das sog. "quadratische Mittel"

$$\|f\| \equiv \Big(\int_a^b |f(x)|^2 \, dx \Big)^{1/2}$$

als Maß für die Güte einer Approximation. d. h.: Gesucht ist ein $g \in S$, so daß

$$\|f - g\| = \min_{\varphi \in S} \|f - \varphi\|. \tag{2.5.31}$$

Ein $g \in S$ mit dieser Eigenschaft heißt dann "Bestapproximation" von f (in S bzgl. $\|\cdot\|$). Durch $\|\cdot\|$ ist auf $C[a,b]$ eine Norm gegeben, d. h. eine Funktion $\|\cdot\| : C[a,b] \to \mathbb{R}_+$ mit analogen Eigenschaften wie die wohlbekannte euklidische Vektornorm auf dem $\mathbb{K}^n$. Es sei an die folgenden Eigenschaften einer "Norm" erinnert:

1. Definitheit: $\|f\| \in \mathbb{R}_+, \quad \|f\| = 0 \;\Rightarrow\; f = 0$.
2. Sublinearität: $\|f+g\| \leq \|f\| + \|g\|$ (Dreiecksungleichung),
3. Homogenität: $\|\alpha f\| = |\alpha| \, \|f\|, \quad \alpha \in \mathbb{K}$.

Das Analogon zum euklidischen Skalarprodukt ist in diesem Fall das sog. L^2-Skalarprodukt

$$(f,g) \equiv \int_a^b f(t)\overline{g(t)} \, dt, \quad (f,f) = \|f\|^2.$$

Hierfür liegen wieder die für ein "Skalarprodukt" charakteristischen Eigenschaften vor:

1. Definitheit: $(f,f) \in \mathbb{R}_+, \quad (f,f) = 0 \;\Rightarrow\; f = 0$.
2. Linearität: $(\alpha f + g, h) = \alpha(f,g) + (h,g), \quad \alpha \in \mathbb{K}$,
3. Symmetrie: $(f,g) = \overline{(g,f)}$.

Nicht jede Norm gehört zu einem Skalarprodukt; z.B.: die sog. L^p-Normen $\|\cdot\|_p$ für $p \in [1,\infty) \setminus \{2\}$, sowie die sog. "Maximumnorm" $\|\cdot\|_\infty$:

$$\|f\|_p := \big(\int_a^b \|f(x)\|^p\,dx\big)^{1/p}, \qquad \|f\|_\infty := \max_{a\le x\le b} |f(x)|.$$

Wir notieren noch die wichtige "Höldersche[11] Ungleichung" für Skalarprodukte (auch "Schwarzsche[12] Ungleichung" im Fall allgemeiner Skalarprodukte)

$$|(f,g)| \le \|f\|\,\|g\|.$$

Versehen mit dem L^2-Skalarprodukt wird $C[a,b]$ zu einem sog. "unitären Raum". Für die Gauß-Approximation in unitären Räumen haben wir die folgende allgemeine Aussage:

Satz 2.14 (Allgemeine Gauß-Approximation): *Seien H ein unitärer Raum und $S \subset H$ ein endlich dimensionaler Teilraum. Dann existiert zu jedem $f \in H$ eine eindeutig bestimmte "beste Approximation" $g \in S$:*

$$\|f-g\| = \min_{\varphi\in S} \|f-\varphi\|. \tag{2.5.32}$$

Beweis: (i) Wir wollen die Eigenschaft der besten Approximation zunächst durch eine etwas handlichere Bedingung charakterisieren. Sei $g \in S$ eine beste Approximation. Dann besitzt für beliebiges, fest gewähltes $\varphi \in S$ die quadratische Funktion

$$F_\varphi(t) := \|f-g-t\varphi\|^2, \quad t\in\mathbb{R},$$

bei $t=0$ ein Minimum. Folglich ist

$$\frac{d}{dt}F(t)_{|t=0} = \frac{d}{dt}\|f-g-t\varphi\|^2_{|t=0} = 0.$$

Ausgeschrieben bedeutet dies $(f-g-t\varphi,\varphi)_{|t=0}=0$ und folglich

$$(f-g,\varphi) = 0 \quad \forall\varphi\in S. \tag{2.5.33}$$

Diese Beziehung kann man so interpretieren, daß der Fehler $f-g$ auf dem approximierenden Teilraum S "orthogonal" ist bzgl. des L^2-Skalarprodukts $(\cdot,\cdot)$.

Genüge nun umgekehrt $g \in S$ der Bedingung (2.5.33). Dann gilt mit einem beliebigen $\varphi \in S$:

$$\|f-g\|^2 = (f-g,f-g) = (f-g,f-\varphi)+(f-g,\varphi-g) \le \|f-g\|\,\|f-\varphi\|$$

[11] Ludwig Otto Hölder (1859-1937): deutscher Mathematiker; Prof. in Tübingen; Beiträge zunächst zur Theorie der Fourier-Reihen und später vor allem zur Gruppentheorie; fand 1884 die nach ihm benannte Ungleichung.

[12] Hermann Schwarz (1843-1921): Deutscher Mathematiker; wirkte in Halle, Göttingen und Berlin; leistete grundlegende Arbeiten zur Funktionentheorie, Differentialgeometrie und Variationsrechnung.

und folglich

$$\|f - g\| \leq \inf_{\varphi \in S} \|f - \varphi\|,$$

d. h.: g ist auch beste Approximation.

(ii) Eindeutigkeit der besten Approximation: Seien g_1, $g_2 \in S$ zwei Bestapproximationen. Dann gilt notwendig

$$(f - g_1, \varphi) = 0 = (f - g_2, \varphi) \quad \forall \varphi \in S,$$

und folglich

$$(g_1 - g_2, \varphi) = 0 \quad \forall \varphi \in S.$$

Wählen wir $\varphi := g_1 - g_2$, ergibt sich $\|g_1 - g_2\|^2 = 0$ und somit $g_1 = g_2$.

(iii) Existenz der besten Approximation: Der endlich dimensionale Teilraum $S \subset H$ besitzt eine Basis $\{\psi_1, \ldots, \psi_n\}$, $n := \dim H$, bzgl. derer sich die gesuchte besten Approximation $g \in S$ darstellen läßt in der Form

$$g = \sum_{k=1}^{n} \alpha_k \psi_k.$$

Einsetzen dieses Ansatzes in die notwendige Orthogonalitätsbedingung (2.5.33) ergibt

$$(f - \sum_{i=k}^{n} \alpha_k \psi_i k, \varphi) = (f, \varphi) - \sum_{k=1}^{n} \alpha_k (\psi_k, \varphi) = 0 \quad \forall \varphi \in S.$$

Dies ist bei sukzessiver Wahl von $\varphi := \psi_i$ für $i = 1, \ldots, n$, äquivalent zu dem linearen $n \times n$-Gleichungssystem

$$\sum_{k=1}^{n} (\psi_k, \psi_i) \alpha_k = (f, \psi_i), \quad i = 1, \cdots, n. \tag{2.5.34}$$

Mit der Notation

$$\alpha := (\alpha_k)_{k=1}^n, \quad b := ((f, \psi_i))_{i=1}^n, \quad A := ((\psi_k, \psi_i))_{i,k=1}^n,$$

läßt sich dies in der kompakten Form $A\alpha = b$ schreiben. Die Matrix A ist als "Gramsche Matrix" der Basis $\{\psi_1, \ldots, \psi_n\}$ regulär. Dies ersieht man aus der Beziehung

$$\bar{\alpha}^T A \alpha = \sum_{i,k=1}^{n} \bar{\alpha}_i \alpha_k (\psi_k, \psi_i) = (g, g),$$

welche die Injektivität von A impliziert. Ferner ist A offenbar "symmetrisch" (im Fall $\mathbb{K} = \mathbb{R}$) bzw. "hermitesch" (im Fall $\mathbb{K} = \mathbb{C}$) und folglich "positiv definit". Das Gleichungssystem $A\alpha = b$ ist also für jede rechte Seite b, d. h. für jedes $f \in H$ eindeutig lösbar. Folglich bestimmt die Orthogonalitätsbedingung (2.5.33) eindeutig ein Element $g \in S$, welches dann notwendig eine Bestapproximation ist. Q.E.D.

Zur Konstruktion der besten Approximation $g \in S$ zu einer Funktion $f \in H$ kann zunächst das Gleichungssystem (2.5.34) dienen:

$$\sum_{k=1}^{n}(\psi_k, \psi_i)\alpha_k = (f, \psi_i), \quad i = 1, \cdots, n. \tag{2.5.35}$$

Es besitzt eine besonders einfache Lösung, wenn die Basis $\{\psi_1, \ldots, \psi_n\}$ ein "Orthonormalsystem" ("ONS") ist, d. h.: $(\psi_k, \psi_i) = \delta_{ki}$. Dann gilt offenbar

$$\alpha_k = (f, \psi_k), \quad k = 1, \ldots, n,$$

d. h.: Die beste Approximation ist explizit bestimmt durch

$$g = \sum_{k=1}^{n}(f, \psi_k)\psi_k. \tag{2.5.36}$$

Hilfssatz 2.1 (Gram-Schmidt-Algorithmus): *Zu jeder Basis $\{\psi_1, \ldots, \psi_n\}$ von S läßt sich ein Orthonormalsystem mit dem "Gram[13]-Schmidt[14]-Algorithmus" konstruieren:*

$$\tilde{\varphi}_1 := \psi_1\,, \quad \varphi_1 := \|\tilde{\varphi}_1\|^{-1}\tilde{\varphi}_1,$$

$$k = 2, \ldots, n: \qquad \tilde{\varphi}_k := \psi_k - \sum_{i=1}^{k-1}(\psi_k, \varphi_i)\varphi_i\,, \quad \varphi_k := \|\tilde{\varphi}_k\|^{-1}\tilde{\varphi}_k.$$

Das Ergebnis ist ein Orthonormalsystem $\{\varphi_1, \ldots, \varphi_n\}$ in S.

Beweis: Dies zeigt man durch Induktion nach $n = \dim S$. Im Fall $\psi_1 \neq 0$ ist φ_1 wohldefiniert. Sei nun $\{\varphi_1, \ldots, \varphi_{n-1}\}$ wohldefiniert und ONS. Im Fall

$$\tilde{\varphi}_n = \psi_n - \sum_{k=1}^{n-1}(\psi_n, \varphi_k)\varphi_k = 0$$

wäre $\{\psi_1, \ldots, \psi_n\}$ linear abhängig, im Widerspruch zur Annahme. Also ist φ_n wohldefiniert. Weiter ist für $k = 1, \ldots, n-1$:

$$(\varphi_n, \varphi_k) = (\psi_n, \psi_k) - \sum_{i=1}^{n-1}(\psi_n, \varphi_i)\underbrace{(\varphi_i, \varphi_k)}_{=\delta_{ik}} = 0\,,$$

[13] *Jørgen Pedersen Gram (1850-1916): dänischer Mathematiker, Mitarbeiter und später Eigentümer einer Versicherungsgesellschaft, Beiträge zur Algebra (Invariantentheorie), Wahrscheinlichkeitstheorie, Numerik und Forstwissenschaft; das u.a. nach ihm benannte Orthogonalisierungsverfahren geht aber wohl auf Laplace zurück und wurde bereits von Cauchy 1836 verwendet.*

[14] *Erhard Schmidt (1876-1959): deutscher Mathematiker, Prof. in Berlin, Gründer des dortigen Instituts für Angewandte Mathematik 1920, nach dem Krieg Direktor des Mathematischen Instituts der Akademie der Wissenschaften der DDR; Beiträge zur Theorie der Integralgleichungen und der Hilbert-Räume sowie später zur Topologie.*

und $\|\varphi_n\| = 1$ nach Konstruktion. Q.E.D.

Im folgenden wollen wir die obigen allgemeinen Aussagen zur Gauß-Approximation mit Polynomen anwenden. O.B.d.A. legen wir das Intervall $[a,b] = [-1,1]$ zugrunde (gegebenenfalls Variablentransformation). Nach Satz 2.14 existiert zu jedem $f \in C[-1,1]$ die eindeutig bestimmte beste Approximation $g \in S = P_n$. Zu ihrer Berechnung sei zunächst die Basis $\{1, x, \ldots, x^n\}$ von P_n herangezogen. Die Koeffizienten in der Darstellung $g = \sum_{k=0}^{n} \alpha_k x^k$ ergeben sich dann als Lösung eines linearen Gleichungssystems mit der Koeffizientenmatrix $A = (a_{ik})_{i,k=0}^{n}$, wobei

$$A = 2 \begin{bmatrix} 1 & 0 & 1/3 & 0 & 1/5 & \cdots \\ 0 & 1/3 & 0 & 1/5 & & \\ 1/3 & 0 & 1/5 & 0 & & \\ 0 & 1/5 & & \ddots & & \\ 1/5 & & & & & \\ \vdots & & & & & \end{bmatrix}.$$

$$a_{ik} = \int_{-1}^{1} x^k x^i \, dx = \begin{cases} 0 & , \text{ falls } \quad i+k \quad \text{ungerade} \\ \dfrac{2}{i+k+1} & , \text{ falls } \quad i+k \quad \text{gerade} \end{cases}$$

Diese Matrix (sog. Hilbert-Matrix) ist zwar regulär, doch ist ihre Invertierung so extrem schlecht konditioniert, daß für große n die Berechnung von g auf diesem Wege unmöglich ist. Statt dessen wird die Basis $\{1, \ldots, x^n\}$ bzgl. des (reellen) Skalarprodukts ($\mathbb{K} = \mathbb{R}$)

$$(f,g) := \int_{-1}^{1} f(x) g(x) \, dx$$

orthonormalisiert. Das Ergebnis fassen wir in folgendem Satz zusammen.

Satz 2.15 (Legendre-Polynome): *Durch Orthogonalisierung der natürlichen Monombasis $\{1, x, ..., x^n\}$, $n \in \mathbb{N}$, mit dem Gram-Schmidt-Algorithmus ergeben sich Polynome $p_k \in P_k$ (nicht normalisiert), welche sich in der Form*

$$p_k(x) = \frac{k!}{(2k)!} \frac{d^k}{dx^k}(x^2-1)^k, \quad k = 0, 1, ..., n, \tag{2.5.37}$$

darstellen lassen. Für sie gilt die zweistufige Rekursionsformel

$$\begin{aligned} p_0(x) &\equiv 1\,, \quad p_1(x) = x\,, \\ p_{k+1}(x) &= x p_k(x) - \frac{k^2}{4k^2-1} p_{k-1}(x), \quad k = 1, 2, \ldots, n-1. \end{aligned} \tag{2.5.38}$$

sowie

$$\|p_k\| = \frac{k!^2}{(2k)!}\sqrt{\frac{2^{2k+1}}{2k+1}}\,, \qquad p_k(1) = \frac{2^k k!^2}{(2k)!}. \tag{2.5.39}$$

Durch Normierung bei $x = 1$ *erhält man die sog. "Legendre*[15]*-Polynome"*

$$L_k(x) := \frac{(2k)!}{2^k k!^2}\, p_k(x), \quad L_k(1) = 1, \quad k = 0, 1, 2, \ldots, n\,. \tag{2.5.40}$$

Beweis: Wir führen den Beweis in mehreren Schritten, wobei einige Rechnungen als Übungsaufgabe gestellt sind.

(i) Wir zeigen zunächst, daß die durch (2.5.37) definierten Polynome $p_k \in P_k$ bzgl. des Skalarprodukts $(\cdot,\cdot)$ orthogonal sind. Dies ergibt sich durch partielle Integration über dem Intervall $[-1, 1]$ (Übungsaufgabe). Analog erschließen wir die Beziehungen (2.5.39).

(ii) Der führende Term von $p_k(x)$ ist gemäß der Definition x^k. Folglich sind die p_k gerade die durch den Gram-Schmidt-Algorithmus aus der Monombasis erzeugten orthogonalen Polynome.

(iii) Das Polynom $(x^2-1)^k$ ist eine gerade Funktion. Seine k-ten Ableitungen sind dann ungerade oder gerade je nachdem, ob k ungerade oder gerade ist. Folglich ist

$$p_k(x) = (-1)^k p_k(-x)\,.$$

(iv) Wegen $p_{k+1}(x) = x^{k+1} + ...$ ist $p_{k+1}(x) - xp_k(x) = \gamma_k x^k + \gamma_{k-1}x^{k-1} + ... + \gamma_0$ mit gewissen Koeffizienten $\gamma_k, ..., \gamma_0$. Ist nun $k+1$ gerade, so ist das Polynom $p_{k+1}(x) - xp_k(x)$ gerade aber x^k ungerade, so daß notwendig $\gamma_k = 0$ sein muß. Es gibt daher eine Darstellung

$$p_{k+1}(x) - xp_k(x) = \sum_{i=0}^{k-1} \gamma_i p_i(x)$$

mit den Polynomen $p_0, ..., p_{k-1}$, die ja als Orthogonalsystem eine Basis von P_{k-1} bilden. Wegen der Orthogonalität der p_k folgt dann für $j = 0, ..., k-2$:

$$0 = (p_{k+1} - xp_k, L_j) = \sum_{i=0}^{k-1} \gamma_i (p_i, p_j) = \gamma_j \|p_j\|^2,$$

bzw. $\gamma_0 = ... = \gamma_{k-2} = 0$. Es besteht also eine zweistufige Rekursion der Form

$$p_{k+1}(x) = xp_k(x) + \gamma_{k-1}p_{k-1}(x).$$

[15] *Adrien-Marie Legendre (1752-1833): französischer Mathematiker; Mitglied der Pariser Akademie der Wissensch.; Beiträge zur Himmelsmechanik, Zahlentheorie und Geometrie.*

Zur Bestimmung des Koeffizienten γ_{k-1} verwenden wir $p_k(1) = \frac{k!^2}{(2k)!}2^k$. Es ergibt sich

$$\gamma_{k-1} = \frac{p_{k+1}(1) - p_k(1)}{p_{k-1}(1)} = \frac{\frac{(k+1)!^2}{(2k+2)!}2^{k+1} - \frac{k!^2}{(2k)!}2^k}{\frac{(k-1)!^2}{(2k-2)!}2^{k-1}}$$
$$= \frac{4k^2(k+1)^2 - 2k^2(2k+2)(2k+1)}{(2k+2)(2k+1)2k(2k-1)} = \frac{k(k+1) - k(2k+1)}{(2k+1)(2k-1)} = -\frac{k^2}{4k^2-1},$$

wie behauptet. Q.E.D.

Die Rekursionsformel (2.5.38) für die Polynome p_n ist ein Spezialfall eines allgemeinen Resultats für "orthogonale Polynome"; siehe den folgenden Satz 2.17.

Die Gauß-Approximation mit orthogonalen Polynomen hat den Vorteil der formal einfachen Berechenbarkeit der besten Approximation

$$g(x) = \sum_{k=0}^{n} \Big(\int_{-1}^{1} |p_k(x)|^2 \, dx \Big)^{-1} \Big(\int_{-1}^{1} f(\xi) p_k(\xi) \, d\xi \Big) \, p_k(x).$$

Die Berechnung der Koeffizienten erfordert i. Allg. numerische Quadratur.

Die Maximalabweichung der Gauß-Approxierenden

$$\|f - g\|_\infty = \max_{-1 \le x \le 1} |f(x) - g(x)|$$

wird jedoch i. Allg. groß; insbesondere in der Nähe der Intervallenden treten große Fehler auf. Zur Unterdrückung dieses Defektes verwendet man das gewichtete Skalarprodukt

$$(f, g)_\omega = \int_{-1}^{1} f(x) g(x) \omega(x) \, dx \,, \quad \omega(x) \equiv \frac{1}{\sqrt{1-x^2}},$$

wodurch eine stärkere Bindung in der zugehörigen Fehlernorm

$$\|f - g\|_\omega = \Big(\int_{-1}^{1} |f(x) - g(x)|^2 \frac{dx}{\sqrt{1-x^2}} \Big)^{1/2}$$

an den Intervallenden impliziert wird. Zur Anwendung von Satz 2.14 wird die Basis $\{1, x, \ldots, x^n\}$ von P_n nun bzgl. dieses neuen Skalarproduktes orthogonalisiert.

Satz 2.16 (Tschebyscheff-Polynome): *Durch Orthogonalisierung der natürlichen Monombasis* $\{1, x, ..., x^n\}$, $n \in \mathbb{N}$, *bzgl. des gewichteten Skalarprodukts* $(\cdot, \cdot)_\omega$ *mit dem Gram-Schmidt-Algorithmus ergeben sich Polynome* $p_k \in P_k$ *(nicht normalisiert), welche sich in der Form*

$$p_0(x) \equiv 1, \quad p_k(x) = 2^{k-1} \cos[k \arccos(x)], \quad k = 1, 2, ..., n, \tag{2.5.41}$$

darstellen lassen. Für sie gilt die zweistufige Rekursionsformel

$$\begin{aligned} p_0(x) &\equiv 1\,, \quad p_1(x) = x\,, \\ p_{k+1}(x) &= 4xp_k(x) - 4p_{k-1}(x), \quad k = 1, 2, \ldots, n, \end{aligned} \tag{2.5.42}$$

sowie

$$\|p_k\|_\omega = \begin{cases} \sqrt{\pi}, & k = 0 \\ \sqrt{\pi/2}, & k \neq 0, \end{cases} \qquad p_k(1) = 2^{k-1}. \tag{2.5.43}$$

Durch Normierung bei $x = 1$ erhält man die sog. "Tschebyscheff [16]-Polynome"

$$T_k(x) = \cos[k \arccos(x)], \quad T_k(1) = 1, \quad k = 0, 1, 2, \ldots. \tag{2.5.44}$$

Beweis: Zunächst gilt für die Funktionen $g_k := \cos[k \arccos(x)]$ im Fall $k = 0, 1$: $g_0 \equiv 1\,,\ g_1(x) = x\,$. Weiter folgt aus der Identität

$$\cos((n+1)x) + \cos((n-1)x) = 2\cos(x)\,\cos(nx)$$

die rekursive Beziehung

$$\begin{aligned} g_{k+1}(x) &= \cos[(k+1)\arccos(x)] \\ &= 2\,\cos[\arccos(x)]\,\cos[k\,\arccos(x)] - \cos[(k-1)\,\arccos(x)] \\ &= 2xg_k - g_{k-1}. \end{aligned}$$

Weiter erhält man mit Hilfe der Variablentransformation $x = \cos(\theta)$ mit

$$dx = -\sin\theta\, d\theta = -\sqrt{1 - \cos^2(\theta)}\, d\theta = -\sqrt{1 - x^2}\, d\theta$$

die Beziehung

$$\int_{-1}^{1} g_k(x) g_j(x) \omega(x)\, dx = -\int_{\pi}^{0} \cos(k\theta)\,\cos(j\theta)\, d\theta = \begin{cases} \pi, & k = j = 0 \\ \pi/2, & k = j \neq 0\,. \\ 0, & k \neq j \end{cases}$$

Hieraus entnehmen wir, daß die g_k tatsächlich Polynome k-ten Grades über $[-1, 1]$ sind, paarweise orthogonal bzgl. $(\cdot, \cdot)_\omega$ sind und den führenden Koeffizienten $g_k(x) = 2^{k-1}x^k + \ldots$ haben. Die skalierten Polynome $p_0 \equiv 1$ und $p_k := 2^{k-1}q_k$ haben dann den führenden Koeffizienten $p_k(x) = x^k + \ldots$ und genügen der zweistufigen Rekursionsformel

$$p_{k+1}(x) = 4xp_k - 4p_{k-1}.$$

Sie sind also gerade die durch das Gram-Schmidtsche Orthogonalisierungsverfahren erzeugten Polynome. Q.E.D.

[16] *Pafnuty Lvovich Tschebyscheff (russ.: Chebyshev) (1821-1894): russischer Mathematiker; Prof. in St. Petersburg; Beiträge zur Zahlentheorie, Wahrscheinlichkeitstheorie und vor allem zur Approximationstheorie; entwickelte eine allgemeine Theorie orthogonaler Polynome.*

Die beste Approximation einer Funktion $f \in C[-1,1]$ in P_n bzgl. des gewichteten Skalarproduktes $(\cdot,\cdot)_\omega$ hat also die Gestalt

$$g = \sum_{k=0}^{n} \alpha_k T_k(x)$$

mit den Koeffizienten

$$\alpha_0 = \frac{1}{\pi}\int_{-1}^{1} f(x)\,\omega(x)\,dx\,, \quad \alpha_k = \frac{2}{\pi}\int_{-1}^{1} f(x)T_k(x)\omega(x)\,dx\,, \quad k = 1,\ldots,n\,.$$

Das Bestehen von zweistufigen Rekusionsformeln für die Gauß-Legendre sowie für die Tschebyscheff-Polynome legt nahe, daß dies vielleicht generell für orthogonale Polynomen bzgl. Skalarprodukten der betrachteten Form $(\cdot,\cdot)_\omega$ gilt. Dazu gilt der folgende Satz.

Satz 2.17 (Orthogonale Polynome): *Das allgemeine Skalarprodukt $(\cdot,\cdot)$ auf $C[-1,1]$ habe die auf dem Vektorraum P der Polynome die Symmetrieeigenschaft*

$$(p, xq) = (xp, q) \quad \forall p, q \in P. \tag{2.5.45}$$

Dann genügen die durch das Gram-Schmidtsche Orthonormalisierungsverfahren aus der Basis $\{1, x, x^2, \ldots\}$ gewonnenen orthogonalen Polynome p_k, $k = 0, 1, 2, \ldots$ (nicht normalisiert), den rekursiven Beziehungen beginnend mit $p_0(x) \equiv 1$, $p_1(x) = x - \beta_0$:

$$p_{k+1}(x) = (x - \beta_k)p_k - \gamma_k p_{k-1}, \quad k = 1, 2, 3, \ldots, \tag{2.5.46}$$

mit den Koeffizienten

$$\beta_k = \frac{(xp_k, p_k)}{\|p_k\|^2}, \quad k = 0, 1, 2, \ldots, \quad \gamma_k = \frac{\|p_k\|^2}{\|p_{k-1}\|^2}\,, \quad k = 1, 2, 3, \ldots.$$

Beweis: Das Gram-Schmidt-Verfahren erzeugt die Polynome p_k nach der Vorschrift

$$p_0 \equiv 1, \quad k = 1, 2, \ldots: \quad p_k = x^k - \sum_{i=0}^{k-1} \frac{(x^k, p_i)}{\|p_i\|^2} p_i.$$

Also ist $p_0 \equiv 1$ und $p_1 = x - \beta_0$. Wir setzen $q_{k+1} := (x - \beta_k)p_k - \gamma_k p_{k-1}$. Dann gilt offenbar wegen $p_k \perp P_{k-1}$ und der Symmetrieeigenschaft von $(\cdot,\cdot)$:

$$(q_{k+1}, p_k) = (xp_k, p_k) - \beta_k\|p_k\|^2 - \gamma_k \underbrace{(p_{k-1}, p_k)}_{=0} = 0,$$

$$(q_{k+1}, p_{k-1}) = (xp_k, p_{k-1}) - \beta_k \underbrace{(p_k, p_{k-1})}_{=0} - \underbrace{\gamma_k\|p_{k-1}\|^2}_{=\|p_k\|^2} = (p_k, \underbrace{xp_{k-1} - p_k}_{\in P_{k-1}}) = 0,$$

sowie für $j = 0, \ldots, k-2$:

$$(q_{k+1}, p_j) = (p_k, \underbrace{xp_j}_{\in P_{k-1}}) - \beta_k \underbrace{(p_k, p_j)}_{=0} - \gamma_k \underbrace{(p_{k-1}, p_j)}_{=0} = 0.$$

Also ist q_{k+1} orthogonal zu $P_k = \mathrm{Span}\{p_0, \ldots, p_k\}$ und hat die Gestalt

$$q_{k+1}(x) = x^{k+1} + r(x), \quad r \in P_k.$$

Entwickelt man r nach $\{p_0, \ldots, p_k\}$,

$$r(x) = \sum_{i=0}^{k} (r, p_i) \|p_i\|^{-2} p_i(x),$$

so ergibt sich mit

$$\begin{aligned} q_{k+1}(x) &= x^{k+1} + \sum_{i=0}^{k} \frac{(r, p_i)}{\|p_i\|^2} p_i(x) \\ &= x^{k+1} + \sum_{i=0}^{k} \{ \underbrace{(q_{k+1}, p_i)}_{=0} - (x^{k+1}, p_i) \} \|p_i\|^{-2} p_i(x) = p_{k+1}(x) \end{aligned}$$

schließlich die Behauptung. Q.E.D.

2.6 Tschebyscheff-Approximation

Im folgenden betrachten wir nur reell-wertige Funktionen. Die Gauß-Approximation hat gewisse Probleme mit der gleichmäßigen Approximation auf dem ganzen zugrunde liegenden Intervall; sie ist aber verhältnismäßig einfach zu realisieren. Die sog. "Tschebyscheff-Approximation" verwendet direkt die "Maximumnorm"

$$\|f\|_\infty = \max_{a\le x\le b} |f(x)|$$

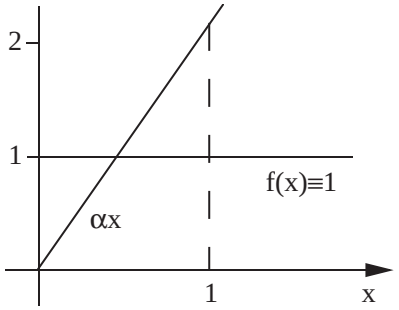
zur Bestimmung der "besten" Approximation $g \in S \subset C[a,b]$.

$$\|f-g\|_\infty = \min_{\varphi\in S} \|f-\varphi\|_\infty. \tag{2.6.47}$$

Die Norm $\|\cdot\|_\infty$ auf $C[a,b]$ wird nicht durch ein Skalarprodukt erzeugt; die Existenz der besten Approximation kann also nicht aus Satz 2.12 erschlossen werden. Tatsächlich ist i. Allg. die beste Approximation nicht einmal eindeutig bestimmt. Ihre Existenz ist jedoch allgemein in normierten Räumen gesichert.

Beispiel 2.9: Das Beispiel belegt die mögliche Mehrdeutigkeit der "besten" Tschebyscheff-Approximation.

$$[a,b] = [0,1]\,, \quad f(x) \equiv 1$$
$$S = \{\, g|\, g(x) = \alpha x\,,\ \alpha \in \mathbb{R}\,\}\,, \quad \dim\, S = 1$$
$$\|f-g\|_\infty \ge 1 \quad \forall g \in S$$
$$\|f-g\|_\infty = 1 \quad \forall g = \alpha x\,, \quad 0 \le \alpha \le 2$$



Satz 2.18 (Allgemeine Tschebyscheff-Approximation): *Sei E ein normierter Vektorraum mit Norm $\|\cdot\|$ und $S \subset E$ ein endlich dimensionaler Teilraum. Dann gibt es zu jedem $f \in E$ eine beste Approximation $g \in S$:*

$$\|f-g\| = \min_{\varphi\in S} \|f-\varphi\|. \tag{2.6.48}$$

Beweis: Ein $g_0 \in S$ mit $\|g_0\| > 2\|f\|$ kann keine beste Approximation sein, da

$$\|f-g_0\| \ge \|g_0\| - \|f\| > \|f\| = \|f-0\| \ge \inf_{\varphi\in S} \|f-\varphi\|.$$

Die optimale Approximation ist also in der beschränkten Teilmenge

$$S_0 := \{\varphi \in S : \|\varphi\| \le 2\|f\|\} \subset S$$

zu suchen. Sie ist abgeschlossen und, da S endlich dimensional ist, kompakt (Satz von Bolzano/Weierstraß). Die auf S stetige Funktion $F(\varphi) := \|f-\varphi\|$ nimmt dann auf S_0

ein Minimum g an, d. h.:

$$\|f-g\| = \min_{\varphi\in S_0} \|f-\varphi\| = \min_{\varphi\in S} \|f-\varphi\|.$$

Q.E.D.

Die Eindeutigkeit der Tschebyscheff-Approximation wird durch die sog. "Haarsche[17] Bedingung" (H) an den Ansatzraum $S \subset C[a,b]$ mit $\dim S = n$ garantiert:

Definition 2.8: *(H) Man sagt, daß der (endlich dimensionale) Teilraum S der "Haarschen Bedingung" genügt, wenn die Lagrangesche Interpolationsaufgabe $g(x_i) = y_i\,,\ i = 1,\ldots,n$ mit beliebigen Stützstellen $a \le x_1 < x_2 < \ldots < x_n \le b$ und Werten $y_1,\ldots,y_n \in \mathbb{R}$ stets durch ein $g \in S$ lösbar ist.*

Für einen Teilraum $S \subset C[a,b]$ ist die Haarsche Bedingung äquivalent zur *eindeutigen* Lösbarkeit der Lagrangeschen Interpolationsaufgabe. Dies sieht man wie folgt:

Sei $\{g_1,\ldots,g_n\}$ eine Basis von S. Die Existenz eines interpolierenden $g = \sum a_i\, g_i \in S$ ist äquivalent zur Lösbarkeit des linearen Gleichungssystems

$$\sum_{i=1}^{n} a_i g_i(x_j) = y_j\,, \quad j = 1,\ldots,n, \tag{2.6.49}$$

für den Koeffizientenvektor $(a_i,\ldots,a_n)^T$. Die Haarsche Bedingung ist äquivalent zur Regularität der Matrix $(g_i(x_j))_{i,j=1,\ldots,n}$, d. h. zur eindeutigen Lösbarkeit der Interpolationsaufgabe.

Beispiel 2.10: *Wir geben Beispiele von Systemen S, für welche die Haarsche Bedingung (H) erfüllt ist.*

1. *Die Polynomräume P_n erfüllen die Haarsche Bedingung auf jedem Intervall $[a,b]$.*
2. *Der Raum $S = \mathrm{Span}\{x,\ldots,x^n\}$ erfüllt die Haarsche Bedingung nicht, wenn $0 \in [a,b]$:*
$$x_1 = 0 \;\Rightarrow\; x_1^k = 0\,, \quad k = 1,\ldots,n \quad\Rightarrow\quad \det(x_i^j)_{i,j=1,\ldots,n} = 0.$$

Satz 2.19 (Tschebyscheffscher Alternantensatz): *Für den Teilraum $S \subset C[a,b]$ mit $\dim S = n$ sei die Haarsche Bedingung (H) erfüllt. Dann ist die Tschebyscheff-Approximation $g \in S$, einer Funktion $f \in C[a,b]$ durch folgende Eigenschaft charakterisiert:*

[17] Alfréd Haar (1885-1933): Ungarischer Mathematiker; Prof. in Kolozsvár (Cluj), Budapest und Szeged; viele wichtige Beiträge zur Approximationstheorie ("Haarsche Bedingung") und Analysis auf Gruppen ("Haar measure").

(A) Es existieren $m \geq n+1$ *Stellen* $a \leq x_0 < \ldots < x_m \leq b$ *(sog. "Alternante", so daß für die Fehlerfunktion* $e(x) = f(x) - g(x)$ *gilt:*

$$|e(x_i)| = \|e\|_\infty\,, \quad e\,(x_i) = -e\,(x_{i+1})\,; \quad i = 1, \ldots, n\,. \tag{2.6.50}$$

Insbesondere ist die beste Approximation eindeutig bestimmt.

Beweis: Für den nicht trivialen Beweis der Alternantenaussage verweisen wir auf die Literatur; z.B. [20]. Q.E.D.

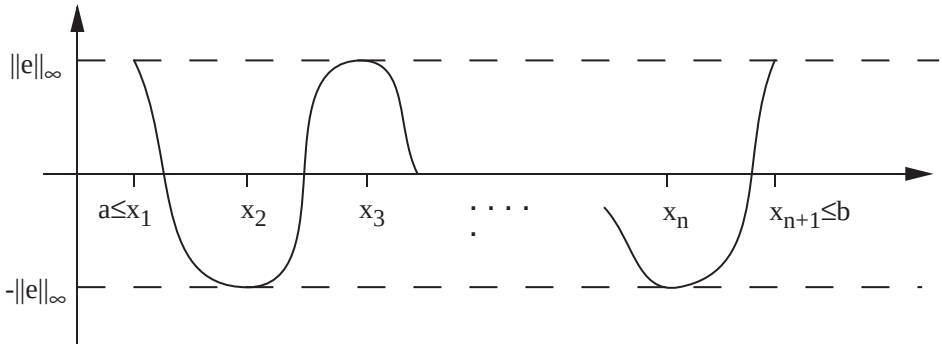


Abbildung 2.9: *Schema der Alternantenregel*

Korollar 2.2: *Der Alternantensatzes impliziert, daß die beste Approximation für den Spezialfall* $S = P_{n-1}$ *eindeutig bestimmt ist.*

Beweis: Seien g_1, g_2 zwei beste Approximationen mit $e_1 = f - g_1\,,\ e_2 = f - g_2\,.$ Für $\lambda \in (0,1)$ ist dann $\|\lambda e_2\|_\infty < \|e_1\|_\infty\,,$ so daß der Graph von $\lambda e_2(x)$ den von $e_1(x)$ mindestens n-mal schneidet (siehe Abbildung 2.9). Jede der Funktionen $\varphi_\lambda(x) = e_1(x) - \lambda e_2(x)$ hat also mindestens n Nullstellen. Durch Grenzübergang $\lambda \to 1$ folgt, daß $\varphi_1(x) = e_1(x) - e_2(x) = g_2(x) - g_1(x)$ mindestens n (ihrer Vielfachheit entsprechend oft gezählte) Nullstellen besitzt. Wegen $g_2 - g_1 \in P_{n-1}$ ergibt sich zwangsläufig $g_1 \equiv g_2\,.$ Q.E.D.

Bemerkung 2.4: Der Alternantensatz ist die Grundlage des sog. "Remez[18]-Algorithmus" zur Konstruktion der Tschebyscheff-Approximation. Wäre eine Alternante $\{x_1, \ldots, x_{n+1}\}$ bekannt, so könnte man bei gegebener Basis $\{\varphi_1, \ldots, \varphi_n\}$ von S die beste Approximation

$$g = \sum_{i=1}^{n} \alpha_i \varphi_i$$

[18]Evgeny Yakovlevich Remez (1896-1975): Russischer Mathematiker; Professor an der Universität Kiew (1935); Beiträge zur konstruktiven Approximationstheorie ("Remez-Algorithmus") und zur numerischen Lösung von Differentialgleichungen.

sowie die Größe $\alpha_{n+1} := \sigma\|f-g\|_\infty$, $\sigma \in \{-1,1\}$, aus dem linearen Gleichungssystem

$$\sum_{i=1}^{n} \alpha_i \varphi_i(x_k) + (-1)^k \alpha_{n+1} = f(x_k), \quad k = 1,\ldots,n+1,$$

berechnen. Der Remez-Algorithmus besteht aus der systematischen iterativen Suche nach der Alternante $\{x_1,\ldots,x_{n+1}\}$. In jedem Schritt wird mit der Näherung $\{x_i^{(t)},\ldots,x_{n+1}^{(t)}\}$ das Gleichungssystem für $\alpha_1^{(t)},\ldots,\alpha_n^{(t)}$ und $\alpha_{n+1}^{(t)}$ gelöst und für

$$g(t) = \sum_{i=1}^{n} \alpha_i^{(t)} \varphi_i$$

das Optimalitätskriterium

$$\|f - g^{(t)}\|_\infty = |\alpha_{n+1}^{(t)}|$$

abgefragt. Im allg. konvergiert $\{x_1^{(t)},\ldots,x_{n+1}^{(t)}\}$ gegen $\{x_1,\ldots,x_{n+1}\}$, allerdings nicht in endlich vielen Schritten.

Beispiel 2.11: $f(x) = \cos(x)$, $[a,b] = [0,\pi/2]$, $S = P_1$

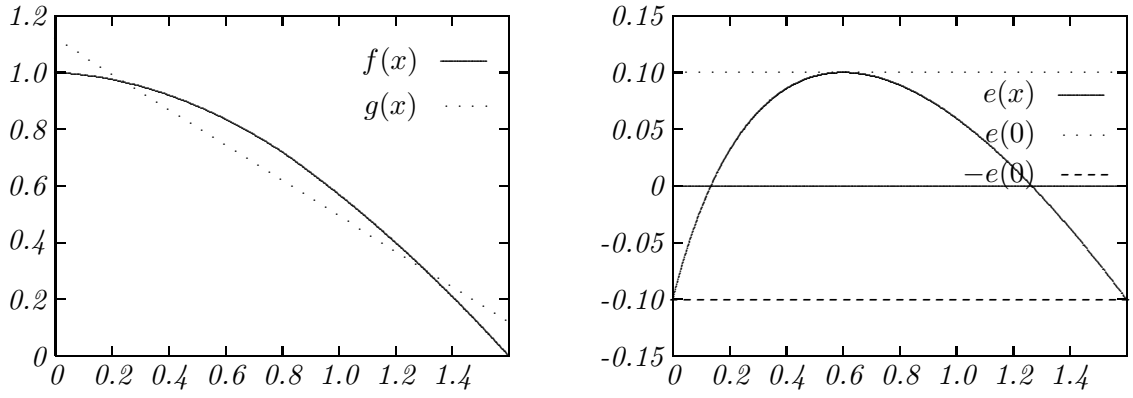


Abbildung 2.10: Anwendung der Alternantenregel

2.6.1 "Optimale" Lagrange-Interpolation

Zur Anwendung der Tschebyscheff-Approximation stellen wir die Frage nach der "optimalen" Wahl der Stützstellen bei der Lagrange-Interpolation. Für das Lagrangesche Interpolationspolynom $p \in P_n$ einer Funktion $f \in C^{n+1}[a,b]$ zu den Stützstellen $a \le x_0 < x_1 < \ldots x_n \le b$ gilt nach Satz 2.3 die Fehlerdarstellung

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\zeta_x) L(x), \quad x \in [x_0,\ldots,x_n],$$

wobei $L(x) = \prod_{j=0}^{n}(x - x_j)$. Die Stützstellen $x_0, \ldots, x_n$ solle so bestimmt werden, daß

$$\max_{a \le x \le b} |L(x)| = \|L\|_\infty \tag{2.6.51}$$

minimal wird. Damit hätte man eine "optimale" Darstellbarkeit von Funktionen aus $C^{n+1}[a,b]$ durch Lagrangesche Interpolationspolynome in P_n. Nun ist $L(x) = x^{n+1} - p$ mit einem $p \in P_n$, d. h.: Die Aufgabe "optimale" Stützstellen zu bestimmen, ist äquivalent zur Konstruktion der Tschebyscheff-Approximation zu $f(x) = x^{n+1}$ bzgl. $S = P_n$. Nach dem Alternantensatz hat die Fehlerfunktion $e = x^{n+1} - p$ mindestens $n+1$ Nullstellen im Intervall $[a,b]$.

Satz 2.20 (Optimale Stützstellen): *Auf dem Intervall $[a,b] = [-1,1]$ ist die Tschebyscheff-Approximation $g \in P_n$ zu $f(x) = x^{n+1}$ gegeben durch*

$$g(x) = x^{n+1} - 2^{-n} T_{n+1}(x) \tag{2.6.52}$$

mit dem $(n+1)$-ten Tschebyscheff-Polynom

$$T_{n+1}(x) = \cos[(n+1)\arccos(x)].$$

Die Nullstellen

$$x_k = \cos\Big(\frac{\pi}{2}\,\frac{2k+1}{n+1}\Big), \quad k = 0, \ldots, n \tag{2.6.53}$$

von T_{n+1} sind gerade die "optimalen" Stützstellen der Lagrange-Interpolation auf $[-1,1]$.

Beweis: Das Polynom $T_{n+1} \in P_{n+1}$ hat die $n+1$ Nullstellen $x_k = \cos(\frac{\pi}{2}\frac{2k+1}{n+1})$, $k = 0, \ldots, n$, und es gilt (Übungsaufgabe)

$$\max_{-1 \le x \le 1} \prod_{k=0}^{n} |x - x_k| = 2^{-n}.$$

Der rekursiven Beziehung

$$T_0 \equiv 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$$

entnimmt man, daß $T_{n+1}(x) = 2^n x^{n+1} + q(x)$ mit einem $q \in P_n$. Also ist

$$2^{-n} T_{n+1}(x) = \prod_{k=0}^{n} (x - x_k) = L(x).$$

Weiter nimmt $T_{n+1}(x) = \cos((n+1)\arccos(x))$ im Intervall $[-1,1]$ offenbar genau $(n+2)$-mal einen Extremwert an, abwechselnd ± 1 (siehe Abb. 2.11). Diese $n+2$ Extremalstellen bilden dann eine Alternante für die Approximation $g(x) = x^{n+1} - 2^{-n}T_{n+1}(x) \in P_n$ zu x^{n+1}. Nach dem Alternantensatz ist g also die eindeutig beste Approximation zu x^{n+1}. Mit den Nullstellen x_k von $T_{n+1}(x)$ gilt folglich

$$\begin{aligned}\max_{-1\le x\le 1} |\prod_{k=0}^{n}(x-x_k)| &= 2^{-n} \max_{-1\le x\le 1} |T_{n+1}(x)| \\ &= \max_{-1\le x\le 1} |x^{n+1} - [x^{n+1} - 2^{-n}T_{n+1}(x)]| \\ &= \min_{p\in P_n} \max_{-1\le x\le 1} |x^{n+1} - p(x)| \\ &= \min_{-1\le\zeta_0<\ldots<\zeta_n\le 1} \max_{-1\le x\le 1} |\prod_{k=0}^{n}(x-\zeta_k)|\end{aligned}$$

und damit die behauptete Optimalitätseigenschaft. Q.E.D.

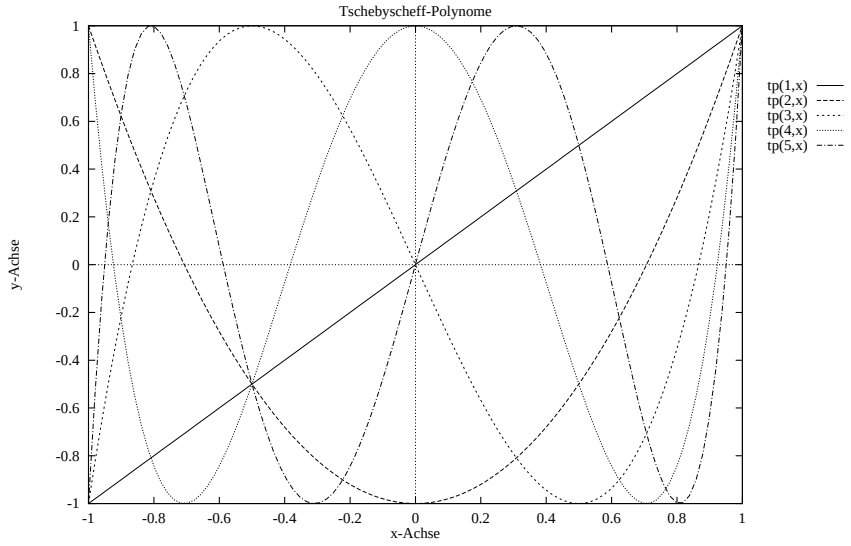


Abbildung 2.11: *Anwendung der Alternantenregel*

Die optimalen Stützstellen der Lagrange-Interpolation auf einem beliebigen Intervall $[a,b]$ gewinnt man aus dem Resultat von Satz 2.20 mit Hilfe der Transformation $\varphi : [a,b] \to [-1,1]$:

$$x = \varphi(y) := \frac{2}{b-a}y - \frac{a+b}{b-a},$$

zu

$$y_k = \frac{b-a}{2}x_k + \frac{a+b}{2}, \quad k = 0,\ldots,n.$$

Für den Interpolationsfehler einer Funktion $f \in C^{n+1}[a,b]$ gilt dann

$$\|f - p_n\|_\infty \le \frac{1}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}\|f^{(n+1)}\|_\infty. \tag{2.6.54}$$

2.7 Übungsaufgaben

Aufgabe 2.7.1: Für verschiedene Orte wurde an einem bestimmten Tag die Tageslänge gemessen:

Ort	Tageslänge	Lage
A	17h 28m	$55,7^o$
B	18h 00m	$57,7^o$
C	18h 31m	$59,3^o$
D	19h 56m	$62,6^o$
E	22h 34m	$65,6^o$

Man bestimme die Tageslänge am Ort F bei $61,7^o$ durch Auswertung des zugehörigen Interpolationspolynoms mit Hilfe des Neville-Algorithmus. (Es genügt 4-stellige Dezimalrechnung.)

Aufgabe 2.7.2: Es soll eine 10-stellige Wertetabelle von

$$f(x) = \int_0^x \sin(t)^2\, dt, \quad x \in [0, \pi],$$

erstellt werden (in Festkommadarstellung), so daß kubische Lagrange-Interpolation einen Fehler kleiner als $5 \cdot 10^{-9}$ für jeden Wert von x im Intervall $[0, \pi]$ ergibt. Reichen dazu die Werte zu 250 äquidistant verteilten Stützstellen aus? (Hinweis: Der Auswertungsfehler setzt sich zusammen aus dem absoluten Interpolationsfehler und dem absoluten Rundungsfehler in den zur Interpolation verwendeten Stützwerten.)

Aufgabe 2.7.3: Gegeben sei die Funktion $f(x) = e^{\lambda x},\ \lambda \in \mathbb{R}$, auf einem Intervall $[a, b]$. Man zeige, daß in diesem Fall der Fehler $f - p_n$ der Lagrange-Interpolation von f über beliebig verteilten $n+1$ (paarweise verschiedenen) Stützstellen aus $[a, b]$ für $n \to \infty$ gleichmäßig gegen Null konvergiert:

$$\max_{x \in [a,b]} |f(x) - p_n(x)| \to 0 \quad (n \to \infty).$$

Was unterscheidet diese Funktion von dem in der Vorlesung angegebenen Beispiel $f(x) = (1 + x^2)^{-1}$, für welches die Lagrange-Interpolation für $n \to \infty$ nicht konvergiert (s. Aufgabe 3.5)?

Aufgabe 2.7.4: Wir betrachten die Hermitesche Interpolationsaufgabe, zu (paarweise verschiedenen) Stützstellen x_i $(i = 0, \ldots, m)$ und zu gegebenen Werten $y_i^{(0)}$, $y_i^{(1)}$ $(i = 0, \ldots, m)$ ein Polynom $p \in P_n$, $n = 2m + 1$, so zu bestimmen, daß

$$p(x_i) = y_i^{(0)}, \quad p'(x_i) = y_i^{(1)} \quad (i = 0, \ldots, m).$$

Man zeige:

a) Die Hermitesche Interpolationsaufgabe besitzt eine eindeutig bestimmte Lösung.

b) Im Falle der Interpolation einer $(n+1)$-mal differenzierbaren Funktion f, d.h. $p(x_i) = f(x_i),\ p'(x_i) = f'(x_i)\ (i = 0, \ldots, m)$ gilt die Fehlerdarstellung

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{m} (x - x_i)^2$$

mit x-abhängigen Zwischenstellen $\xi_x \in [a, b]$. (Hinweis: Man modifiziere die entsprechende Argumentation der Vorlesung oder des Skriptums für die Lagrange-Interpolation.)

Aufgabe 2.7.5: (Praktische Aufgabe) Man berechne die Lagrangeschen Interpolationspolynome der Funktionen

$$f(x) = \frac{1}{1 + 25x^2}, \qquad g(x) = \sqrt{|x|}, \qquad -1 \le x \le 1,$$

in Nevillescher Darstellung, jeweils zu den Stützstellen $x_i = -1 + ih,\ i = 0, ..., n$, mit $h = 2/n$, für $n = 5, 10, 15, 20$. Man stelle die Polynome graphisch dar und vergleiche die Polynomgraphen mit den richtigen Funktionsverläufen.

Aufgabe 2.7.6: Auf einer Zerlegung $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$ des Intervalls $I = [0, 1]$ werde eine Funktion stückweise quintisch interpoliert, d.h.: Auf jedem der Teilintervalle $I_k = [x_{k-1}, x_k],\ k = 1, \ldots, N$, wird f durch ein Polynom $p_5^{(k)} \in P_5$ interpoliert, wobei jeweils einer der folgenden Sätze an Bedingungen verwendet wird:

(i) $\quad p_5^{(k)}(\xi) = f(\xi), \quad \xi \in \{x_{k-1} + jh_k/5,\ j = 0, \ldots, 5\}, \quad h_k = x_k - x_{k-1},$

(ii) $\quad p_5^{(k)}(\xi) = f(\xi)\,,\ (p_5^{(k)})'(\xi) = f'(\xi)\,,\ (p_5^{(k)})''(\xi) = f''(\xi), \quad \xi \in \{x_{k-1}, x_k\}\,.$

Im ersten Fall (Lagrange-Interpolation) ist die resultierende zusammengesetzte Funktion auf dem Intervall I stetig, und im zweiten Fall (Hermite-Interpolation) zweimal stetig differenzierbar. Man zeige, daß für $f \in C^6[0, 1]$ in beiden Fällen die Abschätzung gilt:

$$\max_{x \in I} |f(x) - p(x)| \le \frac{h^6}{720} \max_{x \in I} |f^{(6)}(x)|, \quad h := \max_{k=1,\ldots,N} h_k\,.$$

(Hinweis: Man wende die Fehlerdarstellung der Vorlesung für die Lagrange- bzw. die Hermite-Interpolation auf jedem der Teilintervalle I_k an.)

Aufgabe 2.7.7: Für die Funktion $f(x) = \cosh(x)$ ist die Wertetabelle gegeben

x	$f(x)$
0.52	1, 1382741
0.56	1, 1609408
0.60	1, 1854652
0.64	1, 2118867
0.68	1, 2402474.

Man bestimme durch Extrapolation eines geeigneten Differenzenquotienten möglichst gute Näherungen zum Ableitungswert $f'(0.6) = 0, 63665358...$.

Aufgabe 2.7.8: Welche von den Indexfolgen

$$(i) \quad n_i = 2i - 1,\ i \in \mathbb{N}, \qquad (ii) \quad n_i = 3^i,\ i \in \mathbb{N}, \qquad (iii) \quad n_i = i^2,\ i \in \mathbb{N},$$

für Schrittweiten $h_i = h/n_i$ ist zulässig für die Extrapolation zum Limes?

Aufgabe 2.7.9: Es bezeichne S_0 den Vektorraum der kubischen, natürlichen Spline-Funktionen zu den Stützstellen $x_0 = 0, x_1 = 1, x_2 = 2$.

a) Sind die folgenden Funktionen in S_0 ?

$$\begin{aligned} (i) \quad & f(x) = x^3 - x^2, \\ (ii) \quad & f(x) = x^2(x-6) - (x-2)^3, \\ (iii) \quad & f(x) = \max\{0, x-1\}^3 - \tfrac{1}{2}x^3. \end{aligned}$$

b) Man bestimme den interpolierenden Spline $s_2 \in S_0$ für $f(x) = x^3$. Wie lautet das Ergebnis, wenn die *natürlichen* Randbedingungen durch $s_2''(x_0) = f''(x_0)$, und $s_2''(x_2) = f''(x_2)$ ersetzt werden.

Aufgabe 2.7.10: (Praktische Aufgabe) Zur näherungsweisen Bestimmung des halben Umfangs

$$\pi = 3, 1415\,92653\,58979\,32384\,62643\ldots$$

des Einheitskreises verwendeten schon die "alten Griechen" den Umfang einbeschriebener regulärer Polygone. Mit Hilfe des 96-seitigen Polygons fand z.B. Archimedes den Wert $\pi \approx 3, 142$. Die allgemeine Formel für den Umfang T_n des n-seitigen einbeschriebenen, regulären Polygons ist

$$(*) \qquad T_n = 2n \sin(\pi/n) \quad \to \quad 2\pi \quad (n \to \infty).$$

Man kann die T_n für $n = 6 \cdot 2^i$ mit Hilfe der Rekursionsformel (nachprüfen!)

$$T_6 = 6, \quad T_{2n} = 2\sqrt{2n^2 - n\sqrt{4n^2 - T_n^2}},$$

ohne Auswertung des Sinus berechnen.

i) Man bestimme mit Hilfe der Richardson-Extrapolation aus den Stützwerten

$$\{T_n,\ n = 6 \cdot 2^i,\ i = 0, \ldots, k\},$$

für $k = 1, \ldots, 30$, Approximationen für π und plotte den resultierenden Fehler in Abhängigkeit von k. (Hinweis: Man setze $x_n = 1/n$ und $T(x_n) := T_n$ und extrapoliere die Funktion $T(x) := 2/x \sin(\pi x)$ nach $x = 0$.)

ii) Man wiederhole die Rechnung mit den Stützwerten $\{T_i,\ i = 3, 4, \ldots, k\}$ für $k = 4, \ldots, 33$, wobei die benötigten Werte T_i direkt aus der Definitionsformel $(*)$ bestimmt werden sollen.

Wie sind die beobachteten Phänomene zu erklären?

Übung 2.1: Man zeige, daß die Funktionen

$$\varphi_0(x) = 1/\sqrt{2x}, \quad \varphi_k(x) = \frac{1}{\sqrt{\pi}} \cos(kx), \quad \psi_k(x) = \frac{1}{\sqrt{\pi}} \sin(kx), \quad k = 1, \ldots, n,$$

ein Orthonormalsystem des Teilraums $T_n \subset C[-\pi, \pi]$ der trigonometrischen Polynome vom Grad kleiner oder gleich n bzgl. des L^2-Skalarprodukts über dem Intervall $[-\pi, \pi]$ bilden und bestimme die beste Approximation der Funktion $f(x) = x$ in T_n.

Übung 2.2: Man zeige, daß die durch

$$\varphi_k(x) = \frac{k!}{(2k)!} \frac{d^k}{dx^k}(x^2 - 1)^k, \quad k = 0, 1, \ldots, n,$$

definierten Polynome orthogonal bzgl. des L^2-Skalarprodukts über $[-1, 1]$ sind, und daß

$$\|\varphi_k\| = \frac{k!^2}{(2k)!} \sqrt{\frac{2^{2k+1}}{2k+1}}, \qquad \varphi_k(1) = \frac{k!^2}{(2k)!} 2^k.$$

Durch Normierung erhält man hieraus die sog. Gauß-Legendre-Polynome

$$L_k(x) := \frac{(2k)!}{k!^2 2^k} \varphi_k(x), \qquad L(1) = 1.$$

(Hinweis: Man verwende partielle Integration, leite für die Integrale $I_k := \int_{-1}^{1} (1 - x^2)^k \, dx$ eine einstufige Rekursionsformel her und differenziere die Funktion $(x^2 - 1)^k$.)

Übung 2.3: Man zeige die folgenden Eigenschaften einer Norm $\|\cdot\|$ auf einem Vektorraum E (nicht notwendig endlich dimensional):

a) $\|x - y\| \geq \big|\|x\| - \|y\|\big|, \quad x, y \in E$.

b) Die Funktion $N(\cdot) = \|\cdot\| : E \to \mathbb{R}$ ist stetig.

c) Ist E endlich dimensional, so sind alle Normen auf E äquivalent, d.h.: Zu je zwei Normen $\|\cdot\|_1, \|\cdot\|_2$ auf E gibt es stets Konstanten $M \geq m > 0$, so daß

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1, \quad x \in E.$$

(Hinweis: Für eine Basis $\{e^1, \ldots, e^n\}$ von E betrachte man die Funktion $F(\alpha_1, \ldots, \alpha_n) := \|\sum_{i=1}^n \alpha_i e^i\|$ auf der Einheitssphäre des $\mathbb{R}^n$.)

Übung 2.4: Man bestimme die Gauß-Approximationen der Funktion $f(x) = \sqrt{x}$ bzgl. der L^2-Norm über dem Intervall $[0,1]$ in den Polynomräumen P_0, P_1 und P_2. Man stelle die Ergebnisse graphisch dar.

Übung 2.5: (Praktische Aufgabe) Man berechne rekursiv die Gauß-Legendre- und die Tschebyscheff-Polynome L_k und T_k auf dem Intervall $[-1,1]$ aus den Beziehungen

$$a) \qquad L_k(x) := \frac{(2k)!}{2^k k!^2}\,\varphi_k(x), \quad k = 0,1,2,\ldots,$$
$$\varphi_0(x) = 1,\ \varphi_1(x) = x, \quad \varphi_{k+1}(x) = x\varphi_k(x) - \frac{k^2}{4k^2-1}\varphi_{k-1},$$

$$b) \qquad T_0(x) = 1,\ T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x),$$

für $k = 0,1,2,\ldots,10$ und stelle sie graphisch dar.

3 Numerische Integration

Die Berechnung bestimmter Integrale kann in der Praxis meist nur näherungsweise mit Hilfe von sog. "Quadraturformeln" erfolgen. Dazu macht man für eine Funktion $f \in C[a,b]$ den Ansatz

$$I(f) = \int_a^b f(x)\,dx \;\sim\; I^{(n)}(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$$

mit Stützstellen $a \le x_0 < \ldots < x_n \le b$ und Gewichten $\alpha_i \in \mathbb{R}$. Ein einfaches Beispiel ist die sog. "(summierte) Rechteckregel":

$$I(f) \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i).$$

3.1 Interpolatorische Quadraturformeln

Ein naheliegender Weg zur Konstruktion von Quadraturformeln ist der über die Polynominterpolation. Zu den (paarweise verschiedenen) Stützstellen $a \le x_0 < \ldots < x_n \le b$ wird das Lagrangesche Interpolationspolynom gebildet

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i^{(n)}(x)$$

und dann gesetzt

$$I^{(n)}(f) := \int_a^b p_n(x)\,dx = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b L_i^{(n)}(x)\,dx}_{\equiv \alpha_i}. \tag{3.1.1}$$

Die Quadraturgewichte α_i hängen offenbar nur von $[a,b]$ und den Stützstellen $x_0, \ldots, x_n$ ab. Der Quadraturfehler einer solchen sog. "interpolatorischen" Quadraturformel läßt sich leicht angeben:

Satz 3.1 (Lagrange-Quadratur): *Für interpolatorische Quadraturformeln gilt:*

$$I(f) - I^{(n)}(f) = \int_a^b f[x_0, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j) dx. \tag{3.1.2}$$

Beweis: Die allgemeine Darstellung des Quadraturfehlers folgt aus der Restglieddarstellung der Interpolation in Satz 2.4. Q.E.D.

Aus der Fehlerdarstellung (3.1.2) folgt, daß die interpolatorische Quadraturformel $I^{(n)}(\cdot)$ "exakt" ist für Polynome $p \in P_n$; dies ergibt sich ja bereits aus ihrer Konstruktion.

Definition 3.1: *Eine Quadraturformel $I^{(n)}(\cdot)$ wird "(mindestens) von der Ordnung m" genannt, wenn durch sie wenigstens alle Polynome aus P_{m-1} exakt integriert werden.*

Die interpolatorischen Quadraturformeln $I^{(n)}(\cdot)$ zu $n+1$ Stützstellen sind also mindestens von der Ordnung $n+1$.

Ein wichtiger Spezialfall sind die auf äquidistant verteilten Stützstellen basierenden sog. "Newton-Cotes[1]-Quadraturformeln":

(a) "abgeschlossene" Newton-Cotes-Formeln (a, b sind Stützstellen)

$$x_i = a + iH\,, \quad i = 0, \ldots, n\,, \quad H = \frac{b-a}{n}\,,$$

(b) "offene" Newton-Cotes-Formeln (a, b sind keine Stützstellen)

$$x_i = a + (i+1)H\,, \quad i = 0, \ldots, n\,, \quad H = \frac{b-a}{n+2}\,,$$

Zur Berechnung der Gewichte α_i geht man z.B. im Fall der abgeschlossenen Formeln wie folgt vor: Jedes $x \in [a, b]$ ist darstellbar als $x = a + tH$ mit einem $t \in [0, n]$. Durch Koordinatentransformation $x \to t = (x-a)/H$ erhält man

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{a + tH - a - jH}{a + iH - a - jH} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t-j}{i-j}.$$

Also ist

$$\alpha_i = \int_a^b L_i^{(n)}(x)\, dx = H \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t-j}{i-j}\, dt\,, \quad i = 0, \ldots, n.$$

Diese Gewichte werden ein für allemal berechnet und tabelliert. Für die offenen Newton-Cotes-Formeln verfährt man analog.

[1]Roger Cotes (1682-1716): Englischer Mathematiker; Professor für Astronomy und Experimentalphilosophie an der Universität Cambridge (1706) (zusammen mit Newton); Beiträge zu vielen konkreten Fragen der reellen Analysis, insbesondere zur Numerik, Intzerpolation und Integraltafelnberechnung).

Beispiel 3.1: Als abgeschlossene Newton-Cotes-Formel für $n = 2,\ H := (b-a)/2$:

$$a_0 = H \int_0^2 \frac{t-1}{0-1}\frac{t-2}{0-2}\,dt = \frac{H}{2}\int_0^2 (t^2-3t+2)\,dt = \frac{1}{3}H$$
$$a_1 = H \int_0^2 \frac{t-0}{1-0}\frac{t-2}{1-2}\,dt = -H\int_0^2 (t^2-2t)\,dt = \frac{4}{3}H$$
$$a_2 = H \int_0^2 \frac{t-0}{2-0}\frac{t-1}{2-1}\,dt = \frac{H}{2}\int_0^2 (t^2-t)\,dt = \frac{1}{3}H$$

ergibt sich die sog "Simpson[2]-Regel":

$$I^{(2)}(f) = \frac{H}{3}\Big(f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big).$$

Beispiel 3.2: Wir geben im folgenden einige der einfachsten Newton-Cotes-Formeln an:

(a) Abgeschlossene Formeln $(n = 1,2,3,4)$: $\quad H := (b-a)/n$

$$I^{(1)}(f) = \frac{b-a}{2}\big\{f(a)+f(b)\big\} \quad (\text{"Trapezregel" bzw. "Sehnen-Trapezregel"})$$
$$I^{(2)}(f) = \frac{b-a}{6}\Big\{f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big\} \quad (\text{"Simpson-Regel"})$$
$$I^{(3)}(f) = \frac{b-a}{8}\big\{f(a)+3f(a+H)+3f(b-H)+f(b)\big\} \quad (\text{"}\tfrac{3}{8}\text{-Regel)"}$$
$$I^{(4)}(f) = \frac{b-a}{90}\Big\{7f(a)+32f(a+H)+12f\Big(\frac{a+b}{2}\Big)+32f(b-H)+7f(b)\Big\}.$$

(b) Offene Formeln $(n = 0,1,2,3) \quad H := (b-a)/n$

$$I^{(0)}(f) = (b-a)f\Big(\frac{a+b}{2}\Big) \quad (\text{"Mittelpunktregel" bzw. "Tangenten-Trapezregel"})$$
$$I^{(1)}(f) = \frac{b-a}{2}\big\{f(a+H)+f(b-H)\big\}$$
$$I^{(2)}(f) = \frac{b-a}{3}\Big\{2f(a+H) - f\Big(\frac{a+b}{2}\Big) + 2f(b-H)\Big\}$$
$$I^{(3)}(f) = \frac{b-a}{24}\big\{11f(a+H)+f(a+2H)+f(b-2H)+11f(b-H)\big\}.$$

[2]Thomas Simpson (1710-1761): Englischer Mathematiker; seit 1743 Professor an der Royal Military Academny at Woolwich; neben der nach ihm benannten Quadraturformel Beiträge zur Geometrie, Trogonometrie, Wahrscheinlichkeitstheorie und Astronomie; auf ihn gehen die heutige üblichen Bezeichnungen Sinus, Cosinus, Tangens und Cotangens zurück, auch die differentielle Form Newton-Verfahrens wurde von ihm 1740 eingeführt.

Bemerkung 3.1: Im Gegensatz zu den Newton-Cotes-Formeln verwenden die sog. "Besselschen[3] Formeln" auch Stützstellen außerhalb von $[a,b]$; z.B.:

$$I^{(3)}(f) = \frac{b-a}{24}\big\{ -f(2a-b) + 13f(a) + 13f(b) - f(2b-a)\big\}.$$

Die sog. "Hermiteschen Formeln" verwenden Ableitungswerte; z.B.:

$$I^{(3)}(f) = \frac{b-a}{2}\big\{f(a)+f(b)\big\} + \frac{(b-a)^2}{12}\big\{f'(a)-f'(b)\big\}.$$

Sie basieren auf dem Hermiteschen Interpolationspolynom zu $n+1 = 2m+2$ Stützstellen $a \le x_0 < \ldots < x_m \le b$ und Stützwerten $f(x_i), f'(x_i),\ i = 0,\ldots,m$.

Wir wollen nun für die drei einfachsten Newton-Cotes-Formeln, die Trapezregel, die Simpson-Regel und die Mittelpunktregel, die Restglieddarstellungen ableiten.

Satz 3.2 (Quadraturrestglieder): *Es gelten die folgenden Restglieddarstellungen*
(i) *für die Trapezregel:*

$$I(f) - \frac{b-a}{2}\big\{f(a)+f(b)\big\} = -\frac{(b-a)^3}{12}\,f''(\zeta), \quad f \in C^2[a,b]\,,$$

(ii) *für die Simpson-Regel:*

$$I(f) - \frac{b-a}{6}\Big\{f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big\} = -\frac{(b-a)^5}{2880}\,f^{(4)}(\zeta)\,, \quad f \in C^4[a,b],$$

(iii) *für die Mittelpunktregel:*

$$I(f) - (b-a)f\Big(\frac{a+b}{2}\Big) = \frac{(b-a)^3}{24}\,f''(\zeta), \quad f \in C^2[a,b]\,,$$

mit gewissen Zwischenstellen $\zeta \in [a,b]$.

Beweis: (i) Wegen $(x-a)(x-b) \le 0$ in $[a,b]$ gilt

$$I(f) - I^{(1)}(f) = \frac{f''(\zeta)}{2}\int_a^b (x-a)(x-b)\,dx = -\frac{(b-a)^3}{12}\,f''(\zeta).$$

(ii) Da $(x-a)(x-\frac{a+b}{2})(x-b)$ in $[a,b]$ einen Vorzeichenwechsel hat, kann der in (i) verwendete trick nicht direkt angewendet werden. Mit der Newtonschen Form des Inter-

[3]Friedrich Wilhelm Bessel (1784-1846): deutscher Astromom und Mathematiker; Direktor des Observatoriums in Königsberg und Mitglied der Berliner Akademie; grundlegende Beiträge zur mathematischen Fehlerkorrektur bei astronomischen Beobachtungen und zur Sternpositionierung.

polationsrestglieds gilt

$$\begin{aligned} I(f) - I^{(2)}(f) &= \int_a^b f[a, \tfrac{a+b}{2}, b, x](x-a)(x-\tfrac{a+b}{2})(x-b)\,dx \\ &= \int_a^b \frac{f[a, \frac{a+b}{2}, b, x] - f[a, \frac{a+b}{2}, b, \frac{a+b}{2}]}{x - \frac{a+b}{2}} \underbrace{(x-a)(x-\tfrac{a+b}{2})^2(x-b)}_{\leq 0}\,dx + \\ &+ f[a, \tfrac{a+b}{2}, b, \tfrac{a+b}{2}] \underbrace{\int_a^b (x-a)(x-\tfrac{a+b}{2})(x-b)\,dx}_{=0}. \end{aligned}$$

Nach dem verallgemeinerten Mittelwertsatz der Integralrechnung folgt

$$\begin{aligned} I(f) - I^{(2)}(f) &= \int_a^b f[a, \tfrac{a+b}{2}, \tfrac{a+b}{2}, b, x](x-a)(x-\tfrac{a+b}{2})^2(x-b)\,dx \\ &= \frac{f^{(4)}(\zeta)}{4!} \int_a^b (x-a)(x-\tfrac{a+b}{2})^2(x-b)\,dx. \end{aligned}$$

(iii) Da $x - \frac{a+b}{2}$ in $[a,b]$ einen Vorzeichenwechsel hat, verwenden wir eine analoge Schlußweise wie in (ii):

$$\begin{aligned} I(f) - I^{(0)}(f) &= \int_a^b f[\tfrac{a+b}{2}, x](x - \tfrac{a+b}{2})\,dx \\ &= \int_a^b \frac{f[\frac{a+b}{2}, x] - f[\frac{a+b}{2}, \frac{a+b}{2}]}{x - \frac{a+b}{2}} \underbrace{(x - \tfrac{a+b}{2})^2}_{\geq 0}\,dx + f'(\tfrac{a+b}{2}) \underbrace{\int_a^b (x - \tfrac{a+b}{2})\,dx}_{=0} \\ &= \int_a^b f[\tfrac{a+b}{2}, \tfrac{a+b}{2}, x](x - \tfrac{a+b}{2})^2\,dx = \frac{f''(\zeta)}{2} \int_a^b (x - \tfrac{a+b}{2})^2\,dx. \end{aligned}$$

Analog lassen sich die Restglieddarstellungen der Newton-Cotes-Formeln höherer Ordnung herleiten. Q.E.D.

Bemerkung 3.2: *Besitzen die in den Restgliedern auftretenden Ableitungen von f auf $[a,b]$ festes Vorzeichen, so gestattet der Vergleich der abgeschlossenen und offenen Formeln (unter Vernachlässigung des Rundungsfehlers) eine Einschließung des Integralwertes. Zum Beispiel ergibt sich für "konvexe" Funktionen f ($f'' \geq 0$) mit der (Sehnen)-Trapezregel und der (Tangenten)-Trapezregel (Mittelpunktsregel):*

$$(b-a) f\Big(\frac{a+b}{2}\Big) \leq I(f) \leq \frac{b-a}{2} \big\{ f(a) + f(b) \big\}. \tag{3.1.3}$$

Bei den abgeschlossenen Newton-Cotes-Formeln treten ab $n = 7$ und bei den offenen ab $n = 2$ *negative* Gewichte α_i auf. Dadurch erhöht sich die Rundungsfehleranfälligkeit dieser Formeln (Auslöschungsgefahr). Außerdem kann i.Allg. keine Konvergenz

$$I^{(n)}(f) \to I(f) \quad (n \to \infty)$$

erwartet werden, da die Lagrange-Interpolation kein generell konvergenter Prozeß ist. Man wendet daher zur Berechnung von $I(f)$ die Quadraturformeln nur auf Teilintervalle der Länge h an und summiert die Einzelbeiträge zu den sog. "summierten" Quadraturformeln.

$$I_h^{(n)}(f) := \sum_{i=0}^{N-1} I^{(n)}_{[x_i,x_{i+1}]}(f), \quad h = \frac{b-a}{N}. \tag{3.1.4}$$

Gilt für die verwendete Quadraturformel die Fehlerdarstellung

$$I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f) = \omega_n \, h^{m+2} f^{(m+1)}(\zeta_i)\,, \quad \zeta_i \in [x_i, x_{i+1}],$$

mit einem $m \geq n$, so ergibt sich mit dem Zwischenwertsatz für den Fehler die Darstellung

$$I(f) - I_h^{(n)}(f) = \sum_{i=0}^{N-1} \omega_n \, h^{m+2} \, f^{(m+1)}(\zeta_i) = \omega_n \, h^{m+2} \, N \, f^{(m+1)}(\zeta),$$

mit einem $\zeta \in [a, b]$. Wegen $N = \frac{b-a}{h}$ folgt also

$$I(f) - I_h^{(n)}(f) = \omega_n (b-a) \, h^{m+1} \, f^{(m+1)}(\zeta)\,, \quad \zeta \in [a, b]\,. \tag{3.1.5}$$

Beispiel 3.3: Wir geben die Restglieder für die einfachsten Formeln an:

(1) Summierte Trapezregel $(m = 1)$

$$I_h^{(1)}(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2} \big\{ f(x_i) + f(x_{i+1}) \big\} = \frac{h}{2} \Big\{ f(a) + 2 \sum_{i=1}^{N-1} f(x_i) + f(b) \Big\}.$$

$$I(f) - I_h^{(1)}(f) = -\frac{b-a}{12} \, h^2 \, f''(\zeta)\,, \quad \zeta \in [a, b]. \tag{3.1.6}$$

(2) Summierte Simpson-Regel $(m = 3)$

$$\begin{aligned} I_h^{(2)}(f) &= \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{6} \Big\{ f(x_i) + 4 f\Big(\frac{x_i + x_{i+1}}{2} \Big) + f(x_{i+1}) \Big\} \\ &= \frac{h}{6} \left\{ f(a) + 2 \sum_{i=1}^{N-1} f(x_i) + 4 \sum_{i=0}^{N-1} f\Big(\frac{x_i + x_{i+1}}{2} \Big) + f(b) \right\}. \end{aligned}$$

$$I(f) - I_h^{(2)}(f) = -\frac{b-a}{2880}\, h^4\, f^{(4)}(\zeta)\,, \quad \zeta \in [a,b]. \tag{3.1.7}$$

(3) Summierte Mittelpunktregel $(m = 1)$

$$I_h^{(0)}(f) = \sum_{i=0}^{N-1} (x_{i+1} - x_i) f\Big(\frac{x_i + x_{i+1}}{2}\Big) = h \sum_{i=0}^{N-1} f\Big(\frac{x_i + x_{i+1}}{2}\Big).$$

$$I(f) - I_h^{(0)}(f) = \frac{b-a}{24}\, h^2\, f''(\zeta), \quad \zeta \in [a,b]. \tag{3.1.8}$$

3.2 Gaußsche Quadraturformeln

Die interpolatorischen Quadraturformeln

$$I^{(n)}(f) = \sum_{i=0}^{n} \alpha_i f(x_i) \tag{3.2.9}$$

zu den Stützstellen $x_0, \ldots, x_n \in [a,b]$ sind nach Konstruktion mindestens von der Ordnung $n+1$, d.h.: Für ihr Restglied gilt:

$$R^{(n)}(p) \equiv I(p) - I^{(n)}(p) = 0, \qquad p \in P_n. \tag{3.2.10}$$

Für den Spezialfall der Newton-Cotes-Formeln mit geradem $n > 0$ haben wir gesehen (Übungsaufgabe), daß sogar Polynome aus P_{n+1} exakt integriert werden. Es stellt sich nun die Frage, die Stützstellen $x_0, \ldots, x_n$ und die Gewichte $\alpha_0, \ldots, \alpha_n$ so zu wählen, daß Polynome möglichst hohen Grades exakt integriert werden.

Hilfssatz 3.1: *Eine obere Grenze für die Ordnung einer Quadraturformel der Art $I^{(n)}(\cdot)$ ist $2n+2$.*

Beweis: Wäre $I^{(n)}(\cdot)$ von höherer Ordnung, d.h. insbesondere also exakt für das Polynom

$$p(x) = \prod_{i=0}^{n} (x - x_i)^2 \in P_{2n+2},$$

so ergäbe sich der Widerspruch

$$0 < \int_a^b p(x)\,dx = I^{(n)}(p) = 0.$$

Q.E.D.

Wir wollen im folgenden zeigen, daß es tatsächlich interpolatorische Quadraturformeln zu $n+1$ Stützstellen gibt, welche die Maximalordnung $2n+2$ haben. Sie heißen "Gaußsche Quadraturformeln". Seien $p_n \in P_n$ und $p_{2n+1} \in P_{2n+1}$ die Lagrangeschen Interpolationspolynome einer Funktion $f \in C[a,b]$ zu den $n+1$ bzw. $2n+2$ Stützstellen $x_0, \ldots, x_n, x_{n+1}, \ldots, x_{2n+1} \in [a,b]$. Für die zugehörigen Quadraturformeln $I^{(n)}(\cdot)$ bzw. $I^{(2n+1)}(\cdot)$ gilt dann

$$\begin{aligned} I(f) - I^{(2n+1)}(f) &= I(f) - \sum_{i=0}^{2n+1} f[x_0, \ldots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j)\,dx \\ &= I(f) - I^{(n)}(f) - \sum_{i=n+1}^{2n+1} f[x_0, \ldots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j)\,dx. \end{aligned}$$

Wir schreiben für $i = n+1, \ldots, 2n+1$:

$$\int_a^b \prod_{j=0}^{i-1} (x - x_j)\, dx = \int_a^b \underbrace{\prod_{j=0}^{n} (x - x_j)}_{\in P_{n+1}} \underbrace{\prod_{j=n+1}^{i-1} (x - x_j)}_{\in P_n}\, dx.$$

Die $n+1$ Polynome

$$\Big\{1,\ \ x - x_{n+1},\ \ (x - x_{n+1})\,(x - x_{n+2}),\ \ldots,\ \prod_{j=n+1}^{2n} (x - x_j)\Big\}$$

bilden eine Basis von P_n. Wählen wir nun die ersten $n+1$ Stützstellen $x_0, \ldots, x_n \in [a, b]$ so, daß

$$\int_a^b \prod_{j=0}^{n} (x - x_j) q(x)\, dx = 0 \quad \forall\, q \in P_n, \tag{3.2.11}$$

so folgt

$$I(f) - I^{(n)}(f) = I(f) - I^{(2n+1)}(f),$$

d.h.: Die interpolatorische Quadraturformel $I^{(n)}(\cdot)$ ist exakt für Polynome aus P_{2n+1}, also von der Ordnung $2n+2$.

Auf dem Funktionenraum $C[a, b]$ verwenden wir im folgenden wieder das übliche L^2-Skalarprodukt und die zugehörige Norm

$$(f, g) := \int_a^b f(x) g(x)\, dx, \quad \|f\| := (f, f)^{1/2}.$$

Die obige Bedingung (3.2.11) besagt dann, daß das Polynom

$$p(x) \equiv \prod_{j=0}^{n} (x - x_j) = x^{n+1} + r(x)\,, \quad r \in P_n,$$

bzgl. des Skalarprodukts $(\cdot, \cdot)$ "orthogonal" zum Teilraum $P_n[a, b] \subset C[a, b]$ sein muß. Zur Konstruktion von p und damit seiner Nullstellen $x_0, \ldots, x_n$ wenden wir das Gram-Schmidt-Verfahren auf die Monombasis $\{1, x, \ldots, x^{n+1}\}$ von $P_{n+1}[a, b]$ an:

$$p_0(x) := 1, \quad k = 1, \ldots, n+1: \quad p_k(x) := x^k - \sum_{j=0}^{k-1} \frac{(x^k, p_j)}{\|p_j\|^2}\, p_j(x). \tag{3.2.12}$$

Dann ist $\{p_0, \ldots, p_{n+1}\}$ ein "Orthogonalsystem" in $P_{n+1}[a, b]$. Offenbar ist

$$p_{n+1}(x) = x^{n+1} + r(x)\,, \quad r \in P_n,$$

so daß wir $p(x) := p_{n+1}(x)$ setzen können. Die $n+1$ Nullstellen $\lambda_0, \ldots, \lambda_n$ von $p(x)$ sind dann mögliche Kandidaten für "optimale" Integrationspunkte.

Wir legen im folgenden ein Skalarprodukt der allgemeineren Gestalt

$$(f,g)_\omega := \int_a^b f(x)g(x)\omega(x)\,dx$$

mit einer integrablen Gewichtsfunktion $\omega(x) \geq 0,\ x \in (a,b)$, mit höchstens endlich vielen Nullstellen in $[a,b]$, zugrunde. Seien dann p_n, $n = 0,1,2,\ldots,$ die mit Hilfe des Gram-Schmidt-Verfahren aus $\{1, x, x^2, \ldots,\}$ gewonnenen bzgl. $(\cdot,\cdot)_\omega$ orthogonalen Polynome.

Satz 3.3 (Nullstellen orthogonaler Polynome): *Die bzgl. des Skalarproduktes $(\cdot,\cdot)_\omega$ orthogonalen Polynome p_n besitzen lauter reelle, einfache Nullstellen, die alle im Innern des Intervalls $[a,b]$ liegen.*

Beweis: Wir definieren die Menge

$$N_n := \{\lambda \in (a,b) |\, \lambda \text{ Nullstelle } \textit{ungerader} \text{ Vielfachheit von } p_n\}$$

und setzen

$$q(x) := 1 \quad \text{für } N_n = \emptyset,$$
$$q(x) := \prod_{i=1}^{m} (x - \lambda_i) \quad \text{für } N_n = \{\lambda_1, \ldots, \lambda_m\}.$$

Dann ist $p_n \cdot q \in P_{n+m}$ reell und hat in (a,b) keinen Vorzeichenwechsel. Es gilt

$$(p_n, q)_\omega = \int_a^b p_n(x)q(x)\omega(x)\,dx \neq 0.$$

Für $m < n$ ist dies ein Widerspruch zu $p_n \perp P_{n-1}$. Q.E.D.

Die orthogonalen Polynome p_n bzgl. des Skalarprodukts $(\cdot,\cdot)$ auf $[-1,1]$ sind gerade Vielfache der "Legendre-Polynome" $L_n(x)$ mit der Normierung $p_n(1) = x^n + \ldots$. Aufgrund von Satz 3.3 können wir nun die Nullstellen $\lambda_0, \ldots, \lambda_n$ des $(n+1)$-ten Legendre-Polynoms L_{n+1} als Stützstellen einer interpolatorischen Quadraturformel auf dem Intervall $[-1,1]$ verwenden:

$$I^{(n)}(f) = \sum_{i=0}^{n} \alpha_i f(\lambda_i), \quad \alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - \lambda_j}{\lambda_i - \lambda_j}\,dx. \tag{3.2.13}$$

Wir fassen die Ergebnisse dieser Vorüberlegungen in folgendem Satz zusammen.

Satz 3.4 (Gaußsche-Quadraturformeln): *Es gibt genau eine interpolatorische Quadraturformel zu $n+1$ paarweise verschiedenen Stützstellen über dem Intervall $[-1,1]$ mit*

der Ordnung $2n+2$ *. Ihre Stützstellen sind gerade die Nullstellen* $\lambda_0, \ldots, \lambda_n \in (-1,1)$ *des* $(n+1)$*-ten Legendre-Polynoms* $L_{n+1} \in P_{n+1}$ *, und ihre Gewichte genügen der Beziehung*

$$\alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \Big(\frac{x-\lambda_j}{\lambda_i-\lambda_j}\Big)^2 \, dx > 0, \quad i = 0, \ldots, n. \tag{3.2.14}$$

Für $f \in C^{2n+2}[-1,1]$ *besitzt ihr Restglied die Darstellung*

$$R^{(n)}(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{j=0}^{n} (x-\lambda_j)^2 \, dx\,, \quad \xi \in (-1,1). \tag{3.2.15}$$

Beweis: (i) Das orthogonale Polynom p_{n+1} ist orthogonal zu $P_n[-1,1]$ und hat mit seinen (reellen) Nullstellen $\lambda_0, \ldots, \lambda_n \in (-1,1)$ die Darstellung

$$p_{n+1}(x) = \prod_{i=0}^{n} (x-\lambda_i) = x^{n+1} + \ldots .$$

Aufgrund der obigen Vorbetrachtung ist die zugehörige interpolatorische Quadraturformel dann von $(2n+2)$-ter Ordnung. Zur Bestimmung der Gewichte α_i setzen wir

$$l_i(x) \;:=\; \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-\lambda_j}{\lambda_i-\lambda_j}\,, \quad i = 0, \ldots, n,$$

und erhalten wegen $l_i^2 \in P_{2n}$

$$0 < \int_{-1}^{1} l_i(x)^2 \, dx = \sum_{j=0}^{n} \alpha_j \, \underbrace{l_i(\lambda_j)^2}_{\delta_{ij}} = \alpha_i.$$

(ii) Zum Beweis der Eindeutigkeit der Gaußschen Quadraturformel sei angenommen, es gäbe eine zweite Formel

$$\tilde{I}^{(n)}(f) = \sum_{i=0}^{n} \tilde{\alpha}_i \, f(\tilde{\lambda}_i)$$

der Ordnung $2n+2$. Mit den analog gebildeten Polynomen $\tilde{l}_i \in P_n$ folgte dann ebenfalls $\tilde{\alpha}_i > 0$. Also wäre

$$0 = \int_{-1}^{1} \frac{1}{\tilde{\alpha}_i} \underbrace{\tilde{l}_i(x)}_{\in P_n} p_{n+1}(x) \, dx = \sum_{j=0}^{n} \frac{\tilde{\alpha}_j}{\tilde{\alpha}_i} \underbrace{\tilde{l}_i(\tilde{\lambda}_j)}_{=\delta_{ij}} p_{n+1}(\tilde{\lambda}_j) = p_{n+1}(\tilde{\lambda}_i).$$

Wegen der eindeutigen Bestimmtheit der Nullstellen λ_i von p_{n+1} bzw. L_{n+1} folgte damit $\lambda_i = \tilde{\lambda}_i$ sowie $\alpha_i = \tilde{\alpha}_i$.

(iii) Es bleibt, die Restglieddarstellung herzuleiten. Nach Satz 2.5 und 2.6 gibt es ein Polynom $h \in P_{2n+1}$, welches die Hermitesche Interpolationsaufgabe

$$h(\lambda_i) = f(\lambda_i)\,, \quad h'(\lambda_i) = f'(\lambda_i)\,, \quad i = 0, \ldots, n,$$

löst und für $f \in C^{2n+2}[-1,1]$ die Restglieddarstellung hat:

$$f(x) - h(x) = f[\lambda_0, \lambda_0, \lambda_1, \lambda_1, \ldots, \lambda_n, \lambda_n, x] \prod_{i=0}^{n} (x - \lambda_i)^2.$$

Anwendung der Gaußschen Quadraturformel auf $h(x)$ ergibt dann wegen der Identität $I^{(n)}(h) = I(h)$:

$$\begin{aligned} I(f) - I^{(n)}(f) &= I(f-h) - I^{(n)}(f-h) \\ &= \int_{-1}^{1} f[\lambda_0, \lambda_0, \ldots, \lambda_n, \lambda_n, x] \prod_{i=0}^{n} \underbrace{(x-\lambda_i)^2}_{\geq 0}\, dx - \sum_{i=0}^{n} \alpha_i \underbrace{\{f(\lambda_i) - h(\lambda_i)\}}_{=0} \\ &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{i=0}^{n} (x-\lambda_i)^2\, dx. \end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Die Legendre-Polynome $L_n \in P_n$ bzw. ihre Vielfachen p_n lassen sich auf $[-1,1]$ in der Form

$$p_n(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n} (x^2-1)^n \quad (0! := 1)$$

schreiben und genügen der rekursiven Beziehung

$$p_0(x) \equiv 1\,, \quad p_1(x) \equiv x\,, \quad p_{n+1}(x) = x p_n(x) - \frac{n^2}{4n^2-1} p_{n-1}(x)\,, \quad n \geq 1.$$

Ihre Nullstellen werden analytisch bzw. (für $n > 3$) numerisch bestimmt und können Tabellenwerken entnommen werden; z.B.:

$$\begin{aligned} p_2(x) &= x^2 - \tfrac{1}{3} : \quad \lambda_0 = -\sqrt{1/3}\,, \quad \lambda_1 = \sqrt{1/3} \\ p_3(x) &= x^3 - \tfrac{3}{5}x : \quad \lambda_0 = -\sqrt{3/5}\,, \quad \lambda_1 = 0\,, \quad \lambda_2 = \sqrt{3/5}\,. \end{aligned}$$

Die Gewichte der zugehörigen Quadraturformeln bestimmt man gemäß

$$\alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-\lambda_j}{\lambda_i - \lambda_j}\, dx = \frac{1}{p'_{n+1}(\lambda_i) p_n(\lambda_i)} \cdot \frac{(n!)^4\, 2^{2n+1}}{(2n)!^3 (2n+1)},$$

und für die Restglieder gilt

$$R^{(n)}(f) = \frac{2^{2n+3}\,[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\, f^{(2n+2)}(\zeta)\,, \quad \zeta \in (-1,1).$$

Für $n=1$ und $n=2$ ergeben sich also die Quadraturformeln

$$\begin{aligned} I^{(1)}(f) &= f\big(-\sqrt{1/3}\big) + f\big(\sqrt{1/3}\big) = \int_{-1}^{1} f(x)\,dx - \frac{1}{135}\, f^{(iv)}(\zeta)\,, \\ I^{(2)}(f) &= \frac{1}{9}\big\{5f\big(-\sqrt{3/5}\big) + 8f(0) + 5f\big(\sqrt{3/5}\big)\big\} \\ &= \int_{-1}^{1} f(x)\,dx - \frac{1}{15.750}\, f^{(vi)}(\zeta)\,, \quad \zeta \in (-1,1)\,. \end{aligned}$$

Gaußsche Quadraturformeln über einem beliebigen (beschränkten) Intervall $[a,b]$ gewinnt man durch Anwendung der Koordinatentransformation $\varphi : [-1,1] \to [a,b]\,,$

$$y = \varphi(x) = \frac{b-a}{2}\, x + \frac{b+a}{2}. \tag{3.2.16}$$

Es ist dann

$$\begin{aligned} \int_a^b f(y)\,dy &= \frac{b-a}{2} \int_{-1}^{1} f(\varphi(x))\,dx \\ &= \frac{b-a}{2} \sum_{i=0}^{n} \alpha_i\, f(\varphi(\lambda_i)) + \frac{b-a}{2}\, R^{(n)}(f(\varphi(\cdot))), \end{aligned}$$

wobei

$$R^{(n)}(f(\varphi(\cdot))) = \frac{2^{(2n+3)}\,(n+1)!^4}{(2n+3)(2n+2)!^3} \Big(\frac{b-a}{2}\Big)^{2n+2} f^{(2n+2)}(\varphi(\zeta)),$$

d.h.: Die Stützstellen und Gewichte der Quadraturformel $(2n+2)$-ter Ordnung über $[a,b]\,,$

$$I^{(n)}(f) = \sum_{i=0}^{n} \tilde{\alpha}_i f(\tilde{\lambda}_i),$$

sind gegeben durch

$$\tilde{\lambda}_i = \tfrac{1}{2}(b-a)\,\lambda_i + \tfrac{1}{2}(b+a)\,, \quad \tilde{\alpha}_i = \tfrac{1}{2}(b-a)\,\alpha_i\,, \quad i = 0,\ldots,n.$$

Für $n=1$ und $n=2$ erhalten wir mit $c = \frac{b+a}{2}$ und $h = \frac{b-a}{2}$ die Quadraturformeln

$$\begin{aligned} I^{(1)}(f) &= \frac{b-a}{2}\{f(c-\sqrt{1/3}h) + f(c+\sqrt{1/3}h)\}\,, \\ I^{(2)}(f) &= \frac{b-a}{18}\{5f(c-\sqrt{3/5}h) + 8f(c) + 5f(c+\sqrt{3/5}h)\}. \end{aligned}$$

Die zugehörigen summierten Gaußschen Quadraturformeln haben die Gestalt $(x_j = a + jh\,,\ h = (b-a)/N)$:

$$I_h^{(1)}(f) = \frac{h}{2} \sum_{j=0}^{N-1} \{f(x_j + h') + f(x_{j+1} - h')\}$$

mit $h' = (\frac{1}{2} - \frac{1}{2\sqrt{3}})h \sim 0.2113249\, h$,

$$I^{(2)}(f) = \frac{h}{18} \sum_{j=0}^{N-1} \{5f(x_j + h') + 8f(x_j + \tfrac{1}{2}h) + 5f(x_{j+1} - h')\}$$

mit $h' = \left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{3}{5}}\right) h \sim 0,1127012\, h$.

Satz 3.5 (Konvergenz der Gauß-Quadratur): *Sei $I^{(n)}(f)$ die $(n+1)$-punktige Gauß-Formeln zur Berechnung von*

$$I(f) = \int_{-1}^{1} f(x)\, dx.$$

Für jede Funktion $f \in C[-1,1]$ konvergiert dann

$$I^{(n)}(f) \to I(f) \ (n \to \infty).$$

Beweis: Für die Gewichte der Gauß-Formel gilt

$$I^{(n)}(f) = \sum_{i=0}^{n} \alpha_i^{(n)} f(\lambda_i^{(n)}), \qquad \alpha_i^{(n)} > 0, \quad \sum_{i=0}^{n} \alpha_i^{(n)} = 2.$$

Sei $\varepsilon > 0$ beliebig vorgegeben. Nach dem Weierstraßschen Approximationssatz gibt es ein $p_\varepsilon \in P_m$ (m hinreichend groß), so daß

$$\max_{-1 \le x \le 1} |f(x) - p_\varepsilon(x)| \le \tfrac{\varepsilon}{4}.$$

Es ist $R^{(n)}(p_\varepsilon) = 0$ für $2n + 2 > m$ hinreichend groß. Für solche n ist also

$$|I(f) - I^{(n)}(f)| \le |\underbrace{I(f - p_\varepsilon)}_{\le \frac{\varepsilon}{4} \cdot 2}| + |\underbrace{I(p_\varepsilon) - I^{(n)}(p_\varepsilon)}_{= 0}| + |\underbrace{I^{(n)}(p_\varepsilon - f)}_{\le \frac{\varepsilon}{4} \cdot 2}| \le \varepsilon.$$

Wegen der beliebigen Wahl von $\varepsilon > 0$ muß $I^{(n)}(f) \to I(f)$ konvergieren für $n \to \infty$.
Q.E.D.

Die Methode zur Gewinnung der Gauß-Formeln zur "optimalen" Berechnung von $I(f)$ läßt sich übertragen auf den Fall von Integralen

$$I(f\omega) = \int_a^b f(x)\omega(x)\, dx$$

mit einer (uneigentlich) R-integrabierbaren Gewichtsfunktion $\omega(x) \geq 0$ mit höchstens endlich vielen Nullstellen auf (a,b). Hierbei verwendet man als Stützstellen gerade die Nullstellen der bzgl. des gewichteten Skalarprodukts

$$(p,q)_\omega = \int_a^b p(x)q(x)\omega(x)\,dx$$

orthogonalen Polynome, was durch Satz 3.3 gesichert ist; Satz 3.4 gilt dann sinngemäß.

Beispiel 3.4: Wir betrachten den Fall

$$[a,b] = [-1,1], \qquad \omega(x) = \frac{1}{\sqrt{1-x^2}}.$$

Die orthogonalen Polynome $p_n \in P_n[-1,1]$ sind in diesem Fall Vielfache der "Tschebyscheff-Polynome" $T_n(x) \in P_n[-1,1]$ und sind durch die rekursive Beziehung

$$p_0(x) = 1\,, \quad p_1(x) = x\,, \quad p_{n+1}(x) = 2xp_n(x) - p_{n-1}(x)\,, \quad n \geq 1,$$

bestimmt. Die Stützstellen und Gewichte der zugehörigen Quadraturformeln sind

$$\lambda_i = \cos\Big(\frac{\pi}{2}\,\frac{2i+1}{n+1}\Big)\,, \quad \alpha_i = \frac{\pi}{n+1}\,, \quad i = 0,\ldots,n.$$

Die Restglieder haben die Form

$$R^{(n)}(f) = \frac{2\pi}{2^{2n+2}(2n+2)!}\,f^{(2n+2)}(\zeta)\,, \quad \zeta \in (-1,1).$$

Fall $n = 2$:

$$\int_{-1}^{1} f(x)\omega(x)dx = \frac{\pi}{3}\Big\{f\big(-\tfrac{1}{2}\sqrt{3}\big) + f(0) + f\big(\tfrac{1}{2}\sqrt{3}\big)\Big\} + \frac{\pi}{23.040}\,f^{(vi)}(\zeta).$$

3.3 Das Rombergsche Integrationsverfahren

Die zusammengesetzten Quadraturformeln mit Schrittweite $h = \frac{b-a}{N}$ legen es nahe, das Prinzip der Extrapolation zum Grenzwert $h = 0$ zu verwenden. Die dazu nötige häufige Anwendung der Quadraturformeln erfordert solche mit einfacher Struktur und einer möglichst geringen Anzahl von Funktionsauswertungen. Wir beschränken uns daher im folgenden auf die zusammengesetzte Trapezregel. Das durch Extrapolation der Trapezregel gewonnene Integrationsverfahren geht auf Romberg[4] (1955) zurück und trägt daher auch seinen Namen. Wir setzen $h = (b-a)/N$ und $x_j = a + jh,\ j = 0, \ldots, N$. Für die zusammengesetzte Trapezregel gilt dann

$$\int_a^b f(x)\,dx \le h\left\{\frac{1}{2}f(a) + \sum_{j=1}^{N-1} f(x_j) + \frac{1}{2}f(b)\right\} - h^2\,\frac{b-a}{12}f''(\zeta). \tag{3.3.17}$$

Ist $f \in C[a,b]$, so konvergiert bekanntlich

$$a(h) = h\sum_{j=0}^{N-1} f(x_j) + \underbrace{\frac{h}{2}\big\{f(b) - f(a)\big\}}_{\to\,0} \to \int_a^b f(x)\,dx \quad (h \to 0).$$

Die Grundlage der Berechnung von $\lim_{h\to 0} a(h)$ durch Extrapolation ist wieder eine asymptotische Entwicklung von $a(h)$ nach Potenzen der Gitterweite h.

Satz 3.6 (Euler-Maclaurinsche Summenformel): *Für $f \in C^{2m+2}[a,b]$ gilt die sog. "Euler[5]-Maclaurinsche[6] Summenformel"*

$$a(h) = \int_a^b f(x)\,dx + \sum_{k=1}^{m} h^{2k}\frac{B_{2k}}{(2k)!}\left\{f^{(2k-1)}(b) - f^{(2k-1)}(a)\right\} +$$
$$+\, h^{2m+2}\,\frac{b-a}{(2m+2)!}\,B_{2m+2}f^{(2m+2)}(\zeta)\,, \quad \zeta \in [a,b],$$

mit den Bernoulli[7]-Zahlen B_{2k}.

[4] Werner Romberg (1909-2003): Deutscher Mathematiker; emigrierte 1937 aus polit. Gründen nach Rußland und später nach Norwegen; 1950-1968 Professor in Trondheim und 1968-1977 Inhaber des Lehrstuhls für Mathematische Methoden der Naturwissensch. und Numerik in Heidelberg; Beiträge zur Numerik von Differentialgleichungen und numerischen Integration ("Rombergsches Extrapolationsverfahren").

[5] *Leonhard Euler (1707-1783), geb. in Basel: universeller Mathematiker und Physiker; bedeutendster und produktivster Mathematiker seiner Zeit; wirkte in Berlin und St. Petersburg; Arbeiten zu allen mathemischen Gebieten seiner Zeit.*

[6] *Colin Maclaurin (1698-1746): Schottischer Mathematischer; Professor an den Universitäten Aberdeen (1717) und Edinburgh (1725); Beiträge zur damals neuenNewtonschen Differentialrechnung (erste systematische Darstellung des Newtonschen "Kalküls" und Entwicklung der nach ihm benannten Integralformel (1742)), zur klassischen Mechanik, Geometrie und Algebra.*

[7] *Bernoulli: Schweizer Mathematiker Familie; Jakob Bernoulli (1655-1705) lehrte in Basel; verwendete*

Beweis: Siehe z.B. Stoer I. Die Bernoulli-Zahlen sind z. B. bestimmt als die Koeffizienten in der Potenzreihenentwicklung

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k}{k!}\, x^k, \tag{3.3.18}$$

und genügen der Rekursionsformel

$$B_k = -\sum_{j=0}^{k-1} \frac{k!}{j!(k-j+1)!}\, B_j\,, \quad k = 1, \ldots. \tag{3.3.19}$$

$$B_0 = 1\,, \quad B_1 = -\tfrac{1}{2}\,, \quad B_2 = \tfrac{1}{6}\,, \quad B_4 = -\tfrac{1}{30}\,, \quad B_6 = \tfrac{1}{42}\,, \quad B_8 = -\tfrac{1}{30}\,, \ldots.$$

Für ungerade Indizes gilt $B_{2j+1} = 0$, und für $k \to \infty$ wachsen die Bernoulli-Zahlen sehr schnell an wie

$$B_{2k} \approx (2k)!/(2\pi)^{2k}.$$

Q.E.D.

Die summierte Trapezregel besitzt also eine Entwicklung nach geraden Potenzen von h. Dieser Umstand macht die Extrapolation mit geraden Polynomen, d. h. solchen in h^2, besonders effizient. Zur Berechnung von

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

geht man nach dem Extrapolationsprinzip wie folgt vor:

1. Für eine Folge von Schrittweiten $h_0 > h_1 > h_2 > \ldots > h_m$ wird $a(h_k)$ berechnet. Dabei verwendet man in der Regel die sog. "Romberg-Folge" $h_k = h/2^k$. Diese bietet den Vorteil der Wiederverwendbarkeit der schon berechneten Funktionswerte, führt aber auf eine rasch anwachsende Zahl von Stützstellen.

2. Das Interpolationspolynom in h^2 zu den Stützpunkten $(h_i^2, a(h_i))$, $i = 0, \ldots, m$ wird an der Stelle $h = 0$ nach dem Neville-Schema ausgewertet:

$$\begin{aligned} a_{i0} &= a(h_i)\,, \quad i = 0, \ldots, m\,, \qquad k = 1, \ldots, m\,: \\ a_{ik} &= a_{i,k-1} + \frac{a_{i,k-1} - a_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1}\,, \quad i = k, \ldots, m. \end{aligned}$$

bereits die vollständige Induktion; Entdecker der "Bernoulli-Zahlen" und Mitbegründer der Wahrscheinlichkeitsrechnung; sein jüngerer Bruder Johann Bernoulli (1667-1748) wirkte zuletzt in Basel und galt nach dem Tode seines Bruders Jakob als führender Mathematiker seiner Zeit; er leistete Beiträge über Reihen und Differentialgleichungen; sein Sohn Daniel Bernoulli (1700-1782) setzte diese Arbeiten fort; er wirkte in St. Petersburg und Basel und leistete wichtige Beiträge zur Hydromechanik und Gasdynamik.

Dies ist das sog. "Integrationsverfahren von Romberg". Es baut also sukzessive folgendes Extrapolationsschema auf:

k	0	1	2		$m-1$	m
h_0	$\boxed{a_{00}}$					
h_1	a_{10}	$\boxed{a_{11}}$				
h_2	a_{20}	a_{21}	$\boxed{a_{22}}$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
h_{m-1}	$a_{m-1,0}$	$a_{m-1,1}$	$a_{m-1,2}$	$\cdots$	$\boxed{a_{m-1,m-1}}$	
h_m	$a_{m,0}$	$a_{m,1}$	$a_{m,2}$	$\cdots$	$a_{m,m-1}$	$\boxed{a_{m,m}}$

Die Diagonalelemente $a_{k,k}$ sind gerade die Näherungen zu $a(0)$, die man durch Extrapolation der Stützpunkte $(h_i^2, a(h_i))$, $i = 0, \ldots, k$, gewinnt.

Als Folgerung aus dem allgemeinen Satz 2.7 erhält man die Konvergenzaussage:

Satz 3.7 (Romberg-Integration): *Es sei $f \in C^{2m+2}[a,b]$. Der für die Schrittweitenfolge $h_k = h/2^k$, $k = 0, \ldots, m$, berechnete extrapolierte Wert $a_{m,m}$ konvergiert gegen $a(0)$ für $h \to 0$ mit der Fehlerordnung*

$$a(0) - a_{m,m} = O(h^{2m+2}). \tag{3.3.20}$$

Bemerkung 3.3: Sei $f \in C^{2m+2}(-\infty, \infty)$ mit der Periode $[a,b]$. Dann ist $f^{(k)}(a) = f^{(k)}(b)$, und Satz 3.6 ergibt

$$a(h) = \int_a^b f(x)\,dx + O(h^{2m+2}).$$

Ist sogar $f \in C^\infty(-\infty, \infty)$, so konvergiert die zusammengesetzte Trapezregel schneller als jede Potenz von h gegen das Integral von f über ein ganzes Periodenintervall. Wegen $f(a) = f(b)$ ist in diesem Fall

$$a(h) = h \sum_{j=0}^{N-1} f(x_j), \tag{3.3.21}$$

d.h.: Die Trapezregel fällt mit der summierten Rechteckregel zusammen. Diese primitivste Quadraturregel konvergiert also bereits besser als jede Potenz von h, so daß die Anwendung komplizierter Formeln eher schädlich wäre.

3.4 Praktische Aspekte der Integration

Das Hauptproblem bei der numerischen Integration ist die Gewinnung realistischer Schätzungen für den Fehler. Die z. B. für die summierten Newton-Cotes-Formeln hergeleitete *a priori* Fehlerabschätzung (n ungerade)

$$|I(f) - I_h^{(n)}(f)| \leq \omega_n (b-a) \max_{a \leq x \leq b} |f^{(k)}(x)|\, h^k$$

ist dazu in der Regel ungeeignet, da höhere Ableitungen $f^{(k)}$ des Integranden nur schwer zu berechnen sind. (Man berechne z.B. die 4-te Ableitung von $f(x) = (1+x^2)^{-1/2}\tan(x)$!) Die Verwendung der numerischen Differentiation scheidet wegen des damit verbundenen großen Rundungsfehlers und des erneuten Bedarfs von Fehlerabschätzungen aus.

Bei Quadraturformeln, die wie die summierten Newton-Cotes-Formeln von einem Schrittweitenparameter h abhängen, kann das Restglied näherungsweise aus den tatsächlich berechneten Werten durch eine sog. "*a posteriori*" Fehlerabschätzung bestimmt werden. Für eine Quadraturformel $I_h(f)$ zur Schrittweite h gelte

$$I(f) = I_h(f) + \omega_f h^k + r(f;h)h^{k+1}. \tag{3.4.22}$$

Zur Bestimmung des Restgliedkoeffizienten ω_f berechnet man für ein gewisses h zusätzlich zu $I_h(f)$ noch den Wert $I_{h/2}(f)$ zur halbierten Schrittweite. Für diesen gilt dann

$$I(f) = I_{h/2}(f) + \omega_f \Big(\frac{h}{2}\Big)^k + r\Big(f;\frac{h}{2}\Big)\Big(\frac{h}{2}\Big)^{k+1}. \tag{3.4.23}$$

Durch Elimination von $I(f)$ aus (3.4.22) und (3.4.23) folgt

$$\begin{aligned} I_{h/2}(f) - I_h(f) &= \omega_f\Big\{h^k - \Big(\frac{h}{2}\Big)^k\Big\} + h^{k+1}r(f;h) - \Big(\frac{h}{2}\Big)^{k+1} r\Big(f;\frac{h}{2}\Big) \\ &= \omega_f h^k(1-2^{-k}) + O(h^{k+1}) \end{aligned}$$

bzw.

$$h^k\omega_f = \frac{I_{h/2}(f) - I_h(f)}{1-2^{-k}} + O(h^{k+1}). \tag{3.4.24}$$

Durch den Quotienten

$$\frac{I_{h/2}(f) - I_h(f)}{1-2^{-k}}\, h^{-k} \doteq \omega_f \tag{3.4.25}$$

erhält man also eine Schätzung für den führenden Koeffizienten ω_f im Restglied bis auf einen Fehler der Ordnung $O(h)$, der vernachlässigt wird. Dies kann zu einem heuristischen Abbruchkriterium für die numerische Quadratur ausgebaut werden:

Für eine Funktion $f \in C^{k+1}[a,b]$ soll das Integral $I(f)$ über $[a,b]$ mit Hilfe einer interpolatorischen Quadraturformel der Ordnung k berechnet werden. Die vorgegebene

Fehlertoleranz sei TOL. Das Problem ist also die Bestimmung einer geeigneten Schrittweite h. Wir nehmen wieder an, daß die Quadraturformel eine asymptotische Entwicklung der Art (3.4.22) gestattet. In erster Näherung wollen wir h so bestimmen, daß wenigstens für den führenden Fehlerterm gilt

$$|I(f) - I_h(f)| \dot{\leq} |\omega_f h^k| \leq TOL\,! \tag{3.4.26}$$

Aufgrund der Schätzung für ω_f sollte also gelten:

$$\left| \frac{I_{h/2}(f) - I_h(f)}{1 - 2^{-k}} \right| \leq TOL\,. \tag{3.4.27}$$

Wegen $I_h(f) \to I(f)\ (h \to 0)$ wird nach einer gewissen Anzahl von Schrittweitenhalbierungen diese Bedingung erfüllt sein. Aus (3.4.26) läßt sich die gesuchte Schrittweite näherungsweise bestimmen zu

$$h_{TOL} \approx \left(\frac{TOL}{\omega_f} \right)^{1/k}. \tag{3.4.28}$$

Zur Überprüfung der Gültigkeit dieser Schrittweitenwahl, d. h. der Verläßlichkeit der Schätzung von ω_f, vergleicht man noch h_{TOL} mit der Schätzschrittweite h. Im Falle $h_{TOL} \approx h$ wird das Ergebnis akzeptiert. Ist dagegen $h_{TOL} \ll h$, so wird der ganze Schätzprozeß mit der neuen Schätzschrittweite $h := h_{TOL}$ wiederholt, bis schließlich der erste Fall eintritt. Bei Unterschreiten einer vorgegebenen minimal erlaubten Schrittweite h_{min} wird der Approximationsprozeß abgebrochen, da das Integral offenbar mit dem zur Verfügung stehenden Aufwand nicht verläßlich berechenbar ist.

Bei gleichem Rechenaufwand (gemessen an der Zahl der Funktionsauswertungen) liefert die Gaußsche Integrationsmethode die genauesten Resultate. Wenn man bei einem vorgelegten Integral $I(f)$ und gewünschter Genauigkeit ε wüßte, welche Schrittweite h man zu nehmen hätte, so wären die Gaußschen Formeln den anderen Methoden überlegen. Da dies aber a priori kaum möglich ist, müssen die beschriebenen Methoden zur a posteriori Fehlerabschätzung verwendet werden. Da man im Gegensatz zu den Extrapolationsverfahren beim Übergang von h nach $h/2$ die bis dahin berechneten Funktionswerte von $f(x)$ nicht weiter verwenden kann, gehen die Vorzüge des Gaußschen Verfahrens schnell verloren. Im Fall $[a, b] = [0, 1]$ berechnet sich mit $h = 1/N$ der Rechenaufwand der summierten (abgeschlossene) Newton-Cotes-Formeln (für ungerades n) zu etwa n/h Funktionsauswertungen, d. h.: Zur Erzielung der Ordnung $O(h^{2n+2})$ sind etwa n/h^2 Funktionsauswertungen erforderlich. Die summierten Gauß-Formeln benötigen für dieselbe Genauigkeit nur etwa n/h Funktionsauswertungen. Das Romberg-Verfahren (für $h_i = 2^{-i}h$) liegt mit etwa $2^n/h$ Funktionsauswertungen auch nocht recht gut.

Beispiel 3.5: Für $n = 3$ und $h = 10^{-2}$ ergibt sich ein Fehler der Größe $10^{-16} f^{(8)}(\zeta)$. Die drei Verfahren benötigen hierfür folgende Anzahlen von Funktionsauswertungen:

$$Newton - Cotes:\ 30.000\,, \quad Gauß:\ 400\,, \quad Romberg:\ 800.$$

3.5 Übungsaufgaben

Übung 3.1: Mit wievielen Funktionsauswertungen kann das Integral

$$I = \int_0^1 \frac{dx}{1+2x} = 0,54930614\ldots$$

mit einem Fehler kleiner als 10^{-8} berechnet werden,

a) mit Hilfe der summierten Trapezregel,

b) mit Hilfe der summierten Simpson-Regel ?

(Hinweis: Das Integral soll hierzu nicht explizit berechnet werden!)

Übung 3.2: Man berechne das Integral

$$I = \int_0^{\pi/2} \sin(x)\, dx$$

mit Hilfe des Romberg-Verfahrens (Schrittweitenfolge $h_i = 2^{-i-1}\pi$, $i = 0, 1, 2, \ldots,$) mit einem Fehler kleiner 10^{-4}. Die Genauigkeit kontrolliere man dabei mit der in der Vorlesung angegebenen Methode zur a posteriori Fehlerabschätzung beim Extrapolationsverfahren.

Übung 3.3: Man bestimme eine Gaußsche Quadraturformel, welche das Integral

$$I = \int_{-1}^1 f(x)\sqrt{|x|}\, dx$$

für alle Polynome aus P_3 exakt integriert.

Übung 3.4: Man gebe eine Quadraturformel zur Berechnung des Integrals

$$I = \int_{-1}^1 \frac{\cos(\pi x/2)}{\sqrt{1-x^2}}\, dx$$

mit einem garantierten Fehler $\varepsilon \leq 10^{-4}$ an (bei Vernachlässigung der Rundungsfehler). Die Quadraturformel soll so gewählt sein, daß möglichst wenige Funktionsauswertungen erforderlich sind.

Übung 3.5: (Praktische Aufgabe) Man schreibe ein Programm zur näherungsweisen Berechnung des Integrals $I(f)$ mit dem Romberg-Verfahren und wende dieses an für das Integral

$$I(f) = \int_0^1 \frac{4}{x^2+1}\, dx.$$

Welcher "exakte" Wert ergibt sich für das Integral?

a) Der Extrapolationsprozeß zur Schrittweitenfolge $h_i = 2^{-i}\,(i = 0, 1, 2, \ldots)$ liefert Näherungswerte $R_i(f) := a_{ii}$ (Diagonalelemente im Extrapolationstableau). Man setze den Extrapolationsprozeß fort, bis entweder der Fehler kleiner als 10^{-10} oder $i = 20$ ist. Dabei kontrolliere man die Genauigkeit von $R_i(f)$ jeweils mit Hilfe der in der Vorlesung angegebenen Methode zur a posteriori Fehlerschätzung. Man gebe die Folgen a_{ii} aus und stelle die Entwicklung des geschätzten und des tatsächlichen Fehlers $|b_{ii} - a_{ii}|$ bzw. $|R_i(f) - I(f)|$ graphisch dar.

b) Zur Illustration der Aussagekräftigkeit der Theorie des Extrapolationsverfahrens wiederhole man die Rechnungen (und die graphischen Ausgaben) für die Extrapolation nach Potenzen von h (statt h^2) sowie für die Schrittweitenfolge $h_i = (i+1)^{-1}$, $i = 0, \ldots, 20$.)

4 Lineare Gleichungssysteme I (Direkte Verfahren)

Seien A eine Matrix und b ein Vektor

$$A = (a_{jk})_{j,k=1}^{m,n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \qquad b = (b_j)_{j=1,\ldots,m} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

Gesucht ist ein Vektor $x = (x_k)_{k=1,\ldots,n}$ mit der Eigenschaft

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

oder abgekürzt geschrieben: $Ax = b$.

Definition 4.1: *Das lineare Gleichungssystem $Ax = b$ heißt "unterbestimmt" im Fall $m < n$, "quadratisch" im Fall $m = n$ und "überbestimmt" im Fall $m > n$.*

Das lineare Gleichungssystem ist genau dann lösbar, wenn $\mathrm{Rang}(A) = \mathrm{Rang}[A, b]$, mit der zusammengesetzten Matrix

$$[A, b] = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right].$$

Im "quadratischen" Fall sind die folgenden Aussagen äquivalent:

(i) $Ax = b$ ist für jedes b eindeutig lösbar.

(ii) $\mathrm{Rang}(A) = n$.

(iii) $\det(A) \neq 0$.

(iv) Alle Eigenwerte von A sind ungleich Null.

Wir beschäftigen uns im folgenden hauptsächlich mit der Lösung von quadratischen Gleichungssystemen. Die dazu verwendeten Verfahren lassen sich grob in zwei Klassen einteilen:

Definition 4.2: *Ein "direktes" Verfahren zur Lösung des Gleichungssystems $Ax = b$ ist ein Algorithmus, der (bei Vernachlässigung von Rundungsfehlern) in endlich vielen Schritten die Lösung x liefert. Im Gegensatz dazu erzeugen die "iterativen" Verfahren sukzessive eine Folge von Vektoren $(x^{(t)})_{t\in\mathbb{N}}$, die im Limes für $t \to \infty$ immer bessere Approximationen zur Lösung x sind.*

4.1 Störungstheorie

Wir beschäftigen uns zunächst mit dem Problem der "Konditionierung" von quadratischen linearen Gleichungssystemen. Bei der Lösung eines Gleichungssystems $Ax = b$ treten zwei Fehlereinflüsse auf:

a) Fehler in der "theoretischen" Lösung aufgrund von Eingangsfehlern in den Elementen von A und b,

b) Fehler in der "numerischen" Lösung aufgrund des Rundungsfehlers im Verlaufe des Lösungsprozesses.

4.1.1 Vektor- und Matrizennormen

Zur Erfassung dieser Fehler benötigen wir ein Maß für die "Größe" von Vektoren und Matrizen. Dazu dienen üblicherweise Normen auf dem n-dimensionalen Zahlenraum $\mathbb{K}^n$, $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. (Im Hinblick auf spätere Anwendungen lassen wir im folgenden auch komplexe Vektoren bzw. Matrizen zu.)

Definition 4.3: *Eine Abbildung $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_+$ heißt "Norm", wenn sie folgende Eigenschaften besitzt:*

(N1) $\|x\| > 0\,, \quad x \in \mathbb{K}^n \setminus \{0\}$ *(Definitheit),*

(N2) $\|\alpha x\| = |\alpha|\,\|x\|\,, \quad x \in \mathbb{K}^n,\ \alpha \in \mathbb{K}$ *(positive Homogenität),*

(N3) $\|x + y\| \le \|x\| + \|y\|\,, \quad x, y \in \mathbb{K}^n$ *(Subadditivität).*

Beispiel 4.1: *Gebräuchliche Beispiele von Vektornormen sind:*

$$\begin{aligned} \|x\|_2 &:= \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2} && \text{"Euklidische Norm" } (l_2\text{-Norm)} \\ \|x\|_\infty &:= \max_{i=1,\ldots,n} |x_i| && \text{"Maximumnorm" } (l_\infty\text{-Norm)} \\ \|x\|_1 &:= \sum_{i=1}^{n} |x_i| && \text{"}l_1\text{-Norm"} \end{aligned}$$

Mit Hilfe einer Norm $\|\cdot\|$ auf $\mathbb{K}^n$ läßt sich die Konvergenz einer Folge von Vektoren gegen einen Vektor erklären durch

$$x^{(t)} \to x\ (t \to \infty) \quad :\Longleftrightarrow \quad \|x^{(t)} - x\| \to 0\ (t \to \infty)\,.$$

Die sog. "Dreiecksungleichung" (N3) ergibt über die Beziehung $\|x\| = \|x - y + y\|$ die wichtige Ungleichung

$$\|x - y\| \ge \big|\|x\| - \|y\|\big|\,, \quad x, y \in \mathbb{K}^n\,, \tag{4.1.1}$$

welche u.a. die Stetigkeit von Normen als Funktionen von $\mathbb{K}^n$ in $\mathbb{R}$ impliziert.

Hilfssatz 4.1 (Normäquivalenz): *Auf dem endlich dimensionalen Vektorraum* $\mathbb{K}^n$ *sind alle Normen äquivalent, d.h.: Zu je zwei Normen* $\|\cdot\|$, $\|\cdot\|'$ *gibt es positive Konstanten* m, M, *mit denen gilt:*

$$m\,\|x\| \le \|x\|' \le M\,\|x\|\,, \quad x \in \mathbb{K}^n\,. \tag{4.1.2}$$

Beweis: Es genügt, die Behauptung für den Fall zu zeigen, daß eine der beiden Normen die Maximumnorm $\|\cdot\|_\infty$ ist. Sei $\|\cdot\|$ irgendeine zweite Norm. Bzgl. der kartesischen Einheitsvektoren $e_1, \ldots, e_n$ hat jeder Vektor $x \in \mathbb{K}^n$ die Darstellung $x = \sum_{i=1}^n x_i e_i$. Folglich gilt

$$\|x\| \le \gamma\,\|x\|_\infty\,, \quad \gamma := \sum_{i=1}^n \|e_i\|\,.$$

Die Norm $\|\cdot\|$ ist also auch stetig bzgl. der komponentenweisen Konvergenz von Vektoren. Die Punktmenge

$$S \equiv \{x \in \mathbb{K}^n,\ \|x\|_\infty = 1\} \subset \mathbb{K}^n$$

ist beschränkt und abgeschlossen (und damit kompakt). Die Norm $\|\cdot\|$ nimmt also als stetige Funktion auf S ihr Minimum und Maximum an. Es existieren also $x_0, x_1 \in S$, so daß

$$0 < \|x_0\| \le \|x\| \le \|x_1\| < \infty\,, \quad \forall x \in S\,.$$

Für beliebiges $y \in \mathbb{K}^n \setminus \{0\}$ ist $y/\|y\|_\infty \in S$ und folglich

$$\|x_0\| \le \|y\|/\|y\|_\infty \le \|x_1\|\,.$$

Mit $m \equiv \|x_0\|$ und $M \equiv \|x_1\|$ gilt daher

$$m\,\|y\|_\infty \le \|y\| \le M\,\|y\|_\infty\,, \quad \forall y \in \mathbb{K}^n$$

Q.E.D.

Die Beziehung (4.1.2) impliziert, daß die durch eine beliebige Norm induzierte Konvergenz von Vektoren stets äquivalent zur "komponentenweisen" Konvergenz ist.

Wir betrachten nun den Vektorraum der $n \times n$-Matrizen $A \in \mathbb{K}^{n\times n}$. Offenbar kann dieser mit dem Vektorraum der n^2-Vektoren identifiziert werden. Somit übertragen sich alle Aussagen für Vektornormen auf Normen für Matrizen. Insbesondere sind alle Normen für $n\times n$-Matrizen äquivalent, und die Konvergenz von Matrizen ist die komponentenweise Konvergenz:

$$A^{(t)} \to A\,(t \to \infty) \quad \Longleftrightarrow \quad a_{jk}^{(t)} \to a_{jk}\,(t \to \infty)\,, \quad j,k = 1, \ldots, n\,.$$

Definition 4.4: *Eine Norm* $\|\cdot\|$ *auf* $\mathbb{K}^{n\times n}$ *heißt "verträglich" mit einer Vektornorm* $\|\cdot\|$ *auf* $\mathbb{K}^n$, *wenn gilt:*

$$\|Ax\| \le \|A\|\,\|x\|\,, \quad x \in \mathbb{K}^n\,, \quad A \in \mathbb{K}^{n\times n}.$$

Sie heißt "Matrizennorm", wenn sie submultiplikativ ist:

$$\|AB\| \le \|A\|\,\|B\|\,, \quad A, B \in \mathbb{K}^{n\times n}.$$

Z.B. ist die Quadratsummennorm (sog. "Frobenius[1]-Norm")

$$\|A\|_{\mathrm{Fr}} := \Big(\sum_{j,k=1}^{n} |a_{jk}|^2\Big)^{1/2}$$

eine mit der euklidischen Vektornorm verträgliche Matrizennorm. Für eine beliebige Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$ wird durch

$$\|A\| := \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{K}^n,\|x\|=1} \|Ax\|$$

eine mit $\|\cdot\|$ verträgliche Matrizennorm erklärt (Übungsaufgabe). Diese heißt die von $\|\cdot\|$ erzeugte "natürliche" Matrizennorm. Für natürliche Matrizennormen gilt

$$\|I\| = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ix\|}{\|x\|} = 1.$$

Hilfssatz 4.2: *Die natürlichen Matrizennormen zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind die "maximale Zeilensumme" bzw. die "maximale Spaltensumme":*

$$\|A\|_\infty := \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}|, \qquad \|A\|_1 := \max_{1\le k\le n} \sum_{j=1}^{n} |a_{jk}|\,. \tag{4.1.3}$$

Beweis: Wir geben den Beweis nur für $\|\cdot\|_\infty$. Für $\|\cdot\|_1$ verläuft er analog.
(i) Die maximale Zeilensumme $\|\cdot\|_\infty$ ist eine eine Matrizennorm. Die Normeigenschaften (N1) - (N3) folgen mit Hilfe der entsprechenden Eigenschaften des Absolutbetrags, und für ein Matrizenprodukt AB gilt

$$\begin{aligned}\|AB\|_\infty &= \max_{1\le i\le n} \Big|\sum_{j=1}^{n}\Big(\sum_{k=1}^{n} a_{ik}b_{kj}\Big)\Big| \le \max_{1\le i\le n}\sum_{k=1}^{n}\Big(|a_{ik}|\sum_{j=1}^{n}|b_{kj}|\Big)\\ &\le \max_{1\le i\le n}\sum_{k=1}^{n}|a_{ik}|\ \max_{1\le k\le n}\sum_{j=1}^{n}|b_{kj}| = \|A\|_\infty\|B\|_\infty.\end{aligned}$$

(ii) Weiter ist die maximale Zeilensumme wegen

$$\|Ax\|_\infty = \max_{1\le j\le n} |\sum_{k=1}^{n} a_{jk}x_k| \le \max_{1\le j\le n}\sum_{k=1}^{n}|a_{jk}|\ \max_{1\le k\le n}|x_k| = \|A\|_\infty\|x\|_\infty$$

[1]Ferdinand Georg Frobenius (1849-1917): deutscher Mathematiker; Prof. in Zürich und Berlin; bed. Beiträge zur Theorie der Differentialgleichungen, zu Determinanten und Matrizen sowie zur Gruppentheorie.

verträglich mit $\|\cdot\|_\infty$, und es gilt

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty \le \|A\|_\infty.$$

(iii) Im Falle $\|A\|_\infty = 0$ ist $A = 0$, d.h. trivialerweise

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty.$$

Sei also $\|A\|_\infty > 0$ und $m \in \{1, \ldots, n\}$ ein Index mit der Eigenschaft

$$\|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}| = \sum_{k=1}^{n} |a_{mk}|.$$

Wir setzen für $k = 1, \ldots, n$:

$$z_k := \begin{cases} |a_{mk}|/a_{mk} & \text{für } a_{mk} \neq 0, \\ 0, & \text{sonst,} \end{cases}$$

d.h.: $z = (z_k)_{k=1}^n \in \mathbb{K}^n$, $\|z\|_\infty = 1$. Für $v := Az$ gilt dann

$$v_m = \sum_{k=1}^{n} a_{mk} z_k = \sum_{k=1}^{n} |a_{mk}| = \|A\|_\infty.$$

Folglich ist

$$\|A\|_\infty = v_m \le \|v\|_\infty = \|Az\|_\infty \le \sup_{\|y\|_\infty=1} \|Ay\|_\infty,$$

was zu zeigen war. Q.E.D.

4.1.2 Eigenwerte und Skalarprodukte

Die "Eigenwerte" $\lambda \in \mathbb{K}$ einer Matrix $A \in \mathbb{K}^{n\times n}$ sind definiert als die Nullstellen ihres charakteristischen Polynoms $p(\lambda) = \det(A - \lambda I)$. Folglich existieren genau n (ihrer Vielfachheit als Nullstelle entsprechend oft gezählte) Eigenwerte λ, und zu jedem λ existiert mindestens ein "Eigenvektor" $w \in \mathbb{K}^n \setminus \{0\} : Aw = \lambda w$. Sei nun $\|\cdot\|$ eine beliebige Vektornorm und $\|\cdot\|$ eine damit verträgliche Matrizennorm, (wobei die beiden Normen der Einfachheit halber gleich bezeichnet werden). Mit einem normierten Eigenvektor zum Eigenwert λ gilt

$$|\lambda| = |\lambda|\,\|w\| = \|Aw\| \le \|A\|\,\|w\| = \|A\|, \tag{4.1.4}$$

d.h. alle Eigenwerte von A liegen in einer Kreisscheibe in $\mathbb{C}$ mit Mittelpunkt Null und Radius $\|A\|$. Speziell mit $\|A\|_\infty$ erhält man die Abschätzung

$$\max\ |\lambda| \le \|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}|.$$

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt "hermitesch", wenn gilt:

$$A = \bar{A}^T \quad \text{bzw.} \quad a_{jk} = \overline{a_{kj}},\ j,k = 1,\ldots,n\,.$$

Reelle hermitesche Matrizen werden "symmetrisch" genannt. Hermitesche Matrizen haben nur reelle Eigenwerte und besitzen dazu eine Orthonormalbasis von Eigenvektoren. Der Begriff der Symmetrie ist eng verknüpft mit dem des Skalarprodukts.

Definition 4.5: *Eine Abbildung* $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ *wird ein "Skalarprodukt" genannt, wenn sie folgende Eigenschaften hat:*

(S1) $(x,y) = \overline{(y,x)}\,, \quad x,y \in \mathbb{K}^n$ *(Symmetrie),*

(S2) $(\alpha x + \beta y, z) = \alpha(x,z) + \beta(y,z)\,, \quad x,y,z \in \mathbb{K}^n,\ \alpha,\beta \in \mathbb{K}$ *(Linearität),*

(S3) $(x,x) > 0\,, \quad x \in \mathbb{K}^n \setminus \{0\}$ *(Definitheit).*

Ein Skalarprodukt auf $\mathbb{K}^n \times \mathbb{K}^n$ *erzeugt eine zugehörige Vektornorm durch*

$$\|x\| := (x,x)^{1/2}, \quad x \in \mathbb{K}^n.$$

Im folgenden wird fast ausschließlich das sog. "euklidische" Skalarprodukt verwendet:

$$(x,y)_2 = \sum_{j=1}^{n} x_j \overline{y_j}\,, \quad (x,x)_2 = \|x\|_2^2.$$

Mit Hilfe des euklidischen Skalarprodukts läßt sich die Eigenschaft einer Matrix, hermitesch zu sein, äquivalent ausdrücken durch:

$$A = \bar{A}^T \quad \Longleftrightarrow \quad (Ax,y)_2 = (x,Ay)_2, \quad x,y \in \mathbb{K}^n.$$

Die von der euklidischen Vektornorm erzeugte natürliche Matrizennorm heißt die "Spektralnorm" und wird mit $\|\cdot\|_2$ bezeichnet. Diese Bezeichnung ist durch das folgende Resultat gerechtfertigt:

Hilfssatz 4.3: *Für die Spektralnorm hermitescher Matrizen* $A \in \mathbb{K}^{n\times n}$ *gilt*

$$\|A\|_2 = \max\{|\lambda|,\ \lambda \textit{ Eigenwert von } A\}\,. \tag{4.1.5}$$

Für allgemeine Matrizen $A \in \mathbb{K}^{n\times n}$ *gilt*

$$\|A\|_2 = \max\{|\lambda|^{1/2},\ \lambda \textit{ Eigenwert von } \bar{A}^T A\}. \tag{4.1.6}$$

Beweis: Bekanntlich besitzt eine hermitesche Matrix $A \in \mathbb{K}^{n\times n}$ nur reelle Eigenwerte und zwar genau n Stück (ihrer Vielfachheit entsprechend oft gezählt), $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Ferner existiert ein zugehöriges "Orthonormalsystem" von Eigenvektoren

$$\{w^1, \ldots, w^n\} \subset \mathbb{K}^n : Aw^i = \lambda_i w^i\,, \quad (w^i, w^j)_2 = \delta_{ij}, \quad i,j = 1, \ldots, n.$$

Jedes $x \in \mathbb{K}^n$ besitzt eine Darstellung der Form

$$x = \sum_{i=1}^{n} \alpha_i w^i\,, \quad \alpha_i = (x, w^i)_2,$$

und es gilt

$$\|x\|_2^2 = (x,x)_2 = \sum_{i,j=1}^{n} \alpha_i \bar{\alpha}_j\,(w^i, w^j)_2 = \sum_{i=1}^{n} |\alpha_i|^2,$$

$$\|Ax\|_2^2 = (Ax, Ax)_2 = \sum_{i,j=1}^{n} \lambda_i \alpha_i \overline{\lambda_j \alpha_j}\,(w^i, w^j)_2 = \sum_{i=1}^{n} \lambda_i^2\,|\alpha_i|^2.$$

Hiermit folgt

$$\|A\|_2^2 = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\sum_{i=1}^n \lambda_i^2\,|\alpha_i|^2}{\sum_{i=1}^n\,|\alpha_i|^2} \le \max_{1\le i\le n}\,|\lambda_i|^2.$$

Wegen der Eigenwertschranke (4.1.4) ergibt sich damit die Behauptung. Q.E.D.

Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt "positiv definit", wenn gilt:

$$(Ax, x)_2 \in \mathbb{R}\,, \quad (Ax,x)_2 > 0 \quad \forall x \in \mathbb{K}^n \setminus \{0\}.$$

Lemma 4.1: *Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann positiv definit, wenn alle ihre (reellen) Eigenwerte positiv sind. Alle ihre Hauptdiagonalelemente sind positiv, und ihr betragsmäßig größtes Element liegt auf der Hauptdiagonalen.*

Beweis: Seien $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ die (ihrer Vielfachheiten entsprechend oft gezählten) Eigenwerte der symmetrischen matrix A und $\{w^1, \ldots, w^n\}$ eine zugehörige Orthonormalbasis von Eigenvektoren.

(i) Sei $\lambda \in \mathbb{R}$ Eigenwert und $v \in \mathbb{R}^n$, $\|v\|_2 = 1$, ein zugehöriger Eigenvektor von A,

$$Av = \lambda v.$$

Aus der positiven Definitheit von A folgt $\lambda = \lambda\|v\|_2^2 = (Av, v)_2 > 0$. Sind umgekehrt alle Eigenwerte von A positiv, so folgt für beliebigen Vektor $v = \sum_{i=1}^n \alpha_i w^i \neq 0$:

$$(Av, v)_2 = \sum_{i,j=1}^{n} \lambda_i \alpha_i \alpha_j (w^i, w^j)_2 = \sum_{i=1}^{n} \lambda_i \alpha_i^2 > 0.$$

(ii) Mit dem k-ten kartesischen Einheitsvektor e^k liefert die positive Definitheit von A:

$$a_{kk} = \sum_{i,j=1}^{n} a_{ij}\delta_{ik}\delta_{jk} = (Ae^k, e^k)_2 > 0.$$

(iii) Sei $a_{ij} \neq 0$ ein betragsmäßig größtes Element von A, und sei $i \neq j$. Wir wählen $x = e^i - \operatorname{sign}(a_{ij})e^j \neq 0$ und erhalten wieder aus der positiven Definitheit von A den folgenden Widerspruch:

$$\begin{aligned} 0 < (Ax, x)_2 &= (Ae^i, e^i)_2 - 2\operatorname{sign}(a_{ij})(Ae^i, e^j)_2 + \operatorname{sign}(a_{ij})^2(Ae^j, e^j)_2 \\ &= a_{ii} - 2\operatorname{sign}(a_{ij})a_{ij} + a_{jj} = a_{ii} - 2|a_{ij}| + a_{jj} \leq 0. \end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Im folgenden werden wir in Verbindung mit der Eigenschaft "positiv definit" stets auch die Eigenschaft "hermitesch" (bzw. "symmetrisch" im Reellen) einer Matrix annehmen. Dies ist im Komplexen automatisch gegeben, im Rellen aber eine zusätzliche Bedingung. Wir werden später sehen, daß lineare Gleichungssysteme mit positiv definiten Koeffizientenmatrizen besonders günstige Lösbarkeitseigenschaften besitzen.

4.1.3 Fehleranalyse

Wir kommen nun zur Fehleranalyse für lineare Gleichungssysteme

$$Ax = b \tag{4.1.7}$$

mit regulärer Koeffizientenmatrix $A \in \mathbb{K}^{n\times n}$. Die Matrix A und der Vektor b seien mit Fehlern δA bzw. δb behaftet, so daß ein gestörtes System

$$\tilde{A}\tilde{x} = \tilde{b} \tag{4.1.8}$$

mit $\tilde{A} = A + \delta A$, $\tilde{b} = b + \delta b$ und $\tilde{x} = x + \delta x$ gelöst wird. Wir wollen den Fehler δx in Abhängigkeit von δA und δb abschätzen. Dazu sei im folgenden $\|\cdot\|$ eine beliebige Vektornorm und entsprechend $\|\cdot\|$ die zugehörige natürliche Matrizennorm.

Hilfssatz 4.4: *Die Matrix $B \in \mathbb{K}^{n\times n}$ habe die Norm $\|B\| < 1$. Dann ist die Matrix $I + B$ regulär, und es gilt*

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \tag{4.1.9}$$

Beweis: Für alle $x \in \mathbb{K}^n$ gilt

$$\|(I + B)x\| \geq \|x\| - \|Bx\| \geq (1 - \|B\|)\|x\|.$$

Wegen $1 - \|B\| > 0$ ist also $I + B$ injektiv und folglich regulär. Mit der Abschätzung

$$\begin{aligned} 1 &= \|I\| = \|(I+B)(I+B)^{-1}\| = \|(I+B)^{-1} + B(I+B)^{-1}\| \\ &\geq \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|\,(1 - \|B\|) > 0 \end{aligned}$$

erhält man die behauptete Ungleichung. Q.E.D.

Nach diesen Vorbereitungen können wir den folgenden allgemeinen Störungssatz für lineare Gleichungssysteme beweisen:

Satz 4.1 (Störungssatz): *Die Matrix $A \in \mathbb{K}^{n\times n}$ sei regulär, und es sei*

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}. \tag{4.1.10}$$

Dann ist die gestörte Matrix $\tilde{A} = A + \delta A$ ebenfalls regulär, und für den relativen Fehler der Lösung gilt:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|/\|A\|} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \tag{4.1.11}$$

mit der sog. Konditionszahl $\mathrm{cond}(A) := \|A\|\,\|A^{-1}\|$ *von* A.

Beweis: Aufgrund der Voraussetzung ist

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\,\|\delta A\| < 1,$$

so daß auch $A + \delta A = A[I + A^{-1}\delta A]$ nach Hilfssatz 4.4 regulär ist. Aus

$$(A + \delta A)\tilde{x} = b + \delta b\,, \quad (A + \delta A)x = b + \delta A x$$

folgt dann für $\delta x = \tilde{x} - x$

$$(A + \delta A)\delta x = \delta b - \delta A x,$$

und damit unter Verwendung von Hilfssatz 4.4

$$\begin{aligned} \|\delta x\| &\leq \|(A+\delta A)^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &= \|\big(A(I + A^{-1}\delta A)\big)^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &= \|(I + A^{-1}\delta A)^{-1}A^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \|(I + A^{-1}\delta A)^{-1}\|\,\|A^{-1}\| \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|} \left\{ \|\delta b\| + \|\delta A\|\,\|x\| \right\} \\ &\leq \frac{\|A^{-1}\|\,\|A\|\,\|x\|}{1 - \|A^{-1}\|\,\|\delta A\|\,\|A\|\,\|A\|^{-1}} \left\{ \frac{\|\delta b\|}{\|A\|\,\|x\|} + \frac{\|\delta A\|}{\|A\|} \right\}. \end{aligned}$$

Wegen $\|b\| = \|Ax\| \leq \|A\|\,\|x\|$ folgt schließlich

$$\|\delta x\| \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|\|A\|^{-1}} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\} \|x\|,$$

was zu zeigen war. Q.E.D.

Die Konditionszahl $\mathrm{cond}(A)$ hängt offenbar von der bei ihrer Definition zugrundegelegten Vektornorm ab. Meistens verwendet man die Maximumnorm $\|\cdot\|_\infty$ oder die euklidische Norm $\|\cdot\|_2$. Im ersten Fall ist

$$\mathrm{cond}_\infty(A) := \|A\|_\infty \|A^{-1}\|_\infty$$

mit der maximalen Zeilensumme $\|\cdot\|_\infty$. Speziell für "hermitesche" Matrizen gilt nach Hilfssatz 4.3

$$\mathrm{cond}_2(A) := \|A\|_2 \|A^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

mit den betragsmäßig größten bzw. kleinsten Eigenwerten $\lambda_{\max}$ und $\lambda_{\min}$ von A; die Größe $\mathrm{cond}_2(A)$ wird auch die "Spektralkonditionszahl" von A genannt.

Ist $\mathrm{cond}(A)\|\delta A\|\,\|A\|^{-1} \ll 1$, so wird in Satz 4.1

$$\frac{\|\delta x\|}{\|x\|} \dot{\leq} \mathrm{cond}(A) \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\},$$

d.h.: $\mathrm{cond}(A)$ ist gerade der Verstärkungsfaktor, mit dem sich die relativen Fehler in A und b auf den in x auswirken. Diese Fehlerabschätzung erlaubt folgenden Schluß:

Regel 4.1.1: *Die Kondition von A sei $\mathrm{cond}(A) \sim 10^s$. Sind dann die Elemente von A und b mit einem relativen Fehler der Art*

$$\frac{\|\delta A\|}{\|A\|} \sim 10^{-k}, \quad \frac{\|\delta b\|}{\|b\|} \sim 10^{-k}\ (k > s)$$

behaftet, so muß mit einem relativen Fehler im Ergebnis der Größenordnung

$$\frac{\|\delta x\|}{\|x\|} \sim 10^{s-k}$$

gerechnet werden, d.h.: Im Fall $\|\cdot\| = \|\cdot\|_\infty$ verliert man s Stellen Genauigkeit.

Beispiel 4.2: Wir betrachten die folgende Koeffizientenmatrix A:

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|A\|_\infty = 2.1617, \quad \|A^{-1}\|_\infty = 1.513 \cdot 10^8 \Rightarrow \mathrm{cond}(A) \approx 3.3 \cdot 10^8.$$

Bei der Lösung des Gleichungssystems $Ax = b$ gehen also im ungünstigsten Fall 8 wesentliche Stellen an der Genauigkeit, mit der die Elemente a_{jk} und b_j gegeben sind, verloren. Dieses System ist extrem "schlecht konditioniert".

Wir demonstrieren anhand der Spektralkondition, daß die Abschätzung in Satz 4.1 im wesentlichen scharf ist. Sei A eine positiv definite $n \times n$-Matrix mit kleinstem und größtem Eigenwert λ_1 bzw. λ_n sowie zugehörigen normierten Eigenvektoren w_1 bzw. w_n. Wir wählen

$$\delta A \equiv 0, \quad b \equiv w_n, \quad \delta b \equiv \varepsilon w_1 \, (\varepsilon \neq 0).$$

Dann haben die Gleichungen $Ax = b$ und $A\tilde{x} = b + \delta b$ die Lösungen

$$x = \frac{1}{\lambda_n} w_n \,, \quad \tilde{x} = \frac{1}{\lambda_n} w_n + \varepsilon \frac{1}{\lambda_1} w_1 \,.$$

Folglich ist für $\delta x = \tilde{x} - x$

$$\frac{\|\delta x\|_2}{\|x\|_2} = \varepsilon \frac{\lambda_n}{\lambda_1} \frac{\|w_1\|_2}{\|w_n\|_2} = \mathrm{cond}_2(A) \frac{\|\delta b\|_2}{\|b\|_2} \,.$$

4.2 Eliminationsverfahren

Im folgenden diskutieren wir "direkte" Lösungsmethoden für (reelle) quadratische lineare Gleichungssysteme

$$Ax = b\,. \tag{4.2.12}$$

Besonders leicht lösbar sind gestaffelte Systeme, z.B. solche mit einer oberen Dreiecksmatrix $A = (a_{jk})$ als Koeffizientenmatrix

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n = b_1 \\ & & a_{22}x_2 & + & \ldots & + & a_{2n}x_n = b_2 \\ & & & & & & \vdots \\ & & & & & & a_{nn}x_n = b_n \end{array} \quad .$$

Im Falle $a_{jj} \neq 0$, $j = 1, \ldots, n$, erhält man die Lösung durch sog. "Rückwärtseinsetzen":

$$x_n = \frac{b_n}{a_{nn}}\,, \quad j = n-1, \ldots, 1: \quad x_j = \frac{1}{a_{jj}}\,(b_j - \sum_{k=j+1}^{n} a_{jk}\,x_k)\,.$$

Dazu sind offensichtlich $n^2/2 + O(n)$ *arithmetische Operationen* erforderlich (1 a.Op. := 1 Multiplikation (+ 1 Addition) oder 1 Division).

Das *klassische* direkte Verfahren zur Lösung (regulärer) Gleichungssysteme ist das Gaußsche[2] Eliminationsverfahren Dabei wird das gegebene System $Ax = b$ schrittweise in ein oberes Dreieckssystem $Rx = c$ umgeformt, welches dieselbe Lösung x besitzt und dann durch Rückwärtseinsetzen gelöst wird. Dazu stehen die folgenden elementaren Umformungen zur Verfügung:

- Vertauschung zweier Gleichungen,
- Addition des Vielfachen einer Gleichung zu einer anderen.

(Die Vertauschung zweier Spalten von A ist ebenfalls zulässig, wenn die Unbekannten x_i entsprechend umnumeriert werden.)

In der praktischen Durchführung des Gaußschen Eliminationsverfahrens wendet man die elementaren Umformungen auf die zusammengesetzte Matrix $[A, b]$ an. Im folgenden wird A als regulär angenommen. Zunächst setzt man $A^{(0)} \equiv A$, $b^{(0)} \equiv b$ und bestimmt $a^{(0)}_{r1} \neq 0$, $r \in \{1, \ldots, n\}$. (Solch ein Element existiert, da A sonst singulär wäre). Vertausche die 1-te und die r-te Zeile. Das Resultat sei $[\tilde{A}^{(0)}, \tilde{b}^{(0)}]$. Dann wird für $j = 2, \ldots, n$

[2]Carl Friedrich Gauß (1777-1855): bedeutender deutscher Mathematiker, Astronom und Physiker; wirkte in Göttingen; fundamentale Beiträge zur Arithmetik, Algebra und Geometrie, Begründer der modernen Zahlentheorie, Bestimmung von Planetenbahnen durch "Gauß-Ausgleich", Arbeiten zum Erdmagnetismus und Konstruktion eines elektromagnetischen Telegraphen.

das q_{j1}-fache der 1-ten Zeile von der j-ten Zeile abgezogen:

$$q_{j1} \equiv \tilde{a}_{j1}^{(0)}/\tilde{a}_{11}^{(0)} \; (= a_{r1}^{(0)}/a_{rr}^{(0)}), \quad a_{ji}^{(1)} := \tilde{a}_{ji}^{(0)} - q_{j1}\tilde{a}_{1i}^{(0)}, \quad b_j^{(1)} := \tilde{b}_j^{(0)} - q_{j1}\tilde{b}_1^{(0)}.$$

Das Resultat ist

$$[A^{(1)}, b^{(1)}] = \left[\begin{array}{cccc|c} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \dots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & & \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right].$$

Den Übergang $[A^{(0)}, b^{(0)}] \to [\tilde{A}^{(0)}, \tilde{b}^{(0)}] \to [A^{(1)}, b^{(1)}]$ kann man mit Hilfe von Matrizenmultiplikation beschreiben

$$[\tilde{A}^{(0)}, \tilde{b}^{(0)}] = P_1[A^{(0)}, b^{(0)}], \quad [A^{(1)}, b^{(1)}] \; = \; G_1[\tilde{A}^{(0)}, \tilde{b}^{(0)}],$$

wobei P_1 eine "Permutationsmatrix" und G_1 eine sog. "Frobenius-Matrix" der folgenden Gestalt sind:

$$P_1 = \begin{array}{c} \begin{array}{ccccccc} 1 & & & & \mathrm{r} & & \end{array} \\ \left[\begin{array}{ccccccc} 0 & \cdots & & & 1 & & \\ & 1 & & & & & \\ \vdots & & \ddots & & \vdots & & \\ & & & 1 & & & \\ 1 & \cdots & & & 0 & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{array}\right] \end{array} \begin{array}{c} 1 \\ \\ \\ \\ \mathrm{r} \\ \\ \\ \end{array} \qquad G_1 = \left[\begin{array}{ccccc} & & 1 & & \\ & & 1 & & \\ -q_{21} & 1 & & & \\ \vdots & & & \ddots & \\ -q_{n1} & & & & 1 \end{array}\right] 1$$

Beide Matrizen P_1 und G_1 sind regulär (Determinante $= \pm 1$), und es gilt

$$P_1^{-1} = P_1, \qquad G_1^{-1} \; = \left[\begin{array}{cccc} 1 & & & \\ q_{21} & 1 & & \\ \vdots & & \ddots & \\ q_{n1} & & & 1 \end{array}\right].$$

Die Gleichungssysteme $Ax = b$ und $A^{(1)}x = b^{(1)}$ haben offenbar dieselbe Lösung:

$$Ax = b \quad \Longleftrightarrow \quad A^{(1)}x = G_1P_1Ax = G_1P_1b = b^{(1)}.$$

Definition 4.6: *Das Element* $a_{r1} = \tilde{a}_{11}^{(0)}$ *heißt "Pivotelement" und der ganze Teilschritte seiner Bestimmung "Pivotsuche". Aus Gründen der numerischen Stabilität trifft man gewöhnlich die Wahl*

$$|a_{r1}| = \max_{1 \le j \le n} |a_{j1}| \,. \tag{4.2.13}$$

Der ganze Prozeß inkl. Zeilenvertauschung wird dann "Spaltenpivotierung" genannt. Sind die Elemente der Matrix A *von sehr unterschiedlicher Größenordnung, so empfiehlt es sich, sog. "Totalpivotierung" vorzunehmen. Diese besteht aus der Wahl*

$$|a_{rs}| = \max_{1 \le j,k \le n} |a_{jk}| \tag{4.2.14}$$

und anschließender Vertauschung der 1-ten mit der r*-ten Zeile und der 1-ten mit der* s*-ten Spalte. Entsprechend der Spaltenvertauschung müssen dann die Unbekannten* x_k *umnumeriert werden. Bei großen Gleichungssystemen ist die totale Pivotierung meist zu aufwendig, so daß man sich mit der Spaltenpivotierung begnügt.*

Die im 1. Schritt erzeugte Matrix $A^{(1)}$ ist wieder regulär. Dasselbe gilt auch für die um die erste Zeile und Spalte reduzierte Teilmatrix, so daß auf sie der Eliminationsprozeß analog zu Schritt 1 angewendet werden kann. Durch Weiterführung dieses Eliminationsprozesses erhält man in $n-1$ Schritten eine Kette von Matrizen

$$[A, b] \to [A^{(1)}, b^{(1)}] \to \ldots \to [A^{(n-1)}, b^{(n-1)}] =: [R, c] \,,$$

wobei

$$[A^{(i)}, b^{(i)}] = G_i P_i [A^{(i-1)}, b^{(i-1)}] \,, \quad [A^{(0)}, b^{(0)}] := [A, b] \,,$$

mit Permutationsmatrizen P_i und (regulären) Frobenius-Matrizen G_i der Form

$$P_i = \begin{array}{c} \\ 1 \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & \mathrm{i} & & & \mathrm{r} \end{array} \\ \begin{bmatrix} \ddots & & & & & & \\ & 1 & & & & & \\ & & 0 & \cdots & 1 & & \\ & & & 1 & & & \\ & & \vdots & & \ddots & & \\ & & & & & 1 & \\ & & 1 & \cdots & 0 & & \\ & & & & & & 1 \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \mathrm{i} \\ \\ \\ \\ \mathrm{r} \\ \\ \\ \\ \end{array}$$

$$G_i = \begin{array}{c} \\ 1 \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & \mathrm{i} & & \end{array} \\ \begin{bmatrix} \ddots & & & & \\ & 1 & & & \\ & -q_{i+1,i} & 1 & & \\ & \vdots & & \ddots & \\ & -q_{ni} & & & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \mathrm{i} \\ \\ \\ \end{array}$$

Das Endresultat

$$[R, c] = G_{n-1} P_{n-1} \dots G_1 P_1 [A, b] \tag{4.2.15}$$

entspricht einem oberen Dreieckssystem $Rx = c$, welches dieselbe Lösung wie das Ausgangssystem $Ax = b$ besitzt.

Im i-ten Eliminationsschritt $[A^{(i-1)}, b^{(i-1)}] \to [A^{(i)}, b^{(i)}]$ werden in der i-ten Spalte die Elemente unterhalb der Diagonalen annuliert. Den frei werdenden Platz benutzt man zur Abspeicherung der wesentlichen Elemente $q_{i+1,i}, \dots, q_{n,i}$ der Frobenius-Matrizen G_i^{-1} $(i = 1, \dots, n-1)$. Da im i-ten Eliminationsschritt die vorausgehenden Zeilen 1 bis i nicht verändert werden, arbeitet man also mit Matrizen der Form

$$\left[\begin{array}{ccccccc|c}
r_{11} & r_{12} & \cdots & r_{1i} & r_{1,i+1} & \cdots & r_{1n} & c_1 \\
\lambda_{21} & r_{22} & \cdots & r_{2i} & r_{2,i+1} & \cdots & r_{2n} & c_2 \\
\lambda_{31} & \lambda_{32} & & r_{3i} & r_{3,i+1} & \cdots & r_{3n} & c_3 \\
\vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{i1} & \lambda_{i2} & & r_{ii} & r_{i,i+1} & \cdots & r_{in} & c_i \\
\lambda_{i+1,1} & \lambda_{i+2,2} & & \lambda_{i+1,i} & a^{(i)}_{i+1,i+1} & \cdots & a^{(i)}_{i+1,n} & b^{(i)}_{i+1} \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{n,1} & \lambda_{n,2} & \cdots & \lambda_{n,i} & a^{(i)}_{n,i+1} & \cdots & a^{(i)}_{n,n} & b^{(i)}_{n}
\end{array}\right]$$

Dabei sind die Subdiagonalelemente $\lambda_{k+1,k}, \dots, \lambda_{nk}$ der k-ten Spalte Permutationen der Elemente $q_{k+1,k}, \dots, q_{nk}$ von G_k^{-1}, da die Zeilenvertauschungen (nur diese!) an der gesamten Matrix vorgenommen werden. Als Endresultat erhält man eine Matrix

$$\left[\begin{array}{cccc|c}
r_{11} & & \cdots & r_{1n} & c_1 \\
l_{21} & r_{22} & & r_{2n} & c_2 \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
l_{n1} & \cdots & l_{n,n-1} & r_{nn} & c_n
\end{array}\right] .$$

Satz 4.2 (LR-Zerlegung): *Die Matrizen*

$$L = \begin{bmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{bmatrix}$$

bilden eine sog. "LR-Zerlegung" der Matrix PA:

$$PA = LR, \qquad P = P_{n-1} \cdots P_1 . \tag{4.2.16}$$

Diese Zerlegung ist im Falle $P = I$ eindeutig bestimmt.

Beweis: Wir führen den Beweis für den Fall, daß keine Pivotierung erforderlich ist, d.h.: $P_i = I$. Dann ist $R = G_{n-1} \cdots G_1 A$ bzw.

$$G_1^{-1} \cdots G_{n-1}^{-1} R = A\,.$$

Wegen $L = G_1^{-1} \cdots G_{n-1}^{-1}$ folgt die Behauptung. Zum Nachweis der Eindeutigkeit, seien nun $A = L_1 R_1 = L_2 R_2$ zwei LR-Zerlegungen. Dann ist

$$L_2^{-1} L_1 = R_2 R_1^{-1} = I\,,$$

da $L_2^{-1} L_1$ untere Dreiecksmatrix mit Einsen auf der Hauptdiagonalen und $R_2 R_1^{-1}$ obere Dreiecksmatrix ist. Folglich ist $L_1 = L_2$ und $R_1 = R_2$. Q.E.D.

Lemma 4.2: *Die zur Lösung eines $n \times n$ Gleichungssystems $Ax = b$ mit Hilfe des Gaußschen Eliminationsverfahrens erforderliche Anzahl von arithmetischen Operationen ("a. Op.") ist*

$$N_{\text{Gauß}}(n) = \frac{n^3}{3} + O(n^2).$$

Dasselbe gilt für die Bestimmung der Dreieckszerlegung $PA = LR$.

Beweis: Der k-te Eliminationsschritt

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \quad i,j = k,\ldots,n,$$

erfordert $n-k$ Divisionen sowie $(n-k)+(n-k)^2$ Multiplikationen und Additionen; also zusammen

$$\sum_{k=1}^{n-1} k^2 + O(n^2) = \frac{1}{3} n^3 + O(n^2) \quad \text{a. Op.}$$

für die $n-1$ Schritte der Vorwärtselimination. Damit werden alle Elemente der Zerlegungsmatrizen L und R bestimmt. Q.E.D.

Beispiel 4.3: Mit $\boxed{\cdot}$ wird das Pivotelement markiert.

$$\begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix} \quad \rightarrow$$

Pivotierung

$\boxed{3}$	1	6	2
2	1	3	7
1	1	1	4

Elimination

3	1	6	2
2/3	1/3	−1	17/3
1/3	$\boxed{2/3}$	−1	10/3

$\rightarrow$

Pivotierung

3	1	6	2
1/3	2/3	−1	10/3
2/3	1/3	−1	17/3

Elimination			
3	1	6	2
1/3	2/3	−1	10/3
2/3	1/2	−1/2	4

$$\rightarrow \qquad \begin{aligned} x_3 &= -8 \\ x_2 &= \tfrac{3}{2}(\tfrac{10}{3} - x_3) = -7 \\ x_1 &= \tfrac{1}{3}(2 - x_2 - 6x_3) = 19\,. \end{aligned}$$

LR-Zerlegung:

$$P_1 = I\,, \quad P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\,,$$

$$PA = \begin{bmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} = LR = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{bmatrix}\,.$$

Beispiel 4.4: Zur Demonstration der Bedeutung der Pivotierung beim Gaußschen Eliminationsverfahren betrachten wir das folgende Gleichungssystem

$$\begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{4.2.17}$$

mit der exakten Lösung $x_1 = 1.00010001$, $x_2 = 0.99989999$. Bei 3-stelliger Gleitpunktrechnung mit korrekter Rundung erhält man:

a) ohne Pivotierung:

x_1	x_2	
$0.1 \cdot 10^{-3}$	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
0	$-0.1 \cdot 10^{5}$	$-0.1 \cdot 10^{5}$
$x_2 = 1\,,$	$x_1 = 0$	

b) mit Pivotierung:

x_1	x_2	
$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$	$0.2 \cdot 10^{1}$
0	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
$x_2 = 1\,,$	$x_1 = 1$	

Beispiel 4.5: Der positive Effekt der Spaltenpivotierung ist allerdings nur dann gesichert, wenn die (betragsmäßigen) Zeilensummen der Matrix in etwa gleich groß sind. Als Beispiel betrachte man das Gleichungssystem

$$\begin{bmatrix} 2 & 20000 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 20000 \\ 2 \end{bmatrix}\,,$$

welches aus (4.2.17) durch Multiplikation der ersten Zeile mit dem Faktor 20000 hervorgeht. Da nun in der ersten Spalte das betragsgrößte Element in der Diagonalen steht, liefert der Gauß-Algorithmus mit und ohne Spaltenpivotierung dasselbe inakzeptable Resultat $(x_1, x_2)^T = (0, 1)^T$. Man führt daher vor der Rechnung eine sog. “Äquilibrierung”

durch, das heißt, man multipliziert mit einer Diagonalmatrix D

$$Ax = b \quad \rightarrow \quad DAx = Db\,, \qquad d_i = \Big(\sum_{j=1}^{n} |a_{ij}|\Big)^{-1},$$

die alle Zeilensummen der Matrix auf 1 transformiert. Eine verbesserte Stabilisierung im Fall stark unterschiedlicher Größenordnung der Matrixeinträge ist die "totale" Pivotierung. Vor der Durchführung wird hier eine Äquilibrierung (zeilenweise und spaltenweise) vorgenommen.

4.2.1 Konditionierung der Gauß-Elimination

Wir diskutieren nun noch die Konditionierung des Lösens eines linearen Gleichungssystems $Ax = b$ mit Hilfe des Gaußschen Eliminationsverfahrens. Die (reguläre) Matrix A besitze mit Spaltenpivotierung eine LR-Zerlegung der Form $PA = LR$. Dann gilt

$$R = L^{-1}PA, \quad R^{-1} = (PA)^{-1}L.$$

Wegen der Spaltenpivotierung sind die Elemente der Dreiecksmatrix L alle kleiner gleich eins, und es gilt somit

$$\mathrm{cond}_\infty(L) = \|L\|_\infty \|L^{-1}\|_\infty \le n.$$

Folglich ist

$$\begin{aligned}\mathrm{cond}_\infty(R) &= \|R\|_\infty \|R^{-1}\|_\infty = \|L^{-1}PA\|_\infty \|(PA)^{-1}L\|_\infty \\ &\le \|L^{-1}\|_\infty \|PA\|_\infty \|(PA)^{-1}\|_\infty \|L\|_\infty \le n^2\,\mathrm{cond}_\infty(PA).\end{aligned}$$

Nach dem allgemeinen Störungssatz gilt dann für die Lösung des Systeme $LRx = Pb$:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \mathrm{cond}_\infty(L)\mathrm{cond}_\infty(R)\frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty} \le n^2\,\mathrm{cond}_\infty(PA)\frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty}.$$

Die Konditionierung des Ausgangssystems wird also durch die LR-Zerlegung im schlimmsten Fall nur mit n^2 verschlechtert.

Wir geben zum Abschluß noch ein Resultat von Wilkinson an, das die Fortpflanzung des Rundungsfehlers im Verlaufe der Gaußschen Elimination beschreibt.

Satz 4.3 (Rundungsfehlereinfluß): *Die Matrix $A \in \mathbb{R}^{n\times n}$ sei regulär, und das Gleichungssystem $Ax = b$ werde mit Gaußscher Elimination mit Spaltenpivotierung gelöst. Dann ist die unter dem Einfluß von Rundungsfehlern tatsächlich berechnete Lösung $x+\delta x$ exakte Lösung eines gestörten Systems $(A + \delta A)(x + \delta x) = b$, wobei (*eps *= Maschinengenauigkeit)*

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \le 1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}. \tag{4.2.18}$$

In Verbindung mit der Fehlerabschätzung von Satz 4.1 ergibt das Wilkinsonsche Resultat die folgende Abschätzung für den Rundungsfehlereinfluß:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|_\infty / \|A\|_\infty} \{1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}\}\,. \tag{4.2.19}$$

Diese Abschätzung ist, wie die Praxis zeigt, viel zu pessimistisch, da sie am ungünstigsten Fall ausgerichtet ist und keine Rundungsfehlerauslöschungen berücksichtigt. Zur Erfassung der letzteren wäre eine statistische Theorie erforderlich. Außerdem gilt die obige Abschätzung allgemein für "vollbesetzte" Matrizen. Für "dünnbesetzte" Matrizen sind wesentlich günstigere Resultate zu erwarten. Insgesamt sieht man, daß der Gaußsche Eliminationsprozeß (in Abhängigkeit von der Dimension n) ein gutartiger numerischer Algorithmus ist, d.h.: Der Rundungsfehlereinfluß kann abgeschätzt werden allein in Abhängigkeit von der Kondition $\mathrm{cond}(A)$, die ja auch die Konditionierung der numerischen Aufgabe selber beschreibt.

4.2.2 Nachiteration

Wir diskutieren nun noch einige Varianten und weitere Anwendungsmöglichkeiten des Gaußschen Eliminationsverfahrens. Das Gaußsche Eliminationsverfahren überführt ein Gleichungssystem $Ax = b$ in ein oberes Dreieckssystem $Rx = c$, aus dem sich die Lösung x durch einfaches Rückwärtsauflösen berechnen läßt. Nach Satz 4.2 ist dieser Prozeß gleichbedeutend mit der Erstellung einer Dreieckszerlegung $PA = LR$ und der anschließenden Lösung der beiden gestaffelten Systeme

$$Ly = Pb\,, \quad Rx = y\,. \tag{4.2.20}$$

Diese Variante des Gaußschen Algorithmus ist insbesondere dann vorzuziehen, wenn dasselbe Gleichungssystem nacheinander für verschiedene rechte Seiten b gelöst werden soll. Aufgrund des unvermeidlichen Rundungsfehlers erhält man in der Praxis nur eine fehlerhafte LR-Zerlegung

$$\tilde{L}\tilde{R} \neq PA$$

und damit nur eine Näherungslösung x^0 mit dem (exakten) "Defekt"

$$\hat{d}^0 := b - Ax^0 \neq 0\,.$$

Unter Verwendung der bereits erstellten Dreieckszerlegung $\tilde{L}\tilde{R} \sim PA$ löst man nun (näherungsweise) die sog. "Defektgleichung"

$$Ak = \hat{d}^0, \qquad \tilde{L}\tilde{R}k^1 = \hat{d}^0, \tag{4.2.21}$$

und erhält daraus eine Korrektur k^1 für x^0 :

$$x^1 := x^0 + k^1\,. \tag{4.2.22}$$

Hätte man die Defektgleichung exakt gelöst, d.h. $k^1 \equiv k$, so wäre

$$Ax^1 = Ax^0 + Ak = Ax^0 - b + b + \hat{d}^0 = b\,,$$

d.h.: $x^1 = x$ wäre die exakte Lösung des Systems $Ax = b$. I. Allg. wird x^1 auch bei fehlerhafter Lösung der Defektgleichung eine bessere Näherung zu x als x^0 sein. Dazu ist es jedoch erforderlich, den Defekt d mit *erhöhter* Genauigkeit zu berechnen. Dies wird durch die folgende Fehleranalyse belegt (der Einfachheit halber sei $P = I$):

Wir nehmen an, daß sich der relative Fehler bei der LR-Zerlegung der Matrix A durch eine kleine Zahl ε beschränken läßt. Nach dem allgemeinen Störungssatz 4.1 gilt dann die Abschätzung

$$\frac{\|x^0 - x\|}{\|x\|} \le \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\frac{\|A-\tilde{L}\tilde{R}\|}{\|A\|}} \underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\,\varepsilon}\,.$$

Der Verlust von Stellen entspricht der Größe von $\mathrm{cond}(A)$. Zusätzlich auftretende Rundungsfehler werden vernachlässigt. Den exakten Defekt $\hat{d}^0$ ersetzen wir durch den Ausdruck $d^0 := b - \tilde{A}x^0$, wobei $\tilde{A}$ eine genauere Approximation für A ist,

$$\frac{\|A - \tilde{A}\|}{\|A\|} \le \tilde{\varepsilon} \ll \varepsilon\,.$$

Nach Konstruktion gilt

$$\begin{aligned} x^1 = x^0 + k^1 &= x^0 + (\tilde{L}\tilde{R})^{-1}[b - \tilde{A}x^0] \\ &= x^0 + (\tilde{L}\tilde{R})^{-1}[Ax - Ax^0 + (A - \tilde{A})\,x^0], \end{aligned}$$

und daher

$$\begin{aligned} x^1 - x &= x^0 - x - (\tilde{L}\tilde{R})^{-1}A(x^0 - x) + (\tilde{L}\tilde{R})^{-1}(A - \tilde{A})\,x^0 \\ &= (\tilde{L}\tilde{R})^{-1}[\tilde{L}\tilde{R} - A](x^0 - x) + (\tilde{L}\tilde{R})^{-1}(A - \tilde{A})\,x^0. \end{aligned}$$

Wegen

$$\tilde{L}\tilde{R} = A - A + \tilde{L}\tilde{R} = A\big(I - A^{-1}(A - \tilde{L}\tilde{R})\big)$$

folgt mit Hilfssatz 4.4:

$$\begin{aligned} \|(\tilde{L}\tilde{R})^{-1}\| &\le \|A^{-1}\|\,\|\,[I - A^{-1}(A - \tilde{L}\tilde{R})]^{-1}\| \\ &\le \frac{\|A^{-1}\|}{1 - \|A^{-1}(A - \tilde{L}\tilde{R})\|} \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|A - \tilde{L}\tilde{R}\|} = \frac{\|A^{-1}\|}{1 - \mathrm{cond}(A)\,\frac{\|A-\tilde{L}\tilde{R}\|}{\|A\|}}. \end{aligned}$$

Dies impliziert schließlich

$$\frac{\|x^1 - x\|}{\|x\|} \;\sim\; \mathrm{cond}(A)\,\Big[\underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\,\varepsilon}\;\underbrace{\frac{\|x^0 - x\|}{\|x\|}}_{\sim\,\mathrm{cond}(A)\varepsilon} + \underbrace{\frac{\|A - \tilde{A}\|}{\|A\|}}_{\sim\,\tilde{\varepsilon}}\;\frac{\|x^0\|}{\|x\|}\Big]\,.$$

Diese Korrektur der Lösung kann natürlich iteriert werden, in dem die jeweils neuen Näherungen x^i wieder in die Defektgleichung eingesetzt werden. Diesen Prozeß nennt man "Nachiteration"; in der Praxis wird der Fehler in x schon durch wenige Korrekturschritte (meist $2-3$) auf die Größenordnung der Genauigkeit der Defektauswertung gedrückt, d.h.: $\|x^3 - x\|/\|x\| \sim \tilde{\varepsilon}$.

Beispiel 4.6: Das Gleichungssystem

$$\begin{bmatrix} 1.05 & 1,02 \\ 1.04 & 1,02 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

hat die exakte Lösung $x = (-100, 103.921\ldots)^T$. Das Gaußsche Eliminationsverfahren ergibt bei Verwendung 3-stelliger Gleitpunktarithmetik (mit korrekter Rundung) die genäherten Zerlegungsmatrizen

$$\tilde{L} = \begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix}, \qquad \tilde{R} = \begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix},$$

$$\tilde{L}\tilde{R} - A = \begin{bmatrix} 0 & 0 \\ 5 \cdot 10^{-4} & 2 \cdot 10^{-4} \end{bmatrix} \qquad \text{(im Rahmen der Maschinengenauigkeit korrekt)}.$$

Die damit bestimmte "Lösung" $x^0 = (-97, 1.101)^T$ hat den Defekt

$$d^0 = b - Ax^0 = \begin{cases} (0,0)^T & \text{3-stellige Rechnung} \\ (0,065, 0,035)^T & \text{6-stellige Rechnung}. \end{cases}$$

Die approximative Korrekturgleichung

$$\begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix} \begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} k_1^1 \\ k_2^1 \end{bmatrix} = \begin{bmatrix} 0.065 \\ 0.035 \end{bmatrix}$$

hat (3-stellige Rechnung) die Lösung $k^1 = (,)^T$. Die durch Nachkorrektur verbesserte Lösung ist also

$$x^1 = x^0 + k^1 = (-99.9, 104)^T,$$

welche deutlich genauer als die erste Näherung x^0 ist.

4.2.3 Determinantenberechnung

Für quadratische Matrizen gilt der Determinantensatz

$$\det(AB) = \det(A)\ \det(B). \tag{4.2.23}$$

Für die durch Gaußsche Elimination aus der gegebenen Matrix A gewonnene Dreiecksmatrix

$$R = G_{n-1}P_{n-1}\cdots G_1P_1A$$

folgt somit unter Beachtung von

$$\det(P_i^{-1}) = \det(P_i) = -1\,, \quad \det(G_i^{-1}) = 1\,,$$

die Beziehung

$$\det(A) = \det(P_1^{-1}G_1^{-1}\cdots P_{n-1}^{-1}G_{n-1}^{-1}R) = \pm\det(R) = \pm\prod_{j=1}^{n} r_{jj}\,. \tag{4.2.24}$$

Das Vorzeichen in $\det(A)$ ist $+/-$, je nachdem, ob eine gerade oder ungerade Anzahl von Zeilenvertauschungen vorgenommen wurde. Läßt sich im Verlaufe des Eliminationsprozesses einmal in einer Spalte kein von Null verschiedenes Pivotelement finden, so ist die Matrix A singulär und folglich $\det(A) = 0$. (Man beachte, daß bei Rechnung in Gleitpunktarithmetik aufgrund des Rundungsfehlers durchaus auch im Falle $\det(A) = 0$ der tatsächlich berechnete Wert $\neq 0$ sein kann!)

4.2.4 Rangbestimmung

Ist die Elimination durchführbar, d.h. läßt sich immer ein Pivotelement $\neq 0$ finden, und ist schließlich auch das letzte Diagonalelement $a_{n,n}^{(n-1)} \neq 0$, so ist $\det(A) \neq 0$, d.h.

$$\mathrm{Rang}(A) = n$$

(dies natürlich nur bei Vernachlässigung der Rundungsfehler!). Gilt dagegen im i-ten Eliminationsschritt für alle Elemente in der i-ten Spalte

$$a_{ji}^{(i-1)} = 0\,, \quad j = i, \ldots, n\,,$$

so ist A singulär. In diesem Fall wird zur weiteren Rangberechnung Totalpivotierung vorgenommen:

$$|a_{rs}^{(i-1)}| = \max_{j,k=1,\ldots,n} |a_{jk}^{(i-1)}|\,.$$

(Zeilen- und Spaltenvertauschungen ändern $\mathrm{Rang}(A)$ nicht!) Gilt dann nach dem i-ten Eliminationsschritt

$$a_{jk}^{(i)} = 0\,, \quad j,k = i+1, \ldots, n\,,$$

so ist $\mathrm{Rang}(A) = i$. Dieser Prozeß kann natürlich auch zur Rangbestimmung bei *nicht* quadratischen Matrizen verwendet werden.

4.2.5 Inversenberechnung (Gauß-Jordan-Algorithmus)

Grundsätzlich kann die Inverse A^{-1} einer regulären Matrix A wie folgt berechnet werden:

(i) Berechnung der LR-Zerlegung von PA;

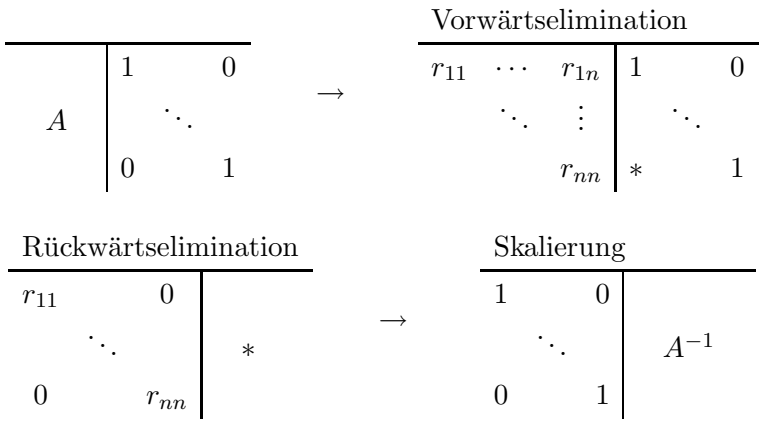
(ii) Lösung der gestaffelten Systeme

$$Ly^{(i)} = Pe^{(i)},\ Rx^{(i)} = y^{(i)}, \quad i = 1, \ldots, n,$$

mit den kartesischen Basisvektoren $e^{(i)}$ des $\mathbb{R}^n$;

(iii) $A^{-1} = [x^{(1)}, \ldots, x^{(n)}]$.

Praktischer ist jedoch eine simultane Elimination (hier ohne Berechnung der Matrizen L und R), die direkt auf die Inverse führt (ohne Zeilenvertauschungen):



Beispiel 4.7: Es markiere $\boxed{\cdot}$ das Pivotelement.

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \quad : \quad \text{Vorwärtselimination} \quad \left(\begin{array}{ccc|ccc} \boxed{3} & 1 & 6 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \rightarrow$$

$$\rightarrow \quad \text{Zeilenvertauschung} \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \\ 0 & \boxed{2/3} & -1 & -1/3 & 0 & 1 \end{array}\right) \rightarrow \quad \text{Vorwärtselimination} \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \end{array}\right) \rightarrow$$

$$\rightarrow \quad \text{Rückwärtselimination} \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array}\right) \rightarrow \quad \text{Rückwärtselimination} \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 0 & -5 & 12 & -6 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array}\right) \rightarrow$$

$$\rightarrow \quad \begin{array}{ccc|ccc} & \text{Skalierung} & & & & \\ 3 & 0 & 0 & -6 & 15 & -9 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow \quad \begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & 5 & -3 \\ 0 & 1 & 0 & 1 & -3 & 3 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}$$

$$\Rightarrow \quad A^{-1} = \begin{bmatrix} -2 & 5 & -3 \\ 1 & -3 & 3 \\ 1 & -2 & 1 \end{bmatrix} .$$

Eine alternative Methode zur Berechnung der Inversen einer Matrix ist das sog. "Austauschverfahren" (manchmal auch "Gauß-Jordan[3]-Algorithmus" oder "Pivotierungsmethode") genannt. Gegeben sei ein (nicht notwendig quadratisches) lineares Gleichungssystem

$$Ax = y \qquad \text{mit} \quad A \in \mathbb{R}^{m\times n}\,, \quad x \in \mathbb{R}^n\,, \quad y \in \mathbb{R}^m\,. \tag{4.2.25}$$

Eine Lösung wird berechnet durch sukzessiven Austausch der Komponenten von x gegen solche von y. Ist ein Matrixelement $a_{pq} \neq 0$, so kann die p-te Gleichung nach x_q aufgelöst werden:

$$x_q = -\frac{a_{p1}}{a_{pq}}x_1 - \ldots - \frac{a_{p,q-1}}{a_{pq}}x_{q-1} + \frac{1}{a_{pq}}y_p - \frac{a_{p,q+1}}{a_{pq}}x_{q+1} - \ldots - \frac{a_{pn}}{a_{pq}}x_n\,.$$

Durch Substitution von x_q in den anderen Gleichungen,

$$a_{j1}x_1 + \ldots + a_{j,q-1}x_{q-1} + a_{jq}\boxed{x_q} + a_{j,q+1}x_{q+1} + \ldots + a_{jn}x_n = y_j\,,$$

erhält man für $j = 1, \ldots, m$, $j \neq p$:

$$\left[a_{j1} - \frac{a_{jq}a_{p1}}{a_{pq}}\right] x_1 + \ldots + \left[a_{j,q-1} - \frac{a_{jq}a_{p,q-1}}{a_{pq}}\right] x_{q-1} + \frac{a_{jq}}{a_{pq}}y_p \quad +$$
$$+ \left[a_{j,q+1} - \frac{a_{jq}a_{p,q+1}}{a_{pq}}\right] x_{q+1} + \ldots + \left[a_{jn} - \frac{a_{jq}a_{pn}}{a_{pq}}\right] x_n \;=\; y_j\,.$$

Das Resultat ist ein zum Ausgangssystem äquivalentes System

$$\tilde{A}\begin{bmatrix} x_1 \\ \vdots \\ y_p \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ x_q \\ \vdots \\ y_m \end{bmatrix}\,, \tag{4.2.26}$$

[3]Marie Ennemond Camille Jordan (1838-1922): französischer Mathematiker; Prof. in Paris; Beiträge zur Algebra, Gruppentheorie, Analysis und Topologie.

wobei die Elemente der Matrix $\tilde{A}$ wie folgt bestimmt sind:

$$
\begin{array}{lll}
\text{Pivotelement} & : & \tilde{a}_{pq} = 1/a_{pq}\,, \\
\text{Pivotzeile} & : & \tilde{a}_{pk} = a_{pk}/a_{pq}\,, \quad k = 1,\ldots,n\,, \quad k \neq q\,, \\
\text{Pivotspalte} & : & \tilde{a}_{jq} = a_{jq}/a_{pq}\,, \quad j = 1,\ldots,m\,, \quad j \neq p\,, \\
\text{sonstige} & : & \tilde{a}_{jk} = a_{jk} - a_{jq}a_{pk}/a_{pq}\,, \quad j = 1,\ldots,m,\ j \neq p,\ k = 1,\ldots,n\,,\ k \neq q.
\end{array}
$$

Gelingt es, durch Fortsetzung des Verfahrens alle Komponenten von x durch solche von y zu ersetzen, so hat man eine explizite Darstellung der Lösung von $y = A^{-1}x$. Im Fall $m = n$ ergibt sich so auch die Inverse A^{-1}, allerdings i. Allg. mit vertauschten Zeilen und Spalten. Bei der Festlegung des Pivotelementes empfiehlt es sich aus Stabilitätsgründen, unter allen in Frage kommenden a_{pq} jeweils eines von möglichst großem Betrag zu wählen.

Satz 4.4 (Gauß-Jordan-Algorithmus): *Es können genau* $r = \mathrm{Rang}(A)$ *Austauschschritte durchgeführt werden.*

Beweis: Das Verfahren breche nach r Austauschschritten ab. O.B.d.A. seien $x_1,\ldots,x_r$ gegen $y_1,\ldots,y_r$ ausgetauscht, so daß das resultierende System die Gestalt hat:

$$
\begin{array}{l}
\left.\begin{array}{l}\\ \\ \\ \end{array}\right\} r \\
\left.\begin{array}{l}\\ \\ \\ \end{array}\right\} m-r
\end{array}
\left[\begin{array}{c|c}
* & * \\
\hline
* & 0
\end{array}\right]
\begin{array}{c}
\underbrace{}_{r}\ \underbrace{}_{n-r}
\end{array}
\left[\begin{array}{c} y_1 \\ \vdots \\ y_r \\ x_{r+1} \\ \vdots \\ x_n \end{array}\right]
= \left[\begin{array}{c} x_1 \\ \vdots \\ x_r \\ y_{r+1} \\ \vdots \\ y_m \end{array}\right].
$$

Wählt man nun $y_1 = \cdots = y_r = 0, \quad x_{r+1} = \lambda_1, \cdots, \quad x_n = \lambda_{n-r}$, so sind die $x_1,\cdots,x_r$ dadurch eindeutig bestimmt, und es folgt $y_{r+1} = \cdots = y_m = 0$. Für beliebige Werte $\lambda_1,\cdots,\lambda_{n-r}$ ist also

$$
A \left[\begin{array}{c} x_1(\lambda_1,\cdots,\lambda_{n-r}) \\ \vdots \\ x_r(\lambda_1,\cdots,\lambda_{n-r}) \\ \lambda_1 \\ \vdots \\ \lambda_{n-r} \end{array}\right]
= \left[\begin{array}{c} 0 \\ \\ \vdots \\ \\ 0 \end{array}\right],
$$

d.h.: $\dim(\mathrm{Kern}(A)) \geq n - r$. Andererseits ist wegen der möglichen freien Wahl von $y_1,\cdots,y_r$ offenbar $\dim(\mathrm{Bild}(A)) \geq r$. Da $\dim(\mathrm{Bild}(A)) + \dim(\mathrm{Kern}(A)) = n$ ist, folgt $\mathrm{Rang}(A) = \dim(\mathrm{Bild}(A)) = r$. Q.E.D.

Für ein quadratisches Gleichungssystem mit regulärer Koeffizientenmatrix A ist das Gauß-Jordan-Verfahren zur Berechnung von A^{-1} also stets durchführbar.

Beispiel 4.8:

$$\begin{bmatrix} 1 & 2 & 1 \\ -3 & -5 & -1 \\ -7 & -12 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Austauschschritte: Mit $\boxed{\cdot}$ wird das Pivotelement markiert.

x_1	x_2	x_3	
1	2	1	y_1
−3	−5	−1	y_2
−7	$\boxed{-12}$	−2	y_3

x_1	y_3	x_3	
−1/6	−1/6	$\boxed{2/3}$	y_1
−1/12	5/12	−1/6	y_2
−7/12	−1/12	−1/6	x_2

x_1	y_3	y_1	
1/4	1/4	3/2	x_3
$\boxed{-1/8}$	3/8	−1/4	y_2
−5/8	−1/8	−1/4	x_2

y_2	y_3	y_1	
−2	1	1	x_3
−8	3	−2	x_1
5	−2	1	x_2

Inverse: $\begin{bmatrix} -2 & -8 & 3 \\ 1 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$

Lemma 4.3: *Die zur Invertierung einer regulären $n \times n$-Matrix mit Hilfe der simultanen Elimination oder des Gauß-Jordan-Algorithmus erforderliche Anzahl von arithmetischen Operationen ("a. Op.") ist*

$$N_{\text{Gauß-Jordan}}(n) = n^3 + O(n^2).$$

Beweis: (i) Die $n-1$ Schritte der Vorwärtselimination an der Matrix A erfordert $\frac{1}{3}n^3 + O(n^2)$ a. Op.. Die simultane Bearbeitung der Spalten der Einheitsmatrix erfordert wegen der Dreicksgestalt von I zusätzliche $\frac{1}{6}n^3 + O(n^2)$ a. Op.. Die Rückwärtselimination zur Erstellung der Einheitsmatrix links erfordert schließlich nochmal

$$(n-1)n + (n-2)n + \ldots + n = \frac{n(n-1)}{2}n = \frac{1}{2}n^3 + O(n^2)$$

Multiplikationen und Additionen und nachfolgend n^2 Divisionen. Für die gesamte Berechnung der Inversen ergibt sich also:

$$\frac{1}{3}n^3 + \frac{1}{6}n^3 + \frac{1}{2}n^3 + O(n^2) = n^3 + O(n^2).$$

(ii) Beim Gauß-Jordan-Verfahren erfordert der k-te Austauschschritt $2n+1$ Divisionen in Pivotzeile und -spalte und $(n-1)^2$ Multiplikationen und Additionen für den Update der Restmatrix, also insgesamt $n^2+O(n)$ a. Op. Zur Berechnung der Inversen sind n Austauschschritte durchzuführen, so daß sich ebenfalls ein Gesamtaufwand von $n^3+O(n^2)$ a. Op. ergibt. Q.E.D.

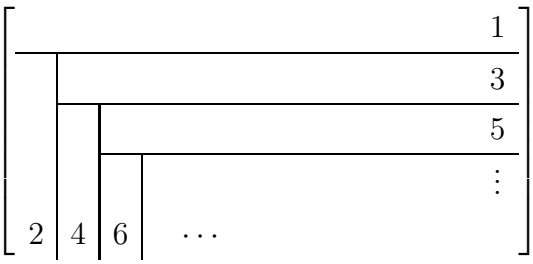
4.2.6 Direkte LR-Zerlegung

Der Gaußsche Algorithmus zur Berechnung der LR-Zerlegung $A=LR$ (falls sie existiert) kann auch in direkter Form geschrieben werden, bei der die Elemente l_{jk} von L und r_{jk} von R rekursiv berechnet werden.

Die Gleichung $A=LR$ ergibt n^2 Bestimmungsgleichungen für die n^2 unbekannten Größen $r_{jk}\,,\ j\le k\,,\ l_{jk}\,,\ j>k\,(l_{jj}=1)$:

$$a_{jk}=\sum_{i=1}^{\min(j,k)} l_{ji}\,r_{ik}. \tag{4.2.27}$$

Die Reihenfolge der Berechnung von $l_{jk},\,r_{jk}$ ist zunächst noch offen. Beim sog. "Algorithmus von Crout[4] " wird die Matrix $A=LR$ wie folgt parkettiert:



Die einzelnen Schritte des Algorithmus sind dann $(l_{ii}\equiv 1)$:

$$\begin{aligned}
&k=1,\cdots,n: && a_{1k}=\sum_{i=1}^{1} l_{1i}r_{ik} \;\Rightarrow\; r_{1k}:=a_{1k},\\
&j=2,\cdots,n: && a_{j1}=\sum_{i=1}^{1} l_{ji}r_{i1} \;\Rightarrow\; l_{j1}:=r_{11}^{-1}a_{j1},\\
&k=2,\cdots,n: && a_{2k}=\sum_{i=1}^{2} l_{2i}r_{ik} \;\Rightarrow\; r_{2k}:=a_{2k}-l_{21}r_{1k},\\
&\vdots
\end{aligned}$$

[4]Prescott D. Crout (????-????): US-Amerikanischer Mathematiker und Ingenieur; Professor am Massachusetts Institute of Technology (MIT); Beiträge zur numerischen Linearen Algebra ("A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients", Trans. Amer. Inst. Elec. Eng. 60, 1235-1241, 1941) und zur numerischen Elektrodynamik.

und allgemein für $j = 1, \cdots, n$:

$$\begin{aligned} r_{jk} &:= a_{jk} - \sum_{i=1}^{j-1} l_{ji} r_{ik}\,, \quad k = j, j+1, \cdots, n\,, \\ l_{kj} &:= r_{jj}^{-1} \left(a_{kj} - \sum_{i=1}^{j-1} l_{ki} r_{ij} \right)\,, \quad k = j+1, j+2, \cdots, n\,. \end{aligned} \tag{4.2.28}$$

Die Gauß-Elimination und die direkte Dreieckszerlegung unterscheiden sich nur in der Reihenfolge der Operationen und sind algebraisch völlig äquivalent.

4.3 Spezielle Gleichungssysteme

4.3.1 Bandmatrizen

Die Anwendung des Gaußschen Eliminationsverfahrens zur Lösung "großer" Gleichungssysteme ($n > 1000$) ist mit großen technischen Schwierigkeiten verbunden, wenn der Kernspeicher des Rechners nicht zur Speicherung der ganzen Koeffizientenmatrix ausreicht. In diesem Fall müssen externe Speicher verwendet werden, was wegen des Datentransfers die Rechenzeit in die Höhe treibt. Viele der in der Praxis auftretenden großen Matrizen besitzen jedoch eine besondere Struktur, welche es erlaubt, bei der Durchführung des Gaußschen Verfahrens Speicherplatz zu sparen.

Definition 4.7: *Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt "Bandmatrix" vom Bandtyp (m_l, m_r) mit $0 \le m_l, m_r \le n-1$, wenn gilt:*

$$a_{jk} = 0 \quad \textit{für} \quad k < j - m_l \quad \textit{oder} \quad k > j + m_r \quad (j,k = 1,\ldots,n)\,.$$

Die Elemente von A sind also bis auf die Hauptdiagonale und höchstens $m_l + m_r$ Nebendiagonalen gleich Null. Die Größe $m = m_l + m_r + 1$ ist dann die "Bandbreite".

Beispiel 4.9: Wir geben einige einfache Beispiele von Bandmatrizen an:

Typ $(n-1, 0)$ untere Dreiecksmatrix

Typ $(0, n-1)$ obere Dreiecksmatrix

Typ $(1, 1)$ Tridiagonalmatrix

Beispiel einer (16×16)-Matrix vom Bandtyp $(4, 4)$:

$$A = \left[\begin{array}{cccc} B & -I & & \\ -I & B & -I & \\ & -I & B & -I \\ & & -I & B \end{array}\right] \Bigg\} 16$$

$$B = \left[\begin{array}{cccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{array}\right] \Bigg\} 4$$

$$I = \left[\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{array}\right] \Bigg\} 4$$

Satz 4.5 (Bandmatrix): *Ist $A \in \mathbb{R}^{n\times n}$ eine Bandmatrix vom Typ (m_l, m_r), für die das Gaußsche Eliminationsverfahren ohne Zeilenvertauschung durchführbar ist, dann sind auch alle reduzierten Matrizen Bandmatrizen desselben Typs, und die Faktoren L und R*

der Dreieckszerlegung von A sind Bandmatrizen vom Typ $(m_l, 0)$ bzw. $(0, m_r)$. Der Aufwand für die Berechnung der LR-Zerlegung einer Bandmatrix vom Typ (m_l, m_r) ist

$$N = \frac{1}{3} n m_l m_r + O(n(m_l + m_r)).$$

Beweis: Man erhält die Behauptung durch Nachrechnen (Übung). Q.E.D.

Zur Durchführung der Gaußschen Elimination genügt es also bei Bandmatrizen, das "Band" zu speichern. Bei Größenordnungen $n \sim 10.000$ und $m \sim 100$ macht dies die Anwendung des Verfahrens erst möglich. Bei der obigen Modellmatrix ergibt sich ein reduzierter Speicherplatzbedarf von $16 \times 9 = 144$ (oder weniger) anstatt der $16 \times 16 = 256$ für die volle Matrix. (Die Ausnutzung der Symmetrie wird später noch diskutiert.)

Eine extreme Ersparnis ergibt sich natürlich bei den besonders einfach strukturierten Tridiagonalmatrizen

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & c_n & a_n \end{bmatrix}.$$

Hier lassen sich die Elemente der LR-Zerlegung

$$L = \begin{bmatrix} 1 & & & \\ \gamma_2 & \ddots & & \\ & \ddots & 1 & \\ & & \gamma_n & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{n-1} & \beta_{n-1} \\ & & & \alpha_n \end{bmatrix}$$

durch einfache, rekursive Beziehungen bestimmen (Beweis durch Probe):

$$\begin{aligned} & & & & \alpha_1 &= a_1 & , \quad \beta_1 &= b_1 , \\ i = 2, \ldots, n-1 : \quad & \gamma_i &= c_i / \alpha_{i-1} & , \quad \alpha_i &= a_i - \gamma_i \beta_{i-1} & , \quad \beta_i &= b_i , \\ & \gamma_n &= c_n / \alpha_{n-1} & , \quad \alpha_n &= a_n - \gamma_n \beta_{n-1} & . & \end{aligned}$$

Hierzu sind offenbar nur $3n - 2$ Speicherplätze und $2n - 2$ a. Op. erforderlich.

Häufig sind Bandmatrizen auch noch "dünn besetzt", d.h.: Die meisten Elemente innerhalb des Bandes sind Null. Dieser Umstand kann beim Gaußschen Eliminationsverfahren jedoch nicht zur Speicherersparnis ausgenutzt werden, da i. Allg. das ganze Band im Verlaufe des Verfahrens mit Elementen ungleich Null aufgefüllt wird.

Wesentlich für Satz 4.5 war, daß das Gaußsche Verfahren ohne Zeilenvertauschungen durchgeführt werden kann, da andernfalls die Bandbreite anwächst. Wir betrachten im folgenden zwei Klassen von Matrizen, bei denen dies der Fall ist.

4.3.2 Diagonaldominante Matrizen

Definition 4.8: *Eine Matrix* $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ *heißt "diagonaldominant", wenn*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \leq |a_{jj}|\,, \quad j = 1,\ldots,n\,. \tag{4.3.29}$$

Satz 4.6 (Existenz der LR-Zerlegung): *Die Matrix* $A \in \mathbb{R}^{n\times n}$ *sei regulär und diagonaldominant. Dann existiert eine LR-Zerlegung* $A = LR$*, die mit Gaußscher Elimination* ohne *Pivotierung bestimmt werden kann.*

Beweis: Da A regulär und diagonaldominant ist, muß $a_{11} \neq 0$ sein. Folglich kann der erste Eliminationsschritt $A := A^{(0)} \to A^{(1)}$ *ohne* Pivotierung durchgeführt werden. Die Elemente $a_{jk}^{(1)}$ erhält man durch $a_{1k}^{(1)} = a_{1k},\ k = 1,\ldots,n$, und

$$j = 2,\ldots,n\,, \quad k = 1,\ldots,n\,: \qquad a_{jk}^{(1)} = a_{jk} - q_{j1}a_{1k}\,, \quad q_{j1} = \frac{a_{j1}}{a_{11}}\,.$$

Also gilt für $j = 2,\ldots,n$:

$$\begin{aligned}
\sum_{k=2,k\neq j}^{n} |a_{jk}^{(1)}| &\leq \sum_{k=2,k\neq j}^{n} |a_{jk}| + |q_{j1}| \sum_{k=2,k\neq j}^{n} |a_{1k}| \\
&\leq \underbrace{\sum_{k=1,k\neq j}^{n} |a_{jk}|}_{\leq |a_{jj}|} - |a_{j1}| + \underbrace{|q_{j1}|}_{= \left|\frac{a_{j1}}{a_{11}}\right|} \underbrace{\sum_{k=2}^{n} |a_{1k}|}_{\leq |a_{11}|} - |q_{j1}||a_{1j}| \\
&\leq |a_{jj}| - |q_{j1}a_{1j}| \leq |a_{jj} - q_{j1}a_{1j}| = |a_{jj}^{(1)}|.
\end{aligned}$$

Die Matrix $A^{(1)} = G_1 A^{(0)}$ ist regulär und offenbar wieder diagonaldominant, und folglich ist $a_{22}^{(1)} \neq 0$. Diese Eigenschaft bleibt also bei Durchführung der Gaußschen Elimination erhalten. Der ganze Prozeß ist somit ohne Zeilenvertauschungen durchführbar. Q.E.D.

Bemerkung 4.1: Gilt in (4.3.29) für alle $j \in \{1,\ldots,n\}$ die strikte Ungleichung, so heißt die Matrix A "strikt diagonaldominant". Der Beweis von Satz 4.6 zeigt, daß für eine solche die Gaußsche Elimination stets ohne Pivotierung durchführbar ist, d.h.: Die Matrix ist notwendig "regulär". Die obige Modellmatrix ist zwar diagonaldominant, aber nicht strikt diagonaldominant. Daß sie trotzdem regulär ist, wird sich später aufgrund eines schärferen Kriteriums ergeben.

4.3.3 Positiv definite Matrizen

Wir erinnern daran, daß eine (symmetrische) Matrix $A \in \mathbb{R}^{n\times n}$ mit der Eigenschaft

$$(Ax, x)_2 > 0\,, \quad x \in \mathbb{R}^n \setminus \{0\}$$

"positiv definit" genannt wird.

Satz 4.7 (Existenz der LR-Zerlegung): *Für positiv definite Matrizen $A \in R^{n\times n}$ ist das Gaußsche Eliminationsverfahren ohne Zeilenvertauschung durchführbar, und die dabei auftretenden Pivotelemente $a_{ii}^{(i)}$ sind positiv.*

Beweis: Da A symmetrisch und positiv definit ist, ist notwendig $a_{11} > 0$, und die Beziehung

$$a_{jk}^{(1)} = a_{jk} - \frac{a_{j1}}{a_{11}} a_{1k} = a_{kj} - \frac{a_{k1}}{a_{11}} a_{1j} = a_{kj}^{(1)}$$

für $j, k = 2, \ldots, n$ zeigt, daß die im ersten Eliminationsschritt erzeugte $(n-1)\times(n-1)$-Matrix $\tilde{A}^{(1)} = (a_{jk}^{(1)})_{j,k=2,\ldots,n}$ ebenfalls symmetrisch ist. Wir wollen zeigen, daß sie auch positiv definit ist, so daß wieder $a_{22}^{(1)} > 0$. Der Eliminationsprozeß kann dann mit positivem Pivotelement fortgesetzt werden, und die Behauptung folgt durch Induktion.

Sei $\tilde{x} = (x_2, \ldots, x_n)^T \in \mathbb{R}^{n-1} \setminus \{0\}$ und $x = (x_1, \tilde{x})^T \in \mathbb{R}^n$ mit

$$x_1 = -\frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k\,.$$

Dann ist

$$\begin{aligned} 0 < \sum_{j,k=1}^{n} a_{jk} x_j x_k &= \sum_{j,k=2}^{n} a_{jk} x_j x_k + 2x_1 \sum_{k=2}^{n} a_{1k} x_k + a_{11} x_1^2 \\ &\quad \underbrace{- \frac{1}{a_{11}} \sum_{j,k=2}^{n} a_{k1} a_{1j} x_k x_j + \frac{1}{a_{11}} \Big(\sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0 \ (a_{jk} = a_{kj})} \\ &= \sum_{j,k=2}^{n} \underbrace{\Big(a_{jk} - \frac{a_{k1} a_{1j}}{a_{11}} \Big)}_{= a_{jk}^{(1)}} x_j x_k + a_{11} \underbrace{\Big(x_1 + \frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0} \end{aligned}$$

und somit $\tilde{x}^T \tilde{A}^{(1)} \tilde{x} > 0$. Q.E.D.

Für positiv definite Matrizen existiert also stets eine LR-Zerlegung $A = LR$ mit positiven Pivotelementen

$$r_{ii} = a_{ii}^{(i)} > 0\,, \quad i = 1, \ldots, n\,.$$

Wegen $A = A^T$ gilt aber auch

$$A = A^T = (LR)^T = (LD\tilde{R})^T = \tilde{R}^T D L^T$$

mit den Matrizen

$$\tilde{R} = \begin{bmatrix} 1 & r_{12}/r_{11} & \cdots & r_{1n}/r_{11} \\ & \ddots & \ddots & \vdots \\ & & 1 & r_{n-1,n}/r_{n-1,n-1} \\ 0 & & & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix}.$$

Mit der Eindeutigkeit der LR-Zerlegung folgt aus

$$A = LR = \tilde{R}^T D L^T$$

notwendig $L = \tilde{R}^T$ bzw. $R = DL^T$. Damit haben wir den folgenden Satz bewiesen.

Satz 4.8: *Positiv definite Matrizen gestatten eine sog. "Cholesky[5]-Zerlegung".*

$$A = LDL^T = \tilde{L}\tilde{L}^T \tag{4.3.30}$$

mit der Matrix $\tilde{L} := LD^{1/2}$. *Bei der Berechnung der Cholesky-Zerlegung genügt es, die Matrizen* D *und* L *zu bestimmen. Dies reduziert die benötigten Operationen auf*

$$N_{\text{Cholesky}}(n) = n^3/6 + O(n^2).$$

Der sog. "Algorithmus von Cholesky" zur Berechnung der Zerlegungsmatrix

$$\tilde{L} = \begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix}$$

geht direkt von der Beziehung $A = \tilde{L}\tilde{L}^T$ aus, die man als ein System von $n(n+1)/2$ Gleichungen für die Größen $\tilde{l}_{jk}$, $k \le j$, auffassen kann. Ausmultiplizieren von

$$\begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix} \begin{bmatrix} \tilde{l}_{11} & \cdots & \tilde{l}_{n1} \\ & \ddots & \vdots \\ 0 & & \tilde{l}_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

ergibt in der ersten Spalte von $\tilde{L}$:

$$\tilde{l}_{11}^2 = a_{11}\,, \quad \tilde{l}_{21}\,\tilde{l}_{11} = a_{21}\,, \quad \ldots\,, \quad \tilde{l}_{n1}\,\tilde{l}_{11} = a_{n1}\,,$$

[5] *André Louis Cholesky (1975-1918): französischer Mathematiker; Militärkarriere; Beiträge zur Numerischen Linearen Algebra.*

woraus sich

$$\tilde{l}_{11} = \sqrt{a_{11}}\,, \quad j = 2, \ldots, n : \quad \tilde{l}_{j1} = \frac{a_{j1}}{\tilde{l}_{11}} = \frac{a_{j1}}{\sqrt{a_{11}}}\,, \tag{4.3.31}$$

berechnet. Seien nun für ein $i \in \{2, \cdots, n\}$ die Elemente $\tilde{l}_{jk}$, $k = 1, \ldots, i-1$, $j = k, \ldots, n$, schon bekannt. Dann erhält man aus

$$\begin{aligned} \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \ldots + \tilde{l}_{ii}^2 &= a_{ii}\,, \quad \tilde{l}_{ii} > 0\,, \\ \tilde{l}_{j1}\tilde{l}_{i1} + \tilde{l}_{j2}\tilde{l}_{i2} + \ldots + \tilde{l}_{ji}\tilde{l}_{ii} &= a_{ji}, \end{aligned}$$

die nächsten Elemente $\tilde{l}_{ii}$ und $\tilde{l}_{ji}$, $j = i+1, \ldots, n$, gemäß

$$\begin{aligned} \tilde{l}_{ii} &= \sqrt{a_{ii} - \tilde{l}_{i1}^2 - \tilde{l}_{i2}^2 - \ldots - \tilde{l}_{i,i-1}^2}\,, \\ \tilde{l}_{ji} &= \tilde{l}_{ii}^{-1}\big\{a_{ji} - \tilde{l}_{j1}\tilde{l}_{i1} - \tilde{l}_{j2}\tilde{l}_{i2} - \ldots - \tilde{l}_{j,i-1}\tilde{l}_{i,i-1}\big\}, \quad j = i+1, \ldots, n, \end{aligned}$$

4.4 Nicht reguläre Systeme

Mit einer (nicht notwendig quadratischen) Matrix $A \in \mathbb{R}^{m\times n}$ und einem Vektor $b \in \mathbb{R}^m$ sei das Gleichungssystem

$$Ax = b \tag{4.4.32}$$

gegeben. Es wird hier auch $\mathrm{Rang}(A) < \mathrm{Rang}[A,b]$ zugelassen, d.h.: Das System muß nicht unbedingt im eigentlichen Sinne lösbar sein. In diesem Fall wird ein geeigneter erweiterter Lösungsbegriff eingeführt. Wir betrachten im folgenden die auf Gauß zurückgehende sog. "Methode der kleinsten Fehlerquadrate". Dabei wird ein Vektor $\bar{x} \in \mathbb{R}^n$ gesucht, dessen Defekt $d \equiv b - A\bar{x}$ bzgl. der euklidischen Norm minimal ist. Dieser Lösungsbegriff fällt natürlich im Falle $\mathrm{Rang}(A) = \mathrm{Rang}[A,b]$ mit dem üblichen zusammen.

Satz 4.9 ("Least-Squares"-Lösung): *Es existiert stets eine "Lösung" $\bar{x} \in \mathbb{R}^n$ von (4.4.32) mit kleinsten Fehlerquadraten ("Least-Squares"-Lösung)*

$$\|A\bar{x} - b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|_2. \tag{4.4.33}$$

Dies ist äquivalent dazu, daß $\bar{x}$ Lösung der sog. "Normalgleichung" ist:

$$A^T A\bar{x} = A^T b. \tag{4.4.34}$$

Im Falle $\mathrm{Rang}(A) = n$ ist $\bar{x}$ eindeutig bestimmt, andernfalls ist jede weitere Lösung von der Form $\bar{x} + y$ mit $y \in \mathrm{Kern}(A)$.

Beweis: (i) Sei $\bar{x}$ Lösung der Normalgleichung. Für ein beliebiges $x \in \mathbb{R}^n$ gilt dann

$$\begin{aligned}\|b - Ax\|_2^2 &= \|b - A\bar{x} + A(\bar{x} - x)\|_2^2 \\ &= \|b - A\bar{x}\|_2^2 + 2\underbrace{(b - A\bar{x}}_{\in \mathrm{Kern}(A^T)}, \underbrace{A[\bar{x} - x])}_{\in \mathrm{Bild}(A)} + \|A(\bar{x} - x)\|_2^2 \geq \|b - A\bar{x}\|_2^2,\end{aligned}$$

d.h.: $\bar{x}$ ist Minimallösung. Für eine Minimallösung $\bar{x}$ gilt umgekehrt notwendig

$$\begin{aligned}0 = \frac{\partial}{\partial x_i}\|Ax - b\|_{2|x=\bar{x}}^2 &= \frac{\partial}{\partial x_i}\Big(\sum_{j=1}^n \Big|\sum_{k=1}^n a_{jk}x_k - b_j\Big|^2\Big)_{|x=\bar{x}} \\ &= 2\sum_{j=1}^n a_{ji}\Big(\sum_{k=1}^n a_{jk}\bar{x}_k - b_j\Big) = 2(A^T A\bar{x} - A^T b)_i,\end{aligned}$$

d.h.: $\bar{x}$ löst die Normalgleichung.

(ii) Wir untersuchen nun die Lösbarkeit der Normalgleichung. Das orthogonale Komplement von $\mathrm{Bild}(A)$ in $\mathbb{R}^m$ ist $\mathrm{Kern}(A^T)$. Also besitzt b eine eindeutige Zerlegung

$$b = s + r\,, \quad s \in \mathrm{Bild}(A)\,, \quad r \in \mathrm{Kern}(A^T).$$

Für ein $\bar{x} \in \mathbb{R}^n$ mit $A\bar{x} = s$ gilt dann

$$A^T A\bar{x} = A^T s \;=\; A^T s + A^T r = A^T b\,,$$

d.h.: $\bar{x}$ löst die Normalgleichung. Im Falle $\mathrm{Rang}(A) = n$ ist $\mathrm{Kern}(A) = \{0\}$ und $\mathrm{Bild}(A) = \mathbb{R}^n$. Aus $A^T Ax = 0$ folgt also wegen $\mathrm{Kern}(A^T) \perp \mathrm{Bild}(A)$ notwendig $Ax = 0$ bzw. $x = 0$. Die Matrix $A^T A \in \mathbb{R}^{n\times n}$ ist regulär und folglich $\bar{x}$ eindeutig bestimmt. Im Falle $\mathrm{Rang}(A) < n$ gilt für jede weitere Lösung x_1 der Normalgleichung

$$b = Ax_1 + (b - Ax_1) \;\in\; \mathrm{Bild}(A) + \mathrm{Kern}(A^T)\,.$$

Wegen der Eindeutigkeit dieser orthogonalen Zerlegung ist notwendig $Ax_1 = A\bar{x}$ bzw. $\bar{x} - x_1 \in \mathrm{Kern}(A)$. Q.E.D.

4.4.1 Gaußsche Ausgleichsrechnung

Im Anschluß an Satz 4.9 betrachten wir als klassische Anwendung der Methode der kleinsten Fehlerquadrate, die sog. "Gaußsche Ausgleichsrechnung" (kurz *Gauß-Ausgleich*). Die Aufgabenstellung ist dabei die folgende:

Zu gegebenen Funktionen $u_1, \ldots, u_n$ und Punkten $(x_j, y_j) \in \mathbb{R}^2$, $j = 1, \ldots, m$, $m > n$, ist eine Linearkombination

$$u(x) = \sum_{k=1}^{n} c_k u_k(x)$$

so zu bestimmen, daß die sog. "mittlere Abweichung"

$$\Delta_2 \;\equiv\; \Big(\sum_{j=1}^{m} |u(x_j) - y_j|^2 \Big)^{1/2}$$

möglichst klein wird. (Die sog. "Tschebyscheffsche Ausgleichsaufgabe", bei der die "maximale Abweichung"

$$\Delta_\infty \;\equiv\; \max_{j=1,\ldots,m} |u(x_j) - y_j|$$

minimiert wird, ist i. Allg. wesentlich schwieriger zu behandeln.)

Zur Lösung der Gaußschen Ausgleichsaufgabe setzen wir

$$\begin{aligned} y &\equiv (y_1, \ldots, y_m)^T\,, \qquad c \equiv (c_1, \ldots, c_n)^T\,, \\ a_k &\equiv (u_k(x_1), \ldots, u_k(x_m))^T\,,\ k = 1, \ldots, n\,, \qquad A \equiv [a_1, \ldots, a_n]. \end{aligned}$$

Zu minimieren ist also bzgl. $c \in \mathbb{R}^n$ das Funktional

$$F(c) = \|Ac - y\|_2.$$

Dies ist gleichbedeutend damit, für das (überbestimmte) Gleichungssystem $Ac = y$ eine "Lösung" mit kleinsten Fehlerquadraten zu ermitteln. Im Falle $\mathrm{Rang}(A) = n$ ist die

eindeutige "Lösung" c dann bestimmt als Lösung der Normalgleichung

$$A^T Ac = A^T y.$$

Ist speziell $u_k(x) = x^{k-1}$, so nennt man die "optimale" Lösung

$$u(x) = \sum_{k=1}^{n} c_k x^{k-1}$$

die "Gaußsche Ausgleichsparabel" zu den Punkten (x_j, y_j), $j = 1, \ldots, m$. Wegen der Regularität der sog. "Vandermondschen[6] Determinante"

$$\det \begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{bmatrix} = \prod_{j,k=1, j<k}^{n} (x_k - x_j) \neq 0$$

für paarweise verschiedene Stützstellen x_j ist dann stets $\mathrm{Rang}(A) = n$, d.h.: Die Ausgleichsparabel ist eindeutig bestimmt.

Beispiel 4.10: Zu den Meßdaten

x_i	-2	-1	0	1	2
y_i	0.5	0.5	2	3.5	3.5

soll mit Hilfe der Gaußschen Ausgleichsrechnung eine lineare Funktion $y(x) = a + bx$ angepaßt werden. Dies ist äquivalent zur Lösung des überbestimmten Gleichungssystems

$$\begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 2 \\ 3.5 \\ 3.5 \end{bmatrix}.$$

[6] Alexandre-Thophile Vandermonde (1735-1796): Französischer Mathematiker; begabter Musiker, kam spät zur Mathematik und publizierte hierzu nur vier Arbeiten (trotzdem Mitglied der Akademie der Wissenschaften in Paris); Beiträge zur Determinantentheorie und kombinatorischer Probleme (kurioserweise taucht die nach ihm benannte "Determinante" in keiner dieser Arbeiten explizit auf).

Die zugehörige Normalgleichung lautet

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \\ 2.0 \\ 3.5 \\ 3.5 \end{bmatrix}$$

$$\rightarrow \quad \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 10 \\ 9 \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 0.9 \end{bmatrix}.$$

Es ergibt sich die Lösung $y(x) = 2 + 0.9x$ mit der mittleren Abweichung:

$$\Delta_2 = \Big(\sum_{i=1}^{5} |y(x_i) - y_i|^2 \Big)^{1/2} = \sqrt{0.9} < 1,$$

und der maximalen Abweichung:

$$\Delta_\infty = \max_{1 \le i \le 5} |y(x_i) - y_i| = 0.6.$$

Durch geometrische Anschauung erhält man in diesem Fall auch die Lösung des Tschebyscheffschen Ausgleichproblems:

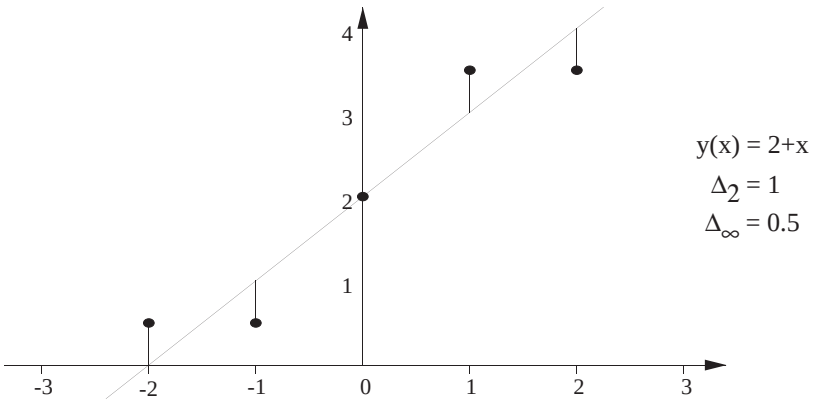


Abbildung 4.1: Lösung der Tschebyscheffschen Ausgleichsaufgabe

Bemerkung 4.2: Wesentlich für die Anwendbarkeit der Gaußschen Ausgleichsrechnung ist, daß für die zu bestimmenden Größen eine "lineare" Beziehung gegeben ist, z. B. $y(x) = a + bx$. Ist die gegebene Beziehung (etwa aus physikalischen Gründen) nichtlinear, so kann man versuchen, aus ihr eine lineare Beziehung für unter Umständen andere Größen

zu gewinnen, aus denen sich dann nachträglich die eigentlich gesuchten Größen bestimmen lassen; z. B.:

$$y(x) = \frac{a}{1+bx}\,.$$

$$\text{Umformung:}\quad \frac{1}{a} + \frac{b}{a}x = \frac{1}{y(x)}\,, \qquad \text{neue Größen:}\quad \tilde{a} = \frac{1}{a}\,,\quad \tilde{b} = \frac{b}{a}\,.$$

Zur Berechnung der Lösung mit kleinsten (Fehler-)Quadraten eines irregulären Systems $Ax = b$ muß die Normalgleichung $A^T Ax = A^T b$ gelöst werden. Dessen Matrix besitzt einige Besonderheiten, die in folgendem Lemma zusammengefaßt sind.

Lemma 4.4: *Für eine Matrix $A \in \mathbb{K}^{m\times n}$ mit $m \geq n$ ist die Matrix $\bar{A}^T A \in \mathbb{K}^{n\times n}$ stets hermitesch (symmetrisch) und positiv semi-definit. Im Fall* $\mathrm{Rang}(A) = n$ *ist $\overline{A}^T A$ sogar positiv definit.*

Beweis: Nach den Regeln der Matrizenrechnung gilt:

$$(\bar{A}^T A)^T = A^T \bar{A} = \overline{\bar{A}^T A}, \qquad \bar{x}^T (\bar{A}^T A)x = \overline{(Ax)}^T Ax = \|Ax\|_2^2 \geq 0,$$

d.h.: $\bar{A}^T A$ ist hermitesch und positiv semi-definit. Im Fall $\mathrm{Rang}(A) = n$ ist die Matrix als Abbildung $A : \mathbb{R}^n \to \mathbb{R}^m$ $(n \leq m)$ injektiv, d.h.: $\|Ax\|_2 = 0$ impliziert $x = 0$. Die Matrix $\bar{A}^T A$ ist also positiv definit. Q.E.D.

Die Lösung des Normalgleichungssystems kann wegen der Symmetrie der Matrix $A^T A$ prinzipiell mit dem Cholesky-Algorithmus erfolgen. Im Allg. ist sie aber sehr schlecht konditioniert; im Fall $m = n$ ist

$$\mathrm{cond}(A^T A) \sim \mathrm{cond}(A)^2\,. \tag{4.4.35}$$

Beispiel 4.11: Bei 3-stelliger Rechnung erhält man

$$A = \begin{bmatrix} 1.07 & 1.10 \\ 1.07 & 1.11 \\ 1.07 & 1.15 \end{bmatrix} \quad \to \quad A^T A = \begin{bmatrix} 3.43 & 3.60 \\ 3.60 & 3.76 \end{bmatrix}.$$

Aber $A^T A$ ist <u>nicht</u> positiv definit: $(-1,1)\cdot A^T A \cdot (-1,1)^T = -0.01$, d.h. Das Cholesky-Verfahren wird i. Allg. keine Lösung liefern!

Wir werden nun eine Methode betrachten, die es gestattet, $\bar{x}$ ohne explizite Aufstellung der Normalgleichung zu berechnen. Für spätere Zwecke wird dabei der Fall komplexer Matrizen zugelassen.

Satz 4.10 (QR-Zerlegung): *Sei* $A \in \mathbb{K}^{m\times n}$ *eine rechteckige Matrix mit* $m \geq n$ *und* $\mathrm{Rang}(A) = n$. *Dann existiert eine eindeutig bestimmte Matrix* $Q \in \mathbb{K}^{m\times n}$ *mit der Eigenschaft*

$$\bar{Q}^T Q = I \quad (\mathbb{K} = \mathbb{C}), \qquad Q^T Q = I \quad (\mathbb{K} = \mathbb{R}), \tag{4.4.36}$$

und eine eindeutig bestimmte obere Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ *mit reellen Diagonalelementen* $r_{ii} > 0$, $i = 1, \ldots, n$, *so daß*

$$A = QR. \tag{4.4.37}$$

Wegen $\bar{Q}^T Q = I$ *sind offenbar die Spalten von* Q *paarweise orthonormal;* Q *wird daher "orthogonale" (genauer "orthonormale") Matrix genannt (im Falle* $m = n$ *"unitäre" Matrix).*

Beweis: Die Matrix Q wird durch sukzessive Orthonormalisierung der Spaltenvektoren a_k, $k = 1, \ldots, n$, von A erzeugt. Nach dem Gram-Schmidt-Verfahren setzt man

$$q_1 \equiv \|a_1\|_2^{-1} a_1$$

$$k = 2, \ldots, n: \qquad \tilde{q}_k \equiv a_k - \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i, \quad q_k \equiv \|\tilde{q}_k\|_2^{-1} \, \tilde{q}_k.$$

Wegen $\mathrm{Rang}(A) = n$ sind die n Spaltenvektoren $\{a_1, \ldots, a_n\}$ linear unabhängig, und der Orthonormalisierungsprozeß kann folglich nicht vorzeitig abbrechen.

Die Matrix $Q \equiv [q_1, \ldots, q_n]$ ist konstruktionsgemäß orthonormal. Ferner gilt für $k = 1, \ldots, n$:

$$a_k = \tilde{q}_k + \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i = \|\tilde{q}_k\|_2 \, q_k + \sum_{i=1}^{k-1} (a_k, q_i)_2 \, q_i$$

bzw.

$$a_k = \sum_{i=1}^{k} r_{ik} \, q_k, \qquad r_{kk} \equiv \|\tilde{q}_k\|_2 \in \mathbb{R}_+, \quad r_{ik} \equiv (a_k, q_i)_2 .$$

Setzt man noch $r_{ik} \equiv 0$ für $i > k$, so ist dies äquivalent zur Gleichung

$$A = QR$$

mit der oberen Dreiecksmatrix $R = (r_{ik}) \in \mathbb{K}^{n\times n}$.

Zum Beweis der Eindeutigkeit der QR-Zerlegung seien $A = Q_1 R_1$ und $A = Q_2 R_2$ zwei solche Zerlegungen. Da R_1 und R_2 regulär sind $(\det(R_i) > 0)$, gilt:

$$\begin{aligned} Q &:= \bar{Q}_2^T Q_1 = R_2 R_1^{-1} && \text{rechte obere Dreiecksmatrix}, \\ \bar{Q}^T &= \bar{Q}_1^T Q_2 = R_1 R_2^{-1} && \text{rechte obere Dreiecksmatrix}. \end{aligned}$$

Wegen $\bar{Q}^T Q = R_1 R_2^{-1} R_2 R_1^{-1} = I$ ist Q *orthonormal* und *diagonal* mit $|\lambda_i| = 1$. Aus $QR_1 = R_2$ folgt $\lambda_i r_{ii}^1 = r_{ii}^2 > 0$ und damit $\lambda_i \in \mathbb{R}$ sowie $\lambda_i = 1$. Also ist $Q = I$, d.h.

$$R_1 = R_2\,, \qquad Q_1 = AR_1^{-1} = AR_2^{-1} = Q_2$$

Dies vervollständigt den Beweis. Q.E.D.

Im Fall $\mathbb{K} = \mathbb{R}$ geht die Normalgleichung $A^T Ax = A^T b$ bei Verwendung der QR-Zerlegung über in

$$A^T Ax = R^T Q^T QRx = R^T Rx = R^T Q^T b,$$

bzw. wegen der Regularität von R^T,

$$Rx = Q^T b. \tag{4.4.38}$$

Dieses System ist nun durch Rückwärtseinsetzen lösbar. Wegen

$$A^T A = R^T R \tag{4.4.39}$$

ist mit R also die Cholesky-Zerlegung von $A^T A$ bestimmt, ohne $A^T A$ explizit berechnen zu müssen. Bei einer quadratischen Matrix $A \in \mathbb{R}^{n\times n}$ erfordert die Berechnung der QR-Zerlegung etwa den **doppelten** Aufwand wie zur Berechnung der LR-Zerlegung mit dem Gaußschen Algorithmus.

4.4.2 Householder-Verfahren

Das in Satz 4.9 verwendete Gram-Schmidtsche Orthogonalisierungsverfahren zum Nachweis der Existenz der QR-Zerlegung ist ungeeignet zur praktischen Berechnung von Q und R, da aufgrund von Rundungsfehlern die Orthonormalität der Spalten von Q rasch verloren geht. Das Gram-Schmidtsche Orthogonalisierungsverfahren ist kein numerisch gutartiger Algorithmus. Eine stabilere Methode zur Erstellung der Zerlegung $A = QR$ ist das sog. "Householder[7]-Verfahren", welches wir nun beschreiben werden. Zur Verwendung an späterer Stelle lassen wir dabei wieder komplexe Matrizen zu. Für einen Vektor $v \in \mathbb{K}^m$ nennt man

$$v\bar{v}^T := \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} [\bar{v}_1, \ldots, \bar{v}_m] = \begin{bmatrix} |v_1|^2 & v_1\bar{v}_2 & \cdots & v_1\bar{v}_m \\ \vdots & & & \\ v_m\bar{v}_1 & v_m\bar{v}_2 & \cdots & |v_m|^2 \end{bmatrix} \in \mathbb{K}^{m\times m}$$

sein "dyadisches Produkt" (nicht zu verwechseln mit dem "skalaren Produkt" $\bar{v}^T v = \|v\|_2^2$).

[7]Alston Scott Householder (1904-1993): US-Amerikanischer Mathematiker; Direktor des Oak Ridge Natioanl Laboratory (1948-1969), danach Professor an der University of Tennessee; Arbeiten zur mathematischen Biologie, aber am besten bekannt durch fundamentalen Beiträge zur Numerik, insbesondere zur numerischen linearen Algebra.

Definition 4.9: *Für einen Vektor* $v \in \mathbb{K}^n$ *mit* $\|v\|_2 = 1$, *heißt die Matrix*

$$S = I - 2v\bar{v}^T \in \mathbb{K}^{m\times m}$$

"Householder-Transformation". Offenbar ist $S = \bar{S}^T = S^{-1}$ *d.h.:* S *(und auch* $\bar{S}^T$*) ist hermitesch und unitär. Ferner ist das Produkt von zwei Householder-Transformationen wieder unitär.*

Zur geometrischen Interpretation der Householder-Transformation S beschränken wir uns auf den $\mathbb{R}^2$ und betrachten dort für irgendeinen normierten Vektor v, $\|v\|_2 = 1$, die Basis $\{v, v^\perp\}$, wobei $v^T v^\perp = 0$. Für einen Vektor $u = \alpha v + \beta v^\perp \in \mathbb{R}^2$ ist dann

$$\begin{aligned} Su &= (I - 2vv^T)(\alpha v + \beta v^\perp) \\ &= \alpha v + \beta v^\perp - 2\alpha (v \underbrace{v^T) v}_{=1} - 2\beta (v \underbrace{v^T) v^\perp}_{=0} = -\alpha v + \beta v^\perp . \end{aligned}$$

Die Anwendung von $S = I - 2vv^T$ auf einen Vektor u bewirkt also in der Ebene $\mathrm{Span}\{v, u\}$ eine Spiegelung von u an der orthogonalen Achse $\mathrm{Span}\{v^\perp\}$.

Ausgehend von einer Matrix $A \in \mathbb{K}^{m\times n}$ erzeugt nun das Householder-Verfahren in n Schritten eine Folge von Matrizen

$$A := A^{(0)} \to \cdots \to A^{(i-1)} \to \cdots \to A^{(n)} := \tilde{R},$$

wobei $A^{(i-1)}$ die Gestalt hat:

$$A^{(i-1)} = \left[\begin{array}{ccc|ccc} * & & & & & * \\ & * & & & & \\ & & \ddots & & & \vdots \\ & & & * & & \\ \hline & & & * & \cdots & * \\ & & 0 & & & \\ & & & * & \cdots & * \end{array}\right] \begin{matrix} \\ \\ \\ \\ i \\ \\ \\ \end{matrix} \quad i$$

Im i-ten Schritt wird eine Householder-Transformation $S_i \in \mathbb{K}^{m\times m}$ so bestimmt, daß

$$S_i A^{(i-1)} = A^{(i)}.$$

Nach n Schritten ist dann

$$\tilde{R} = A^{(n)} = S_n S_{n-1} \cdots S_1 A =: \bar{\tilde{Q}}^T A,$$

wobei $\bar{\tilde{Q}} \in \mathbb{K}^{m\times m}$ als Produkt unitärer Matrizen selbst unitär ist, und $\tilde{R} \in \mathbb{K}^{m\times n}$ die Gestalt hat:

$$\tilde{R} = \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline 0 & \cdots & 0 \end{array}\right] \begin{matrix} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} \text{n} \\ \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} \text{m.} \end{matrix}$$

Dies ergibt die Darstellung

$$A = \bar{S}_1^T \cdots \bar{S}_n^T \tilde{R} = \tilde{Q}\tilde{R}.$$

Hieraus erhält man die gewünschte QR-Zerlegung von A einfach durch Streichen der letzten $m-n$ Spalten in $\tilde{Q}$ sowie der letzten $m-n$ Zeilen in $\tilde{R}$:

$$A = \underbrace{\left[\begin{array}{c|c} Q & * \end{array}\right]}_{\text{n}\quad\text{m - n}} \cdot \left[\begin{array}{c} R \\ \hline 0 \end{array}\right] \begin{matrix} \left.\vphantom{\begin{array}{c}a\\a\end{array}}\right\} \text{n} \\ \left.\vphantom{\begin{array}{c}a\\a\end{array}}\right\} \text{m} - \text{n} \end{matrix} = QR$$

Man beachte, daß hier die Diagonalelemente von R nicht notwendig positiv sein müssen, d.h.: Der Householder-Algorithmus liefert in der Regel nicht die durch Satz 4.10 gegebene "eindeutig bestimmte" QR-Zerlegung.

Wir beschreiben nun den Transformationsprozeß im Detail. Seien a_k die Spaltenvektoren der Matrix A.

1. Schritt: S_1 wird so gewählt, daß $S_1 a_1 \in \text{Span}\{e_1\}$.

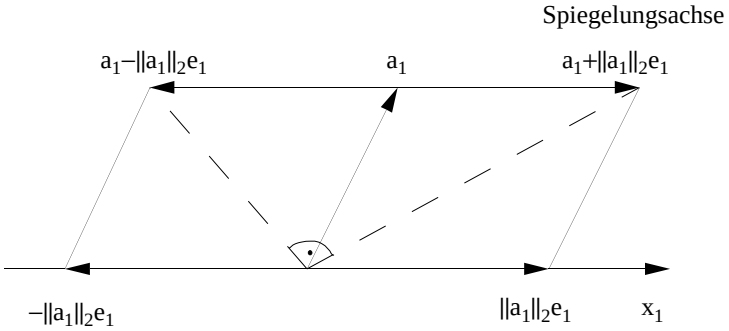


Abbildung 4.2: Schema der Householder-Transformation

Im Folgenden werden euklidische Norm und Skalarprodukt zur Abkürzung mit $\|\cdot\| = \|\cdot\|_2$ und $(\cdot,\cdot) = (\cdot,\cdot)_2$ bezeichnet. Der Vektor a_1 wird an der Achse $\text{Span}\{a_1 + \|a_1\| e_1\}$ (oder

$\mathrm{Span}\{a_1 - \|a_1\|e_1\}$) in die x_1-Achse gespiegelt. (Zur Vermeidung von Rundungsfehlern durch Auslöschung wählt man gewöhnlich das Vorzeichen entsprechend $\mathrm{sign}(a_{11})$!) Im Falle $a_{11} \geq 0$ ist also

$$v_1 = \frac{a_1 + \|a_1\|e_1}{\|a_1 + \|a_1\|e_1\|}\,, \qquad v_1^{\perp} = \frac{a_1 - \|a_1\|e_1}{\|a_1 - \|a_1\|e_1\|}.$$

Die Matrix $A^{(1)} = (I - 2v_1\bar{v}_1^T)A$ hat dann die Spaltenvektoren

$$a_1^{(1)} = -\|a_1\|e_1\,, \quad a_k^{(1)} = a_k - 2(a_k, v_1)v_1, \quad k = 2, \ldots, n\,.$$

Sei nun die transformierte Matrix $A^{(i-1)}$ schon berechnet.

<u>i-ter Schritt</u>: Für S_i machen wir den folgenden Ansatz:

$$S_i = \underbrace{\left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2\tilde{v}_i\bar{\tilde{v}}_i^T \end{array}\right]}_{\text{i - 1}} = I - 2v_i\bar{v}_i^T, \qquad v_i = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{v}_i \end{array}\right] \begin{array}{l} \Big\} \text{i - 1} \\ \Big\} \text{ m} \end{array}$$

Die Anwendung der (unitären) Matrix S_i von links auf $A^{(i-1)}$ läßt die ersten $i-1$ Zeilen- und Spalten von $A^{(i-1)}$ unverändert. Zur Konstruktion von v_i wenden wir die Überlegung vom 1. Schritt auf die Teilmatrix:

$$\tilde{A}^{(i-1)} = \left[\begin{array}{ccc} \tilde{a}_{ii}^{(i-1)} & \cdots & \tilde{a}_{in}^{(i-1)} \\ \vdots & & \vdots \\ \tilde{a}_{mi}^{(i-1)} & \cdots & \tilde{a}_{mn}^{(i-1)} \end{array}\right] = \left[\tilde{a}_i^{(i-1)}, \ldots, \tilde{a}_n^{(i-1)}\right]$$

an. Es ist demnach

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i-1)} - \|\tilde{a}_i^{(i-1)}\|\tilde{e}_i}{\|\quad\cdots\quad\|}\,, \quad \tilde{v}_i^{\perp} = \frac{\tilde{a}_i^{(i-1)} + \|\tilde{a}_i^{(i-1)}\|\tilde{e}_i}{\|\quad\cdots\quad\|}\,,$$

und die Matrix $A^{(i)}$ hat die Spaltenvektoren

$$\begin{aligned} a_k^{(i)} &= a_k^{(i-1)}\,, \quad k = 1, \ldots, i-1\,, \\ a_i^{(i)} &= (a_{1i}^{(i-1)}\,,\ \ldots\ ,\ a_{i-1,i}^{(i-1)},\ \|\tilde{a}_i^{(i-1)}\|\,, 0, \ldots, 0)^T\,, \\ a_k^{(i)} &= a_k^{(i-1)} - 2(\tilde{a}_k^{(i-1)}, \tilde{v}_i)v_i, \quad k = i+1, \ldots, n. \end{aligned}$$

4.5 Die Singulärwertzerlegung

Die in den vorhergehenden Abschnitten vorgestellten Methoden zur Lösung linearer Gleichungssysteme oder Ausgleichsprobleme (Methode der kleinsten Quadrate) werden numerisch unzuverlässig, wenn die Matrizen sehr schlecht konditioniert sind. Es kann sein, daß eine eigentlich reguläre Matrix für die numerische Rechnung singulär erscheint. Die Bestimmung des Ranges einer Matrix ist mit der LR- oder QR-Zerlegung oft nicht mit genügender Sicherheit zu entscheiden. Die derzeit zuverlässigste Technik zur Behandlung nahezu rang-defizienter linearer Gleichungs- und Ausgleichsprobleme verwendet die sog. "Singulärwertzerlegung" ("singular value decomposition", "SVD") einer Matrix. Dabei handelt es sich um einer spezielle "orthogonale" Zerlegung, welche die Matrix von beiden Seiten transformiert.

Es sei $A \in \mathbb{R}^{m\times n}$ gegeben. Weiter seien $Q \in \mathbb{R}^{m\times m}$ und $Z \in \mathbb{R}^{n\times n}$ orthogonal. Dann gilt zunächst

$$\|QAZ\|_2 = \|A\|_2, \tag{4.5.40}$$

so daß auch solche beidseitigen Transformationen die Kondition der Matrix A nicht verschlechtern. Für geeignete Matrizen Q und Z erhält man nun präzise Informationen über den Rang einer Matrix. Außerdem läßt sich das Ausgleichsproblem auch im Fall reduzierten Ranges befriedigend lösen. Allerdings ist die numerisch stabile Berechnung solcher Transformationen recht aufwendig, wie aus der Tabelle am Ende dieses Abschnittes hervorgeht.

Satz 4.11 (Singulärwertzerlegung): *Es sei $A \in \mathbb{R}^{m\times n}$. Dann existieren orthogonale Matrizen $V \in \mathbb{R}^{n\times n}$ und $U \in \mathbb{R}^{m\times m}$, so daß*

$$U^T A V = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min(m, n), \tag{4.5.41}$$

wobei $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.

Je nachdem, ob $m \leq n$ oder $m \geq n$ ist, erhält Σ die Gestalt

$$\left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right) \qquad \text{oder} \qquad \left(\begin{array}{c} \sigma_1 \\ & \ddots \\ & & \sigma_n \\ \hline & 0 & \end{array}\right).$$

Man nennt die Werte σ_i die "singulären Werte" der Matrix A. Aus (4.5.41) liest man unmittelbar ab, daß mit den Spaltenvektoren u_i, v_i von U, V gilt:

$$Av_i = \sigma_i u_i, \quad A^T u_i = \sigma_i v_i,$$

für $i = 1, \ldots, \min(m,n)$. Daraus ergibt sich

$$A^T A v_i = \sigma_i^2 v_i, \quad AA^T u_i = \sigma_i^2 u_i.$$

Die singulären Werte σ_i, $i = \ldots, \min(m,n)$ sind also gerade die Wurzeln der Eigenwerte von $A^T A$ bzw. AA^T.

Die Existenz einer Zerlegung der Form (4.5.41) läßt sich mit der letzten Überlegung unmittelbar darauf zurückführen, daß $A^T A$ sich durch orthogonale Matrizen auf Diagonalgestalt transformieren läßt,

$$Q^T (A^T A) Q = D.$$

Wir geben hier einen anderen Beweis.

Beweis: Es sei $\sigma = \|A\|_2$. Wegen $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$ existiert ein $x \in \mathbb{R}^n$ mit

$$Ax = \sigma y, \qquad \|x\|_2 = \|y\|_2 = 1.$$

Wir ergänzen x und y zu Orthonormalbasen des $\mathbb{R}^n$ und $\mathbb{R}^m$:

$$U = [y, \tilde{y}], \qquad V = [x, \tilde{x}].$$

Damit ergibt sich

$$A_1 \equiv U^T A V = \begin{pmatrix} \sigma & w^T \\ 0 & B \end{pmatrix}$$

mit einem Vektor $w \in \mathbb{R}^{n-1}$ und einer Matrix B. Da U und V orthogonal sind, folgt aus (4.5.40)

$$\|A_1\|_2 = \|A\|_2 = \sigma.$$

Andererseits gilt

$$\|A_1 (\sigma, w)^T\|_2^2 = \|(\sigma^2 + \|w\|_2^2, Bw)^T\|_2^2 \geq (\sigma^2 + \|w\|_2^2)^2 = (\sigma^2 + \|w\|_2^2) \|(\sigma, w)^T\|_2^2$$

und somit $w \equiv 0$. Der Rest folgt mit vollständiger Induktion. Q.E.D.

Wir stellen nun einige einfache Folgerungen aus (4.5.41) zusammen. Die singulären Werte seien geordnet in der Form $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, $p = \min(m,n)$.

- $\mathrm{Rang}(A) = r$,
- $\mathrm{Kern}(A) = \mathrm{Span}\{v_{r+1}, \ldots, v_n\}$,
- $\mathrm{Bild}(A) = \mathrm{Span}\{u_1, \ldots, u_r\}$,
- $A = U_r \Sigma_r V_r^T \equiv \sum_{i=1}^r \sigma_i u_i v_i^T$ (Singulärwertzerlegung von A),
- $\|A\|_2 = \sigma_1 = \sigma_{\max}$,
- $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$.

Wir betrachten nun das Problem der Bestimmung des "numerischen Rangs" einer Matrix. Wir definieren

$$\mathrm{Rang}(A,\varepsilon) = \min_{\|A-B\|_2 \le \varepsilon} \mathrm{Rang}(B)\,.$$

Man bezeichnet eine Matrix als "numerisch rang-defizient", falls

$$\mathrm{Rang}(A,\varepsilon) < \min(m,n)\,, \qquad \varepsilon = \mathrm{eps}\|A\|_2\,.$$

Stammen die Einträge der Matrix z.B. aus Meßreihen, so ist statt dessen ε an die Genauigkeit der Meßergebnisse zu knüpfen.

Satz 4.12 (Fehlerabschätzung): *Es seien* $A\,,\, U\,,\, V\,,\, \Sigma$ *wie in Satz* 4.11. *Falls* $k < r = \mathrm{Rang(A)}$*, so gilt mit der abgeschnittenen Singulärwertzerlegung*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

die Abschätzung

$$\min_{\mathrm{Rang}(B)=k} \|A-B\|_2 = \|A-A_k\|_2 = \sigma_{k+1}\,.$$

Als Konsequenz ergibt sich für $r_\varepsilon = \mathrm{Rang}(A,\varepsilon)$ *die Beziehung*

$$\sigma_1 \ge \cdots \ge \sigma_{r_\varepsilon} > \varepsilon \ge \sigma_{r_\varepsilon+1} \ge \cdots \ge \sigma_p\,, \quad p = \min(m,n)\,.$$

Beweis: Wegen

$$U^T A_k V = \mathrm{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)$$

folgt $\mathrm{Rang}(A_k) = k$. Weiter erhält man

$$U^T(A-A_k)V = \mathrm{diag}(0,\ldots,0,\sigma_{k+1},\ldots,\sigma_p)$$

und wegen der Orthogonalität von U und V somit

$$\|A-A_k\|_2 = \sigma_{k+1}.$$

Es bleibt zu zeigen, daß für jede andere Matrix B mit Rang k die Ungleichung

$$\|A-B\|_2 \ge \sigma_{k+1}$$

gilt. Dazu wählt man eine Orthonormalbasis $\{x_1,\ldots,x_{n-k}\}$ von $\mathrm{Kern}(B)$. Aus Dimensionsgründen gilt offensichtlich

$$\mathrm{Span}\{x_1,\ldots,x_{n-k}\} \cap \mathrm{Span}\{v_1,\ldots,v_{k+1}\} \neq \emptyset.$$

Sei z mit $\|z\|_2 = 1$ aus dieser Menge. Es gilt dann

$$Bz = 0\,, \qquad Az = \sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i$$

und somit

$$\|A - B\|_2^2 \geq \|(A-B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2\,.$$

Hier wurde ausgenutzt, daß $z = \sum_{i=1}^{k+1} (v_i^T z) v_i$ und deshalb

$$1 = \|z\|_2^2 = \sum_{i=1}^{k+1} (v_i^T z)^2\,.$$

Q.E.D.

Mit Hilfe der Singulärwertzerlegung läßt sich auch das Ausgleichsproblem elegant lösen. Es sei im folgenden wieder $m \geq n$. Wir haben bereits gesehen, daß jede Minimallösung,

$$\|Ax - b\|_2 = \min!$$

notwendig der Normalgleichung

$$A^T A x = A^T b$$

genügt. Die Lösung ist jedoch nur im (numerisch nicht unbedingt eindeutig feststellbaren) Fall, daß $\mathrm{Rang}(A) = n$ maximal ist, eindeutig bestimmt. In diesem Fall ist $A^T A$ invertierbar und es gilt

$$x = (A^T A)^{-1} A^T b.$$

Im Fall $\mathrm{Rang}(A) < n$ besitzen die Normalgleichungen unendlich viele Lösungen. Eindeutigkeit erzielt man durch die Zusatzforderung, daß diejenige Lösung gesucht wird, die minimale euklidische Norm besitzt. Sie heißt die "Minimallösung" des Ausgleichsproblems.

Satz 4.13 (Minimallösung): *Es sei $A = U\Sigma V^T$ die Singulärwertzerlegung der Matrix $A \in \mathbb{R}^{m\times n}$ und es sei $r = \mathrm{Rang}(A)$. Dann ist*

$$\bar{x} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i$$

die eindeutig bestimmte Lösung der Normalgleichung mit minimaler euklidischer Norm. Der Fehler genügt der Beziehung

$$\rho^2 = \|A\bar{x} - b\|_2^2 = \sum_{i=r+1}^{m} (u_i^T b)^2.$$

Beweis: Für jedes $x \in \mathbb{R}^n$ gilt die Identität

$$\|Ax - b\|_2^2 = \|AVV^T x - b\|_2^2 = \|U^T AVV^T x - U^T b\|_2^2 = \|\Sigma V^T x - U^T b\|_2^2.$$

Mit der Abkürzung $z = V^T x$ liefert dies

$$\|Ax - b\|_2^2 = \|\Sigma z - U^T b\|_2^2 = \sum_{i=1}^{r} (\sigma_i z_i - u_i^T b)^2 + \sum_{i=r+1}^{m} (u_i^T b)^2.$$

Ein Minimum erfüllt also notwendig

$$\sigma_i z_i = u_i^T b\,, \qquad i = 1, \ldots, r.$$

Unter allen z mit dieser Eigenschaft hat dasjenige mit $z_i = 0\,,\ i = r+1, \ldots, m\,,$ die minimale euklidische Norm. Die Identität für den Fehler ist offensichtlich. Q.E.D.

Die eindeutig bestimmte Minimallösung des Ausgleichsproblems läßt sich kompakt wie folgt darstellen: Es sei

$$\Sigma^+ = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{R}^{n \times m}.$$

Wir nennen die Matrix

$$A^+ = V \Sigma^+ U^T \tag{4.5.42}$$

die "Pseudo-Inverse" der Matrix A (oder auch die die "Penrose[8]-Inverse" (1955)). Der letzte Satz besagt

$$\bar{x} = A^+ b\,, \qquad \rho = \|(I - AA^+)b\|_2. \tag{4.5.43}$$

Die Pseudo-Inverse ist die eindeutige Lösung von

$$\min_{X \in \mathbb{R}^{n \times m}} \|AX - I\|_F,$$

mit der Frobenius-Norm $\|\cdot\|_F$. Da die Identität in (4.5.43) für alle b gilt, folgt

$$\begin{aligned} \mathrm{Rang}(A) = n \quad &\Rightarrow \quad A^+ = (A^T A)^{-1} A^T \\ \mathrm{Rang}(A) = n = m \quad &\Rightarrow \quad A^+ = A^{-1}\,. \end{aligned}$$

In der numerischen Praxis ist bei der Definition der Pseudoinversen natürlich der (geeignet definierte) numerische Rang zu benutzen. Die numerisch stabile Berechnung der Singulärwertzerlegung ist recht aufwendig. Auf Einzelheiten kann hier nicht eingegangen werden; es sei auf das Buch von Golub/van Loan: Matrix Computations, verwiesen.

[8]Roger Penrose (1931-): Englischer Mathematiker; Professor am Birkbeck College in London (1964) und seit 1973 Professor an der Universität Oxford; fundamentale Beiträge in der Mathematik zur Theorie von Halbgruppen, zur Matrix-Analysis und zur Theorie von "Kachelungen" sowie in der Theoretischen Physik zur Kosmologie, Relativitäts- und Quantentheorie.

4.6 Übungsaufgaben

Übung 4.1: Man zeige, daß für jede Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$ durch

$$\|A\| := \sup\left\{\frac{\|Ax\|}{\|x\|},\, x \in \mathbb{K}^n, x \neq 0\right\} = \sup\left\{\|Ax\|,\, x \in \mathbb{K}^n, \|x\| = 1\right\}$$

eine mit ihr "verträgliche" Matrizennorm erklärt ist. Diese wird als die von $\|\cdot\|$ erzeugte "natürliche" Matrizennorm bezeichnet. Warum kann die Quadratsummennorm (sog. "Frobenius-Norm") keine natürliche Matrizennorm sein?

$$\|A\|_{\mathrm{FR}} = \Big(\sum_{i,j=1}^{n} |a_{ij}|^2\Big)^{1/2}$$

Übung 4.2: Man betrachte das lineare Gleichungssystem

$$\begin{pmatrix} 0,5 & 0,5 \\ 0,5 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Wie groß sind die relativen Fehler $\|\delta x\|_1/\|x\|_1$ und $\|\delta x\|_\infty/\|x\|_\infty$, wenn der relative Fehler in den Matrixelementen höchstens $\pm 1\%$ und der in den Komponenten der rechten Seite höchstens $\pm 3\%$ beträgt? Man zeichne die Punktmenge im $\mathbb{R}^2$, in denen die Lösung $x+\delta x$ des gestörten Systems liegt. (Hinweis: Man berechne die Inverse der Koeffizientenmatrix und bestimme damit die l_1- und die l_∞-Kondition.)

Übung 4.3: a) Man löse durch Gaußsche Elimination (ohne Pivotierung) das lineare Gleichungssystem $Ax = b$, wobei (Hinweis: Der Lösungsvektor ist ganzzahlig.)

$$A = \begin{bmatrix} -\frac{1}{2} & 9 & -2 & 1 \\ -\frac{3}{2} & 30 & -12 & 0 \\ 1 & -15 & 0 & -4 \\ 0 & -6 & 18 & 8 \end{bmatrix}, \qquad b = \begin{bmatrix} 3 \\ 3 \\ 2 \\ -4 \end{bmatrix}.$$

b) Man bestimme die LR-Zerlegung von A und berechne die Determinante $\det(A)$. c) Man bestimme die Inverse A^{-1} und die Konditionszahl $\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$.

Übung 4.4: Sei $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix. Der Gaußsche Zerlegungsalgorithmus (ohne Pivotierung) erzeugt bei Anwendung auf A eine Folge von Matrizen $A = A^{(0)} \to \ldots \to A^{(k)} \to \ldots \to A^{(n-1)} = R$ mit einer oberen Dreiecksmatrix $R = (r_{ij})_{i,j=1}^n$ als Resultat. Man zeige, daß dieser Algorithmus im folgenden Sinne "stabil" ist:

$$k = 1, \ldots, n-1: \qquad a_{ii}^{(k)} \le a_{ii}^{(k-1)}, \;\; i = 1, \ldots, n, \quad \max_{1\le i,j\le n} |r_{ij}| \le \max_{1\le i,j\le n} |a_{ij}|.$$

(Hinweis: Man gehe von den Rekursionsformeln der Eliminationsprozesses aus.)

Übung 4.5: (Praktische Aufgabe) Das folgende MATLAB-Programm leistet die Berechnung der LR-Zerlegung $A = LR$ (sofern sie existiert) einer regulären Matrix:

```
function [L,R] = LR(A)
%--------------------------------------------------------------------
% Berechnet eine LR-Zerlegung (Gauss-Elimination).
% Eingabe: A nxn-regulre Matrix.
% Ausgabe: L nxn-untere Dreiecksmatrix, R nxn-obere Dreiecksmatrix.
%--------------------------------------------------------------------
[m,n] = size(A);
if (m ~= n), error('Matrix A ist nicht quadratisch'), end
for k=1:n-1
  A(k+1:n,k) = A(k+1:n,k)/A(k,k);
  A(k+1:n,k+1:n) = A(k+1:n,k+1:n)-A(k+1:n,k)*A(k,k+1:n);
end
L = eye(n,n) + tril(A,-1); R = triu(A); return;
```

(i) Man berechne die LR-Zerlegung der symmetrischen, positiv definiten Blockmatrix

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

($n = m^2$) mit der m-dimensionalen Einheitsmatrix I_m, für $m = 2^k$, $k = 1, \ldots, 7$. Die Leerstellen sind dabei mit Nullen besetzt gedacht. Mit Hilfe eines selbst erstellten Programms bestimme man mit Hilfe der gewonnenen LR-Zerlegung $A_n = L_n R_n$ noch die Cholesky-Zerlegung $A_n = L_n L_n^T$ und die Inverse A_n^{-1}. Die Genauigkeit überprüfe man jeweils durch Berechnung der Fehlernormen

$$\|A_n - L_n R_n\|_\infty, \quad \|A_n - L_n L_n^T\|_\infty, \quad \|A_n A_n^{-1} - I_n\|_\infty .$$

(ii) Was läßt sich über die l_∞-Konditionszahl $\mathrm{cond}_\infty(A_n) = \|A_n\|_\infty \|A_n^{-1}\|_\infty$ von A_n in Abhängigkeit von der Dimension n sagen?

Übung 4.6: Man zeige für allgemeine Matrizen $A \in \mathbb{K}^{n\times n}$ die Beziehung

$$\|A\|_2 := \sup\left\{\frac{\|Ax\|_2}{\|x\|_2},\, x \in \mathbb{K}^n, x \neq 0\right\} = \sup\left\{\sqrt{|\lambda|},\, \lambda \text{ Eigenwert von } \bar{A}^T A\right\}.$$

(Hinweis: Siehe den Beweis für hermitesches A in der Vorlesung. Man beachte, daß für allgemeines A die Matrix $\bar{A}^T A$ stets hermitesch ist und somit eine Orthonormalbasis von Eigenvektoren besitzt.)

Übung 4.7: Unter einer "LR-Zelegung" einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ versteht man allgemein eine Produktzerlegung der Form $A = LR$ mit einer unteren Dreiecksmatrix L, mit Einsen auf der Hauptdiagonalen, und einer (regulären) oberen Dreiecksmatrix R.

(i) Man verifiziere, daß die regulären unteren Dreiecksmatrizen $L \in \mathbb{K}^{n\times n}$, mit Einsen auf der Hauptdiagonalen, und ebenso die allgemeinen regulären oberen Dreiecksmatrizen $R \in \mathbb{K}^{n\times n}$ bezüglich der üblichen Matrizenmultiplikation "Gruppen" bilden. Sind diese Gruppen abelsch?

(ii) Man zeige damit, daß die mit dem Gaußschen Verfahren erzeugte LR-Zerlegung einer regulären Matrix $A \in \mathbb{K}^{n\times n}$ (falls sie existiert) eindeutig bestimmt ist.

Übung 4.8: Gegeben sei das Gleichungssystem $Ax = b$ mit

$$A = \begin{bmatrix} 5 & -5 & 0 & 0 \\ -5 & 7 & -2 & 0 \\ 0 & -2 & 20 & -18 \\ 0 & 0 & -18 & 19 \end{bmatrix}, \qquad b = \begin{bmatrix} 5 \\ -7 \\ 20 \\ -17 \end{bmatrix}$$

und der Lösung $x = (2, 1, 2, 1)^T$.

a) Man bestimme eine Näherungslösung mit dem Cholesky-Algorithmus unter Verwendung 4-stelliger Arithmetik mit korrekter Rundung.

b) Man versuche, das Ergebnis durch einen Nachiterationsschritt unter Verwendung 8-stelliger Arithmetik für den Defekt zu verbessern.

Übung 4.9: Sei $A \in \mathbb{R}^{n\times n}$ eine reguläre Matrix, für die eine LR-Zerlegung existiert. In der Vorlesung wurde gezeigt, daß sich diese mit dem Gaußschen Eliminationsverfahren (ohne Pivotierung) in $\frac{1}{3}n^3 + O(n^2)$ a. Op. berechnen läßt. Im Falle einer symmetrischen Matrix reduziert sich dieser Aufwand zu $\frac{1}{6}n^3 + O(n^2)$ a. Op. Dabei entspricht eine "a. Op." gerade einer Multiplikation (mit einer Addition) oder einer Division.

Frage: Wie sehen diese Aufwandszahlen für Band-Matrizen vom Typ (m_l, m_r) mit $m_l = m_r$ aus? Man konkretisiere dies anhand der Modellmatrix aus der Vorlesung (s. Aufgabe 8.5a) für die Werte $m = 10^2$ bzw. $n = m^2 = 10^4$.

Übung 4.10: (Praktische Aufgabe) a) Man schreibe ein MATLAB-Programm zur Berechnung der Cholesky-Zerlegung von symmetrischen positiv definiten Matrizen mit Hilfe des Algorithmus von Cholesky und wende es an für die Modellmatrix

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

($n = m^2$) mit der m-dimensionalen Einheitsmatrix I_m, für $m = 2, \ldots, 20$. Man überprüfe die Genauigkeit wieder durch die Probe $\|A_n - L_n^T L_n\|_\infty$. Welche Einsparungen an Speicherplatz und a. Op. ließen sich hier durch Ausnutzen der Matrixstruktur erzielen?

b) Man wende das Programm auf die Hilbert-Matrix

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix},$$

an für $n = 2, \ldots, 20$ und plotte die Residuennorm $\|A_n - L_n^T L_n\|_\infty$. Welche Ergebnisse liefert hier das MATLAB-interne Programm zur Cholesky-Zerlegung?

Übung 4.11: Betrachtet werde das Gleichungssystem $Ax = b$ der Form

$$\begin{bmatrix} 1 & 3 & -4 \\ 3 & 9 & -2 \\ 4 & 12 & -6 \\ 2 & 6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

a) Man untersuche, ob das System lösbar ist (mit Begründung).

b) Man bestimme eine Lösung nach der Methode der kleinsten Fehlerquadrate.

c) Ist diese Lösung eindeutig?

d) Ist die Matrix $A^T A$ positiv definit?

Übung 4.12: Wenn in dem Gleichungssystem von Aufgabe 9.1 einzelnen der Gleichungen bei der Lösung mehr Gewicht beibemessen werden soll, z.B. weil die zugehörigen Meßwerte zuverlässiger als die anderen sind, so kann dies dadurch berücksichtigt werden, daß statt $\|Ax - b\|_2$ eine gewichtete Quadratsumme $\|D(Ax - b)\|_2$ minimiert wird. Dabei ist $D = \mathrm{diag}(d_{ii})$ eine Diagonalmatrix mit Elementen $d_{ii} > 0$. Wie lautet in diesem Fall das zugehörige Normalgleichungssystem?

Übung 4.13: Man berechne die QR-Zerlegung der Matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 2 \\ -2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}$$

mit Hilfe des Householder-Verfahrens.

Übung 4.14: Nach dem ersten Keplerschen Gesetz bewegt sich ein Komet im Sonnensystem auf einer ebenen Bahn von Ellipsen- oder Hyperbelform, wenn Störungen durch die Planeten vernachlässigt werden. Bezüglich eines in der Sonne zentrierten polaren (r, φ)-Koordinatensystems wird diese Bahn durch die sog. "Kegelschnittgleichung"

$$r = \frac{p}{1 - e\,\cos(\varphi)}$$

beschrieben mit der sog. "Exzentrizität" e und einem Parameter p. Für $0 \le e < 1$ liegt eine Ellipse, für $e = 1$ eine Parabel und für $e > 1$ eine Hyperbel vor. Für einen neu entdeckten Kometen wurden die folgenden Beobachtungen gemacht:

Meßtag	15. Jan.	15. April	15. Juni	15. MAug.	15. Sept.	
r	10	5	2.5	1.3	1	(Einheiten)
$\cos(\varphi)$	$\sim 0,63$	$\sim 0,39$	$\sim 0,12$	$\sim -0,31$	$\sim -0,59$	

Man bestimme mit Hilfe der Gaußschen Ausgleichsrechnung den Typ der Kometenbahn. (Hinweis: Man schreibe die Kegelschnittgleichung zunächst in der Form $1/p - e/p\cos(\varphi) = 1/r$, die linear in $1/p$ und $1/e$ ist. Es genügt zweistellige Rechnung).

Übung 4.15: (Praktische Aufgabe) a) Man schreibe ein Programm zur Berechnung der QR-Zerlegung einer Matrix $A \in \mathbb{R}^{n\times n}$ mit dem Householder-Verfahren und teste es für die Matrix aus Aufgabe 9.3.

b) Die numerische Stabilität des Algorithmus untersuche man anhand der Hilbert-Matrix

$$H_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix},$$

für $n = 2^k$, $k = 1, 2, \ldots, 8$. Man überprüfe die Genauigkeit der QR-Zerlegung anhand der Defektnorm $\|A - QR\|_\infty$.

c) Die QR-Zerlegung einer Matrix $A \in \mathbb{R}^{n\times n}$ liefert die Cholesky-Zerlegung der Matrix A^TA gemäß $A^TA = R^TR$. Man überprüfe die Qualität dieser Zerlegung für die Hilbert-Matrix H_n anhand der Defektnorm $\|A^TA - R^TR\|_\infty$ und vergleiche dies mit der Cholesky-Zerlegung von A^TA, welche mit Hilfe des Programms aus Aufgabe 7.5 erzeugt wird.

5 Nichtlineare Gleichungen

Es sei f eine (reellwertige) stetige Funktion auf einem Intervall $I = [a, b]$. Das einfachste Verfahren zur Bestimmung von Nullstellen von f beruht auf der folgenden Konsequenz des Zwischenwertsatzes für stetige Funktionen: *Existiert ein Teilintervall* $I_0 = [a_0, b_0] \subset I$ *mit* $f(a_0)f(b_0) < 0$, *so hat* f *in* I_0 *mindestens eine Nullstelle.* Die sog. "Intervallschachtelung" erzeugt nun ausgehend von einem solchen I_0 eine Folge von Intervallen $I_t = [a_t, b_t]$, $t = 1, 2, \ldots$, welche jeweils mindestens eine Nullstelle von f enthalten, durch die Iteration

$$x_t := \tfrac{1}{2}(a_t + b_t)\,, \qquad (\, f(x_t) = 0 \;\Rightarrow\; \text{STOP}\,),$$

mit der Auswahlvorschrift

$$\begin{aligned} f(a_t)f(x_t) < 0 \quad &\Rightarrow \quad a_{t+1} := a_t\,, \quad b_{t+1} := x_t\,, \\ f(a_t)f(x_t) > 0 \quad &\Rightarrow \quad a_{t+1} := x_t\,, \quad b_{t+1} := b_t. \end{aligned}$$

Offenbar ist dann $a_t \le a_{t+1} \le b_{t+1} \le b_t$ und

$$|b_{t+1} - a_{t+1}| = \tfrac{1}{2}|b_t - a_t| = 2^{-t-1}|b_0 - a_0|. \tag{5.0.1}$$

Die monotonen Zahlenfolgen $(a_t)_{t\in\mathbb{N}}$, $(b_t)_{t\in\mathbb{N}}$ konvergieren gegen ein $z \in I_0$, welches wegen $f(z)^2 = \lim_{t\to\infty} f(a_t)f(b_t) \le 0$ notwendig Nullstelle von f ist. Dieses Verfahren ist numerisch sehr stabil, aber auch sehr langsam; für $b_0 - a_0 = 1$ erhält man z.B. aus der obigen a priori Abschätzung ($2^{-10} \le 10^{-3}$):

$$|x_9 - z| < 10^{-3}\,, \quad |x_{19} - z| < 10^{-6}\,, \quad |x_{29} - z| < 10^{-9}.$$

Die Intervallschachtelung für stetige Funktionen liefert stets eine Nullstelle, sofern für das Startintervall ein Vorzeichenwechsel vorliegt. Dieses Vorgehen ist naturgemäß auf *reelle* Funktionen beschränkt. Die im folgenden betrachteten Verfahren sind dagegen teilweise auch für komplexwertige Funktionen anwendbar.

5.1 Das Newton-Verfahren im $\mathbb{R}^1$

Ist die gegebene Funktion f auf dem Intervall $[a, b]$ stetig differenzierbar, so kann diese Zusatzinformation zur effizienteren Berechnung einer Nullstelle verwendet werden. Das (klassische) "Newton-Verfahren" (auch "Newton-Raphson[1]-Verfahren" genannt) ist motiviert durch die folgende graphische Überlegung (s. Abb. 5.1).

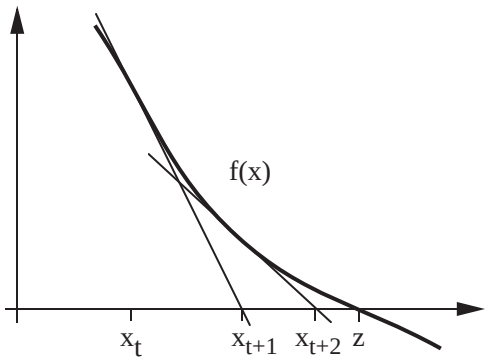


Abbildung 5.1: *Geometrische Interpretation des Newton-Verfahrens*

Im Punkt x_t wird die Tangente an $f(x)$ berechnet und deren Schnittpunkt mit der x-Achse als neue Näherung x_{t+1} für die Nullstelle z von f genommen. Die Tangente ist gegeben durch die Gleichung

$$T(x) = f'(x_t)(x - x_t) + f(x_t).$$

Ihre Nullstelle x_{t+1} ist bestimmt durch

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}. \tag{5.1.2}$$

Diese Iteration ist offenbar möglich, wenn die Ableitungswerte $f'(x_t)$ nicht zu klein werden. In dieser Form gestattet das Newton-Verfahren es also, *einfache* Nullstellen zu approximieren.

Satz 5.1 (Newton-Verfahren): *Die Funktion $f \in C^2[a, b]$ habe im Innern des Intervalls $[a, b]$ eine Nullstelle z, und es sei*

$$m := \min_{a \le x \le b} |f'(x)| > 0\,, \quad M := \max_{a \le x \le b} |f''(x)|.$$

[1] Joseph Raphson (1648-1715): Englischer Mathematiker; wirkte an der Universität Cambride; sein Buch "Analysis Aequationum Universalis" (1660) enthält bereits die Newton-Methode (50 Jahre vor Newton selbst); übersetzte einige Werke Newtons (von Latein nach Englisch); wichtige eigene Beiträge zur Analysis.

Sei $\rho > 0$ *so gewählt, daß*

$$q := \frac{M}{2m}\,\rho < 1, \qquad K_\rho(z) := \{x \in \mathbb{R} \,|\, |x - z| \le \rho\} \subset [a, b]. \tag{5.1.3}$$

Dann sind für jeden Startpunkt $x_0 \in K_\rho(z)$ *die Newton-Iterierten* $x_t \in K_\rho(z)$ *definiert und konvergieren gegen die Nullstelle* z *. Dabei gelten die a priori Fehlerabschätzung*

$$|x_t - z| \le \frac{2m}{M}\, q^{(2^t)}\,, \quad t \in \mathbb{N}, \tag{5.1.4}$$

und die a posteriori Fehlerabschätzung

$$|x_t - z| \le \frac{1}{m}\,|f(x_t)| \le \frac{M}{2m}\,|x_t - x_{t-1}|^2, \quad t \in \mathbb{N}. \tag{5.1.5}$$

Beweis: Der Beweis erfordert einige Vorbereitungen. Für Punkte $x, y \in [a, b]$, $x \neq y$, gilt aufgrund des Mittelwertsatzes der Differentialrechnung mit einem $\zeta \in [x, y]$:

$$\left|\frac{f(x) - f(y)}{x - y}\right| = |f'(\zeta)| \;\ge\; m,$$

und folglich

$$|x - y| \le \frac{1}{m}\,|f(x) - f(y)|.$$

(Die Nullstelle z von f ist also die einzige in $[a, b]$.) Weiter gilt die Taylor-Formel mit Restglied zweiter Ordnung:

$$f(y) = f(x) + (y - x)f'(x) + \underbrace{(y - x)^2 \int_0^1 f''(x + s(y - x))(1 - s)\, ds}_{=:\, R\,(y;\,x)}\,.$$

Mit Hilfe der Voraussetzung erhalten wir

$$|R(y; x)| \le M\,|y - x|^2 \int_0^1 (1 - s)\, ds = \frac{M}{2}|y - x|^2.$$

Für $x \in K_\rho(z)$ setzen wir $g(x) := x - \frac{f(x)}{f'(x)}$ und finden

$$g(x) - z = x - \frac{f(x)}{f'(x)} - z = -\frac{1}{f'(x)}\,\underbrace{\{f(x) + (z - x)f'(x)\}}_{=\,-R\,(z;\,x)}\,.$$

Also ist

$$|g(x) - z| \leq \frac{M}{2m} |x - z|^2 \leq \frac{M}{2m} \rho^2 < \rho, \tag{5.1.6}$$

d. h.: $g(x) \in K_\rho(z)$. Die Abbildung g bildet die Menge $K_\rho(z)$ in sich ab. Für $x_0 \in K_\rho(z)$ bleiben also alle Newton-Iterierten in $K_\rho(z)$. Setzt man

$$\rho_t := \frac{M}{2m} |x_t - z|,$$

so impliziert (5.1.6), daß

$$\rho_t \leq \rho_{t-1}^2 \leq \ldots \leq \rho_0^{2^t}, \quad |x_t - z| \leq \frac{2m}{M} \rho_0^{2^t}.$$

Für $\rho_0 = \frac{M}{2m} |x_0 - z| \leq \frac{M}{2m} \rho < 1$ liegt also die Konvergenz $x_t \to z (t \to \infty)$ vor mit der behaupteten a priori Fehlerabschätzung. Zum Beweis der a posteriori Fehlerabschätzung setzt man in der Taylor-Formel $y = x_t,\ x = x_{t-1}$, und erhält

$$f(x_t) = \underbrace{f(x_{t-1}) + (x_t - x_{t-1}) f'(x_{t-1})}_{=0} + R(x_t; x_{t-1})$$

bzw.

$$|x_t - z| \leq \frac{1}{m} |f(x_t) - \underbrace{f(z)}_{=0}| \leq \frac{M}{2m} |x_t - x_{t-1}|^2.$$

Dies vervollständigt den Beweis. Q.E.D.

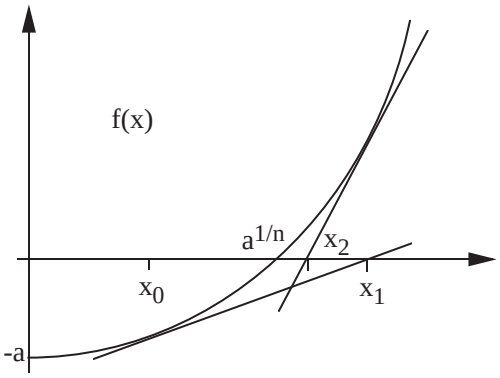
Für eine zweimal stetig differenzierbare Funktion f existiert zu jeder einfachen Nullstelle z ($f(z) = 0,\ f'(z) \neq 0$) stets eine (möglicherweise sehr kleine) Umgebung $K_\rho(z)$, für welche die Voraussetzungen von Satz 5.1 erfüllt sind. Das Problem beim Newton-Verfahren ist also die Bestimmung eines im "Einzugsbereich" der Nullstelle z gelegenen Startpunktes x_0. Ist ein solcher einmal gefunden, so konvergiert das Newton-Verfahren enorm schnell gegen die Nullstelle z: Im Fall $q \leq \frac{1}{2}$ gilt z.B. nach nur 10 Iterationsschritten bereits ($2^{10} > 1.000$)

$$|x_{10} - z| \leq \frac{2m}{M} q^{1.000} \sim \frac{2m}{M} 10^{-300}.$$

Beispiel 5.1: Newton-Verfahren zur Wurzelberechnung:
Die n-te Wurzel einer Zahl $a > 0$ ist Nullstelle der Funktion $f(x) = x^n - a$. Das Newton-Verfahren zur Berechnung von $z = \sqrt[n]{a} > 0$ hat die Gestalt

$$x_{t+1} = x_t - \frac{x_t^n - a}{n x_t^{n-1}} = \frac{1}{n} \left\{ (n-1)\, x_t + \frac{a}{x_t^{n-1}} \right\}. \tag{5.1.7}$$

Aufgrund von Satz 5.1 konvergiert $x_t \to z\,(t \to \infty)$, wenn nur x_0 nahe genug bei z gewählt wird. Bei diesem einfachen Beispiel kann aber mit Hilfe der folgenden geometrischen Betrachtung die Konvergenz für jeden Startpunkt $x_0 > 0$ gesichert werden (s. Abb. 5.1) Die monoton fallende Folge $(x_t)_{t\in\mathbb{N}}$ konvergiert notwendig gegen $\sqrt[n]{a}$. Für hinreichend großes t ist dann x_t i.m Einzugsbereich der Nullstelle z, und die Fehlerabschätzung von Satz 5.1 gelten mit diesem x_t als Startpunkt. Auf diese Weise wird auf vielen Rechenern die Wurzel $\sqrt[n]{a}$ berechnet.



$$x_0 > 0 \;\Rightarrow\; \begin{cases} x_t > \sqrt[n]{a}\,, & t \in \mathbb{N}\,. \\ \sqrt[n]{a} < x_{t+1} < x_t \end{cases}$$

Abbildung 5.2: *Newton-Iteration zur Wurzelberechnung*

Für den Spezialfall $n = 2$ wollen wir den Einzugsbereich der quadratischen Konvergenz des Newton-Verfahrens bestimmen. Es gilt

$$x_{t+1} - \sqrt{a} = \frac{1}{2}\left\{x_t + \frac{a}{x_t}\right\} - \sqrt{a} = \frac{1}{2x_t}\left\{x_t^2 + a - 2x_t\sqrt{a}\right\} = \frac{1}{2x_t}\,(x_t - \sqrt{a})^2,$$

also für $t \geq 1$ (wegen $x_t > \sqrt{a}$)

$$|x_{t+1} - \sqrt{a}| \leq \frac{1}{2\sqrt{a}}\,|x_t - \sqrt{a}|^2.$$

Quadratische Konvergenz liegt vor für Startwerte x_0 mit der Eigenschaft

$$\frac{1}{2\sqrt{a}}\,|x_0 - \sqrt{a}| < 1 \quad \text{bzw.} \quad |x_0 - \sqrt{a}| < 2\sqrt{a}.$$

Aus der Ungleichung $\sqrt{a} \leq x_t\,,\ t \in \mathbb{N}$, ergibt sich noch die Beziehung

$$\frac{a}{x_t} \leq \sqrt{a} \leq x_t,$$

was für ein Abbruchkriterium verwendet werden kann:

$$0 \leq e_t := x_t - \frac{a}{x_t} \leq \varepsilon \quad \Longrightarrow \quad \text{STOP}.$$

Die folgende Tabelle zeigt das Konvergenzverhalten der Newton-Iteration zur Berechnung von $x = \sqrt{2} = 1.414213562373095\ldots$ (16-stellige Rechnung). In jedem Iterationsschritt verdoppelt sich die Anzahl der richtigen Dezimalen:

$$\begin{aligned} x_0 &= 2 \\ x_1 &= \underline{1}.5 \\ x_2 &= \underline{1.416}\,, & e_2 &\le 5 \cdot 10^{-3} \\ x_3 &= \underline{1.41421}568627451\,, & e_3 &\le 5 \cdot 10^{-6} \\ x_4 &= \underline{1.41421356137}469\,, & e_4 &\le 5 \cdot 10^{-12}. \end{aligned}$$

Bemerkung 5.1: Die Bedingungen von Satz 5.1 lassen sich so modifizieren, daß auf die Voraussetzung der Existenz einer Nullstelle verzichtet werden kann, und, ähnlich wie beim Banachschen Fixpunktsatz, die Konvergenz der Newton-Folge gegen eine (lokal eindeutige) Nullstelle folgt. Diese Variante von Satz 5.1, der sog. "Satz von Newton-Kantorowitsch", wird im Rahmen der Diskussion des Newton-Verfahrens im $\mathbb{R}^n$ bewiesen werden.

Bemerkung 5.2: Das Hauptproblem bei der Durchführung des Newton-Verfahrens ist die Bestimmung eines geeigneten Startwertes x^0, da der Einzugsbereich der quadratischen Konvergenz in der Praxis häufig sehr klein ist. Deshalb arbeitet man stets mit dem sog. "gedämpften Newton-Verfahren"

$$x_{t+1} = x_t - \lambda_t \frac{f(x_t)}{f'(x_t)}, \tag{5.1.8}$$

mit einem "Dämpfungsparameter" $\lambda_t \in (0, 1]$. Die geeignete Wahl dieses Dämpfungsparameters ist eine Wissenschaft für sich. Sie wird später im Zusammenhang mit dem Newton-Verfahren im $\mathbb{R}^n$ diskutiert werden.

Mehrfache Nullstellen

Wir betrachten nun den kritischen Fall, daß mit dem Newton-Verfahren eine mehrfache Nullstelle berechnet werden soll. Sei dazu zunächst z eine zweifache Nullstelle der Funktion f, d. h.: $f(z) = f'(z) = 0,\ f''(z) \neq 0$. Für die Newton-Iteration gilt dann

$$x_{t+1} = x_t - \frac{f(x_t) - f(z)}{f'(x_t) - f'(z)} = x_t - \frac{f'(\zeta_t)}{f''(\eta_t)}$$

mit Zwischenpunkten $\zeta_t, \eta_t \in [x_t, z]$. Der Quotient $f(x_t)/f'(x_t)$ bleibt also für $x_t \to z$ wohl definiert. Sei nun allgemein z eine p-fache Nullstelle der Funktion $f \in C^{p+1}[a, b]$:

$$f(z) = \ldots = f^{(p-1)}(z) = 0\,, \quad f^{(p)}(z) \neq 0.$$

Aus der Taylor-Formel um z

$$f(x) = \underbrace{f(z)}_{=0} + \ldots + \frac{1}{(p-1)!}(x-z)^{p-1}\underbrace{f^{(p-1)}(z)}_{=0} + (x-z)^p \underbrace{\frac{1}{p!}f^{(p)}(\zeta_x)}_{=:\,Q(z;x)}$$

folgt durch Ableiten

$$f'(x) = Q'(z;x)(x-z)^p + p\,Q(z;x)(x-z)^{p-1}.$$

Also ist für $f'(x) \neq 0$:

$$\begin{aligned}\frac{f(x)}{f'(x)} &= \frac{(x-z)Q(z;x)}{Q'(z;x)(x-z) + p\,Q(z;x)} \\ &= \frac{x-z}{p} - \frac{1}{p}(x-z)^2\frac{Q'(z;x)}{Q'(z;x)(x-z) + p\,Q(z;x)}.\end{aligned}$$

Für den Iterationsansatz

$$x_{t+1} = x_t - \alpha\,\frac{f(x_t)}{f'(x_t)} \tag{5.1.9}$$

folgt dann

$$\begin{aligned}x_{t+1} - z &= x_t - z - \alpha\,\frac{f(x_t)}{f'(x_t)} \\ &= (x_t - z)\Big(1 - \frac{\alpha}{p}\Big) + (x_t - z)^2\,\frac{\alpha Q'(z;x_t)}{pQ'(z;x_t)(x_t - z) + p^2 Q(z;x_t)}.\end{aligned}$$

Bei der Wahl von $\alpha = p$ erhält man für das so modifizierte Newton–Verfahren

$$x_{t+1} = x_t - p\frac{f(x_t)}{f'(x_t)}. \tag{5.1.10}$$

ein analoges "quadratisches" Konvergenzverhalten wie im Fall einer einfachen Nullstelle.

Vereinfachtes Newton-Verfahren

Ist z Nullstelle einer stetig differenzierbaren Funktion f, so konvergiert die Newton-Iteration

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} \to z \quad (t \to \infty),$$

wenn x_0 hinreichend nahe bei z gewählt war. Jeder Iterationsschritt erfordert die Auswertung der Ableitung $f'(x_t)$, was bei komplizierten (möglicherweise auch nur implizit definierten) Funktionen f unter Umständen zuviel Aufwand erfordert. In solchen Fällen

geht man zum sog. "vereinfachten Newton-Verfahren" über

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(c)} \tag{5.1.11}$$

mit einem festen, geeignet gewählten Punkt c. Diese Iteration ist Spezialfall der allgemeineren "Fixpunktiteration"

$$x_{t+1} = x_t + \sigma\, f(x_t) \tag{5.1.12}$$

mit einer geeigneten Zahl $\sigma \in \mathbb{R}$, $\sigma \neq 0$, zur Berechnung einer Nullstelle von f. Konvergiert hier $x_t \to z\ (t \to \infty)$, so gilt im Limes

$$\begin{array}{ccccccc} x_{t+1} & = & x_t & + & \sigma f(x_t) & & \\ \downarrow & & \downarrow & & \downarrow & & (t \to \infty) \\ z & = & z & + & \sigma f(z) & & \end{array}$$

d. h.: z ist "Fixpunkt" der Abbildung $g(x) = x + \sigma f(x)$ und wegen $\sigma \neq 0$ notwendig Nullstelle von f. Der Vorteil der obigen Fixpunktiteration besteht in ihrer ableitungsfreien Form. Kriterien für die Konvergenz einer Fixpunktiteration $x_{t+1} = g(x_t)$, $t = 0, 1, 2, \ldots$, werden wir später in einem etwas allgemeineren Rahmen herleiten.

Wir wollen nun noch das Newton-Verfahren zur Berechnung von Nullstellen von Polynomen

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n\,, \quad a_n \neq 0,$$

spezialisieren. Zunächst verschafft man sich etwa (im Reellen) mit der Intervallschachtelung einen groben Überblick über die Lage der Nullstellen. Nach Vorgabe einer Fehlertoleranz $\varepsilon \gg \text{eps}$ lautet der Newton-Algorithmus dann wie folgt:

1. Wahl eines Startwertes x;

2. Auswertung von $p(x)$ und $p'(x)$ mit dem Horner Schema:

$$\begin{aligned} i = n, n-1, \ldots, 0 : \quad & \alpha_i = a_i + \alpha_{i+1} x\,, \quad \beta_i = \alpha_i + \beta_{i+1} x \\ & (\alpha_{n+1} = \beta_{n+1} = 0) \\ & p(x) = \alpha_0\,, \quad p'(x) = \beta_1\,, \\ (\beta_1 = 0 : \quad \text{Startwert ändern}) \quad & \text{sei} \quad \beta_1 \neq 0\,; \end{aligned}$$

3. Newton-Korrektur $\quad q = \dfrac{\alpha_0}{\beta_1}\,, \quad |q| \leq \varepsilon \begin{cases} \text{ja :} & x \quad \text{wird akzeptiert;} \\ \text{nein :} & \text{Iterationsschritt;} \end{cases}$

4. Iterationsschritt $\quad x := x - q\,, \quad$ weiter mit (2).

5.2 Das Konvergenzverhalten iterativer Verfahren

Das Newton-Verfahren besitzt lokal in der Umgebung einer Nullstelle die charakteristische Konvergenzeigenschaft

$$|x_t - z| \;\le\; c|x_{t-1} - z|^2. \tag{5.2.13}$$

Man nennt es daher "quadratisch konvergent" oder auch "von 2-ter Ordnung".

Definition 5.1: *Allgemein spricht man bei einem Iterationsverfahren zur Berechnung einer Größe z von Konvergenz mit der "Ordnung" α, $\alpha \ge 1$, wenn gilt*

$$|x_t - z| \;\le\; c|x_{t-1} - z|^{\alpha}, \tag{5.2.14}$$

mit einer festen Konstante $c > 0$. Im Fall $\alpha = 1$, d. h. "linearer" Konvergenz, heißt die "beste" Konstante c "lineare Konvergenzrate". Gilt die Abschätzung

$$|x_t - z| \le c_t|x_{t-1} - z| \tag{5.2.15}$$

mit einer Nullfolge $c_t \to 0$ $(t \to \infty)$, so spricht man von "superlinear" Konvergenz.

Im Fall $\alpha > 1$ impliziert die Beziehung (5.2.14) wiederum Konvergenz $x_t \to z$ $(t \to \infty)$, wenn der Startwert x_0 hinreichend nahe bei z liegt:

$$c^{\frac{1}{\alpha-1}}|x_t - z| \;\le\; \left[c^{\frac{1}{\alpha-1}}|x_{t-1} - z|\right]^{\alpha} \;\le\; \ldots \;\le\; \underbrace{\left[c^{\frac{1}{\alpha-1}}|x_0 - z|\right]}_{<1\,!}^{\alpha^t} \;\to\; 0.$$

Im Fall $\alpha = 1$ folgt Konvergenz für $c < 1$:

$$|x_t - z| \le c|x_{t-1} - z| \le \ldots \le c^t|x_0 - z| \to 0 \quad (t \to \infty).$$

Bei Fixpunktiterationen $x_{t+1} = g(x_t)$ mit stetig differenzierbarer Abbildung g gilt

$$\left|\frac{x_{t+1} - z}{x_t - z}\right| = \left|\frac{g(x_t) - g(z)}{x_t - z}\right| \;\to\; |g'(z)| \quad (t \to \infty),$$

d. h.: Die lineare Konvergenzrate ist asymptotisch (für $t \to \infty$) gerade gleich $|g'(z)|$. Im Falle $g'(z) = 0$ liegt also (mindestens) superlineare Konvergenz der Fixpunktiteration vor.

Definition 5.2: *Ein Fixpunkt z einer stetig differenzierbaren Abbildung g heißt "anziehend", wenn $|g'(z)| < 1$ ist, da dann die sukzessive Approximiert für jeden hinreichend nahe bei z gelegenen Startwert gegen ihn konvergiert. Im Fall $|g'(z)| > 1$ heißt er "abstoßend", da er durch sukzessive Approximation i. Allg. nicht angenähert werden kann.*

Einen Hinweis zur Konstruktion von Verfahren höherer Ordnung gibt der folgende Satz.

Satz 5.2 (Iterative Verfahren): *Die Funktion g sei in einer Umgebung des Fixpunktes z p-mal stetig differenzierbar mit $p \geq 2$. Genau dann hat die Fixpunktiteration $x_{t+1} = g(x_t)$ die genaue Ordnung p, wenn*

$$g'(z) = \ldots = g^{(p-1)}(z) = 0 \quad \textit{und} \quad g^{(p)}(z) \neq 0. \tag{5.2.16}$$

Beweis: Sei $g'(z) = \ldots = g^{(p-1)}(z) = 0$. Die Taylor-Formel mit dem Restglied p-ter Ordnung erhält dann im Punkt z die Form

$$x_{t+1} - z = g(x_t) - g(z) = \sum_{i=1}^{p-1} \frac{(x_t - z)^i}{i!}\, g^{(i)}(z) + \frac{(x_t - z)^p}{p!}\, g^{(p)}(\zeta_t),$$

und folglich

$$|x_{t+1} - z| \leq \frac{1}{p!} \max |g^{(p)}|\, |x_t - z|^p.$$

Sei nun umgekehrt die Iteration von p-ter Ordnung, d. h.: $|x_{t+1} - z| \leq c|x_t - z|^p$. Gäbe es ein minimales $m \leq p-1$ mit $g^{(m)}(z) \neq 0$, aber $g^{(i)}(z) = 0$, $i = 1, \ldots, m-1$, so konvergierte jede Iteriertenfolge $(x_t)_{t\in\mathbb{N}}$ mit hinreichend kleinem $|x_0 - z| \neq 0$ notwendig gegen z wie

$$|x_t - z| = |\frac{1}{m!}\, g^{(m)}(\zeta_t)|\, |x_{t-1} - z|^m$$

Dies impliziert aber im Widerspruch zur Annahme:

$$|g^{(m)}(z)| = \lim_{t\to\infty} |g^{(m)}(\zeta_t)| \leq c\, m! \lim_{t\to\infty} |x_t - z|^{p-m} = 0.$$

Hieraus folgt auch, daß im Fall $g'(z) = \ldots = g^{(p-1)}(z) = 0$, aber $g^{(p)}(z) \neq 0$, die Iteration nicht von höherer als p-ter (ganzzahliger) Ordnung sein kann. Q.E.D.

Beispiel 5.2: Beim Newton-Verfahren zur Bestimmung einer einfachen Nullstelle der Funktion f ist $g(x) = x - f(x)/f'(x)$ also

$$g'(z) = 1 - \frac{f'(z)^2 - f(z)f''(z)}{f'(z)^2} = 0,$$

und i.allg. $g''(z) \neq 0$. Die Newton-Iteration ist also, wie wir schon gesehen haben, von 2-ter Ordnung.

Beispiel 5.3: Bei einer Fixpunktiteration von mindestens 3-ter Ordnung muß $g'(z) = g''(z) = 0$ gelten. Zur Konstruktion eines solchen Verfahrens zur Nullstellenbestimmung machen wir den Ansatz

$$g(x) = x - r(x) + s(x)r(x)^2 \quad \text{mit} \quad r(x) = \frac{f(x)}{f'(x)}.$$

Wegen $r(z) = 0$ und $r'(z) = 1$ ist hier automatisch $g'(z) = 0$. Die zusätzliche Forderung $g''(z) = 0$ wird z.B. erfüllt für

$$s(x) = \frac{r''(x)}{2r'(x)^2}.$$

Dieses Verfahren erfordert also die Auswertung der Ableitungen bis zur Ordnung 3 der Funktion f.

Zur Klärung der numerischen Bedeutung des Ordnungsbegriffes definieren wir für eine Iterationsfolge $(x_t)_{t\in\mathbb{N}}$

$$e_t := x_t - z \quad \text{(absoluter Fehler)}, \quad \bar{e}_t := \frac{e_t}{z} \quad \text{(relativer Fehler)}.$$

Haben x_t und z die dezimalen Gleitpunktdarstellungen (mit gemeinsamen Exponenten und m gleichen Mantissenstellen)

$$\begin{aligned} z &= a_m \dots a_1.a_{-1} \dots \cdot 10^s, \quad a_m \neq 0, \\ x_t &= a_m \dots a_1.\tilde{a}_{-1} \dots \cdot 10^s, \end{aligned}$$

so gilt

$$|\bar{e}_t| = \left|\frac{x_t - z}{z}\right| \leq 10^{-m},$$

d. h.: Die Größe

$$\rho_t := -\log_{10} |\bar{e}_t| = m$$

gibt ungefähr die Anzahl der richtigen Mantissendezimalen von x_t an. Wegen

$$|\bar{e}_{t+1}| = \left|\frac{x_{t+1} - z}{z}\right| = |g'(\zeta_t)| \left|\frac{x_t - z}{z}\right|, \quad \zeta_t \in [x_t, x_{t+1}],$$

gilt

$$\rho_{t+1} = -\log_{10} |\bar{e}_{t+1}| = -\log_{10} |g'(\zeta_t)| \underbrace{-\log_{10} |\bar{e}_t|}_{\rho_t}$$

und im Limes

$$\rho_{t+1} - \rho_t \to -\log_{10} |g'(z)| \quad (t \to \infty). \tag{5.2.17}$$

Die numerische Bedeutung der "asymptotischen" linearen Konvergenzrate $|g'(z)|$ einer Fixpunktiteration ist also, daß sich in jedem Iterationsschritt (für große t) die Anzahl der richtigen Mantissendezimalen um $-\log_{10} |g'(z)|$ erhöht (für $|g'(z)| \neq 0$).

Für eine Iteration p-ter Ordnung mit $p \geq 2$ gilt

$$|x_{t+1} - z| = \frac{1}{p!} |g^{(p)}(\zeta_t)| \, |x_t - z|^p, \quad t \geq 1,$$

mit $\zeta_t \to z\,(t\to\infty)$. Also ist in diesem Fall

$$|\bar{e}_{t+1}| = \left|\frac{x_{t+1}-z}{z}\right| = \underbrace{\left|\frac{1}{p!}\,g^{(p)}(\zeta_t)\right|\,|z|^{p-1}}_{=:\,\sigma_t}\;\underbrace{\left|\frac{x_t-z}{z}\right|^p}_{=\,|\bar{e}_t|^p}$$

und

$$\sigma_t \to \left|\frac{1}{p!}\,g^{(p)}(z)\right|\,|z|^{p-1} \quad (t\to\infty).$$

Es folgt die Beziehung $(\rho_t = -\log_{10}|\bar{e}_t|)$

$$\rho_{t+1} = p\,\rho_t - \log_{10}\sigma_t,$$

und hieraus wegen $\rho_t \to \infty\,(t\to\infty)$

$$\lim_{t\to\infty}\frac{\rho_{t+1}}{\rho_t} = p. \tag{5.2.18}$$

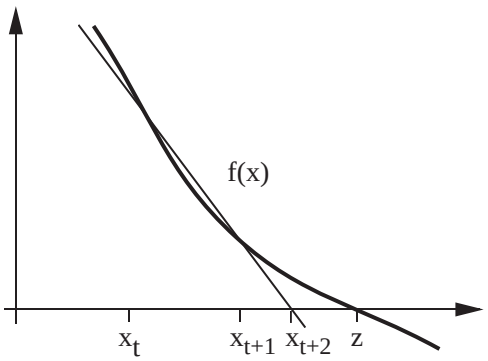
Dies läßt sich so interpretieren, daß sich bei einer Iteration p-ter Ordnung (für große t) die Anzahl der richtigen Mantissendezimalen in jedem Schritt etwa ver-p-facht. Dies wird durch unser obiges Beispiel beim Newton-Verfahren bestätigt. Wir fassen die bisher abgeleiteten reglen zusammen:

Regel 5.2.1: *Bei einem "linear" konvergenten Iterationsverfahren erhöht sich in jedem Schritt die Anzahl von exakten Dezimalstellen in der Näherung in etwa um den Summanden* $|\log_{10}(g'(z))|$*; bei einer Iteration der Ordnung* $p>1$ *ver-p-facht sich in jedem Schritt die Anzahl der exakten Dezimalstellen.*

5.3 Interpolationsmethoden

Das Ziel ist die iterative Berechnung von Nullstellen ohne Auswertung von Ableitungen, aber effizienter als mit Intervallschachtelung oder einfacher sukzessiver Approximation. Dabei werden wir auf ein Iterationsverfahren mit nicht ganzzahliger Ordnung geführt.

Die "Sekantenmethode" berechnet ausgehend von einem Paar von Werten x_{t-1}, x_t die neue Iterierte x_{t+1} als Nullstelle der Geraden (Sekante) durch die Punkte $(x_{t-1}, f(x_{t-1}))$, $(x_t, f(x_t))$ (s. Abb. 5.3):



$$s(x) = f(x_t) + (x - x_t)\frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$$

Iteration:

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}\,.$$

Abbildung 5.3: *Geometrische Interpretation des Sekanten-Verfahrens*

Im Gegensatz zu den bisher betrachteten Verfahren handelt es sich hierbei um ein sog. "Zweischrittverfahren", d. h.: Die Iterierte x_{t+1} wird jeweils aus den beiden vorausgehenden Iterierten $x_t\,,\, x_{t-1}$ berechnet. Zum Starten der Sekantenmethode sind zwei Anfangsschätzungen $x_0\,,\, x_1$ für die Nullstelle erforderlich. Analog zum Konvergenzsatz 5.1 für das Newton-Verfahren haben wir auch eine Konvergenzaussage für die Sekantenmethode. Dabei spielen die durch die Vorschrift

$$\gamma_0 = \gamma_1 = 1\,, \quad \gamma_{t+1} = \gamma_t + \gamma_{t-1}\,, \quad t \in \mathbb{N},$$

definierten sog. "Fibonacci[2]-Zahlen" γ_t eine wichtige Rolle.

Satz 5.3 (Sekanten-Methode): *Die Funktion $f \in C^2[a,b]$ habe im Innern des Intervalls $[a,b]$ eine Nullstelle z, und es sei*

$$m := \min_{a \le x \le b} |f'(x)| > 0\,, \quad M := \max_{a \le x \le b} |f''(x)| < \infty. \tag{5.3.19}$$

[2]Leonardo Pisano (aus Pisa), genannt Fibonacci (um 1170 - um 1250): "erster" bedeutender Mathematiker des Abendlandes; gehörte zum Gelehrtenkreis um Kaiser Friedrich II; brachte von ausgedehnten Reisen eine systematische Einführung in das indisch-arabische Zahlensystem nach Europa; in seinem Rechenbuch "Liber abacci" untersuchte er u.a. die nach ihm benannte Folge als einfaches Modell für das Wachstum von Populationen.

Sei ferner $\rho > 0$ *so gewählt, daß*

$$q \equiv \frac{M}{2m}\rho < 1, \qquad K_\rho(z) = \{x \in \mathbb{R} \,|\, |x - z| \le \rho\} \subset [a, b].$$

Dann sind für jedes Paar von Startwerten $x_0, x_1 \in K_\rho(z)$, $x_0 \neq x_1$, *die Iterierten* $x_t \in K_\rho(z)$ *der Sekantenmethode wohl definiert und konvergieren gegen die Nullstelle* z. *Dabei gelten die a priori Fehlerabschätzung*

$$|x_t - z| \le \frac{2m}{M}\, q^{\gamma_t}, \quad t \in \mathbb{N}, \tag{5.3.20}$$

und die a posteriori Fehlerabschätzung

$$|x_t - z| \le \frac{1}{m}\,|f(x_t)| \le \frac{M}{2m}\,|x_t - x_{t-1}|\,|x_t - x_{t-2}|, \quad t \in \mathbb{N}. \tag{5.3.21}$$

Beweis: Die Argumentation ist ähnlich wie im Beweis von Satz 5.1 für das Newton-Verfahren. Für je zwei Punkte $x, y \in [a, b]$, $x \neq y$, gilt wieder

$$|x - y| \le \frac{1}{m}\,|f(x) - f(y)|,$$

woraus u.a. die Eindeutigkeit der Nullstelle z folgt. Weiter ist

$$\frac{f(x) - f(y)}{x - y} = -\int_0^1 \frac{d}{dr}\, f(x + r(y - x))\, \frac{dr}{x - y} = \int_0^1 f'(x + r(y - x))\, dr.$$

Mit einem dritten Punkt $\zeta \in [a, b]$, $\zeta \neq x$, ergibt sich hiermit

$$\begin{aligned}
\frac{f(x) - f(y)}{x - y} - \frac{f(x) - f(\zeta)}{x - \zeta} &= \int_0^1 \Big\{ f'(x - r(y - x)) - f'(x + r(\zeta - x)) \Big\}\, dr \\
&= -\int_0^1 \Big\{ \int_0^r \frac{d}{ds}\, f'(x + r(y - x) + s(\zeta - y))\, ds \Big\}\, dr \\
&= \int_0^1 \Big\{ \int_0^r f''(x + r(y - x) + s(\zeta - y))\, ds \Big\}\, dr\,(y - \zeta),
\end{aligned}$$

bzw.

$$\Big|\frac{f(x) - f(y)}{x - y} - \frac{f(x) - f(\zeta)}{x - \zeta}\Big| \le \frac{M}{2}\,|y - \zeta|.$$

Für Punkte $x, y \in K_\rho(z)$, $x \neq y$, $x \neq z$, $y \neq z$ definieren wir

$$g(x, y) := x - f(x)\,\frac{x - y}{f(x) - f(y)}.$$

Dann gilt

$$\begin{aligned} g(x,y) - z &= x - z - f(x)\frac{x-y}{f(x)-f(y)} \\ &= \frac{x-y}{f(x)-f(y)} \left\{ (x-z)\frac{f(x)-f(y)}{x-y} - f(x) + \underbrace{f(z)}_{=0} \right\} \end{aligned}$$

und folglich

$$\begin{aligned} |g(x,y) - z| &\le |x-z| \, \Big| \frac{f(x)-f(y)}{x-y} - \frac{f(x)-f(z)}{x-z} \Big| \\ &\le \frac{M}{2m} \, |x-z| \, |y-z| \, \le \, \frac{M}{2m} \, \rho^2 < \rho. \end{aligned}$$

Die Iterierten x_t der Sekantenmethode bleiben also in der Menge $K_\rho(z)$, und es gilt

$$|x_{t+1} - z| \le \frac{M}{2m} |x_t - z| \, |x_{t-1} - z|.$$

Setzt man $\rho_t := \frac{M}{2m} |x_t - z|$, so folgt

$$\rho_{t+1} \, \le \, \rho_t \rho_{t-1} \, , \quad t \in \mathbb{N},$$

d. h. mit $\rho_0 \le q$, $\rho_1 \le q$ gilt $\rho_t \le q^{\gamma_t}$, $t \in \mathbb{N}$. Wegen $\gamma_t \to \infty \, (t \to \infty)$ und $q < 1$ konvergiert also

$$|x_t - z| = \frac{2m}{M} \, \rho_t \, \le \, \frac{2m}{M} \, q^{\gamma_t} \to 0 \quad (t \to \infty).$$

Zum Nachweis der a posteriori Fehlerabschätzung setzen wir oben $x = x_{t-1}$, $y = x_t$ und $\zeta = x_{t-2}$ ($x_{t-2} \ne x_{t-1}$, da sonst bereits $f(x_{t-1}) = 0$) und finden

$$\begin{aligned} |x_t - z| &\le \frac{1}{m} \, |f(x_t) - f(z)| \\ &\le \frac{1}{m} \, \Big| f(x_{t-1}) + (x_t - x_{t-1}) \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}} \Big| \\ &\le \frac{1}{m} \, |x_t - x_{t-1}| \, \Big| \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}} - \frac{f(x_{t-1}) - f(x_{t-2})}{x_{t-1} - x_{t-2}} \Big| \\ &\le \frac{M}{2m} \, |x_t - x_{t-1}| \, |x_t - x_{t-2}|. \end{aligned}$$

Q.E.D.

Zur Beurteilung der Konvergenzgeschwindigkeit der Sekantenmethode benötigen wir Informationen über das Anwachsen der Fibonacci-Zahlen γ_t für $t \to \infty$.

Hilfssatz 5.1: *Die Fibonacci-Zahlen verhalten sich asymptotisch wie*

$$\gamma_t \sim \frac{\lambda_1}{\sqrt{5}}\,\lambda_1^t \sim 0.723 \cdot (1.618)^t, \tag{5.3.22}$$

wobei $\lambda_1 := \frac{1}{2}(1 \pm \sqrt{5})$ *gerade der sog. "goldene Schnitt" ist.*

Beweis: Die Fibonacci-Zahlen genügen nach Konstruktion der (linearen) homogenen Differenzengleichung

$$\gamma_{t+2} - \gamma_{t+1} - \gamma_t = 0\,, \quad t \geq 0. \tag{5.3.23}$$

Zu ihrer Lösung machen wir den Ansatz $\gamma_t = \lambda^t$ und erhalten die Gleichung

$$\lambda^t(\lambda^2 - \lambda - 1) = 0$$

zur Bestimmung von λ. Die Wurzeln $\lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{5})$ der quadratischen Gleichung $\lambda^2 - \lambda - 1 = 0$ ergeben durch

$$\gamma_t = c_1\lambda_1^t + c_2\lambda_2^t\,, \quad c_1, c_2 \quad \text{beliebig}, \tag{5.3.24}$$

die allgemeine Lösung der Differenzengleichung. Durch Berücksichtigung der Anfangsbedingungen $\gamma_0 = \gamma_1 = 1$ werden die Konstanten c_1, c_2 festgelegt:

$$\left.\begin{array}{rcl} c_1 + c_2 & = & 1 \\ c_1\lambda_1 + c_2\lambda_2 & = & 1 \end{array}\right\} \Rightarrow c_1 = \frac{1-\lambda_2}{\lambda_1 - \lambda_2} = \frac{\lambda_1}{\sqrt{5}}\,, \quad c_2 = \frac{\lambda_1 - 1}{\lambda_1 - \lambda_2} = -\frac{\lambda_2}{\sqrt{5}}.$$

Die Fibonacci-Zahlen haben also die Gestalt

$$\gamma_t = \frac{1}{\sqrt{5}}\left\{\lambda_1^{t+1} - \lambda_2^{t+1}\right\}, \quad \lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{5}). \tag{5.3.25}$$

Asymptotisch für $t \to \infty$ verhält sich γ_t wie

$$\gamma_t \sim \frac{\lambda_1}{\sqrt{5}}\,\lambda_1^t \sim 0.723 \cdot (1.618)^t,$$

was zu zeigen war. Q.E.D.

Die Sekantenmethode konvergiert also asymptotisch mindestens so schnell wie ein Ein-Schrittverfahren der Ordnung $p = 1.6$. In jedem Schritt ist dabei nur eine neue Funktionsauswertung, nämlich die von $f(x_t)$, erforderlich. *Ein* Schritt des Newton-Verfahrens (Auswertung von $f(x_t)$ und $f'(x_t)$) ist also mindestens so aufwendig wie *zwei* Schritte der Sekantenmethode. Faßt man jedoch zwei Schritte der Sekantenmethode zu einem

Makroschritt zusammen, so erhält man wegen

$$|x_{2t} - z| \;\leq\; \frac{2m}{M}\, q^{\gamma_{2t}}\,, \quad \gamma_{2t} \sim 0.723\,(2.618)^t \quad (\lambda_1^2 = \lambda_1 + 1), \tag{5.3.26}$$

ein Verfahren der Ordnung $p \geq 2.6$. Bei gleichem Arbeitsaufwand konvergiert also die Sekantenmethode asymptotisch (für große t) schneller als das Newton-Verfahren. Dieser theoretische Vorteil wird aber in der Praxis oft durch eine große Rundungsfehleranfälligkeit der Sekantenmethode relativiert. Konvergiert nämlich hier $f(x_t) \to 0$ monoton (mit nicht alternierenden Vorzeichen), so tritt im Sekantenschritt

$$x_{t+1} = x_t - f(x_t)\,\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}. \tag{5.3.27}$$

Auslöschung auf. Zur Stabilisierung der Methode kombiniert man sie mit der Intervallschachtelungsidee zur sog. "regula falsi".

Definition 5.3: *Werden im Sekanten-Verfahren die Intervallendpunkte $a_t < b_t$ so gewählt, daß $f(a_t)f(b_t) < 0$ ist, d. h. daß f eine Nullstelle $z \in (a_t, b_t)$ hat, so spricht man von der "Regula falsi".*

Beim Sekantenschritt unter berücksichtigung der Regula falsi,

$$x_t \;:=\; a_t - f(a_t)\,\frac{a_t - b_t}{f(a_t) - f(b_t)}, \tag{5.3.28}$$

tritt dann keine Auslöschung im Term $f(a_t) - f(b_t)$ auf, solange $b_t - a_t \gg \text{eps}$. Offenbar ist $a_t \leq x_t \leq b_t$. Das neue Intervall $[a_{t+1}, b_{t+1}]$ wird bestimmt durch die Vorschrift: $(f(x_t) = 0 \Rightarrow \text{STOP})$

$$\begin{aligned} f(x_t)f(a_t) > 0 \;&\Longrightarrow\; a_{t+1} = x_t\,, \quad b_{t+1} = b_t\,, \\ f(x_t)f(a_t) < 0 \;&\Longrightarrow\; a_{t+1} = a_t\,, \quad b_{t+1} = x_t. \end{aligned} \tag{5.3.29}$$

Die regula falsi ist offensichtlich numerisch stabiler als die ihr zugrunde liegende Sekantenmethode, doch konvergiert sie i.allg. linear. In Extremfällen ist sie sogar langsamer (größere Konvergenzrate) als das einfache Intervallschachtelungsverfahren.

5.4 Methode der sukzessiven Approximation im $\mathbb{R}^n$

Im folgenden betrachten wir iterative Verfahren zur Lösung nichtlinearer Gleichungssysteme im $\mathbb{R}^n$

$$f_i(x_1, \ldots, x_n) = 0\,, \quad i = 1, \ldots, n, \tag{5.4.30}$$

bzw. $f(x) = 0$ mit $f = (f_1, \ldots, f_n)^T$ und $x = (x_1, \ldots, x_n)^T$. Zur Berechnung einer Lösung z verwendet die sog. "Methode der sukzessiven Approximation" die Iteration (in Anlehnung an das vereinfachte Newton-Verfahren in einer Dimension und unter Hochstellung des Iterationsindexes bei vektor- oder matrix-wertigen Größen)

$$x^{t+1} = x^t + C^{-1} f(x^t)\,, \quad t = 0, 1, 2, \ldots \tag{5.4.31}$$

mit einer geeigneten regulären Matrix $C \in \mathbb{R}^{n\times n}$. Konvergiert dann $x^t \to z\,(t \to \infty)$, so ist für stetiges f im Limes $z = z + C^{-1} f(z)$ bzw. $f(z) = 0$. Die Lösung z des Gleichungssystems ist also ein sog. "Fixpunkt" der Abbildung $g(x) := x + C^{-1} f(x)$.

Wir wollen nun die Konvergenz von Fixpunktiterationen der Form

$$x^{t+1} = g(x^t)\,, \quad t = 0, 1, 2, \ldots, \tag{5.4.32}$$

untersuchen. Dazu sei im folgenden $\|\cdot\|$ eine beliebige Vektornorm auf $\mathbb{R}^n$, $\|\cdot\|$ die zugehörige natürliche Matrizennorm

Definition 5.4: *Sei $G \subset \mathbb{R}^n$ eine (nichtleere) abgeschlossene Menge. Eine Abbildung $g : G \to \mathbb{R}^n$ heißt "Lipschitz[3]-stetig" (kurz "L-stetig"), wenn mit einem $q > 0$ gilt:*

$$\|g(x) - g(y)\| \le q\,\|x - y\|, \quad x, y \in G. \tag{5.4.33}$$

Ist die sog. "Lipschitz-Konstante" $q < 1$, so nennt man g eine "Kontraktion" auf G.

Beispiel 5.4: a) Die Funktion $f(x) = |x|$ ist L-stetig auf ganz $\mathbb{R}$, wegen

$$|\,|x| - |y|\,| \le |x - y|.$$

b) Die Funktion $f(x) = \sqrt{|x|}$ ist nicht L-stetig bei $x = 0$:

$$|\,|x|^{1/2} - |0|^{1/2}\,| = |x|^{1/2} \ge |x|^{-1/2}|x - 0|.$$

Der folgende fundamentale "Banachsche[4]Fixpunktsatz" sichert die Existenz von Fixpunkten von Kontraktionen.

[3]Rudolf O.S. Lipschitz (1832-1903): Deutscher Mathematiker aus Königsberg; seit 1864 Prof. in Bonn; arbeitete auf verschiedenen Gebieten der Mathematik.

[4]Stefan Banach (1892-1945): Polnischer Mathematiker; Prof. in Lvov; begründete die Funktionalanalysis.

Satz 5.4 (Sukzessive Approximation): *Sei $G \subset \mathbb{R}^n$ eine nichtleere, abgeschlossene Punktmenge und $g : G \to G$ eine Kontraktion. Dann existiert genau ein Fixpunkt $z \in G$ von g, und für jeden Startpunkt $x^0 \in G$ konvergiert die Folge der durch (5.4.32) erzeugten sukzessiven Approximationen $x^t \to z$ $(t \to \infty)$. Es gelten die a posteriori und a priori Fehlerabschätzungen*

$$\|x^t - z\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\| \le \frac{q^t}{1-q}\,\|x^1 - x^0\|. \tag{5.4.34}$$

Beweis: Da g die Menge G in sich abbildet, sind für $x^0 \in G$ die Iterierten $x^t = g(x^{t-1}) = \ldots = g^t(x^0)$ definiert, und es gilt

$$\begin{aligned}\|x^{t+1} - x^t\| &= \|g(x^t) - g(x^{t-1})\| \\ &\le q\,\|x^t - x^{t-1}\| \le \ldots \le q^t\,\|x^1 - x^0\|.\end{aligned}$$

Wir wollen zeigen, daSS $(x^t)_{t\in\mathbb{N}}$ eine Cauchy-Folge ist. Seien dazu $\varepsilon > 0$ und $m \ge 1$ beliebig vorgegeben:

$$\begin{aligned}\|x^{t+m} - x^t\| &\le \|x^{t+m} - x^{t+m-1}\| + \ldots + \|x^{t+1} - x^t\| \\ &\le \underbrace{\{q^{t+m-1} + \ldots + q^t\}}\,\|x^1 - x^0\| \\ &= q^t \sum_{i=0}^{m-1} q^i = q^t\,\frac{1-q^m}{1-q} \le \varepsilon \quad \text{für} t \ge t(\varepsilon).\end{aligned}$$

Also existiert $z = \lim_{t\to\infty} x^t \in G$ (wegen der Abgeschlossenheit von G) mit $z = g(z)$. Die Eindeutigkeit des Fixpunktes z folgt sofort aus der Kontraktionseigenschaft von g. Zum Nachweis von (5.4.34) schreiben wir

$$\begin{aligned}\|x^{t+m} - x^t\| &\le \|x^{t+m} - x^{t+m-1}\| + \ldots + \|x^{t+1} - x^t\| \\ &\le \underbrace{\{q^m + \ldots + q\}}\,\|x^t - x^{t-1}\|\,, \quad m \ge 1. \\ &\qquad \le q/(1-q)\end{aligned}$$

Durch Grenzübergang $m \to \infty$ folgt daraus

$$\|z - x^t\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\| \le \frac{q^t}{1-q}\,\|x^1 - x^0\|.$$

Q.E.D.

Zur Anwendung des Banachschen Fixpunktsatzes auf eine Abbildung $g : G \to \mathbb{R}^n$ muß gezeigt werden, daß es eine abgeschlossene (nichtleere) Teilmenge von G gibt, die von g in sich abgebildet wird, und auf der g eine Kontraktion ist. Sei g eine Kontraktion auf der Kugel

$$K_\rho(c) \equiv \{\, x \in \mathbb{R}^n \mid \|x - c\| \le \rho \,\}\,, \quad \rho > 0,$$

um einen Punkt $c \in \mathbb{R}^n$ mit Lipschitz-Konstante $q < 1$. Für $x \in K_\rho(c)$ gilt dann

$$\|g(x) - c\| \leq \underbrace{\|g(x) - g(c)\|}_{\leq q\rho} + \|g(c) - c\|.$$

Unter der Bedingung

$$\|g(c) - c\| \leq (1 - q)\rho \tag{5.4.35}$$

bildet dann g die Menge $K_\rho(c)$ in sich ab. Ist g differenzierbar, so wird die Matrix

$$g'(x) \equiv \Big(\frac{\partial g_i}{\partial x_j}\Big)_{i,j=1,\ldots,n} \in \mathbb{R}^{n\times n}$$

der partiellen Ableitungen die "Jacobi[5]-Matrix" genannt.

Hilfssatz 5.2 (L-Stetigkeit): *Die Abbildung $g : G \to \mathbb{R}^n$ sei stetig differenzierbar, und die Menge G sei konvex. Dann gilt*

$$\|g(x) - g(y)\| \leq \sup_{\zeta\in G} \|g'(\zeta)\|\, \|x - y\|, \quad x, y \in G, \tag{5.4.36}$$

d. h.: Im Falle $\sup_{\zeta\in G} \|g'(\zeta)\| < 1$ ist g eine Kontraktion auf G.

Beweis: Seien $x, y \in G$. Wir setzen für $i = 1, \ldots, n$:

$$\varphi_i(s) := g_i(x + s(y - x)), \quad 0 \leq s \leq 1,$$

und haben damit

$$g_i(y) - g_i(x) = \varphi_i(1) - \varphi_i(0) = \int_0^1 \varphi_i'(s)\, ds.$$

Wegen

$$\varphi_i'(s) = \sum_{j=1}^n \frac{\partial g_i}{\partial x_j}(x + s(y - x))(y - x)_j$$

und den Stetigkeitseigenschaften der Vektornorm folgt

$$\begin{aligned}\|g(y) - g(x)\| &= \Big\| \int_0^1 g'(x + s(y - x)) \cdot (y - x)\, ds\Big\| \\ &\leq \int_0^1 \|g'(x + s(y - x))\|\, ds\, \|y - x\| \leq \sup_{\zeta\in G} \|g'(\zeta)\|\, \|y - x\|.\end{aligned}$$

Dies impliziert die Bahauptung. Q.E.D.

[5]Carl Gustav Jakob Jacobi (1804-1851): Deutscher Mathematiker; schon als Kind hochbegabt; wirkte in Königsberg und Berlin; Beiträge zu vielen Bereichen der Mathematik: Zahlentheorie, elliptische Funktionen, partielle Differentialgleichungen, theoretische Mechanik.

Korollar 5.1: *Mit Hilfe der Abschätzung aus Hilfssatz 5.2 und (5.4.35) ergibt sich, daß es zu jedem Fixpunkt $z \in G$ von g, in dem $\|g'(z)\| < 1$ gilt, eine Umgebung*

$$K_\rho(z) = \{x \in \mathbb{R}^n | \, \|x - z\| \le \rho\} \subset G$$

gibt, so daß g eine Kontraktion von $K_\rho(z)$ in sich ist.

Wir betrachten nun wieder die Lösung der Gleichung $f(x) = 0$ mit Hilfe der sukzessiven Approximation

$$x^{t+1} = x^t + C^{-1} f(x^t)\,, \quad t = 0, 1, 2, \ldots . \tag{5.4.37}$$

Nach den obigen Überlegungen ist die Konvergenz dieser Iteration z.B. gesichert, wenn f auf einer geeigneten Kugel $K_\rho(c) \subset \mathbb{R}^n$ stetig differenzierbar ist, und wenn dort gilt

$$\sup_{\zeta \in K_\rho(c)} \|I + C^{-1} f'(\zeta)\| =: q < 1\,, \quad \|C^{-1} f(c)\| \le (1 - q)\rho\,. \tag{5.4.38}$$

Beispiel 5.5: Seien $A \in \mathbb{R}^{n \times n}$ und $b \in \mathbb{R}^n$ gegeben. Das lineare Gleichungssystem $Ax = b$ ist äquivalent zur Nullstellenaufgabe $f(x) := b - Ax = 0$. Zu deren iterativen Lösung betrachten wir mit einer regulären Matrix $C \in \mathbb{R}^{n \times n}$ die Fixpunktaufgabe

$$x = g(x) := x + C^{-1} f(x) = x + C^{-1}(b - Ax) = \underbrace{(I - C^{-1}A)}_{= B}\, x + \underbrace{C^{-1} b}_{= c}\,.$$

Die Matrix $B := I - C^{-1}A$ wird die "Iterationsmatrix" der zugehörigen Fixpunktiteration ("sukzessive Approximation") genannt:

$$x^{t+1} = Bx^T + c, \quad t = 1, 2, \ldots .$$

Die Abbildung g ist wegen

$$\|g(x) - g(y)\| = \|B(x - y)\| \le \|B\| \, \|x - y\|$$

für $\|B\| < 1$ eine Kontraktion auf ganz $\mathbb{R}^n$. Dabei ist $\|\cdot\|$ eine geeignete (natürliche) Matrizennorm. Nach dem Banachschen Fixpunktsatz konvergiert daher die sukzessive Approximation gegen den (eindeutig bestimmten) Fixpunkt der Abbildung g bzw. die Lösung des Gleichungssystems $Ax = b$.

Beispiel 5.6: Die Funktion $f(x) = \cosh(x) - 2x = \frac{1}{2}(e^x + e^{-x}) - 2x$ hat genau zwei Nullstellen $z_1 \sim 0.59$, $z_2 \sim 2.1$. Zu ihrer Approximation machen wir den Ansatz

$$g(x) = x + \tfrac{1}{2}\{\cosh(x) - 2x\} = \tfrac{1}{2}\cosh(x), \quad g'(x) = \tfrac{1}{2}\sinh(x),$$

Offensichtlich bildet g das Intervall $[0, z_2]$ in sich ab. Da die Beziehung

$$\max_{0 \le x \le b} |g'(x)| = \tfrac{1}{2}\sinh(b) < 1 \quad (\operatorname{arcsinh}(2) = 1.44\ldots)$$

notwendig $b < 2$ voraussetzt, muß $|g'(z_2)| > 1$ sein. Für alle Startwerte $x^0 \in (z_2, \infty)$ divergieren die Iterierten $x^t \to \infty$ für $t \to \infty$. Tatsächlich konvergiert aber $x^t \to z_1\,(t \to \infty)$ sogar für alle $x^0 \in [0, z_2]$ (geometrische Überlegung). Die Bedingung, daß g überall eine Kontraktion sein muß, ist nicht notwendig für die Konvergenz der sukzessiven Approximation. Der Fixpunkt z_1 mit $|g'(z_1)| < 1$ in Beispiel 5.4 ist also "anziehend", der Fixpunkt z_2 dagegen "abstoßend", da hier wegen $|g'(z_2)| > 1$ in jeder Umgebung von z_2 Startpunkte x^0 existieren, für die die sukzessive Approximation nicht gegen z_2 konvergiert.

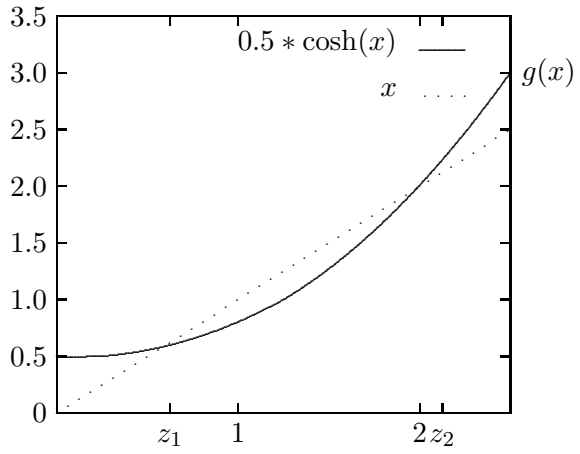


Abbildung 5.4: *Nullstellenproblem*

Nach Wahl einer Fehlertoleranz $\varepsilon >$ eps (z.B.: $\varepsilon = 10^{-4}$) wird ausgehend von $x^0 = 0$ iteriert gemäß $x^{t+1} = \frac{1}{2}\cosh(x^t)(t = 0, 1, 2, \ldots)$ bis

$$\left|\frac{x^{t+1} - x^t}{x^{t+1}}\right| \leq \varepsilon.$$

$$x^1 = 0.\underline{5},\ x^2 = 0.\underline{5}63,\ \ldots,\ x^7 = 0.\underline{589}31,\ \ x^8 = 0.\underline{58936},\ \ldots,\ x^{19} = 0.\underline{5893877633}.$$

$$\left|\frac{x^8 - x^7}{x^8}\right| \leq 0.8532 \cdot 10^{-4}.$$

Auf dem Intervall $[0, 1]$ gilt

$$q = \max_{0 \leq x \leq 1} |g'(x)| = \tfrac{1}{2}\sinh(1) \sim 0.6.$$

Die a priori Fehlerabschätzungen von Satz 5.2 ergibt dann

$$|x^8 - z_1| \leq \frac{0.6^8}{1 - 0.6}\,|x^1 - x^0| \sim 2 \cdot 10^{-2}.$$

Beispiel 5.7: Zur Bestimmung der Quadratwurzel $A^{1/2} \in \mathbb{R}^{n\times n}$ einer positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ betrachtet man die Abbildung

$$g(X) = \tfrac{1}{2}(X^2 + B)$$

mit der Matrix $B = I - A$. Dies wird motiviert durch die Äquivalenz

$$Z = g(Z) = \tfrac{1}{2}(Z^2 + B) \quad \Leftrightarrow \quad (I - Z)^2 = I - B = A,$$

bzw. $I - Z = A^{1/2}$. Im Falle $\|B\| = q < 1$ gilt für $X, Y \in K_q(0) = \{C \in \mathbb{R}^{n\times n} \mid \|C\| \le q\}$

$$\|g(X)\| \le \frac{1}{2}(\|X\|^2 + \|B\|) \le \frac{1}{2}(q^2 + q) \le q$$

und

$$\begin{aligned}\|g(X) - g(Y)\| &= \frac{1}{2}\|X^2 - Y^2\| = \frac{1}{2}\|X(X - Y) + (X - Y)Y\| \\ &\le \frac{1}{2}(\|X\| + \|Y\|)\,\|X - Y\| \le q\|X - Y\|,\end{aligned}$$

d. h.: g ist eine Kontraktion der abgeschlossenen Teilmenge $K_q(0) \subset \mathbb{R}^{n\times n}$ in sich. Nach dem Banachschen Fixpunktsatz existiert also genau ein Fixpunkt $Z \in K_q(0)$ von g, und die Folge der sukzessiven Iterierten $X^t = g(X^{t-1})$, $t \in \mathbb{N}$, konvergiert für jeden Startwert $X^0 \in K_q(0)$: $X^t \to Z$ $(t \to \infty)$. Wegen der obigen Äquivalenz ist dann $I - Z = A^{1/2}$. Alle Iterierten X^t und damit auch der Fixpunkt Z sind symmetrisch. Wegen $\|Z|| \le q$ ist daher $I - Z$ auch positiv definit, so daß mit $A^{1/2} := I - Z$ die eindeutug bestimmte, "positive" Wurzel von A bestimmt ist.

5.5 Das Newton-Verfahren im $\mathbb{R}^n$

Wir betrachten nun das Newton-Verfahren zur Lösung nichtlinearer Gleichungssysteme mit stetig differenzierbaren Abbildungen $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$. Formal lautet die Newton-Iteration

$$x^{t+1} = x^t - f'(x^t)^{-1} f(x^t), \quad t = 0, 1, 2 \ldots, \tag{5.5.39}$$

mit der Jacobi-Matrix $f'(\cdot)$ von f. In jedem Iterationsschritt ergibt sich ein lineares $(n \times n)$-Gleichungssystem mit $f'(x^t)$ als Koeffizientenmatrix:

$$f'(x^t)x^{t+1} = f'(x^t)x^t - f(x^t), \quad t = 0, 1, 2 \ldots . \tag{5.5.40}$$

Dies macht das Newton-Verfahren wesentlich aufwendiger als die einfache Fixpunktiteration; dafür konvergiert es aber auch sehr viel schneller. Das Newton-Verfahren wird meist in Form einer Defektkorrekturiteration durchgeführt (mit dem "Defekt" $d^t := -f(x^t)$):

$$f'(x^t)\delta x^t = -f(x^t), \quad x^{t+1} = x^t + \delta x^t, \quad t = 0, 1, 2 \ldots . \tag{5.5.41}$$

Dies spart gegenüber (5.5.40) pro Iterationsschritt eine Matrix-Vektor-Multiplikation.

Im Folgenden geben wir ein Konvergenzresultat für das Newton-Verfahren, welches nebenbei auch die Existenz einer Nullstelle sichert. Mit $\|\cdot\|$ seien die euklidische Vektornorm und ebenso die zugehörige natürliche Matrizennorm bezeichnet. Sei $f : G \subset \mathbb{R}^n \to \mathbb{R}^n$ eine differenzierbare Abbildung, für die eine Nullstelle z gesucht ist. Die Jacobi-Matrix $f'(\cdot)$ sei auf der Niveaumenge

$$D_* := \{x \in G | \, \|f(x)\| \leq \|f(x^*)\|\}$$

zu einem festen Punkt $x^* \in G$ regulär mit gleichmäßig beschränkter Inverser:

$$\|f'(x)^{-1}\| \leq \beta, \quad x \in D_*.$$

Ferner sei $f'(\cdot)$ auf D_* gleichmäßig L-stetig:

$$\|f'(x) - f'(y)\| \leq \gamma \|x - y\|, \quad x, y \in D_*.$$

Mit diesen Bezeichnungen haben wir den folgenden Satz von Newton-Kantorovich[6]

Satz 5.5 (Newton-Kantorovich): *Unter den vorausgehenden Voraussetzungen gelte für den Startpunkt $x^0 \in D_*$ mit $\alpha := \|f'(x^0)^{-1} f(x^0)\|$ die Bedingung*

$$q := \tfrac{1}{2}\alpha\beta\gamma < 1.$$

[6]Leonid Vitalyevich Kantorovich (1912-1986): Russischer Mathematiker; Professor an der Universität Leningrad (1934-1960), an der Akademie der Wissenschaften (1961-1971) und an der Universität Moskau (1971-1976); fundamentale Beiträge zur Anwendung der linearen Optimierung in der Ökonomie, zur Funktionalanalysis und Numerik.

Dann erzeugt die Newton-Iteration

$$f'(x^t)x^{t+1} = f'(x^t)x^t - f(x^t), \quad t \geq 1,$$

eine Folge $(x^t)_{t\in\mathbb{N}} \subset D$*, welche quadratisch gegen eine Nullstelle* $z \in D$ *von* f *konvergiert, mit der a priori Fehlerabschätzung*

$$\|x^t - z\| \leq \frac{\alpha}{1-q}\, q^{(2^t-1)}, \quad t \geq 1. \tag{5.5.42}$$

Beweis: Zum Startpunkt $x^0 \in D_*$ gehört die abgeschlossene, nicht leere Niveaumenge

$$D_0 := \{x \in G |\, \|f(x)\| \leq \|f(x^0)\|\} \subset D_*.$$

Wir betrachten die stetige Abbildung $g : D_0 \to \mathbb{R}^d$: $g(x) := x - f'(x)^{-1}f(x)$.

(i) Wir wollen zunächst einige Hilfsresultate ableiten. Für $x \in D_0$ sei

$$x_r := x - rf'(x)^{-1}f(x), \;\; 0 \leq r \leq 1,$$

und $R := \max\{r |\, x_s \in D_0, 0 \leq s \leq r\} = \max\{r |\, \|f(x_s)\| \leq \|f(x^0)\|,\; 0 \leq s \leq r\}$. Für die Vektorfunktion $h(r) := f(x_r)$ gilt

$$h'(r) = -f'(x_r)f'(x)^{-1}f(x), \qquad h'(0) = -h(0).$$

Für $0 \leq r \leq R$ ergibt dies

$$\begin{aligned}\|f(x_r)\| - (1-r)\|f(x)\| \leq \|f(x_r) - (1-r)f(x)\| &= \|h(r) - (1-r)h(0)\| \\ &= \Big\| \int_0^r h'(s)\, ds + rh(0) \Big\| = \Big\| \int_0^r \{h'(s) - h'(0)\}\, ds \Big\| \\ &\leq \int_0^r \|h'(s) - h'(0)\|\, ds,\end{aligned}$$

und ferner wegen $x_s - x = -sf'(x)^{-1}f(x)$:

$$\begin{aligned}\|h'(s) - h'(0)\| &= \|\{f'(x_s) - f'(x)\}f'(x)^{-1}f(x)\| \\ &\leq \gamma\|x_s - x\|\|f'(x)^{-1}f(x)\| \leq \gamma s\|f'(x)^{-1}f(x)\|^2.\end{aligned}$$

Dies ergibt

$$\|f(x_r)\| - (1-r)\|f(x)\| \;\leq\; \tfrac{1}{2}r^2\gamma\|f'(x)^{-1}f(x)\|^2. \tag{5.5.43}$$

Mit der Größe $\alpha_x := \|f'(x)^{-1}f(x)\|$ und $\|f'(x)^{-1}\| \leq \beta$ folgt

$$\|f(x_r)\| \leq (1 - r + \tfrac{1}{2}r^2\alpha_x\beta\gamma)\|f(x)\|.$$

Im Falle $\alpha_x \leq \alpha$ gilt dann wegen der Voraussetzung $\frac{1}{2}\alpha\beta\gamma < 1$:

$$\|f(x_r)\| \leq (1 - r + r^2)\|f(x)\|.$$

Folglich ist in diesem Fall $R = 1$, d.h.: $g(x) \in D_0$. Für solche $x \in D_0$ gilt weiter

$$\|g(x) - g^2(x)\| = \|g(x) - g(x) + f'(g(x))^{-1} f(g(x))\| \le \beta \|f(g(x))\|.$$

Mit Hilfe der Abschätzung (5.5.43) für $r = 1$ folgt bei Beachtung von $g(x) = x_1$:

$$\|g(x) - g^2(x)\| \le \tfrac{1}{2}\beta\gamma \|f'(x)^{-1} f(x)\|^2 = \tfrac{1}{2}\beta\gamma \|x - g(x)\|^2. \tag{5.5.44}$$

(ii) Nach diesen Vorbereitungen kommen wir nun zum Beweis des Satzes. Zunächst wollen wir zeigen, daß die Newton-Iterierten $(x^t)_{t\in\mathbb{N}}$ in D_0 existieren und die Ungleichung

$$\|x^t - g(x^t)\| = \|f'(x^t)^{-1} f(x^t)\| \le \alpha$$

erfüllen. Dies erfolgt durch vollständige Induktion. Für $t = 0$ ist die Aussage trivialerweise richtig; insbesondere ist wegen $\alpha_{x^0} = \alpha$ nach dem oben gezeigten $g(x^0) \in D_0$. Sei nun $x^t \in D_0$ eine Iterierte mit $g(x^t) \in D_0$ und $\|x^t - g(x^t)\| \le \alpha$. Dann folgt

$$\|x^{t+1} - g(x^{t+1})\| = \|g(x^t) - g^2(x^t)\| \le \tfrac{1}{2}\beta\gamma \|x^t - g(x^t)\|^2 \le \tfrac{1}{2}\alpha^2\beta\gamma \le \alpha$$

und somit nach dem oben Gezeigten $g(x^{t+1}) \in D_0$. Also existiert $(x^t)_{t\in\mathbb{N}} \subset D_0$. Als nächstes zeigen wir, daß diese Folge Cauchy-Folge ist. Mit Hilfe von (5.5.44) ergibt sich

$$\|x^{t+1} - x^t\| = \|g^2(x^{t-1}) - g(x^{t-1})\| \le \tfrac{1}{2}\beta\gamma \|g(x^{t-1}) - x^{t-1}\|^2 = \tfrac{1}{2}\beta\gamma \|x^t - x^{t-1}\|^2,$$

und bei Iteration dieser Abschätzung:

$$\begin{aligned}\|x^{t+1} - x^t\| &\le \tfrac{1}{2}\beta\gamma\big(\tfrac{1}{2}\beta\gamma \|x^{t-1} - x^{t-2}\|^2\big)^2 \le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)} \|x^{t-1} - x^{t-2}\|^{(2^2)} \\ &\le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)}\big(\tfrac{1}{2}\beta\gamma \|x^{t-2} - x^{t-3}\|^2\big)^{(2^2)} = (\tfrac{1}{2}\beta\gamma)^{(2^3-1)} \|x^{t-2} - x^{t-3}\|^{(2^3)}.\end{aligned}$$

Fortsetzung der Iteration bis $t = 0$ ergibt mit $q = \frac{1}{2}\alpha\beta\gamma$:

$$\|x^{t+1} - x^t\| \le (\tfrac{1}{2}\beta\gamma)^{(2^t-1)} \|x^1 - x^0\|^{(2^t)} \le (\tfrac{1}{2}\beta\gamma)^{(2^t-1)} \alpha^{(2^t)} \le \alpha q^{(2^t-1)}.$$

Für beliebiges $m \in \mathbb{N}$ folgt damit wegen $q < 1$:

$$\begin{aligned}\|x^{t+m} - x^t\| &\le \|x^{t+m} - x^{t+m-1}\| + \cdots + \|x^{t+2} - x^{t+1}\| + \|x^{t+1} - x^t\| \\ &\le \alpha q^{(2^{t+m-1}-1)} + \cdots + \alpha q^{(2^{t+1}-1)} + \alpha q^{(2^t-1)} \\ &\le \alpha q^{(2^t-1)}\big\{(q^{(2^t)})^{(2^{m-1}-1)} + \cdots + q^{(2^t)} + 1\big\} \\ &\le \alpha q^{(2^t-1)} \sum_{j=0}^{\infty} (q^{(2^t)})^j \;\le\; \frac{\alpha q^{(2^t-1)}}{1 - q^{(2^t)}}.\end{aligned}$$

Dies besagt, daß $(x^t)_{t\in\mathbb{N}} \subset D_0$ Cauchy-Folge ist. Deren Limes $z \in D_0$ ist dann notwendig ein Fixpunkt von g bzw. Nullstelle von f:

$$z = \lim_{t\to\infty} x^t = \lim_{t\to\infty} g(x^{t-1}) = g(z).$$

Durch Grenzübergang $m \to \infty$ erhalten wir auch die Fehlerabschätzung

$$\|z - x^t\| \leq \frac{\alpha}{1-q} q^{(2^t-1)},$$

was den Beweis vervollständigt. Q.E.D.

Bemerkung 5.3: Unter der Annahme, daß eine Nullstelle $z \in G$ von f existiert kann die Aussage von Satz 5.1 für das Newton-Verfahren im $\mathbb{R}^1$ sinngemäß auf den $\mathbb{R}^n$ mit der Maximumnorm $\|\cdot\|_\infty$ verallgemeinert werden. Dabei sind die auftretenden Konstanten sind gemäß $m = 1/\beta, M = \gamma$ zu identifizieren. Inbesondere gilt neben der a priori Fehlerabschätzung (5.5.42) auch die folgende a posteriori Fehlerabschätzung (Übungsaufgabe):

$$\|x^t - z\|_\infty \leq \frac{1}{m} \|f(x^t)\|_\infty \leq \frac{M}{2m} \|x^t - x^{t-1}\|_\infty^2, \quad t \in \mathbb{N}. \tag{5.5.45}$$

Beispiel 5.8: Zur Bestimmung der Inversen $Z = A^{-1}$ einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ wird gesetzt

$$f(X) := X^{-1} - A,$$

für $X \in \mathbb{R}^{n\times n}$ regulär. Eine Nullstelle dieser Abbildung $f(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ ist gerade die Inverse $Z = A^{-1}$. Diese soll mit dem Newton-Verfahren berechnet werden. Dazu ist zunächst eine Umgebung von A bzw. von A^{-1} zu bestimmen, auf der $f(\cdot)$ definiert und differenzierbar ist. Für $X \in K_\rho(A)$ mit $\rho < \|A^{-1}\|^{-1}$ folgt aus $X = A - A + X = A(I - A^{-1}(A - X))$ die Beziehung

$$\|A^{-1}(A - X)\| \leq \|A^{-1}\| \, \|A - X\| \leq \rho \|A^{-1}\| < 1,$$

d. h.: $I - A^{-1}(A - X)$ und damit auch X sind regulär. Als nächstes ist die Jacobi-Matrix $f'(\cdot)$ von $f(\cdot)$ als Abbildung von $\mathbb{R}^{n\times n}$ in sich zu bestimmen. Für die Durchführung des Newton-Verfahrens genügt es offensichtlich, die Wirkung von $f'(\cdot)$ auf Matrizen $Y \in \mathbb{R}^{n\times n}$ zu bestimmen. Wir wollen zeigen , daß

$$f'(X)Y = -X^{-1}YX^{-1}, \quad Y \in \mathbb{R}^{n\times n}.$$

Dies sieht man wie folgt: Aus $f(X) = X^{-1} - A$ folgt $Xf(X) = I - XA$. Für die Jacobi-Matrizen der rechten und linken Seite gilt

$$\begin{aligned}
\big([Xf(X)]'Y\big)_{j,k} &= \sum_{pq} \frac{\partial}{\partial x_{pq}} \sum_l x_{jl} f_{lk}(X)\, y_{pq} \\
&= \sum_{p,q} \sum_l \Big\{ \underbrace{\frac{\partial x_{jl}}{\partial x_{pq}}}_{\delta_{jp}\cdot\delta_{lq}} f_{lk}(X) + x_{jl} \frac{\partial f_{lk}}{\partial x_{pq}}(X) \Big\} y_{pq} \\
&= \sum_q f_{qk}(X) y_{jq} + \sum_{p,q} \sum_l x_{jl} \frac{\partial f_{lk}}{\partial x_{pq}}(X) y_{pq} \\
&= \big(Yf(X) + Xf'(X)Y \big)_{jk}.
\end{aligned}$$

Analog finden wir

$$[I - XA]'Y = -YA\,.$$

Also ist

$$-YA = Yf(X) + Xf'(X)Y = YX^{-1} - YA - Xf'(X)Y$$

bzw.

$$f'(X)Y = -X^{-1}YX^{-1}\,.$$

Das Newton-Verfahren

$$f'(X^t)X^{t+1} = f'(X^t)X^t - f(X^t)$$

erhält in diesem Fall also die Gestalt

$$-X^{t^{-1}}X^{t+1}X^{t^{-1}} = -X^{t^{-1}}\underbrace{X^tX^{t^{-1}}}_{=I} - X^{t^{-1}} + A$$

bzw.

$$X^{t+1} = 2X^t - X^tAX^t = X^t\{2I - AX^t\}. \tag{5.5.46}$$

Diese Iteration ist das mehrdimensionale Analogen der Iteration $x_{t+1} = x_t(2 - ax_t)$ im skalaren Fall zur divisionsfreien Berechnung des Kehrwertes $1/a$ einer Zahl $a \neq 0$. Über die Identität

$$X^{t+1} - Z = 2X^t - X^tAX^t - Z = -(X^t - Z)A(X^t - Z) \tag{5.5.47}$$

gewinnt man die Fehlerabschätzung

$$\|X^{t+1} - Z\| \leq \|A\|\,\|X^t - Z\|^2. \tag{5.5.48}$$

Der Einzugsbereich der quadratischen Konvergenz für das Newton-Verfahren ist in diesem Fall also die Menge

$$\{X \in \mathbb{R}^{n\times n} |\, \|X - Z\| < \|A\|^{-1}\}.$$

5.5.1 Gedämpftes Newton-Verfahren

Bei der Durchführung des Newton-Verfahrens zur Lösung nichtlinearer Gleichungssysteme treten zwei Hauptschwierigkeiten auf:

(i) hoher Aufwand pro Iterationsschritt,

(ii) "guter" Startpunkt x^0 erforderlich.

Zur Überwindung dieser Probleme verwendet man gegebenenfalls das sog. "vereinfachte Newton-Verfahren"

$$f'(c)\delta x^t = -f(x^t), \quad x^{t+1} = x^t + \delta x^t, \tag{5.5.49}$$

mit einem geeigneten $c \in \mathbb{R}^n$, etwa $c = x^{(0)}$, welches nahe bei der Nullstelle z liegt. Dabei haben alle zu lösenden Gleichungssysteme dieselbe Koeffizientenmatrix und können mit Hilfe einer einmal berechneten LR-Zerlegung von $f'(c)$ effizient gelöst werden. Andererseits führt man zur Vergrößerung des Konvergenzbereiches des Newton-Verfahrens eine "Dämpfung" ein,

$$f'(x^t)\delta x^t = -f(x^t), \quad x^{t+1} = x^t + \lambda_t \delta x^t, \tag{5.5.50}$$

wobei der Parameter $\lambda_t \in (0, 1]$ zu Beginn klein gewählt wird und dann nach endlich vielen Schritten gemäß einer geeigneten Dämpfungsstrategie $\lambda_t = 1$ gesetzt wird. Der folgende Satz gibt ein konstruktives Kriterium für die a posteriori Wahl des Dämpfungsparameters λ_t.

Satz 5.6 (gedämpftes Newton-Verfahren): *Unter den Voraussetzungen von Satz 5.5 erzeugt für jeden Startpunkt $x^0 \in D_*$ die gedämpfte Newton-Iteration (5.5.50) mit*

$$\lambda_t := \min\Big\{1, \frac{1}{\alpha_t \beta\gamma}\Big\}, \qquad \alpha_t := \|f'(x^t)^{-1} f(x^t)\|,$$

eine Folge $(x^t)_{t\in\mathbb{N}}$, für welche nach t_* Schritten $q_* := \frac{1}{2}\alpha_{t_*}\beta\gamma < 1$ erfüllt ist, so daß ab dann x^t quadratisch konvergiert, mit der a priori Fehlerabschätzung

$$\|x^t - z\| \le \frac{\alpha}{1 - q_*}\, q_*^{(2^t-1)}, \quad t \ge t_*. \tag{5.5.51}$$

Beweis: Wir verwenden wieder die Bezeichnungen aus dem Beweis von Satz 5.5. Für ein $x \in D_0$ gilt mit $x_r := x - r f'(x)^{-1} f(x)$, $0 \le r \le 1$, und $\alpha_x := \|f'(x)^{-1} f(x)\|$ die Abschätzung

$$\|f(x_r)\| \le (1 - r + \tfrac{1}{2} r^2 \alpha_x \beta\gamma)\|f(x)\|, \quad 0 \le r \le R = \max\{r |\, x_s \in D_0, 0 \le s \le r \le 1\}.$$

Der Vorfaktor wird minimal für

$$r_* = \min\Big\{1, \frac{1}{\alpha_x \beta\gamma}\Big\} > 0 : \quad 1 - r_* + \tfrac{1}{2} r_*^2 \alpha_x \beta\gamma \;\le\; 1 - \frac{1}{2\alpha_x\beta\gamma} < 1.$$

Bei Wahl von

$$r_t := \min\Big\{1, \frac{1}{\alpha_t \beta\gamma}\Big\}$$

ist also $(x^t)_{t\in\mathbb{N}} \subset D_0$, und die Norm $\|g(x^t)\|$ fällt streng monoton, d.h.:

$$\|f(x^{t+1})\| \le \Big(1 - \frac{1}{2\alpha_t\beta\gamma}\Big)\|f(x^t)\|.$$

Nach endlich vielen, $t_* \ge 1$, Iterationsschritten ist dann $\frac{1}{2}\alpha_{t_*}\beta\gamma < 1$, und die quadratische Konvergenz der weiteren Folge $(x^t)_{t\ge t_*}$ folgt aus Satz 5.5. Q.E.D.

5.6 Übungsaufgaben

Übung 5.1: Man berechne mit einem Fehler kleiner 10^{-6} die Nullstelle $z = \pi$ der Funktion $f(x) = \sin(x)$:

a) mit der Intervallschachtelung zum Startintervall $[2, 4]$;

b) mit der Fixpunktiteration $x_t = x_{t-1} + f(x_{t-1})$ zum Startwert $x_0 = 4$;

c) mit dem Newton-Verfahren $x_t = x_{t-1} - f'(x_{t-1})^{-1} f(x_{t-1})$ zum Startwert $x_0 = 4$.

Warum konvergiert in diesem Fall die Fixpunktiteration (b) genauso schnell wie das Newton-Verfahren ?

Übung 5.2: Zur Berechnung der Lösung $z \in [0.5, 0.6]$ der Gleichung $x + \ln(x) = 0$ werden folgende Fixpunktiterationen vorgeschlagen:

a) $x_t = -\ln(x_{t-1})$;

b) $x_t = e^{-x_{t-1}}$;

c) $x_t = \frac{1}{2}(x_{t-1} + e^{-x_{t-1}})$.

Welche dieser Iterationen kann man verwenden, welche sollte man verwenden, und läßt sich vielleicht eine noch "bessere" Iteration angeben?

Übung 5.3: Es sei $a > 0$ gegeben. Man zeige, daß für beliebigen Startwert $x_0 > 0$ die Fixpunktiteration

$$x_t = \frac{x_{t-1}^3 + 3a x_{t-1}}{3x_{t-1}^2 + a}, \quad t = 1, 2, \ldots,$$

monoton gegen $z = \sqrt{a}$ konvergiert. Wie groß ist die lokale Konvergenzordnung? Man überprüfe durch einen numerischen Test das theoretische Ergebnis.

Übung 5.4: Für zweimal stetig differenzierbare Funktionen f konvergiert das Newton-Verfahren lokal quadratisch gegen eine Nullstelle z. Man zeige, daß es für (nur) stetig differenzierbare Funktionen immer noch "super-linear" konvergiert,

$$\left| \frac{x_t - z}{x_{t-1} - z} \right| \to 0 \quad (t \to \infty),$$

d.h.: Es ist asymptotisch schneller als die einfache Fixpunktiteration.

Übung 5.5: (Praktische Aufgabe) Man schreibe ein Programm zur Berechnung der Nullstellen eines Polynoms

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n, \quad a_n \neq 0,$$

mit Hilfe des Newton-Verfahrens, wobei zur Auswertung von $p(x)$ und $p'(x)$ das Horner Schema zu verwenden ist. Startwerte sollen etwa durch Intervallschachtelung ermittelt

werden. Als Abbruchkriterium frage man ab, ob gilt:

$$\frac{|x_{t+1} - x_t|}{|x_t|} = \left| \frac{p(x_t)}{p'(x_t)x_t} \right| \leq 10^{-8}.$$

Dieses Kriterium wird auch im Fall einer mehrfachen Nullstelle, d. h. $p'(x_t) \to 0 \ (t \to \infty)$, verwendet. Es muß allerdings abgefragt werden, ob $p'(x_t)x_t \neq 0$ ist.

Man berechne mit diesem Programm sämtliche Nullstellen der Legendre- und der Tschebyscheff-Polynome vom Grad $p = 4$ und $p = 5$, d. h. die Stützstellen der entsprechenden Gaußschen Quadraturformeln. Dabei sollen jeweils alle Iterierte inkl. der Startwertberechnung bis zur Erfüllung des Abbruchkriteriums ausgegeben werden.
(Hinweis: Die Polynome $L_k(x)$ und $T_k(x)$ erhält man mit Hilfe der Rekursionsformeln für die Legendre- und die Tschebyscheff-Polynome.)

Übung 5.6: Zur Berechnung der Inversen A^{-1} einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ werden die beiden Fixpunktiterationen

$$\begin{aligned} &a) \quad X_t = X_{t-1}(I - AC) + C, \quad t = 1, 2, \ldots, \quad C \in \mathbb{R}^{n\times n} \text{ regulär}, \\ &b) \quad X_t = X_{t-1}(2I - AX_{t-1}), \quad t = 1, 2, \ldots, \end{aligned}$$

betrachtet. Man gebe hinreichende Kriterien für die Konvergenz dieser Iterationen an. Wie würde in diesem Fall das Newton-Verfahren lauten?

Übung 5.7: Es sollen die Schnittpunkte des durch $x_1^2 + x_2^2 = 2$ gegebenen Kreises und der durch $x_1^2 - x_2^2 = 1$ gegebenen Hyperbel bestimmt werden. Wie lauten die exakten Lösungen?

a) Man schreibe die Aufgabenstellung als Nullstellenproblemn einer geeigneten Abbildung $f = f(x_1, x_2) : \mathbb{R}^2 \to \mathbb{R}^2$ und iteriere ausgehend von dem Startwert $x^{(0)} = (1, 1)^T$ mit dem Newton-Verfahren bis das Inkrement $\|x^{(t)} - x^{(t-1)}\|_\infty$ kleiner als 2×10^{-3} ist.

b) Man bestimme zu der Abbildung f aus (a) eine Matrix $C \in \mathbb{R}^{2\times 2}$ in der Form

$$C = \begin{bmatrix} c & c \\ c & -c \end{bmatrix}, \quad c \neq 0,$$

so daß die Fixpunktiteration

$$x^{(t+1)} = x^{(t)} - Cf(x^{(t)})$$

ausgehend vom Startwert $x^{(0)} = (1, 1)^T$ garantiert gegen die Nullstelle z von f im ersten Quadranten der (x_1, x_2)-Ebene konvergiert. Wie viele Schritte müßte man mit der gewählten Fixpunktiteration machen, damit $\|x^{(t)} - z\|_\infty$ kleiner als 2×10^{-3} ist? (Hinweis: Bei den Abschätzungen verwende man die Maximumnorm.)

Übung 5.8: Die Eigenwertaufgabe $Ax = \lambda x$ einer Matrix $A \in \mathbb{R}^{n\times n}$ ist äquivalent zu dem nichtlinearen Gleichungssystem

$$\begin{aligned} Ax - \lambda x &= 0, \\ \|x\|_2^2 - 1 &= 0, \end{aligned}$$

von $n+1$ Gleichungen in den $n+1$ Unbekannten $x_1, \ldots, x_n, \lambda$.

a) Man gebe die Newton-Iteration zur Lösung dieses Gleichungssystems an.

b) Man führe zwei (oder bei Interesse auch mehr) Newton-Schritte durch für die Matrix

$$A = \begin{bmatrix} 4 & 0 \\ -1 & 4 \end{bmatrix}$$

mit den Startwerten $x_1^0 = 0$, $x_2^0 = 1.5$, $\lambda^0 = 3.5$. Man berechne die Eigenwerte und Eigenvektoren dieser Matrix und stelle fest, ob das Newton-Verfahren in diesem Fall quadratisch konvergiert.

Übung 5.9: Man untersuche die Konvergenz der Fixpunktiteration

$$x^t = Bx^{t-1} + c$$

für die Matrizen

$$(i) \quad B = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.7 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}, \qquad (ii) \quad B = \begin{bmatrix} 0 & 0.5 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Was ist der Limes der Folgen im Falle der Konvergenz? (Hinweis: Es sind die Eigenwerte der Matrizen abzuschätzen. Dazu kann eine geeignete Norm oder auch der Zusammenhang zwischen den Eigenwerten und der Determinante einer Matrix dienen.)

Übung 5.10: (Praktische Aufgabe) Man schreibe ein Programm zur Realisierung der Iterationsverfahren aus Aufgabe 11.1 für die (positiv definite) Tridiagonalmatrix

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$$

für $n = 2^k$, $k = 2, \ldots, 10$. Zur Vermeidung zu langer Rechenzeiten sollte eine obere Schranke (etwa $k \leq 10^5$ für die Anzahl der Iterationsschritte gesetzt werden. Mit der Matrix $C = \frac{1}{8} I \in \mathbb{R}^{n\times n}$ ist in diesem Fall nach Aufgabe 11.1 die Konvergenz der Iteration

(a) für jeden Startwert garantiert. Man verwende daher versuchsweise für beide Iterationen (a) und (b) die Startmatrix $X_0 = \frac{1}{8}I$. Als Abbruchkriterium wähle man die Größe des Residuums $AX_t - I$ gemäß

$$\|AX_t - I\|_\infty = \max_{i=1,\ldots,n} \Big(\sum_{j=1}^{n} |(AX_t)_{ij} - \delta_{ij}| \Big) \leq 10^{-8}.$$

Man gebe die Anzahl der benötigten Iterationen in Abhängigkeit von n an. Sind diese Verfahren konkurrenzfähig mit der direkten Berechnung der Inversen mit Hilfe des Gaußschen Eliminationsverfahrens?

6 Lineare Gleichungssysteme II (Iterative Verfahren)

Für sehr große Gleichungssysteme mit $n \gg 1.000$ ist das Gaußsche Eliminationsverfahren nur sehr schwer zu realisieren, da es zu viel Speicherplatz erfordert. Für eine $n \times n$-Matrix A mit $n = 10^6$ und Bandbreite $m = 10^2$ sind dies bereits 10^8 Speicherplätze, was die Kernspeicherkapazität der meisten zur Zeit im Einsatz befindlichen Rechenanlagen übersteigt. Zur Durchführung der Elimination müßte man in diesem Fall also mit externen Speichern arbeiten, was wegen des aufwendigen Datentransfers die Rechenzeit stark verlängert. Bei vielen in der Praxis auftretenden großen Gleichungssystemen hat man es jedoch mit sehr dünn besetzten Bandmatrizen mit $5 - 25$ von Null verschiedene Elemente pro Zeile zu tun. Die im folgenden betrachteten "iterativen Verfahren" benötigen zur näherungsweisen Lösung des Gleichungssystems $Ax = b$ nicht viel mehr Speicherplatz, als zur Speicherung von A erforderlich ist.

Als erstes betrachten wir Fixpunktiterationen zur Lösung des Gleichungssystems $Ax = b$ mit einer regulären $n \times n$-Matrix A und einem n-Vektor b. Zur Konstruktion solcher Iterationsvorschriften geht man etwa wie folgt vor:

Das Gleichungssystem $Ax = b$ lautet ausgeschrieben

$$a_{jj}x_j + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k = b_j, \quad j = 1, \ldots, n.$$

Im Falle $a_{jj} \neq 0$ ist dies äquivalent zu

$$x_j = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k \Big\}, \quad j = 1, \ldots, n.$$

Das sog. "Gesamtschritt-Verfahren" (oder auch "Jacobi-Verfahren") erzeugt Iterierte $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$, durch die Iterationsvorschrift

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{6.0.1}$$

Zum Zeitpunkt der Berechnung von x_j^t sind die vorausgehenden neuen Komponenten x_r^t, $r < j$, bereits berechnet. Zur Beschleunigung der Konvergenz liegt es also nahe, diese Zusatzinformation schon zur Berechnung von x_j^t auszunutzen. Dies ist die Grundlage des "Einzelschritt-Schritt-Verfahren (oder auch "Gauß-Seidel[1]-Verfahrens"):

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} a_{jk}x_k^t - \sum_{k>j} a_{jk}x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{6.0.2}$$

[1]Philipp Ludwig von Seidel (1821-1896): deutscher Mathematiker; Prof. in München; Beiträge zur Analysis (u.a. Methode der kleinsten Fehlerquadrate) owie Himmelsmechanik und Astronomie.

6.1 Fixpunktiterationen

Zur kompakteren Schreibweise der betrachteten Iterationsverfahren führen wir die Aufspaltung $A = D + L + R$ ein, wobei

$$D = \begin{bmatrix} a_{11} & & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & & a_{nn} \end{bmatrix} \quad L = \begin{bmatrix} 0 & & \cdots & 0 \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & \cdots & & 0 \end{bmatrix}$$

Damit schreibt sich das Jacobi-Verfahren in der Form

$$x^t = D^{-1}\{b - (L+R)x^{t-1}\} = \underbrace{-D^{-1}(L+R)}_{J} x^{t-1} + D^{-1}b,$$

mit der sog. "Jacobi-Matrix" J und das Gauß- Seidel-Verfahren in der Form

$$x^t = D^{-1}\{b - Lx^t - Rx^{t-1}\} = \underbrace{-(D+L)^{-1}R}_{H_1} x^{t-1} + (D+L)^{-1}b.$$

mit der sog. "Gauß-Seidel-Matrix" H_1 (Die Notation H_1 für die Gauß-Seidel-Matrix wird später klar werden.). Beide Verfahren besitzen also die Gestalt

$$x^t = Bx^{t-1} + c \tag{6.1.3}$$

mit einer sog. "Iterationsmatrix" B. Konvergiert nun die Folge der Iterierten $(x^{(t)})_{t\in\mathbb{N}}$ gegen einen Vektor $x \in \mathbb{R}^n$, so gilt für diesen offenbar

$$x = Bx + c, \tag{6.1.4}$$

d. h. er ist ein "Fixpunkt" der Abbildung $g : x \to Bx + c$; daher auch die Bezeichnung "Fixpunktiteration". Ein sinnvolles iteratives Verfahren dieser Art muß also so gebaut sein, daß die Fixpunkte von g automatisch Lösungen des ursprünglichen Gleichungssystems $Ax = b$ sind. Dies ist beim Jacobi- und beim Gauß-Seidel-Verfahren aufgrund ihrer Konstruktion der Fall. Zur Konstruktion allgemeinerer iterativer Verfahren dieses Typs wählt man etwa eine reguläre $n\times n$-Matrix C und iteriert ausgehend von der Beziehung

$$Ax = b \quad \leftrightarrow \quad Cx = Cx - Ax + b \quad \leftrightarrow \quad x = x + C^{-1}(b - Ax)$$

in der Form

$$x^t = \underbrace{(I - C^{-1}A)}_{=:\,B} x^{t-1} + \underbrace{C^{-1}b}_{=:\,c}. \tag{6.1.5}$$

Dies wird in der Praxis auf dem Rechner als sog. "Defektkorrekturiteration" realisiert, bei der in jedem Schritt im wesentlichen ein lineares Gleichungssystem mit der gewählten Matrix C gelöst werden muß:

$$d^{t-1} = b - Ax^{t-1}\,, \qquad C\delta x^t = d^{t-1}\,, \qquad x^t = x^{t-1} + \delta x^t.$$

Ein hinreichendes Kriterium für die Konvergenz der Iteration (6.1.3) ist nach dem Banachschen Fixpunktsatz, daß

$$\|B\| < 1$$

für irgendeine Matrizennorm $\|\cdot\|$ auf $\mathbb{R}^{n\times n}$. Die Gültigkeit dieser Beziehung kann aber für eine konkrete Matrix sehr wohl von der speziellen Wahl der Norm abhängen. Daher verwendet man zur Charakterisierung der Fixpunktiteration besser den sog. "Spektralradius" der Iterationsmatrix:

$$\mathrm{spr}\,(B) := \max\{\,|\lambda| : \lambda \in \sigma(B)\,\}.$$

Hierbei bezeichnet $\sigma(B) \subset \mathbb{C}$ das "Spektrum" der Matrix B, d. h.: die Menge ihrer Eigenwerte. Offenbar ist $\mathrm{spr}\,(B)$ der Radius der kleinsten Kreisscheibe um den Nullpunkt in der komplexen Zahlenebene, die alle Eigenwerte von B enthält. Mit einer beliebigen natürlichen Matrizennorm gilt

$$\mathrm{spr}\,(B) \le \|B\|. \tag{6.1.6}$$

Für symmetrisches B ist sogar

$$\mathrm{spr}\,(B) = \|B\|_2 = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Bx\|_2}{\|x\|_2}; \tag{6.1.7}$$

jedoch ist $\mathrm{spr}\,(\cdot)$ keine Norm auf $\mathbb{R}^{n\times n}$, da im allgemeinen die Dreiecksungleichung nicht gilt.

Satz 6.1 (Fixpunktiteration): *Die durch*

$$x^t = Bx^{t-1} + c \tag{6.1.8}$$

erzeugten Iterierten $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$, konvergieren genau dann für jeden Startwert $x^0 \in \mathbb{R}^n$ gegen die Lösung $x \in \mathbb{R}^n$ der Fixpunktgleichung $x = Bx + c$, wenn

$$spr(B) < 1. \tag{6.1.9}$$

Im Falle der Konvergenz ist das asymptotische Konvergenzverhalten bzgl. einer beliebigen Vektornorm $\|\cdot\|$ charakterisiert durch

$$\sup_{x^0\in\mathbb{R}^n} \limsup_{t\to\infty} \Big(\frac{\|x^t - x\|}{\|x^0 - x\|}\Big)^{1/t} = \mathrm{spr}(B). \tag{6.1.10}$$

Beweis: Wir führen die Fehlervektoren $e^t := x^t - x$ ein und finden (wegen $x = Bx + c$)

$$e^t = x^t - x = Bx^{t-1} + c - \underbrace{(Bx + c)}_{= x} = Be^{t-1},$$

d. h. $e^t = B^t e^0$, $t \in \mathbb{N}$.

(i) Im Falle $\operatorname{spr}(B) < 1$ existiert gemäß Hilfssatz 6.1 eine natürliche Matrizennorm $\|\cdot\|_\varepsilon$, so daß

$$\|B\|_\varepsilon \leq \operatorname{spr}(B) + \varepsilon < 1$$

für ein $\varepsilon < 1 - \operatorname{spr}(B)$. Folglich konvergiert in der zugehörigen Vektornorm $\|\cdot\|_\varepsilon$ für $t \to \infty$:

$$\|e^t\|_\varepsilon = \|B^t e^0\|_\varepsilon \leq \|B^t\|_\varepsilon \, \|e^0\|_\varepsilon \leq \|B\|_\varepsilon^t \, \|e^0\|_\varepsilon \to 0.$$

Aufgrund der Äquivalenz aller Normen auf $\mathbb{R}^n$ konvergiert also $x^t \to x \, (t \to \infty)$.

(ii) Aus der Konvergenz der Iteration (für jeden Startwert) folgt bei Wahl von $x^0 = w + x$ mit einem Eigenvektor $w \in \mathbb{R}^n \setminus \{0\}$ zum betragsgrößten Eigenwert λ von B:

$$\lambda^t w = B^t w = B^t e^0 = e^t \to 0 \quad (t \to \infty).$$

Dies impliziert notwendig $|\lambda| < 1$ für $\lambda \in \sigma(B)$, d. h. $\operatorname{spr}(B) < 1$. Als Nebenprodukt erhalten wir noch die Beziehung

$$\Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = |\lambda|.$$

(iii) Zu beliebig kleinen $\varepsilon > 0$ sei wieder $\|\cdot\|_\varepsilon$ eine natürliche Matrizennorm mit $\|B\|_\varepsilon \leq \operatorname{spr}(B) + \varepsilon$. Dann existieren Zahlen $m, M > 0$, so daß für die gegebene (beliebige) Vektornorm $\|\cdot\|$ gilt

$$m\|x\| \leq \|x\|_\varepsilon \leq M\|x\|, \quad x \in \mathbb{R}^n,$$

Damit erhalten wir

$$\begin{aligned} \|e^t\| &\leq \frac{1}{m}\|e^t\|_\varepsilon = \frac{1}{m}\|B^t e^0\|_\varepsilon \leq \frac{1}{m}\|B\|_\varepsilon^t \|e^0\|_\varepsilon \\ &\leq \frac{M}{m}(\operatorname{spr}(B) + \varepsilon)^t \|e^0\|, \end{aligned}$$

bzw. wegen $\big(\frac{M}{m}\big)^{1/t} \to 1 \, (t \to \infty)$:

$$\limsup_{t\to\infty} \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} \leq \operatorname{spr}(B) + \varepsilon.$$

Da $\varepsilon > 0$ beliebig klein gewählt werden kann, ergibt sich die Behauptung. Q.E.D.

Wir tragen noch den im obigen Beweis verwendeten Hilfssatz nach:

Hilfssatz 6.1 (Spektralradius): *Für jede Matrix $B \in \mathbb{R}^{n\times n}$ gibt es zu jedem beliebig*

kleinen $\varepsilon > 0$ *eine natürliche Matrizennorm* $\|\cdot\|_\varepsilon$, *so daß*

$$\operatorname{spr}(B) \le \|B\|_\varepsilon \le \operatorname{spr}(B) + \varepsilon. \tag{6.1.11}$$

Beweis: Die Matrix B ist ähnlich zu einer Dreiecksmatrix

$$B = T^{-1}RT\,, \quad R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix},$$

mit den Eigenwerten von B auf der Hauptdiagonalen, d. h.:

$$\operatorname{spr}(B) = \max_{1\le i\le n} |r_{ii}|.$$

Für ein beliebiges $\delta \in (0,1]$ setzen wir

$$S_\delta = \begin{bmatrix} 1 & & & 0 \\ & \delta & & \\ & & \ddots & \\ 0 & & & \delta^{n-1} \end{bmatrix}, \qquad R_0 = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix},$$

$$Q_\delta = \begin{bmatrix} 0 & r_{12} & \delta r_{13} & \cdots & \delta^{n-2} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta r_{n-2,n} \\ & & & \ddots & r_{n-1,n} \\ & & & & 0 \end{bmatrix},$$

und haben damit

$$R_\delta := S_\delta^{-1} R S_\delta = \begin{bmatrix} r_{11} & \delta r_{12} & \cdots & \delta^{n-1} r_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \delta r_{n-1,n} \\ 0 & & & r_{nn} \end{bmatrix} = R_0 + \delta Q_\delta.$$

Wegen der Regularität von $S_\delta^{-1}T$ wird durch

$$\|x\|_\delta := \|S_\delta^{-1} T x\|_2\,, \quad x \in \mathbb{R}^n,$$

eine Vektornorm erklärt. Dann ist wegen $R = S_\delta R_\delta S_\delta^{-1}$:

$$B = T^{-1}RT = T^{-1} S_\delta R_\delta S_\delta^{-1} T$$

für alle $x \in \mathbb{R}^n$ und $y = S_\delta^{-1} T x$:

$$\begin{aligned} \|Bx\|_\delta &= \|T^{-1} S_\delta R_\delta S_\delta^{-1} T x\|_\delta = \|R_\delta y\|_2 \\ &\leq \|R_0 y\|_2 + \delta \|Q_\delta y\|_2 \leq \{\max_{1 \leq i \leq n} |r_{ii}| + \delta\mu\} \, \|y\|_2 \\ &\leq \{\mathrm{spr}\,(B) + \delta\mu\} \|x\|_\delta \end{aligned}$$

mit der Konstante

$$\mu = \Big(\sum_{i,j=1}^{n} |r_{ij}|^2 \Big)^{1/2}.$$

Also ist

$$\sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Bx\|_\delta}{\|x\|_\delta} \leq \mathrm{spr}\,(B) + \mu\delta,$$

und die Behauptung folgt mit $\delta = \varepsilon/\mu$. Q.E.D.

Der Spektralradius der Iterationsmatrix B bestimmt also das asymptotische Konvergenzverhalten der Iterierten x^t bzgl. jeder Vektornorm. Zu jedem $\varepsilon > 0$ existiert ein $t_\varepsilon \in \mathbb{N}$, so daß

$$\|x^t - x\| \leq (\mathrm{spr}\,(B) + \varepsilon)^t \|x^0 - x\| \quad (t \geq t_\varepsilon).$$

Dies läßt sich wie folgt interpretieren: Ist etwa $\mathrm{spr}\,(B) \leq \rho < 1$, so erhält man nach t_0 Schritten die zur weiteren Reduktion des Fehlers $\|x^{t_0} - x\|$ um den Faktor 10^{-1} (d. h. zur Gewinnung einer Dezimalstelle Genauigkeit) erforderliche Anzahl von Iterationsschritten aus der Beziehung $\rho^t \leq 10^{-1}$ zu

$$t \sim -\frac{1}{\log_{10} \rho} = -\frac{\ln(10)}{\ln(\rho)}. \tag{6.1.12}$$

In ungünstigsten Fällen mit z.B. $\mathrm{spr}\,(B) \sim 0.99$ ist $t_1 \sim 230$. Für Gleichungssysteme der Größenordnung $n > 1000$ bedeutet dies einen beträchtlichen Rechenaufwand zur Erlangung einer akzeptablen Genauigkeit.

Abbruchkriterien

Bei iterativen Verfahren ist es erforderlich, ein Abbruchkriterium anzugeben. Zunächst erhalten wir direkt durch Anwendung des Banachschen Fixpunktsatzes die Fehlerabschätzung

$$\|x^t - z\| \leq \frac{q}{1-q} \, \|x^t - x^{t-1}\|, \tag{6.1.13}$$

mit der "Kontraktionskonstante" $q = \|B\| < 1$. Bei vorgegebener Fehlertoleranz $\varepsilon > 0$ könnte man das Verfahren dann abbrechen, sobald für die relative Änderung gilt:

$$\frac{\|\delta^t\|}{\|x^t\|} \leq \frac{\|B\|}{1 - \|B\|} \, \varepsilon. \tag{6.1.14}$$

Zur Realisierung dieser Strategie wird aber eine Schätzung für die Norm $\|B\|$ bzw. für $\mathrm{spr}\,(B)$ benötigt. Diese muß indirekt aus den berechneten Iterierten x^t, d. h. *a posteriori* im Verlauf der Rechnung, gewonnen werden. In der Regel kann die Iterationsmatrix $B = I - C^{-1}A$ mit vertretbarem Aufwand gar nicht explizit gebildet werden. Methoden zur Bestimmung von $\mathrm{spr}\,(B)$ werden im Kapitel über Eigenwertaufgaben diskutiert.

Alternativ könnte man auch das Residuum $\|Ax^t - b\|$ abfragen. Über die Argumentation

$$e^t = x^t - x = A^{-1}(Ax^t - b)\,, \quad b = Ax$$
$$\|e^t\| \le \|A^{-1}\|\,\|Ax^t - b\|\,, \quad \frac{1}{\|b\|} \ge \frac{1}{\|A\|\,\|x\|}$$

erhält man

$$\frac{\|e^t\|}{\|x\|} \le \mathrm{cond}(A)\,\frac{\|Ax^t - b\|}{\|b\|}.$$

Dies hat allerdings den Nachteil, daß dazu extra Ax^t berechnet werden müßte, und daß im Falle $\mathrm{cond}(A) \gg 1$ eine starke Unterschätzung des tatsächlichen Fehlers erfolgt. Zudem ist $\mathrm{cond}(A)$ selbst natürlich wieder nur schwer schätzbar (noch schwieriger als $\mathrm{spr}\,(B)$). Wir verweisen hierfür auch auf das Kapitel über Eigenwertaufgaben.

Konstruktion von Iterationsverfahren

Bei der Konstruktion der Iterationsverfahren etwa auf dem ersten der angegebenem Wege, d. h. bei der Wahl der Matrix C, müssen zwei wesentliche Ziele berücksichtigt werden:

- $\mathrm{spr}\,(I - C^{-1}A)$ soll möglichst klein sein.
- Die Gleichungssysteme $Cx^t = (C - A)x^{t-1} + b$ sollen möglichst leicht (und mit wenig zusätzlichem Speicherplatzbedarf!) lösbar sein.

Leider widersprechen sich diese beiden Prämissen; die extremen Lösungen sind:

$$C = A \quad \Rightarrow \quad \mathrm{spr}\,(I - C^{-1}A) = 0$$
$$C = D \quad \Rightarrow \quad \mathrm{spr}\,(I - D^{-1}A) \sim 1.$$

Man wird also einen gewissen Kompromiß eingehen. Inwieweit dies beim Jacobi- und beim Gauß-Seidel-Verfahren gelungen ist, wollen wir jetzt untersuchen. Zunächst ist festzustellen, daß Punkt (ii) in beiden Fällen gut erfüllt ist, denn in jedem Iterationsschritt ist beim Jacobi-Verfahren nur ein Diagonalsystem und beim Gauß-Seidel-Verfahren ein unteres Dreieckssystem zu lösen. Es wird außerdem nicht mehr Speicherplatz benötigt, als zur Speicherung der Matrix A erforderlich ist. Dies läßt vermuten, daß der Spektralradius von $I - C^{-1}A$ nicht besonders klein sein wird. Trotzdem läßt sich für eine große Klasse von Matrizen wenigstens die Konvergenz der Verfahren garantieren, wenn diese auch oft sehr langsam ist.

6.1.1 Jacobi- und Gauß-Seidel-Verfahren

Satz 6.2 (Starkes Zeilensummenkriterium): *Genügen die Zeilensummen der Matrix* $A \in \mathbb{R}^{n\times n}$ *der Bedingung (strikte Diagonaldominanz)*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \;<\; |a_{jj}|\,, \quad j = 1,\ldots,n, \tag{6.1.15}$$

so ist $\mathrm{spr}(J) < 1$ *und* $\mathrm{spr}(H_1) < 1$ *, d.h. Jacobi- und Gauß-Seidel-Verfahren konvergieren.*

Beweis: Seien $\lambda \in \sigma(J)$ bzw. $\mu \in \sigma(H_1)$ und v bzw. w zugehörige Eigenvektoren (beachte $a_{jj} \neq 0$), d. h.:

$$\lambda v = Jv = -D^{-1}(L+R)v$$

bzw.

$$\mu w = H_1 w = -(D+L)^{-1}Rw \iff \mu w = -D^{-1}(\mu L+R)w.$$

Hieraus folgt zunächst im Falle $\|v\|_\infty = \|w\|_\infty = 1$

$$|\lambda| \le \|D^{-1}(L+R)\|_\infty = \max_{j=1,\ldots,n} \Big\{ \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\} < 1.$$

Ferner ist

$$|\mu| \le \|D^{-1}(\mu L+R)\|_\infty \le \max_{1\le j\le n} \Big\{ \frac{1}{|a_{jj}|} \Big[\sum_{k<j} |\mu|\,|a_{jk}| + \sum_{k>j} |a_{jk}| \Big] \Big\}.$$

Im Falle $|\mu| \ge 1$ ergäbe sich der Widerspruch

$$|\mu| \le |\mu|\,\|D^{-1}(L+R)\|_\infty < |\mu|,$$

so daß auch $|\mu| < 1$ sein muß. Q.E.D.

Matrizen mit der Eigenschaft aus Satz 6.2 heißen "strikt diagonaldominant". Für die Bedürfnisse der Praxis ist die Bedingung zu einschränkend; die einfache Modellmatrix aus Abschnitt 4.3

$$A = \left.\begin{bmatrix} B & -I_4 & & \\ -I_4 & B & -I_4 & \\ & -I_4 & B & -I_4 \\ & & -I_4 & B \end{bmatrix}\right\} 16\,, \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{bmatrix}\right\} 4$$

ist z.B. zwar diagonaldominant, aber nicht <u>strikt</u> diagonaldominant. Sie ist jedoch in einigen Zeilen (z.B. der ersten) strikt diagonaldominant. Dieser Umstand kann nun zum Nachweis der Konvergenz des Jacobi- und des Gauß-Seidel-Verfahrens verwendet werden.

Definition 6.1: *Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt "irreduzibel", wenn es keine Permutationsmatrix P gibt, so daß*

$$PAP^T = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}$$

(simultane Zeilen- und Spaltenvertauschung) mit Matrizen $\tilde{A}_{11} \in \mathbb{R}^{p\times p}$, $\tilde{A}_{22} \in \mathbb{R}^{q\times q}$, $\tilde{A}_{21} \in \mathbb{R}^{q\times p}$, $p, q > 0$, $p + q = n$.

Hilfssatz 6.2 (Irreduzibilität): *Eine Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann irreduzibel, wenn der zugehörige gerichtete Graph*

$$G(A) := \left\{ \textit{Knoten } P_1, ..., P_n\,,\ \textit{Kanten } \overline{P_j P_k} \Leftrightarrow a_{jk} \neq 0,\ j, k = 1, ..., n \right\}$$

zusammenhängend ist, d. h.: wenn zu jedem Knotenpaar $\{P_j, P_k\}$ eine gerichtete Kantenverbindung zwischen P_j und P_k existiert.

Beweis: Die Reduzibilität von A läßt sich auch wie folgt formulieren: Es existiert eine (nicht–triviale) Zerlegung $N_n = J \cup K$ der Indexmenge $N_n = \{1, ..., n\}$, $J, K \neq \emptyset$, $J \cap K = \emptyset$, so daß $a_{jk} = 0$ für alle Paare $\{j, k\} \in J \times K$. Der Zusammenhang des Graphen $G(A)$ bedeutet nun, daß es zu je zwei Indizes j, k stets eine Kette von Indizes $i_1, \ldots, i_m \in \{1, \ldots, n\}$ gibt, so daß

$$a_{ji_1} \neq 0\,,\ a_{i_1 i_2} \neq 0\,, \ldots\,, a_{i_{m-1} i_m} \neq 0\,,\ a_{i_m k} \neq 0.$$

Hieraus liest man direkt die behauptete Charakterisierung ab (Übungsaufgabe). Q.E.D.

Für irreduzible Matrizen kann die Bedingung des starken Zeilensummenkriteriums entscheidend abgemildert werden.

Satz 6.3 (Schwaches Zeilensummenkriterium): *Genügen die Zeilensummen einer irreduziblen Matrix $A \in \mathbb{R}^{n\times n}$ den Bedingungen*

$$\sum_{k=1, k\neq j}^{n} |a_{jk}| \leq |a_{jj}| \quad \textit{für} \quad j = 1, \ldots, n, \tag{6.1.16}$$

$$\sum_{k=1, k\neq r}^{n} |a_{rk}| < |a_{rr}| \quad \textit{für ein} \quad r \in \{1, \ldots, n\}, \tag{6.1.17}$$

so ist A regulär und $\mathrm{spr}(J) < 1$ sowie $\mathrm{spr}(H_1) < 1$, d.h. Jacobi- und Gauß-Seidel-Verfahren konvergieren.

Beweis: Wegen der Irreduzibilität von A ist notwendig

$$\sum_{k=1}^{n} |a_{jk}| > 0\,, \quad j = 1, \ldots, n,$$

und wegen der Diagonaldominanz folglich $a_{jj} \neq 0$, $j = 1, \ldots, n$. Jacobi- und Gauß-Seidel-Verfahren sind also durchführbar. Mit Hilfe der Diagonaldominanz erschließt man analog zum Beweis von Satz 6.2:

$$\operatorname{spr}(J) \leq 1\,, \quad \operatorname{spr}(H_1) \leq 1.$$

Angenommen, es gibt einen Eigenwert $\lambda \in \sigma(J)$ mit $|\lambda| = 1$. Sei $v \in \mathbb{C}^n$ ein zugehöriger normierter Eigenvektor, so daß $|v_s| = \|v\|_\infty = 1$. Es gilt dann

$$|\lambda|\,|v_i| \leq \frac{1}{|a_{ii}|} \sum_{k \neq i} |a_{ik}|\,|v_k|\,, \quad i = 1, \ldots, n. \tag{6.1.18}$$

Aufgrund der Irreduzibilität von A gibt es nun Indizes $i_1, \ldots, i_m$, so daß $a_{si_1} \neq 0\,, \ldots,$ $a_{i_m r} \neq 0$. Durch mehrfache Anwendung von (6.1.18) folgt so der Widerspruch ($|\lambda| = 1$)

$$\begin{aligned}
|v_r| &= |\lambda v_r| \leq \frac{1}{|a_{rr}|} \sum_{k \neq r} |a_{rk}|\,\|v\|_\infty &&< \|v\|_\infty \\
|v_{i_m}| &= |\lambda v_{i_m}| \leq \frac{1}{|a_{i_m i_m}|} \Big\{ \sum_{k \neq i_m, r} |a_{i_m k}|\,\|v\|_\infty + \underbrace{|a_{i_m r}|}_{\neq 0}\,|v_r| \Big\} &&< \|v\|_\infty \\
&\vdots \\
|v_{i_1}| &= |\lambda v_{i_1}| \leq \frac{1}{|a_{i_1 i_1}|} \Big\{ \sum_{k \neq i_1, i_2} |a_{i_1 k}|\,\|v\|_\infty + \underbrace{|a_{i_1 i_2}|}_{\neq 0}\,|v_{i_2}| \Big\} &&< \|v\|_\infty \\
\|v\|_\infty &= |\lambda v_s| \leq \frac{1}{|a_{ss}|} \Big\{ \sum_{k \neq s, i_1} |a_{sk}|\,\|v\|_\infty + \underbrace{|a_{s i_1}|}_{\neq 0}\,|v_{i_1}| \Big\} &&< \|v\|_\infty .
\end{aligned}$$

Also muß $\operatorname{spr}(J) < 1$ sein. Analog erschließt man unter Verwendung von (6.1.18) auch $\operatorname{spr}(H_1) < 1$. Wegen $A = D(I - J)$ muß A regulär sein. Q.E.D.

6.1.2 SOR-Verfahren

Für die in der Praxis auftretenden großen aber dünn besetzten Matrizen ist $\operatorname{spr}(J)$ bzw. $\operatorname{spr}(H_1)$ nahe bei Eins, so daß Jacobi- und Gauß-Seidel-Verfahren viel zu langsam konvergieren. Man versucht daher, die Konvergenz durch Einführung eines (oder mehrerer) sog. "Relaxationsparameter" zu beschleunigen. Beim "SOR-Verfahren" (**S**uccessive **O**ver**R**elaxation method) berechnet man im t-ten Iterationsschritt ausgehend von dem Gauß-Seidel-Zwischenwert

$$\tilde{x}_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} x_k^t - \sum_{k>j} x_k^{t-1} \Big\}$$

die nächste Iterierte x_j^t als Linearkombination

$$x_j^t = \omega \tilde{x}_j^t + (1 - \omega)\, x_j^{t-1}$$

mit einem Parameter $\omega \geq 1$. Im Falle $\omega = 1$ ist dies gerade das Gauß-Seidel-Verfahren. Im Falle $\omega < 1$ spricht man von "Unterrelaxation". Die Iterationsmatrix des Relaxationsverfahrens erhält man aus den Beziehungen

$$x^t = \omega D^{-1} \{ b - Lx^t - Rx^{t-1} \} + (1-\omega)\, x^{t-1}$$

als

$$H_\omega = (D + \omega L)^{-1} [\, (1-\omega)\, D - \omega R \,],$$

d. h.: Der Iterationsschritt lautet

$$x^t = H_\omega x^{t-1} + \omega\, (D + \omega L)^{-1} b. \tag{6.1.19}$$

Der folgende Hilfssatz zeigt, daß man sich beim Relaxationsverfahren auf den Parameterbereich $0 < \omega < 2$ beschränken muß.

Hilfssatz 6.3 (Relaxation): *Für eine beliebige Matrix $A \in \mathbb{R}^{n\times n}$ mit regulärem Diagonalanteil D gilt*

$$\operatorname{spr}(H_\omega) \geq |\omega - 1|\,, \quad \omega \in \mathbb{R}. \tag{6.1.20}$$

Beweis: Wir formen um

$$H_\omega = (D+\omega L)^{-1}[\, (1-\omega)\, D - \omega R \,] = (I + \omega \underbrace{D^{-1}L}_{=:\, L'})^{-1} \underbrace{D^{-1}D}_{=\, I} [\, (1-\omega)\, I - \omega \underbrace{D^{-1}R}_{=:\, R'}]$$

Dann gilt

$$\det(H_\omega) = \underbrace{\det(I+\omega L')}_{=\,1})^{-1} \cdot \underbrace{\det((1-\omega)\, I - \omega R')}_{=\,(1-\omega)^n} = (1-\omega)^n.$$

Wegen $\det(H_\omega) = \prod_{i=1}^n \lambda_i$ ($\lambda_i \in H_\omega$) folgt notwendigerweise

$$\operatorname{spr}(H_\omega) = \max_{1\leq i\leq n} |\lambda_i| \geq \Big(\prod_{i=1}^n |\lambda_i| \Big)^{1/n} = |1-\omega|.$$

Q.E.D.

Für positiv definite Matrizen läßt sich diese Aussage in gewisser Weise umkehren.

Satz 6.4 (SOR-Verfahren): *Für eine positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ gilt*

$$\operatorname{spr}(H_\omega) < 1\,, \quad \textit{für} \quad 0 < \omega < 2\,; \tag{6.1.21}$$

insbesondere ist das Gauß-Seidel-Verfahren konvergent.

Beweis: Wegen der Symmetrie von A ist $R = L^T$, d. h. $A = L+D+L^T$. Sei $\lambda \in \sigma(H_\omega)$ beliebig für $0 < \omega < 2$, mit einem Eigenvektor $v \in \mathbb{R}^n$, d. h. $H_\omega v = \lambda v$. Es gilt also

$$\left((1-\omega)\, D - \omega L^T\right) v = \lambda \left(D+\omega L\right) v$$

bzw.

$$\omega \left(D+L^T\right) v = (1-\lambda)\, Dv - \lambda\omega Lv.$$

Hiermit erschließt man

$$\begin{aligned} \omega Av &= \omega \left(D+L^T\right) v + \omega Lv \\ &= (1-\lambda)\, Dv - \lambda\omega Lv + \omega Lv \\ &= (1-\lambda)\, Dv + \omega\,(1-\lambda)\, Lv\,, \end{aligned}$$

und

$$\begin{aligned} \lambda\omega Av &= \lambda\omega \left(D+L^T\right) v + \lambda\omega Lv \\ &= \lambda\omega \left(D + L^T\right) v + (1-\lambda)\, Dv - \omega \left(D+L^T\right) v \\ &= (\lambda-1)\omega(D+L^T)\, v + (1-\lambda)\, Dv \\ &= (1-\lambda)(1-\omega)\, Dv - (1-\lambda)\, \omega L^T v. \end{aligned}$$

Wegen $v^T Lv = v^T L^T v$ folgt

$$\begin{aligned} \omega v^T Av &= (1-\lambda)\, v^T Dv + \omega\,(1-\lambda)\, v^T Lv \\ \lambda\omega v^T Av &= (1-\lambda)(1-\omega)\, v^T Dv - (1-\lambda)\, \omega v^T Lv, \end{aligned}$$

und hieraus durch Addition

$$\omega\,(1+\lambda)\, v^T Av = (1-\lambda)\,(2-\omega)\, v^T Dv.$$

Da mit A auch D positiv definit ist, gilt

$$v^T Av > 0, \quad v^T Dv > 0.$$

Folglich ist $(0<\omega<2)$ $\lambda \neq \pm 1$, und es gilt

$$\mu := \frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{\omega}\, \frac{v^T Dv}{v^T Av} > 0.$$

Durch Auflösen nach λ erhalten wir schließlich

$$|\lambda| = \left|\frac{\mu-1}{\mu+1}\right| < 1,$$

was zu zeigen war. Q.E.D.

Definition 6.2: *Die qualitative Konvergenzaussagen der letzten Sätze lassen sich für eine gewisse Klasse von Matrizen wesentlich verschärfen. Man nennt die Matrix* $A \in \mathbb{R}^{n\times n}$ *mit der additiven Aufspaltung* $A = L+D+R$ *"konsistent geordnet", wenn die Eigenwerte der Matrizen*

$$J(\alpha) = -D^{-1}\{\alpha L + \alpha^{-1} R\}\,, \quad \alpha \in \mathbb{C},$$

unabhängig vom Parameter α *also stets gleich denen der Jacobi–Matrix* $J = J(1)$ *sind.*

Man kann zeigen, daß neben anderen die oben eingeführte Modellmatrix "konsistent geordnet" ist. Die Bedeutung dieser Eigenschaft besteht darin, daß man in diesem Fall explizit angeben kann, wie die Eigenwerte von J mit denen von H_ω zusammenhängen.

Satz 6.5 (Optimales SOR-Verfahren): *Die Matrix* $A \in \mathbb{R}^{n\times n}$ *sei konsistent geordnet und* $0 \le \omega \le 2$. *Dann besteht zwischen den Eigenwerten* $\mu \in \sigma(J)$ *und* $\lambda \in \sigma(H_\omega)$ *die Beziehung*

$$\lambda^{1/2}\omega\mu = \lambda + \omega - 1. \tag{6.1.22}$$

Beweis: Seien λ, $\mu \in \mathbb{C}$ zwei Zahlen, welche der Gleichung (6.1.22) genügen. Im Falle $0 \neq \lambda \in \sigma(H_\omega)$ ist dann $H_\omega v = \lambda v$ äquivalent zu

$$\left((1-\omega)I - \omega D^{-1}R\right) v = \lambda(I + \omega D^{-1}L)v$$

bzw.

$$(\lambda + \omega - 1)v = -\lambda^{1/2}\omega\left(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\right) v = \lambda^{1/2}\omega J(\lambda^{1/2})\, v.$$

Also ist v Eigenvektor von $J(\lambda^{1/2})$ zum Eigenwert

$$\mu = \frac{\lambda + \omega - 1}{\lambda^{1/2}\omega}.$$

Mit der Voraussetzung an A folgt auch $\mu \in \sigma(J)$. Umgekehrt erhält man für $\mu \in \sigma(J)$ auf diese Weise auch $\lambda \in \sigma(H_\omega)$. Q.E.D.

Als direkte Folgerung aus diesem Resultat erhalten wir für konsistent geordnete Matrizen für das Gauß-Seidel-Verfahren (Fall $\omega = 1$) alternativ $\operatorname{spr}(H_1) = 0$ oder die Beziehung

$$\operatorname{spr}(H_1) = \operatorname{spr}(J)^2. \tag{6.1.23}$$

Im Falle $\operatorname{spr}(J) < 1$ konvergiert das Jacobi-Verfahren. Zur Reduzierung des Fehlers um den Faktor $1/10$ sind dann mit dem Gauß-Seidel-Verfahren nur halb so viel Iterationen erforderlich. Im allgemeinen ist das Gauß-Seidel-Verfahren dem Jacobi-Verfahren vorzuziehen. (Dies darf jedoch nicht generalisiert werden, da man Beispiele konstruieren kann, bei denen jeweils das eine, aber nicht das andere Verfahren konvergiert.)

Für konsistent geordnete Matrizen läßt sich aus der Identität (6.1.22) der "optimale" Relaxationsparameter $\omega_{\text{opt}} \in (0,2)$ mit

$$\operatorname{spr}(H_{\omega_{\text{opt}}}) \le \operatorname{spr}(H_\omega)\,, \quad \omega \in (0,2),$$

explizit ableiten. Im Falle $\rho := \mathrm{spr}\,(J) < 1$ gilt für $0 < \omega < 2$:

$$\mathrm{spr}\,(H_\omega) = \begin{cases} \omega - 1 & , \quad \omega_{\mathrm{opt}} \le \omega \\ \frac{1}{4}\left(\rho\,\omega + \sqrt{\rho^2\omega^2 - 4(\omega-1)}\right)^2 & , \quad \omega \le \omega_{\mathrm{opt}}\,, \end{cases}$$

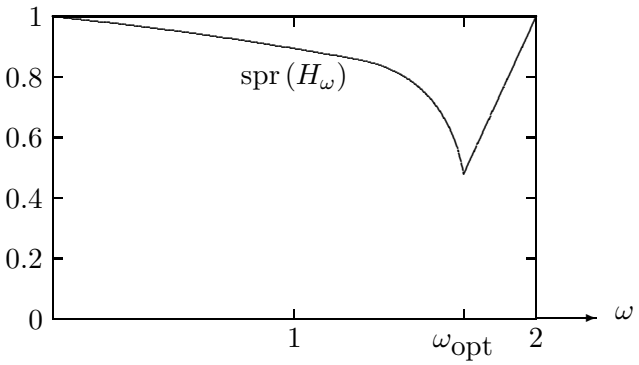


Abbildung 6.1: *Spektralradius der SOR-Matrix als Funktion von* ω

Dann ist

$$\omega_{\mathrm{opt}} = \frac{2}{1+\sqrt{1-\rho^2}}, \qquad \mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) = \omega_{\mathrm{opt}} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} < 1. \tag{6.1.24}$$

Im allgemeinen ist der genaue Wert für $\mathrm{spr}\,(J)$ nicht bekannt. Da die linksseitige Ableitung der Funktion $f(\omega) = \mathrm{spr}\,(H_\omega)$ für $\omega \to \omega_{\mathrm{opt}}$ singulär wird, ist es bei Schätzungen von ω_{opt} besser, einen etwas zu großen als zu kleinen Wert zu nehmen. Mit Hilfe von Einschließungssätzen oder auch nur der Schranke $\rho \le \|J\|_\infty$ erhält man obere Schätzungen $\bar\rho \ge \rho$. Im Falle $\bar\rho < 1$ erhält man damit durch

$$\bar\omega := \frac{2}{1+\sqrt{1-\bar\rho^2}} \ge \frac{2}{1+\sqrt{1-\rho^2}} = \omega_{\mathrm{opt}}$$

eine obere Schätzung $\bar\omega \ge \omega_{\mathrm{opt}}$ mit

$$\mathrm{spr}\,(H_{\bar\omega}) = \bar\omega - 1 = \frac{1-\sqrt{1-\bar\rho^2}}{1+\sqrt{1-\bar\rho^2}} < 1\,. \tag{6.1.25}$$

Dies setzt natürlich voraus, daß die Formel (6.1.24) überhaupt anwendbar ist.

Beispiel 6.1: Konvergenzverbesserung durch Überrelaxation

$$\mathrm{spr}\,(H_1) = \mathrm{spr}\;(J)^2 = \begin{cases} 0.81 \\ 0.99 \end{cases} \implies \mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) = \begin{cases} 0.39 \\ 0.8 \end{cases}.$$

6.2 Abstiegsverfahren

Im folgenden betrachten wir eine Klasse von Verfahren zur Lösung linearer Gleichungssysteme, die primär auf positiv definite Koeffizientenmatrizen zugeschnitten sind.

Sei $A \in \mathbb{R}^{n\times n}$ eine (symmetrische) positiv definite Matrix, d. h.:

$$\begin{aligned} (Ax,y) &= (x,Ay)\,, \quad \forall\, x\,,\, y \in \mathbb{R}^n \\ (Ax,x) &> 0\,, \quad \forall\, x \in \mathbb{R}^{n\times n} \setminus \{0\}. \end{aligned} \tag{6.2.26}$$

Es bezeichnet wieder $(\cdot,\cdot)$ das euklidische Skalarprodukt auf $\mathbb{R}^n$ und $\|\cdot\|$ ist die euklidische Vektornorm. Zugehörig zur Matrix A werden das sog. "A-Skalarprodukt" und die zugehörige "A-Norm" definiert:

$$(x,y)_A := (Ax,y)\,, \quad \|x\|_A := (Ax,x)^{1/2}. \tag{6.2.27}$$

Wir haben früher schon einige wichtige Eigenschaften positiv definiter Matrizen kennengelernt: Ihre Eigenwerte sind reell und positiv, $\lambda := \lambda_1 \le \ldots \le \lambda_n =: \Lambda$ und es existiert eine zugehörige Orthonormalbasis $\{w_1,\ldots,w_n\}$ von Eigenvektoren. Für den Spektralradius und die Spektralkonditonszahl gilt

$$\mathrm{spr}(A) = \Lambda\,, \quad \mathrm{cond}_2(A) = \frac{\Lambda}{\lambda}. \tag{6.2.28}$$

Grundlegend für das Folgende ist die Charakterisierung der Lösung $x \in \mathbb{R}^n$ des Gleichungssystems $Ax = b$ als Minimum eines quadratischen Funktionals auf $\mathbb{R}^n$.

Satz 6.6 (Minimierungseigenschaft): *Die Matrix A sei (symmetrisch) positiv definit. Die eindeutige Lösung des Gleichungssystems $Ax = b$ ist charakterisiert durch die Eigenschaft*

$$Q(x) < Q(y) \qquad \forall\, y \in \mathbb{R}^n\,,\, y \neq x, \tag{6.2.29}$$

mit dem quadratischen Funktional

$$Q(y) := \tfrac{1}{2}\,(Ay,y) - (b,y)\,. \tag{6.2.30}$$

Beweis: Sei zunächst $Ax = b$. Für $y \neq x$ ist dann

$$\begin{aligned} Q(y) - Q(x) &= \tfrac{1}{2}\,\{\,(Ay,y) - 2(b,y) - (Ax,x) + 2(b,x)\,\} \\ &= \tfrac{1}{2}\,\{\,(Ay,y) - 2(Ax,y) + (Ax,x)\,\} \\ &= \tfrac{1}{2}\,(A[x-y],x-y) > 0, \end{aligned}$$

wegen der Definitheit von A. Ist umgekehrt $Q(x) < Q(y)$, für $x \neq y$, d. h. ist x ein Minimum der Funktion Q auf $\mathbb{R}^n$, so muß notwendig $\mathrm{grad}\,Q(x) = 0$ sein. Dies bedeutet

gerade, daß gilt:

$$\frac{\partial Q}{\partial x_i}(x) = \frac{1}{2}\,\frac{\partial}{\partial x_i}\,\sum_{j,k=1}^{n} a_{jk}x_jx_k - \frac{\partial}{\partial x_i}\,\sum_{k=1}^{n} b_kx_k = \sum_{k=1}^{n} a_{ik}x_k - b_i = 0\,,$$

für $i = 1, \ldots, n$; man beachte $a_{jk} = a_{kj}$. Also ist $Ax = b$. Q.E.D.

Wir halten fest, daß der Gradient von Q in einem Punkt $y \in \mathbb{R}^n$ gegeben ist durch

$$\text{grad } Q(y) = \tfrac{1}{2}\,(A + A^T)y - b = Ay - b. \tag{6.2.31}$$

(Dies ist gerade der negative "Defekt" im Punkt y.) Die sog. "Abstiegsverfahren" bestimmen nun ausgehend von einem geeigneten Startvektor $x^{(0)} \in \mathbb{R}^n$ eine Folge von Iterierten x^t, $t \in \mathbb{N}$, durch

$$x^{t+1} = x^t + \alpha_t r^t. \tag{6.2.32}$$

Dabei sind die r^t vorgegebene oder auch erst im Verlauf der Iteration berechnete "Abstiegsrichtungen", und die "Schrittweiten" $\alpha_t \in \mathbb{R}$ sind durch die Vorschrift bestimmt (sog. "line search"):

$$Q(x^{t+1}) = \min_{\alpha\in\mathbb{R}}\, Q(x^t + \alpha r^t). \tag{6.2.33}$$

Wegen

$$\frac{d}{d\alpha}\,Q(x^t + \alpha r^t) = \text{grad}\,Q(x^t + \alpha r^t)\cdot r^t = (Ax^t - b, r^t) + \alpha(Ar^t, r^t)$$

ist notwendig

$$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}\,, \quad g^t := Ax^t - b = \text{grad}\,Q(x^t)\,.$$

Definition 6.3: *Das allgemeine Abstiegsverfahren bestimmt ausgehend von einem Startwert $x^0 \in \mathbb{R}^n$ eine Folge von Iterierten $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$ nach der Vorschrift:*

$$\begin{aligned} &\textit{Gradient } g^t = Ax^t - b, \quad \textit{Abstiegsrichtung } r^t, \\ &\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t. \end{aligned}$$

Praktisch günstiger ist die folgende Schreibweise, bei der man eine Matrix-Vektor-Multiplikation spart, wenn man den Vektor Ar^t abspeichert:

$$\begin{aligned} &\text{Startwerte:} && x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b \\ &\text{für } t \geq 0: && g^t = Ax^t - b, \quad \text{Abstiegsrichtung} \quad r^t \\ & && \alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t, \quad g^{t+1} = g^t + \alpha_t Ar^t. \end{aligned}$$

Unter Verwendung der Notation $\|y\|_B := (By, y)^{1/2}$ gilt

$$2Q(y) = \|Ay - b\|^2_{A^{-1}} - \|b\|^2_{A^{-1}} = \|y - x\|^2_A - \|x\|^2_A, \tag{6.2.34}$$

d. h.: Die Minimierung des Funktionals $Q(\cdot)$ ist äquivalent zur Minimierung der Defektnorm $\|Ay - b\|_{A^{-1}}$ bzw. der Fehlernorm $\|y - x\|_A$.

6.2.1 Gradienten-Verfahren

Die verschiedenen Abstiegsverfahren unterscheiden sich im wesentlichen durch die jeweilige Wahl der Abstiegsrichtungen r^t. Die einfachste Möglichkeit wäre, die Richtungen r^t zyklisch die kartesischen Einheitsvektoren $\{e^1, \ldots, e^n\}$ durchlaufen zu lassen. Die so erhaltene iterative Methode wird "Koordinatenrelaxation" genannt; sie ist äquivalent zum Gauß-Seidel-Verfahren (Übungsaufgabe). Naheliegender ist die Wahl der Richtung des stärksten Abfalls von $Q(\cdot)$ im Punkt x^t als Suchrichtung:

$$r^t = -\mathrm{grad}\, Q(x^t) = -g^t. \tag{6.2.35}$$

Definition 6.4: *Das "Gradientenverfahren" bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

$$\begin{aligned} &\textit{Startwerte:} && x^0 \in \mathbb{R}^n\,, \quad g^0 := Ax^0 - b \\ &\textit{für } t \geq 0\textit{:} && \alpha_t = \frac{\|g^t\|^2}{(Ag^t, g^t)} \\ & && x^{t+1} = x^t - \alpha_t g^t, \quad g^{t+1} = g^t - \alpha_t A g^t. \end{aligned}$$

Im Falle $(Ag^t, g^t) = 0$ ist notwendig auch $g^t = 0$, d. h.: Die Iteration kann nur mit $Ax^t = b$ abbrechen.

Satz 6.7 (Gradientenverfahren): *Ist die Matrix $A \in \mathbb{R}^{n\times n}$ positiv definit, so konvergiert das Gradientenverfahren für jeden Startvektor $x^0 \in \mathbb{R}^n$ gegen die Lösung des Gleichungssystems $Ax = b$.*

Beweis: Wir führen das "Fehlerfunktional" ein

$$E(y) \;:=\; \|y - x\|^2_A = (y - x, A[y-x])\,, \quad y \in \mathbb{R}^n,$$

und setzen zur Abkürzung $e^t := x^t - x$. Mit diesen Bezeichnungen gilt dann

$$\begin{aligned} \frac{E(x^t) - E(x^{t+1})}{E(x^t)} &= \frac{(e^t, Ae^t) - (e^{t+1}, Ae^{t+1})}{(e^t, Ae^t)} \\ &= \frac{(e^t, Ae^t) - (e^t - \alpha_t g^t, A[e^t - \alpha_t g^t])}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t (e^t, Ag^t) - \alpha_t^2 (g^t, Ag^t)}{(e^t, Ae^t)} \end{aligned}$$

und folglich, wegen $Ae^t = Ax^t - Ax = Ax^t - b = g^t$,

$$\frac{E(x^t) - E(x^{t+1})}{E(x^t)} = \frac{2\alpha_t \|g^t\|^2 - \alpha_t^2 (g^t, Ag^t)}{(g^t, A^{-1}g^t)} = \frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)}.$$

Für die positiv definite Matrix A gilt

$$\lambda \|y\|^2 \le (y, Ay) \le \Lambda \|y\|^2, \quad \Lambda^{-1}\|y\|^2 \le (y, A^{-1}y) \le \lambda^{-1}\|y\|^2,$$

mit $\lambda = \lambda_{\min}(A)$ und $\Lambda = \lambda_{\max}(A)$. Im Falle $x^t \neq x$, d. h. $E(x^t) \neq 0$ und $g^t \neq 0$, erschließt man damit

$$\frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)} \ge \frac{\|g^t\|^4}{\Lambda \|g^t\|^2 \, \lambda^{-1} \|g^t\|^2} = \frac{\lambda}{\Lambda},$$

bzw.

$$E(x^{t+1}) \le \{\, 1 - \kappa^{-1} \,\} \, E(x^t), \quad \kappa := \mathrm{cond}_{\mathrm{nat}}(A).$$

Wegen $0 < 1 - 1/\kappa < 1$ konvergiert somit für jedes $x^0 \in \mathbb{R}^n$ das Fehlerfunktional $E(x^t) \to 0$ $(t \to \infty)$, d. h.: $x^t \to x$ $(t \to \infty)$. Q.E.D.

Für eine verschärfte Abschätzung der Konvergenzgeschwindigkeit des Gradientenverfahrens benötigen wir das folgende Resultat von Kantorowitsch:

Hilfssatz 6.4 (Lemma von Kantorowitsch): *Für die positiv definite Matrix $A \in \mathbb{R}^n$ mit kleinstem Eigenwert λ und größtem Eigenwert Λ gilt*

$$4\,\frac{\lambda\Lambda}{(\lambda - \Lambda)^2} \le \frac{\|y\|^4}{(y, Ay)(y, A^{-1}y)}, \quad \forall\, y \in \mathbb{R}^n. \tag{6.2.36}$$

Beweis: Seien $\lambda = \lambda_1 \le \ldots \le \lambda_n = \Lambda$ die Eigenwerte von A und $\{w_1, \ldots, w_n\}$ eine zugehörige Orthonormalbasis von Eigenvektoren. Ein beliebiger Vektor $y \in \mathbb{R}^n$ gestattet die Entwicklung $y = \sum_{i=1}^n y_i w_i$ mit den Koeffizienten $y_i = (y, w_i)$. Dann gilt

$$\frac{\|y\|^4}{(y, Ay)(y, A^{-1}y)} = \frac{(\sum_{i=1}^n y_i^2)^2}{(\sum_{i=1}^n \lambda_i y_i^2)\,(\sum_{i=1}^n \lambda_i^{-1} y_i^2)} = \frac{1}{(\sum_{i=1}^n \lambda_i \zeta_i)\,(\sum_{i=1}^n \lambda_i^{-1} \zeta_i)} = \frac{\varphi(\zeta)}{\psi(\zeta)}$$

mit den Bezeichnungen

$$\zeta = (\zeta_i)_{i=1,\ldots,n}, \quad \zeta_i = y_i^2 (\sum_{i=1}^n y_i^2)^{-1},$$

$$\psi(\zeta) = \sum_{i=1}^n \lambda_i^{-1} \zeta_i, \quad \varphi(\zeta) = (\sum_{i=1}^n \lambda_i \zeta_i)^{-1}.$$

Da die Funktion $f(\lambda) = \lambda^{-1}$ konvex ist, folgt aus $0 \le \zeta_i \le 1$ und $\sum_{i=1}^n \zeta_i = 1$, daß gilt

$$\sum_{i=1}^{n} \lambda_i^{-1}\zeta_i \ge (\sum_{i=1}^{n} \lambda_i\zeta_i)^{-1}.$$

Wir setzen $g(\lambda) := (\lambda_1 + \lambda_n - \lambda)/(\lambda_1\lambda_n)$.

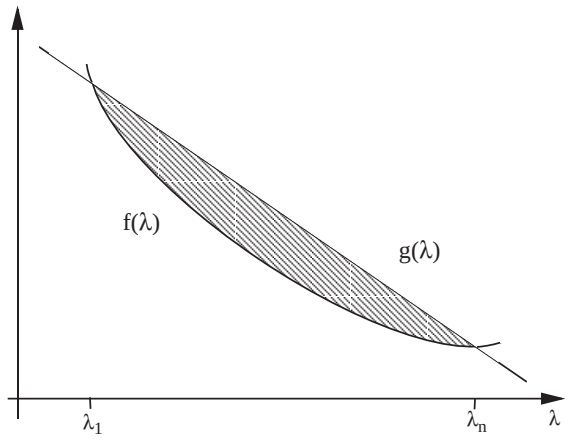


Abbildung 6.2: *Skizze zu Beweis des Lemma von Kantorowitsch*

Offenbar liegt $\varphi(\zeta)$ für alle Argumente ζ stets auf der Kurve $f(\lambda)$, und $\psi(\zeta)$ liegt stets zwischen den Kurven $f(\lambda)$ und $g(\lambda)$ (schraffierter Bereich). Folglich gilt

$$\frac{\varphi(\zeta)}{\psi(\zeta)} \ge \min_{\lambda_1 \le \lambda \le \lambda_n} \frac{f(\lambda)}{g(\lambda)} = \frac{f([\lambda_1 + \lambda_n]/2)}{g([\lambda_1 + \lambda_n]/2)} = \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}.$$

Q.E.D.

Satz 6.8 (Fehlerabschätzung): *Für das Gradientenverfahren gilt die Fehlerabschätzung*

$$\|x^t - z\|_A \le \Big(\frac{1 - 1/\kappa}{1 + 1/\kappa}\Big)^t \|x^0 - z\|_A\,, \quad t \in \mathbb{N}, \tag{6.2.37}$$

mit der Spektralkonditionszahl $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ *von* A.

Beweis: Im Beweis von Satz 6.7 wurde die folgende Identität gezeigt:

$$E(x^{t+1}) = \Big\{ 1 - \frac{\|g^t\|^4}{(g^t, Ag^t)\,(g^t, A^{-1}g^t)} \Big\} E(x^t).$$

Diese ergibt mit der Ungleichung von Kantorowitsch

$$E(x^{t+1}) \le \Big\{ 1 - 4\frac{\lambda\Lambda}{(\lambda + \Lambda)^2} \Big\} E(x^t) = \Big(\frac{\lambda - \Lambda}{\lambda + \Lambda} \Big)^2 E(x^t).$$

Daraus folgt dann durch sukzessive Anwendung

$$\|x^t - x\|_A^2 \leq \Big(\frac{\lambda - \Lambda}{\lambda + \lambda}\Big)^{2t} \|x^0 - x\|_A^2\,, \quad t \in \mathbb{N}.$$

Q.E.D.

Der Beziehung

$$\begin{aligned}(g^{t+1}, g^t) &= (g^{(t)} - \alpha_t A g^t, g^t) \\ &= \|g^t\|^2 - \alpha_t (A g^t, g^t) = 0.\end{aligned} \tag{6.2.38}$$

entnehmen wir, daß die Abstiegsrichtungen $r^{(t)} = -g^{(t)}$ des Gradientenverfahrens in jeweils direkt aufeinanderfolgenden Schritten orthogonal sind. Dagegen kann $g^{(t+2)}$ weit von Orthogonalität zu $g^{(t)}$ abweichen. Dies führt zu einem stark oszillatorischen Konvergenzverhalten des Gradientenverfahrens besonders bei Matrizen A mit weit auseinander liegenden Eigenwerten. Dies bedeutet etwa in Fall $n = 2$, daß das Funktional $Q(\cdot)$ stark exzentrische Niveaulinien hat, und sich die Iterierten in einem Zickzackkurs der Lösung annähern (siehe Abb. 6.2.1).

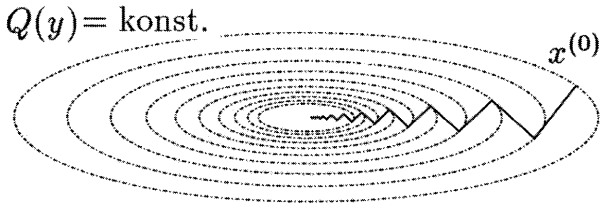


Abbildung 6.3: *Oszillierende Konvergenz des Gradientenverfahrens*

6.2.2 CG-Verfahren

Das Gradientenverfahren nutzt die Struktur des quadratischen Funktionals $Q(\cdot)$, d. h. die Verteilung der Eigenwerte der Matrix A, nur lokal von einem Schritt zum nächsten aus. Es wäre günstiger, wenn bei der Wahl der Abstiegsrichtungen auch die bereits gewonnenen Informationen über die globale Struktur von $Q(\cdot)$ berücksichtigt würde, d. h. wenn etwa die Abstiegsrichtungen paarweise orthogonal wären. Dies ist die Grundidee des "Verfahrens der konjugierten Richtungen" ("conjugate gradient method"; "CG-Verfahren") nach Hestenes[2] und Stiefel[3] (1992), welches sukzessive eine Folge von Abstiegsrichtungen $d^{(t)}$

[2]Magnus R. Hestenes (1906-1991): US-Amerikanischer Mathematiker; arbeitete am National Bureau of Standards (NBS)und an der University of California at Los Angeles (UCLA); Beiträge zur Optimierung und Kontrolltheorie und zur Numerischen Linearen Algebra.

[3]Eduard Stiefel (1909-1978): Schweizer Mathematiker; seit 1943 Professor für Angewandte Mathematik an der ETH Zürich; wichtige Beiträge zu Topologie, Gruppentheorie, Numerische Lineare Algebra (CG-

erzeugt, die bzgl. des Skalarprodukts $(\cdot,\cdot)_A$ orthogonal sind "A-orthogonal").

Zur Herleitung des CG-Verfahrens verwenden wir den Ansatz

$$B_t := \text{span}\{d^0, \cdots, d^{t-1}\} \tag{6.2.39}$$

mit einem noch zu bestimmenden linear unabhängigen Satz von Vektoren d^i und suchen die Iterierten in der Form

$$x^t = x^0 + \sum_{i=0}^{t-1} \alpha_i d^i \ \in \ x^0 + B_t \tag{6.2.40}$$

so zu bestimmen, daß

$$Q(x^t) = \min_{y \in x^0 + B_t} Q(y) \quad \leftrightarrow \quad \|Ax^t - b\|_{A^{-1}} = \min_{y \in x^0 + B_t} \|Ay - b\|_{A^{-1}}. \tag{6.2.41}$$

Durch Nullsetzten der Ableitungen von $Q(\cdot)$ nach den α_i sehen wir, daß dies äquivalent ist zu den sog. "Galerkin[4]-Gleichungen":

$$(Ax^t - b, d^i) = 0\,, \quad i = 0, \ldots, t-1. \tag{6.2.42}$$

bzw. in Kurzschreibweise $Ax^t - b = g^t \perp B_t\,.$

Bemerkung 6.1: Wir bemerken, daß (6.2.42) nicht von der Symmetrie der Matrix A abhängt. Ausgehend von dieser Beziehung lassen sich auch CG-artige Verfahren für unsymmetrische und sogar indefinite Gleichungssysteme ableiten. Diese werden allgemein "Projektionsverfahren" genannt.

Setzt man den obigen Ansatz für x^t in die Orthogonalitätsbedingung (6.2.42) ein, so erhält man ein reguläres Gleichungssystem für die Koeffizienten $\alpha_i\,(i = 0, \ldots, t-1)$. Es sei nochmals daran erinnert, daß die Galerkin-Gleichungen (6.2.42) äquivalent sind zur Minimierung der Defektnorm $\|Ax^t - b\|_{A^{-1}}$ oder der Fehlernorm $\|x^t - x\|_A$ über $x^0 + B_t$. Eine natürliche Wahl für B_t sind die sog. Krylow[5]-Räume

$$B_t = K_t(d^0; A) := \text{span}\{d^0, Ad^0, \ldots, A^{t-1}d^0\}, \tag{6.2.43}$$

mit einem Vektor d^0, etwa dem Anfangsdefekt $d^0 = b - Ax^0$ zu irgend einem Startvektor x^0. Dies ist motiviert durch die Beobachtung, daß aus $A^t d^0 \in K_t(d^0; A)$ notwendig $-g^t = b - Ax^t = d^0 + A(x^0 - x^t) \in d^0 + AK_t(d^0; A) \in K_t(d^0; A)$ folgt. Wegen $g^t \perp K_t(d^0; A)$ impliziert dies dann $g^t = 0$ gemäß Konstruktion.

verfahren) und Approximationstheorie sowie zur Himmelsmechanik.

[4]Boris Grigorievich Galerkin (1871-1945): russischer Bauingenieur und Mathematiker; Prof. in St. Petersburg; Beiträge zur Struktur-Mechanik, insbesondere zur Plattentheorie.

[5]Aleksei Nikolaevich Krylov (1863-1945): russischer Mathematiker; Prof. an der Sov. Akademie der Wissensch. in St. Petersburg; Beiträge zu Fourier-Analysis und Differentialgleichungen, Anwendungen in der Schiffstechnik.

Das CG-Verfahren erzeugt nun Abstiegsrichtungen, die eine A-orthogonale Basis des Krylow-Raumes $K_t(d^0;A)$ bilden. Dazu wird induktiv vorgegangen: Ausgehend von einem Startpunkt x^0 mit Defekt $d^0 = b - Ax^0$ seien Iterierte x^i und zugehörige Abstiegsrichtungen $d^i (i = 0, ..., t-1)$ bereits bestimmt, so daß $\{d^0, ..., d^{t-1}\}$ eine A-orthogonale Basis von $K_t(d^0;A)$ ist. Zur Konstruktion des nächsten $d^t \in K_{t+1}(d^0;A)$ mit der Eigenschaft $d^t \perp_A K_t(d^0;A)$ machen wir den Ansatz

$$d^t = -g^t + \sum_{j=0}^{t-1} \beta_j^{t-1} d^j \ \in \ K_{t+1}(d^0;A) \tag{6.2.44}$$

Dabei wird o.B.d.A. angenommen, daß $g^t = Ax^t - b \notin K_t(d^0;A)$ ist, da andernfalls $g^t = 0$ bzw. $x^t = x$ wäre. Dann gilt für $i = 0, ..., t-1$:

$$(d^t, Ad^i) = (-g^t, Ad^i) + \sum_{j=0}^{t-1} \beta_j^{t-1} (d^j, Ad^i) = (-g^t + \beta_i^{t-1}\, d^i, Ad^i)\,. \tag{6.2.45}$$

Für $i < t-1$ ist $(g^t, Ad^i) = 0$ wegen $Ad^i \in K_t(d^0;A)$ und demnach $\beta_i^{t-1} = 0$. Für $i = t-1$ führt die Bedingung

$$0 = (-g^t, Ad^{t-1}) + \beta_{t-1}^{t-1} (d^{t-1}, Ad^{t-1}) \tag{6.2.46}$$

zu den Formeln

$$\beta_{t-1} := \beta_{t-1}^{t-1} = \frac{(g^t, Ad^{t-1})}{(d^{t-1}, Ad^{t-1})}\,, \quad d^t = -g^t + \beta_{t-1} d^{t-1}. \tag{6.2.47}$$

Die nächsten Iterierten x^{t+1} und $g^{t+1} = Ax^{t+1} - b$ sind dann bestimmt durch

$$\alpha_t = -\frac{(g^t, d^t)}{(d^t, Ad^t)}\,, \quad x^{t+1} = x^t + \alpha_t d^t\,, \quad g^{t+1} = g^t + \alpha_t A d^t\,. \tag{6.2.48}$$

Dies sind die Rekursionsformeln des klassischen CG-Verfahrens. Nach Konstruktion gilt

$$(d^t, Ad^i) = (g^t, d^i) = 0, \quad i \le t-1\,, \quad (g^t, g^{t-1}) = 0. \tag{6.2.49}$$

Damit folgern wir, daß

$$\|g^t\|^2 = (d^t - \beta_{t-1} d^{t-1}, -g^{t+1} + \alpha_t A d^t) = \alpha_t (d^t, Ad^t), \tag{6.2.50}$$

$$\|g^{t+1}\|^2 = (g^t + \alpha_t A d^t, g^{t+1}) = \alpha_t (Ad^t, g^{t+1}). \tag{6.2.51}$$

Dies gestattet einige Vereinfachungen in den Formeln, nämlich

$$\alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \tag{6.2.52}$$

solange die Iteration nicht mit $g^t = 0$ abbricht.

Definition 6.5: *Das CG-Verfahren bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

$$\begin{aligned} &\textit{Startwerte:} && x^0 \in \mathbb{R}^n\,, \quad d^0 = -g^0 = b - Ax^0, \\ &\textit{für } t \geq 0\textit{:} && \alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^{(t)} + \alpha_t A d^t, \\ & && \beta_t = \frac{\|g^{(t+1)}\|^2}{\|g^{(t)}\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t. \end{aligned}$$

Nach Konstruktion erzeugt das CG-Verfahren eine Folge von Abstiegsrichtungen d^t, welche automatisch paarweise A-orthogonal sind. Dies impliziert, daß die Vektoren $d^0, \ldots, d^t$ jeweils linear unabhängig sind, und daß folglich gilt

$$\operatorname{span}\{d^0, \ldots, d^{n-1}\} = \mathbb{R}^n. \tag{6.2.53}$$

Wir fassen die bisher abgeleiteten Eigenschaften des CG-Verfahrens zusammen:

Satz 6.9 (CG-Verfahren): *Das CG-Verfahren bricht für jeden Startvektor $x^0 \in \mathbb{R}^n$ (bei rundungsfreier Rechnung) nach spätestens n Schritten mit $x^{t_n} = x$ ab. Dabei gilt in jedem Schritt*

$$Q(x^t) = \min_{\alpha \in \mathbb{R}} Q(x^{t-1} + \alpha d^{t-1}) = \min_{y \in x^0 + B_t} Q(y) \tag{6.2.54}$$

mit $B_t := \operatorname{span}\{d^0, \ldots, d^{t-1}\}$.

Die Methode der konjugierten Richtungen gehört also im Gegensatz zum Gradientenverfahren eigentlich zur Klasse der "direkten" Verfahren. In der Praxis wird sie jedoch wie ein iteratives Verfahren angewendet, da

1. aufgrund von Rundungsfehlern die Richtungen d^t nicht wirklich A-orthogonal sind, und die Iteration nicht abbricht;
2. bei großen Matrizen auch mit deutlich weniger als n Iterationsschritten schon brauchbare Näherungen erzielbar sind.

Zur Vorbereitung des Hauptsatzes über die Konvergenzgeschwindigkeit des CG-Verfahrens beweisen wir zunächst den folgenden Hilfssatz.

Hilfssatz 6.5 (Polynomiale Normschranke): *Für ein Polynom $p \in P_t$, $p(0) = 1$, gelte auf einer Menge $S \subset \mathbb{R}$, welche alle Eigenwerte von A enthält,*

$$\sup_{\mu \in S} |p(\mu)| \leq M. \tag{6.2.55}$$

Dann gilt

$$\|x^t - x\|_A \leq M\|x^0 - x\|_A. \tag{6.2.56}$$

Beweis: Unter Beachtung der Beziehung

$$\|x^t - x\|_A = \min_{y \in x^0 + B_t} \|y - x\|_A,$$

$$B_t = \operatorname{span}\{d^0, \ldots, d^{t-1}\} = \operatorname{span}\{A^0 g^{(0)}, \ldots, A^{t-1} g^0\}$$

finden wir

$$\|x^t - x\|_A = \min_{p \in P_{t-1}} \|x^0 - x + p(A) g^0\|_A.$$

Wegen $g^0 = Ax^0 - b = A(x^0 - x)$ folgt weiter

$$\begin{aligned} \|x^t - x\|_A &= \min_{p \in P_{t-1}} \|[I + Ap(A)](x^0 - x)\|_A \\ &\leq \min_{p \in P_{t-1}} \|I + Ap(A)\|_A \, \|x^0 - x\|_A \leq \min_{p \in P_t,\, p(0)=1} \|p(A)\|_A \, \|x^0 - x\|_A, \end{aligned}$$

mit der von $\|\cdot\|_A$ erzeugten natürlichen Matrizennorm $\|\cdot\|_A$. Für beliebiges $y \in \mathbb{R}^n$ gilt mit einer Orthonormalbasis $\{w_1, \ldots, w_n\}$ aus Eigenvektoren von A:

$$y = \sum_{i=1}^{n} \gamma_i w_i\,, \quad \gamma_i = (y, w_i),$$

und folglich

$$\|p(A)y\|_A^2 = \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \gamma_i^2 \leq M^2 \sum_{i=1}^{n} \lambda_i \gamma_i^2 = M^2 \, \|y\|_A^2.$$

Dies impliziert

$$\|p(A)\|_A = \sup_{y \in \mathbb{R}^n,\, y \neq 0} \frac{\|p(A)y\|_A}{\|y\|_A} \leq M$$

und damit die Behauptung. Q.E.D.

Als Folgerung von Hilfssatz 6.5 erhalten wir die folgende a priori Fehlerabschätzung:

Satz 6.10 (CG-Konvergenz): *Für das CG-Verfahren gilt die Fehlerabschätzung*

$$\|x^t - x\|_A \leq 2 \left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} \right)^t \|x^0 - x\|_A\,, \quad t \in \mathbb{N}\,, \tag{6.2.57}$$

mit der Spektralkonditionszahl $\kappa = \operatorname{cond}_2(A) = \Lambda/\lambda$ *von* A*. Zur Reduzierung des Anfangsfehlers um den Faktor* ε *sind höchstens*

$$t(\varepsilon) \leq \frac{1}{2} \sqrt{\kappa} \, \ln\left(\frac{2}{\varepsilon}\right) + 1 \tag{6.2.58}$$

Iterationsschritte erforderlich.

Beweis: Setzt man $S := [\lambda, \Lambda]$ in Hilfssatz 6.5, so folgt

$$\|x^t - x\|_A \leq \min_{p \in P_t,\, p(0)=1} \Big\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \Big\} \, \|x^0 - x\|_A \, .$$

Dies ergibt die Behauptung, wenn wir zeigen können, daß

$$\min_{p \in P_t,\, p(0)=1} \Big\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \Big\} \leq 2 \Big(\frac{1 - \sqrt{\lambda/\Lambda}}{1 + \sqrt{\lambda/\Lambda}} \Big)^t .$$

Dabei handelt es sich um ein Problem der Bestapproximation mit Polynomen bzgl. der Maximumnorm (Tschebyscheff-Approximation). Die Lösung $\bar{p}$ ist gegeben durch

$$\bar{p}(\mu) = T_t\Big(\frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda} \Big) \, T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda} \Big)^{-1},$$

mit dem t-ten Tschebyscheff-Polynom T_t auf $[-1, 1]$. Dabei ist

$$\sup_{\lambda \leq \mu \leq \Lambda} \bar{p}(\mu) = T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda} \Big)^{-1}.$$

Aus der Darstellung

$$T_t(\mu) = \tfrac{1}{2} \left[\big(\mu + \sqrt{\mu^2 - 1} \big)^t + \big(\mu - \sqrt{\mu^2 - 1} \big)^t \right], \quad \mu \in [-1, 1],$$

für die Tschebyscheff-Polynome folgt über die Identität

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

die Abschätzung nach unten

$$T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda} \Big) = T_t\Big(\frac{\kappa + 1}{\kappa - 1} \Big) = \frac{1}{2} \left[\Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \Big)^t + \Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \Big)^t \right] \geq \frac{1}{2} \Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \Big)^t .$$

Also wird

$$\sup_{\lambda \leq \mu \leq \Lambda} \bar{p}(\mu) \leq 2 \Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \Big)^t,$$

was (6.2.57) impliziert. Zur Herleitung von (6.2.58) fordern wir

$$2 \Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \Big)^{t(\varepsilon)} < \varepsilon$$

bzw.

$$t(\varepsilon) > \ln \Big(\frac{2}{\varepsilon} \Big) \ln \Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \Big)^{-1}.$$

Wegen

$$\ln \frac{x+1}{x-1} = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \dots\right\} \geq \frac{2}{x}$$

ist dies erfüllt für $t(\varepsilon) \geq \frac{1}{2}\sqrt{\kappa}\,\ln(2/\varepsilon)$. Q.E.D.

Wegen $\kappa = \mathrm{cond}_{\mathrm{nat}}(A) > 1$ ist $\sqrt{\kappa} < \kappa$. Da die Funktion $f(\lambda) = (1-\lambda^{-1})\,(1+\lambda^{-1})^{-1}$ für $\lambda > 0$ streng monoton wachsend ist $(f'(\lambda) > 0)$, folgt

$$\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} < \frac{1 - 1/\kappa}{1 + 1/\kappa},$$

d. h.: Die Methode der konjugierten Richtungen sollte schneller konvergieren als das Gradientenverfahren. Dies ist auch praktisch der Fall. Beide Verfahren konvergieren offenbar umso schneller, je näher die Kondition $\mathrm{cond}_{\mathrm{nat}}(A)$ bei 1 liegt. Ist aber $\Lambda \gg \lambda$, was in der Praxis leider häufig der Fall ist, konvergiert auch die Methode der konjugierten Richtungen nur sehr langsam. Eine Beschleunigung der Konvergenz kann durch sog. "Vorkonditionierung" erreicht werden, die wir weiter unten beschreiben werden.

6.2.3 Allgemeinere CG-Verfahren und Vorkonditionierung

Zur Lösung allgemeiner Gleichungssysteme $Ax = b$ mit einer regulären, aber nicht notwendig positiv definiten Matrix $A \in \mathbb{R}^n$ mit Hilfe des CG-Verfahrens kann man etwa zu dem äquivalenten System

$$A^T A x = A^T b \tag{6.2.59}$$

mit der positiv definiten Matrix $A^T A$ übergehen. Hierauf angewendet, schreibt sich das CG-Verfahren wie folgt:

$$\begin{aligned}
&\text{Startwerte:} && x^0 \in \mathbb{R}^n\,, \quad d^0 = A^T(b - Ax^0)\\
&\text{für } t \geq 0: && \alpha_t = \frac{\|g^t\|^2}{\|Ad^t\|^2}\,, \quad (g^t = A^T Ax - A^T b)\\
& && x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A^T A d^t\\
& && \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t\,.
\end{aligned}$$

Die Konvergenzgeschwindigkeit ist dabei charakterisiert durch $\kappa(A^T A)$. Das ganze Verfahren beruht offenbar auf der Minimierung des Funktionals

$$Q(y) := \tfrac{1}{2}\,(A^T Ay, y) - (A^T b, y) = \|Ay - b\|^2 - \tfrac{1}{2}\,\|b\|^2. \tag{6.2.60}$$

Da $\kappa(A^T A) \sim \kappa(A)^2$ ist, muß man mit einer recht langsamen Konvergenz des CG-Verfahrens für nicht symmetrische Systeme rechnen.

Offensichtlich funktioniert das Verfahren der konjugierten Richtungen um so besser,

je näher die Kondition der Matrix A bei Eins liegt:

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1. \tag{6.2.61}$$

Daher wird eine "Vorkonditionierung" vorgenommen, d. h.: Das System $Ax = b$ wird in ein äquivalentes umgeformt, $\tilde{A}\tilde{x} = \tilde{b}$, dessen Matrix $\tilde{A}$ näher bei I liegt. Sei C eine symmetrische, positiv definite Matrix, welche sich als Produkt darstellen läßt,

$$C = KK^T \tag{6.2.62}$$

mit einer regulären Matrix K. Das System $Ax = b$ wird dann in der äquivalenten Form geschrieben

$$\underbrace{K^{-1}A\,(K^T)^{-1}}_{\tilde{A}}\ \underbrace{K^T x}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}. \tag{6.2.63}$$

Das Verfahren der konjugierten Richtungen wird nun auf das System $\tilde{A}\tilde{x} = \tilde{b}$ angewendet. Die Idee dabei ist, die Matrix C so zu wählen, daß $\kappa(\tilde{A}) \ll \kappa(A)$ wird. Die Beziehung

$$(K^T)^{-1}\tilde{A}K^T = (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T = C^{-1}A \tag{6.2.64}$$

zeigt, daß bei der Wahl $C = uivA$ die Matrix $\tilde{A}$ ähnlich zu I, d. h. $\kappa(\tilde{A}) = \kappa(I) = I$ wäre. Folglich wird man C als möglichst gute Approximation von A wählen, wobei natürlich die Zerlegung $C = KK^T$ bekannt sein muß. Das CG-Verfahren sieht dann wie folgtaus:

Startwerte: $\tilde{x}^{(0)} \in \mathbb{R}^n\,, \quad \tilde{d}^0 = -\tilde{g}^0 = \tilde{b} - \tilde{A}\tilde{x}^0$

für $t \geq 0$:
$$\alpha_t = \frac{\|\tilde{g}^t\|^2}{(\tilde{d}^t, \tilde{A}\tilde{d}^t)}$$
$$\tilde{x}^{t+1} = \tilde{x}^t + \alpha_t\tilde{d}^t, \quad \tilde{g}^{t+1} = \tilde{g}^t + \alpha_t\tilde{A}\tilde{d}^t$$
$$\beta_t = \frac{\|\tilde{g}^{t+1}\|^2}{\|\tilde{g}^t\|^2}, \quad \tilde{d}^{t+1} = -\tilde{g}^{t+1} + \beta_t\tilde{d}^t.$$

Diesen Algorithmus schreibt man üblicherweise bezogen auf die ursprüngliche Matrix A und erhält so das sog. "PCG-Verfahren" ("Preconditioned PC method").

Definition 6.6: *Das PCG-Verfahren mit (regulärer) Vorkonditionierungsmatrix $C = KK^T$ bestimmt eine Folge von Iterierten $x^t \in \mathbb{R}^n$ gemäß der Vorschrift:*

Startwerte: $x^0 \in \mathbb{R}^n\,, \quad g^0 = Ax^0 - b\,, \quad C\rho^0 = g^0\,, \quad d^0 = -\rho^0$

für $t \geq 0$:
$$\alpha_t = \frac{(g^t, \rho^t)}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t Ad^{(t)}$$
$$C\rho^{t+1} = g^{t+1}$$
$$\beta_t = \frac{(g^{t+1}, \rho^{t+1})}{(g^t, \rho^t)}, \quad d^{t+1} = -\rho^{t+1} + \beta_t d^t.$$

Beim PCG-Verfahren ist in jedem Iterationsschritt ist also ein lineares Gleichungssystem mit der Koeffizientenmatrix $C = KK^T$ zu lösen. Dies bedingt, daß K etwa eine Dreiecksmatrix sein sollte, so daß die Lösung von $C\rho^t = g^t$ durch einfaches Vorwärts- und Rückwärtseinsetzen erfolgen kann.

Beispiel 6.2: Wir geben einige einfache Beispiele von Vorkonditionierungen für das CG-Verfahren.

a) Skalierung: Mit der üblichen Zerlegung $A = D + L + R,\ R = L^T$ wird gesetzt:

$$C = D\,,\ K = D^{1/2} \quad \text{Skalierungsmatrix}$$
$$\tilde{A} = D^{-1/2}AD^{-1/2} \Rightarrow \tilde{a}_{ii} = 1 \quad (1 \le i \le n).$$

Die Skalierung bewirkt, daß die Elemente von A auf etwa gleiche Größenordnung gebracht werden. Dies reduziert die Kondition, denn es gilt folgender Satz: Der kleinste (größte) Eigenwert einer symmetrischen, positiv definiten Matrix ist höchstens (mindestens) so groß wie das kleinste (größte) Diagonalelement, und die Kondition der Matrix ist mindestens so groß wie der Quotient aus dem größten und dem kleinsten Diagonalelement.

b) SSOR-Vorkonditionierung: Mit einem Parameter ω wird gesetzt

$$C = (D + \omega L)D^{-1}(D + \omega R) = (\underbrace{D^{1/2} + \omega LD^{-1/2}}_{K})(\underbrace{D^{1/2} + \omega D^{-1/2}R}_{K^T}).$$

Offenbar besitzt die Dreiecksmatrix K dieselbe schwache Besetzung wie A. Pro Iterationsschritt erfordert das so vorkonditionierte Verfahren etwa doppelt so viel Aufwand wie das einfache Verfahren. Dagegen gilt für die Modellmatrix (vgl. Abschnitt 6.3) bei optimaler Wahl des Parameters ω (i.a. nicht leicht zu bestimmen!)

$$\kappa(\tilde{A}) = \sqrt{\kappa(A)}.$$

c) ICCG-Vorkonditionierung (Incomplete Cholesky Conjugate Gradient): Die symmetrische, positiv definite Matrix A besitzt eine Cholesky-Zerlegung $A = LL^T$ mit einer unteren Dreiecksmatrix

$$L = \begin{bmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{bmatrix}.$$

Die Elemente von L sind bestimmt durch die Rekursionsformeln

$$l_{ii} = [\,a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\,]^{1/2}\,, \quad i = 1, \ldots, n\,,$$
$$l_{ji} = [a_{ji} - \sum_{k=1}^{i-1} l_{jk}l_{ik}]/l_{ii}\,, \quad j = i+1, \ldots, n.$$

Die Matrix L hat i.a. innerhalb der Hülle von A von Null verschiedene Elemente, erfordert also in der Regel weit mehr Speicherplatz als A selbst. Dies wird jedoch dadurch ausgeglichen, daß man nur eine "unvollständige Cholesky-Zerlegung" vornimmt, d. h.: Im Cholesky-Algorithmus werden einige der l_{ji} willkürlich Null gesetzt, z.B.:

$$\tilde{l}_{ji} = 0\,, \quad \text{wenn} \quad a_{ji} = 0. \tag{6.2.65}$$

Dies ergibt dann eine Zerlegung

$$A = \tilde{L}\tilde{L}^T + E \tag{6.2.66}$$

mit einer unteren Dreiecksmatrix $\tilde{L} = (\tilde{l}_{ij})_{i,j=1,\ldots,n}$, welche eine ähnliche (dünne) Besetzungsstruktur wie A besitzt. Man spricht von der $ICCG(0)$-Variante, wenn (6.2.65) gefordert wird. Werden im Fall einer Bandmatrix A weitere p Nebendiagonalen mit von Null verschiedenen Elementen in $\tilde{L}$ hinzugefügt bzw. weggestrichen, so nennt man dies die ICCG(+p) bzw. ICCG(-p)-Variante.

Zur Vorkonditionierung verwendet die ICCG-Methode die Matrix

$$C = KK^T = \tilde{L}\tilde{L}^T. \tag{6.2.67}$$

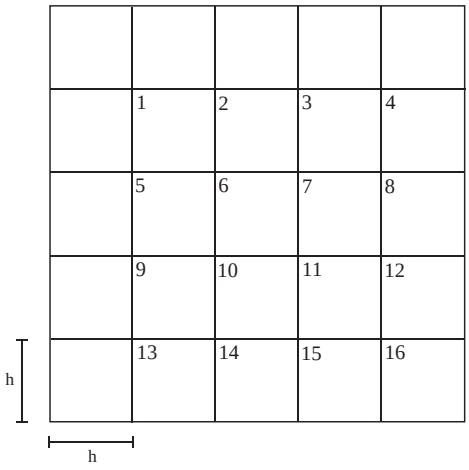
Obwohl keine strenge theoretische Begründung für den Erfolg dieses Ansatzes vorliegt, so zeigen doch numerische Tests an Modellproblemen, welchen Einfluß diese Konditionierung auf die Verteilung der Eigenwerte der Matrix $\tilde{A}$ hat. Zwar wird die Konditionszahl $\kappa(\tilde{A})$ nicht deutlich kleiner als $\kappa(A)$, doch die Eigenwerte von $\tilde{A}$ häufen sich im Gegensatz zu denen von A stark bei $\lambda = 1$. Dies bewirkt, wie eine feinere Analyse in (6.2.55) zeigt, eine deutliche Beschleunigung der Konvergenz.

6.3 Ein Modellproblem

Wir wollen im folgenden die Leistungsfähigkeit der bisher untersuchten Verfahren zur Lösung linearer Gleichungssysteme anhand eines Modellproblems vergleichen. Dazu betrachten wir zunächst das sog. "1. Randwertproblem des Laplace[6]-Operators"

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) &= f(x,y) \quad \text{für } (x,y) \in Q \\ u(x,y) &= 0 \quad \text{für } (x,y) \in \partial Q, \end{aligned} \tag{6.3.68}$$

auf dem Einheitsquadrat $Q = (0,1) \times (0,1) \subset \mathbb{R}^2$. Die Lösung u beschreibt z.B. die Auslenkung einer (idealisierten) elastischen Membran, die über dem Gebiet $\mathbb{R}$ horizontal gespannt und mit einer Kraftdichte f vertikal belastet wird. Eine Lösung $u(x,y)$ ist i.allg. nicht geschlossen angebbar, so daß man sich numerisch eine Näherungslösung verschafft. Dazu wird zunächst das Gebiet Q mit einem Quadratgitter überdeckt.



$h = \frac{1}{m+1}$ Gitterweite

$n = m^2$ Anzahl der gesuchten Knotenwerte

Abbildung 6.4: *Zur Diskretisierung des Modellproblems*

Die "inneren" Gitterpunkte werden zeilenweise durchnumeriert. Ersetzt man dann in der obigen Differentialgleichung die 2. Ableitungen durch die entsprechenden zentralen Differenzenquotienten 2. Ordnung und fordert die Gleichung nur in den inneren Gitterpunkten, so erhält man

$$-h^{-2}\big\{u(x+h,y) - 2u(x,y) + u(x-h,y) + u(x,y+h) - 2u(x,y) + u(x,y-h)\big\} \cong f(x,y)$$

Durch Berücksichtigung der Randbedingung $u(x,y) = 0$ für $(x,y) \in \partial Q$ ist dies äquiva-

[6]Pierre Simon Marquis de Laplace (1749-1827): französischer Mathematiker und Astronom; Prof. in Paris; begründete u.a. die Wahrscheinlichkeitsrechnung.

lent zu einem linearen Gleichungssystem

$$Ax = b \tag{6.3.69}$$

für den Vektor $x \in \mathbb{R}^n$ der unbekannten Knotenwerte

$$x_i \sim u(P_i)\,, \quad P_i \quad \text{Gitterpunkt.}$$

Die Matrix hat die schon bekannte Gestalt

$$A = \left.\begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} n \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m$$

mit der $m \times m$-Einheitsmatrix I. Die rechte Seite ist

$$b = h^2(f(P_1), \ldots, f(P_n))^T.$$

Die Matrix A ist

- eine dünn besetzte Bandmatrix mit der Bandbreite $2m + 1$;
- symmetrisch, irreduzibel;
- schwach diagonal dominant und positiv definit;
- derartig, daß die Theorie für das SOR-Verfahren anwendbar ist.

Die Eigenwerte und zugehörigen Eigenvektoren von A lassen sich explizit angeben. Für $k, l = 1, \ldots, m$ ergibt sich $(h = 1/(m+1))$:

$$\begin{aligned} \lambda_{kl} &= 4 - 2\,(\cos(kh\pi) + \cos(lh\pi)) \\ w^{kl} &= (\sin(ikh\pi)\ \sin(jlh\pi))_{i,j=1,\ldots,m}. \end{aligned}$$

Also ist (für $h \ll 1$)

$$\begin{aligned} \Lambda := \lambda_{\max} &= 4 - 4\,\cos\,(1-h)\pi \;\approx\; 8 \\ \lambda := \lambda_{\min} &= 4 - 4\,\cos\,(h\pi) = 4 - 4\,(1 - \frac{\pi^2}{2}\,h^2 + O(h^4)) \;\approx\; 2\pi^2 h^2 \end{aligned}$$

und somit

$$\kappa \;:=\; \mathrm{cond}_{\mathrm{nat}}(A) \;\approx\; \frac{4}{\pi^2 h^2} \tag{6.3.70}$$

Die Eigenwerte der Jacobi-Matrix $J = -D^{-1}(L+R)$ sind

$$\mu_{kl} = \tfrac{1}{2}\left(\cos(kh\pi) + \cos(lh\pi)\right) \quad (k,l = 1,\ldots,m)$$

Folglich wird

$$\mu_{\max} = \cos(h\pi) = 1 - \frac{\pi^2}{2}\,h^2 + O(h^4),$$

bzw.

$$\rho := \operatorname{spr}(J) = \mu_{\max} \approx 1 - \frac{\pi^2}{2}\,h^2. \tag{6.3.71}$$

Für die Iterationsmatrizen H_1 und $H_{\omega_{\text{opt}}}$ des Gauß-Seidel-Verfahrens und des optimalen SOR-Verfahrens gilt dann

$$\operatorname{spr}(H_1) = \rho^2 = 1 - \pi^2 h^2 + O(h^4),$$

$$\operatorname{spr}(H_{\omega_{\text{opt}}}) = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \frac{1-\pi h + O(h^2)}{1+\pi h + O(h^2)} = 1 - 2\,\pi h + O(h^2)\,.$$

Wir kommen nun zum Leistungsvergleich der verschiedenen Verfahren. Um den Anfangsfehler $\|x^{(0)} - x\|_2$ durch Anwendung der Fixpunktiterationen um den Faktor $\varepsilon \ll 1$ zu reduzieren, sind etwa

$$T(\varepsilon) \approx \frac{\ln(\varepsilon)}{\ln \operatorname{spr}(B)}\,, \quad B = I - C^{-1}A \quad \text{Iterationsmatrix},$$

Iterationsschritte erforderlich. Es ergibt sich somit

$$\begin{aligned}
T_J(\varepsilon) &\sim -\frac{\ln(1/\varepsilon)}{\ln(1-\frac{\pi^2}{2}h^2)} \sim 2\frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{2}{\pi^2} n\, \ln(1/\varepsilon),\\
T_{GS}(\varepsilon) &\sim -\frac{\ln(1/\varepsilon)}{\ln(1-\pi^2 h^2)} \sim \frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{1}{\pi^2}\, n\, \ln(1/\varepsilon),\\
T_{SOR}(\varepsilon) &\sim -\frac{\ln(1/\varepsilon)}{\ln(1-2\pi h)} \sim \frac{\ln(1/\varepsilon)}{2\pi h} = \frac{1}{2\pi}\sqrt{n}\, \ln(1/\varepsilon).
\end{aligned}$$

Das Gradientenverfahren und das CG-Verfahren benötigen zur Reduzierung des Anfangsfehlers $\|x^0 - x\|_2$ um den Faktor $\varepsilon \ll 1$ die folgenden Iterationszahlen:

$$\begin{aligned}
T_G(\varepsilon) &= \frac{1}{2}\,\kappa\, \ln(2/\varepsilon) \sim \frac{2}{\pi^2 h^2}\, \ln(\frac{1}{\varepsilon}) \sim \frac{2}{\pi^2} n\, \ln(\frac{1}{\varepsilon}),\\
T_{CG}(\varepsilon) &= \frac{1}{2}\,\sqrt{\kappa}\, \ln(2/\varepsilon) \sim \frac{1}{\pi h}\, \ln(\frac{2}{\varepsilon}) \sim \frac{1}{\pi}\sqrt{n}\, \ln(\frac{2}{\varepsilon}).
\end{aligned}$$

Wir sehen, daß das Jacobi-Verfahren und das Gradientenverfahren ungefähr gleich schnell sind. Das CG-Verfahren ist zwar nur halb so schnell wie das "optimale" SOR-Verfahren, erfordert aber nicht die Bestimmung eines Iterationsparameters.

Für die spezielle rechte Seite $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$ ist die exakte Lösung der obigen Randwertaufgabe gerade

$$u(x,y) = \sin(\pi x)\sin(\pi y). \tag{6.3.72}$$

Für den Diskretisierungsfehler der Differenzenapproximation läßt sich folgende Darstellung zeigen

$$\max_{P_i} |u(P_i) - x_i| = \frac{\pi^2}{12}\, h^2 + O(h^4). \tag{6.3.73}$$

Zur Erzielung einer Genauigkeit von $\varepsilon = 10^{-4}$ (vier Stellen) ist also die Gitterweite

$$h \;\sim\; \frac{\sqrt{12}}{\pi}\, 10^{-2} \;\sim\; 10^{-2}$$

erforderlich. Die Anzahl von Unbekannten ist dann $n \sim 10^4$. Für die Spektralradien bzw. Konditionszahlen der betrachteten Iterationsverfahren und für die Anzahl der Iterationsschritte, die zur Erzielung einer Fehlergröße von etwa 10^{-4} erforderlich sind, ergibt sich in diesem Fall $(\ln(1/\varepsilon) \sim 10)$:

$$\begin{array}{rlrll}
\operatorname{spr}(J) & \sim 0,9995 & T_{\mathrm{J}}(\varepsilon) & \sim 20.000 & \\
\operatorname{spr}(H_1) & \sim 0.999 & T_{\mathrm{GS}}(\varepsilon) & \sim 10.000 & \\
\operatorname{spr}(H_{\omega *}) & \sim 0,995 & T_{\mathrm{SOR}}(\varepsilon) & \sim 170 & \\
\operatorname{cond}_{\mathrm{nat}}(A) & \sim 5.000 & T_G(\varepsilon) & \sim 20.000, & T_{\mathrm{CG}}(\varepsilon) \sim 340
\end{array}$$

Zum Vergleich der Effizienz der Iterationsverfahren muß natürlich auch der Aufwand pro Iterationsschritt berücksichtigt werden. Für die Anzahl "OP" der arithmetischen Operationen (1 Multiplikation + 1 Addition) pro Iterationsschritt gilt

$$\begin{aligned}
&\mathrm{OP_J} \;\approx\; \mathrm{OP_{H_1}} \;\approx\; \mathrm{OP_{H_\omega}} \;\approx\; 6\,\mathrm{n}\,, \\
&\mathrm{OP_G} \;\approx\; \mathrm{OP_{CG}} \;\approx\; 10\,\mathrm{n}\,.
\end{aligned}$$

Als Endresultat finden wir, daß zur Bestimmung der Lösung des durch Diskretisierung der Randwertaufgabe (6.3.68) entstehenden $(n \times n)$-Gleichungssystems $Ax = b$ das Jacobi-Verfahren, das Gauß-Seidel-Verfahren und das Gradientenverfahren $O(n^2)$ a. Op. benötigen. Zur Lösung des Gleichungssystems $Ax = b$ mit einem direkten Verfahren würde man das Cholesky-Verfahren verwenden. Bei Berücksichtigung der speziellen Struktur der Modellmatrix erfordert dies $O(n^2) = O(m^2 n)$ a. Op. zur Berechnung der Zerlegung $A = LL^T$ und weitere $O(n^{3/2}) = O(mn)$ a. Op. für Vorwärts- und Rückwärtseinsetzen. Damit scheint das direkte Verfahren z. B. dem Gauß-Seidel-Verfahren überlegen zu sein. Es ist jedoch zu berücksichtigen, daß letzteres nur $O(n)$ Speicherplätze benötigt im Gegensatz zu den $O(n^{3/2}) = O(mn)$ für das Cholesky-Verfahren. In den letzten Jahren wurden sehr effiziente Verfahren zur Lösung von Problemen des obigen Typs entwickelt, die im wesentlichen die n Unbekannten mit $O(n)$ Operationen berechnen.

6.4 Übungsaufgaben

Übung 6.1: Für die Matrizen

$$A_1 = \begin{bmatrix} 2 & -1 & 2 \\ 1 & 2 & -2 \\ 2 & 2 & 2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 5 & 5 & 0 \\ -1 & 5 & 4 \\ 2 & 3 & 8 \end{bmatrix},$$

untersuche man, ob das Jacobi- und das Gauß-Seidel-Verfahren für die Gleichungssysteme $A_i x = b$ $(i = 1, 2)$ konvergiert. (Hinweis: Man wende die Konvergenzkriterien der Vorlesung an bzw. schätze den Spektralradius ab.)

Übung 6.2: Das Gleichungssystem

$$\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

soll mit dem Jacobi- und dem Gauß-Seidel-Verfahren gelöst werden. Wieviele Iterationen sind jeweils ungefähr erforderlich, um den Iterationsfehler $\|x^{(t)} - x\|_2$ um den Faktor 10^{-6} zu reduzieren? (Hinweis: Man verwende die Fehlerabschätzung der Vorlesung.)

Übung 6.3: Zur Lösung des linearen (2×2)-Gleichungssystems

$$\begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} x = b, \quad x, b \in \mathbb{R}^2,$$

sei das folgende Iterationsverfahren angesetzt

$$\begin{bmatrix} 1 & 0 \\ -\omega a & 1 \end{bmatrix} x^t = \begin{bmatrix} 1-\omega & \omega a \\ 0 & 1-\omega \end{bmatrix} x^{t-1} + \omega b, \quad \omega \in \mathbb{R}.$$

a) Für welche $a \in \mathbb{R}$ ist diese Methode mit $\omega = 1$ konvergent?

b) Man bestimme für $a = 0.5$ den Wert

$$\omega \in \{0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4\},$$

für den der Spektralradius der Iterationsmatrix B_ω minimal wird und skizziere den Graphen der Funktion $f(\omega) = \mathrm{spr}(B_\omega)$.

Übung 6.4: Man zeige, daß die in der Vorlesung angegebenen beiden Definitionen der "Irreduzibilität" einer Matrix $A \in \mathbb{R}^{n \times n}$ äquivalent sind.

(Hinweis: Die Definition der "Nichtzerlegbarkeit" des Gleichungssystems in einer Form

$$PAP^T = \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{12} & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{A}_{11} \in \mathbb{R}^{p\times p}, \quad \tilde{A}_{22} \in \mathbb{R}^{q\times q}, \quad n = p+q,$$

läßt sich auch wie folgt ausdrücken: Es gibt keine (nicht-triviale) Partitionierung $\{J, K\}$ von $N_n = \{1, \ldots, n\}$, $J \cup K = N_n$, $J \cap K = \emptyset$, so daß $a_{jk} = 0$ für $j \in J$, $k \in K$.

Übung 6.5: Man betrachte das lineare Gleichungssystem $A_n x = b$ mit der $(n \times n)$-Blockmatrix $(n = m^2,\ h := (m+1)^{-1})$

$$A_n = \begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & B_m \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{m\times m}$$

mit der Einheitsmatrix $I_m \in \mathbb{R}^{m\times m}$ und dem Vektor $b = h^2(1, \ldots, 1)^T \in \mathbb{R}^n$. Man schreibe ein Programm zur Lösung dieses Gleichungssystems mit Hilfe

a) des Jacobi-Verfahrens;

b) des Gauß-Seidel-Verfahrens;

c) des SOR-Verfahrens mit "optimalem" Relaxationsparameter ω_{opt} gemäß der in der Vorlesung angegebenen Theorie:

$$1 < \omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \text{spr}(J)^2}} < 2, \qquad \text{spr}(J) = \cos(h\pi) < 1.$$

Als Startvektoren verwende man jeweils $x^{(0)} = 0$ und als Abbruchkriterium

$$\frac{\|Ax^t - b\|_\infty}{\|x^t\|_\infty} \le 10^{-8} \qquad \text{oder} \qquad t_{\max} \le 20000.$$

Für $m = 2^k$, $k = 1, \ldots, 6$, vergleiche man das Konvergenzverhalten dieser Verfahren. (Das Programm soll möglichst sparsam im Speicherverbrauch sein!)

7 Matrizeneigenwertaufgaben

Im folgenden betrachten wir quadratische Matrizen $A \in \mathbb{K}^{n\times n}$ mit Elementen aus $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$.

Definition 7.1: *Eine Zahl $\lambda \in \mathbb{C}$ ist "Eigenwert" von A, wenn es einen zugehörigen "Eigenvektor" $w \in \mathbb{C}^n$, $w \neq 0$, gibt mit $Aw = \lambda w$. Die Eigenwerte sind gerade die Nullstellen des "charakteristischen Polynoms" χ_A von A*

$$\chi_A(z) = \det(A - zI) = (-1)^n z^n + b_1 z^{n-1} + \ldots + b_n.$$

Es gibt genau n (ihrer Vielfachheit als Nullstelle entsprechend oft gezählte) Eigenwerte, welche sich unabhängig voneinander bestimmen lassen.
Man spricht vom "partiellen Eigenwertproblem", wenn nur einzelne Eigenwerte (etwa der kleinste oder der größte) und gegebenenfalls die zugehörigen Eigenvektoren gesucht sind, und vom "vollständigen Eigenwertproblem", wenn alle Eigenwerte und eventuell zugehörige Eigenvektoren berechnet werden sollen.

Ist ein Eigenwert λ bekannt, so erhält man die zugehörigen Eigenvektoren als Lösung des homogenen Gleichungssystems $(A-\lambda I)\, w = 0$. Umgekehrt bestimmt ein Eigenvektor w eindeutig den zugehörigen Eigenwert etwa durch den sog. "Rayleigh[1]-Quotienten"

$$\lambda = \frac{(Aw, w)_2}{\|w\|_2^2}.$$

Dabei bezeichnen wieder $(\cdot,\cdot)_2$ das euklidische Skalarprodukt und $\|\cdot\|_2$ die zugehörige Vektornorm. Im folgenden stellen wir einige Tatsachen und Definitionen der linearen Algebra zusammen: Das charakteristische Polynom einer Matrix $A \in \mathbb{K}^{n\times n}$ besitzt mit seinen paarweise verschiedenen Nullstellen λ_i (den Eigenwerten von A) die Darstellung

$$\chi_A(z) = \prod_{i=1}^{m} (z - \lambda_i)^{\sigma_i}, \quad \sum_{i=1}^{m} \sigma_i = n.$$

Die Eigenvektoren zum Eigenwert λ_i bilden einen Unterraum von $\mathbb{C}^n$, den sog. "geometrischen Eigenraum" zu λ_i, mit der Dimension $\rho_i = \dim(\mathrm{Kern}(A - \lambda_i I))$. Es heißen σ_i die "'algebraisch" und ρ_i die "geometrische" Vielfachheit von λ_i. Es ist stets $\rho_i \leq \sigma_i$.

Beispiel 7.1: Die Bedeutung der folgenden Matrizen $C_m(\lambda)$ liegt darin, daß aus ihnen die sog. "Jordansche[2] Normalform" einer Matrix A aufgebaut ist (siehe Abschnitt 7.3):

[1] John William Strutt (Lord Rayleigh) (1842-1919): Englischer Mathematiker und Physiker; forschte zunächst als (adliger) Privatgelehrter, 1879-1884 Professor für experimentelle Physik in Cambridge; fundamentale Beiträge zur theoretischen Physik: Streutheorie, Akustik, Elektro-Magnetismus, Gasdynamik.

[2] Marie Ennemond Camille Jordan (1838-1922): Französischer Mathematiker; Prof. in Paris; Beiträge zur Algebra, Gruppentheorie, Analysis und Topologie.

$$C_m(\lambda) = \begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ 0 & & & & \lambda \end{bmatrix} \in \mathbb{K}^{m\times m}, \quad \text{Eigenwerte:} \quad \lambda \in \mathbb{C}$$

$$\chi_{C_m(\lambda)}(z) = (\lambda - z)^m \;\Rightarrow\; \sigma = m, \quad \text{Rang}(C_m(\lambda) - \lambda I) = m-1 \;\Rightarrow\; \rho = 1\,.$$

Der naheliegende Weg zur Berechnung der Eigenwerte einer Matrix $A \in \mathbb{K}^{n\times n}$ wäre die Bestimmung der Koeffizienten ihres charakteristischen Polynoms $p_A(z)$ und die anschließende Berechnung der Nullstellen von $\chi_A(z)$ mit einem geeigneten numerischen Verfahren. Dieses Vorgehen ist jedoch i. Allg. nicht angeraten, da die Nullstellenbestimmung bei Polynomen ein hochgradig schlecht konditioniertes Problem ist, obwohl das ursprüngliche Eigenwertproblem, wie wir sehen werden, meist gut konditioniert ist.

Beispiel 7.2: $A \in \mathbb{R}^{20\times 20}$ symmetrisch mit Eigenwerten $\lambda_j = j$, $j = 1, \ldots, 20$:

$$\chi_A(z) = \prod_{j=1}^{20} (z-j) = z^{20} \underbrace{-210}_{b_1} z^{19} + \ldots + \underbrace{20!}_{b_{20}}$$

Der Koeffizient b_1 sei gestört: $\tilde{b}_1 = -210 + 2^{-23} \sim -210,000000119\ldots$,

$$\text{relativer Fehler} \quad \left|\frac{\tilde{b}_1 - b_1}{b_1}\right| \sim 10^{-10}.$$

Das gestörte Polynom $\tilde{\chi}_A(z)$ hat dann u. a. die Wurzeln: $\lambda_\pm \sim 16,7 \pm 2,8i$.

Über das charakteristische Polynom werden die Eigenwerte nur für sehr einfach strukturierte Matrizen berechnet, bei denen dies möglich ist, ohne die Koeffizienten von $\chi_A(z)$ auszurechnen. Dies sind "Tridiagonalmatrizen" und allgemeiner sog. "Hessenberg[3] -Matrizen".

Tridiagonalmatrix

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & & b_{n-1} \\ & & c_n & a_n \end{bmatrix}$$

Hessenberg-Matrix

$$\begin{bmatrix} a_{11} & \cdots & & a_{1n} \\ a_{21} & \ddots & & \vdots \\ & \ddots & & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{nn} \end{bmatrix}$$

[3]Karl Hessenberg (1904-1959): Deutscher Mathematiker; Dissertation "Die Berechnung der Eigenwerte und Eigenlösungen linearer Gleichungssysteme", TU Darmstadt 1942.

7.1 Konditionierung des Eigenwertproblems

Hilfssatz 7.1 (Stabilität): *Seien $A, B \in \mathbb{K}^{n\times n}$ beliebige Matrizen und $\|\cdot\|$ eine natürliche Matrizennorm. Dann gilt für jeden Eigenwert λ von A, der nicht zugleich auch Eigenwert von B ist, die Beziehung*

$$\|(\lambda I - B)^{-1}(A - B)\| \geq 1. \tag{7.1.1}$$

Beweis: Ist w Eigenvektor zum Eigenwert λ von A, so folgt aus der Identität

$$(A - B)\, w = (\lambda I - B)\, w,$$

wenn λ kein Eigenwert von B ist,

$$(\lambda I - B)^{-1}(A - B)\, w = w.$$

Also ist

$$1 \leq \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|(\lambda I - B)^{-1}(A - B)\, x\|}{\|x\|} = \|(\lambda I - B)^{-1}(A - B)\|,$$

was zu zeigen war. Q.E.D.

Als Folgerung aus Hilfssatz 7.1 erhalten wir zunächst den wichtigen Einschließungssatz von Gerschgorin[4] (1931).

Satz 7.1 (Satz von Gerschgorin): *Alle Eigenwerte einer Matrix $A \in \mathbb{K}^{n\times n}$ liegen in der Vereinigung der sog. "Gerschgorin-Kreise"*

$$K_j := \Big\{ z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\}, \quad j = 1, \ldots, n. \tag{7.1.2}$$

Sind die Mengen $U \equiv \cup_{i=1}^{m} K_{j_i}$ und $V \equiv \overline{\cup_{j=1}^{n} K_j \setminus U}$ disjunkt, so liegen in U genau m und in V genau $n - m$ Eigenwerte von A (mehrfache Eigenwerte ihrer algebraischen Vielfachheiten entsprechend oft gezählt).

Beweis: (i) Wir setzen $B \equiv D = \mathrm{diag}(a_{jj})$ in Hilfssatz 7.1 und nehmen die "maximale Zeilensumme" als natürliche Matrizennorm. Damit folgt dann für $\lambda \neq a_{jj}$:

$$\|(\lambda I - D)^{-1}(A - D)\|_\infty = \max_{j=1,\ldots,n} \frac{1}{|\lambda - a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \geq 1,$$

d. h.: λ liegt in einem der Gerschgorin-Kreise.

[4]Semyon Aranovich Gershgorin (1901-1933): Russischer Mathematiker; seit 1930 Professor in Leningrad (St. Petersburg); arbeitete über Algebra, Funktionenthorie, Differentialgleichungen und Numerik.

(ii) Zum Beweis der zweiten Behauptung setzen wir $A_t \equiv D + t(A - D)$. Offenbar liegen genau m Eigenwerte von $A_0 = D$ in U und $n - m$ Eigenwerte in V. Dasselbe folgt dann auch für $A_1 = A$, da die Eigenwerte von A_t stetige Funktionen von t sind. Q.E.D.

Der Satz von Gerschgorin liefert sehr viel genauere Informationen über die Lage der Eigenwerte λ von A als die uns schon bekannte grobe Abschätzung $|\lambda| \leq \|A\|_\infty$. Die Matrizen A und A^T haben dieselben Eigenwerte. Durch Anwendung von Satz 7.1 auf A^T erhält man oft eine weitere Verschärfung der Abschätzung für die Eigenwerte.

Beispiel 7.3:

$$A = \begin{bmatrix} 1 & 0.1 & -0.2 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{bmatrix} \quad \|A\|_\infty = 3.2\,, \quad \|A\|_1 = 3.6.$$

$$\begin{aligned} K_1 &= \{z \in \mathbb{C} : |z-1| \leq 0.3\} & K_1^T &= \{z \in \mathbb{C} : |z-1| \leq 0.2\} \\ K_2 &= \{z \in \mathbb{C} : |z-2| \leq 0.4\} & K_2^T &= \{z \in \mathbb{C} : |z-2| \leq 0.1\} \\ K_3 &= \{z \in \mathbb{C} : |z-3| \leq 0.2\} & K_3^T &= \{z \in \mathbb{C} : |z-3| \leq 0.6\} \end{aligned}$$

$$|\lambda_1 - 1| \leq 0.2\,, \quad |\lambda_2 - 2| \leq 0.1\,, \quad |\lambda_3 - 3| \leq 0.2$$

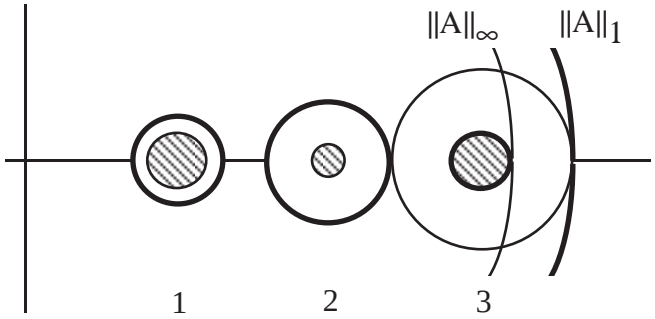


Abbildung 7.1: Gerschgorin-Kreise

Weiter erhalten wir mit Hilfssatz 7.1 den folgenden Störungssatz:

Satz 7.2 (Stabilitätssatz): *Sei $A \in \mathbb{K}^{n\times n}$ eine Matrix, zu der es n linear unabhängige Eigenvektoren $\{w_1, \ldots, w_n\}$ gibt, und sei $B \in \mathbb{K}^{n\times n}$ eine zweite Matrix. Dann gibt es zu jedem Eigenwert $\lambda(B)$ von B einen Eigenwert $\lambda(A)$ von A, so daß mit der Matrix $W = [w_1, \ldots, w_n]$ gilt:*

$$|\lambda(A) - \lambda(B)| \;\leq\; \mathrm{cond}_2(\mathrm{W})\, \|\mathrm{A} - \mathrm{B}\|_2\,. \tag{7.1.3}$$

Beweis: Die Eigenwertgleichungen $Aw^{(i)} = \lambda_i(A)w^{(i)}$ lassen sich in der Form $AW = W \cdot \mathrm{diag}(\lambda_i(A))$ schreiben mit der regulären Matrix $W = [w_1, \ldots, w_n]$. Also ist

$$A = W\,\mathrm{diag}(\lambda_i(A))\,W^{-1},$$

d. h.: A ist "ähnlich" zu der Diagonalmatrix $\Lambda = \mathrm{diag}(\lambda_i(A))$. Wenn $\lambda = \lambda(B)$ nun kein Eigenwert von A ist, so gilt

$$\begin{aligned}\|(\lambda I - A)^{-1}\|_2 &= \|W(\lambda I - \Lambda)^{-1}W^{-1}\|_2 \\ &\leq \|W^{-1}\|_2\|W\|_2\|(\lambda I - \Lambda)^{-1}\|_2 \\ &= \mathrm{cond}_2(W) \max_{i=1,\ldots,n} |\lambda - \lambda_i(A)|^{-1}.\end{aligned}$$

Hilfssatz 7.1 ergibt dann die Behauptung. Q.E.D.

Für hermitesche Matrizen $A \in \mathbb{K}^{n\times n}$ existiert eine Orthonormalbasis des $\mathbb{K}^n$ von Eigenvektoren, so daß die Matrix W in Satz 7.2 als unitär angenommen werden kann: $W\bar{W}^T = I$. In diesem Fall ist

$$\mathrm{cond}_2(W) = \|\bar{W}^T\|_2\,\|W\|_2 = 1. \tag{7.1.4}$$

Wir fassen die gefundenen Ergebnisse zusammen.

Regel 7.1.1: *Das Eigenwertproblem hermitescher Matrizen ist gut konditioniert, während das allgemeine Eigenwertproblem je nach Größe von* $\mathrm{cond}_2(W)$ *beliebig schlecht konditioniert sein kann.*

7.2 Iterative Verfahren

Im folgenden betrachten wir ein iteratives Verfahren zur Lösung des partiellen Eigenwertproblems einer Matrix $A \in \mathbb{K}^{n\times n}$.

Definition 7.2: *Die "Potenzmethode" nach v. Mises*[5] *erzeugt ausgehend von einem Startvektor* $z^0 \in \mathbb{C}^n$ *mit* $\|z^0\| = 1$ *eine Folge von Iterierten* $z^t \in \mathbb{C}^n$, $t = 1, 2, \ldots,$ *durch*

$$\tilde{z}^t = Az^{t-1}, \quad z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|}. \tag{7.2.5}$$

Für einen beliebigen Index $k \in \{1, \ldots, n\}$ *(etwa den der "maximalen" Komponente von* z^t*) wird gesetzt*

$$\lambda^{(t)} := \frac{(Az^t)_k}{z^t_k}. \tag{7.2.6}$$

Zur Normierung wird üblicherweise $\|\cdot\| = \|\cdot\|_\infty$ oder $\|\cdot\| = \|\cdot\|_2$ verwendet. Zur Analyse dieses Verfahrens nehmen wir an, daß die Matrix A "diagonalisierbar", d. h. ähnlich zu einer Diagonalmatrix, ist. Dies ist, äquivalent dazu, daß A eine Basis von Eigenvektoren $\{w_1, \ldots, w_n\}$ besitzt (Übungsaufgabe). Diese Eigenvektoren w_i seien normiert. Wir nehmen weiter an, daß $z^{(0)}$ eine nichttriviale Komponente bzgl. w_n besitzt. Dies ist keine wesentliche Einschränkung, da aufgrund des unvermeidbaren Rundungsfehlers dieser Fall im Verlauf der Iteration sicher einmal eintritt

Satz 7.3 (Potenzmethode): *Die Matrix* A *sei diagonalisierbar und ihr betragsgrößter Eigenwert sei separiert von den anderen Eigenwerten, d. h.:* $|\lambda_n| > |\lambda_i|$, $i \neq n$. *Ferner habe der Startvektor* $z^{(0)}$ *eine nichttriviale Komponente bzgl. des zugehörigen Eigenvektors* w_n. *Dann gibt es Zahlen* $\sigma_t \in \mathbb{C}$, $|\sigma_t| = 1$, *so daß*

$$\|z^t - \sigma_t w_n\| \to 0 \quad (t \to \infty), \tag{7.2.7}$$

und es gilt

$$\lambda^{(t)} - \lambda_n = O\Big(\Big|\frac{\lambda_{n-1}}{\lambda_n}\Big|^t\Big) \quad (t \to \infty). \tag{7.2.8}$$

Beweis: Sei $z^0 = \sum_{i=1}^n \alpha_i w_i$ die Basisdarstellung der Startvektoren. Für die Iterierten z^t gilt

$$z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|} = \frac{Az^{t-1}}{\|Az^{t-1}\|} = \frac{A\tilde{z}^{t-1}}{\|\tilde{z}^{t-1}\|}\,\frac{\|\tilde{z}^{t-1}\|}{\|A\tilde{z}^{t-1}\|} = \ldots = \frac{A^t z^0}{\|A^t z^0\|}.$$

[5] *Richard von Mises (1883-1953): Östereichischer Mathematiker; Professor für Angewandte Mathematik und Mechanik in Straßburg (1909-1918), in Dresden und dann Gründer des neuen Instituts für Angewandte Mathematik in Berlin (1919-1933), danach Emigration in die Türkei (Istambul) und schließlich in die USA (1938); Professor an der Harward University; wichtige Beiträge zur theoretischen Strömungsmechanik (Einführung des "Spannungstensors"), Aerodynamik, Numerik, Statistik und Wahrscheinlichkeitstheorie.*

Ferner ist

$$A^t z^0 = \sum_{i=1}^{n} \alpha_i \lambda_i^t w_i = \lambda_n^t \alpha_n \Big\{ w_n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n} \Big(\frac{\lambda_i}{\lambda_n} \Big)^t w_i \Big\}$$

und folglich wegen $|\lambda_i/\lambda_n| < 1$, $i = 1, \ldots, n-1$,

$$A^t z^0 = \lambda_n^t \alpha_n \{ w_n + o(1) \} \quad (t \to \infty).$$

Dies ergibt

$$z^t = \frac{\lambda_n^t \alpha_n \{ w_n + o(1) \}}{|\lambda_n^t \alpha_n| \, \| w_n + o(1) \|} = \underbrace{\frac{\lambda_n^t \alpha_n}{|\lambda_n^t \alpha_n|}}_{=:\ \sigma_t} w_n + o(1).$$

Die Iterierten z^t konvergieren also gegen $\operatorname{span}\{w_n\}$. Weiter gilt (wegen $\alpha_n \neq 0$)

$$\begin{aligned} \lambda^{(t)} &= \frac{(Az^t)_k}{z_k^t} = \frac{(A^{t+1} z^0)_k}{\|A^t z^0\|} \, \frac{\|A^t z^0\|}{(A^t z^0)_k} \\ &= \frac{\lambda_n^{t+1} \big\{ \alpha_n w_{n,k} + \sum_{i=1}^{n-1} \alpha_i \big(\frac{\lambda_i}{\lambda_n} \big)^{t+1} w_{i,k} \big\}}{\lambda_n^t \big\{ \alpha_n w_{n,k} + \sum_{i=1}^{n-1} \alpha_i \big(\frac{\lambda_i}{\lambda_n} \big)^{t} w_{i,k} \big\}} = \lambda_n + O\Big(\Big| \frac{\lambda_{n-1}}{\lambda_n} \Big|^t \Big) \quad (t \to \infty). \end{aligned}$$

Q.E.D.

Die Konvergenz der Potenzmethode ist um so besser, je mehr der betragsgrößte Eigenwert λ_n von den übrigen betragsmäßig separiert ist. Der Konvergenzbeweis läßt sich verallgemeinern für diagonalisierbare Matrizen, bei denen der betragsgrößte Eigenwert zwar mehrfach sein kann, aber $|\lambda_n| = |\lambda_i|$ notwendig $\lambda_n = \lambda_i$ impliziert. Für noch allgemeinere Matrizen ist die Konvergenz nicht mehr gesichert.

Bei hermiteschen Matrizen erhält man im Rahmen der Potenzmethode bessere Eigenwertnäherungen mit Hilfe des Rayleigh-Quotienten:

$$\lambda^{(t)} := (Az^t, z^t), \quad \|z^t\| = 1. \tag{7.2.9}$$

In diesem Fall kann $\{w_1, \ldots, w_n\}$ als Orthonormalsystem gewählt werden, so daß gilt:

$$\begin{aligned} \lambda^{(t)} &= \frac{(A^{t+1} z^0, \, A^t z^0)}{\|A^t z^0\|^2} = \frac{\sum_{i=1}^{n} |\alpha_i|^2 \lambda_i^{2t+1}}{\sum_{i=1}^{n} |\alpha_i|^2 \lambda_i^{2t}} \\ &= \frac{\lambda_n^{2t+1} \big\{ |\alpha_n|^2 + \sum_{i=1}^{n-1} |\alpha_i|^2 \big(\frac{\lambda_i}{\lambda_n} \big)^{2t+1} \big\}}{\lambda_n^{2t} \, \big\{ |\alpha_n|^2 + \sum_{i=1}^{n} |\alpha_i|^2 \, (\lambda_i | \lambda_n)^{2t} \big\}} = \lambda_n + O\Big(\Big| \frac{\lambda_{n-1}}{\lambda_n} \Big|^{2t} \Big). \end{aligned}$$

Die Konvergenz der Eigenwertnäherungen ist hier also doppelt so schnell wie im nicht hermiteschen Fall.

Für die praktische Rechnung ist die einfache Potenzmethode nur bedingt brauchbar, da sie schlecht konvergiert, wenn $|\lambda_{n-1}/\lambda_n| \sim 1$ ist, und auch nur den betragsgrößten Eigenwert liefert.

Eine Weiterentwicklung der Potenzmethode ist die sog. "Inverse Iteration" nach Wielandt[6] Hier wird davon ausgegangen, daß man bereits eine gute Näherung $\tilde{\lambda}$ für einen Eigenwert λ_k der Matrix A kennt (durch Einschließungssätze oder andere Verfahren), so daß gilt:

$$|\lambda_k - \tilde{\lambda}| \ll |\lambda_i - \tilde{\lambda}|, \quad i = 1, \ldots, n\,, \quad i \neq k. \tag{7.2.10}$$

Im Falle $\tilde{\lambda} \neq \lambda_k$ hat $(A - \tilde{\lambda}I)^{-1}$ die Eigenwerte $\mu_i = (\lambda_i - \tilde{\lambda})^{-1}$, $i = 1, \ldots, n$, und es gilt

$$\Big|\frac{1}{\lambda_k - \tilde{\lambda}}\Big| \gg \Big|\frac{1}{\lambda_i - \tilde{\lambda}}\Big|, \quad i = 1, \ldots, n\,, \quad i \neq k. \tag{7.2.11}$$

Definition 7.3: *Die "inverse Iteration" besteht in der Anwendung der Potenzmethode auf die Matrix $(A - \tilde{\lambda}I)^{-1}$ mit einer a priori Schätzung $\tilde{\lambda}$ zum gesuchten Eigenwert λ_k. Ausgehend von einem Startvektor z^0 werden Iterierte z^t bestimmt als Lösungen der Gleichungssysteme*

$$(A - \tilde{\lambda}I)z^t = z^{t-1}\,, \quad z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|}\,, \quad t = 1, 2, \ldots. \tag{7.2.12}$$

Die zugehörige Eigenwertnäherung wird bestimmt durch

$$\mu^{(t)} := \frac{z_k^t}{((A - \tilde{\lambda}I)z^t)_k}, \quad ((A - \tilde{\lambda}I)z^t)_k \neq 0, \tag{7.2.13}$$

oder im symmetrischen Fall wieder mit Hilfe des Rayleigh-Quotienten.

Aufgrund der Aussagen über die Potenzmethode liefert die inverse Iteration also für eine diagonalisierbare Matrix jeden Eigenwert, zu dem bereits eine hinreichend gute Näherung bekannt ist.

Beispiel 7.4: Wir wollen die eben beschriebenen Verfahren auf die Modellmatrix aus Abschnitt 6.3 anwenden. Zur Bestimmung der Schwingungsformen und Frequenzen einer über dem Gebiet Ω gespannten Membran (Trommel) hat man das Eigenwertproblem des Laplace-Operators

$$\begin{aligned} -\frac{\partial^2 w}{\partial x^2}(x,y) - \frac{\partial^2 w}{\partial y^2}(x,y) &= \mu w(x,y) \quad \text{für} \quad (x,y) \in \Omega, \\ w(x,y) &= 0 \quad \text{für} \quad (x,y) \in \partial\,\Omega, \end{aligned} \tag{7.2.14}$$

zu lösen. Man kann zeigen, daß dieses Problem eine abzählbar unendliche Folge von reellen, positiven Eigenwerten besitzt. Der kleinste von diesen $\mu_{\min} > 0$, beschreibt ge-

[6] Helmut Wielandt (1910-2001): Deutscher Mathematiker; Professor in Mainz (1946-1951) und Tübingen (1951-1977); Beiträge zu Gruppentheorie, Lineare Algebra und Matrix-Theorie.

rade den Grundton der Trommel und die zugehörige Eigenfunktion $w_{\min}$ die zugehörige Grundschwingungsform. Die Diskretisierung dieses Problems mit Hilfe des 5-Punkte-Differenzensterns führt auf eine Matrizeneigenwertaufgabe

$$Az = \lambda z\,, \quad \lambda = h^2\mu \tag{7.2.15}$$

mit derselben Blocktridiagonalmatrix A wie beim Randwertproblem. Unter Verwendung der Bezeichnungen des Abschnitts 6.3 lassen sich deren Eigenwerte explizit angeben durch

$$\lambda_{kl} = 4 - 2(\cos(kh\pi) + \cos(lh\pi))\,, \quad k,l = 1,\ldots,m.$$

Von Interesse ist nun offenbar insbesondere der kleinste Eigenwert $\lambda_{\min}$ von A, der mit $h^{-2}\lambda_{\min} \approx \mu_{\min}$ eine Approximation zum kleinsten Eigenwert des Ausgangsproblems (3.2.9) liefert. Für $\lambda_{\min}$ und den darauffolgenden Eigenwert $\lambda^* > \lambda_{\min}$ gilt offenbar

$$\begin{aligned} \lambda_{\min} &= 4 - 4\cos(h\pi) = 2\pi^2h^2 + O(h^4) \\ \lambda^* &= 4 - 2(\cos(2h\pi) + \cos(h\pi)) = 5\pi^2h^2 + O(h^4). \end{aligned}$$

Zur Berechnung von $\lambda_{\min}$ könnte die inverse Iteration (mit Shift $\lambda = 0$) verwendet werden. Dies erfordert in jedem Iterationsschritt die Lösung eines Gleichungssystems

$$Az^t = z^{t-1} \tag{7.2.16}$$

Für die zugehörige Eigenwertnäherung

$$\lambda^t = \frac{(Az^t, z^t)}{\|z^t\|^2} \tag{7.2.17}$$

gilt dann mit dem auf $\lambda_{\min}$ folgenden Eigenwert $\lambda^* > \lambda_{\min}$ die Konvergenzaussage

$$|\lambda^t - \lambda_{\min}| \approx (\lambda_{min}/\lambda^*)^{2t} \approx (0,4)^{2t}, \tag{7.2.18}$$

d. h.: Die Konvergenzgeschwindigkeit ist unabhängig von der Gitterweite h bzw. der Dimension $n = m^2 \approx h^{-2}$ der Matrix A. Allerdings müßte zur Erzielung einer vorgegebenen Genauigkeit für die Approximation $\mu^{(t)} = h^{-2}\lambda^{(t)}$ die Toleranz mit h^{-2} skaliert werden, was wiederum eine logarithmische h-Abhängigkeit der erforderlichen Iterationszahl einführt.

$$t(\varepsilon) \approx \frac{\log(\varepsilon h^2)}{\log(0,4)} \approx \log(n). \tag{7.2.19}$$

Dieser Weg zur Berechnung von $\mu_{\min}$ wäre damit sehr aufwendig, wenn die Lösung der Probleme (3.2.11) etwa mit Hilfe des PCG-Verfahrens auf maximale Genauigkeit erfolgen würde. Zur Aufwandsreduktion könnte man die Abbruchgenauigkeit der CG-Iteration am Anfang niedrig ansetzen und sie erst im Laufe der äußeren Iteration sukzessive erhöhen.

7.3 Reduktionsmethoden

Wir rekapitulieren einige für das Folgende wichtigen Eigenschaften "ähnlicher" Matrizen.

Definition 7.4: *Zwei Matrizen $A, B \in \mathbb{C}^{n\times n}$ heißen "ähnlich", in Symbolen $A \sim B$, wenn es eine reguläre Matrix $T \in \mathbb{C}^{n\times n}$ gibt, so daß gilt:*

$$A = T^{-1}BT. \tag{7.3.20}$$

Wegen

$$\begin{aligned} \det(A - zI) &= \det(T^{-1}[B - zI]T) \\ &= \det(T^{-1})\,\det(B - zI)\,\det(T) = \det(B - zI) \end{aligned} \tag{7.3.21}$$

haben ähnliche Matrizen A, B dasselbe charakteristische Polynom und folglich dieselben Eigenwerte. Ist λ ein Eigenwert von A mit Eigenvektor w, so ist wegen

$$Aw = T^{-1}BTw = \lambda w$$

offenbar Tw ein Eigenvektor von B zum Eigenwert λ. Algebraische und geometrische Vielfachheiten von Eigenwerten ähnlicher Matrizen stimmen also überein. Es liegt nun nahe, eine gegebene Matrix A durch eine Folge von Ähnlichkeitstransformationen

$$A = A^{(0)} = T_1^{-1}A^{(1)}T_1 = Q\ldots = T_i^{-1}A^{(i)}T_i = \ldots \tag{7.3.22}$$

auf eine Form zu bringen, für welche Eigenwerte und zugehörige Eigenvektoren leicht zu bestimmen sind. Dies ist die Vorgehensweise der sog. "Reduktionsmethoden". Wir geben dazu (ohne Beweis) eine Reihe von grundlegenden Resultaten an:

Satz 7.4 (Jordansche Normalform): *Die Matrix $A \in \mathbb{C}^{n\times n}$ habe die (paarweise verschiedenen) Eigenwerte λ_i, $i = 1,\ldots,m$, mit Vielfachheiten σ_i und ρ_i. Dann gibt es Zahlen $r_k^{(i)} \in \mathbb{N}$ $k = 1,\ldots,\rho_i$, $\sigma_i = r_1^{(i)} + \ldots + r_{\rho_i}^{(i)}$, so daß A ähnlich ist zu einer Jordanschen Normalform*

$$J_A = \begin{bmatrix} C_{r_1^{(1)}}(\lambda_1) & & & & & & \\ & \ddots & & & & 0 & \\ & & C_{r_{\rho_1}^{(1)}}(\lambda_1) & & & & \\ & & & \ddots & & & \\ & & & & C_{r_1^{(m)}}(\lambda_m) & & \\ & 0 & & & & \ddots & \\ & & & & & & C_{r_{\rho_m}^{(m)}}(\lambda_m) \end{bmatrix}.$$

Die Zahlen $r_k^{(i)}$ sind dabei bis auf die Reihenfolge eindeutig bestimmt.

Läßt man als Ähnlichkeitstransformation nur solche mit unitären Matrizen zu, so gilt der folgende Satz von Schur[7]

Satz 7.5 (Schursche Normalform): *Die Matrix $A \in \mathbb{C}^{n\times n}$ habe die (ihrer algebraischen Vielfachheiten entsprechend oft gezählten) Eigenwerte λ_i, $i = 1, \ldots, n$. Dann gibt es eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ mit*

$$\bar{U}^T A U = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}. \tag{7.3.23}$$

Ist $A \in \mathbb{C}^{n\times n}$ hermitesch, $A^T = \bar{A}$, so ist auch $\bar{U}^T A U$ hermitesch. Folglich sind hermitesche Matrizen $A \in \mathbb{C}^{n\times n}$ stets "unitär-ähnlich" zu einer Diagonalmatrix, $\bar{U}^T A U = diag(\lambda_i)$, d. h. "diagonalisierbar".

Hilfssatz 7.2 (Diagonalisierbarkeit): *Für eine Matrix $A \in \mathbb{C}^{n\times n}$ sind die folgenden Aussagen äquivalent:*

(i) A ist diagonalisierbar.

(ii) Es gibt eine Basis des $\mathbb{C}^n$ aus Eigenvektoren von A.

(iii) Für alle Eigenwerte von A ist die algebraische gleich der geometrischen Vielfachheit.

Beweis: Übungsaufgabe. Q.E.D.

Die direkte Transformation einer gegebenen Matrix auf Normalform ist i. Allg. nur bei vorheriger Kenntnis der Eigenvektoren in *endlich* vielen Schritten möglich. Daher transformiert man in der Praxis eine Matrix zunächst nur in eine einfachere Form (z. B. Hessenberg-Form) und wendet auf diese dann andere Verfahren an:

$$A = A^{(0)} \to A^{(1)} = T_1^{-1} A^{(0)} T_1 \to \ldots A^{(m)} = T_m^{-1} A^{(m-1)} T_m.$$

Die Transformationsmatrizen T_i sollten dabei explizit durch die Elemente von $A^{(i-1)}$ angebbar sein. Ferner sollte das Eigenwertproblem der Matrix $A^{(i)} = T_i^{-1} A^{(i-1)} T_i$ nicht wesentlich schlechter konditioniert sein als das von $A^{(i-1)}$.

Sei $\|\cdot\|$ eine natürliche Matrizennorm zur Vektornorm $\|\cdot\|$ auf $\mathbb{C}^n$. Für zwei ähnliche Matrizen $B \sim A$ gilt dann

$$B = T^{-1} A T, \quad B + \delta B = T^{-1}(A + \delta A) T, \quad \delta A = T \delta B T^{-1},$$

[7] Issai Schur (1875-1941): Russisch-Deutscher Mathematiker; Professor in Bonn (1911-1916) und in Berlin (1916-1935), wo er eine berühmte mathematische Schule begründete; wegen seiner jüdischen Herkunft verfolgt emmigrierte er 1939 nach Palestina; fundamentale Arbeiten insbesondere zur Darstellungstheorie von Gruppen und zur Zahlentheorie.

bzw.

$$\|B\| \le \mathrm{cond}(T)\,\|A\|\,, \quad \|\delta A\| \le \mathrm{cond}(T)\,\|\delta B\|.$$

Folglich ist

$$\frac{\|\delta A\|}{\|A\|} \le \mathrm{cond}(T)^2\,\frac{\|\delta B\|}{\|B\|}. \tag{7.3.24}$$

Für große $\mathrm{cond}(T) \gg 1$ werden also kleine Änderungen in B die Eigenwerte unter Umständen wesentlich stärker verfälschen als solche in A. Um die Gutartigkeit der Reduktionsmethode zu garantieren, hat man wegen

$$\mathrm{cond}(T) = \mathrm{cond}(T_1 \dots T_m) \le \mathrm{cond}(T_1) \cdot \ \dots \ \cdot \mathrm{cond}(T_m)$$

die Matrizen T_i so zu wählen, daß $\mathrm{cond}(T_i)$ nicht zu groß wird. Dies ist insbesondere bei den folgenden drei Typen von Transformationen der Fall:

a) Drehungen:

$$T = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos(\varphi) & & & -\sin(\varphi) & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin(\varphi) & & & \cos(\varphi) & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} \implies \mathrm{cond}_{\mathrm{nat}}(T) = 1.$$

b) Spiegelungen (Householder-Transformationen):

$$T = I - 2u\bar{u}^T \implies \mathrm{cond}_{\mathrm{nat}}(T) = 1.$$

Die Householder-Transformationen sind *unitär* und folglich ist ihre Spektralkondition gleich $\mathrm{cond}_{\mathrm{nat}}(T) = 1$.

c) Eliminationen

$$T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,i} & & 1 \end{bmatrix}, \quad |l_{jk}| \le 1 \implies \mathrm{cond}_\infty(T) \le 4.$$

Im folgenden betrachten wir nun das Eigenwertproblem reeller Matrizen. Die Grundlage des sog. "Householder-Verfahrens" ist der folgende Satz:

Satz 7.6 (Hessenberg-Normalform): *Zu jeder Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine Folge von Householder-Matrizen T_i, $i = 1,\ldots,n-2$, so daß TAT^T mit $T = T_{n-2}\ldots T_1$ eine Hessenberg-Matrix ist. Für symmetrisches A ist TAT^T eine Tridiagonalmatrix.*

Beweis: $A = [a_1,\ldots,a_n]$, a_k Spaltenvektoren von A. Im ersten Schritt wird $u_1 = (0,u_{12},\ldots,u_{1n})^T \in \mathbb{R}^n$, $\|u_1\|_2 = 1$, so bestimmt, daß mit $T_1 = I - 2u_1u_1^T$ gilt:

$$T_1 a_1 \in \operatorname{span}\{e_1, e_2\}$$

Damit gilt dann

$$A^{(1)} = T_1AT_1 = \underbrace{\left[\begin{array}{c|ccc} a_{11} & a_{12} & \ldots & a_{1n} \\ \hline //// & * & & \\ \hline 0 & & & \end{array}\right]}_{T_1A} \underbrace{\left[\begin{array}{c|ccc} 1 & 0 & \ldots & \\ \hline 0 & * & & \\ \vdots & & & \end{array}\right]}_{T_1^T} = \left[\begin{array}{c|c} a_{11} & * \\ \hline //// & \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right].$$

Im nächsten Schritt wendet man eine analoge Prozedur auf die reduzierte Matrix $\tilde{A}^{(1)}$ an. Nach $n-2$ Schritten erhält man so eine Matrix $A^{(n-2)}$, welche offenbar Hessenberg-Gestalt hat. Mit A ist auch $A^{(1)} = T_1AT_1$ symmetrisch und folglich auch $A^{(n-2)}$. Dann ist $A^{(n-2)}$ als symmetrische Hessenberg-Matrix eine Tridiagonalmatrix. Q.E.D.

Für eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ erfordert das Householder-Verfahren zur Reduktion von A auf Tridiagonalform $\frac{2}{3}n^3 + O(n^2)$ Operationen. Zur Reduktion einer allgemeinen Matrix auf Hessenberg-Form sind $\frac{5}{3}n^3 + O(n^2)$ Operationen nötig. In diesem Fall erweist sich die folgende Methode von Wilkinson[8] als günstiger; sie erfordert nur etwa halb soviele Operationen. Zur Reduktion einer Matrix $A \in \mathbb{R}^{n\times n}$ auf Hessenberg-Form werden dabei Ähnlichkeitstransformationen mit Eliminationsmatrizen

$$E_{p-1} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{c} \text{p} \end{array} \\ \left[\begin{array}{ccccc} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{p+1,p} & 1 & \\ & & \vdots & & \ddots \\ & & l_{np} & & & 1 \end{array}\right] \end{array}\!\!\begin{array}{l} \\ \\ \\ \\ \\ \\ \text{p} \end{array}, \quad E_{p-1}^{-1} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{c} \text{p} \end{array} \\ \left[\begin{array}{ccccc} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{p+1,p} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{np} & & & 1 \end{array}\right] \end{array}\!\!\begin{array}{l} \\ \\ \\ \\ \\ \\ \text{p} \end{array}$$

[8] James Hardy Wilkinson (1919-1986): Englischer Mathematiker; Wissenschaftler National Physical Laboratory in London (seit 1946); fundamentale Beiträge zur numerischen linearen Algebra, insbes. zur Rundungsfehleranalyse; Mitbegründer der NAG Library (1970).

und Permutationsmatrizen

$$P_{pq} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} & & & & \mathrm{p} & & & & \mathrm{q} \end{array} \\ \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & \cdots & 1 & & & \\ & & & & 1 & & & & \\ & & & \vdots & & \ddots & \vdots & & \\ & & & & & & 1 & & \\ & & & 1 & \cdots & 0 & & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \mathrm{p} \\ \\ \\ \mathrm{q} \\ \\ \\ \end{array} , \qquad P_{pq}^{-1} = P_{pq},$$

verwendet:

$$A^{(i)} = T_i^{-1} A^{(i-1)} T_i \,, \quad T_i = P_{p_i q_i} E_{p_i}. \tag{7.3.25}$$

Wie wir schon bei der Analyse des Gaußschen Eliminationsverfahrens (mit Spaltenpivotierung) gesehen haben, bewirkt die Multiplikation $E_p^{-1}A$ bzw. AE_p die Subtraktion des l_{jp}-fachen $(j = p+1, \ldots, n)$ der p-ten Zeile von der j-ten Zeile von A bzw. die Addition des l_{jp}-fachen der j-ten Spalte zur p-ten Spalte von A. Die Multiplikation $P_{pq}A$ bzw. AP_{pq} bewirkt die Vertauschung der p-ten mit der q-ten Zeile bzw. der p-ten mit der q-ten Spalte. Die Matrizen P_{pq} sind offenbar unitär und haben folglich die Kondition $\mathrm{cond}_{\mathrm{nat}}(P_{pq}) = 1$. Die Eliminationsmatrizen E_p haben zwar eine günstige Kondition, $\mathrm{cond}_\infty(E_p) \leq 4$, doch zerstören sie die eventuell vorhandene Symmetrie von A

Satz 7.7 (Reduktion auf Hessenbergmatrix): *Zu jeder Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine Folge von Permutations- und Eliminationsmatrizen $P_{i+1,q_i}E_i$, $i = 1, \ldots, n-2$, so daß $T^{-1}AT$ mit $T = P_{2,q_1}E_1 \cdots P_{n-1,q_{n-2}}E_{n-2}$ eine Hessenberg-Matrix ist.*

Beweis: Die Permutationsmatrizen P_{i+1,q_i} dienen allein der Stabilisierung des Eliminationsprozesses durch Spaltenpivotierung; dadurch wird garantiert, daß stets $|l_{j,p_i}| \leq 1$ ist. Im ersten Schritt ermittelt man zunächst das Pivotelement in der 1-sten Spalte

$$|a_{p_1 1}| = \max_{j=2,\ldots,n} |a_{j1}|,$$

und vertauscht die 2-te mit der p_1-ten Zeile sowie die 2-te mit der p_1-ten Spalte:

$$A' := P_{2p_1} A P_{2p_1}.$$

Dann werden durch den Eliminationsschritt die Elemente in der 1-sten Spalte unterhalb von a'_{21} annuliert: $A'' := E_1^{-1} P_{2p_1} A P_{2p_1}$ mit

$$E_1^{-1} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{32} & 1 & & \\ & \vdots & & \ddots & \\ & -l_{n2} & & & 1 \end{bmatrix}, \quad l_{j2} = \frac{a'_{j1}}{a'_{21}}, \quad j = 3, \ldots, n.$$

Die Ähnlichkeitstransformation erfordert noch die Multiplikation mit E_1 von rechts:

$$A^{(1)} = A'' E_1 = \left[\begin{array}{c|ccc} a'_{11} & a'_{12} & \cdots & a'_{1n} \\ a'_{21} & a'_{22} & \cdots & a'_{2n} \\ \hline 0 & & * & \end{array}\right]$$

Dies bewirkt eine Addition von Vielfachen der k-ten Spalten, $k = 3, \ldots, n$, zur 2-ten Spalte, d. h.: Die Elemente der 1-ten Spalte werden nicht mehr verändert. Nach $n-2$ solcher Transformationen erhält man eine zu A ähnliche Hessenberg-Matrix. Q.E.D.

Das älteste Verfahren zur Reduktion einer gegebenen (hier reell symmetrischen) Matrix auf Tridiagonalform stammt von Givens[9] (1958). Es verwendet Drehmatrizen der Form

$$U_{pq} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} & & & & \mathrm{p} & & & & \mathrm{q} & & \\ \end{array} \\ \left[\begin{array}{ccccccccccc} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \cos(\alpha) & & \cdots & & \sin(\alpha) & & & \\ & & & & & 1 & & & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & & & & 1 & & & & \\ & & & -\sin(\alpha) & & \cdots & & \cos(\alpha) & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{array}\right] \end{array} \begin{array}{l} \\ \\ \\ \\ \mathrm{p} \\ \\ \\ \\ \mathrm{q} \\ \\ \\ \end{array}.$$

Dieser Algorithmus benötigt aber etwa doppelt so viele Operationen wie die Methode von Householder, so daß wir hier nicht weiter darauf eingehen.

[9] James Wallace Givens, 1910-1993: US-Amerikanischer Mathematiker: arbeitete am Oak Ridge National Laboratory; bekannt durch die nach ihm benannte Matrixtransformation "Givens-Rotation" ("Computation of plane unitary rotations transforming a general matrix to triangular form", SIAM J. Anal. Math. 6, 26-50, 1958).

7.4 Tridiagonal- und Hessenberg-Matrizen

Im folgenden behandeln wir Verfahren zur Lösung des Eigenwertproblems symmetrischer Tridiagonalmatrizen und von Hessenberg-Matrizen, die durch Anwendung einer Reduktionsmethode aus einer allgemeinen Matrix erzeugt werden.

7.4.1 LR- und QR-Verfahren

Sei $A \in \mathbb{R}^{n\times n}$ zunächst eine beliebige Matrix. Wir betrachten zwei iterative Verfahren zur Lösung des vollständigen Eigenwertproblems von A:

(I) Das "LR-Verfahren" nach Rutishauser[10] (1958) erzeugt ausgehend von $A^{(1)} := A$ eine Folge von Matrizen $A^{(t)}$, $t \in \mathbb{N}$, durch die Vorschrift

$$A^{(t)} = L^{(t)}R^{(t)} \ (LR\text{-Zerlegung}), \quad A^{(t+1)} := R^{(t)}L^{(t)}. \tag{7.4.26}$$

Wegen

$$A^{(t+1)} = R^{(t)}L^{(t)} = [L^{(t)}]^{-1}\underbrace{L^{(t)}R^{(t)}}_{=A^{(t)}}L^{(t)}$$

sind alle Iterierten $A^{(t)}$ ähnlich zu A und haben somit dieselben Eigenwerte. Unter geeigneten Voraussetzungen an A läßt sich zeigen, daß

$$\lim_{t\to\infty} A^{(t)} = \lim_{t\to\infty} R^{(t)} = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}, \quad \lim_{t\to\infty} L^{(t)} = I, \tag{7.4.27}$$

wobei λ_i die Eigenwerte von A sind.

Das LR-Verfahren erfordert in jedem Schritt die Berechnung einer LR-Zerlegung mit Hilfe des Gaußschen Algorithmus und ist folglich viel zu aufwendig für allgemeine vollbesetzte Matrizen. Bei Hessenberg-Matrizen ist der Aufwand jedoch vertretbar. Der schwerwiegende Nachteil des LR-Verfahrens ist die notwendige Voraussetzung der Existenz der LR-Zerlegungen $A^{(t)} = L^{(t)}R^{(t)}$. Hat man nur $P^{(t)}A^{(t)} = L^{(t)}R^{(t)}$ mit einer Permutationsmatrix $P^{(t)} \neq I$, so braucht keine Konvergenz vorzuliegen.

(II) Das "QR-Verfahren" nach Francis[11] (1961) gilt als das derzeit effizienteste zur Lösung des Eigenwertproblems von Hessenberg-Matrizen. Zur Umgehung der Hauptschwierigkeit beim LR-Verfahren liegt es nahe, eine analoge Iteration mit Hilfe der stets existierenden

[10] Heinz Rutishauser (1918-1970): Schweizer Mathematiker und Informatiker; seit 1962 Professor an der ETH Zürich; Beiträge zu Numerische Lineare Algebra (LR-Verfahren: Solution of eigenvalue problems with the LR transformation, Appl. Math. Ser. nat. Bur. Stand. 49, 47-81(1958).) und Analysis sowie zu Grundlagen der Computer-Arithmetik.

[11] J. F. G. Francis: the QR transformation. A unitary analogue to the LR transformation, Computer J. 4, 265-271 (1961/1962).

QR-Zerlegung anzusetzen:

$$A^{(t)} = Q^{(t)} R^{(t)}\,, \quad A^{(t+1)} := R^{(t)} Q^{(t)}\,, \quad t \in \mathbb{N}, \tag{7.4.28}$$

wobei $Q^{(t)}$ unitäre und $R^{(t)}$ rechte obere Dreiecksmatrizen mit positiven Diagonalelementen sind. Die QR-Zerlegung wird etwa mit Hilfe des Householder-Verfahrens vorgenommen. Auch hier kommt aus Ökonomiegründen nur eine Anwendung auf einfach strukturierte Matrizen wie Hessenberg-Matrizen in Frage. Wegen

$$A^{(t+1)} = R^{(t)} Q^{(t)} = {Q^{(t)}}^T \underbrace{Q^{(t)} R^{(t)}}_{= A^{(t)}} Q^{(t)}$$

sind wieder alle Iterierten $A^{(t)}$ ähnlich zu A. Zum Nachweis der Konvergenz des QR-Verfahrens benötigen wir den folgenden Hilfssatz:

Hilfssatz 7.3: *Es seien $E^{(t)} \in \mathbb{R}^{n\times n}$, $t \in \mathbb{N}$, reguläre Matrizen mit $\lim E^{(t)} = I$ und $E^{(t)} = Q^{(t)} R^{(t)}$ zugehörige QR-Zerlegungen. Dann gilt notwendig*

$$\lim_{t\to\infty} Q^{(t)} = I = \lim_{t\to\infty} R^{(t)}. \tag{7.4.29}$$

Beweis: Wegen

$$\|E^{(t)} - I\|_2 = \|Q^{(t)} R^{(t)} - Q^{(t)} {Q^{(t)}}^T\|_2 = \|R^{(t)} - {Q^{(t)}}^T\|_2 \to 0$$

konvergiert $q_{jk}^{(t)} \to 0$ $(t \to \infty)$, $j < k$. Dies erzwingt wegen

$$I = Q^{(t)} {Q^{(t)}}^T = \begin{bmatrix} \square & & & \to & 0 \\ & \square & & * & \\ & & \ddots & & \\ & * & & \square & \\ & & & & \square \end{bmatrix} \begin{bmatrix} \square & & & & \\ & \square & & * & \\ & * & \ddots & & \\ & & & \square & \\ \to & 0 & & & \square \end{bmatrix}$$

notwendig

$$q_{jj}^{(t)} \to \pm 1\,, \quad q_{jk}^{(t)} \to 0 \quad (t \to \infty)\,, \quad j > k\,.$$

Also konvergiert $Q^{(t)} \to \mathrm{diag}(\pm 1)$ $(t \to \infty)$. Wegen

$$Q^{(t)} R^{(t)} = E^{(t)} \to I \quad (t \to \infty)\,, \quad r_{jj} > 0$$

ist also $\lim_{t\to\infty} Q^{(t)} = I$. Dann ist aber auch

$$\lim_{t\to\infty} R^{(t)} = \lim_{t\to\infty} {Q^{(t)}}^T E^{(t)} = I,$$

was zu zeigen war. Q.E.D.

Satz 7.8 (QR-Algorithmus): *Die Eigenwerte der Matrix $A \in \mathbb{R}^{n\times n}$ seien betragsmäss ig separiert: $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Dann gilt für die durch das QR-Verfahren erzeugten Matrizen $A^{(t)} = (a_{jk}^{(t)})_{j,k=1,\ldots,n}$:*

$$\{\lim_{t\to\infty} a_{jj}^{(t)} \mid j = 1,\ldots,n\} = \{\lambda_1,\ldots,\lambda_n\}. \tag{7.4.30}$$

Beweis: Es gilt

$$A^{(t+1)} = R^{(t)}Q^{(t)} = {Q^{(t)}}^T \underbrace{Q^{(t)}R^{(t)}}_{=A^{(t)}} Q^{(t)} =$$

$$= \ldots = [Q^{(1)}\ldots Q^{(t)}]^T A \underbrace{[Q^{(1)}\ldots Q^{(t)}]}_{=:\,P^{(t)}}$$

Die normierten Eigenvektoren w_i, $\|w_i\| = 1$, zu den Eigenwerten λ_i sind linear unabhängig. Die Matrix $W = [w_1,\ldots,w_n]$ ist also regulär und genügt der Beziehung $AW = W\Lambda$ mit der Diagonalmatrix $\Lambda = \mathrm{diag}(\lambda_i)$. Folglich gilt

$$A = W\Lambda W^{-1}.$$

Sei $QR = W$ die QR-Zerlegung von W und $LS = PW^{-1}$ eine LR-Zerlegung von PW^{-1} (P geeignete Permutationsmatrix). Wir betrachten im folgenden den Fall $P = I$. Es ist

$$A^t = [W\Lambda W^{-1}]^t = W\Lambda^t W^{-1} = W\Lambda^t LS = W[\Lambda^t L(\Lambda^{-1})^t]\Lambda^t S$$

$$= W \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ l_{jk}\left(\frac{\lambda_j}{\lambda_k}\right)^t & & 1 \end{bmatrix} \Lambda^t S = W[I + N^{(t)}]\Lambda^t S$$

$$= QR[I + N^{(t)}]\Lambda^t S = Q[R + RN^{(t)}]\Lambda^t S,$$

und somit

$$A^t = Q[I + RN^{(t)}R^{-1}]R\Lambda^t S.$$

Da die Eigenwerte λ_i betragsmäßig der Größe nach geordnet sind, ist $|\lambda_j/\lambda_k| < 1$, $j > k$, d. h.

$$N^{(t)} \to 0, \quad RN^{(t)}R^{-1} \to 0 \quad (t \to \infty).$$

Für die QR-Zerlegungen $I + RN^{(t)}R^{-1} = \tilde{Q}^{(t)}\tilde{R}^{(t)}$ folgt dann nach Hilfssatz 7.3

$$\tilde{Q}^{(t)} \to I, \quad \tilde{R}^{(t)} \to I \quad (t \to \infty).$$

Weiter ist

$$A^t = Q\tilde{Q}^{(t)}[\tilde{R}^{(t)}R\Lambda^t S]$$

offenbar eine QR-Zerlegung von A^t (mit nicht notwendig positiven Diagonalelementen

von R !). Aus

$$\begin{aligned}\underbrace{Q^{(1)}\dots Q^{(t)}}_{=P^{(t)}}\,\underbrace{R^{(t)}\dots R^{(1)}}_{=:S^{(t)}} &= \underbrace{Q^{(1)}\dots Q^{(t-1)}}_{=P^{(t-1)}}\,A^{(t)}\,\underbrace{R^{(t-1)}\dots R^{(1)}}_{=:S^{(t-1)}}\\ &= P^{(t-1)}\,\underbrace{[P^{(t-1)^T}\,A\,P^{(t-1)}]}_{\text{s. Beweisbeginn}}\,S^{(t-1)}\\ &= A\,P^{(t-1)}S^{(t-1)}\end{aligned}$$

folgt

$$P^{(t)}S^{(t)} = AP^{(t-1)}S^{(t-1)} = \ldots = A^{t-1}P^{(1)}S^{(1)} = A^t\,.$$

Also ist $P^{(t)}S^{(t)} = A^t$ eine zweite QR-Zerlegung von A^t (mit positiven Diagonalelementen von R!). Es gibt folglich Diagonalmatrizen $D^{(t)} = \mathrm{diag}(\pm 1)$, so daß

$$P^{(t)} = Q\,\underbrace{\tilde{Q}^{(t)}D^{(t)}}_{=:T^{(t)}},\quad (|t_{jk}^{(t)}|)_{j,k=1,\ldots,n} \to I \quad (t\to\infty).$$

Damit finden wir, daß

$$\begin{aligned}A^{(t+1)} &= P^{(t)^T}AP^{(t)} = [QT^{(t)}]^T AQT^{(t)} = T^{(t)^T}Q^TAQT^{(t)}\\ &= T^{(t)^T}R\Lambda R^{-1}T^{(t)} \quad (\text{wegen } W^{-1}AW = \Lambda \Leftrightarrow R^{-1}Q^TAQR = \Lambda).\\ &= T^{(t)^T}\begin{bmatrix}\lambda_1 & & *\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}\tilde{Q}^{(t)}D^{(t)}.\end{aligned}$$

Wegen $\tilde{Q}^{(t)} \to I\ (t\to\infty)$ konvergiert also

$$D^{(t)}A^{(t+1)}D^{(t)} \to \begin{bmatrix}\lambda_1 & & *\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}\quad (t\to\infty).$$

Besitzt W^{-1} keine LR-Zerlegung, so erscheinen die Eigenwerte λ_i nicht mehr betragsmäßig der Größe nach geordnet. Q.E.D.

Die Konvergenzgeschwindigkeit des QR-Verfahrens wird bestimmt durch die Größen

$$\left|\frac{\lambda_j}{\lambda_k}\right| < 1, \quad j > k,$$

d. h.: Die Konvergenz ist um so schneller, je besser die Eigenwerte von A betragsmäßig separiert sind. Für positiv definite Matrizen kann man zeigen, daß das QR-Verfahren

doppelt so schnell konvergiert wie das entsprechende LR-Verfahren; es benötigt jedoch pro Iterationsschritt auch etwa die doppelte Anzahl von Operationen. Unter gewissen Bedingungen kann für da QR-Verfahren sogar kubische Konvergenz erreicht werden, d. h.: $|\lambda^{(t)} - \lambda| \leq c|\lambda^{(t-1)} - \lambda|^3$. Wie das LR-Verfahren wendet man das QR-Verfahren nur auf bereits reduzierte Matrizen an, für die eine QR-Zerlegung leichter zu berechnen ist: Hessenberg-Matrizen, symmetrische Tridiagonalmatrizen oder allgemeiner Bandmatrizen der Bandbreite $2m + 1 \ll n$. Dies ist gerechtfertigt aufgrund der folgenden Aussage:

Hilfssatz 7.4: *Ist A eine Hessenberg-Matrix (oder eine symmetrische $2m+1$-Bandmatrix), so gilt dasselbe für alle vom QR-Algorithmus erzeugten Matrizen $A^{(t)}$.*

Praktische Erfahrungen zeigen, daß das QR-Verfahren in Verbindung mit der Reduktionsmethode allen anderen bekannten Verfahren zur Lösung des vollständigen Eigenwertproblems überlegen ist.

7.4.2 Verfahren von Hyman

Die klassische Methode zur Berechnung der Eigenwerte einer Tridiagonal- oder Hessenberg-Matrix basiert auf der Bestimmung des charakteristischen Polynoms, allerdings ohne dessen Koeffizienten explizit auszurechnen. Das "Verfahren von Hyman[12] (1957) berechnet das charakteristische Polynom $p_A(z)$ einer Hessenberg-Matrix $A \in \mathbb{R}^{n\times n}$. Wir nehmen an, daß die Matrix A nicht in zwei Teilmatrizen vom Hessenberg-Typ zerfällt, d. h.: $a_{j+1,j} \neq 0$, $j = 1, \ldots, n-1$. Mit einer noch zu wählenden Funktion $c(z)$ wird das Gleichungssystem betrachtet:

$$\begin{aligned}
(a_{11} - z)x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= -c(z) \\
a_{21}x_1 + (a_{22} - z)x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n &= 0 \\
&\vdots \\
a_{n,n-1}x_{n-1} + (a_{nn} - z)x_n &= 0
\end{aligned}$$

Setzt man $x_n = 1$, so lassen sich $x_{n-1}, \ldots, x_1$ sowie $c(z)$ sukzessive bestimmen. Nach der Cramerschen Regel gilt

$$1 = x_n = \frac{(-1)^n c(z) a_{21}a_{32}\ldots a_{n,n-1}}{\det(A - zI)}.$$

Folglich ist $c(z) = \text{konst.} \cdot \det(A - zI)$, und man erhält eine rekursive Formel zur Bestimmung des charakteristischen Polynoms $p_A(z) = \det(A - zI)$.

[12] M. A. Hyman: Eigenvalues and eigenvectors of general matrices, Twelfth National Meeting A.C.M., Houston, Texas, 1957.

Sei nun $A \in \mathbb{R}^{n\times n}$ eine symmetrische Tridiagonalmatrix mit $b_i \neq 0\,,\; i = 1, \dots, n-1:$

$$A = \begin{bmatrix} a_1 & b_1 & & 0 \\ b_1 & & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & b_{n-1} & a_n \end{bmatrix}.$$

Zur Berechnung des charakteristischen Polynoms $p_A(z)$ dienen die Rekursionsformeln

$$\begin{aligned} &p_0(z) = 1\,, \quad p_1(z) = a_1 - z\,, \\ &p_i(z) = (a_i - z)\,p_{i-1}(z) - b_{i-1}^2 p_{i-2}(z), \quad i = 2, \dots, n. \end{aligned}$$

Die Polynome $p_i \in P_i$ sind gerade die i-ten Hauptminoren von $\det(A - zI)$, d. h.: $p_n = p_A$. Um dies einzusehen, entwickle man den $(i+1)$-ten Hauptminor nach der $(i+1)$-ten Spalte:

$$\left[\begin{array}{cccc|c|c|c} a_1 - z & b_1 & & & & & \\ b_1 & \ddots & \ddots & & & & \\ & \ddots & & & & & \\ \hline & & & & b_{i-1} & & \\ & & & b_{i-1} & a_i - z & b_i & \\ \hline & & & & b_i & a_{i+1} - z & \ddots \\ \hline & & & & & \ddots & \ddots \end{array}\right] = \underbrace{(a_{i+1} - z)p_i(z) - b_i^2 p_{i-1}(z)}_{=:\, p_{i+1}(z)}$$

$$\qquad\qquad i-1 \qquad i \qquad i+1$$

Es ist oft nützlich, die Ableitung p_A' zu kennen (z. B. bei der Anwendung des Newton-Verfahrens). Diese erhält man durch die Rekursionsformeln

$$\begin{aligned} &q_0(z) = 0\,, \quad q_1(z) = -1 \\ &q_i(z) = -p_{i-1}(z) + (a_i - z)q_{i-1}(z) - b_{i-1}^2 q_{i-2}(z)\,, \quad i = 2, \dots, n\,, \\ &\quad q_n = p_A'\,. \end{aligned}$$

Hat man eine Nullstelle λ von p_A, d. h. einen Eigenwert von A bestimmt, so erhält man einen zugehörigen Eigenvektor durch $w\,(\lambda)$, wobei

$$w\,(z) = \begin{bmatrix} w_0\,(z) \\ \vdots \\ w_{n-1}(z) \end{bmatrix}, \qquad \begin{aligned} w_0(z) \;&\equiv\; 1 \quad (b_n := 1) \\ w_i(z) \;&:=\; \frac{(-1)^i p_i(z)}{b_1 \dots b_i}\,, \quad i = 1, \dots, n\,. \end{aligned} \tag{7.4.31}$$

Um dies zu verifizieren, berechnet man $(A - zI)\, w(z)$. Für $i = 1, \ldots, n-1$ $(b_0 := 0)$ ist

$$\begin{aligned}
&b_{i-1}w_{i-2}(z) + a_i w_{i-1}(z) + b_i w_i(z) - z\, w_{i-1}(z) = \\
&= b_{i-1}(-1)^{i-2}\,\frac{p_{i-2}(z)}{b_1 \ldots b_{i-2}} + a_i(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \ldots b_{i-1}} + b_i(-1)^i\,\frac{p_i(z)}{b_1 \ldots b_i} - z(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \ldots b_{i-1}} \\
&= b_{i-1}^2(-1)^{i-2}\,\frac{p_{i-2}(z)}{b_1 \ldots b_{i-1}} + a_i(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \ldots b_{i-1}} + (-1)^i\,\frac{(a_i - z)\,p_{i-1}(z) - b_{i-1}^2 p_{i-2}(z)}{b_1 \ldots b_{i-1}} \\
&\quad - z(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \ldots b_{i-1}} = 0.
\end{aligned}$$

Für $i = n$ ist $(b_n := 1)$

$$\begin{aligned}
b_{n-1}w_{n-2}(z) + a_n w_{n-1}(z) - z\, w_{n-1}(z) = \\
&= b_{n-1}(-1)^{n-2}\,\frac{p_{n-2}(z)}{b_1 \ldots b_{n-2}} + (a_n - z)(-1)^{n-1}\,\frac{p_{n-1}(z)}{b_1 \ldots b_{n-1}} \\
&= -b_{n-1}^2(-1)^{n-1}\,\frac{p_{n-2}(z)}{b_1 \ldots b_{n-1}} + (a_n - z)(-1)^{n-1}\,\frac{p_{n-1}(z)}{b_1 \ldots b_{n-1}} \\
&= (-1)^{n-1}\,\frac{p_n(z)}{b_1 \ldots b_{n-1}} = -w_n(z)\,.
\end{aligned}$$

Wir finden also

$$(A - zI)w(z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -w_n(z) \end{bmatrix}. \tag{7.4.32}$$

Für einen Eigenwert λ von A ist $w_n(\lambda) = \text{konst.} \cdot p_A(\lambda) = 0$, d. h.

$$(A - \lambda I)\, w\,(\lambda) = 0.$$

7.4.3 Verfahren der Sturmschen Kette

Wir werden nun ein Verfahren zur Bestimmung der Nullstellen des charakteristischen Polynoms P_A einer symmetrischen (unzerlegbaren) Tridiagonalmatrix $A \in \mathbb{R}^{n\times n}$ beschreiben. Ableitung der Identität (7.4.32) ergibt

$$[(A - zI)w(z)]' = -w(z) + (A - zI)w'(z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -w_n'(z) \end{bmatrix}.$$

Wir setzen $z = \lambda$ mit einem Eigenwert λ von A und multiplizieren mit $-w(\lambda)$:

$$\begin{aligned} 0 &< \|w(\lambda)\|_2^2 - (\underbrace{[A-\lambda I]w(\lambda)}_{=0}, w'(\lambda)) \\ &= w_{n-1}(\lambda)w_n'(\lambda) = -\frac{p_{n-1}(\lambda)p_n'(\lambda)}{b_1^2 \dots b_{n-1}^2}. \end{aligned}$$

Folglich ist $p_n'(\lambda) \neq 0$, d. h. wir haben allgemein:

(S1) *Alle Nullstellen von p_n sind einfach.*

(S2) *Für jede Nullstelle λ von p_n gilt*

$$p_{n-1}(\lambda)p_n'(\lambda) < 0.$$

Weiter gilt

(S3) *Für jede reelle Nullstelle ζ von p_{i-1} ist*

$$p_i(\zeta)p_{i-2}(\zeta) < 0\,, \quad i = 2, \dots, n;$$

denn in diesem Fall ist $p_i(\zeta) = -b_{i-1}^2 p_{i-2}(\zeta)$ und wäre $p_i(\zeta) = 0$, so folgte der Widerspruch

$$0 = p_i(\zeta) = p_{i-1}(\zeta) = p_{i-2}(\zeta) = \ldots = p_0(\zeta) = 1.$$

Schließlich gilt trivialerweise

(S4) *$p_0 \neq 0$ hat keinen Vorzeichenwechsel.*

Definition 7.5: *Eine Folge von Polynomen $p = p_n, p_{n-1}, \dots, p_0$ (oder allgemeiner stetiger Funktionen $f_n, f_{n-1}, \dots, f_0$) mit den Eigenschaften (S1)-(S4) heißt eine "Sturmsche Kette*[13] *von p.*

Die vorausgegangene Überlegung hat also zu folgendem Resultat geführt:

Satz 7.9 (Sturmsche Kette): *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, unzerlegbare Tridiagonalmatrix. Dann bilden die Hauptminoren $p_i(z)$ der Matrix $A - zI$ eine Sturmsche Kette des charakteristischen Polynoms $p_A(z) = p_n(z)$ von A.*

Der Wert der Existenz einer Sturmschen Kette zu einem Polynom p beruht auf dem folgenden Resultat:

Satz 7.10 (Intervallschachtelung): *Es sei p ein Polynom und $p = p_n, p_{n-1}, \dots, p_0$ eine zugehörige Sturmsche Kette. Dann ist die Anzahl der reellen Nullstellen von p in*

[13] *Jacques Charles Franois Sturm (1803-1855): Französisch-Schweizer Mathematiker; Professor an der École Polytechnique in Paris seit 1840; Beiträge zur Mathematischen Physik, Differentialgleichungen, ("Sturm-Liouville-Problem") und Differentialgeometrie.*

einem Intervall $[a,b]$ *gleich* $N(b)-N(a)$, *wobei* $N(\zeta)$ *die Anzahl der Vorzeichenwechsel der Kette* $p_n(\zeta),\dots,p_0(\zeta)$ *bezeichnet.*

Beweis: Wir betrachten die Zahl der Vorzeichenwechsel $N(a)$ für wachsendes a. $N(a)$ bleibt konstant, solange a keine Nullstelle eines der p_i passiert. Sei nun a Nullstelle eines der p_i; wir unterscheiden zwei Fälle:

(i) Fall $p_i(a)=0$ für $i\neq n$: In diesem Fall ist $p_{i+1}(a)\neq 0$, $p_{i-1}(a)\neq 0$. Die Vorzeichen von $p_j(a)$, $j\in\{i-1,i,i+1\}$ zeigen daher für genügend kleines $h>0$ ein Verhalten, das durch eines der zwei folgenden Tableaus skizziert werden kann:

	$a-h$	a	$a+h$
$i-1$	$-$	$-$	$-$
i	$+/-$	0	$-/+$
$i+1$	$+$	$+$	$+$

	$a-h$	a	$a+h$
$i-1$	$+$	$+$	$+$
i	$+/-$	0	$-/+$
$i+1$	$-$	$-$	$-$

In jedem Fall ist $N(a-h)=N(a)=N(a+h)$, und die Anzahl der Vorzeichenwechsel ändert sich nicht.

(ii) Fall $p_n(a)=0$: In diesem Fall kann das Verhalten von $p_j(a)$, $j\in\{n-1,n\}$, durch eines der folgenden beiden Tableaus beschrieben werden (wegen (S2)):

	$a-h$	a	$a+h$
n	$-$	0	$+$
$n-1$	$-$	$-$	$-$

	$a-h$	a	$a+h$
n	$+$	0	$-$
$n-1$	$+$	$+$	$+$

Also ist $N(a-h)=N(a)=N(a+h)-1$, d. h.: Beim Passieren einer Nullstelle von p_n kommt genau ein Vorzeichenwechsel hinzu. Für $a<b$ gibt daher $N(b)-N(a)=N(b+h)-N(a-h)$, $h>0$ genügend klein, die Anzahl der Nullstellen von p_n im Intervall $[a-h,b+h]$ an. Da h beliebig klein gewählt werden kann, folgt die Behauptung. Q.E.D.

Satz 7.10 führt auf ein einfaches "Bisektionsverfahren" zur Bestimmung der Nullstellen des charakteristischen Polynoms p_A einer symmetrischen, irreduziblen Tridiagonalmatrix $A\in\mathbb{R}^{n\times n}$. Offenbar besitzt A nur reelle, einfache Eigenwerte

$$\lambda_1<\lambda_2<\cdots<\lambda_n.$$

Für $x\to-\infty$ besitzt die Kette

$$\begin{aligned}&p_0(x)=1\,,\quad p_1(x)=a_1-x\\&i=2,\dots,n:\quad p_i(x)=(a_i-x)p_{i-1}(x)-b_i^2p_{i-2}(x)\,,\end{aligned}$$

die Vorzeichenverteilung $+,\dots,+$; also ist $N(x)=0$. Folglich gibt $N(\zeta)$ gerade die

Anzahl der Nullstellen λ von p_A mit $\lambda < \zeta$ an. Für die Eigenwerte λ_i von A gilt also:

$$\lambda_i < \zeta \quad \Longleftrightarrow \quad N(\zeta) \geq i. \tag{7.4.33}$$

Zur Bestimmung des i-ten Eigenwertes λ_i startet man mit einem Intervall $[a_0, b_0]$, das λ_i sicher enthält, z. B.: $a_0 < \lambda_1 < \lambda_n < b_0$. Dann halbiert man sukzessiv das Intervall und testet mit Hilfe der Sturmschen Kette, in welchem der beiden neuen Teilintervalle λ_i liegt, d. h.: Man bildet für $t = 0, 1, 2, \ldots$:

$$\mu_t := \frac{a_t + b_t}{2}, \qquad \begin{aligned} a_{t+1} &:= \begin{cases} a_t\,, & \text{falls} \quad N(\mu_t) \geq i \\ \mu_t\,, & \text{falls} \quad N(\mu_t) < i \end{cases} \\ b_{t+1} &:= \begin{cases} \mu_t\,, & \text{falls} \quad N(\mu_t) \geq i \\ b_t\,, & \text{falls} \quad N(\mu_t) < i \end{cases} \end{aligned} \tag{7.4.34}$$

Es gilt dann stets $\lambda_i \in [a_{t+1}, b_{t+1}]$,

$$[a_{t+1}, b_{t+1}] \subset [a_t, b_t]\,, \quad |a_{t+1} - b_{t+1}| = \tfrac{1}{2}|a_t - b_t|,$$

und a_t konvergiert monoton wachsend, b_t monoton fallend gegen λ_i. Dies Verfahren ist zwar langsam, aber sehr genau (geringe Rundungsfehleranfälligkeit) und gestattet die Bestimmung eines jeden beliebigen Eigenwertes unabhängig von den übrigen.

7.5 Übungsaufgaben

Übung 7.1:

8 Lineare Optimierung

8.1 Lineare Programme

"Lineares Programm" ist die historisch bedingte Bezeichnung für lineare Optimierungsaufgaben vom folgenden Typ:

Beispiel 8.1: Eine Fabrik kann zwei Typen A und B eines Produkts unter folgenden Bedingungen herstellen:

Produkt	Typ A	Typ B	maximal möglich
Stück pro Tag	x_1	x_2	100 Stück
Arbeitszeit pro Stück	4	1	160 Stunden
Kosten pro Stück	20	10	1100 DM
Gewinn pro Stück	120	40	? DM

Wie müssen x_1 und x_2 gewählt werden, damit der Gewinn maximal wird? Dabei muß offenbar der lineare Ausdruck

$$Q(x_1, x_2) := 120x_1 + 40x_2$$

zu einem Maximum gemacht werden unter den linearen Nebenbedingungen

$$\begin{aligned} x_1 + \quad x_2 &\leq 100 \\ 4x_1 + \quad x_2 &\leq 160 \quad , \quad x_1 \geq 0\,, \quad x_2 \geq 0. \\ 20x_1 + 10x_2 &\leq 1100 \end{aligned} \tag{8.1.1}$$

Dies ist ein lineares Programm in der sog. "Standardform".

Beispiel 8.2: Die Produktion von 7 Zuckerfabriken soll so auf 300 Verbrauchsorte verteilt werden, daß der Bedarf befriedigt wird und die Transportkosten minimiert werden.

Fabrik	F_j $(j = 1, \ldots, 7)$,	Verbrauchsort	G_k $(k = 1, \ldots, 300)$
Produktion	a_j (pro Monat),	Verbrauch	r_k (pro Monat)

transportierte Menge $F_j \to G_k$: $x_{j,k}$, Kosten $c_{j,k}$ (pro Einheit).

Es sei vorausgesetzt, daß Bedarf und Produktionsmenge gleich sind:

$$\sum_{k=1}^{300} r_k = \sum_{j=1}^{7} a_j.$$

Zu minimieren sind die Gesamtkosten

$$Q(x_{1,1},\ldots,x_{7,300}) := \sum_{j=1}^{7}\sum_{k=1}^{300} c_{j,k}x_{j,k}$$

unter den Nebenbedingungen $x_{j,k} \geq 0$ und

$$\sum_{j=1}^{7} x_{j,k} = r_k \quad (k=1,\ldots,300)\,, \qquad \sum_{k=1}^{300} x_{j,k} = a_j \quad (j=1,\ldots,7).$$

Diese Aufgabenstellungen lassen sich in folgenden allgemeinen Rahmen einordnen: (Für einen Vektor $x=(x_1,\ldots,x_n)^T$ bedeutet die Schreibweise $x \geq 0$, daß alle Komponenten $x_i \geq 0$ sind.)

Für $1 \leq m \leq n$ seien eine Matrix $A \in \mathbb{R}^{m\times n}$ vom Rang m sowie Vektoren $b \in \mathbb{R}^m$, $b \geq 0$, und $c \in \mathbb{R}^n$ gegeben.

Definition 8.1: *Als "lineares Programm in Normalform" (bzw. "kanonischer" Form), abgekürzt LP, bezeichnet man die Aufgabe, unter den Gleichungsnebenbedingungen und Vorzeichenbedingungen*

$$Ax = b\,, \quad x \geq 0$$

ein Minimum der Zielfunktion $Q(x) := c^T x$ zu bestimmen.

Anders ausgedrückt sucht ein lineares Programm im "zulässigen Bereich"

$$M := \{\, x \in \mathbb{R}^n \mid Ax = b\,, \quad x \geq 0 \,\} \tag{8.1.2}$$

ein $x^* \in M$ zu bestimmen, so daß

$$c^T x^* = \min_{x\in M} c^T x. \tag{8.1.3}$$

Zur Einordnung der obigen Beispiele in diesen Rahmen können folgende Umformungen herangezogen werden:

- Ein Ungleichung mit $\geq$ wird durch Multiplikation mit -1 in eine mit $\leq$ überführt.
- Eine Ungleichung $a_1x_1+\ldots+a_nx_n \leq \beta$ wird durch Einführung einer sog. "Schlupfvariablen" y in eine Gleichung und eine Vorzeichenbedingung überführt:

$$a_1x_1 + \ldots + a_nx_n + y = \beta\,, \quad y \geq 0.$$

- Für jede Gleichung $a_1x_1 + \ldots + a_nx_n = \beta$ kann (eventuell nach Multiplikation mit -1) stets $\beta \geq 0$ vorausgesetzt werden.
- Fehlt für eine Variable, etwa für x_1, die Vorzeichenbedingung, so wird x_1 durch die Differenz $y_1 - y_2$ zweier neuer Variablen ersetzt, und man fordert $y_1 \geq 0$, $y_2 \geq 0$.

– Gleichungen, die Linearkombinationen anderer Gleichungen sind, werden weggelassen, so daß für die Matrix $A \in \mathbb{R}^{m \times n}$ stets Rang $A = m$ angenommen werden kann.

– Wegen $\max c^T x = -\min(-c^T x)$ kann man alle linearen Programme auf die Bestimmung eines Minimums zurückführen.

Der zulässige Bereich M eines LP ist der Durchschnitt einer linearen Mannigfaltigkeit mit Halbräumen und folglich abgeschlossen. Weiter ist M *"konvex"*:

$$x, y \in M \Longrightarrow \lambda x + (1 - \lambda)\, y \in M \quad \forall\, \lambda \in [0, 1].$$

Ist $M = \emptyset$, so besitzt das LP keine Lösung. Im Fall $M \neq \emptyset$ existiert immer eine Lösung, wenn M beschränkt (und damit kompakt) ist; für unbeschränktes M braucht keine Lösung zu existieren.

Beispiel 8.3: In einfachen Fällen lassen sich lineare Programme graphisch lösen: Der zulässige Bereich M in Beispiel 1 (Standardformulierung) ist der Durchschnitt von 5 Halbebenen des $\mathbb{R}^2$, deren Begrenzungsgeraden durch die Gleichungen

$$\begin{aligned} x_1 \geq 0, \quad x_2 \geq 0, \\ x_1 + x_2 &= 100 \\ 4x_1 + x_2 &= 160 \\ 20x_1 + 10x_2 &= 1100 \end{aligned}$$

gegeben sind:

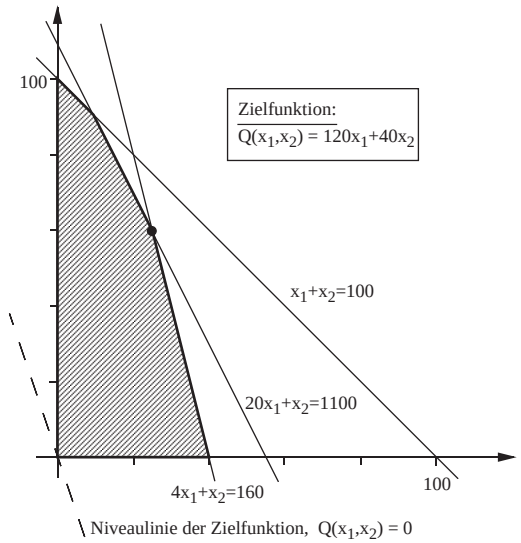


Abbildung 8.1: *Grafische Lösung eines Linearen Programms*

Parallelverschiebung der Niveaulinie bis an den Rand von M ergibt als maximalen Wert

den der Niveaulinie durch den Punkt (x_1^*, x_2^*) mit

$$\left.\begin{array}{rcl} 4x_1^* + x_2^* &=& 160 \\ 20x_1^* + 10x_2^* &=& 1100 \end{array}\right\} \Longrightarrow \begin{array}{rcl} x_2^* &=& 60 \\ x_1^* &=& 25 \end{array},$$

$$Q_{\max} = 120x_1^* + 40x_2^* = 5400.$$

Der optimale Punkt ist in diesem Fall eine Ecke des Polygongebiets M. Dies ist kein Zufall und wird sich als wesentlicher Punkt bei der Behandlung allgemeinerer Probleme dieses Typs erweisen. Die maximal mögliche Stückzahl von $x_1 + x_2 = 100$ wird unter dem Kriterium der Gewinnmaximierung also nicht erreicht; dafür wird die zur Verfügung stehende Arbeitszeit voll genutzt.

Wir betrachten im folgenden die linearen Programmierungsaufgaben stets in Normalform. Zunächst studieren wir die Struktur der Lösungsmenge eines LP.

Definition 8.2: *Ein Vektor $x \in M$ heißt "Ecke" (oder "Extremalpunkt") der zulässigen Menge M, wenn er keine Darstellung*

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

mit $x^{(1)}$, $x^{(2)} \in M$, $x^{(1)} \neq x^{(2)}$, mit einem $\lambda \in (0, 1)$ zuläßt. Für $x \in M$ setzen wir $I(x) := \{i \in \{1, \ldots, n\} \,|\, x_i > 0\}$. Weiter bezeichnen wir im folgenden mit a_k die Spaltenvektoren der Matrix A.

Hilfssatz 8.1 (Sekantensatz): *Sind für ein $x \in M$ die Spaltenvektoren in*

$$B(x) := \{\, a_k \,|\, k \in I(x) \,\} \tag{8.1.4}$$

linear abhängig, so besitzt x eine Darstellung

$$x = \tfrac{1}{2}\,(x^{(1)} + y) \tag{8.1.5}$$

mit $x^{(1)}$, $y \in M$ und $I(x^{(1)}) \subsetneq I(x)$.

Beweis: O.B.d.A. sei $I(x) = \{1, \ldots, k\}$, so daß

$$\sum_{i=1}^{k} x_i a_i = b.$$

Ist $B(x)$ linear abhängig, so gibt es Zahlen d_i, $i = 1, \ldots, k$, nicht alle Null, so daß

$$\sum_{i=1}^{k} d_i a_i = 0.$$

Der Vektor

$$x(\lambda) = (x_1 + \lambda d_1, \ldots, x_k + \lambda d_k, \underbrace{0, \ldots, 0}_{n-k})^T$$

erfüllt dann für jedes $\lambda \in \mathbb{R}$ die Gleichung $Ax(\lambda) = b$. Wegen $x_i > 0,\ i = 1, \ldots, k$, ist für hinreichend kleines $|\lambda|$ auch $x(\lambda) \geq 0$ und somit $x(\lambda) \in M$. Läßt man nun λ ausgehend von Null wachsen oder fallen, so gelangt man in einem der beiden Fälle zu einem λ^*, für das mindestens eine der Komponenten $x_i(\lambda^*)$, $i = 1, \ldots, k$, verschwindet. Ferner ist $x(\lambda) \in M$ für $|\lambda| \leq |\lambda^*|$. Mit $x^{(1)} = x(\lambda^*)$ und $y = x(-\lambda^*)$ gilt dann $x = \frac{1}{2}(x^{(1)} + y)$ und $I(x^{(1)}) \subsetneqq I(x)$. Q.E.D.

Hilfssatz 8.2 (Eckensatz): *Ein Vektor $x \in M$ ist genau dann Ecke, wenn die Spaltenvektoren in $B(x)$ linear unabhängig sind.*

Beweis: Aus Hilfssatz 8.1 folgt, daß für eine Ecke $x \in M$ notwendig $B(x)$ linear unabhängig sein muß. Sei nun $B(x)$ linear unabhängig, aber $x \in M$ keine Ecke, d. h. $x = \lambda x^{(1)} + (1 - \lambda)x^{(2)}$ mit $x^{(1)}, x^{(2)} \in M$, $x^{(1)} \neq x^{(2)}$, $\lambda \in (0, 1)$. Dann impliziert $x_i = 0$ auch $x_i^{(1)} = x_i^{(2)} = 0$, so daß gilt:

$$\sum_{i \in I(x)} x_i^{(1)} a_i = \sum_{i \in I(x)} x_i^{(2)} a_i = b.$$

Wegen $x^{(1)} \neq x^{(2)}$ ist also $B(x)$ linear abhängig im Widerspruch zur Annahme. Q.E.D.

Definition 8.3: *Wegen Rang $A = m$ besteht $B(x)$ für eine Ecke $x \in M$ aus höchstens m Vektoren. Ist für eine Ecke $x \in M$ aber* dim $B(x) < m$, *so heißt x "entartete Ecke"; in diesem Fall kann $B(x)$ zu einer Basis $\hat{B}(x)$ aus Spaltenvektoren von A ergänzt werden. In jedem Fall heißt eine solche Basis $\hat{B}(x)$ "Basis der Ecke x".*

Durch eine Basis ist eine Ecke $x \in M$ über das Gleichungssystem $Ax = b$ eindeutig bestimmt. Andererseits gibt es höchstens

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

Systeme von m linear unabhängigen Spaltenvektoren der Matrix A, d. h. Ecken von M.

Hilfssatz 8.3 (Eckenlösung): *Besitzt das LP eine Lösung $x \in M$, so gibt es eine Ecke $x^* \in M$, die ebenfalls Lösung ist.*

Beweis: Ist x selbst keine Ecke, so wenden wir Hilfssatz 8.1 an, d. h.: Das Minimum von $c^T x$ wird im Mittelpunkt $x = \frac{1}{2}(x^{(1)} + y)$ der Verbindungsgeraden zwischen $x^{(1)}$ und y angenommen. Folglich ist $c^T x$ dort konstant, d. h.: $x^{(1)}$ ist auch Lösung, aber mit $I(x^{(1)}) \subsetneqq I(x)$. Auf diese Weise erreicht man in endlich vielen Schritten eine Ecke von M (im Extremfall $\tilde{x} = 0$. Q.E.D.

8.2 Das Simplex-Algorithmus

Im Folgenden entwickeln wir den sog. "Simplex-Algorithmus" nach G. B. Dantzig[1] (1947) zur Lösung von Linearen Programmen. Wir verwenden weiter die Bezeichnungen des vorherigen Abschnitts. Sei x^0 eine Ecke des zulässigen Bereichs M der kanonischen Programmierungsaufgabe

$$Ax = b\,, \qquad x \geq 0\,, \qquad c^T x = \min!$$

mit zugehöriger Basis $\hat{B}(x^0) = \{\, a_i\,,\, i \in I^0 \,\}\,, I^0 \supseteq I(x^0)$. Dann gilt

$$\sum_{i \in I^0} x_i^0\, a_i = b. \tag{8.2.6}$$

Für ein beliebiges $x \in M$ ist $Ax = b$ und folglich

$$\sum_{i \in I^0} \{x_i - x_i^0\} a_i = -\sum_{i \notin I^0} x_i a_i.$$

(Die Schreibweise "$i \notin I^0$" bedeutet $i \in \{1, \ldots, n\} \setminus I^0$.) Wegen der linearen Unabhängigkeit von $\hat{B}(x^0)$ kann nach den Differenzen $x_i - x_i^0$ aufgelöst werden, und man erhält Gleichungen der Form

$$x_i = \sum_{k \notin I^0} \alpha_{ik} x_k + x_i^0\,, \quad i \in I^0. \tag{8.2.7}$$

Der zugehörige Zielfunktionswert

$$c^T x = c^T x^0 + c^T (x - x^0) = c^T x^0 + \sum_{i \in I^0} c_i (x_i - x_i^0) + \sum_{i \notin I^0} c_i x_i.$$

ergibt sich nach Substitution von $x_i - x_i^0$ $(i \in I^0)$ in der Form

$$c^T x = \sum_{k \notin I^0} \gamma_k x_k + c^T x^0 \tag{8.2.8}$$

$$\gamma_k = \sum_{i \in I^0} \alpha_{ik} c_i + c_k\,, \quad k \notin I^0. \tag{8.2.9}$$

Die Gleichungsnebenbedingung und die Zielfunktion $c^T x$ sind offenbar (bzgl. der Basis $\hat{B}(x^0)$) durch das folgende $(m+1) \times (n-m+1)$-Gleichungssystem wiedergegeben:

[1] George B. Dantzig (1914-): US-Amerikanischer Mathematiker; entwickelte 1947 den Simplex-Algorithmus wärend seiner Tätigkeit in einem Forschungslabor der U.S. Air Force; s. sein Buch "Lineare Programmierung und Erweiterungen", Springer 1966 (Übersetzung aus dem Englischen); seit 1966 Professor an der Stanford University.

$$\begin{aligned} x_i &= \sum_{k \notin I^0} \alpha_{ik} x_k + x_i^0, \quad i \in I^0 \\ c^T x &= \sum_{k \notin I^0} \gamma_k x_k + c^T x^0. \end{aligned} \tag{8.2.10}$$

Die Komponenten $x_i\,,\ i \in I^0\,,$ und der zugehörige Zielfunktionswert z eines Vektors $x \in \mathbb{R}^n$ sind durch Vorgabe von $x_k \geq 0\,(k \notin I^0)$ in (8.2.10) eindeutig bestimmt. Gilt dabei $x_i \geq 0\,(i \in I^0)\,,$ so ist $x \in M\,.$ Für die speziellen Werte $x_k = 0\,(k \notin I^0)$ ergibt sich gerade die Ausgangsecke $x^0\,.$

Wir betrachten nun die umgekehrte Situation, daß der zulässige Bereich M gerade aus denjenigen Vektoren $x \geq 0$ besteht, deren Komponenten einem System der Gestalt (8.2.10) genügen, mit gewissen Zahlen $x_i^0 \geq 0\,(i \in I^0)\,.$ Dann ist der Vektor $x^0 \in \mathbb{R}^n$ mit den Komponenten $x_i^0\,(i \in I^0)\,,\ x_i^0 = 0\,(i \notin I^0)$ automatisch Ecke von $M\,,$ denn er erfüllt offensichtlich (8.2.10), d. h.: $Ax^0 = b\,,$ und jede Darstellung $x^0 = \lambda x + (1-\lambda)\tilde{x}$ mit $x, \tilde{x} \in M\,,\ 0 < \lambda < 1$ impliziert zunächst notwendig $x_k = \tilde{x_k} = x_k^0\,(k \notin I^0)$ und damit auch $x_i = \tilde{x}_i = x_i^0\,(i \in I^0)$.

Das System (8.2.10) charakterisiert also zu einer festen Ecke x^0 bzw. der zugehörigen Basis $\hat{B}(x^0)$ den zulässigen Bereich M. Der Simplexalgorithmus sucht nun eine (zu x^0 benachbarte) Ecke x^1 von $M\,,$ wobei möglichst $c^T x^1 < c^T x^0$ gelten soll. Der zugehörige Basiswechsel in der Darstellung (8.2.10) wird mit Hilfe des sog. "Gauß-Jordan-Algorithmus" bewerkstelligt (s. Kapitel 4.2). Der Vollständigkeit halber rekapitulieren wir im Folgenden den Gauß-Jordan-Algorithmus. Dieser dient zur Lösung linearer Gleichungssysteme $Ax = y$ mit $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m\,,$ durch sukzessiven Austausch der Komponenten von x gegen solche von $y\,.$ Ist ein Matrixelement $a_{pq} \neq 0\,,$ so kann die p-te Gleichung nach x_q aufgelöst werden:

$$x_q = -\frac{a_{p1}}{a_{pq}}\, x_1 - \ldots - \frac{a_{p,q-1}}{a_{pq}}\, x_{q-1} + \frac{1}{a_{pq}}\, y_p - \frac{a_{p,q+1}}{a_{pq}}\, x_{q+1} - \ldots - \frac{a_{pn}}{a_{pq}}\, x_n\,.$$

Durch Substitution von x_q in den anderen Gleichungen

$$a_{j1}x_1 + \ldots + a_{j,q-1}x_{q-1} + a_{jq}\boxed{x_q} + a_{j,q+1}x_{q+1} + \ldots + a_{jn}x_n = y_j$$

erhält man für $j = 1, \ldots, m,\ j \neq p$:

$$\left[a_{j1} - \frac{a_{jq}a_{p1}}{a_{pq}} \right] x_1 + \ldots + \left[a_{j,q-1} - \frac{a_{jq}a_{p,q-1}}{a_{pq}} \right] x_{q-1} + \frac{a_{jq}}{a_{pq}}\, y_p +$$

$$+ \left[a_{j,q+1} - \frac{a_{jq}a_{p,q+1}}{a_{pq}} \right] x_{q+1} + \ldots + \left[a_{jn} - \frac{a_{jq}a_{pn}}{a_{pq}} \right] x_n = y_j\,.$$

Das Resultat ist ein zum Ausgangssystem äquivalentes System

$$\tilde{A} \begin{bmatrix} x_1 \\ \vdots \\ y_p \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ x_q \\ \vdots \\ y_m \end{bmatrix}, \tag{8.2.11}$$

wobei die Elemente der Matrix $\tilde{A}$ wie folgt bestimmt sind:

– Pivotelement: $\tilde{a}_{pq} = \frac{1}{a_{pq}}$,

– Pivotzeile: $\tilde{a}_{pk} = -\frac{a_{pk}}{a_{pq}}, \quad k = 1, \ldots, n, \quad k \neq q\,,$

– Pivotspalte: $\tilde{a}_{jq} = \frac{a_{jq}}{a_{pq}}\,, \quad j = 1, \ldots, m, \quad j \neq p\,,$

– sonstige: $\tilde{a}_{jk} = a_{jk} - a_{jq}\frac{a_{pk}}{a_{pq}}, \quad \begin{matrix} j = 1, \ldots, m, & j \neq p \\ k = 1, \ldots, n, & k \neq q. \end{matrix}$

Gelingt es, durch Fortsetzung des Verfahrens alle Komponenten von x durch solche von y zu ersetzen, so hat man eine explizite Darstellung der Lösung von $Ax = y$. Im Fall $m = n$ ergibt sich so auch die Inverse A^{-1}, allerdings im allgemeinen mit vertauschten Zeilen und Spalten. Bei der Festlegung des Pivotelementes empfiehlt es sich aus Stabilitätsgründen, unter allen in Frage kommenden a_{pq} jeweils eines von möglichst großem Betrag zu wählen. Für ein quadratisches Gleichungssystem mit regulärer Koeffizientenmatrix A ist das Gauß-Jordan-Verfahren zur Berechnung von A^{-1} stets durchführbar.

Beispiel 8.4:

$$\begin{bmatrix} 1 & 2 & 1 \\ -3 & -5 & -1 \\ -7 & -12 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Austauschschritte: $\boxed{\cdot}$ Pivotelement

x_1	x_2	x_3	
1	2	1	y_1
−3	−5	−1	y_2
−7	$\boxed{-12}$	−2	y_3

x_1	y_3	x_3	
−1/6	−1/6	$\boxed{2/3}$	y_1
−1/12	5/12	−5/6	y_2
−7/12	−1/12	−1/6	x_2

x_1	y_3	y_1	
1/4	1/4	3/2	x_3
$\boxed{-1/8}$	3/8	−1/4	y_2
−5/8	−1/8	−1/4	x_2

y_2	y_3	y_1	
−2	1	1	x_3
−8	3	−2	x_1
5	−2	1	x_2

Inverse: $\begin{bmatrix} -2 & -8 & 3 \\ 1 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$.

Der Simplexalgorithmus besteht aus zwei "Phasen". In Phase I wird eine Ausgangsecke x^0 konstruiert und das zugehörige *Tableau* erstellt:

	$x_k\ (k \notin I^0)$	
x_i $(i \in I^0)$	$\alpha_{ik}\ (i \in I^0,\ k \notin I^0)$	x_i^0 $(i \in I^0)$
z	$\gamma_k\ (k \notin I^0)$	$c^T x^0$

In Phase II werden dann unter Verwendung des Gauß-Jordan-Algorithmus Basiswechsel vollzogen, wobei der Zielfunktionswert jeweils möglichst stark verkleinert wird. Wir beginnen mit der Beschreibung dieses Basisaustausches.

Phase II (Basisaustausch)

1. Gilt $\gamma_i \geq 0\,(i \notin I^0)$, so folgt für beliebige Vorgabe von $x_i \geq 0\,(i \notin I^0)$, d. h.: für beliebige Punkte $x \in M$, stets

$$z = \sum_{k \notin I^0} \gamma_k x_k + c^T x^0 \geq c^T x^0,$$

   d. h. Die Startecke x^0 ist bereits optimal.

2. Gibt es ein $q \notin I^0$ mit $\gamma_q < 0$, so sind zwei Fälle zu unterscheiden:

   (a) Im Falle $\alpha_{iq} \geq 0\,(i \in I^0)$ erhält man durch Vorgabe von $x_q := \lambda$, $\lambda \in \mathbb{R}_+$ und

$x_k = 0\,(k \notin I^0,\ k \neq q)$ Vektoren $x \in M$,

$$x_i \;=\; \alpha_{iq}\lambda + x_i^0 \geq 0,$$

mit Zielfunktionswert

$$z \;=\; \gamma_q\lambda + c^T x^0 \to -\infty \quad (\lambda \to \infty).$$

Die Aufgabe ist also unlösbar.

(b) Gibt es Indizes $p \in I^0$ mit $\alpha_{pq} < 0$, so wird die Variable x_q gegen eine noch auszuwählende x_p ausgetauscht. Die Elemente des Tableaus sind dabei gemäß dem Gauß-Jordan-Algorithmus wie folgt zu transformieren:

Pivotelement:	$\alpha_{pq} \to \alpha'_{qp} = \dfrac{1}{\alpha_{pq}}$;
Pivotspalte:	$\alpha_{iq} \to \alpha'_{ip} = \dfrac{\alpha_{iq}}{\alpha_{pq}} \quad (i \in I^0 \setminus \{p\})$
$(\gamma_q \to \gamma_p)$	$\gamma_q \to \gamma'_p = \dfrac{\gamma_q}{\alpha_{pq}}$;
Pivotzeile:	$\alpha_{pk} \to \alpha'_{qk} = -\dfrac{\alpha_{pk}}{\alpha_{pq}} \quad (k \notin I^0,\ k \neq q)$
$(x_p \to x_q)$	$x_p^0 \to x_q^1 = -\dfrac{x_p^0}{\alpha_{pq}}$;
sonstige:	$\alpha_{ik} \to \alpha'_{ik} = \alpha_{ik} - \dfrac{\alpha_{iq}\alpha_{pk}}{\alpha_{pq}}$
$(x_k \to x_k,\ x_i \to x_i)$	$x_i^0 \to x_i^1 = x_i^0 - \dfrac{\alpha_{iq}x_p^0}{\alpha_{pq}}$
	$\gamma_k \to \gamma'_k = \gamma_k - \dfrac{\gamma_q\alpha_{pk}}{\alpha_{pq}}$
	$c^T x^0 \to c^T x^1 = c^T x^0 - \dfrac{\gamma_q x_p^0}{\alpha_{pq}}$.

Auswahlregel (R): Der Index $p \in I^0$ wird dabei gemäß der folgenden Regel ausgewählt:

$$\alpha_{pq} < 0\,, \quad \frac{x_p^0}{\alpha_{pq}} = \max_{\alpha_{iq}<0, i\in I^0} \frac{x_i^0}{\alpha_{iq}}. \tag{8.2.12}$$

Satz 8.1 (Simplex-Algorithmus): *Wird der Basisaustausch gemäß der Regel* (R) *vorgenommen, so ist der Vektor* $x^1 \in \mathbb{R}^n$ *mit den Komponenten* $x_i^1 > 0$ $(i \in I^1 := [I^0 \setminus \{p\}] \cup \{q\})$, $x_i^1 = 0$ $(i \notin I^1)$ *wieder eine Ecke von* M *mit der zugehörigen Basis*

$$\hat{B}(x^1) \;=\; \left[\hat{B}(x^0) \setminus \{a_p\}\right] \cup \{a_q\}, \tag{8.2.13}$$

und es gilt $c^T x^1 \le c^T x^0$ *. Im Falle* $x_p^0 > 0$ *ist sogar* $c^T x^1 < c^T x^0$ *.*

Beweis: Wegen $\alpha_{pq} < 0$ folgt $x_q^1 = -x_p^0/\alpha_{pq} \ge 0$. Ist weiter $\alpha_{iq} \ge 0$, so folgt $x_i^1 = x_i^0 - \alpha_{iq} x_p^0/\alpha_{pq} \ge 0$. Im Falle $\alpha_{iq} < 0$ gilt ebenfalls wegen der Auswahlregel (R):

$$\frac{x_i^1}{\alpha_{iq}} = \frac{x_i^0}{\alpha_{iq}} - \frac{x_p^0}{\alpha_{pq}} \le 0 \qquad \Longrightarrow \qquad x_i^1 \ge 0\,..$$

Nach dem oben gesagten ist x^1 also Ecke von M. Q.E.D.

Der Eckenaustausch nach (2b) kann solange fortgesetzt werden, bis Fall (1) oder Fall (2a) eintritt. Eine *nicht* entartete Ecke kann dabei nie ein zweites Mal erreicht werden, da ihr Austausch je zu einer Verkleinerung des Zielfunktionswertes führt. Das Auftreten entarteter Ecken wird weiter unten diskutiert werden. Hier könnten sich (theoretisch) im Laufe des Verfahrens verschiedene Basen zu einer entarteten Ecke zyklisch wiederholen, so daß der Algorithmus nicht abbricht.

Beispiel 8.5: Beispiel 8.1.1 aus Abschnitt 8.1 erhält nach Einführung von Schlupfvariablen x_3, x_4, x_5 die Form

$$-120x_1 - 40x_2 = \min!\,, \quad x_i \ge 0\,, \quad i = 1, \ldots, 5,$$

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & + & x_3 & & & = & 100 \\ 4x_1 & + & x_2 & + & & + & x_4 & = & 160 \\ 20x_1 & + & 10x_2 & & & & + \; x_5 & = & 1100 \end{array}$$

Offensichtlich ist $x^0 = (0, 0, 100, 160, 1100)^T$ eine *nicht* entartete Ecke mit der Basis

$$B(x^0) = \{a_3, a_4, a_5\} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Das Ausgangstableau ist also:

I	x_1	x_2	$x_6 = 1$	x_i^0/α_{iq}
x_3	-1	-1	100	-100
x_4	$\boxed{-4}$	-1	160	-40
x_5	-20	-10	1100	-55
z	$\boxed{\text{-120}}$	-40	0	

$\boxed{\text{Fall (2b)}}$

Wahl $q = 1$ (oder $q = 2$)

Regel (R) $\Longrightarrow p = 4$.

Eckentausch (Pivotelement α_{41})

II	x_4	x_2	$x_6 = 1$	x_i^0/α_{iq}
x_3	$1/4$	$-3/4$	60	-80
x_1	$-1/4$	$-1/4$	40	-160
x_5	5	$\boxed{-5}$	300	-60
z	30	$\boxed{-10}$	-4800	

$\boxed{\text{Fall (2b)}}$

Wahl $q = 2$

Regel (R) $\Longrightarrow p = 5$.

Eckentausch (Pivotelement α_{52})

III	x_4	x_5	$x_6 = 1$
x_3	$-1/2$	$-3/20$	15
x_1	$-1/2$	$1/20$	25
x_2	1	$-1/5$	60
z	20	2	-5400

$\boxed{\text{Fall (1)}}$

Eckenlösung $x^3 = (25, 60, 15, 0, 0)^T$

Extremwert $z = -5400$.

Wie wir schon gesehen haben, wird der maximale Gewinn von 5400 DM erreicht, wenn 25 Produkte des Typs A und 60 des Typs B hergestellt werden.

Bemerkung 8.1: Zur Kontrolle der Rechnung sollten die Größen $\gamma_k \, (k \notin I^0)$ zusätzlich auch aus der folgenden Formel berechnet werden:

$$\gamma_k = \sum_{i \in I^0} \alpha_{ik} c_i + c_k. \tag{8.2.14}$$

Phase I (Konstruktion einer Startecke)

Wir diskutieren nun die Phase I des Simplexalgorithmus, d. h. die Konstruktion einer Ausgangsecke x^0. Im Falle eines in Standardform gegebenen Programms mit $b \geq 0$ ist dies, wie obiges Beispiel zeigt, sehr einfach:

$$\text{(I)} \qquad c^T x = \max!\,, \quad Ax \leq b\,, \quad x \geq 0. \tag{8.2.15}$$

Durch Einführung von Schlupfvariablen $v \in \mathbb{R}^m$ geht (I) in die kanonische Form über:

$$(\tilde{\text{I}}) \qquad \tilde{c}^T \tilde{x} = \min!\,, \quad \tilde{A}\tilde{x} = \tilde{b}\,, \quad \tilde{x} \geq 0, \tag{8.2.16}$$

mit

$$\tilde{A} = [A, I_m]\,, \quad \tilde{b} = b\,, \quad \tilde{c} = (-c, 0_m)^T, \quad \tilde{x} = (x, v)^T.$$

Wegen $b \geq 0$ ist der Vektor $\tilde{x}^0 = (0_n, \tilde{b}) \in \mathbb{R}^{n+m}$ automatisch eine Ecke von $\tilde{M}$, denn die zugehörigen Spaltenvektoren von $\tilde{A}$ bilden gerade die Einheitsmatrix I_m.

Ist die Programmierungsaufgabe in kanonischer Form gestellt,

$$\text{(II)} \qquad c^T x = \min!\,, \quad Ax = b\,, \quad x \geq 0, \tag{8.2.17}$$

oder ist in (I) $b \geq 0$ nicht erfüllt, so ist i. Allg. keine Ecke von M (oft nicht einmal ein zulässiger Vektor) ersichtlich. Zu ihrer Konstruktion betrachte man das Hilfsproblem (o.B.d.A.: $\tilde{b} \geq 0$)

$$\text{(}\tilde{\text{II}}\text{)} \qquad \tilde{c}^T \tilde{x} = \min!\,, \quad \tilde{A}\tilde{x} = \tilde{b}\,, \quad \tilde{x} \geq 0, \tag{8.2.18}$$

mit $v \in \mathbb{R}^m$,

$$\tilde{A} = [A, I_m]\,, \quad \tilde{b} = b\,, \quad \tilde{c} = (0_n, 1, \ldots, 1)^T\,, \quad \tilde{x} = (x, v)^T.$$

Hier ist mit $\tilde{x}^0 = (0_n, \tilde{b})$ eine Ausgangsecke von $\tilde{M}$ bekannt. Wegen $\tilde{x} \geq 0$ auf M besitzt das Hilfsproblem stets eine Lösung $\tilde{x}^* = (x^*, v^*)$, die man mit Hilfe des Simplexverfahrens bestimmen kann. Ist $v^* = 0$, so liefern die ersten n Komponenten der Lösung $\tilde{x}^*$ eine Ausgangsecke des ursprünglichen LP:

$$x_i^0 := \tilde{x}_i^*\,, \quad i = 1, \ldots, n. \tag{8.2.19}$$

Im Fall $v_i^* > 0$ für ein i muß $M = \emptyset$ sein, d. h.: Das LP besitzt keine Lösung. Mit dem Simplexverfahren kann also auch die Frage nach der Lösbarkeit des LP entschieden werden.

Nach den bisherigen Überlegungen ist der Simplexalgorithmus (mit der Auswahlregel (R)) grundsätzlich geeignet zur Lösung linearer Programmierungsaufgaben, bzw. zur Entscheidung ihrer Unlösbarkeit, vorausgesetzt, es treten keine entarteten Ecken auf. Das Erscheinen einer entarteten Ecke x^1 ist dadurch gekennzeichnet, daß im vorangehenden Austauschschritt das Kriterium (R) nicht zu einem eindeutig bestimmten Index $p \in I^0$ führt:

$$x^1 \quad \textit{nicht} \text{ entartet} \iff \begin{array}{c} \exists!\, p \in I^0 : \alpha_{pq} < 0 \\ \dfrac{x_p^0}{\alpha_{pq}} = \max\limits_{\alpha_{iq}<0} \dfrac{x_i^0}{\alpha_{iq}}; \end{array}$$

andernfalls folgte mit $x_p^0/\alpha_{pq} = x_{p'}^0/\alpha_{p'q}$

$$\left.\begin{array}{rcl} x_p^1 &=& x_p^0 - \alpha_{pq} x_p^0/\alpha_{pq} \;=\; 0 \\ x_{p'}^1 &=& x_{p'}^0 - \alpha_{p'q} x_p^0/\alpha_{pq} \;=\; 0 \end{array}\right\} \iff \begin{array}{l} x^1 \text{ hat weniger als } m \\ \text{positive Komponenten.} \end{array}$$

Tritt im Verlaufe des Simplexverfahrens eine entartete Ecke x^0 auf, so kann es passieren, daß für den Index $p \in I^0$ gerade $x_p^0 = 0$ ist. Dann bewirkt der Austauschschritt offenbar keine Veränderung des Vektors x^0, insbesondere also keine Reduzierung des Zielfunktionswertes, sonderen nur den Übergang zu einer anderen Basis zur Ecke x^0. Wiederholt sich dann dieselbe Basis zyklisch, so führt das Verfahren nicht zum Ziel. Obwohl in der Praxis häufig entartete Ecken auftreten, sind derartige Zyklen noch nicht beobachtet worden (nur bei eigens zu diesem Zweck konstruierten pathologischen Beispielen). Für Belange

der Praxis erscheint der Simplexalgorithmus mit der Auswahlregel (R) (ergänzt um eine geeignete Regel für den Fall einer entarteten Ecke) als hinreichend robust. Vom theoretischen Standpunkt ist diese Situation aber unbefriedigend, und man sucht nach einer Auswahlregel, mit der der Algorithmus grundsätzlich zum Ziel führt.

Definition 8.4 (Lexikographische Ordnung): *Ein Vektor $u \in \mathbb{R}^n$ heißt "lexikographisch positiv", $u \overset{\rightarrow}{>} 0$, wenn $u \neq 0$ ist und die erste nicht verschwindende Komponente positiv ist. Ein Vektor $u \in \mathbb{R}^n$ heißt "lexikographisch kleiner (größer)" als ein $v \in \mathbb{R}^n$, wenn $v-u \overset{\rightarrow}{>} 0$ $(u-v \overset{\rightarrow}{>} 0)$. Damit ist auf dem $\mathbb{R}^n$ eine "Ordnung" erklärt.*

Das Simplexverfahren sei gestartet mit einer Ecke x^{start} mit der Basis $\hat{B}(x^{\text{start}}) = \{a_1, \ldots, a_m\}$ (gegebenenfalls nach Umbenennung der Variablen). Damit ist $I^{\text{start}} = \{1, \ldots, m\}$. Zur Einführung einer erweiterten Auswahlregel werden die Parameter α_{ik} und γ_k auch für $k \in I^{\text{start}}$ erklärt durch

$$\alpha_{ik} := -\delta_{ik}\,, \quad \gamma_k := 0\,, \quad i\,, k \in I^{\text{start}},$$

Hilfssatz 8.4 (Hilfssatz): *Für die durch die Darstellungen*

$$a_i = \sum_{k \in I^0} c_{ik} a_k\,, \quad i \in \{1, \ldots, n\} \tag{8.2.20}$$

eindeutig bestimmten Zahlen c_{ik} gilt

$$c_{ki} = -\alpha_{ik}\,, \quad k \in \{1, \ldots, n\}\,, \quad i \in I^{\text{start}}. \tag{8.2.21}$$

Beweis: Für $i, k \in I^{\text{start}}$ ist nach Definition

$$\alpha_{ik} = -\delta_{ki} = -c_{ki}.$$

Sei nun $x \in M$ beliebig. Dann gilt

$$\sum_{i \in I^{\text{start}}} \{x_i - x_i^0\} a_i = -\sum_{i \notin I^{\text{start}}} x_i a_i = -\sum_{i \notin I^{\text{start}}} x_i \sum_{k \in I^{\text{start}}} c_{ik} a_k = -\sum_{i \in I^{\text{start}}} \Big(\sum_{k \notin I^{\text{start}}} c_{ki} x_k \Big) a_i$$

und folglich wegen der linearen Unabhängigkeit von $\hat{B}(x^0)$

$$x_i - x_i^0 = -\sum_{k \notin I^{\text{start}}} c_{ki} x_k\,, \quad i \in I^{\text{start}}.$$

Da die α_{ik} in der Darstellung (8.2.10) eindeutig bestimmt sind, ergibt sich notwendig $c_{ki} = -\alpha_{ik}$ $(k \in I^{\text{start}})$. Q.E.D.

Sei nun x^0 eine im Verlaufe des Verfahrens erreichte Ecke und $q \in I^0$ der Austausch-

index. Zur Bestimmung des Index $p \in I^0$ bilde man für alle $r \in I^0$ mit

$$\frac{x_r^0}{\alpha_{rq}} = \max_{\alpha_{jq}<0} \frac{x_j^0}{\alpha_{jq}}\,, \quad \alpha_{rq} < 0, \tag{8.2.22}$$

die Vektoren

$$u^r = \left(\frac{x_r^0}{\alpha_{rq}}\,, -\frac{\alpha_{r1}}{\alpha_{rq}}\,, \ldots, -\frac{\alpha_{rm}}{\alpha_{rq}}\right)^T \in \mathbb{R}^{m+1}.$$

Auswahlregel ($\tilde{\text{R}}$): Der Index $p \in I^0$ wird dann als derjenige mit der Eigenschaft (8.2.22) gewählt, so daß u^p der lexikographisch größte unter den u^r ist.

Im Falle, daß der "maximale" Index in (8.2.22) eindeutig bestimmt ist, stimmt die Auswahlregel ($\tilde{\text{R}}$) offenbar mit (R) überein. Ansonsten ist durch ($\tilde{\text{R}}$) eindeutig ein $p \in I^0$ festgelegt; denn gäbe es keines, so wären für zwei $p, p' \in I^0$ die Vektoren u^p, $u^{p'}$ identisch. Dies bedeutete aber, daß die quadratische Matrix $(\alpha_{ik})_{i\in I^0, k=1,\ldots,m}$ zwei zueinander proportionale Zeilen hätte und somit singulär wäre. Nach Hilfssatz 8.4 wäre dann auch $(c_{ik})_{i=1,\ldots,m,k\in I^0}$ singulär im Widerspruch zur linearen Unabhängigkeit der Vektoren in $\hat{B}(x^0)$ und $\hat{B}(x^{\text{start}})$.

Der Ecke x^0 ordnen wir nun den folgenden Vektor zu:

$$v^0 = (c^T x^0, c_1 - \gamma_1, \ldots, c_m - \gamma_m)^T \in \mathbb{R}^{m+1}$$

Hilfssatz 8.5 (Reduktionssatz): *Beim Eckenaustausch $x^0 \to x^1$ unter der Bedingung der Auswahlvorschrift ($\tilde{\text{R}}$) wird der Vektor v^0 durch einen lexikographisch kleineren Vektoren v^1 ersetzt.*

Beweis: Die Indexmenge I^0 wird ersetzt durch $I^1 = (I^0 \backslash \{p\}) \cup \{q\}$. Die γ_i werden nach den folgenden Regeln transformiert:

$$\begin{aligned}
k \in I^0\,, \quad k \neq q\,: &\qquad \gamma_k \to \gamma_k - \gamma_q \tfrac{\alpha_{pk}}{\alpha_{pq}} \\
k = q\,: &\qquad \gamma_q \to 0 \,=\, \gamma_q - \gamma_q \tfrac{\alpha_{pq}}{\alpha_{pq}} \\
k \notin I^0\,, \quad k \neq p\,: &\qquad \gamma_k \to 0 \\
k = p\,: &\qquad \gamma_p \to \gamma_q \tfrac{1}{\alpha_{pq}} = \gamma_p - \gamma_q \tfrac{\alpha_{pp}}{\alpha_{pq}} \quad (\gamma_p = 0\,,\ \alpha_{pp} = -1).
\end{aligned}$$

Ferner gilt: $\qquad c^T x^0 \to c^T x^0 - \gamma_k \tfrac{x_p^0}{\alpha_{pq}}\,.$

Für den zur neuen Ecke x^1 gehörenden Vektor $v^1 \in \mathbb{R}^{m+1}$ gilt also:

$$v^1 = v^0 - \gamma_q \begin{bmatrix} x_p^0/\alpha_{pq} \\ -\alpha_{p1}/\alpha_{pq} \\ \vdots \\ -\alpha_{pm}/\alpha_{pq} \end{bmatrix} = v^0 - \gamma_q u^p.$$

Da $\gamma_q < 0$, $\alpha_{pq} < 0$ bleibt zu zeigen, daß der Vektor $w^p = (x_p^0, -\alpha_{p1}, \ldots, -\alpha_{pm})^T \in \mathbb{R}^{m+1}$ lexikographisch positiv ist. Dies geschieht durch Induktion bzgl. der Zahl der durchgeführten Verfahrensschritte.

(i) Die zur Ausgangsecke x^{start} gehörenden Vektoren w^k $(k = 1, \ldots, m)$ sind trivialerweise lexikographisch positiv, denn es ist $x_k^0 \geq 0$ und $-\alpha_{ki} = \delta_{ki}$ $(i = 1, \ldots, m)$.

(ii) Sei x^0 eine im Verlaufe des Verfahrens auftretende Ecke, und alle zu x^0 gebildeten Vektoren w^k $(k \in I^0)$ seien lexikographisch positiv. Beim Übergang von x^0 zur Ecke x^1 ergeben sich die zugehörigen Vektoren $\tilde{w}^k$ $(k \in I^1)$ wie folgt:

$$\begin{aligned} k \in I^1, \quad k \neq q: \qquad w^k &= \left(x_k^0 - \frac{\alpha_{kq} x_p^0}{\alpha_{pq}}, -\alpha_{k1} + \frac{\alpha_{kq}\alpha_{p1}}{\alpha_{pq}}, \ldots, -\alpha_{km} + \frac{\alpha_{kq}\alpha_{pm}}{\alpha_{pq}} \right)^T \\ &= w^k - \frac{\alpha_{kq}}{\alpha_{pq}} w^p \\ k = q: \qquad \tilde{w}^q &= \left(\frac{x_p^0}{\alpha_{pq}}, +\frac{\alpha_{p1}}{\alpha_{pq}}, \ldots, +\frac{\alpha_{pm}}{\alpha_{pq}} \right)^T = -\frac{1}{\alpha_{pq}} w^p. \end{aligned}$$

Hieraus entnehmen wir mit der Induktionsannahme:

$$\begin{aligned} k \in I^1, \quad k \neq q: \quad \text{a)} \quad & \alpha_{kq} \geq 0 \Longrightarrow \tilde{w}^k = w^k + \left| \frac{\alpha_{kq}}{\alpha_{pq}} \right| w^p \overset{\rightarrow}{>} 0, \\ \text{b)} \quad & \alpha_{kq} < 0 \Longrightarrow \text{Auswahlregel } (\tilde{R}): \; u^p \overset{\rightarrow}{>} u^k \\ & \Longrightarrow \tilde{w}^k = \alpha_{kq} u^k - \frac{\alpha_{kq}}{\alpha_{pq}} \alpha_{pq} u^p \overset{\rightarrow}{>} 0, \\ k = q: \qquad & \tilde{w}^q = \left| \frac{1}{\alpha_{pq}} \right| w^p \overset{\rightarrow}{>} 0. \end{aligned}$$

Dies vervollständigt den Beweis. Q.E.D.

Wir fassen die Ergebnisse der bisherigen Überlegungen zusammen:

Satz 8.2 (Erweitertes Simplex-Verfahren): *Unter der obigen Voraussetzung an die Ausgangsecke liefert der Simplexalgorithmus mit der Auswahlvorschrift* $(\tilde{\text{R}})$ *in endlich vielen Schritten eine Lösung des kanonischen Problems* (II) *oder die Bestätigung seiner Unlösbarkeit.*

Beweis: Nach Hilfssatz 8.5 kann aufgrund der Auswahlregel $(\tilde{\text{R}})$ keine Basis von Spaltenvektoren von A zweimal auftreten, denn durch die Ecke x^0 und eine zugehörige Basis $\hat{B}(x^0)$ ist der Vektor v^0 eindeutig bestimmt. Zyklen werden also vermieden. Q.E.D.

Index