NUMERISCHE MATHEMATIK I
Wintersemester 2009/10

G. Lube
Georg-August-Universität Göttingen, NAM

2. Februar 2010

0

Inhaltsverzeichnis

Kapitel 0

Einleitung

Die Beschreibung eines Modells in der Natur oder Ökonomie durch Relationen zwischen bekannten und unbekannten Größen führt auf ein *mathematisches Modell.* Dies ist oft eine algebraische, häufiger eine Differential- oder Integralgleichung (eventuell auch -ungleichung) bzw. ein System derartiger Relationen. Mathematische Disziplinen wie Algebra und Analysis befassen sich mit Fragen der Existenz und Eindeutigkeit derartiger mathematischer Modelle. Bei der Konstruktion der Lösung gibt man nicht selten einen – möglicherweise unendlichen – *Algorithmus* an, durch den sich die Lösung zumindest theoretisch beliebig genau ermitteln läßt.

Die *Numerische Mathematik* befaßt sich mit der zahlenmäßigen Berechnung von Lösungen mathematischer Modelle, die durch Formeln, Gleichungen, als Grenzwerte oder auch in anderer Form gegeben sind. Sie hat zu sichern, daß das entsprechende numerische Verfahren (bzw. Algorithmus) tatsächlich effizient auf einer Rechenanlage durchführbar ist und die Lösung des mathematischen Modells auch zuverlässig mit der gewünschten Genauigkeit ermittelt.

Wir veranschaulichen das Problem an einfachen Beispielen:

- Der Fundamentalsatz der Algebra zeigt, daß ein reelles Polynom vom Grad n auch n Nullstellen in der Menge der komplexen Zahlen besitzt. Der Existenzbeweis ist jedoch nicht konstruktiv, d.h. man erhält kein Verfahren zur expliziten Nullstellenbestimmung. Diese liefert die Numerische Mathematik.

- Die Lösung linearer Gleichungssysteme mit nicht verschwindender Determinante kann explizit durch die Cramersche Regel aufgeschrieben werden. Für die praktische Berechnung ist sie aber bei mehr als drei Unbekannten unbrauchbar.

- Für Anfangswertprobleme einer gewöhnlichen Differentialgleichung (oder eines entsprechenden Systems) liefert der Existenzbeweis für den Satz von Picard-Lindelöf unter bestimmten Glattheitsvoraussetzungen an die rechte Seite ein konstruktives Iterationsverfahren für die Lösung. Bei der Realisierung auf dem Computer ist aber dieses Verfahren viel zu uneffektiv.

Unter einem *konstruktiven Verfahren* wollen wir in der Numerischen Mathematik verstehen, daß die numerische Lösung einer mathematischen Aufgabe in endlich vielen Rechenschritten auf vorzugebende Genauigkeit ermittelt wird. In der Regel hängt dabei die Zahl notwendiger Rechenschritte von der geforderten Genauigkeit ab. Drei Fragestellungen sind neben der *Wohldefiniertheit* eines Verfahrens grundlegend für die Numerische Mathematik:

Fehlerabschätzungen, numerische Stabilität und *Aufwand des Verfahrens.*

Wir wollen diese Punkte nachfolgend kurz beleuchten:

- Die Mehrzahl der numerischen Algorithmen ist nicht endlich, d.h. man ist auf Näherungslösungen angewiesen. Zur Absicherung eines numerischen Verfahrens gehören somit *Fehlerabschätzungen*, also die Bestimmung gesicherter Schranken für die Abweichung der ermittelten Näherungslösung von der exakten Lösung der mathematischen Aufgabe, sofern deren Existenz nachweisbar ist. Dabei unterscheidet man einerseits zwischen *Verfahrens-* oder *Diskretisierungsfehlern*, d.h. den Fehlern, die durch die Approximation der exakten Lösung durch die numerische Näherungslösung entstehen. Andererseits erhält man *Rundungsfehler* durch die jeweilige Maschinengenauigkeit, d.h. durch die Ersetzung reeller Zahlen durch Dezimalzahlen mit fixierter Stellenzahl auf dem Computer.

- Der Aspekt der *Stabilität* bezieht sich auf die Empfindlichkeit gegenüber Fehlern der Eingangsgrößen (z.B. Datenfehler) oder von Zwischenergebnissen. Bei der Stabilitätsfrage unterscheidet man zwischen Aufgaben mit *guter oder schlechter Kondition.* Dabei bedeutet gute Kondition eines Problems, daß kleine Änderungen der Ausgangsgrößen nur kleine Lösungsänderungen bewirken. Für schlecht konditionierte Probleme muß man geeignete numerische Verfahren wählen. Hinzu kommt, daß auch bei der diskreten Approximation einer (im oben genannten Sinne gut konditionierten) kontinuierlichen Aufgabenstellung ein schlecht konditioniertes diskretes Problem entstehen kann. Letzteres kann z.B. bei der Diskretisierung von Differentialgleichungen durch Differenzen- oder Finite-Elemente-Verfahren der Fall sein.

- Fragen der *Effektivität* bzw. des *Rechenaufwandes* spielen vor allem deshalb eine ständig wachsende Rolle, da sich aufgrund der sich immer schneller entwickelnden Rechentechnik zunehmend komplexere mathematische Modelle numerisch berechnen lassen. Daher ist der *Komplexität* eines Verfahrens, d.h. der Aufwand an wesentlichen Rechenoperationen in Abhängigkeit von der Zahl der zu bestimmenden Unbekannten, zu betrachten.

Im Rahmen dieser Vorlesung wird eine Einführung in folgende Grundgebiete der Numerischen Mathematik gegeben:

- Lineare und nichtlineare Gleichungssysteme (**Teil I**)

- Interpolation und Numerische Integration (**Teil II**).

In dem sich in folgenden Semester anschließenden Kurs *Numerische Mathematik II* werden Einführungen in folgende Gebiete angeboten:

- Approximationstheorie

- Optimierung (insbesondere lineare Optimierung)

- Numerische Lineare Algebra (Eigenwertaufgaben, Krylov-Methoden für lineare Gleichungssysteme)

- Numerische Lösung gewöhnlicher Differentialgleichungen.

Die folgenden Bemerkungen beziehen sich auf die gedankliche Einstellung eines Teilnehmers auf diese Lehrveranstaltung:

- Es sei betont, daß in einer einführenden Vorlesung nur eine Heranführung an grundlegende Fragestellungen erfolgen kann. Es sollen Grundaufgaben und tragende Ideen der Numerischen Mathematik anhand ausgewählter Verfahren dargestellt werden, keineswegs geht es um die Darbietung einer ganzen Palette möglicher Verfahren. Bei der Behandlung der hier ausgewählten Methoden geht es um die Einführung in und die solide Vermittlung von grundlegenden Methoden und Denkweisen der Numerischen Mathematik. Dies gebieten die stürmische Entwicklung der Rechentechnik und der Numerischen Mathematik selbst. So hat die Entwicklung paralleler Rechenanlagen eine grundlegende Neubewertung numerischer Verfahren erzwungen.

- Diese einführende Vorlesung wendet sich vorwiegend an Student(inn)en der Mathematik, Physik und Informatik. Für Lehramtskandidat(inn)en mit Fach Mathematik wird eine gesonderte, für diesen Kreis verpflichtende Lehrveranstaltung zum Integrationsgebiet *Schulbezogene angewandte Mathematik, Modellbildung und Informatik* angeboten. Vorliegende Vorlesung sollte ergänzend absolviert werden, wenn eine Vertiefung auf dem Gebiet der Numerischen und Angewandten Mathematik beabsichtigt ist.

- Ohne eigenständige Erfahrungen bei der Umsetzung numerischer Verfahren auf einem Computer bleibt die Beschäftigung mit Numerischer Mathematik wenig sinnvoll. Erst dadurch werden Begriffe wie Stabilität, Konvergenz oder Effizienz eines Verfahrens transparent. Auch ist das Erfolgserlebnis einer tatsächlich selbst ermittelten Lösung nicht zu unterschätzen. Die Beschäftigung mit der Numerischen Mathematik gestattet darüber hinaus den Blick in das "Innenleben" eines Verfahrens und schärft den Blick für inhaltliche Vorzüge, Nachteile und auch Fehler eines Programms. Der eigenständige Umgang mit dem Computer ist schließlich eine normale Anforderung der Praxis an den Absolventen eines Universitätsstudiums.

- Für Student(inn)en) der Mathematik, Physik und Informatik ist eine solide Fertigkeit beim Programmieren sowie beim Umgang mit entsprechender Software unumgänglich. Wir werden im Rahmen der Veranstaltung das System MATLAB nutzen. Zahlreiche Übungsaufgaben sprechen daher diesen Punkt an.

Es sei hier nochmals hervorgehoben:

Numerische Mathematik ist eine wesentliche Disziplin der Mathematik.

Sie schlägt nicht nur die Brücke von Modellen, die in anderen mathematischen Grundgebieten besprochen werden, zur Anwendung in Ökonomie, Natur- und Geisteswissenschaften. Nur durch die Nutzung typisch mathematischer Methoden und einer entsprechenden Sprache (z.B. der Funktionalanalysis) auch in der Numerischen Mathematik ist die Entwicklung und Begründung brauchbarer Verfahren für die sich immer schneller entwickelnde Rechentechnik und die Bedürfnisse der Praxis möglich.

Als ergänzende Literatur zur Vorlesung können u.a. folgende Empfehlungen gegeben werden:

- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner, 2. Auflage, 2006.
- R. Kreß: *Numerical Analysis*, New York: Springer 1998.
- R. Schaback, H. Wendland: *Numerische Mathematik*, Springer-Verlag, 5. Auflage, 2004.
- H.R. Schwarz, N. Köckler: *Numerische Mathematik*, Teubner-Verlag Stuttgart, 2006.

Kapitel 1

Beispiele für Gleichungssysteme

Lineare und nichtlineare Gleichungssysteme nehmen innerhalb der Numerischen Mathematik eine zentrale Stellung ein. Einerseits führt die *Modellbildung* verschiedenartiger Vorgänge aus der Praxis (z.B. Berechnung von Schaltkreisen in der Elektronik, von Stabwerken in der Statik oder Bilanzierungsrechnungen in der Ökonomie) oft direkt auf derartige Aufgaben. Andererseits erhält man bei der numerischen Behandlung mathematischer Modelle (z.B. der Physik, Chemie, Biologie oder im Ingenieurwesen) oft über geeignete Diskretisierungsstrategien (nicht)lineare Gleichungssysteme, die in der Regel von großer Dimension sind.

Wir betrachten zunächst exemplarisch zwei Beispiele aus der erstgenannten Gruppe. Zugleich führen wir in diesen Beispielen die erforderliche Notation ein.

Beispiel 1.1. *Berechnung elektrischer Schaltungen*
Dieses Beispiel wurde entnommen aus G. Maeß: *Vorlesungen über Numerische Mathematik I*, Akademie-Verlag, Berlin 1984. Für die elektrische Schaltung in Abbildung 1.1 sollen die Spannungen U_j, j=1,...,n zwischen den Eckpunkten j und dem Mittelpunkt 0, d.h. dem gemeinsamen Rückleiter ermittelt werden.

Nach den Kirchhoffschen Gesetzen gilt für die Ströme in den Knotenpunkten

$$I_j = I_{j,0} + I_{j,j+1} - I_{j-1,j}$$

sowie für die Spannungen in den Maschen

$$U_{j,j+1} = U_j - U_{j+1}.$$

Nach dem Ohmschen Gesetz hat man für jeden Zweigstrom $U_j = R_{j,0} I_{j,0}$ bzw. $U_{j,j+1} = R_{j,j+1} I_{j,j+1}$. Der Index j wird zyklisch modulo n durchlaufen. Die Ersetzung der Größen I_{jl} in der Stromgleichung führt auf

$$\frac{U_j}{R_{j,0}} + \frac{U_{j,j+1}}{R_{j,j+1}} - \frac{U_{j-1,j}}{R_{j-1,j}} = I_j.$$

Schließlich erhält man im Falle $R_{jl} = R$ und nach Ersetzung der Größen U_{jl} mit $a = \frac{3}{R}, c = -\frac{1}{R}$ das System

$$\begin{array}{ccccccccccccc}
aU_1 & + & cU_2 & & & & & & & + & cU_n & = & I_1 \\
cU_1 & + & aU_2 & + & cU_3 & & & & & & & = & I_2 \\
 & & cU_2 & + & aU_3 & + & cU_4 & & & & & = & I_3 \\
 & & & & & & \vdots & & & & & \vdots & \vdots \\
 & & & & & & cU_{n-2} & + & aU_{n-1} & + & cU_n & = & I_{n-1} \\
cU_1 & & & & & & & + & cU_{n-1} & + & aU_n & = & I_n
\end{array}$$

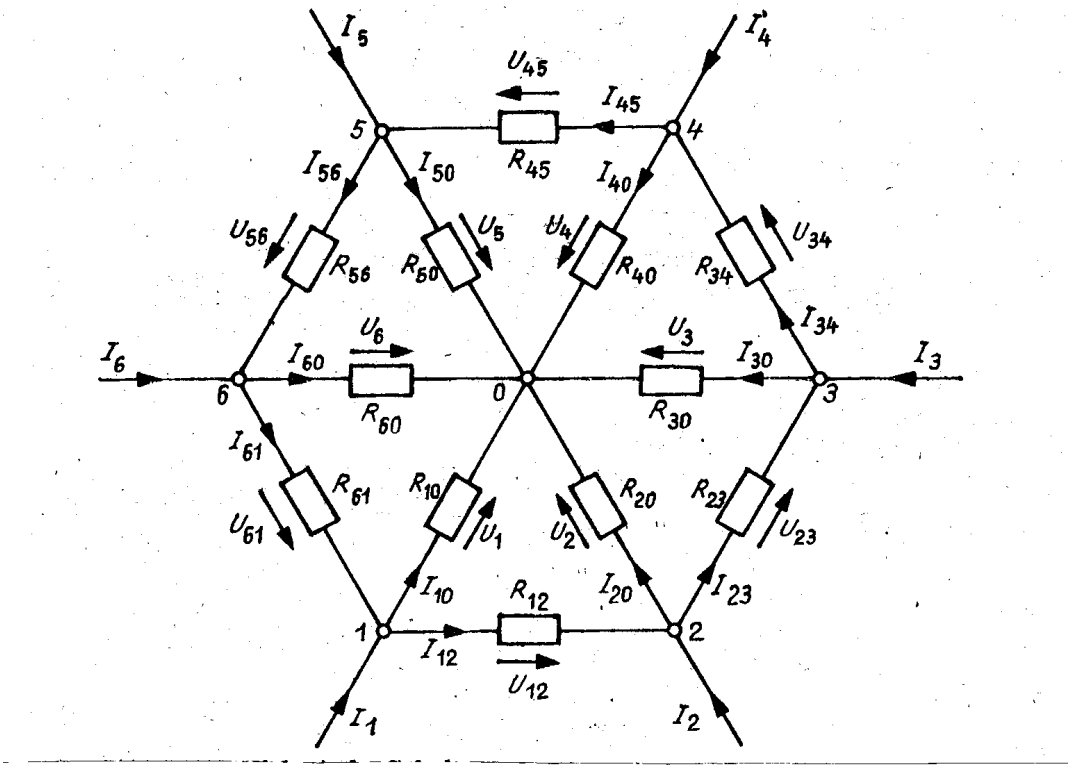

Abbildung 1.1: Elektrische Schaltung ($n = 6$)

oder in Kurzform

$$Au = b \tag{1.1}$$

mit symmetrischer Matrix

$$A = \begin{pmatrix} a & c & \cdot & \cdot & \cdot & \cdot & \cdot & c \\ c & a & c & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & c & a & c & \cdot & \cdot & \cdot & \cdot \\ & & & \ddots & \ddots & \ddots & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & c & a & c \\ c & \cdot & \cdot & \cdot & \cdot & \cdot & c & a \end{pmatrix} \in \mathbb{R}^{n\times n}$$

und den Vektoren $u = (U_1, ..., U_n)^T \in \mathbb{R}^n$ und $b = (I_1, I_2, ..., I_n)^T \in \mathbb{R}^n$.

Bei elektrischen Netzwerken in der Praxis kann die Anzahl n der Eckpunkte auch sehr groß sein. Die Matrix hat dann sehr viele Nullelemente, d.h. man erhält eine *schwachbesetzte* Matrix. □

Für eine $m \times n-$ Matrix $B = (b_{ij}),\ i = 1, ..., m,\ j = 1, ..., n$ mit Koeffizienten $b_{ij} \in \mathbb{K}^{n\times n}, \mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ bezeichnen wir mit B^T die *transponierte Matrix*, d.h.

$$B^T = (b_{ij})^T = (b_{ji}), \quad j = 1, ..., n, \quad i = 1, ..., m.$$

Ferner bezeichnet $B^* = \overline{B}^T$ die *adjungierte Matrix*, d.h. die zur Matrix $\overline{B}$ mit konjugiert komplexen Einträgen transponierte Matrix. Speziell sind die Transponierte bzw. Adjungierte zu einem Zeilenvektor damit Spaltenvektoren.

Beispiel 1.2. *Interpolation. Methode der kleinsten Quadrate*

Wir nehmen an, daß eine bestimmte Größe u in Abhängigkeit von der Zeit t und ausgewählten Parametern $a = (a_0, ..., a_p)^T \in \mathbb{R}^{p+1}$ durch die gegebene Funktion

$$g(t, a) = g(t; a_0, \ldots, a_p) \tag{1.2}$$

modelliert werde. Durch eine Meßreihe für die Größe u zu den Zeitpunkten $t_0, , ..., t_n$ soll der Parametervektor a ermittelt werden. Dies führt zu den *Interpolationsbedingungen*

$$u(t_j) = g(t_j; a), \qquad j = 0, ..., n. \tag{1.3}$$

Im Fall $p = n$ ist dies ein in der Regel nichtlineares System von $n + 1$ Gleichungen für die Parameter $a_0, ..., a_n$. Im Rahmen dieser Vorlesung behandeln wir insbesondere den Spezialfall der *polynomialen Interpolation*, d.h. für den Spezialfall

$$g(t; a) := \sum_{i=0}^{n} a_i l_i(t) \tag{1.4}$$

mit einer geeigneten Basis $\{l_0(t), \ldots, l_n(t)\}$ für die Menge $\mathbb{P}_n$ der Polynome vom maximalen Grad $n \in \mathbb{N}$. Speziell gilt $l_i(t) = t^i$ für die Monom-Basis. Die Interpolationsbedingungen (1.3) führt dann auf ein lineares Gleichungssystem

$$\sum_{i=0}^{n} l_i(t_j) a_i = u(t_j), \quad j = 0, \ldots, n$$

für die Koeffizienten $a_0, \ldots, a_n$. Wir gehen hierauf in Teil II der Vorlesung ein.

Zur Einschränkung des Einflusses von Meßfehlern wählt man oft $n > p$. Ein sinnvolles Bestimmungskriterium für a ist nun die Minimierung der Abweichungen $u(t_j) - g(t_j; a)$ im quadratischen Mittel, d.h.

$$f(a) := \frac{1}{2} \sum_{j=0}^{n} [u(t_j) - g(t_j, a)]^2 \ \rightarrow \text{Min.} \ \ !.$$

Daraus erhält man als notwendige Bedingungen die Normalgleichungen der *Methode der kleinsten Quadrate*

$$\frac{\partial f}{\partial a_k} = \sum_{j=0}^{n} [g(t_j; a) - u(t_j] \frac{\partial g(t_j; a)}{\partial a_k} \ = \ 0, \quad k = 0, ..., p \tag{1.5}$$

d.h. ein $(p + 1)$-dimensionales, in der Regel nichtlineares Gleichungssystem für den Parametervektor $(a_0, , ..., a_p)$. Im Fall $n \rightarrow \infty$ erhält man statt der Summe in (1.5) ein Integral. □

Wir wenden uns jetzt Beispielen aus der oben genannten zweiten Gruppe zu. Speziell gelangt man zu (nicht)linearen Gleichungssystemen bei der Diskretisierung von *Differentialgleichungsmodellen*, insbesondere von Randwertaufgaben gewöhnlicher und partieller Differentialgleichungen. Nachfolgend beschränken wir uns auf einige elementare Fälle.

Beispiel 1.3. *Diskretisierung eines Zweipunkt-Randwertproblems*

Gesucht wird eine zweimal stetig differenzierbare Funktion $u : [0, 1] \mapsto \mathbb{R}$, die bei gegebener Funktion $F : [0, 1] \times \mathbb{R} \mapsto \mathbb{R}$ und Konstante $a \in \mathbb{R}^+$ dem Randwertproblem

$$-au''(x) \ = \ F(x, u(x)), \quad x \in (0, 1), \tag{1.6}$$

$$u(0) \ = \ u(1) = 0 \tag{1.7}$$

genügt. Ein derartiges Problem erhält man zum Beispiel bei der eindimensional modellierten Diffusionsgleichung für die Konzentration $u(\cdot)$ einer Substanz unter Beachtung einer chemischen Reaktion. Der links stehende Term der Differentialgleichung (1.6) modelliert die Konzentrationsänderung durch Diffusion, speziell ist a der materialabhängige Diffusionskoeffizient. Der nichtlineare Term $F(x,\cdot)$ steht für die chemische Reaktionsrate, zum Beispiel

$$F(x,u) = \alpha(x)u^\beta - \gamma(x) \quad \text{oder} \quad F(x,u) = \alpha(x)\exp(-\frac{\beta(x)}{u}).$$

Auch die Modellierung der Schwingungen eines fest eingespannten Stabes oder der Wärmeleitung in einem endlichen Stab führen auf die oben genannte Aufgabe.

Zur Näherungslösung von (1.6),(1.7) mit einem Finite-Differenzen Verfahren gehen wir von einer äquidistanten Zerlegung des Intervalls $[0,1]$ gemäß

$$x_j = jh, \quad j = 0,...,n+1, \quad n \in \mathbb{N}$$

mit der Schrittweite $h = \frac{1}{n+1}$ aus. Wir ersetzen in den inneren Gitterpunkten $x_j, j = 1,...,n$ den Differentialquotienten durch den zentralen Differenzenquotienten

$$u''(x_j) \approx \frac{1}{h^2}\{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})\}$$

und bestimmen die Näherungswerte u_j an die Werte $u(x_j)$ aus dem System

$$-u_{j-1} + 2u_j - u_{j+1} = \frac{h^2}{a}F(x_j,u_j), \quad j = 1,...,n. \tag{1.8}$$

Hinzu kommen gemäß (1.7) die Randbedingungen

$$u_0 = u_{n+1} = 0. \tag{1.9}$$

In verkürzter Form schreiben wir das nichtlineare Gleichungssystem als

$$Au = b(u) \tag{1.10}$$

mit der symmetrischen $n \times n-$Tridiagonal-Matrix

$$A = \begin{pmatrix} 2 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & -1 & 2 & -1 & \cdot & \cdot & \cdot \\ & & \ddots & \ddots & \ddots & & \\ \cdot & \cdot & \cdot & \cdot & -1 & 2 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 2 \end{pmatrix}$$

und den Vektoren

$$u = (u_1,...,u_n)^T, \qquad b(u) = \frac{h^2}{a}(F(x_1,u_1),...,F(x_n,u_n))^T. \qquad \square$$

Beispiel 1.4. *Diskretisierung eines elliptischen Randwertproblems*

Wir betrachten noch exemplarisch ein Beispiel, das aus der Diskretisierung eines Randwertproblems einer sehr einfachen partiellen Differentialgleichung (vom elliptischen Typ) resultiert.

Eine in einem Gebiet $\Omega \subset \mathbb{R}^2$ zweifach stetige differenzierbare Funktion $u : \overline{\Omega} \mapsto \mathbb{R}$ genüge dem Randwertproblem für die sogenannte *Poisson-Gleichung*

$$-(\Delta u)(x) = -\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right)(x) = f(x), \quad x \in \Omega \subset \mathbb{R}^2 \tag{1.11}$$

mit der Randbedingung

$$u(x) = 0, \quad x \in \partial\Omega. \tag{1.12}$$

Dieses Problem erhält man bei verschiedenen physikalischen Modellen (z.B. Schwingungen einer fest eingespannten Membran oder Wärmeleitung in einer Platte).

Wir betrachten vereinfachend die numerische Lösung des Problems im Einheitsquadrat $\Omega = (0,1) \times (0,1)$ auf einem äquidistantem Gitter der Schrittweite $h = 1/(n+1)$ und den Gitterpunkten

$$x_{ij} = (ih, jh), \quad i,j = 0, ..., n+1, \quad n \in \mathbb{N}.$$

Wie in Beispiel 1.3 approximieren wir die Ableitungen durch zentrale Differenzenquotienten:

$$\begin{aligned} \frac{\partial^2 u}{\partial x_1^2}(x_{ij}) &\approx \frac{1}{h^2}\{u(x_{i+1,j}) - 2u(x_{i,j}) + u(x_{i-1,j})\} \\ \frac{\partial^2 u}{\partial x_2^2}(x_{ij}) &\approx \frac{1}{h^2}\{u(x_{i,j-1}) - 2u(x_{i,j}) + u(x_{i,j+1})\}; \end{aligned}$$

damit ergibt sich als Approximation an den Laplace-Operator

$$(\Delta u)(x_{ij}) \approx \frac{1}{h^2}\{u(x_{i+1,j}) + u(x_{i-1,j}) + u(x_{i,j-1}) + u(x_{i,j+1}) - 4u(x_{i,j})\}.$$

Recht anschaulich ist die Bezeichnung *Differenzenstern* über die Gewichte der Differenzenquotienten in den zu x_{ij} benachbarten Gitterpunkten

$$\begin{array}{ccc} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{array}.$$

Für die inneren Knotenpunkte $x_{ij}, \quad i,j = 1, ..., n$ bestimmt man nun aus dem folgenden linearen Gleichungssystem Näherungswerte u_{ij} von $u(x_{ij})$

$$-u_{i-1,j} - u_{i+1,j} + 4u_{i,j} - u_{i,j-1} - u_{i,j+1} = h^2 f(x_{ij}), \quad i,j = 1, ..., n. \tag{1.13}$$

Hinzufügen muß man die Randbedingungen

$$u_{0,j} = u_{n+1,j} = 0, \quad j = 0, ..., n+1; \qquad u_{i,0} = u_{i,n+1} = 0, \quad i = 1, ..., n \tag{1.14}$$

für Knotenpunkte auf dem Rand $\partial\Omega$ des Gebietes.

Bei zeilenweiser (bzw. lexikographischer) Numerierung der gesuchten Werte u_{ij} mit

$$u_1 = u_{11},\ u_2 = u_{21},\ \ldots,u_n = u_{n,1},\ u_{n+1} = u_{1,2},\ \ldots,u_{n^2} = u_{n,n}$$

gelangt man zu einer kompakten Schreibweise des Systems mit Hilfe der Blocktridiagonalmatrix

$$A = \begin{pmatrix} B & -I & & \cdots & & & \\ -I & B & -I & & \cdots & & \\ & -I & B & -I & & & \\ & & \ddots & \ddots & \ddots & & \\ & & \cdots & & & -I & B & -I \\ & & \cdots & & & & -I & B \end{pmatrix} \in \mathbb{R}^{n^2 \times n^2}$$

mit der Einheitsmatrix $I \in \mathbf{R}^{n\times n}$ und der Tridiagonalmatrix

$$B = \begin{pmatrix} 4 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & 4 & -1 & & \cdot & \cdot & \cdot \\ \cdot & -1 & 4 & -1 & \cdot & \cdot & \cdot \\ & & \ddots & \ddots & \ddots & & \\ \cdot & \cdot & \cdot & \cdot & -1 & 4 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 4 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Ferner benutzen wir die Vektoren $u = (u_1, \ldots, u_{n^2})^T \in \mathbb{R}^{n^2}$ sowie $b = h^2(f_1, \ldots, f_{n^2})^T \in \mathbb{R}^{n^2}$ mit sinngemäßer Definition von f_i. Dann lautet das System unter Beachtung der Randbedingungen

$$Au = b.$$

Die Diskretisierung von Randwertproblemen partieller Differentialgleichungen führt oft auf Gleichungssysteme hoher Dimension. Im Beispiel ist für $n = 10^2$ die Dimension bereits ca. 10^4. Das entsprechende Problem in einem Einheitswürfel mit $n = 10^2$ hat schon ca. 10^6 Unbekannte. Trotz in der Regel vorliegender schwacher Besetztheit der Matrix A, d.h. es gibt im Verhältnis zur Gesamtzahl der Matrixeinträge sehr viele Nullelemente, sind effektive Lösungsverfahren erforderlich. □

Wir werden in Teil II der Vorlesung bei weiteren Grundaufgaben der Numerischen Mathematik (zum Beispiel bei Interpolationsproblemen oder Quadraturverfahren zur Bestimmung von Integralen) immer wieder auf Gleichungssysteme stoßen. Abschließend skizzieren wir den *Plan* für den Teil I dieser Vorlesung: Zunächst behandeln wir für *lineare* Gleichungssysteme

- *direkte* Auflösungsverfahren mit exakter Berechnung des Lösungsvektors in endlich vielen Schritten (abgesehen von Rundungsfehlern) (vgl. Kapitel 2),
- *Kleinste-Quadrate-Methode* für überbestimmte lineare Systeme (vgl. Kapitel 3),
- *iterative* Auflösungsverfahren mit schrittweiser Näherungsrechnung des Lösungsvektors durch wiederholte Anwendung einer bestimmten Rechenvorschrift (bis zum Abbruch mittels eines geeigneten Kriteriums) (vgl. Kapitel 6).

Zur Analyse iterativer Lösungsverfahren für lineare (und nichtlineare) Gleichungssysteme stellen wir in den Kapiteln 4 und 5 elementare Mittel der Funktionalanalysis bereit. Die numerische Lösung *nichtlinearer* Gleichungssysteme behandeln wir in den Kapiteln 7 und 8.

Kapitel 2

Faktorisierungsverfahren für lineare Gleichungssysteme

Ein *lineares Gleichungssystem* besteht aus m linearen Gleichungen in n Unbekannten $x_1, \ldots, x_n$. Bei gegebenen Koeffizienten a_{ij}, y_i mit $i = 1, \ldots, m,\ j = 1, \ldots, n$ hat ein derartiges System die allgemeine Form

$$\sum_{j=1}^{n} a_{ij} x_j = y_i, \quad i = 1, ..., m. \tag{2.1}$$

In Kurzform benutzt man die Matrixnotation

$$Ax = y \tag{2.2}$$

mit dem gesuchten Lösungsvektor $x = (x_1, ..., x_n)^T \in \mathbb{K}^n$, der Matrix $A = (a_{ij}) \in \mathbb{K}^{m \times n}$ und der rechten Seite $y = (y_1, ..., y_m)^T \in \mathbb{K}^m$. Hierbei ist entweder $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$.

Im Standardfall $m = n$ heißt die quadratische Matrix $A \in \mathbb{K}^{n \times n}$ *regulär* (bzw. *nichtsingulär*), falls $\text{rang}(A) = n$. Für eine reguläre Matrix A hat man aus der Linearen Algebra die folgenden grundlegenden Aussagen:
Es existiert die inverse Matrix A^{-1} mit $A^{-1}A = I_n = AA^{-1}$ und die Determinante $\det(A)$ ist von Null verschieden. Weiterhin sind die Spalten- und Zeilenvektoren von A linear unabhängig. Das Gleichungssystem (2.2) ist für jede rechte Seite y eindeutig lösbar. Das *homogene* System mit $y = 0$ hat nur die triviale Lösung.
Hat im allgemeinen Fall das System (2.2) mit $A \in \mathbb{K}^{m \times n}$ und $y \in \mathbb{K}^m$ keine (eindeutige) Lösung, so kann man einen Vektor $x \in \mathbb{K}^n$ so suchen, daß Ax möglichst dicht bei y liegt. Dazu könnte man die Funktion $f(x) := \|Ax - y\|_2$ minimieren, wobei die Euklidische Norm durch

$$\|z\|_2 := \left(\sum_{j=1}^{m} |z_j|^2 \right)^{\frac{1}{2}}, \qquad z = (z_1, \ldots, z_m)^T \in \mathbb{K}^m$$

gegeben ist. Man spricht dann von einem *linearen Ausgleichsproblem*

$$\|Ax - y\|_2 \to \min, \qquad x \in \mathbb{K}^n. \tag{2.3}$$

Dabei heißt x Lösung von (2.3), falls

$$\|Ax - y\|_2 \leq \|A\tilde{x} - y\|_2, \qquad \forall \tilde{x} \in \mathbb{K}^n.$$

In den folgenden Abschnitten besprechen wir *Faktorisierungsverfahren* für den Standardfall $m = n$. In Kapitel 3 behandeln wir dann lineare Ausgleichsprobleme für den allgemeinen Fall.

2.1 Gauß-Eliminationsverfahren. LU-Zerlegung

Das *Eliminationsverfahren* nach Gauß ist neben dem *Householder-Verfahren* das wichtigste direkte Auflösungsverfahren. Das Grundprinzip besteht darin, das System durch fortlaufende Elimination in ein äquivalentes Gleichungssystem in Dreiecksgestalt, ein sogenanntes *gestaffeltes System* zu überführen. Sei vorläufig $m = n$.

2.1.1 LU-Zerlegung

Im einfachsten Fall ist $A \in \mathbb{K}^{n\times n}$ bereits eine Dreiecksmatrix.

Definition 2.1. *Eine Matrix* $L = (l_{ij})_{i,j=1,\dots,n} \in \mathbb{K}^{n\times n}$ *heißt* linke untere Dreiecksmatrix, *falls* $l_{ij} = 0$ *für* $j > i$. *Eine Matrix* $U = (u_{ij})_{i,j=1,\dots,n} \in \mathbb{K}^{n\times n}$ *heißt* rechte obere Dreiecksmatrix, *falls* $u_{ij} = 0$ *für* $j < i$.

Das System $Lx = y$ bzw.

$$\begin{array}{ccccccccc} l_{11}x_1 & & & & & & & = & y_1 \\ l_{21}x_1 & + & l_{22}x_2 & & & & & = & y_2 \\ \vdots & & \vdots & & \ddots & & & & \vdots \\ l_{n1}x_1 & + & l_{n1}x_2 & + & \dots & + & l_{nn}x_n & = & y_n. \end{array}$$

mit $l_{ii} \neq 0$ für $i = 1, \dots, n$ löst man durch

Vorwärtselemination:

Initialisierung: $L \in \mathbb{K}^{n\times n}$ ist reguläre linke untere Dreiecksmatrix, $y \in \mathbb{K}^n$.

for $i = 1, \dots, n$ **do**

$$x_i = \left(y_i - \textstyle\sum_{j=1}^{i-1} l_{ij}x_j\right)/l_{ii}$$

end

Ergebnis: $x = L^{-1}y$

Andererseits löst man das System $Ux = y$ bzw.

$$\begin{array}{ccccccccc} u_{11}x_1 & + & u_{12}x_2 & + & \dots & + & u_{1n}x_n & = & y_1 \\ & & u_{22}x_2 & + & \dots & + & u_{2n}x_n & = & y_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & u_{nn}x_n & = & y_n. \end{array}$$

im Fall $u_{ii} \neq 0$ für $i = 1, \dots, n$ durch

Rückwärtselimination:

Initialisierung: $U \in \mathbb{K}^{n\times n}$ ist reguläre rechte obere Dreiecksmatrix, $y \in \mathbb{K}^n$.

for $i = n, n-1, \dots, 1$ **do**

$$x_i = \left(y_i - \textstyle\sum_{j=i+1}^{n} u_{ij}x_j\right)/u_{ii}$$

end

Ergebnis: $x = U^{-1}y$

Definition 2.2. *Eine Zerlegung einer Matrix $A \in \mathbb{K}^{n\times n}$ der Form*

$$A = LU \tag{2.4}$$

mit einer linken unteren Dreiecksmatrix $L \in \mathbb{K}^{n\times n}$ und einer rechten oberen Dreiecksmatrix $U \in \mathbb{K}^{n\times n}$ heißt LU-Zerlegung von A.

Ist eine LU-Zerlegung von A bekannt, so gewinnt man die Lösung von

$$Ax = (LU)x = L(Ux) = y$$

durch

- Lösung von $Lz = y$ durch Vorwärtselimination
- Lösung von $Ux = z$ durch Rückwärtselimination.

Nachfolgend beschreiben wir die Bestimmung der Matrizen L und U durch Gauß-Elimination.

2.1.2 Gauß-Elimination ohne Spaltenpivotisierung

Wir betrachten zuerst die Matrix-Version der Gauß-Elimination. Sie ist aus theoretischer Sicht interessant, jedoch für die praktische Implementierung ungeeignet.

Definition 2.3. *Für einen gegebenen Vektor $l^{(k)} = (0,\dots,0,t_{k+1},\dots,t_n)^T \in \mathbb{K}^n$ mit $k \in \{1,\dots,n\}$ und den $k-$ten Einheitsvektor $e_k = (0,\dots,0,1,0,\dots,0)^T \in \mathbb{K}^n$ ist die* Gauß-Matrix *M_k definiert durch*

$$M_k := I_n - l^{(k)}e_k^T = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -t_{k+1} & 1 & \\ & & \vdots & & \ddots \\ & & -t_n & & 1 \end{pmatrix}. \tag{2.5}$$

Eine Gauß-Matrix M_k unterscheidet sich von der Einheitsmatrix I_n lediglich in der $k-$ten Spalte, hier auch nur von den Elementen unterhalb des Diagonaleintrags. Gauß-Matrizen sind nichtsingulär. Wegen $[l^{(k)}]^T e_k = 0$ ist

$$M_k^{-1} = I_n + l^{(k)}e_k^T. \tag{2.6}$$

Bei Linksmultiplikation einer Matrix mit n Zeilen mit M_k bleiben die ersten k Zeilen invariant. Für $i = k+1,\dots,n$ erhält man die neue $i-$te Zeile durch Subtraktion der mit t_i multiplizierten k-ten Zeile von der alten $i-$ten Zeile.

Den für den Gauß-Algorithmus wesentlichen Schritt erkennt man bei Linksmultiplikation eines Vektors $x \in \mathbb{K}^n$ mit $x_i \neq 0$ für $i = 1,\dots,n$ mit der Gauß-Matrix M_k zum gegebenen Vektor $l^{(k)}$ mit $t_i := x_i/x_k$ für $i = k+1,\dots,n$:

$$M_k x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

d.h. die Werte des Vektors x unterhalb des Diagonaleintrags werden annulliert. Nun kann die Matrix-Version der Gauß-Elimination unter Verwendung von Gauß-Matrizen wie folgt formuliert werden:

Gauß-Elimination ohne Spaltenpivotisierung (Matrix-Version):

Initialisierung: $A \in \mathbb{K}^n$;
$A^{(1)} := A$;

for $k = 1, \ldots, n-1$ **do**

$$l^{(k)} := \left(\underbrace{0, \ldots, 0}_{k-mal}, \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}}, \ldots, \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} \right)^T;$$

$M_k := I_n - l^{(k)} e_k^T$;

$A^{(k+1)} := M_k A^{(k)}$

end

Ergebnis: $U := A^{(n)}$ ist rechte obere Dreiecksmatrix, $L := M_1^{-1} \cdots M_{n-1}^{-1}$ ist linke untere Dreiecksmatrix. Es gilt $A = LU$.

Die Durchführbarkeit dieses Verfahrens rechtfertigt der folgende Satz.

Satz 2.4. *Unter der Voraussetzung*

$$det \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} \neq 0, \quad k = 1, \ldots, n-1 \tag{2.7}$$

gelten folgende Aussagen:

1. *Die Gauß-Elimination ist wegen $a_{kk}^{(k)} \neq 0$ für $k = 1, \ldots, n-1$ ohne Spaltenpivotisierung durchführbar. Es gilt*
$$a_{ij}^{(k+1)} = 0 \text{ für } j < k+1 \text{ und } i > j. \tag{2.8}$$
ferner ist U rechte obere Dreiecksmatrix.

2. *Es ist $A = LU$.*

3. *$L = I_n + \sum_{k=1}^{n-1} l^{(k)} e_k^T$ ist die linke untere Dreiecksmatrix aus der Zerlegung.*

Beweis: (1) Nach Voraussetzung (2.7) ist $a_{11}^{(1)} \neq 0$, d.h. der erste Schritt für $k = 1$ ist wohldefiniert. Nach Wahl von M_1 ist (2.8) erfüllt wegen

$$A^{(2)} = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}.$$

Insbesondere ist

$$\begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ 0 & a_{22}^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -t_2^{(1)} & 1 \end{pmatrix} \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{22}^{(1)} & a_{22}^{(1)} \end{pmatrix}$$

und daher

$$a_{11}^{(2)} a_{22}^{(2)} = \det \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{22}^{(1)} & a_{22}^{(1)} \end{pmatrix} \neq 0.$$

Damit ist $a_{22}^{(2)} \neq 0$, d.h. der erste Schritt für $k = 2$ ist ebenfalls wohldefiniert. Aus der Wahl von M_2 und Voraussetzung (2.7) für $k = 1$ folgert man (2.8) für $k = 2$. Für $k \geq 3$ geht man analog per Induktion vor.

(2) Die gesuchte Aussage $A = LU$ ersieht man sofort aus dem Algorithmus wegen

$$A^{(n)} = M_{n-1} M_{n-2} \cdots M_1 A.$$

(3) Wir wissen bereits, daß $M_k^{-1} = I_n + l^{(k)} e_k^T$. Nun zeigen wir noch mittels Induktion nach j, daß

$$M_{n-j}^{-1} \cdots M_{n-1}^{-1} = I_n + \sum_{k=n-j}^{n-1} l^{(k)} e_k^T.$$

Für $j = 1$ hatten wir die Behauptung schon gesehen. Der Induktionsschluß ergibt sich aus

$$\begin{aligned} M_{n-j}^{-1} \left(M_{n-j+1}^{-1} \cdots M_{n-1}^{-1} \right) &= \left(I_n + l^{(n-j)} e_{n-j}^T \right) \left(I_n + \sum_{k=n-j+1}^{n-1} l^{(k)} e_k^T \right) \\ &= I_n + \sum_{k=n-j}^{n-1} l^{(k)} e_k^T + l^{(n-j)} \sum_{k=n-j+1}^{n-1} \underbrace{(e_{n-j}^T l^{(k)})}_{=0} e_k^T. \end{aligned}$$

Damit ist der Beweis ausgeführt. □

Lemma 2.5. *Die reguläre Matrix $A \in \mathbb{K}^{n \times n}$ besitze eine LU-Zerlegung $A = LU$. Die Matrix L sei so normiert, daß auf der Hauptdiagonale nur Einseinträge stehen. Dann ist die LU-Zerlegung eindeutig bestimmt.*

Beweis: Seien $A = L_1 U_1 = L_2 U_2$ zwei LU-Zerlegungen. In der hieraus folgenden Form

$$L_2^{-1} L_1 = U_2 U_1^{-1}$$

steht links eine linke untere Dreiecksmatrix mit Einseinträgen auf der Hauptdiagonalen und rechts eine obere Dreiecksmatrix. Gleichheit kann nur gelten bei

$$L_2^{-1} L_1 = U_2 U_1^{-1} = I_n.$$

daraus folgt aber über $L_2 = L_1$ und $U_2 = U_1$ die Eindeutigkeit der Zerlegung. □

Bemerkung 2.6. Unter den Voraussetzungen von Lemma 2.5 erhält man nach dem Multiplikationssatz für Determinanten folgende einfache Möglichkeit zur Determinantenberechnung:

$$\det(A) = \det(L) \cdot \det(U) = u_{11} \cdots u_{nn}. \qquad \square$$

2.1.3 Gauß-Elimination mit Spaltenpivotisierung

Offenbar ist die bisherige Form der Gauß-Elimination selbst bei regulärer Matrix nicht immer durchführbar. Man vertausche nur zwei Zeilen in $A = I_n$. Aber auch im Falle der Wohldefiniertheit kann man ungewünschte Effekte erzielen.

Beispiel 2.7. Für beliebige $\epsilon > 0$ gilt die LU-Zerlegung

$$A_\epsilon := \begin{pmatrix} \epsilon & 1 \\ 1 & \epsilon \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{pmatrix} \begin{pmatrix} \epsilon & 1 \\ 0 & \epsilon - 1/\epsilon \end{pmatrix}.$$

Die Einträge von L und U werden für $\epsilon \to +0$ beliebig groß; im Ergebnis treten erhebliche Rundungsfehler auf. □

Bei der *Spaltenpivotisierung* vertauscht man Zeilen derart, daß die Gauß-Elimination bei regulärer Matrix A wohldefiniert ist. Konkret vertauscht man im k-ten Schritt die k-te mit der j-ten Zeile nach der Regel

$$\text{Wähle } j \in \{k, \ldots, n\}: \quad |a_{jk}^{(k)}| \geq |a_{lk}^{(k)}|, \quad l = k, \ldots, n.$$

Dann heißt $a_{jk}^{(k)}$ auch *Pivotelement*.

Bei der Gauß-Elimination mit Spaltenpivotsuche wird eine reguläre Matrix A in $n-1$ Schritten abwechselnd von links mit *Vertauschungs-* und *Gauß-Matrizen* multipliziert und so in eine obere Dreiecksmatrix überführt. Man erhält bei der LU-Zerlegung von A eine Permutationsmatrix P, eine untere Dreiecksmatrix L mit Eins-Einträgen auf der Diagonalen und eine obere Dreiecksmatrix U mit

$$PA = LU. \tag{2.9}$$

Eine *Permutationsmatrix* entsteht aus der Einheitsmatrix I_n durch Vertauschung von Zeilen bzw. Spalten. Für eine Permutation $p = (p(1), ..., p(n))$ der Zahlen $1, ..., n$ hat sie die Gestalt

$$P = \begin{pmatrix} e_{p(1)}^T \\ \vdots \\ e_{p(n)}^T \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{2.10}$$

Es gilt

$$\begin{pmatrix} e_{p(1)}^T \\ \vdots \\ e_{p(n)}^T \end{pmatrix} = \begin{pmatrix} e_{q(1)} & \ldots & e_{q(n)} \end{pmatrix}$$

mit der zu p inversen Permutation $q := p^{-1}$. Offenbar sind Permutationsmatrizen orthogonal, d.h. es gilt

$$P^T P = P P^T = I.$$

Werden lediglich zwei Zeilen bzw. Spalten der Einheitsmatrix vertauscht, so spricht man von einer *Vertauschungsmatrix*. Vertauscht man genauer die $r-$te und $s-$te Zeile der Einheitsmatrix mit $1 \leq r \leq s \leq n$, so gilt dies ebenso für die Spalten. Die zugehörige Vertauschungsmatrix hat die Gestalt

$$P_{rs} = \begin{pmatrix} e_1 & \ldots & e_{r-1} & e_s & e_{r+1} & \ldots & e_{s-1} & e_r & e_{s+1} & \ldots & e_n \end{pmatrix} = P_{rs}^T.$$

Somit ist P_{rs} eine symmetrische Permutationsmatrix, die man schreiben kann als

$$P_{rs} = I - (e_r - e_s)(e_r - e_s)^T. \tag{2.11}$$

Bei Linksmultiplikation von $A \in \mathbb{K}^{n \times n}$ mit P_{rs} erfolgt eine Vertauschung der $r-$ten und $s-$ten Zeile von A. Bei Rechtsmultiplikation AP_{rs} vertauscht man die $r-$te und $s-$te Spalte.

Formal kann man die LU-Zerlegung wie folgt in Pseudo-Code beschreiben:

Gauß-Elimination mit Spaltenpivotisierung (Matrix-Version)

Initialisierung: $A \in \mathbb{K}^{n\times n}$ regulär;
$A^{(0)} := A;\ M^{(0)} = I;$

for $k = 1, \ldots, n$:

$\tilde{A}^{(k)} := M^{(k-1)} A^{(k-1)};$

Berechne Pivotindex $r(k) \in \{k, \ldots, n\}$ mit $|\tilde{a}_{rk}| = \max_{i=k,\ldots,n} |\tilde{a}_{ik}|;$

$P^{(k)} := P_{k,r(k)}$

$A^{(k)} := P^{(k)} \tilde{A}^{(k)};$

$l^{(k)} := (\underbrace{0, \ldots, 0}_{k-mal}, a^{(k)}_{k+1,k}, \ldots, a^{(k)}_{nk})^T / a^{(k)}_{kk};$

$M^{(k)} := I - l^{(k)} e_k^T;$

end

Ergebnis: Es gilt $PA = LU$ mit

$P := P^{(n)} \cdots P^{(1)},$

$U := A^{(n)}$ - rechte obere Dreiecksmatriix,

$L := I + \sum_{k=1}^{n-1} \theta^{(k)} e_k^T$ - linke untere Dreiecksmatrix mit

$\theta^{(k)} := P^{(n)} \cdots P^{(k+1)} l^{(k)}.$

Die Durchführbarkeit dieses Verfahrens rechtfertigt der folgende Satz.

Satz 2.8. *Für eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ existieren eine Permutationsmatrix $P \in \mathbb{R}^{n\times n}$, eine untere Dreiecksmatrix $L \in \mathbb{K}^{n\times n}$ mit Eins-Einträgen auf der Diagonalen sowie eine obere Dreiecksmatrix $U \in \mathbb{K}^{n\times n}$, so daß $PA = LU$.*

Beweis: Per Induktion nach $k = 0, 1, \ldots, n-1$ zeigen wir

(i) $A^{(k)} = P^{(k)} M^{(k-1)} P^{(k-1)} \cdots P^{(1)} A$ und $M^{(k)}$ sind wohldefiniert.

(ii) Für $l = k+1, \ldots, n$ gilt $|a^{(k+1)}_{k+1,k+1}| \geq |a^{(k+1)}_{l,k+1}|$.

(iii) Für $j < k+1$ und $i > j$ gilt $a^{(k+1)}_{ij} = 0$.

Für den Induktionsanfang mit $k = 0$ ist (i) erfüllt wegen $A^{(0)} = A,\ M^{(0)} = I$. Behauptung (ii) ist für $A^{(1)} = P^{(1)} A$ wegen der Wahl von $P^{(1)}$ erfüllt. Behauptung (iii) ist klar.

Wir führen den Induktionsschritt von $k-1$ nach k durch:
Nach Induktionsvoraussetzung ist $a^{(k)}_{kk} \neq 0$. Anderenfalls wäre $A^{(k)}$ singulär. Wegen $\det P^{(j)} = \pm 1$, $\det M^{(j)} = 1$ sowie der Induktionsvoraussetzung wäre dann $0 = \det A^{(k)} = \pm \det A$ im Widerspruch zur vorausgesetzten Regularität von A. Somit ist $M^{(k)}$ wohldefiniert, d.h. (i) ist gezeigt. Aussage (ii) gilt wegen der Wahl von $P^{(k+1)}$. Schließlich gilt (iii) wegen der Wahl von $M^{(k)}$, da $P^{(k+1)}$ die ersten k Zeilen nicht vertauscht. Damit ist der Induktionsbeweis geführt.

Gezeigt wurde, daß

$$U = P^{(n)} M^{(n-1)} P^{(n-1)} \cdots M^{(1)} P^{(1)} A$$

obere Dreiecksmatrix ist. Aus dem Beweis von Satz 2.4 wissen wir bereits, daß $[M^{j)}]^{-1} = I + l^{(j)} e_j^T$ gilt. Damit ist

$$P^{(1)}(I + l^{(1)} e_1^T) P^{(2)} (I + l^{(2)} e_2^T) P^{(3)} \cdots P^{(n-1)} (I + l^{(n-1)} e_{n-1}^T) P^{(n)} U = A.$$

Dies vereinfacht man wegen $e_j^T P^{(j+1)} = e_j^T$ zu

$$P^{(1)}(P^{(2)} + l^{(1)} e_1^T)(P^{(3)} + l^{(2)} e_2^T) \cdots (P^{(n)} + l^{(n-1)} e_{n-1}^T) U = A.$$

Linksmultiplikation mit $P = P^{(n)} \cdots P^{(1)}$ ergibt dann (nach einigen Umformungen)

$$(I + \theta^{(1)} e_1^T) \cdots (I + \theta^{(n-1)} e_{n-1}^T) U = PA.$$

Schließlich sieht man wie beim Beweis von Satz 2.4, daß der linke Faktor in dieser Gleichung mit L übereinstimmt. □

In der nachfolgenden matrixfreien Version des Gauß-Verfahrens mit Spaltenpivotisierung werden die Vektoren $l^{(k)}$ in die frei werdenden Spalten von A geschrieben. Bei Anwendung der Vertauschungen auf die Gesamtmatrix stehen am Schluß die Vektoren $\theta^{(k)}$ in der unteren Hälfte von A. Ferner wird ein Regularitätstest eingebaut.

Gauß-Elimination mit Spaltenpivotisierung

Initialisierung: $A \in \mathbb{K}^{n \times n}$;
$p_i = i$ für $i = 1, \ldots, n$.

for $k = 1, \ldots, n$ **do**

Berechne Pivotindex $r(k) \in \{k, \ldots, n\}$ mit $|a_{rk}| = \max_{i=k,\ldots,n} |a_{ik}|$;

if $a_{rk} = 0$

STOP: A ist singulär;

else

Vertausche p_r und p_k;
for $l = 1, \ldots, n$ **do**
Vertausche a_{kl} und a_{rl};
end;

end

for $i = k+1, \ldots, n$ **do**

$a_{ik} := a_{ik}/a_{kk}$;
for $j = k+1, \ldots, n$ **do**
$a_{ij} := a_{ij} - a_{ik} a_{kj}$;
end

end

end

if $a_{nn} = 0$

STOP: A ist singulär.

end

Ergebnis: Das Programm bricht mit entsprechender Fehlermeldung ab, wenn A singulär ist. Sonst wird A mit einer LU-Zerlegung $PA = LU$ wie folgt überschrieben:

$$l_{ij} = \begin{cases} \delta_{ij}, & \text{falls } i \le j, \\ a_{ij}, & \text{falls } i > j \end{cases}, \quad u_{ij} = \begin{cases} a_{ij}, & \text{falls } i \le j, \\ 0\,, & \text{falls } i > j \end{cases}, \quad p_{ij} = \delta_{p_i,j}.$$

2.1.4 Komplexität der LU-Zerlegung

Wir wollen nun den *Rechenaufwand* bzw. die *Komplexität* der LU-Zerlegung abschätzen.

Lemma 2.9. *Die LU-Zerlegung einer regulären Matrix $A \in \mathbb{K}^{n\times n}$ erfordert insgesamt $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ wesentliche Rechenoperationen (d.h. Multiplikationen und Divisionen, ohne Beachtung von Additionen und Subtraktionen sowie Vergleichen und Vertauschungen).*

Beweis: Die Zahl wesentlicher Rechenoperationen ohne Beachtung von Vergleichen und Vertauschungen ermittelt man aus

$$\begin{aligned} G(n) &= \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \left(1 + \sum_{j=k+1}^{n} 1 \right) = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} (n-k+1) \\ &= \sum_{k=1}^{n-1} (n-k)(n-k+1) = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n. \qquad \square \end{aligned}$$

Bemerkungen 2.10.
(i) Die Aufwandsabschätzung ist wesentlich bei Lösung (sehr) großer vollbesetzter Gleichungssysteme. So führt eine Verdopplung der Unbekannten zur Erhöhung der Rechenzeit um den Faktor 8. Bei großen schwachbesetzten Matrizen kann man bei der LU-Zerlegung die Besetzungsstruktur geschickt ausnutzen (vgl. Abschnitt 2.3). Bei sehr großer Dimension muß man iterative Verfahren benutzen.

(ii) In obigem Algorithmus wurde zur Vermeidung von Rundungsfehlern das betragsmäßig größte Element der Eliminationsspalte (sinngemäß der -zeile) gewählt (*Spalten- bzw. Zeilen-Pivotisierung*). Diese Methode erfordert zusätzlich $O(n^2)$ Operationen bei vollbesetzter Matrix. Man kann natürlich auch das betragsmäßig größte Element der noch zu verändernden Restmatrix ermitteln (*vollständige Pivotisierung*). Dies erfordert jedoch sogar $O(n^3)$ Operationen bei vollbesetzter Matrix.

(iii) Teilweise oder vollständige Pivotsuche sind nur für *äquilibrierte* Matrizen A sinnvoll, d.h. wenn alle Zeilen- und Spaltenbetragssummen etwa gleich groß sind. Sonst könnte man jede Zeile durch Multiplikation mit einem hinreichend großen Faktor zur Pivot- oder Eliminationszeile machen, sofern das potentielle Pivotelement nur von Null verschieden ist. Theoretisch ist die *Äquilibrierung* (oder *Skalierung*) durch Äquivalenztransformation $\hat{A} = D_1 A D_2$ mit regulären Diagonalmatrizen D_1 und D_2 erreichbar. Bei vollbesetzter Matrix A erhöht sich der Rechenaufwand dadurch um $2n^2$ Multiplikationen.

(iv) Wir betrachten noch den Fall, daß zugleich r rechte Seiten des Gleichungssystems betrachtet werden. Dieser Fall tritt zum Beispiel auf, wenn man die inverse Matrix zu A berechnet. Als rechte Seiten wählt man dann jeweils die entsprechenden Spalten der Einheitsmatrix. Ein anderer Anwendungsfall tritt etwa im Beispiel 1.1 auf, wenn verschiedene Lastfälle (d.h. für verschiedene Werte der rechten Seite) berechnet werden müssen. Dies erfordert rn^2 zusätzliche wesentliche Rechenoperationen. □

2.2 Cholesky-Zerlegung für symmetrische positiv definite Matrizen

Wir behandeln nun den wichtigen Spezialfall einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$, d.h. neben $A = A^T$ gilt die Aussage $x^T Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$. Äquivalente Bedingungen hierfür sind die Positivität aller Eigenwerte bzw. aller Hauptabschnittsdeterminanten von A.

Beispiele für symmetrische, positiv definite Matrizen hatten wir bereits in Kapitel 1 bei der Diskretisierung von Randwertproblemen gesehen. Ein anderer Anwendungsfall sind die *Normalengleichungen* bei linearen Ausgleichsproblemen, die wir in Kapitel 3 behandeln werden.

Die zentrale Aussage über die LU-Zerlegung derartiger Matrizen gibt der folgende Satz.

Satz 2.11. (*Cholesky-Zerlegung)*
Für eine symmetrische und positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ existiert genau eine untere Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen derart, daß $A = LL^T$ ist.

Beweis: Der Beweis erfolgt mittels vollständiger Induktion nach n. Zunächst ist die Behauptung für $n = 1$ korrekt. Sei jetzt angenommen, daß jede symmetrische, positiv definite Matrix aus $\mathbb{R}^{(m-1)\times(m-1)}$ eine Cholesky-Zerlegung besitzt. Eine Matrix aus $\mathbb{R}^{m\times m}$ werde zerlegt gemäß

$$A_m = \begin{pmatrix} A_{m-1} & a \\ a^T & \alpha \end{pmatrix} \tag{2.12}$$

mit symmetrischer, positiv definiter Matrix $A_{m-1} \in \mathbb{R}^{(m-1)\times(m-1)}$. Ein geeigneter Ansatz für die Dreiecksmatrix ist

$$L_m = \begin{pmatrix} L_{m-1} & 0 \\ l^T & \beta \end{pmatrix} \tag{2.13}$$

mit unterer Dreiecksmatrix $L_{m-1} \in \mathbb{R}^{(m-1)\times(m-1)}$. Dann gilt

$$\begin{aligned} L_m L_m^T &= \begin{pmatrix} L_{m-1} & 0 \\ l^T & \beta \end{pmatrix} \begin{pmatrix} L_{m-1}^T & l \\ 0^T & \beta \end{pmatrix} \\ &= \begin{pmatrix} L_{m-1}L_{m-1}^T & L_{m-1}l \\ (L_{m-1}l)^T & l^T l + \beta^2 \end{pmatrix} = \begin{pmatrix} A_{m-1} & a \\ a^T & \alpha \end{pmatrix} = A_m \end{aligned}$$

genau dann, wenn

$$L_{m-1}L_{m-1}^T = A_{m-1}, \qquad L_{m-1}l = a, \qquad l^T l + \beta^2 = \alpha. \tag{2.14}$$

Dann ist L_{m-1} der nach Induktionsannahme eindeutig bestimmte Cholesky-Faktor von A_{m-1}. Wegen der Regularität ist der Vektor $l \in \mathbb{R}^{m-1}$ aus $L_{m-1}l = a$ eindeutig bestimmt. Weiter ist

$$\begin{aligned} \alpha - l^T l &= \alpha - a^T L_{m-1}^{-T} L_{m-1}^{-1} a \;=\; \alpha - a^T A_{m-1}^{-1} a \\ &= \begin{pmatrix} -A_{m-1}^{-1}a \\ 1 \end{pmatrix}^T \begin{pmatrix} A_{m-1} & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} -A_{m-1}^{-1}a \\ 1 \end{pmatrix} \;>\; 0. \end{aligned}$$

Damit ist genau eine Zahl $\beta > 0$ bestimmt mit $l^T l + \beta^2 = \alpha$. Dies beweist die Induktionsbehauptung. □

Bemerkung 2.12. Man kommt bei der Cholesky-Zerlegung ohne Pivotisierung aus, da alle Elemente der Hauptdiagonale von A stets positiv sind. Der Speicherplatzbedarf kann gegenüber der LU-Zerlegung auf die Hälfte, der Rechenaufwand auf etwa $n^3/6$ reduziert werden. □

Man kann nun die untere Dreiecksmatrix L recht einfach aus $A = LL^T$ durch Koeffizientenvergleich ermitteln. Für $i \geq j$ ist

$$a_{ij} = (A)_{ij} = (LL^T)_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk} = \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj}$$

und damit

$$i = j: \quad a_{jj} = l_{jj}^2 + \textstyle\sum_{k=1}^{j-1} l_{jk}^2 \quad \Rightarrow \quad l_{jj} := \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right)^{\frac{1}{2}}, \tag{2.15}$$

$$i > j: \quad a_{ij} = l_{ij} l_{jj} + \textstyle\sum_{k=1}^{j-1} l_{ik} l_{jk} \quad \Rightarrow \quad l_{ij} := \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) / l_{jj}. \tag{2.16}$$

Hieraus folgt die Möglichkeit der sukzessiven Berechnung von L.

Cholesky-Zerlegung für symmetrische, positiv definite Matrizen

Initialisierung: Gespeichert ist die untere Hälfte der symmetrischen, positiv definiten Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, d.h. die Elemente a_{ij} mit $i \geq j$.

for $j = 1, \ldots, n$ **do**

  $a_{jj} := (a_{jj} - \sum_{k=1}^{j-1} a_{jk}^2)^{\frac{1}{2}}$;

  for $i = j+1, \ldots, n$ **do**

    $a_{ij} := (a_{ij} - \sum_{k=1}^{j-1} a_{ik} a_{jk}) / a_{jj}$;

  end

end

Ergebnis: Die untere Hälfte von A (einschließlich Diagonale) wird mit der Matrix L überschrieben. Es gilt $A = LL^T$.

Bemerkung 2.12. Oft ist es wünschenswert zu testen, ob eine gegebene symmetrische Matrix auch "numerisch" positiv definit ist. Weit weniger aufwendig als die Berechnung der Eigenwerte von A und dazu numerisch stabil ist die Anwendung des Cholesky-Verfahrens. Man prüft dabei, ob der Radikand $a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2$ größer als eine gegebene kleine Zahl ϵ ist. Als Indiz für die numerische Stabilität des Cholesky-Verfahrens mag gelten, daß sich die Einträge der unteren Dreiecksmatrix L einfach durch $|l_{ij}| \leq \sqrt{a_{ii}},\ i = 1, \ldots, n$ beschränken lassen. □

2.3 Schwachbesetzte Matrizen

Die bisher beschriebenen Dreieckszerlegungen sind bei sehr großen Matrizen A in der Regel ineffizient. Man kann jedoch versuchen, die LU-Zerlegung an Matrizen mit spezieller Struktur anzupassen. Ein in Anwendungen sehr häufig auftretender (und auch angestrebter) Fall

ist der der *schwachbesetzten* Matrizen. Das sind Matrizen, bei denen die Zahl der Nullelemente gegenüber der Zahl der nichtverschwindenden Einträge sehr stark überwiegt. Sie entstehen zum Beispiel bei der numerischen Lösung von linearen Randwertproblemen (vgl. Kapitel 1) mit Finite-Differenzen- oder Finite-Elemente-Verfahren.

Die Hoffnung, daß sich dann für schwachbesetzte Matrizen auch eine bezüglich des Rechenaufwandes effiziente Form der LU-Zerlegung angeben läßt, erfüllt sich leider i.a. Fall nicht. Praktisch erhält man in der Regel eine sehr starke Auffüllung der Matrizen L und U mit Nichtnullelementen, das sogenannte *"fill in"*.

Beispiel 2.13. Für

$$A = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 1 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

sind die Dreiecksmatrizen L und U voll besetzt. Man benutze hierzu z.B. den MATLAB-Befehl `[L,R,P]=lu(A)`. □

Bemerkung 2.14. Es gibt Pivotstrategien, die das "fill in "reduzieren. Eine wichtige Variante, die im Zusammenhang mit der *Vorkonditionierung* iterativer Verfahren auftritt, ist eine approximative $LU-$Zerlegung (ILU – *incomplete LU factorization*). Dann besteht die Hoffnung, daß bei geeigneten $\tilde{L}$ und $\tilde{U}$ die Matrix $\tilde{A} := (\tilde{L}\tilde{U})^{-1}LU$ nicht sehr stark von der Einheitsmatrix abweicht und eine wesentlich günstigere Kondition als A besitzt. Dies ist die Idee der *Vorkonditionierung* mittels $ILU-$Verfahren. □

2.3.1 Bandmatrizen

Ein wichtiger Spezialfall schwachbesetzter Matrizen sind *Bandmatrizen*. Hierzu gehören etwa die Matrizen aus den Beispielen 1.3 und 1.4 der Vorlesung.

Definition 2.15. *Eine Matrix $A = (a_{ij}) \in \mathbb{K}^{n\times n}$ besitzt eine* (p, q)- Bandstruktur, *falls $a_{ij} = 0$ für alle Indizes mit $i > j + p$ und $j > i + q$ gilt. Dann heißt die Zahl $p + q + 1$ auch* Bandbreite.

Man kann bei der LU-Zerlegung von Bandmatrizen Operationen mit Elementen außerhalb des Bandes einsparen. Wir zeigen, daß die Dreiecksfaktoren einer LU-Zerlegung einer Bandmatrix wieder Bandstruktur haben.

Satz 2.16. *Die Matrix $A \in \mathbb{K}^{n\times n}$ habe eine LU-Zerlegung mit unterer Dreiecksmatrix L mit Eins-Diagonalenträgen und oberer Dreiecksmatrix U. Ist A Bandmatrix mit unterer bzw. oberer Bandbreite p bzw. q, so hat L die untere Bandbreite p und U die obere Bandbreite q.*

Beweis: Wir verwenden vollständige Induktion nach n. Der Induktionsanfang für $n = \min(p, q)+1$ ist trivial. Sei nun die Aussage für eine Bandmatrix in $\mathbf{K}^{(n-1)\times(n-1)}$ mit unterer bzw. oberer Bandbreite p bzw. q richtig.

Für eine (p, q)-Bandmatrix $A \in \mathbb{K}^{n\times n}$ mit LU-Zerlegung schreiben wir

$$A = \begin{pmatrix} \alpha & w^T \\ v & B \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha}v & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & B - \frac{1}{\alpha}vw^T \end{pmatrix} \begin{pmatrix} \alpha & w^T \\ 0 & I_{n-1} \end{pmatrix}.$$

Dann ist $B - \frac{1}{\alpha}vw^T$ eine (p, q)-Bandmatrix vom Format $(n-1)\times(n-1)$, denn nur die ersten p Komponenten von v und die ersten q Komponenten von w können nicht verschwinden. Weiterhin

hat diese Matrix eine LU-Zerlegung $B - \frac{1}{\alpha} v w^T = L_1 U_1$, so daß

$$L = \begin{pmatrix} 1 & 0 \\ * & L_1 \end{pmatrix}, \qquad U = \begin{pmatrix} * & * \\ 0 & U_1 \end{pmatrix}.$$

Nach Induktionsannahme ist L_1 untere Dreiecksmatrix mit unterer Bandbreite p und U_1 obere Dreiecksmatrix mit oberer Bandbreite q. Dann gilt aber

$$A = \begin{pmatrix} 1 & 0 \\ \frac{1}{\alpha} v & L_1 \end{pmatrix} \begin{pmatrix} \alpha & w^T \\ 0 & U_1 \end{pmatrix}.$$

Beide Faktoren haben die geforderte Bandstruktur. Dies beweist die Induktionsbehauptung. □

Eine Formulierung als Pseudo-Code lautet wie folgt.

Berechnung der LU-Zerlegung einer Bandmatrix:

Initialisierung: (p,q)-Bandmatrix $A = (a_{ij}) \in \mathbb{K}^{n\times n}$ mit existierender LU-Zerlegung;

for $k = 1, \ldots, n-1$ **do**

  for $i = k+1, \ldots, \min(k+p, n)$ **do**

    $a_{ik} := a_{ik}/a_{kk}$;

    for $j = k+1, \ldots, min(k+q, n)$ **do**

      $a_{ij} := a_{ij} - a_{ik}a_{kj}$

    end

  end

end

Ergebnis: LU-Zerlegung von A, wobei a_{ij} mit l_{ij} für $i > j$ und mit u_{ij} für $i \le j$ überschrieben wird.

Vorwärts- und Rückwärtselimination vereinfachen sich für Bandmatrizen sinngemäß. Der folgende Satz formuliert, wie die Bandstruktur bei der Gauß-Elimination mit Spaltenpivotsuche ausgenutzt werden kann.

Satz 2.17. *Sei $A \in \mathbb{K}^{n\times n}$ eine reguläre (p,q)-Bandmatrix. Bei der Gauß-Elimination mit Spaltenpivotsuche werden Vertauschungsmatrizen $P_1, \ldots, P_{n-1}$ und Gauß-Matrizen $M_1, \ldots, M_{n-1}$ mit $M_k = I - l^{(k)} e_k^T$ für $k = 1, \ldots, n-1$ berechnet, so daß*

$$M_{n-1} P_{n-1} \ldots M_1 P_1 A = U \tag{2.17}$$

obere Dreiecksmatrix ist. Dann hat U eine obere Bandbreite $p+q$, ferner ist $(l^{(k)})_i = 0$ für $i \le k$ und $i > k+p$, $k = 1, \ldots, n-1$. Somit hat der Vektor $l^{(k)}$ maximal p Nichtnull-Komponenten und die Matrix L spaltenweise maximal $p+1$ Nichtnull-Komponenten.

Beweis: Sei $PA = LU$ die durch Gauß-Elimination mit Spaltenpivotisierung ermittelte Zerlegung. Dabei war $P = P_{n-1} \ldots P_1$. Mit einer geeigneten Permutation $\{s_1, \ldots, s_n\}$ von $\{1, \ldots, n\}$ sei

$$P = (\, e_{s_1} \;\; \ldots \;\; e_{s_n})^T.$$

Wir nehmen an, daß $s_i > i + p$ für ein $i \in \{s_1, \ldots, s_n\}$ gelte. Dann wäre aber wegen

$$(PA)_{ij} = a_{s_i, j} = 0, \qquad j = 1, \ldots, i \le s_i - p - 1$$

die $i \times i$-Hauptuntermatrix von PA singulär, da sie eine Nullzeile enthielte. Dann folgt aber ein Widerspruch, da auch der entsprechende $i \times i$-Block von U und damit auch A singulär wäre. Da nun A die obere Bandbreite q hat, ist $a_{s_i,j} = 0$ für $j > s_i + q$ und somit $(PA)_{ij} = 0$ für $j > i + p + q$. Damit hat PA die obere Bandbreite $p + q$, ferner existiert nach Konstruktion die LU-Zerlegung. Nach dem vorhergehenden Satz hat U die obere Bandbreite $p + q$.

Wir nehmen an, daß $M_{k-1}P_{k-1} \dots M_1P_1A$ schon berechnet sind. Dann stehen im unteren $(n - k + 1) \times (n - k + 1)$-Block in permutierter Reihenfolge die $k, ..., n$-Komponenten von Zeilen der Matrix A. In der ersten Spalte des Blocks können Nichtnull-Elemente nur in den Positionen $(k+1, k), \dots, (k+p, k)$ auftreten, denn alle folgenden Zeilen verschwinden in den ersten $(k-1)$ Komponenten. Sie können daher nicht zur Pivotisierung herangezogen worden sein. Dies beweist die letzte Aussage des Satzes. □

Es ist sehr wichtig, eine gegebene schwachbesetzte Matrix unter Angabe der Position und des Eintrags der Nichtnullelemente (NNE) so effizient wie möglich zu speichern. Bei Bandmatrizen muß man etwa nur die besetzten "Bänder" ablegen. Es sei hier auf auf Berechnungen anhand der Poisson-Matrix aus Beispiel 1.4 am Ende des Kapitels verwiesen, die deutlich den Vorteil der "sparse"-Speicherung von Matrizen und -Abarbeitung von Faktorisierungsverfahren zeigt.

2.3.2 Tridiagonal-Matrizen

Einen Extremfall bildet die LU-Zerlegung bei Tridiagonalmatrizen mit $p = q = 1$, die bei Anwendungen durchaus von Bedeutung ist (z.B. Diskretisierung von Randwertproblemen gewöhnlicher Differentialgleichungen oder Berechnung von Interpolations-Splines). Wir vereinbaren folgende Schreibweise mit $b_1 = c_n = 0$:

$$A = \text{tridiag}(b_i, a_i, c_i) := \begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & b_n & a_n \end{pmatrix}. \tag{2.18}$$

Der Ansatz

$$L = \begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} u_1 & c_1 & & & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & c_{n-1} \\ & & & & u_n \end{pmatrix}.$$

für die LU-Zerlegung führt nach dem oben genannten Algorithmus auf die Rekursion:

Finde $u_1, l_2, u_2, l_3, ..., l_n, u_n$ aus

$$u_1 := a_1$$

$$l_i := \frac{b_i}{u_{i-1}}, \quad u_i := a_i - l_i c_{i-1}, \quad i = 2, ..., n.$$

Hier kommt man mit $\mathcal{O}(n)$ wesentlichen Rechenoperationen aus. Das ist bei n Unbekannten ein optimales Ergebnis, ist aber an die sehr spezielle Struktur der Matrix gebunden.

Für die Realisierbarkeit des Verfahrens ist offenbar wesentlich, daß $u_i \neq 0, i = 1, ..., n-1$ und zusätzlich $u_n \neq 0$ (wegen der Nichtsingularität von U) gilt. Eine hinreichende Bedingung gibt

Lemma 2.18. *Für $A = tridiag(b_i, a_i, c_i) \in \mathbb{K}^{n\times n}$ mit $b_1 = c_n = 0$ gelte, daß*

$$|c_j| < |a_j|, \qquad |b_j| + |c_j| \le |a_j|, \quad j = 1, ..., n. \tag{2.19}$$

Dann ist die Zerlegung $A = LU$ ausführbar. Speziell gilt $u_1, ..., u_n \neq 0$ und $\det A = \prod_{i=1}^{n} u_i$.

Beweis: Wir beweisen die Behauptung induktiv. Mit $|c_1| < |a_1| = |u_1|$ ist $u_1 \neq 0$ und $|\frac{c_1}{u_1}| < 1$. Wir zeigen nun

$$u_{i-1} \neq 0, \quad |\frac{c_{i-1}}{u_{i-1}}| < 1 \quad \Longrightarrow \quad u_i \neq 0, \quad |\frac{c_i}{u_i}| < 1, \quad i = 2, ..., n.$$

Wegen

$$|u_i| = |a_i - l_i c_{i-1}| = |a_i - \frac{b_i}{u_{i-1}} c_{i-1}| \ge |a_i| - |b_i| \left|\frac{c_{i-1}}{u_{i-1}}\right|$$

gilt unter Beachtung der Voraussetzung die Fallunterscheidung

$$\begin{array}{llllll} (i) & b_i \neq 0 & \Longrightarrow & |u_i| > |a_i| - |b_i| \ge |c_i| & \Longrightarrow & u_i \neq 0, \quad \left|\frac{c_i}{u_i}\right| < 1 \\ (ii) & b_i = 0 & \Longrightarrow & |u_i| \ge |a_i| > |c_i| & \Longrightarrow & u_i \neq 0, \quad \left|\frac{c_i}{u_i}\right| < 1. \end{array}$$

Die $LU-$Zerlegung war bereits gezeigt worden. Speziell folgt

$$\det(A) = \det(L)\det(U) = 1 \cdot \prod_{i=1}^{n} u_i. \qquad \square$$

Den entsprechenden Algorithmus für die vollständige Lösung des Gleichungssystems $Ax = y$ bzw. nach erfolgter $LU-$Zerlegung mit $A = LU$ nennt man den *Thomas-Algorithmus*. Seine Herleitung und Implementierung ist Gegenstand der Übungen.

2.3.3 UMFPACK in MATLAB

Leider ist die Besetzungsstruktur einer schwachbesetzten Matrix A oft erheblich von einer Bandstruktur mit gegenüber der Dimension n nicht sehr großen Bandbreiten (p, q) entfernt. Durch Vertauschung von Zeilen und Spalten kann man die Bandbreite verringern. Bei großer Dimension der Matrix setzt man verfeinerte graphentheoretisch orientierte Algorithmen ein. Einer der ersten Algorithmen stammt von CUTHILL/MCKEE (1969).

Es gibt seit einiger Zeit eine sehr starke Entwicklung von effizienten Programmen zur LU-Zerlegung schwachbesetzter Matrizen, die sich auch graphentheoretischer Methoden bedienen. Von DEMMEL/GILBERT/LI (1999) stammt der SuperLU-Algorithmus. Die Idee der zugrunde liegenden Technik der *Superknoten* (super nodes) ist die Zusammenfassung von Spalten mit gleichem Besetzungsmuster (d.h. gleicher Position von Nullelementen) zu einer Gruppe. Man kann sie dann bei Abspeicherung und Berechnung als dichtbesetzte Matrix behandeln. Der Algorithmus selbst besteht grob gesagt aus einer *symbolischen* Bestimmung der Besetzungsstruktur der LU-Zerlegung und einer *numerischen* Faktorisierung und Lösung. Hinsichtlich einer genaueren Beschreibung verweisen wir auf die Web-Seite `http://crd.lbl.gov/~xiaoye/SuperLU`.

Unter MATLAB ist seit einiger Zeit das Paket UMFPACK (*Unsymmetric MultiFrontal Package*) von T. Davies verfügbar, vgl. `http://www.cise.ufl.edu/research/sparse/umfpack/`. Es wird zur Matrix PAQ oder sogar zu $PRAQ$ eine LU-Zerlegung erzeugt. Dabeit realisiert die Permutationsmatrix Q eine geeignete Spaltenvertauschung mit einer guten oberen Schranke für

das "fill-in". Diese obere Schranke wird während der Rechnung gesichert und ggf. verfeinert. Die Permutationsmatrix P bewirkt eine geeignete Zeilenvertauschung bei der numerischen Pivotisierung unter Beibehaltung der numerischen Stabilität. Die Diagonalmatrix R dient ggf. einer geeigneten Zeilenskalierung von A.

UMFPACK wird auch im Rahmen der Übungen benutzt. Nach bisherigen Erfahrungen ist dieser bei schwachbesetzten Gleichungssystemen mit bis zu $10^5 - 10^6$ Unbekannten erfolgreich einsetzbar. (Der Algorithmus stellt somit in diesem Bereich eine echte Alternative zu iterativen Lösungsverfahren dar, die wir später in der Vorlesung behandeln.)

Tabelle 1: Poisson-Matrix (voll besetzt):

n	$N = n^2$	LU	LAPACK	LU_{bw}	LU MATLAB	LU UMFPACK
10	100	0.02	<0.01	0.01		
20	400	1.04	0.02	0.04		
30	900	12.70	0.17	0.08		
40	1.600	63.00	0.80	0.19		
50	2.500	245,00	3.00	0.35	0.09	0.04
60	3.600	744.00	8.60	0.63		
70	4.900	1946.00	20.60	1.02		
80	6.400	—	—	—		
100	10.000				1.09	0.19
150	22.500				5.15	0.49
200	40.000				17.00	0.92
250	62.500				39.00	1.65
300	90.000				83.00	2.54
400	160.000				—	5.40
500	250.000					9.30
600	360.000					15.30
800	640.000					34.70

Beispiel 2.19. Wir betrachten Beispiel 1.4 mit der sogenannten Poisson-Matrix. Tabelle 1 zeigt Rechenzeiten (in Sekunden). Für eine "naive" Implementierung des LU-Verfahrens (LU) ersieht man die bereits gezeigte kubische Komplexität bezüglich $N = n^2$. Für das mit den sehr schnellen BLAS-Routinen geschriebene Paket LAPACK ergeben sich deutlich bessere Werte, jedoch stösst das Verfahren ohne Beachtung der Bandstruktur schnell an seine Grenzen. Bei einer "naiven" LU-Implementierung unter Benutzung der Bandstruktur (LU_{bw}) sieht man einen deutlich geringeren Aufwand. Die in MATLAB implementierte LU-Zerlegung (LU MATLAB) erlaubt die Berechnung weit grösserer Matrizen, wobei hier aber die Bandstruktur nicht ausgenutzt wird. Schliesslich wird bei der in MATLAB verfügbaren Routine UMFPACK (LU UMFPACK) ganz offenbar die Bandstruktur erkannt und ausgenutzt. Erst bei $N = 10^6$ trat ein Speicherfehler auf.

Bemerkung 2.20. Das unter MATLAB verfügbare Paket UMFPACK ist als leistungsfähiges Lösungsverfahren für die bei der Finite-Elemente-Lösung vieler physikalischer Modelle mit dem Programmsystem COMSOL entstehenden großen linearen Gleichungssysteme verfügbar. Dies gilt wenigstens für räumliche zweidimensionale Modelle, jedoch leider (noch) nicht für den wichtigeren dreidimensionalen Fall. Das Programm COMSOL wird in Spezialvorlesungen zur numerischen Lösung partieller Differentialgleichungen intensiv benutzt. Es erlaubt dem Nutzer einen sehr zügigen Einstieg in diese Problematik. □

Kapitel 3

Lineare Ausgleichsprobleme

Im vorliegenden Kapitel befassen wir uns mit *linearen Ausgleichsproblemen*. Gegeben sind eine Matrix $A \in \mathbb{K}^{m\times n}$ und ein Vektor $y \in \mathbb{K}^m$. Gesucht ist dann eine Lösung des Problems

$$\text{Minimiere} \quad \|Ax - y\|_2, \qquad x \in \mathbb{K}^n. \tag{3.1}$$

Eines der wichtigsten Lösungsverfahren von (3.1) ist das *QR-Verfahren*, das zugleich ein stabiles Faktorisierungsverfahren für Matrizen ist.

3.1 Problemstellung. Grundlagen

Wir greifen nochmals die schon in Beispiel 1.2 beschriebene Situation auf.

Beispiel 3.1. Ein bestimmtes Modell werde beschrieben durch den Ansatz

$$u(t;x) \equiv u(t;x_1,\ldots,x_n) := \sum_{j=1}^{n} x_j \phi_j(t) \tag{3.2}$$

mit Funktionen $\phi_j : [a,b] \to \mathbb{K},\ j = 1,\ldots,n$. Das Modell (3.2) wird also gerade durch die (geeignet gewählten und nicht zwingend polynomialen) Funktionen $\phi_j,\ j = 1,\ldots,n$ charakterisiert. Durch m Messungen seien Paare $(t_i, y_i),\ i = 1,\ldots m$ ermittelt worden. In der Praxis ist die Zahl m der Messungen (viel) größer als die Zahl n der zu bestimmenden Parameter des Ansatzes.

Wir nehmen für den Moment an, daß das Modell (3.2) exakt ist und keine Meßfehler auftreten. Dann wäre der gesuchte Parametervektor $x = (x_j) \in \mathbb{K}^n$ Lösung des linearen Systems

$$\sum_{j=1}^{n} \phi_j(t_i) x_j = y_i, \qquad i = 1,\ldots,m. \tag{3.3}$$

Die beiden genannten Annahmen sind natürlich unrealistisch. Man versucht daher, eine möglichst gute Lösung des überbestimmten Systems (3.3) zu finden. Bei der *Methode der kleinsten Quadrate* sucht man eine Lösung des Optimierungs- bzw. Approximations-Problems

$$\text{Minimiere} \quad F(x) := \frac{1}{2}\sum_{i=1}^{m} \left[y_i - \sum_{j=1}^{n} \phi_j(t_i) x_j \right]^2, \qquad x \in \mathbb{K}^n, \tag{3.4}$$

d.h. F ist quadratisch in x. Wir formulieren das Problem (3.4) etwas um und definieren

$$A = (a_{ij}) \in \mathbb{K}^{m\times n}, \qquad a_{ij} := \phi_j(t_i); \qquad y = (y_i) \in \mathbb{K}^m.$$

So erhalten wir bei Verwendung der euklidischen Norm $\|\cdot\|_2$ das sogenannte *lineare Ausgleichsproblem*

$$\text{Minimiere} \quad F(x) := \frac{1}{2}\|Ax - y\|_2^2, \qquad x \in \mathbb{K}^n. \qquad \square \tag{3.5}$$

Die Aufgabe (3.5) ist natürlich äquivalent zum Ausgangsproblem (3.1). Wir wollen uns nun mit der Lösbarkeit des linearen Ausgleichsproblems (3.4) befassen. Dazu sei daran erinnert, dass A^* die zu A adjungierte Matrix ist. Ferner ist $\|x\|_2^2 = x^* \cdot x$ für Vektoren $x \in \mathbb{K}^n$.

Satz 3.2. *(Lösbarkeit des linearen Ausgleichsproblems)*
Bei gegebenen Daten $A = (a_{ij}) \in \mathbb{K}^{m\times n}$ mit $m \geq n$ und $y = (y_i) \in \mathbb{K}^m$ gelten für das lineare Ausgleichsproblem (3.5) folgende Aussagen:

(i) *Es existiert eine Lösung $\tilde{x} \in \mathbb{K}^n$ des Problems (3.5), d.h. $F(\tilde{x}) \leq F(x)$ für alle $x \in \mathbb{K}^n$.*

(ii) *Ein Vektor $\tilde{x} \in \mathbb{K}^n$ ist Lösung von (3.5) genau dann, wenn er Lösung des linearen Gleichungssystems $A^*Ax = A^*y$ (Normalengleichungen) ist.*

(iii) *Das Problem (3.5) ist genau dann eindeutig lösbar im Fall $Rang(A) = n$, d.h. wenn die Spalten von A linear unabhängig sind.*

(iv) *In der Lösungsmenge von (3.5) gibt es genau ein Element mit minimaler euklidischer Norm.*

Beweis: (i) Wegen der Äquivalenz der Minimierungsprobleme (3.5) und (3.1) betrachten wir die Minimierung von $M(x) := \|Ax - y\|_2$. Sei $(x_k)_k$ eine sogenannte Minimalfolge, d.h.

$$\|Ax_k - y\|_2 \to \sigma := \inf_{x\in\mathbb{K}^n} \|Ax - y\|_2, \qquad k \to \infty.$$

Für hinreichend große Zahlen k wird $\|Ax_k - y\|_2 \leq 2\sigma$ und somit

$$\|Ax_k\|_2 = \|(Ax_k - y) + y\|_2 \leq \|Ax_k - y\|_2 + \|y\|_2 \leq 2\sigma + \|y\|_2.$$

Daher ist die Folge $(Ax_k)_k \subset \text{Bild}(A) \subset \mathbb{K}^m$ beschränkt. Nach dem Satz von Bolzano/Weierstraß kann man daher eine gegen ein Element $\tilde{y} \in \mathbb{K}^m$ konvergente Teilfolge auswählen. Da der lineare Unterraum $\text{Bild}(A) \subset \mathbb{K}^m$ abgeschlossen ist, gibt es ein Element $\tilde{x} \in \mathbb{K}^n$ mit $\tilde{y} = A\tilde{x}$. Dann ist

$$\|A\tilde{x} - y\|_2 = \inf_{x\in\mathbb{K}^n} \|Ax - y\|_2$$

und somit $\tilde{x}$ eine Lösung von (3.1) bzw. (3.5).

(ii) Jede Lösung $\tilde{x}$ von (3.5) ist auch Lösung des Systems der Normalengleichungen, denn

$$0 = \nabla F(\tilde{x}) = A^*(A\tilde{x} - y).$$

Sei umgekehrt $\tilde{x}$ Lösung von $\nabla F(\tilde{x}) = 0$. Die Taylor-Entwicklung von F ergibt

$$\begin{aligned} F(x) &= F(\tilde{x}) + \underbrace{\nabla F(\tilde{x})^*(x - \tilde{x})}_{=0} + \frac{1}{2}(x - \tilde{x})^* A^* A(x - \tilde{x}) \\ &= F(\tilde{x}) + \frac{1}{2}\|A(x - \tilde{x})\|_2^2 \;\geq\; F(\tilde{x}) \end{aligned}$$

für beliebiges $x \in \mathbb{K}^n$, d.h. $\tilde{x}$ löst (3.5).

(iii) Unter der Bedingung $\text{Rang}(A) = n$ ist $A^*A \in \mathbb{K}^{n\times n}$ symmetrisch und positiv definit. Das System der Normalengleichungen und damit (3.5) ist somit eindeutig lösbar. Ist andererseits $\text{Rang}(A) < n$, so ist $\text{Kern}(A) \neq \{0\}$. Somit sind die Normalengleichungen und auch (3.5) nicht eindeutig lösbar.

(iv) Die Lösungsmenge $\mathcal{M}$ von (3.5) bzw. der Normalengleichungen ist ein affin linearer Unterraum des $\mathbb{K}^n$. Ein derartiger Raum enthält genau ein Element mit minimaler euklidischer Norm, die orthogonale Projektion des Nullpunktes auf $\mathcal{M}$. □

Bemerkung 3.3. Das Modell in Beispiel 3.1 besitzt somit eine Lösung. Zur Eindeutigkeitsuntersuchung untersucht man die Rangbedingung nach Satz 3.2 (iii). Für die Matrix $A = (a_{ij})$ mit $a_{ij} := \phi_j(t_i)$ ist zu prüfen, ob das homogene Problem

$$\sum_{j=1}^{n} \phi_j(t_i)x_j = 0, \qquad i = 1, \ldots, m$$

nur die triviale Lösung $x = 0$ hat. Dies ist der Fall, wie man leicht nachprüft. □

3.2 QR-Zerlegung für lineare Ausgleichsprobleme

Wir wollen jetzt das lineare Ausgleichsproblem (3.1) numerisch lösen. Vereinfachend setzen wir zunächst den Eindeutigkeitsfall $\text{Rang}(A) = n$ mit $m \geq n$ voraus. Der einfachste Weg zur Anwendung unserer Kenntnisse aus Kapitel 2 besteht in der Multiplikation mit A^*. Dann erhalten wir

$$A^*Ax = A^*y. \tag{3.6}$$

Die Matrix A^*A ist symmetrisch und im Fall vollen Ranges, d.h. $\text{Rang}(A) = n$, auch positiv definit. Man könnte dann das Problem (3.6) mit dem Cholesky-Verfahren (vgl. Abschnitt 2.2) lösen. Diese Methode ist die schnellste Lösungsvariante. Sie ist jedoch oft zu ungenau, wie wir in Kapitel 4 etwas genauer diskutieren werden.

Es ist also ratsam, nach einer numerisch stabileren Variante zu suchen. Dazu betrachten wir statt der $LU-$Zerlegung der Matrix A die sogenannte $QR-$Zerlegung.

Definition 3.4. *Sei $A \in \mathbb{K}^{m\times n}$. Unter einer* $QR-$Zerlegung von *A versteht man eine Faktorisierung*

$$A = QR,$$

wobei $Q \in \mathbb{K}^{m\times m}$ unitär ist und $R \in \mathbb{K}^{m\times n}$ die Form

$$R = \left(\begin{array}{cccccc} r_{11} & r_{12} & \cdots & \cdots & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & \cdots & \cdots & r_{2n} \\ \vdots & & \ddots & & & \vdots \\ 0 & \cdots & 0 & r_{kk} & \cdots & r_{kn} \\ \hline 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{array}\right)$$

mit $r_{11}, \cdots, r_{kk} \neq 0$ und $k \leq \min(m, n)$ hat. Dabei heißt Q unitär *(oder im Fall $\mathbb{K} = \mathbb{R}$* orthogonal*), falls $QQ^* = Q^*Q = I$.*

Die im Vergleich zur LU-Zerlegung geringere Anfälligkeit der QR-Zerlegung gegenüber Rundungsfehlern basiert wesentlich auf folgenden Eigenschaften, die wir wiederholen.

Lemma 3.5. *Für eine unitäre Matrix $Q \in \mathbb{K}^{m\times m}$ sind folgende Bedingungen äquivalent:*

(i) *Die Spalten von $Q = (\ q_1\ \cdots\ q_n\ )$ und analog die Zeilen von Q bilden ein Orthonormalsystem bezüglich des euklidischen Skalarproduktes auf $\mathbb{K}^m$.*

(ii) *Die euklidische Norm bzw. das euklidische Skalarprodukt sind invariant unter der Transformation $x \mapsto Qx$, d.h. es gilt $\|Qx\|_2 = \|x\|_2$ für alle $x \in \mathbb{K}^m$ bzw. $(Qx)^*(Qy) = x^*y$ für alle $x, y \in \mathbb{K}^m$.*

Im nächsten Abschnitt beweisen wir, daß zu jeder Matrix $A \in \mathbb{K}^{m\times n}$ mit $\mathrm{Rang}(A) = n$ eine QR-Zerlegung existiert. Wir wollen jetzt zeigen, wie man bei bekannter QR-Zerlegung von A zu einer Lösung des linearen Ausgleichsproblems (3.1) gelangt:
Sei eine QR-Zerlegung von A bekannt, d.h. eine unitäre Matrix $Q \in \mathbb{K}^{m\times m}$ und eine obere Dreiecksmatrix $R \in \mathbb{K}^{m\times n}$ mit

$$A = QR = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}. \tag{3.7}$$

Hierbei ist die obere Dreiecksmatrix $\hat{R} \in \mathbb{K}^{n\times n}$ regulär und sie besitzt nichtverschwindende Diagonalelemente, falls $\mathrm{Rang}(A) = n$. Unter Benutzung von Lemma 3.5 (ii) gilt für beliebige Vektoren $x \in \mathbb{K}^n$

$$\begin{aligned} \|Ax - y\|_2^2 &= \|QRx - y\|_2^2 = \|Rx - Q^*y\|_2^2 \\ &= \left\| \begin{pmatrix} \hat{R}x \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 \\ &= \|\hat{R}x - c\|_2^2 + \|d\|_2^2. \end{aligned}$$

Dabei ist

$$\begin{pmatrix} c \\ d \end{pmatrix} := Q^*y$$

eine Zerlegung von Q^*y in die Vektoren $c \in \mathbb{K}^n$ und $d \in \mathbb{K}^{m-n}$. Dann findet man die eindeutig bestimmte Lösung $x = \hat{R}^{-1}c$ des linearen Ausgleichsproblems durch Rückwärtselimination des gestaffelten linearen Gleichungssystems $\hat{R}x = c$. Der Wert des Minimums ist $\|d\|_2^2$.

Bemerkung 3.6. Die numerisch stabilste, aber bei weitem aufwendigste Methode für lineare Ausgleichsprobleme ist die Methode der *Singulärwertzerlegung*. Leider können wir im Rahmen dieser Einführungsvorlesung hierauf nicht eingehen. □

3.3 Householder-Matrizen

Zum Beweis des Existenzsatzes für die QR-Zerlegung verwenden wir *Householder-Matrizen*.

Definition 3.7. *Eine Matrix H der Form*

$$H = I - 2hh^* \tag{3.8}$$

mit Einheitsvektoren h, d.h. $h^*h = 1$, heißt Householder-Matrix.

Bemerkung 3.8. Man verwechsle für $u, v \in \mathbb{K}^n$ den Ausdruck $uv^* \in \mathbb{K}^{n\times n}$ mit $(uv^*)_{ij} := u_i\overline{v}_j$

nicht mit dem Skalarprodukt $v^*u = \sum_{i=1}^n u_i \overline{v}_i$. □

Lemma 3.9. *Householder-Matrizen sind unitär. Außerdem gilt $H = H^*$.*

Beweis:

Es gilt

$$H^* = I^* - 2(hh^*)^* = I - 2hh^* = H,$$

damit

$$HH^* = H^*H = (I - 2hh^*)(I - 2hh^*) = I - 4hh^* + 4hh^*hh^* = I. \qquad \square$$

Bemerkung 3.10. Im Fall $\mathbb{K} = \mathbb{R}$ kann die Transformation mittels H geometrisch anschaulich charakterisiert werden. eine Spiegelung des Im Fall $\mathbb{K} = \mathbb{R}$ beschreibt die Transformation mittels H geometrisch eine Spiegelung des $\mathbb{R}^n$ an der Hyperebene durch den Koordinatenursprung senkrecht zu h. Dazu zerlegen wir den Vektor x in seine Komponenten in Richtung von h und den dazu orthogonalen Anteil, d.h.

$$x = (hh^*)x + y, \qquad y \perp h.$$

Daraus folgt

$$Hx = x - 2hh^*x = -hh^*x + y,$$

d.h. Hx hat die gleiche Komponente $y \perp h$ und die entgegengesetzte Komponente $-hh^*x$ in $h-$Richtung. □

Householder-Matrizen spielen für die QR-Zerlegung eine ähnliche Rolle wie Gauß-Matrizen bei der LU-Zerlegung. Ähnlich wie bei der Gauß-Elimination bringt man die Matrix spaltenweise durch Multiplikation mit Householder-Matrizen auf obere Dreiecksgestalt. Die Grundlage hierfür gibt folgendes Resultat.

Lemma 3.11. *Für $x \in \mathbb{K}^n \setminus \{0\}$ sei $u := x + sign(x_1)\|x\|_2 e_1$ und $H = I - \frac{2}{u*u} uu*$. Dann gilt*

$$Hx = -\text{sign}(x_1)\|x\|_2 e_1. \tag{3.9}$$

Beweis: Gesucht ist also ein Vektor $u \in \mathbb{K}^n \setminus \{0\}$, so daß $Hx = \sigma e_1$ mit $\sigma \in \mathbb{K}$ bzw.

$$x - u\frac{2u^*x}{u^*u} = \sigma e_1.$$

Hinreichend hierfür ist

$$\frac{2u^*x}{u^*u} = 1, \qquad u = x + ce_1$$

mit einer geeigneten Zahl $c \in \mathbb{K}$. Beide Bedingungen führen auf

$$0 = u^*(2x - u) = (x + ce_1)^*(x - ce_1) = x^*x - c^2$$

mit den Lösungen $c = \pm\|x\|_2$. Die Wahl $c = \text{sign}(x_1)\|x\|_2$ mit $\text{sign}(a) := 1$ bei $a \geq 0$ und $\text{sign}(a) := -1$ bei $a < 0$ sichert $u \neq 0$, falls x ein Vielfaches von e_1 ist. Dies verhindert auch das Aufschaukeln von Rundungsfehlern bei Subtraktion von Zahlen nahezu gleicher Größe. □

3.4 QR-Zerlegung mit Householder-Verfahren

Wir wollen nun das Hauptresultat über die Existenz einer QR-Zerlegung beweisen.

Satz 3.12. *Für eine Matrix $A \in \mathbb{K}^{m\times n}$ mit Rang$(A) = n$ existiert eine QR-Zerlegung.*

Beweis: Unter Beachtung von Lemma 3.11 ist die Idee, die Matrix A spaltenweise durch Multiplikation mit Householder-Matrizen auf obere Dreiecksgestalt zu bringen. Seien bereits nach k Schritten unitäre Matrizen $H^{(1)}, \ldots, H^{(k)}$ so bestimmt, daß für

$$A^{(k)} := H^{(k)} \cdots H^{(1)} A$$

die Aussage $a_{ij}^{(k)} = 0$ für $j \le k$ und $i > j$ gilt. Somit wäre $A^{(k)}$ eine Blockmatrix der Form

$$A^{(k)} = \begin{pmatrix} R^{(k)} & B^{(k)} \\ 0 & C^{(k)} \end{pmatrix} \tag{3.10}$$

mit oberer Dreiecksmatrix $R^{(k)} \in \mathbb{K}^{k\times k}$ sowie $B^{(k)} \in \mathbb{K}^{(m-k)\times k}$ und $C^{(k)} \in \mathbb{K}^{(m-k)\times(m-k)}$.

Im Fall $k = n$ hätte man mit

$$Q = H^{(1)} \cdots H^{(n)}, \qquad R = A^{(n)}$$

eine QR-Zerlegung ermittelt, denn es gilt $(H^{(j)})^{-1} = H^{(j)}$ für $j = 1, \ldots, n$.

Anderenfalls geht man wie folgt vor: Sei

$$\tilde{x}^{(k+1)} := (c_{11}^{(k)}, \ldots, c_{m-k,1}^{(k)})^T \in \mathbb{K}^{m-k}.$$

Für $\tilde{x}^{(k+1)} = 0$ wären die ersten $k+1$ Spalten von $A^{(k)}$ linear abhängig und somit Rang$(A^{(k)}) < n$. Dies ist aber ein Widerspruch zur Voraussetzung Rang $(A) = n$ wegen der Unitarität der $H^{(j)}$. Wir setzen nun

$$\begin{aligned} \tilde{u}^{(k+1)} &:= \tilde{x}^{(k+1)} + \operatorname{sign}(\tilde{x}_1^{(k+1)}) \|\tilde{x}^{(k+1)}\|_2 e_1, \\ \tilde{H}^{(k+1)} &:= I_{m-k} - \frac{2}{(\tilde{u}^{(k+1)})^* \tilde{u}^{(k+1)})} \tilde{u}^{(k+1)} (\tilde{u}^{(k+1)})^*. \end{aligned}$$

Nach Lemma 3.11 ist die erste Spalte von $\tilde{H}^{(k+1)} C^{(k)}$ ein Vielfaches des Einheitsvektors e_1. Schließlich erweitern wir $\tilde{H}^{(k+1)}$ mittels

$$H^{(k+1)} := I_m - \frac{2}{(u^{k+1)})^* u^{(k+1)}} u^{(k+1)} (u^{k+1)})^*, \quad u^{(k+1)} := (\underbrace{0, \ldots, 0}_{k-mal}, (\tilde{u}^{(k+1)})^T)^T$$

zu einer Blockmatrix

$$H^{(k+1)} = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}^{(k+1)} \end{pmatrix} \in \mathbb{K}^{m\times n}.$$

Dann hat

$$A^{(k+1)} = H^{(k)} A^{(k)} = \begin{pmatrix} R^{(k)} & B^{(k)} \\ 0 & \tilde{H}^{(k+1)} C^{(k)} \end{pmatrix} \in \mathbb{K}^{m\times n}$$

die Form (3.10). Der Beweis ist damit geführt. □

Aus dem konstruktiven Beweis gewinnt man folgende Berechnungsvorschrift.

QR-Zerlegung mittels Householder-Verfahren (Matrixversion)

Initialisierung: $A \in \mathbb{K}^{m\times n}$ mit Rang $(A) = n$;
$A^{(0)} := A$;

for $k = 1, \ldots, n$ **do**

$d := a_{k,k}^{(k-1)} + \text{sign } (a_{k,k}^{(k-1)}) \sqrt{\sum_{i=k}^{m} |a_{i,k}^{(k-1)}|^2}$;

$u^{(k)} := (\underbrace{0, \ldots, 0}_{k-1 \ mal}, d, a_{k+1,k}^{(k-1)}, \ldots, a_{m,k}^{(k-1)})^T$;

$H^{(k)} := I_m - \frac{2}{(u^{(k)})^* u^{(k)}} u^{(k)} (u^{k})^*$;

$A^{(k)} := H^{(k)} A^{(k-1)}$;

end

Ergebnis: $A = QR$ mit $R := A^{(n)}$ und $Q := H^{(1)} \cdots H^{(n)}$.

In dieser Form kann das Verfahren noch nicht implementiert werden. Insbesondere ist die Anwendung der Householder-Matrix auf einen Vektor zu programmieren. Seien $H = I - \frac{2}{u^* u} uu^*$ mit $u := x + \text{sign } (x_1)\|x\|_2 e_1$ für $x \in \mathbb{K}^m$. Dann haben wir

$$H = I - \beta uu^*, \qquad \beta = \frac{2}{u^* u} = \frac{1}{\|x\|_2(\|x\|_2 + 1)}$$

sowie für $v \in \mathbb{K}^n$ schließlich

$$Hv = v - su, \qquad s := \beta u^* v.$$

Bei Berechnung von β kann ein Exponentenunter- bzw. -überlauf auftreten, wenn $\|x\|_2$ sehr klein oder sehr groß ist. Zur besseren Skalierung berechnet man zunächst $y := x/\|x\|_\infty$ und dann

$$u := y + \text{sign } (y_1)\|y\|_2 e_1.$$

Die Matrix H ändert sich nicht, jedoch ist nun $\beta := \frac{1}{\|y\|_2(\|y\|_2 + |y_1|)}$ zu setzen.

QR-Zerlegung nach Householder (Implementations-Version)

Initialisierung: $A \in \mathbb{K}^{m \times n}$ mit Rang $(A) = n$;
$u_{ik} := 0$ für $i = 1, \ldots, m$ und $k = 1, \ldots, n$

for $k = 1, \ldots, n$ **do**

(Berechnung der k-ten Householder-Matrix)

$\|a_k\|_\infty := \max_{i=k,\ldots,m} |a_{ik}|$; $\alpha := 0$;

for $i = k, \ldots, m$ **do**

$u_{ik} := a_{ik}/\|a_k\|_\infty$;
$\alpha := \alpha + |u_{ik}|^2$;

end

$\alpha := \sqrt{\alpha}$; $\beta_k := 1/(\alpha(\alpha + |u_{kk}|))$; $u_{kk} := u_{kk} + \text{sign } (a_{kk})\alpha$;

(Multiplikation der k-ten Householder-Matrix mit A)

$a_{kk} := -\text{sign}(a_{kk})\|a_k\|_\infty \alpha$;

for $i = k + 1, \ldots, m$ **do**

$a_{ik} = 0$;

end

for $j = k+1, \ldots, n$ **do**

$s := \beta_k \sum_{i=k}^{m} \overline{u_{ik}} a_{ij}$;

for $i = k, \ldots, m$ **do**

$a_{ij} := a_{ij} - s u_{ik}$;

end

end

end

Ergebnis: A ist mit der Matrix R überschrieben. Ferner ist $Q := H^{(1)} \cdots H^{(n)}$ mit $H^{(k)} := I - \frac{2}{u_k^* u_k} u_k u_k^*$ und $u_k := (u_{ik})_{i=1,\ldots m}$ für $k = 1, \ldots, n$, so daß $A = QR$ gilt.

Bemerkungen 3.13. (i) Die Berechnung der Matrix Q ist oft nicht sinnvoll, da nur die Wirkung von Q bzw. Q^* auf einen Vektor benötigt wird. Für $m \gg n$ erfordert die Speicherung von Q erheblich mehr Speicherplatz als die Speicherung von (u_{ik}) und die Multiplikation von Q mit einem Vektor ist sehr viel billiger als n Matrix-Vektor Multiplikationen mit $m \times m$-Matrizen.

(ii) Eine Abspeicherung der orthogonalen Matrix Q ist nur sinnvoll, wenn das Ausgleichsproblem $\|Ax - y\|_2 = \min\ !$ mit verschiedenen rechten Seiten y (jedoch gleicher Matrix A) gerechnet wird. Sonst multipliziert man y schrittweise mit den Matrizen $H^{(k)}$ und verwirft letztere dann.

(iii) Man kann die Spaltenvektoren u_k der Householder-Matrizen weitgehend in den frei werdenden Speicherplätzen von A ablegen. Lediglich ein Vektor aus $\mathbb{K}^n$, zum Beispiel die Diagonalelemente von R, müssen gesondert gespeichert werden. Allerdings ist dies im obigen Algorithmus nicht realisiert. □

Schließlich wollen wir den *Rechenaufwand* des $QR-$Verfahrens ermitteln.

Lemma 3.14. *Die QR-Zerlegung einer Matrix $A \in \mathbb{K}^{m \times n}$ mit Rang $(A) = n$ erfordert $n^2(m - \frac{1}{3}n) + \mathcal{O}(mn)$ wesentliche Rechenoperationen.*

Beweis: Dies sei als Übungsaufgabe gestellt. □

Für quadratische Matrizen mit $m = n$ sind dies im Prinzip $\frac{2}{3}n^3$ flops. Diese Zahl ist asymptotisch (für $n \to \infty$) doppelt so hoch wie bei der Gauß-Elimination. Jedoch ist es weniger empfindlich gegenüber Rundungsfehlern als das $LU-$Verfahren. Diese Tatsache rechtfertigt bei Matrizen mit schlechter Kondition den zusätzlichen Aufwand.

3.5 QR-Verfahren im rang-defizienten Fall (Exkurs)

Eine Zeilen- oder Spaltenvertauschung ist (im Gegensatz zur LU-Zerlegung) bei der QR-Zerlegung im Fall Rang $(A) = n$ nicht erforderlich. Dies ist jedoch im *rang-defizienten* Fall, d.h. für $\text{Rang}(A) < \min(m, n)$, im allgemeinen Fall notwendig.

Vor Multiplikation mit der k-ten Householder-Matrix tauscht man die Restspalte mit größter euklidischer Norm an die k-te Position. Dies entspricht einer Rechtsmultiplikation mit einer Permutationsmatrix, die man sich durch einen Integer-Vektor merkt.

Im Fall Rang $(A) = k$ wären bei exakter Arithmetik die untersten $m - k$ Zeilen von A Nullvektoren, durch Rundungsfehler können jedoch (sehr kleine) Einträge entstehen. Ferner ist der Rang von A in der Regel nicht bekannt, so daß man numerisch über die Vernachlässigbarkeit des unteren Blocks von $A^{(k)}$ entscheiden möchte. Dazu bietet sich ein Vergleich mit dem Element a_{11} an, da vor dem ersten Schritt die normmäßig größte Spalte nach vorn getauscht wurde und sich die Norm der Spalten bei Multiplikation mit einer unitären Matrix nicht ändern. Hieraus ergibt sich der folgende modifizierte Algorithmus.

Implementationsversion der QR-Zerlegung mit Spaltenvertauschungen

Initialisierung: $A \in \mathbb{K}^{m\times n}$;
$\epsilon > 0$; *Toleranzparameter für numerische Rangentscheidung*
$u_{ik} := 0$ für $i = 1, \ldots, m$ und $k = 1, \ldots, n$
$p_i = i$ für $i = 1, \ldots, n$;

$k = 0$; exitflag =0;
while (exitflag=0 und $k < \min(m, n)$) **do**

for $j = k + 1, \ldots, n$ **do**

$n_j := \max_{i=k+1,\ldots,m} |a_{ij}|$;

end

Bestimme einen Index $r \in \{k + 1, \ldots, n\}$ mit $n_r = \max_{i=k+1,\ldots,n} n_i$;

if $n_r \leq \epsilon |a_{11}|$

exitflag=1;

else

$k = k + 1$;
Vertausche p_k und p_r;
Vertausche $k-$te und $r-$te Spalte von A;
(Berechnung der k-ten Householder-Matrix)
$\alpha := 0$;
for $i = k, \ldots, m$ **do**;
$u_{ik} := a_{ik}/\|a_k\|_\infty$;
$\alpha := \alpha + |u_{ik}|^2$;
end
$\alpha := \sqrt{\alpha}$; $\beta_k := 1/(\alpha(\alpha + |u_{kk}|))$; $u_{kk} := u_{kk} + \text{sign } (a_{kk})\alpha$;

(Multiplikation der k-ten Householder-Matrix mit A)
$a_{kk} := -\text{sign}(a_{kk})\|a_k\|_\infty \alpha$;
for $i = k + 1, \ldots, m$ **do**
$a_{ik} = 0$;
end
for $j = k + 1, \ldots, n$ **do**

```
            s := β_k Σ_{i=k}^{m} conj(u_ik) a_ij;
            for i = k, ..., m do
                a_ij := a_ij − s u_ik;
            end
        end
    end
end
```

Ergebnis: Der numerische Rang von A ist gleich k. A ist mit der Matrix R überschrieben. Ferner ist $P := (\delta_{i,p_j})_{i,j=1,\ldots,n}$ sowie $Q := H^{(1)} \cdots H^{(k)}$ mit $H^{(j)} := I - \frac{2}{u_j^* u_j} u_j u_j^*$ und $u_j := (u_{ij})_{i=1,\ldots,m}$ für $j = 1, \ldots, n$, so daß $A = QR$ gilt.

Bemerkung 3.15. Wir vermerken noch, daß die bei Berechnung der Lösung des linearen Ausgleichsproblems (3.1) im rang-defizienten Fall entstehende Lösung nicht zwingend diejenige mit minimaler euklidischer Norm ist. □

Kapitel 4

Funktionalanalytische Grundlagen I

Ab Kapitel 6 wollen wir iterative Verfahren zur Lösung von linearen und nichtlinearen Gleichungssystemen untersuchen. Dazu stellen wir in diesem und dem folgenden Kapitel elementare Grundlagen der Funktionalanalysis zusammen. Diese sind auch bei der quantitativen Fehleranalyse bei der Lösung von Gleichungssystemen (z.B. über den Begriff *Kondition*) nützlich.

Im vorliegenden Kapitel 4 betrachten wir zunächst *normierte Räume* und Räume mit *Skalarprodukt*. Dabei wird Basiswissen über lineare Räume vorausgesetzt. Dann betrachten wir *lineare beschränkte Operatoren* zwischen normierten Räumen. Dabei legen wir besonderen Wert auf den Fall linearer Operatoren zwischen endlichdimensionalen Räumen. Diese können durch Matrizen charakterisiert werden.

4.1 Normierte Räume. Prä-Hilbert-Räume

Nachfolgend führen wir mit den Begriffen *Norm* bzw. *Skalarprodukt* wesentliche Strukturen in der Klasse der linearen Räume ein. Sei wieder $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

Definition 4.1. *Sei X linearer Raum über $\mathbb{K}$. Dann heißt eine Abbildung $\|\cdot\| : X \mapsto \mathbb{R}$* Norm auf X, *falls folgende Eigenschaften für alle $x, y \in X$ und alle $\gamma \in \mathbb{K}$ gelten:*

(N1)	$\|x\|$	$\geq$	0	(Positivität)
(N2)	$\|x\|$	$=$	$0 \iff x = 0$	(Definitheit)
(N3)	$\|\gamma x\|$	$=$	$\|\gamma\|\|x\|$	(Homogenität)
(N4)	$\|x + y\|$	$\leq$	$\|x\| + \|y\|$	(Dreiecksungleichung).

Ein linearer Raum X mit Norm heißt normierter Raum.

Im Fall $X = \mathbb{K}^m$ verwenden wir speziell den Begriff *Vektornorm*. Wichtige Spezialfälle gibt

Satz 4.2. *Auf $\mathbb{K}^m$ sind für $x = (x_1, ..., x_m)^T$ spezielle Vektornormen gegeben durch :*

$$\|x\|_p := \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty \tag{4.1}$$

$$\|x\|_\infty := \max_{i=1,...,m} |x_i| \qquad \text{(Maximum-Norm)}. \tag{4.2}$$

Beweis: Übungsaufgabe □

Wichtige Spezialfälle sind

$$\|x\|_1 := \sum_{i=1}^{m} |x_i|, \qquad \text{(Betragssummen-Norm)} \tag{4.3}$$

$$\|x\|_2 := \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2}, \quad \text{(Euklidische Norm).} \tag{4.4}$$

Satz 4.3. *In einem normierten Raum X gilt für beliebige $x, y \in X$ die Ungleichung* (2. Dreiecksungleichung)

$$| \, \|x\| - \|y\| \, | \leq \|x - y\|. \tag{4.5}$$

Beweis. Nullergänzung und Dreiecksungleichung (N4) liefern

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|,$$

damit

$$\|x\| - \|y\| \leq \|x - y\|$$

und durch Vertauschung von x und y

$$\|y\| - \|x\| \leq \|y - x\|. \qquad \square$$

Wir betrachten nun lineare Räume mit Skalarprodukt.

Definition 4.4. *Sei X linearer Raum über $\mathbb{K}$. Eine Abbildung $(\cdot, \cdot) : X \to \mathbb{K}$ mit den Bedingungen*

(H1)	(x, x)	$\geq$	0		(Positivität)
(H2)	(x, x)	$=$	0	$\iff \quad x = 0$	(Definitheit)
(H3)	(x, y)	$=$	$\overline{(y, x)}$		(Symmetrie)
(H4)	$(\alpha x + \beta y, z)$	$=$	$\alpha(x, z) + \beta(y, z)$		(Linearität)

für alle $x, y, z \in X$ und $\alpha, \beta \in \mathbb{K}$ heißt Skalarprodukt *auf X. Ein linearer Raum mit Skalarprodukt heißt* Prä-Hilbert Raum.

Aus den Axiomen (H3) und (H4) folgt unmittelbar die Eigenschaft

(H4') $\quad (x, \alpha y + \beta z) = \overline{\alpha}(x, y) + \overline{\beta}(x, z) \qquad$ (Antilinearität)

sowie

Beispiel 4.5. *Auf $\mathbb{K}^m$ bildet*

$$(x, y) := \sum_{i=1}^{m} x_i \overline{y}_i \; = \; y^* x \tag{4.6}$$

mit $x = (x_1, ..., x_m)^T$ bzw. $y = (y_1, ..., y_m)^T$ ein Skalarprodukt. □

Satz 4.6. *Ein Skalarprodukt genügt der* Ungleichung von Cauchy-Schwarz

$$|(x, y)|^2 \leq (x, x)(y, y) \tag{4.7}$$

für alle $x, y \in X$. Genau für linear abhängige x und y gilt die Gleichheit.

Beweis. Für $x = 0$ ist die Ungleichung richtig. Im Falle $x \neq 0$ findet man mit

$$\alpha = -(x,x)^{-1/2}\overline{(x,y)}, \quad \beta = (x,x)^{1/2},$$

daß

$$\begin{aligned} 0 \leq (\alpha x + \beta y, \alpha x + \beta y) &= |\alpha|^2(x,x) + 2\mathrm{Re}\{\alpha\overline{\beta}(x,y)\} + |\beta|^2(y,y) \\ &= (x,x)(y,y) - |(x,y)|^2. \end{aligned} \tag{4.8}$$

Aus (4.8) folgt bereits die Behauptung. Insbesondere hat man Gleichheit genau dann, wenn $\alpha x + \beta y = 0$, d.h. im Falle linearer Abhängigkeit von x und y. □

Satz 4.7. *In jedem Prä-Hilbert Raum X ist durch*

$$\|x\| := (x,x)^{1/2}, \quad x \in X, \tag{4.9}$$

eine Norm erklärt. Damit ist jeder Prä-Hilbert Raum auch normierter Raum.

Beweis. Die Normeigenschaften (N1), (N2) und (N3) folgen jeweils aus den Axiomen (H1), (H2) sowie (H4) und (H4'). Die Dreiecksungleichung (N4) ist Folgerung aus Satz 4.6 wegen

$$\|x+y\|^2 = (x+y, x+y) \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

□

Damit läßt sich die Ungleichung von Cauchy-Schwarz auch schreiben als

$$|(x,y)| \leq \|x\| \ \|y\|. \tag{4.10}$$

Das in Beispiel 4.5 gegebene euklidische Skalarprodukt erzeugt gerade die euklidische Norm $\|\cdot\|_2$ (vgl. Satz 4.2).

4.2 Äquivalente Normen

Wir erinnern zunächst an einige Grundbegriffe zur Konvergenz von Folgen.

Definition 4.8. *Für beliebige Elemente x, y des normierten Raumes X heißt die Zahl $\|x-y\|$* Abstand.

Definition 4.9. *Eine Folge (x_n) von Elementen eines normierten Raumes X heißt* konvergent, *falls es ein Element x in X gibt mit*

$$\lim_{n\to\infty} \|x_n - x\| = 0, \tag{4.11}$$

d.h. zu jedem $\epsilon > 0$ existiert eine Zahl $N(\epsilon) \in \mathbb{N}$, so daß für alle $n \geq N(\epsilon)$ gilt $\|x_n - x\| \leq \epsilon$. Wir verwenden die Kurzschreibweise

$$\lim_{n\to\infty} x_n = x \quad \textit{oder} \quad x_n \to x, \ n \to \infty. \tag{4.12}$$

Eine nicht konvergente Folge heißt divergent.

Satz 4.10. *Das Grenzelement einer konvergenten Folge ist eindeutig bestimmt.*

Beweis. Wir nehmen an, daß x und y Grenzelemente der Folge (x_n) sind. Mittels Dreiecksungleichung (N4) folgt über Nullergänzung

$$\|x-y\| = \|x-x_n+x_n-y\| \le \|x-x_n\| + \|y-x_n\| \to 0, \quad n\to\infty.$$

Nach Axiom (N2) folgt damit $x=y$. □

Definition 4.11. *Zwei Normen auf einem linearen Raum heißen* äquivalent, *falls jede bezüglich der ersten Norm konvergente Folge auch bezüglich der zweiten Norm konvergent ist und umgekehrt.*

Satz 4.12. *Zwei Normen $\|\cdot\|_1$ und $\|\cdot\|_2$ auf einem linearen Raum X sind genau dann äquivalent, wenn positive Zahlen c und C existieren, so daß*

$$c\|x\|_1 \le \|x\|_2 \le C\|x\|_1 \quad \forall x\in X. \tag{4.13}$$

Die Grenzelemente bezüglich beider Normen sind gleich.

Beweis. (i) Wir nehmen zunächst an, daß die im Satz angegebene Ungleichung gilt. Dann folgt aus $\|x-x_n\|_1\to 0, n\to\infty$ die Aussage $\|x-x_n\|_2\to 0, n\to\infty$ und umgekehrt.

(ii) Sei nun die Äquivalenz beider Normen vorausgesetzt. Wir nehmen an, es existiert keine Zahl $C>0$ mit $\|x\|_2\le C$ für alle $x\in X$ mit $\|x\|_1=1$. Dann kann eine Folge (x_n) gewählt werden mit $\|x_n\|_1=1$ und $\|x_n\|_2>n^2$.
Mit $y_n = x_n/n$ folgt $\|y_n\|_2>n$, $\|y_n\|_1 = 1/n$. Dann konvergiert die Folge (y_n) bezüglich der Norm $\|\cdot\|_1$ und divergiert bezüglich $\|\cdot\|_2$ im Widerspruch zur angenommenen Normäquivalenz. Somit gibt es eine Zahl $C>0$ derart, daß $\|x\|_2\le C \quad \forall x\in X,\ \|x\|_1=1$. Aus der Forderung der Homogenität (N3) ergibt sich folglich

$$\|x\|_2 = \left\| \ \|x\|_1 \frac{x}{\|x\|_1} \ \right\|_2 \le C\|x\|_1 \quad \forall x\in X.$$

Die zweite Ungleichung folgt durch Vertauschung der Rolle beider Normen. □

Satz 4.13. *Auf einem endlich-dimensionalen Raum sind alle Normen äquivalent.*

Beweis. Sei X m-dimensionaler Raum mit der Basis $u_1,...,u_m$. Jedes Element von X besitzt dann die eindeutige Darstellung

$$x = \sum_{k=1}^{m} \alpha_k u_k.$$

Dann erklären wir durch den Ausdruck

$$\|x\|_\infty := \max_{k=1,...,m} |\alpha_k|$$

wie in Satz 4.2 die Maximum-Norm auf X.
Sei nun $\|\cdot\|$ eine beliebige andere Norm auf X. Die Idee des Beweises besteht darin, die Äquivalenz dieser Norm zur Maximum-Norm zu zeigen. Die Dreiecksungleichung liefert für beliebige Elemente x aus X

$$\|x\| \le \sum_{k=1}^{m} |\alpha_k|\ \|u_k\| \le C\|x\|_\infty, \quad C := \sum_{k=1}^{m} \|u_k\|.$$

Für die andere Richtung sei nun angenommen, daß keine Zahl $c > 0$ existiert mit $c\|x\|_\infty \leq \|x\|$ für alle $x \in X$. Dann findet man eine Folge $(x_n),\ \|x_n\| = 1$ mit $\|x_n\|_\infty > n$. Für die Folge $(y_n),\ y_n := x_n/\|x_n\|_\infty$ liefert die Basisdarstellung in X

$$y_n = \sum_{k=1}^{m} \alpha_{kn} u_k.$$

Jede der Folgen $(\alpha_{kn}), k = 1, ..., m$ ist wegen $\|y_n\|_\infty = 1$ beschränkt in $\mathbb{K}$. Nach dem Satz von Bolzano-Weierstraß können wir somit für jede Zahl $k = 1, ..., m$ konvergente Teilfolgen $\alpha_{k,n_{(j)}} \to \alpha_k, j \to \infty$ auswählen. Für das Element

$$y := \sum_{k=1}^{m} \alpha_k u_k$$

ergibt sich daraus $\|y_{n(j)} - y\|_\infty \to 0,\ j \to \infty$ und somit $\|y_{n(j)} - y\| \leq C\|y_{n(j)} - y\|_\infty \to 0,\ j \to \infty$. Andererseits hatten wir aber $\|y_n\| = 1/\|x_n\|_\infty \to 0, n \to \infty$. Folglich ist $y = 0$ und auch $\|y_{n(j)}\|_\infty \to 0,\ j \to \infty$. Das steht aber im Widerspruch zur Konstruktion mit $\|y_n\|_\infty = 1$ für beliebige Zahlen n. Die Annahme ist somit falsch. Daraus ergibt sich sich die zu beweisende Normäquivalenz. □

Der gerade bewiesene Satz wird sehr oft in Beweisen benutzt. Wichtig ist, daß er sich im allgemeinen Fall nicht auf Räume mit unendlicher Dimension (z.B. Funktionenräume) übertragen läßt. Ferner verdeckt die Formulierung, daß die Äquivalenzkonstanten c und C in der Regel von der Raumdimension m abhängen.

4.3 Lineare Operatoren auf normierten Räumen

Wir betrachten jetzt Abbildungen (bzw. Funktionen oder Operatoren) $A: X \to Y$ zwischen normierten Räumen X und Y, bei denen jedem Element aus X eindeutig ein Element $Ax \in Y$ zugeordnet wird. Man bezeichnet auch die Menge $A(X) := \{Ax |\ x \in X\}$ als Bildbereich von A.

Definition 4.14. *Für normierte Räume X und Y heißt der Operator $A : X \to Y$* linear, *falls gilt*

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay, \quad \forall x, y \in X, \quad \alpha, \beta \in \mathbb{K}. \tag{4.14}$$

Der Operator $A : X \to Y$ heißt stetig im Punkt $x \in X$, *wenn aus $x_n \to x,\ n \to \infty$ in X auch folgt, daß $Ax_n \to Ax,\ n \to \infty$ in Y.*

Der Operator A heißt stetig auf X, *falls er in jedem Punkt in X stetig ist.*

Ein linearer Operator $A : X \to Y$ heißt beschränkt, *falls eine positive Zahl C existiert, daß*

$$\|Ax\| \leq C\|x\|, \quad \forall x \in X. \tag{4.15}$$

Jede derartige Zahl C heißt Schranke für A.

Satz 4.15. *Ein linearer Operator $A : X \to Y$ ist genau dann beschränkt, wenn*

$$\|A\| := \sup_{\|x\|=1} \|Ax\| < \infty. \tag{4.16}$$

Die Zahl $\|A\|$ ist die kleinste Schranke des Operators A und heißt Norm von A. *Es gilt damit*

$$\|Ax\| \leq \|A\|\ \|x\|, \quad \forall x \in X. \tag{4.17}$$

Beweis. Sei zunächst A beschränkter Operator mit Schranke C, folglich

$$\sup_{\|x\|=1} \|Ax\| \leq C.$$

Damit ist speziell die Zahl $\|A\|$ nicht größer als jede Schranke von A.
Ist andererseits $\|A\| < \infty$, so ergeben für $x \neq 0$ die Linearität von A und die Homogenität der Norm

$$\|Ax\| = \left\| A\left(\|x\| \frac{x}{\|x\|} \right) \right\| = \left\| \|x\| A\left(\frac{x}{\|x\|} \right) \right\| = \|x\| \left\| A\left(\frac{x}{\|x\|} \right) \right\| \leq \|A\| \ \|x\|$$

wegen $\|x\|/\|x\| = 1$. Damit ist der Operator A beschränkt mit Schranke $\|A\|$. □

Satz 4.16. *Für lineare Operatoren $A : X \to Y$ sind die Begriffe "Stetigkeit" und "Beschränktheit" äquivalent.*

Beweis. Sei zunächst A stetig. Wir nehmen an, daß keine Konstante C existiert, so daß $\|Ax\| \leq C\|x\|$ für alle $x \in X$. Dann finden wir eine Folge (x_n) in X mit $\|x_n\| = 1, \|Ax_n\| > n$. Für die Folge $y_n := x_n/\|Ax_n\|$ gilt $y_n \to 0, n \to \infty$. Dann zieht die Stetigkeit von A nach sich, daß $Ay_n \to A(0) = 0, n \to \infty$ im Widerspruch zur Konstruktion $\|Ay_n\| = 1$ für alle n. Folglich muß A beschränkt sein.
Sei nun A beschränkt. Dann sei $x_n \to 0, n \to \infty$. Mit

$$\|Ax_n\| \leq \|A\|\|x_n\|$$

ergibt sich dann $Ax_n \to A(0) = 0, n \to \infty$, d.h. die Stetigkeit von A im Punkt 0. Die Stetigkeit in einem beliebigen Punkt sieht man wie folgt: Für $x_n \to x, n \to \infty$ ergibt die Linearität des Operators A, daß

$$A(x_n) = A(x_n - x) + A(x) \to A(0) + A(x) = A(x), \quad n \to \infty. \qquad \square$$

Künftig bezeichnen wir die Klasse linearer und stetiger Operatoren mit

$$\mathcal{L}(X, Y) := \{A : \ X \to Y \mid A \text{ linear}, \ \text{stetig}\}.$$

Satz 4.17. *Für normierte Räume X, Y und Z sowie lineare stetige Operatoren $A \in \mathcal{L}(X, Y)$ und $B \in \mathcal{L}(Y, Z))$ ist auch der durch die Vorschrift $(BA)x := B(Ax)$, $\forall x \in X$ definierte Operator $BA : X \to Z$ ein linearer stetiger Operator mit*

$$\|BA\| \leq \|A\|\|B\|.$$

Beweis. Dies folgt wegen $\|(BA)x\| = \|B(Ax)\| \leq \|B\|\|A\|\|x\|$. □

4.4 Matrixoperatoren. Matrix-Normen

Wir befassen uns im Rest des Kapitels mit linearen Operatoren $A : \ X \to X$ mit $X = \mathbb{K}^m$, die durch Matrizen charakterisiert werden. Einer Matrix $A = (a_{ik}) \in \mathbb{K}^{m \times m}$ wird durch die Vorschrift

$$(Ax)_i := \sum_{k=1}^{m} a_{ik} x_k, \ \ i = 1, ..., m \tag{4.18}$$

ein linearer Operator (Matrixoperator) $A : \mathbb{K}^m \to \mathbb{K}^m$ zugeordnet. Der folgende Satz zeigt die Aussage $A \in \mathcal{L}(\mathbb{K}^m, \mathbb{K}^m)$ und erlaubt für wichtige Normen deren Berechnung aus Matrixdaten.

Satz 4.18. *Der oben definierte Matrixoperator A ist in jeder Norm auf $\mathbb{K}^m$ beschränkt. Insbesondere gilt für spezielle Normen (in Anlehnung an Satz 4.2)*

$$\|A\|_1 := \sup_{\|x\|_1=1} \|Ax\|_1 = \max_{k=1,\dots,m} \sum_{i=1}^{m} |a_{ik}|, \qquad (Spaltensummennorm) \tag{4.19}$$

$$\|A\|_\infty := \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{i=1,\dots,m} \sum_{k=1}^{m} |a_{ik}|, \qquad (Zeilensummennorm) \tag{4.20}$$

$$\|A\|_2 := \sup_{\|x\|_2=1} \|Ax\|_2 \le (\sum_{i,k=1}^{m} |a_{ik}|^2)^{1/2}. \tag{4.21}$$

Die angegebenen Ausdrücke heißen Matrix-Normen. *In der letzten Ungleichung gilt im allgemeinen Fall nicht die Gleichheit.*

Beweis. Wegen der Äquivalenz aller Normen auf dem endlich-dimensionalen Raum $\mathbb{K}^m$ reicht es aus, die Beschränktheit des Matrixoperators in einer Norm nachzuweisen.

(i) Für $\|\cdot\|_2$ ist

$$\|Ax\|_2^2 = \sum_{i=1}^{m} |(Ax)_i|^2 = \sum_{i=1}^{m} |\sum_{k=1}^{m} a_{ik}x_k|^2$$

und nach der Ungleichung von Cauchy-Schwarz

$$\|Ax\|_2^2 \le \sum_{i=1}^{m} \left(\sum_{k=1}^{m} |a_{ik}|^2 \sum_{k=1}^{m} |x_k|^2 \right) = \sum_{i,k=1}^{m} |a_{ik}|^2 \underbrace{\sum_{k=1}^{m} |x_k|^2}_{=\|x\|_2^2=1}.$$

Supremums-Bildung ergibt

$$\|A\|_2 \le \left(\sum_{i,k=1}^{m} |a_{ik}|^2 \right)^{1/2}.$$

Man sieht am Fall $A = I$, daß die Gleichheit im allgemeinen Fall nicht gilt.

(ii) Wir zeigen nun die im Satz angegebene Identität für die Spaltensummen-Norm $\|\cdot\|_1$. Einerseits gilt

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^{m} |(Ax)_i| = \sum_{i=1}^{m} |\sum_{k=1}^{m} a_{ik}x_k| \\ &\le \sum_{k=1}^{m} |x_k| \sum_{i=1}^{m} |a_{ik}| \le \max_{k=1,\dots,m} \sum_{i=1}^{m} |a_{ik}| \textstyle\sum_{k=1}^{m} |x_k|, \end{aligned}$$

damit

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

Andererseits sei jetzt der Index j so gewählt, daß

$$\sum_{i=1}^{m} |a_{ij}| = \max_{k=1,\dots,m} \sum_{i=1}^{m} |a_{ik}|.$$

Mit der Wahl von $z \in \mathbb{R}^m$ mit $z_j = 1$ und $z_k = 0,\ k \neq j$ ergibt sich $\|z\|_1 = 1$ und

$$\|Az\|_1 = \sum_{i=1}^m |(Az)_i| = \sum_{i=1}^m \left| \sum_{k=1}^m a_{ik} z_k \right| = \sum_{i=1}^m |a_{ij}| = \max_{k=1,\ldots,m} \sum_{i=1}^m |a_{ik}|.$$

Damit folgt aber

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 \geq \|Az\|_1 = \max_{k=1,\ldots,m} \sum_{i=1}^m |a_{ik}|.$$

(iii) Für die Maximum-Norm $\|\cdot\|_\infty$ erhalten wir analog

$$\begin{aligned} \|Ax\|_\infty &= \max_{i=1,\ldots,m} |(Ax)_i| = \max_{i=1,\ldots,m} |\sum_{k=1}^m a_{ik} x_k| \\ &\leq \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}||x_k| \leq \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}| \underbrace{\max_{k=1,\ldots,m} |x_k|}_{=\|x\|_\infty}, \end{aligned}$$

also

$$\|A\|_\infty \leq \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}|.$$

Schließlich wählen wir den Index j so, daß

$$\sum_{k=1}^m |a_{jk}| = \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}|$$

und $z \in \mathbb{K}^m$ mit

$$z_k = \overline{a}_{jk}/a_{jk}, \ \text{ falls } \ a_{jk} \neq 0, \qquad z_k = 1, \ \text{ falls } \ a_{jk} = 0.$$

Dann ist $\|z\|_\infty = 1$ sowie

$$\begin{aligned} \|Az\|_\infty &= \max_{i=1,\ldots,m} |(Az)_i| = \max_{i=1,\ldots,m} \left| \sum_{k=1}^m a_{ik} z_k \right| \\ &\geq \left| \sum_{k=1}^m a_{jk} z_k \right| = \sum_{k=1}^m |a_{jk}| = \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}|. \end{aligned}$$

Daraus können wir aber folgern, daß

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty \geq \|Az\|_\infty = \max_{i=1,\ldots,m} \sum_{k=1}^m |a_{ik}|.$$

Damit ist der Beweis erbracht. □

4.5 Eigenwerte und Eigenvektoren

Unser Ziel ist weiterhin die Ableitung einer geeigneten Darstellung der Matrix-Norm $\|A\|_2$. Dazu benötigen wir einige Begriffe aus der Linearen Algebra.

Definition 4.19. *Eine komplexe Zahl λ heißt* Eigenwert der Matrix A, *falls es einen Vektor $x \in \mathbb{C}^m, x \neq 0$ derart gibt, daß $Ax = \lambda x$. Der Vektor x heißt zum Eigenwert λ gehörender* Eigenvektor.

Jede $m \times m$-Matrix A hat mindestens einen und höchstens m Eigenwerte. Dies folgt aus dem Fundamentalsatz der Algebra für das charakteristische Polynom $\det(A - \lambda I)$. Ferner sind Eigenvektoren zu verschiedenen Eigenwerten linear unabhängig.

In den nächsten Abschnitten benötigen wir die folgenden Aussagen.

Satz 4.20. *(Lemma von Schur)*
Jeder Matrix A läßt sich eine unitäre Matrix Q zuordnen, so daß $\tilde{A} = Q^*AQ$ obere Dreiecksmatrix ist.

Beweis. Sei λ Eigenwert einer Matrix $A_m := A \in \mathbb{K}^{m\times m}$ mit dem (o.B.d.A. orthonormierten) Eigenvektor u, d.h. $(u, u) = 1$. Man kann u (zum Beispiel unter Nutzung des Orthogonalisierungsverfahrens von Gram-Schmidt) zu einer orthonormalen Basis $u, v_2, ..., v_m$ des Raumes $\mathbb{K}^m$ erweitern. Offenbar ist die Matrix

$$U := (u\ v_2\ ...\ v_m) \tag{4.22}$$

unitär. Unter Beachtung von $(u, v_i) = 0,\ \ i = 2, ..., m$ finden wir

$$U^*A_mU = U^*(\lambda u\ A_mv_2\ ...\ A_mv_m) = \begin{pmatrix} \lambda & * \\ 0 & A_{m-1} \end{pmatrix} \tag{4.23}$$

mit einer Matrix $A_{m-1} \in \mathbb{K}^{(m-1)\times(m-1)}$. Nun verfährt man induktiv. □

Satz 4.21. *Sei A^* die adjungierte Matrix zur Matrix $A \in \mathbb{K}^{m\times m}$. Dann gilt*

$$(Ax, y) = (x, A^*y), \quad \forall x, y \in \mathbb{K}^m. \tag{4.24}$$

Beweis. Mit der Notation ${a_{ik}}^* = \overline{a_{ki}}$ gilt

$$\begin{aligned}(Ax, y) &= \sum_{i=1}^{m}(Ax)_i\overline{y}_i = \sum_{i=1}^{m}\sum_{k=1}^{m} a_{ik}x_k\overline{y}_i \\ &= \sum_{k=1}^{m}\sum_{i=1}^{m} x_k\overline{{a_{ki}}^*y_i} = \sum_{k=1}^{m} x_k\overline{A^*y_k} = (x, A^*y).\end{aligned}$$

. □

Wir betrachten nun den wichtigen Spezialfall einer *hermiteschen* Matrix A, d.h. $A = A^*$. Dann ist auch die Matrix $\tilde{A} := Q^*AQ$ aus dem Satz 4.20 hermitesch, denn

$$\tilde{A}^* = (Q^*AQ)^* = Q^*A^*Q^{**} = Q^*AQ = \tilde{A}.$$

Folglich ist die obere Dreiecksmatrix $\tilde{A}$ sogar Diagonalmatrix

$$\tilde{A} = D = \mathrm{diag}(\lambda_1, ..., \lambda_m).$$

Wegen $Q^*AQ = D$ ist $AQ = QD$. Daher gilt für die Spalten der Matrix $Q = (u_1\ ....\ u_m)$, daß $Au_i = \lambda_iu_i, i = 1, ..., m$. Somit bilden die Eigenvektoren hermitescher Matrizen eine orthonormale Basis des $\mathbb{K}^m$. Darüber hinaus sind deren Eigenwerte reell wegen

$$\lambda_i = (Au_i, u_i) = (u_i, Au_i) = \overline{(Au_i, u_i)} = \overline{\lambda_i}.$$

Für eine hermitesche und *positiv semidefinite* Matrix, d.h. mit

$$(Ax, x) \geq 0, \quad \forall x \in \mathbb{K}^m,$$

sind speziell alle Eigenwerte nicht negativ. Im Fall einer hermiteschen und *positiv definiten* Matrix, d.h. mit

$$(Ax, x) > 0, \quad \forall x \in \mathbb{K}^m, \quad x \neq 0,$$

sind sogar alle Eigenwerte positiv.

4.6 Spektralradius einer Matrix

Wir charakterisieren nun über den Begriff "Spektralradius" die Matrixnorm $\|A\|_2$.

Definition 4.22. *Die nichtnegative Zahl*

$$\rho(A) := \max\{|\lambda| : \lambda \ \textit{Eigenwert von} \ A\} \tag{4.25}$$

heißt Spektralradius der Matrix A.

Satz 4.23. *Für eine beliebige Matrix A gilt*

$$\|A\|_2 = \sqrt{\rho(A^*A)}, \tag{4.26}$$

speziell für hermitesche Matrizen

$$\|A\|_2 = \rho(A). \tag{4.27}$$

Beweis. Nach Satz 4.21 gilt

$$0 \leq \|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax), \quad \forall x \in \mathbb{K}^m.$$

Damit ist die hermitesche Matrix A^*A positiv semidefinit und sie hat reelle, nichtnegative Eigenwerte σ_i^2 und zugehörige orthonormale Eigenvektoren $u_i,\ i = 1, \ldots, m$ mit

$$A^*Au_i = \sigma_i^2 u_i, \ \ i = 1, ..., m.$$

Bei Darstellung in der Orthonormalbasis

$$x = \sum_{i=1}^{m} \alpha_i u_i$$

folgt

$$\|x\|_2^2 = (x, x) = \left(\sum_{i=1}^{m} \alpha_i u_i, \sum_{j=1}^{m} \alpha_j u_j \right) = \sum_{i=1}^{m} |\alpha_i|^2$$

bzw.

$$\|Ax\|_2^2 = (x, A^*Ax) = \left(\sum_{i=1}^{m} \alpha_i u_i, \sum_{j=1}^{m} \sigma_j^2 \alpha_j u_j \right) = \sum_{i=1}^{m} \sigma_i^2 |\alpha_i|^2.$$

Daraus ergibt sich

$$\|Ax\|_2^2 \leq \rho(A^*A)\|x\|_2^2$$

und

$$\|A\|_2^2 \leq \rho(A^*A).$$

Wir wählen nun den Index j so , daß $\sigma_j^2 = \rho(A^*A)$. Dann folgt

$$\|A\|_2^2 = \left\{ \sup_{\|x\|_2=1} \|Ax\|_2 \right\}^2 \geq \|Au_j\|_2^2 = (u_j, A^*Au_j) = \sigma_j^2 = \rho(A^*A).$$

Falls A hermitesch ist, so folgt $A^*A = A^2$ und $\rho(A^*A) = \rho(A^2) = \{\rho(A)\}^2$. □

Im Abschnitt über Iterationsverfahren für lineare Gleichungssysteme wollen wir Konvergenzkriterien ableiten. Dazu benutzen wir wesentlich folgende Aussage.

Satz 4.24. *Für jede Matrixnorm auf $\mathbb{K}^m$ und jede Matrix $A \in \mathbb{K}^{m\times m}$ gilt*

$$\rho(A) \leq \|A\|. \tag{4.28}$$

Andererseits existiert zu jeder Matrix A und jedem positivem Wert ϵ eine Norm $\|\cdot\|_\epsilon$ auf $\mathbb{K}^m$ so, daß

$$\|A\|_\epsilon \leq \rho(A) + \epsilon. \tag{4.29}$$

Beweis. Sei λ Eigenwert zur Matrix A mit dem o.B.d.A. normierten Eigenvektor u. Die erste Behauptung ergibt sich dann aus

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \geq \|Au\| = \|\lambda u\| = |\lambda|.$$

Zum Beweis des zweiten Teils des Satzes konstruieren wir nach Satz 4.20 eine unitäre Matrix Q so, daß

$$B = Q^{-1}AQ = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \cdots & b_{1m} \\ & b_{22} & b_{23} & \cdots & b_{2m} \\ & & b_{33} & \cdots & b_{3m} \\ & & & \cdots & \vdots \\ & & & & b_{mm} \end{pmatrix}$$

obere Dreiecksgestalt hat. Wegen $\det(\lambda I - A) = \det(\lambda I - B)$ sind die Diagonalelemente der Matrix B gerade die Eigenwerte $\lambda_i = b_{ii}, i = 1, ..., m$ von A. Mit den Festlegungen

$$b := \max_{i,k=1,...,m} |b_{ik}|, \quad \delta := \min\left(1, \frac{\epsilon}{(m-1)b}\right)$$

bilden wir die Diagonalmatrix

$$D := \mathrm{diag}(1, \delta, \delta^2, ..., \delta^{m-1})$$

und deren inverse Matrix

$$D^{-1} = \mathrm{diag}(1, \delta^{-1}, \delta^{-2}, ..., \delta^{-m+1}).$$

Matrixmultiplikation ergibt

$$BD = \begin{pmatrix} b_{11} & \delta b_{12} & \delta^2 b_{13} & \cdots & \delta^{m-1} b_{1m} \\ & \delta b_{22} & \delta^2 b_{23} & \cdots & \delta^{m-1} b_{2m} \\ & & \delta^2 b_{33} & \cdots & \delta^{m-1} b_{3m} \\ & & & \cdots & \vdots \\ & & & & \delta^{m-1} b_{mm} \end{pmatrix}$$

bzw.

$$C := D^{-1}BD = \begin{pmatrix} b_{11} & \delta b_{12} & \delta^2 b_{13} & \cdots & \delta^{m-1} b_{1m} \\ & b_{22} & \delta b_{23} & \cdots & \delta^{m-2} b_{2m} \\ & & b_{33} & \cdots & \delta^{m-3} b_{3m} \\ & & & \cdots & \vdots \\ & & & & b_{mm} \end{pmatrix}.$$

Satz 4.18 liefert nun wegen $\delta \leq 1$

$$\|C\|_\infty \leq \max_{i=1,\ldots,m} |b_{ii}| + (m-1)\delta b \leq \rho(A) + \epsilon.$$

Mit der Festsetzung $V := QD$ definieren wir auf $\mathbb{K}^m$ eine Norm gemäß

$$\|x\|_\epsilon := \|V^{-1}x\|_\infty.$$

In dieser Norm gilt dann wegen $C = V^{-1}AV$

$$\|Ax\|_\epsilon = \|V^{-1}Ax\|_\infty = \|CV^{-1}x\|_\infty \leq \|C\|_\infty \|V^{-1}x\|_\infty = \|C\|_\infty \|x\|_\epsilon,$$

folglich

$$\|A\|_\epsilon \leq \|C\|_\infty \leq \rho(A) + \epsilon.$$

Damit ist die Behauptung des Satzes bewiesen. □

4.7 Kondition von Matrizen

Definition 4.25. *Für eine regulläre Matrix A heißt die nicht negative Zahl*

$$\mathrm{cond}(A) := \|A\|\|A^{-1}\| \tag{4.30}$$

Konditionszahl von A.

Dieser Begriff ist natürlich abhängig von der in der Definition benutzten Norm auf $\mathbb{K}^m$, jedoch gilt

$$\mathrm{cond}(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1.$$

Von praktischer Bedeutung ist nun das folgende *Störungslemma*, das wir im Kapitel 5 noch verallgemeinern werden.

Satz 4.26. *Bei regulärer Matrix A seien x sowie $x + x_\delta$ Lösungen des Gleichungssystems $Ax = y$ bzw. des gestörten Problems $A(x + x_\delta) = y + y_\delta$. Dann gilt die Stabilitätsabschätzung (bei Störungen der rechten Seite)*

$$\frac{\|x_\delta\|}{\|x\|} \leq \mathrm{cond}(A) \ \frac{\|y_\delta\|}{\|y\|}. \tag{4.31}$$

Beweis. Durch Subtraktion der beiden Gleichungssysteme folgen die Aussagen $A(x_\delta) = y_\delta$ und

$$\|x_\delta\| = \|A^{-1}(y_\delta)\| \leq \|A^{-1}\|\|y_\delta\|.$$

Daraus ergibt sich über

$$\|y\| = \|Ax\| \leq \|A\|\|x\|$$

schließlich die Behauptung des Satzes. □

Dieses Resultat erlaubt folgende Interpretation: Bei kleiner Konditionszahl ist die Störung der Lösung bei kleiner Änderung der rechten Seite des Gleichungssystems klein, d.h. das Problem ist gut konditioniert. Andererseits beschreibt eine große Konditionszahl auch ein Problem mit schlechter Kondition.

Lemma 4.27. *Für eine hermitesche Matrix $A = A^*$ gilt bezüglich der euklidischen Norm $\|\cdot\|_2$ folgende Charakterisierung:*

$$\mathrm{cond}_2(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|}. \tag{4.32}$$

Dabei sind λ_{max} und λ_{min} der betragsmäßig größte bzw. kleinste Eigenwert der Matrix A.

Beweis: Man benutzt Satz 4.21 *Übungsaufgabe !*). □

Bei schlecht konditionierten Gleichungssystemen können beim Gauß-Algorithmus trotz vollständiger Pivotisierung aufgrund von Daten- und/oder Rundungsfehlern auch große Fehler im Lösungsvektor auftreten. In Anlehnung an das Beispiel 1.2 (im Grenzfall $n \to \infty$) betrachten wir ein derartiges Problem.

Beispiel 4.28. (*Ein schlecht konditioniertes Gleichungssystem*)

Wir approximieren eine gegebene Funktion $u(t) : [0,1] \mapsto \mathbb{R}$ durch ein Polynom $g(t,a) = \sum_{k=0}^{n} a_k t^k$ mit $a = (a_0, a_1, ..., a_n)^T$ im Sinne der kleinsten Quadrate, d.h. wir minimieren den Ausdruck

$$f(a) := \int_0^1 \{u(t) - g(t,a)\}^2 dt.$$

Daraus erhält man als notwendige Bedingungen die Normalgleichungen der Methode der kleinsten Quadrate

$$\frac{\partial f}{\partial a_k} = 2\int_0^1 \{\sum_{i=0}^{n} a_i t^i - u(t)\} t^k dt = 0, \quad k = 0, ..., n \tag{4.33}$$

d.h. nach Integration ein $(n+1)$-dimensionales, lineares Gleichungssystem

$$\sum_{i=0}^{n} \frac{1}{k+i+1} a_i = \int_0^1 u(t) t^k dt, \quad k = 0, ..., n \tag{4.34}$$

für den Parametervektor $a = (a_0, , ..., a_n)^T$. Die Matrix $A = \left(\frac{1}{i+k+1}\right)_{i,k=0}^{n}$ heißt auch *Hilbert-Matrix*. Man kann über die Untersuchung des homogenen Systems zeigen, daß das System (4.34) eindeutig lösbar ist.

Sei nun speziell $u(t) = \frac{1}{1+t}$. Man erhält für die rechte Seite

$$y_i = \int_0^1 \frac{t^i}{t+1} dt = \int_0^1 t^{i-1} dt - y_{i-1} = \frac{1}{i} - y_{i-1}, \quad i = 1, ..., n$$

mit dem Anfangswert $y_0 = \ln 2$.

In der folgenden Tabelle wird die Lösung des Gleichungssystems mittels des Eliminationsverfahrens nach Gauß bei exakter Verarbeitung der auftretenden rationalen Zahlen und Verwendung eines Näherungswertes für $\ln 2$ mit einer Genauigkeit von 10 Dezimalstellen angegeben.

n	a_0	a_1	a_2	a_3	a_4	a_5	a_6
1	0.93147	-0.47664					
2	0.98603	-0.80405	0.32740				
3	0.99727	-0.93892	0.66459	-0.22479			
4	0.99948	-0.98302	0.86302	-0.53344	0.15432		
5	0.99990	-0.99563	0.95129	-0.76886	0.41916	-0.10593	
6	0.99998	-0.99894	0.98436	-0.90114	0.66720	-0.32421	0.07275

Bei Verwendung einer 5-stelligen Genauigkeit für $\ln 2$, d.h. bei Änderung der rechten Seite um maximal den Wert 0.000005, ergeben sich in der nächsten Tabelle überraschend große Abweichungen gegenüber den zuvor ermittelten Werten.

n	a_0	a_1	a_2	a_3	a_4	a_5	a_6
1	0.93	-0.47					
2	0.98	-0.80	0.32				
3	0.99	-0.95	0.70	-0.24			
4	1.00	-1.16	1.63	-1.69	0.72		
5	1.06	-2.74	12.68	-31.16	33.87	-13.25	
6	1.39	-16.58	151.09	-584.79	1071.93	-926.75	304.49

Dieses Resultat ist Folge der schlechten Kondition der Matrix des Gleichungssystems. Zur Illustration veranschaulicht die nachfolgende Tabelle die schlechte Kondition der Matrix.

n	2	3	4	5	6
λ_{max}	1.27	1.41	1.50	1.57	1.62
λ_{min}	$6.57 \cdot 10^{-2}$	$2.69 \cdot 10^{-3}$	$9.67 \cdot 10^{-5}$	$3.29 \cdot 10^{-6}$	$1.08 \cdot 10^{-7}$
cond_2	19.3	$5.24 \cdot 10^{2}$	$1.55 \cdot 10^{4}$	$4.77 \cdot 10^{5}$	$1.50 \cdot 10^{7}$

Wir werden im Rahmen dieser Vorlesung der Frage, wie schlecht konditionierte Gleichungssysteme (näherungsweise) gelöst werden können, nicht weiter nachgehen können. Es sei hierzu auf das Lehrbuch von R. Kreß *Numerical Analysis*, Springer-Verlag 1998 hingewiesen. Man findet dort im Kapitel 5 dazu Aussagen unter den Themen *Singulärwert-Zerlegung* einer Matrix und *Regularisierung nach Tikhonov*.

Kapitel 5

Funktionalanalytische Grundlagen II

Im vorliegenden Abschnitt erweitern wir unsere funktionalanalytischen Grundkenntnisse (vgl. auch Kapitel 4). Grundlegend für die Untersuchungen zu Iterationsverfahren bei linearen Gleichungssystemen (vgl. Kapitel 6) sowie für Auflösungsverfahren bei nichtlinearen Gleichungssystemen (vgl. Kapitel 7 und 8) ist vor allem der *Fixpunktsatz von S. Banach.* Ferner beweisen wir einen Existenzsatz sowie einen allgemeinen *Störungssatz* für lineare Operatorgleichungen.

5.1 Banach-Räume

Wir erinnern zunächst an den Begriff der *Cauchy-Folge* und Aussagen über derartige Folgen in normierten Räumen.

Definition 5.1. *Eine Folge (x_n) von Elementen eines normierten Raumes X heißt* Cauchy-Folge, *falls*

$$\lim_{m,n\to\infty} \|x_m - x_n\| = 0, \tag{5.1}$$

d.h. zu jedem Wert $\epsilon > 0$ existiert eine Zahl $N(\epsilon) \in \mathbb{N}$ mit

$$\|x_n - x_m\| < \epsilon, \quad \forall n, m \geq N(\epsilon). \tag{5.2}$$

Satz 5.2. Jede konvergente Folge in einem normierten Raum ist dort auch Cauchy-Folge.

Beweis. Die Folge x_n konvergiere gegen x. Dann findet man zu jedem Wert $\epsilon > 0$ eine Zahl $N(\epsilon) \in \mathbb{N}$ mit $\|x_n - x\| < \epsilon/2 \quad \forall n \geq N(\epsilon)$. Die Dreiecksungleichung ergibt

$$\|x_n - x_m\| \leq \|x_n - x\| + \|x - x_m\| < \epsilon, \ \ \forall n, m \geq N(\epsilon). \qquad \square$$

Im allgemeinen Fall gilt nicht die Umkehrung des Satzes 5.2. Dies motiviert die folgende

Definition 5.3. *Eine Teilmenge U eines normierten Raumes heißt* vollständig, *falls jede Cauchy-Folge aus der Menge U gegen ein Element in U konvergiert. Vollständige normierte Räume werden als* Banach-Räume *bezeichnet.*

Wir bemerken, daß die Menge der rationalen Zahlen nicht vollständig ist in der Menge $\mathbb{R}$. Weiterhin gilt der folgende wichtige

Satz 5.4. *Jeder endlich-dimensionale normierte Raum ist Banach-Raum.*

Beweis. Wir betrachten einen endlich-dimensionalen linearen Raum X mit der Basis $u_1, ..., u_m$ und eine Cauchy-Folge (x_n) in X. In Basisdarstellung hat man

$$x_n = \sum_{k=1}^{m} \alpha_{nk} u_k.$$

Nach dem Satz 4.13 über die Äquivalenz aller Normen auf Räumen mit endlicher Dimension findet man eine Zahl $C > 0$ so, daß

$$\max_{k=1,...,m} |\alpha_{nk} - \alpha_{lk}| =: \|x_n - x_l\|_\infty \leq C\|x_n - x_l\|, \quad \forall n, l \in \mathbb{N}.$$

Für $k = 1, ..., m$ sind dann die Folgen (α_{nk}) auf $\mathbb{K}$ Cauchy-Folgen. Nach dem Cauchyschen Konvergenzkriterium impliziert dies $\alpha_{nk} \to \alpha_k, n \to \infty$ und

$$x_n \to x := \sum_{k=1}^{m} \alpha_k u_k, \quad n \to \infty. \qquad \square$$

Die Aussage des vorgehenden Satzes überträgt sich im allgemeinen Fall nicht auf unendlichdimensionale normierte Räume. Für unsere späteren Betrachtungen benötigen wir den Raum $C[a, b]$ der stetigen Funktionen $f : [a, b] \to \mathbb{R}$. Mit der punktweise vorgenommenen Addition und Skalarmultiplikation

$$(f + g)(x) := f(x) + g(x), \quad (\alpha f)(x) := \alpha f(x)$$

ist $C[a, b]$ linearer Raum. Ferner ist der Raum nicht von endlicher Dimension, da die Monome $x \mapsto x^n, n = 0, 1, 2, ...$ linear unabhängig sind. Die folgenden Betrachtungen zeigen, daß die Charakterisierung dieses Raums als Banach-Raum von der verwendeten Norm abhängt.

Satz 5.5. *Der Vektorraum $C[a, b]$ ist mit der* Maximum-Norm

$$\|f\|_\infty := \max_{x\in[a,b]} |f(x)| \tag{5.3}$$

ein Banach-Raum.

Beweis. Wir zeigen die Normeigenschaften: Die Eigenschaften $(N1) - (N3)$ sind offensichtlich erfüllt. Die Gültigkeit der Dreiecksungleichung ersehen wir aus der folgenden Abschätzung:

$$\begin{aligned} \|f + g\|_\infty &= \max_{x\in[a,b]} |(f + g)(x)| = |(f + g)(x_0)| \leq |f(x_0)| + |g(x_0)| \\ &\leq \max_{x\in[a,b]} |f(x)| + \max_{x\in[a,b]} |g(x)| = \|f\|_\infty + \|g\|_\infty. \end{aligned}$$

Bei diesen Abschätzungen wurde angenommen, daß (nach dem Satz von Weierstraß) das Maximum von $f + g$ an der Stelle x_0 angenommen wird.

Schließlich untersuchen wir die Eigenschaft der Vollständigkeit. Bekanntlich entspricht die Konvergenz einer Funktionenfolge (f_n) aus dem Raum $C[a, b]$ gegen eine Funktion f bezüglich der Maximum-Norm gerade der gleichmäßigen Konvergenz, denn

$$\|f - f_n\|_\infty < \epsilon \quad \Longleftrightarrow \quad |f(x) - f_n(x)| < \epsilon, \quad \forall x \in [a, b].$$

Da nun das Cauchysche Konvergenzkriterium hinreichend für die gleichmäßige Konvergenz einer Folge stetiger Funktionen gegen eine stetige Grenzfunktion ist, folgt damit die Vollständigkeit des Raumes $C[a, b]$ bezüglich der Maximum-Norm. $\square$

Satz 5.6. *Der lineare Raum $C[a, b]$ ist in Verbindung mit der L_1-Norm*

$$\|f\|_1 := \int_a^b |f(x)|dx \tag{5.4}$$

normierter Raum, jedoch nicht vollständig.

Beweis. Die Gültigkeit der Normeigenschaften $(N1) - (N4)$ ist wieder offensichtlich. Bezüglich der Vollständigkeit konstruieren wir ein Gegenbeispiel. Sei $[a, b] = [0, 2]$ und

$$f_n(x) := \begin{cases} x^n, & 0 \le x \le 1, \\ 1, & 1 \le x \le 2. \end{cases}$$

Dann ist (f_n) Cauchy-Folge, denn für $n < m$ gilt

$$\|f_n - f_m\|_1 = \int_0^1 (x^n - x^m)dx \le \frac{2}{n+1} \to 0, \quad n \to \infty.$$

Wir nehmen nun an, daß die Folge (f_n) gegen eine stetige Funktion f konvergiert, d.h.

$$\|f_n - f\|_1 \to 0, \ \ n \to \infty.$$

Durch Nullergänzung finden wir

$$\int_0^1 |f(x)|dx \le \int_0^1 |f(x) - x^n|dx + \int_0^1 x^n dx \le \|f - f_n\|_1 + \frac{1}{n+1} \to 0, n \to \infty.$$

Folglich verschwindet $f(x)$ identisch auf dem Intervall $[0, 1]$. Andererseits ist

$$\int_1^2 |f(x) - 1|dx = \int_1^2 |f(x) - f_n(x)|dx \le \|f - f_n\|_1 \to 0, \quad n \to \infty.$$

Das impliziert $f(x) = 1, \ \ 1 \le x \le 2$. Damit ist aber die Grenzfunktion nicht stetig auf $[0, 2]$ im Widerspruch zur Annahme. Daraus folgt die Behauptung. □

Satz 5.7. *Der lineare Raum $C[a, b]$ ist in Verbindung mit der L_2–Norm*

$$\|f\|_2 := \left(\int_a^b |f(x)|^2 dx \right)^{1/2} \tag{5.5}$$

normierter Raum, jedoch nicht vollständig.

Beweis. Wir hatten bereits in Kapitel 4 gesehen, daß das Skalarprodukt $(f, g) := \int_a^b f(x)g(x)dx$ eine Norm erzeugt. Unter Verwendung der gleichen Folge wie im vorhergehenden Satz zeigt man auch hier die Nichtvollständigkeit. □

Die Aussage der Sätze 5.6 und 5.7 gilt sogar allgemeiner für den Raum $C[a, b]$ mit der L_p-Norm

$$\|f\|_p := \left(\int_a^b |f(x)|^p \ dx \right)^{1/p}, \quad 1 \le p < \infty.$$

Man ersieht also aus den vorhergehenden Aussagen eine Sonderrolle des Falles $p = \infty$. Für die funktionalanalytische Untersuchung von Differential- und Integralgleichungsprobleme erweist es sich oft als sehr sinnvoll, den Raum der stetigen Funktionen in geeigneter Weise bezüglich der Integralnormen $\| \cdot \|_p$ mit $1 \le p < \infty$ zu vervollständigen.

5.2 Fixpunktsatz von Banach

Wir führen zunächst einige Begriffe ein.

Definition 5.8. *Eine Teilmenge U eines normierten Raumes X heißt* abgeschlossen, *wenn sie alle Grenzelemente konvergenter Folgen aus U enthält.*

Per Definition des Vollständigkeits-Begriffes gilt dann

Satz 5.9. *Jede abgeschlossene Teilmenge einer vollständigen Teilmenge eines normierten Raumes ist vollständig.*

Definition 5.10. *Sei U Teilmenge eines normierten Raumes X. Ein Operator $A : U \to X$ heißt* Kontraktionsoperator, *falls eine Zahl $q \in [0, 1)$ existiert, so daß*

$$\|Ax - Ay\| \leq q\|x - y\| \quad \forall x, y \in U. \tag{5.6}$$

Jede derartige Zahl q heißt Kontraktionszahl *von A.*

Man beachte, daß der Operator A nicht zwingend linear sein muß. Aus der Ungleichung in dieser Definition folgt der

Satz 5.11. *Ein Kontraktionsoperator ist stetig.*

Definition 5.12. *Jedes Element x eines normierten Raumes X mit der Eigenschaft*

$$Ax = x \tag{5.7}$$

heißt Fixpunkt *des Operators $A : U \subset X \to X$.*

Wir können nun den zentralen Satz dieses Abschnittes formulieren.

Satz 5.13. (Fixpunktsatz von S. Banach)
Ein Kontraktionsoperator, der eine vollständige Teilmenge U eines normierten Raumes X in sich abbildet, besitzt genau einen Fixpunkt.

Beweis. Sei $A : U \to U \subset X$ Kontraktionsoperator mit der Kontraktionszahl $q \in [0, 1)$ und U vollständig. Wir wählen ein beliebiges Startelement $x_0 \in U$ und erklären die Folge (x_n) in U durch die Iterationsvorschrift *(sukzessive Approximation)*

$$x_{n+1} := Ax_n, \quad n = 0, 1, 2, ... \tag{5.8}$$

Per Definition gilt

$$\|x_{n+1} - x_n\| = \|Ax_n - Ax_{n-1}\| \leq q\|x_n - x_{n-1}\|$$

und damit durch vollständige Induktion

$$\|x_{n+1} - x_n\| \leq q^n\|x_1 - x_0\|, \quad n = 1, 2, ...$$

Wir können nun folgern, daß (x_n) Cauchy-Folge ist, denn für $m \geq n$

$$\begin{aligned} \|x_n - x_m\| &\leq \|x_n - x_{n+1}\| + \|x_{n+1} - x_{n+2}\| + \cdots + \|x_{m-1} - x_m\| \\ &\leq (q^n + q^{n+1} + \cdots + q^{m-1})\|x_1 - x_0\| \\ &\leq \frac{q^n}{1-q}\|x_1 - x_0\| \to 0, \quad n \to \infty. \end{aligned} \tag{5.9}$$

Wegen der Vollständigkeit der Menge U findet man ein Element $x \in U$ mit $x_n \to x, n \to \infty$. Aufgrund der Stetigkeit des Kontraktionsoperators A nach Satz 5.11 folgern wir

$$x = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} Ax_n = Ax,$$

d.h. x ist Fixpunkt des Operators.
Es bleibt der Nachweis der Eindeutigkeit des Fixpunktes: Wir nehmen an, daß x und y Fixpunkte von A mit $x \neq y$ sind. Dann ergibt sich wegen

$$0 \neq \|x - y\| = \|Ax - Ay\| \leq q\|x - y\|$$

die Forderung $q \geq 1$ im Widerspruch zur Annahme der Kontraktivität von A. □

5.3 Verfahren der sukzessiven Approximation

Die zentrale Rolle des Fixpunktsatzes von Banach ergibt sich wesentlich aus dem konstruktiven Existenzbeweis für den Fixpunkt eines Operators. Das dem Beweis zugrunde liegende Iterationsverfahren oder *Verfahren der sukzessiven Approximation* liefert einen Algorithmus zur näherungsweisen Bestimmung des Fixpunktes. Im Fall einer endlich-dimensionalen Menge U ist dieser Algorithmus in der Regel auch elementar programmierbar. Zugleich erhält man Fehlerabschätzungen für die Güte der Approximation.

Satz 5.14. *Der Operator $A : U \to U \subset X$ mit Kontraktionszahl $q \in [0, 1)$ bilde die vollständige Teilmenge U eines normierten Raumes X in sich ab. Dann konvergiert das Verfahren der sukzessiven Approximation*

$$x_{n+1} := Ax_n, \quad n = 0, 1, 2, ...$$

für beliebige Startelemente $x_0 \in U$ gegen den eindeutig bestimmten Fixpunkt x des Operators A. Man hat ferner für beliebige Zahlen $n \in \mathbf{N}_0$ die a-priori Fehlerabschätzung

$$\|x - x_n\| \leq \frac{q^n}{1-q}\|x_1 - x_0\| \tag{5.10}$$

sowie die a-posteriori Fehlerabschätzung

$$\|x - x_n\| \leq \frac{q}{1-q}\|x_n - x_{n-1}\|. \tag{5.11}$$

Beweis. Die a-priori Aussage über den Fehler folgt aus der Formel (5.9) durch Grenzübergang $m \to \infty$. Die a-posteriori Fehleraussage folgt aus der a-priori Analyse durch Wahl von x_{n-1} als Startelement. □

Die a-priori Fehlerabschätzung kann benutzt werden, um vorab bei vorgegebener Fehlertoleranz ϵ eine obere Schranke für die Zahl der notwendigen Iterationsschritte zu bestimmen. Genauer sind zur Gewährleistung von

$$\|x - x_n\| \leq \epsilon$$

aufgrund der a-priori Abschätzung

$$n \geq \frac{\ln \tilde{\epsilon}}{\ln q}, \quad \tilde{\epsilon} := \frac{(1-q)\epsilon}{\|x_1 - x_0\|}$$

Iterationsschritte erforderlich. Mit der Kontraktionszahl q verringert sich natürlich die Zahl der notwendigen Schritte.

Die a-posteriori Abschätzung basiert auf aus dem Verfahren berechenbaren Größen. Sie zeigt die aktuelle Verbesserung der Näherung gegenüber dem letzten Schritt und ist daher für praktische Zwecke nützlich.

Wir vermerken noch, daß die (gegenüber der Kontraktionseigenschaft) abgeschwächte Forderung der *Nichtexpansitivität*

$$\|Ax - Ay\| < \|x - y\| \quad x, y \in U, x \neq y$$

im allgemeinen Fall nicht die Existenz eines Fixpunktes garantiert.

5.4 Spezialfall linearer Operatoren

Nachfolgend formulieren wir die entsprechenden Resultate aus dem Fixpunktsatz von Banach gesondert für den Fall linearer Operatoren, der speziell zur Analyse von Iterationsverfahren bei linearen Gleichungssystemen benötigt wird.

Satz 5.15. *Sei $B : X \to X$ beschränkter linearer Operator im Banach-Raum X mit $\|B\| < 1$. Dann ist der Operator $I - B$ mit dem Einheitsoperator I invertierbar, d.h. die Gleichung*

$$x - Bx = y \tag{5.12}$$

hat für jeden Wert $y \in X$ genau eine Lösung $x \in X$. Der inverse Operator $(I - B)^{-1}$ ist beschränkt mit

$$\|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \tag{5.13}$$

Das Verfahren der sukzessiven Approximation

$$x_{n+1} := Bx_n + y, \quad n = 0, 1, 2, ... \tag{5.14}$$

konvergiert bei beliebigem Startelement x_0 gegen die Lösung x.
Ferner hat man für beliebige Zahlen $n \in \mathbb{N}_0$ die a-priori bzw. a-posteriori Fehlerabschätzung

$$\|x - x_n\| \leq \frac{\|B\|^n}{1 - \|B\|}\|x_1 - x_0\| \tag{5.15}$$

bzw.

$$\|x - x_n\| \leq \frac{\|B\|}{1 - \|B\|}\|x_n - x_{n-1}\|. \tag{5.16}$$

Beweis. Für ein fixiertes (jedoch beliebiges) Element $y \in X$ definieren wir den Operator $A : X \to X$ durch

$$Ax := Bx + y, \qquad x \in X.$$

Wegen

$$\|Ax - A\tilde{x}\| = \|B(x - \tilde{x})\| \leq \|B\|\|x - \tilde{x}\|$$

hat dieser Operator die Kontraktionszahl $q = \|B\| < 1$. Satz 5.14 ergibt nun die angegebenen Aussagen über die Konvergenz des Iterationsverfahrens.

Die Methode der sukzessiven Approximation mit $x_0 = y$ ergibt

$$x_n = \sum_{k=0}^{n} B^k y$$

und damit

$$\|x_n\| \leq \sum_{k=0}^{n} \|B^k y\| \leq \|y\| \sum_{k=0}^{n} \|B\|^k \leq \frac{\|y\|}{1 - \|B\|}.$$

Wegen $x_n \to (I - B)^{-1}y,\ n \to \infty$ bedeutet das gerade

$$\|(I - B)^{-1}y\| \leq \frac{\|y\|}{1 - \|B\|}, \quad \forall y \in X. \qquad \square$$

Die hinreichenden Bedingungen aus Satz 5.15 werden im endlich-dimensionalen Fall sogar notwendig.

Satz 5.16. *Sei $B \in \mathbb{K}^{m\times m}$. Das Verfahren der sukzessiven Approximation*

$$x_{n+1} := Bx_n + y, \quad n = 0, 1, 2, ...$$

konvergiert für beliebige Startwerte $x_0 \in \mathbb{K}^m$ und jedes $y \in \mathbb{K}^m$ genau dann, wenn $\rho(B) < 1$.

Beweis. Zum Beweis der Hinlänglichkeit gelte $\rho(B) < 1$. Nach Satz 4.24 existiert dann eine Norm $\|\cdot\|$ auf $\mathbb{K}^m$ derart, daß $\|B\| \le \rho(B) + \epsilon$. Dann ist Satz 5.15 anwendbar, denn für hinreichend kleines ϵ wird $\|B\| < 1$. Die Folge (x_n) konvergiert somit.

Zum Beweis der Notwendigkeit sei angenommen, daß $\rho(B) \ge 1$. Folglich gibt es einen Eigenwert λ von B mit $\lambda \ge 1$ sowie den zugehörigen Eigenvektor x. Das Verfahren der sukzessiven Approximation mit $y = x$ und dem Startwert $x_0 = x$ liefert die divergente Folge

$$x_n = x \sum_{k=0}^{n} \lambda^k$$

im Widerspruch zur vorausgesetzten Konvergenz des Verfahrens. □

In Verallgemeinerung von Satz 5.26 über die Stabilität der Lösung linearer Gleichungssysteme zeigen wir jetzt einen allgemeineren *Störungssatz*.

Satz 5.17. *Seien X und Y Banach-Räume. Es existiere zu $A \in \mathcal{L}(X, Y)$ der inverse Operator mit $A^{-1} \in \mathcal{L}(Y, X)$. Ferner gelte für den gestörten Operator $A + A_\delta \in \mathcal{L}(X, Y)$ die Bedingung $\|A^{-1}\|\ \|A_\delta\| < 1$. Weiterhin seien x und $x + x_\delta$ Lösungen der Gleichungen $Ax = y$ bzw. $(A + A_\delta)(x + x_\delta) = y + y_\delta$. Dann gilt die Stabilitätsabschätzung*

$$\frac{\|x_\delta\|}{\|x\|} \le \frac{cond(A)}{1 - cond(A)\frac{\|A_\delta\|}{\|A\|}} \left\{ \frac{\|A_\delta\|}{\|A\|} + \frac{\|y_\delta\|}{\|y\|} \right\}. \tag{5.17}$$

wobei die Kondition des Operators A definiert wird durch $cond(A) := \|A\|\ \|A^{-1}\|$.

Beweis. Ausgangspunkt ist die Darstellung

$$A + A_\delta = A\left(I + A^{-1}A_\delta\right).$$

Für den Operator $B := A^{-1}A_\delta$ gilt nach Voraussetzung $\|B\| < 1$. Dann existiert nach Satz 5.15 der inverse Operator $(I - B)^{-1}$. Daraus folgern wir unter Beachtung der vorausgesetzten Existenz von $A^{-1} \in \mathcal{L}(Y, X)$ wiederum auf die Existenz des inversen Operators

$$(A + A_\delta)^{-1} = \left(I + A^{-1}A_\delta\right)^{-1} A^{-1}$$

und dessen Beschränktheit mit

$$\|\,(A + A_\delta)^{-1}\,\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\ \|A_\delta\|}. \tag{5.18}$$

Die Fehlergleichung lautet

$$(A + A_\delta)x_\delta = y_\delta - A_\delta x,$$

daraus folgt

$$x_\delta = (A + A_\delta)^{-1}\left(y_\delta - A_\delta x\right).$$

Dies liefert die Abschätzung

$$\|x_\delta\| \leq \| (A + A_\delta)^{-1} \| \{\|y_\delta\| + \|A_\delta\| \, \|x\|\} .$$

Unter Nutzung von (5.18) erhalten wir

$$\frac{\|x_\delta\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\| \, \|A_\delta\|} \left\{ \frac{\|A_\delta\|}{\|A\|} + \frac{\|y_\delta\|}{\|A\| \, \|x\|} \right\} .$$

Dies ergibt die gesuchte Aussage (5.17) bei Beachtung von $\|A\|\|x\| \geq \|y\|$. □

Der Satz 5.17 zeigt, wie sich Störungen der Daten A und y des Problems $Ax = y$ auf die Lösung auswirken. Insbesondere verstärken sich Datenfehler in Abhängigkeit von der Kondition des Operators.

5.5 Ausblick (Exkurs)

Wir schließen unsere Betrachtungen zu funktionalanalytischen Grundlagen mit einigen weiterreichenden Bemerkungen ab:

Zunächst benutzen wir das Instrumentarium des Fixpunktsatzes von Banach zur Untersuchung der grundlegenden Iterationsverfahren für lineare bzw. für nichtlineare Gleichungssysteme (vgl. Kapitel 6 bzw. 7/8).

Es sei noch auf folgenden Punkt hingewiesen: Das wesentliche (in der Regel nicht triviale) technische Problem bei Anwendung des Fixpunktsatzes auf (nichtlineare) Aufgaben besteht in der Konstruktion einer geeigneten Operatorgleichung in Fixpunktform

$$\text{Finde } \; x \in X : \quad x = A(x).$$

Für dieses Problem muß man die Eigenschaften der Selbstabbildung $A : U \to U$ und der Kontraktivität auf einer geeigneten Umgebung U eines potentielen Fixpunktes nachweisen.

Eine grundlegende Idee bei der Behandlung nichtlinearer Operatorgleichungen in (in der Regel unendlich-dimensionalen) normierten Räumen X ist dabei die Approximation des Problems in Unterräumen $X_n \subset X$ mit endlicher Dimension $n \in \mathbb{N}$. Oft kann man für die Lösung dieser Probleme

$$\text{Finde } \; x_n \in X_n : \quad x_n = A(x_n)$$

eine sogenannte a-priori Abschätzung beweisen

$$\exists C \neq C(n) : \quad \|x_n\|_X \leq C,$$

die gleichmäßig bezüglich $n \in \mathbb{N}$ gilt. Ist der Raum X sogar Hilbert-Raum, d.h. ein vollständiger Prä-Hilbert-Raum, kann man auf die Existenz einer in X konvergenten Teilfolge $x_n \to \tilde{x}$ schließen. Oft gelingt dann der Nachweis, daß $\tilde{x}$ Lösung des Ausgangsproblems ist. Neben dem Existenzbeweis hat man dann mit der Methode der sukzessiven Approximation auch ein konstruktives Verfahren zur (numerischen) Lösung der Probleme in X_n an der Hand.

Die hier eingeführten Grundlagen der Funktionalanalysis erweisen sich nicht nur für die Vorlesungen *Numerische Mathematik I/ II* als grundlegend. Sie dienen auch (natürlich in erweiterter Form) als Handwerkszeug für weitergehende Vorlesungen der Numerischen und Angewandten Mathematik, etwa zur Approximationstheorie, Optimierung oder Numerischen Linearen Algebra sowie zur Behandlung von Integral- und Differentialgleichungen. Eine systematische Vertiefung wird in Vorlesungen zur linearen bzw. nichtlinearen Funktionalanalysis geboten.

Kapitel 6

Elementare Iterationsverfahren für lineare Gleichungssysteme

Gegenstand des vorliegenden Abschnitts sind grundlegende *Iterationsverfahren* für lineare Gleichungssysteme. Derartige Verfahren finden vor allem bei Systemen großer Dimension Anwendung. Der Nachteil *direkter Auflösungsverfahren* (vgl. Kapitel 2) war, daß der erforderliche Speicherbedarf sowie die Zahl der benötigten Rechenoperationen sehr stark mit der Dimension steigen. Dies ist für die hier zu behandelnden Verfahren nicht der Fall. Zu ihrer Begründung werden wir die elementaren funktionalanalytischen Kenntnisse aus den Kapiteln 4 und 5 heranziehen.

6.1 Fixpunktform linearer Gleichungssysteme

Für $A \in \mathbb{K}^{m\times m}$ und $y \in \mathbb{K}^m$ schreiben wir das lineare Gleichungssystem

$$\text{Finde } \; x \in \mathbb{K}^m : \qquad Ax = y \tag{6.1}$$

in der äquivalenten *Fixpunktform*

$$x = x + M^{-1}(y - Ax) \tag{6.2}$$

mit einer zu wählenden regulären Matrix $M \in \mathbb{K}^{m\times m}$. Zur Fixpunktberechnung betrachten wir das *Verfahren der sukzessiven Approximation*

$$x_{n+1} = x_n + M^{-1}(y - Ax_n), \qquad n \in \mathbb{N}. \tag{6.3}$$

Dies entspricht einer zu lösenden Folge linearer Gleichungssysteme

$$Mw_{n+1} = r_n := y - Ax_n; \qquad x_{n+1} := x_n + w_{n+1}, \qquad n \in \mathbb{N}. \tag{6.4}$$

Man wird nun M so wählen, daß diese Systeme effizient lösbar sind. Satz 5.16 liefert für die Konvergenz von (6.3) das notwendige und hinreichende Kriterium

$$\rho(I - M^{-1}A) < 1.$$

Bei der Zerlegung $A = M - N$ erhält man die äquivalente Forderung $\rho(M^{-1}N) < 1$.

In diesem Kapitel gehen wir aus von der Darstellung

$$A = A_D + A_L + A_U \tag{6.5}$$

mit der Diagonalmatrix

$$A_D = \mathrm{diag}(a_{11}, ..., a_{mm}), \tag{6.6}$$

der unteren Dreiecksmatrix

$$A_L = \begin{pmatrix} 0 & & & & & \\ a_{21} & 0 & & & & \\ a_{31} & a_{32} & 0 & & & \\ \vdots & \vdots & & & \ddots & \\ a_{m1} & a_{m2} & \cdot & \cdot & a_{m,m-1} & 0 \end{pmatrix} \tag{6.7}$$

sowie der oberen Dreiecksmatrix

$$A_U = \begin{pmatrix} 0 & a_{12} & \cdot & \cdot & \cdot & a_{1m} \\ & 0 & a_{23} & \cdot & \cdot & a_{2m} \\ & & \ddots & \cdot & \cdot & \cdot \\ & & & & 0 & a_{m-1,m} \\ & & & & & 0 \end{pmatrix}. \tag{6.8}$$

Für unsere Betrachtungen in diesem Kapitel setzen wir voraus, daß (evt. nach Vertauschung von Zeilen und Spalten der Matrix A) die inverse Matrix $(A_D)^{-1}$ existiert. Dies ist äquivalent dazu, daß kein Element der Hauptdiagonale von A verschwindet.

6.2 Gesamtschritt- bzw. Jacobi-Verfahren

Die einfachste Wahl $M = A_D$ führt auf das *Jacobi-* bzw. *Gesamtschritt-Verfahren* (GSV). Die entsprechende Fixpunktgleichung ist äquivalent zu

$$x = -A_D^{-1}(A_L + A_U)x + A_D^{-1}y,$$

das Verfahren der sukzessiven Approximation ergibt

$$x_{n+1} = -A_D^{-1}(A_L + A_U)x_n + A_D^{-1}y, \quad n = 0, 1, 2, ... \tag{6.9}$$

mit beliebigem Startvektor x_0. Diese Iterationsvorschrift des GSV lautet komponentenweise

$$x_{n+1,i} = -\sum_{\substack{k=1 \\ k \neq i}}^{m} \frac{a_{ik}}{a_{ii}} x_{n,k} + \frac{y_i}{a_{ii}}, \quad i = 1, ..., m. \tag{6.10}$$

Auskunft über hinreichende Konvergenzbedingungen des Verfahrens gibt der

Satz 6.1. *Die Matrix $A = (a_{ik}) \in \mathbb{K}^{m \times m}$ genüge einer der folgenden Bedingungen.*
(starkes) Zeilensummenkriterium:

$$q_\infty := \max_{i=1,...,m} \sum_{\substack{k=1 \\ k \neq i}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| < 1 \tag{6.11}$$

(starkes) Spaltensummenkriterium:

$$q_1 := \max_{k=1,...,m} \sum_{\substack{i=1 \\ i \neq k}}^{m} \left| \frac{a_{ik}}{a_{kk}} \right| < 1 \tag{6.12}$$

Quadratsummenkriterium:

$$q_2 := \left(\sum_{\substack{i,k=1 \\ i \neq k}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right|^2 \right)^{1/2} < 1. \tag{6.13}$$

Dann konvergiert das GSV bezüglich jeder Norm auf $\mathbb{K}^m$, für jede rechte Seite $y \in \mathbb{K}^m$ und bei beliebigem Startvektor x_0 gegen die eindeutig bestimmte Lösung des linearen Gleichungssystems $Ax = y$. Für $\mu = 1, 2, \infty$ mit q_μ gelten bei beliebigem $n \in \mathbb{N}$ die a-priori bzw. a-posteriori Fehlerabschätzung

$$\|x_n - x\|_\mu \leq \frac{q_\mu^n}{1 - q_\mu} \|x_1 - x_0\|_\mu$$

bzw.

$$\|x_n - x\|_\mu \leq \frac{q_\mu}{1 - q_\mu} \|x_n - x_{n-1}\|_\mu.$$

Beweis. Die Matrix $B = -A_D^{-1}(A_L + A_U)$ hat verschwindende Elemente auf der Hauptdiagonale sowie die Außerdiagonalelemente $-a_{ik}/a_{ii}$. Dann folgern wir aus Satz 4.18

$$\| - A_D^{-1}(A_L + A_U)\|_\mu \leq q_\mu, \qquad \mu \in \{1, 2, \infty\}$$

Die Behauptung folgt dann aus dem Satz 5.15. □

Beispiel 6.2. Wir betrachten das Beispiel 1.1 (elektrisches Netzwerk). Für die entstehende schwachbesetzte Matrix

$$A = \begin{pmatrix} a & c & 0 & \cdots & 0 & c \\ c & a & c & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & c & a & c \\ c & 0 & & 0 & c & a \end{pmatrix}$$

gilt mit $a = 3/R, c = -1/R$

$$q_\infty = q_1 = \frac{2}{3} < 1, \qquad q_2 = \frac{m\sqrt{2}}{3}$$

und damit konvergiert das GSV nach Satz 6.1. □

Beispiel 6.3. Für die bei Beispiel 1.3 (Zweipunkt-Randwertproblem) entstehende Tridiagonalmatrix

$$A = \text{tridiag}(-1,\ 2,\ -1) \ \in \ \mathbb{R}^{m \times m}$$

gilt

$$q_\infty = q_1 = 1, \qquad q_2 = \frac{m}{\sqrt{2}}.$$

Die Konvergenz des GSV kann also offenbar mittels Satz 6.1 nicht entschieden werden. Eine analoge Situation zeigt sich bei Betrachtung des Beispiels 1.4 (Randwertproblem für die Poissonsche Differentialgleichung). □

Die Beispiele zeigen, daß die hinreichenden Konvergenzkriterien aus Satz 6.1 nicht äquivalent sind.

6.3 Einzelschrit- bzw. Gauß-Seidel-Verfahren

Durch die Wahl $M = A_D + A_L$ erhalten wir das *Gauß-Seidel-* bzw. *Einzelschritt-Verfahren* (ESV). Die zugehörige Fixpunktgleichung ist äquivalent zu

$$x = -(A_D + A_L)^{-1} A_U x + (A_D + A_L)^{-1} y,$$

das Verfahren der sukzessiven Approximation ergibt

$$x_{n+1} = -(A_D + A_L)^{-1} A_U x_n + (A_D + A_L)^{-1} y, \quad n = 0, 1, 2, ... \tag{6.14}$$

mit beliebigem Startvektor x_0. Rechnerisch geht man aus von der äquivalenten Formulierung

$$(A_D + A_L) x_{n+1} = -A_U x_n + y, \quad n = 0, 1, 2, ...$$

Durch Auflösung dieses Systems mittels Vorwärtseinsetzen erhalten wir die komponentenweise Iterationsvorschrift

$$x_{n+1,i} = -\sum_{k=1}^{i-1} \frac{a_{ik}}{a_{ii}} x_{n+1,k} - \sum_{k=i+1}^{m} \frac{a_{ik}}{a_{ii}} x_{n,k} + \frac{y_i}{a_{ii}}, \quad i = 1, ..., m. \tag{6.15}$$

Während beim GSV alle Komponenten des neuen Iterationsvektors x_{n+1} aus den Komponenten der vorhergehenden Näherung x_n ermittelt werden, berücksichtigt man im ESV bei Berechnung der Komponenten $x_{n+1,i}$ bereits die neuen Werte $x_{n+1,j}$, $j < i$. Hierdurch spart man durch Überschreiben Speicherplatz.

Hinreichende Konvergenzaussagen für das ESV gibt

Satz 6.4. *Die Matrix $A = (a_{ik}) \in \mathbb{K}^{m \times m}$ genüge dem* Kriterium nach Sassenfeld:

$$p := \max_{i=1,...,m} p_i < 1, \tag{6.16}$$

mit den rekursiv zu berechnenden Werten

$$p_1 := \sum_{k=2}^{m} \left| \frac{a_{1k}}{a_{11}} \right|, \quad p_i := \sum_{k=1}^{i-1} \left| \frac{a_{ik}}{a_{ii}} \right| p_k + \sum_{k=i+1}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| \quad i = 2, ..., m. \tag{6.17}$$

Dann konvergiert das ESV bezüglich jeder Norm auf $\mathbb{K}^m$, für jede rechte Seite $y \in \mathbb{K}^m$ und bei beliebigem Startvektor x_0 gegen die eindeutig bestimmte Lösung des linearen Gleichungssystems $Ax = y$. Es gelten bei beliebigem $n \in \mathbb{N}$ die a-priori bzw. a-posteriori Fehlerabschätzung

$$\|x_n - x\|_\infty \leq \frac{p^n}{1-p} \|x_1 - x_0\|_\infty$$

bzw.

$$\|x_n - x\|_\infty \leq \frac{p}{1-p} \|x_n - x_{n-1}\|_\infty.$$

Beweis. Wir betrachten ein System der Form

$$(A_D + A_L) x = -A_U z, \quad \|z\|_\infty = 1,$$

um die Zeilensummen-Norm der Iterationsmatrix $(A_D + A_L)^{-1} A_U$ abschätzen zu können. Vorwärtselimination ergibt

$$x_i = -\sum_{k=1}^{i-1} \frac{a_{ik}}{a_{ii}} x_k - \sum_{k=i+1}^{m} \frac{a_{ik}}{a_{ii}} z_k, \quad i = 1, ..., m.$$

Unter Beachtung der Definition von p_i ermitteln wir damit durch vollständige Induktion, daß $|x_i| \leq p_i, i = 1, ..., m$. Damit folgt $\|x\|_\infty \leq p$ und somit

$$\|(A_D + A_L)^{-1} A_U\|_\infty \leq p.$$

Satz 5.15. ergibt dann wieder die Behauptung. □

Ein Vergleich von Sassenfeld- und starkem Zeilensummenkriterium führt auf die

Folgerung 6.5. *Das starke Zeilensummenkriterium für die Matrix A ist ebenfalls hinreichend für die Konvergenz des ESV.*

Wir kehren noch einmal zurück zum Beispiel 6.3 (bzw. 1.3), das die Diskretierung einer Zweipunkt-Randwertaufgabe behandelt.

Beispiel 6.6. *Die im Beispiel 6.3 entstehende Tridiagonalmatrix A genügt dem Sassenfeld-Kriterium, jedoch nicht dem starken Zeilensummenkriterium.*

Beweis. Aus Beispiel 6.3 wissen wir, daß $q_\infty = 1$ ist. Durch Induktion ermittelt man nun über die Beziehung

$$p_{i+1} = \frac{1}{2} p_i + \frac{1}{2}, \quad i = 2, ..., m-2,$$

daß

$$p_i = 1 - \frac{1}{2^i}, \quad i = 1, ..., m-1; \qquad p_m = \frac{1}{2} - \frac{1}{2^m}.$$

Damit gilt aber

$$p = 1 - \frac{1}{2^{m-1}} < 1.$$

Die Kontraktionszahl p strebt für $m \to \infty$ gegen 1. Das bedeutet bei Systemen mit sehr großer Dimension m sehr stark absinkende Konvergenzgeschwindigkeit. Dieses Beispiel ist typisch für die Diskretisierung von Randwertaufgaben von Differentialgleichungen. Für kleiner werdende Schrittweite h der Diskretisierung verschlechtert sich die Kondition der Matrix stark. □

6.4 Zerlegbare Matrizen (Exkurs)

Wir schwächen nun das „starke Zeilensummenkriterium" ab. Dazu benötigen wir

Definition 6.7. *Eine Matrix $A = (a_{ik}) \in \mathbb{K}^{m \times m}$ heißt* zerlegbar (oder reduzibel), *falls nichtleere Indexmengen N_1, N_2 existieren mit $N_1 \cap N_2 = \emptyset$, $N_1 \cup N_2 = \{1, ..., m\}$,* und

$$a_{ik} = 0, \quad \forall i \in N_1, \ \forall k \in N_2.$$

Anderenfalls heißt A unzerlegbar (oder irreduzibel).

Durch Umordnung von Zeilen und Spalten einer zerlegbaren Matrix erreicht man stets, daß die folgende entkoppelbare Struktur mit quadratischen Matrizen $A_{ii}, \ i = 1, 2$ entsteht:

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}.$$

Beispiel 6.8. *Tridiagonalmatrizen, bei denen die Elemente der Hauptdiagonale und der beiden Nebendiagonalen sämtlich nicht verschwinden, sind unzerlegbar. Insbesondere betrifft das die Matrix im Beispiel 6.6 (bzw. 1.3).*

Beweis. A sei zerlegbar. Für $i \in N_1$ folgt wegen $a_{i,i\pm1} \neq 0$ auch $i \pm 1 \in N_1$. Daraus folgt induktiv, daß $N_1 = \{1, ..., m\}$ und ein Widerspruch zur Annahme über A. □

Die angestrebte Begriffsabschwächung gibt der

Satz 6.9. *Die Matrix $A = (a_{ik})$ sei unzerlegbar und genüge dem* schwachen Zeilensummenkriterium:

$$\sum_{\substack{k=1 \\ k \neq i}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| \leq 1, \quad i = 1, ..., m \tag{6.18}$$

mit Ungleichheit in mindestens einer der Zeilen der Matrix. Dann konvergiert das GSV bezüglich jeder Norm auf $\mathbb{K}^m$ für jede rechte Seite $y \in \mathbb{K}^m$ und jeden Startvektor x_0 gegen die eindeutig bestimmte Lösung von $Ax = y$.

Beweis. Nach Voraussetzung und Satz 4.18 erfüllt die Iterationsmatrix $B = -A_D^{-1}(A_L + A_U)$ des GSV die Bedingung $\|B\|_\infty \leq 1$. Nach Satz 4.24 folgt für den Spektralradius $\rho(B) \leq 1$.

Wir treffen die Annahme, daß λ Eigenwert von B mit $|\lambda| = 1$ sei. O.B.d.A. gelte für den zugehörenden Eigenvektor $\|x\|_\infty = 1$. Mit $\lambda x = Bx$ ergibt sich

$$|\lambda||x_i| \leq \sum_{\substack{k=1 \\ k \neq i}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| |x_k| \leq \sum_{\substack{k=1 \\ k \neq i}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| \leq 1, \quad i = 1, ..., m. \tag{6.19}$$

Die Menge $N_1 := \{i : |x_i| = 1\}$ ist wegen $\|x\|_\infty = 1$ nichtleer. Für $i \in N_1$ gilt wegen $|\lambda||x_i| = 1$ in (6.19) sogar

$$\sum_{\substack{k=1 \\ k \neq i}}^{m} \left| \frac{a_{ik}}{a_{ii}} \right| = 1, \quad i \in N_1.$$

Das schwache Zeilensummenkriterium für A impliziert $N_2 := \{i : i \notin N_1\} \neq \emptyset$. Aus der Unzerlegbarkeit von A ergibt sich die Existenz von $i_1 \in N_1$ und $i_2 \in N_2$ mit $a_{i_1 i_2} \neq 0$. Wegen $|a_{i_1 i_2}||x_{i_2}| < |a_{i_1 i_2}|$ schließen wir auf einen Widerspruch wegen

$$1 = |x_{i_1}| = |\lambda||x_{i_1}| = \sum_{\substack{k=1 \\ k \neq i_1}}^{m} \left| \frac{a_{i_1 k}}{a_{i_1 i_1}} \right| |x_k| < \sum_{\substack{k=1 \\ k \neq i_1}}^{m} \left| \frac{a_{i_1 k}}{a_{i_1 i_1}} \right| \leq 1.$$

Daher ist $\rho(B) < 1$ und die Behauptung folgt nach Anwendung von Satz 4.16. □

Beispiel 6.10. *Die Tridiagonalmatrix aus Beispiel 6.3 (bzw. 1.3) ist unzerlegbar und erfüllt das schwache Zeilensummenkriterium.*

6.5 Relaxations-Verfahren

Ziel der sogenannten *Relaxationsvarianten* des GSV bzw.ESV ist, durch geeignete Einführung eines Parameters in der Verfahrensvorschrift die Konvergenz des Verfahrens zu beschleunigen. Dazu muß der Spektralradius des Verfahrens verkleinert werden.

Die entsprechende Modifikation des Gesamtschritt-Verfahrens mit $M = A_D$ zeigt

Definition 6.11. *Das Iterationsverfahren*

$$x_{n+1} = x_n + \omega A_D^{-1}(y - Ax_n), \quad n = 0, 1, 2, ... \tag{6.20}$$

bzw. in Komponentenschreibweise

$$x_{n+1,i} = x_{n,i} + \frac{\omega}{a_{ii}} \left\{ y_i - \sum_{k=1}^{m} a_{ik} x_{n,k} \right\}, \quad i = 1, 2, ..., m \tag{6.21}$$

heißt Gesamtschritt-Relaxationsverfahren.

Auskunft über die geeignete Wahl des Relaxationsparameters ω und die Konvergenz gibt

Satz 6.12. *Die zum GSV gehörende Matrix*

$$B := -A_D^{-1}(A_L + A_U)$$

habe nur reelle Eigenwerte und einen Spektralradius $\rho(B) < 1$. Dann wird der Spektralradius der Relaxationsmatrix

$$I - \omega A_D^{-1} A = (1 - \omega) I - \omega A_D^{-1}(A_L + A_U)$$

minimal bei

$$\omega_0 = \frac{2}{2 - \lambda_{min} - \lambda_{max}}. \tag{6.22}$$

Dabei sind λ_{min} bzw. λ_{max} der kleinste bzw. größte Eigenwert der Matrix B. Speziell konvergiert bei $\lambda_{min} \neq -\lambda_{max}$ das Gesamtschrit-Relaxationsverfahren schneller als das GSV.

Beweis. Mit $\omega > 0$ ist die Gleichung $Bu = \lambda u$ äquivalent zu

$$[(1 - \omega) I + \omega B]\, u = [(1 - \omega) + \omega\lambda]\, u.$$

Daher lassen sich den Eigenwerten λ von B die Eigenwerte $(1-\omega)+\omega\lambda$ der Matrix $(1-\omega)I+\omega B$ zuordnen. Speziell sind $(1-\omega)+\omega\lambda_{min}$ bzw. $(1-\omega)+\omega\lambda_{max}$ der kleinste bzw. größte Eigenwert von $(1 - \omega)I + \omega B$. Der Spektralradius wird minimal, falls kleinster und größter Eigenwert gleichen Betrag haben, d.h. $(1 - \omega_0) + \omega_0\lambda_{min} = -(1 - \omega_0) - \omega_0\lambda_{max}$. Hieraus folgt (6.22). □

Wir besprechen jetzt die entsprechende Relaxationsvariante für das ESV. Dazu schreiben wir die Verfahrensvorschrift $(A_D + A_L)x_{n+1} = -A_U x_n + y$ um in

$$x_{n+1} = x_n + A_D^{-1}\left[y - A_L x_{n+1} - (A_D + A_U)x_n\right].$$

Definition 6.13. *Das Iterationsverfahren*

$$x_{n+1} = x_n + \omega A_D^{-1}\left[y - A_L x_{n+1} - (A_D + A_U)x_n\right], \quad n = 0, 1, 2, ... \tag{6.23}$$

bzw. in Komponentenschreibweise

$$x_{n+1,i} = x_{n,i} + \frac{\omega}{a_{ii}} \left\{ y_i - \sum_{k=1}^{i-1} a_{ik} x_{n+1,k} - \sum_{k=i}^{m} a_{ik} x_{n,k} \right\}, \quad i = 1, 2, ..., m \tag{6.24}$$

heißt Einzelschritt-Relaxationsverfahren *oder* SOR-Verfahren („successive overrelaxation").

Aus der Darstellung

$$(A_D + \omega A_L)x_{n+1} = \omega y + \{(1 - \omega)A_D - \omega A_U\}x_n$$

findet man die Gestalt der Iterationsmatrix zu

$$B(\omega) = (A_D + \omega A_L)^{-1}\{(1-\omega)A_D - \omega A_U\}.$$

Offenbar hängt die Iterationsmatrix nicht linear ab vom Parameter ω, im Unterschied zum Gesamtschritt-Relaxationsverfahren. Daher bereitet die Wahl von ω auch größere Probleme.

Satz 6.14. *Das SOR-Verfahren konvergiert höchstens im Parameterintervall $0 < \omega < 2$. Im Falle einer hermiteschen und positiv definiten Matrix A konvergiert das SOR-Verfahren für alle Werte aus diesem Intervall.*

Beweis. Für die Eigenwerte $\mu_1, ..., \mu_m$ der Iterationsmatrix $B(\omega)$ gilt (bei Beachtung ihrer algebraischen Vielfachheit)

$$\prod_{i=1}^{m} \mu_i = \det B(\omega).$$

Da nun $A_D + \omega A_L$ und $(1-\omega)A_D - \omega A_U$ Dreiecksmatrizen sind, folgt mit den Rechenregeln für Determinanten

$$\begin{aligned} \prod_{i=1}^{m} \mu_i &= \det B(\omega) = \det (A_D + \omega A_L)^{-1} \det [(1-\omega)A_D - \omega A_U] \\ &= (\det A_D)^{-1}(1-\omega)^m \det A_D = (1-\omega)^m, \end{aligned}$$

damit

$$\rho(B(\omega)) = \max_{i=1,...,m} |\mu_i| \geq |1-\omega|.$$

Satz 5.16 liefert sogar die notwendige Bedingung $\rho(B(\omega)) < 1$, d.h. $0 < \omega < 2$.

Für den Spezialfall einer hermiteschen, positiv definiten Matrix A sei μ Eigenwert von $B(\omega)$ mit Eigenvektor x, also

$$[(1-\omega)A_D - \omega A_U]\, x = \mu(A_D + \omega A_L)x.$$

Die aus $A = A_D + A_L + A_U$ folgenden Identitäten

$$(2-\omega)A_D - \omega A - \omega(A_U - A_L) = 2(1-\omega)A_D - 2\omega A_U$$

und

$$(2-\omega)A_D + \omega A - \omega(A_U - A_L) = 2A_D + 2\omega A_L$$

zeigen

$$[(2-\omega)A - \omega A - \omega(A_U - A_L)]\, x = \mu\, [(2-\omega)A_D + \omega A - \omega(A_U - A_L)]\, x.$$

Über Skalarproduktbildung mit x ergibt sich daraus

$$\mu = \frac{(2-\omega)d - \omega a + i\omega s}{(2-\omega)d + \omega a + i\omega s}$$

mit den Abkürzungen

$$a := (Ax, x), \quad d := (A_D x, x), \quad s := i(A_U x - A_L x, x).$$

Aus der positiven Definitheit von A folgt $a, d > 0$. Ferner ist $s \in \mathbb{R}$, da $A = A^*$. Schließlich ist wegen $0 < \omega < 2$

$$|(2-\omega)d - \omega a| < |(2-\omega)d + \omega a|,$$

daher $|\mu| < 1$. Satz 5.16 ergibt die Konvergenz des Verfahrens. □

Nachfolgend wird die Parameterwahl im Fall von *konsistent geordneten Matrizen* behandelt. Dieser Fall schließt den wichtigen Spezialfall regulärer Tridiagonalmatrizen ein.

Definition 6.15. *Eine Matrix $A = A_D + A_L + A_U$ heißt* konsistent geordnet, *falls die Eigenwerte der Matrix*

$$C(\alpha) := -\alpha A_D^{-1} A_L - \frac{1}{\alpha} A_D^{-1} A_U, \;\; \alpha \neq 0 \tag{6.25}$$

unabhängig vom Parameterwert α sind.

Satz 6.16. *Tridiagonalmatrizen mit regulärer Diagonalmatrix sind konsistent geordnet.*

Beweis. Mit der Diagonalmatrix

$$S(\alpha) := \mathrm{diag}(1, \alpha, \alpha^2, ..., \alpha^{m-1})$$

können wir für Tridiagonalmatrizen $A = A_D + A_L + A_U$ schreiben

$$S(\alpha) C(1) S(\alpha)^{-1} = C(\alpha).$$

Daraus folgt die Ähnlichkeit der Matrizen $C(\alpha)$ und damit die Gleichheit der Eigenwerte. □

Das gesuchte Resultat für das SOR-Verfahren im Fall konsistent geordneter Matrizen gibt der

Satz 6.17. *Die Matrix A sei konsistent geordnet, und die Eigenwerte der GSV-Iterationsmatrix $-A_D^{-1}(A_L + A_U)$ seien sämtlich reell. Für deren Spektralradius gelte $\Lambda := \rho(-A_D^{-1}(A_L + A_U)) < 1$. Dann konvergiert das SOR-Verfahren für alle Parameterwerte $0 < \omega < 2$. Der Spektralradius der Iterationsmatrix $B(\omega) = (A_D + \omega A_L)^{-1}\{(1-\omega)A_D - \omega A_U\}$ wird minimiert im Fall*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \Lambda^2}} > 1, \tag{6.26}$$

und es gilt

$$\rho(B(\omega_0)) = \frac{1 - \sqrt{1 - \Lambda^2}}{1 + \sqrt{1 - \Lambda^2}}.$$

Beweis. Wir notieren die Identität

$$\begin{aligned} (I + \omega A_D^{-1} A_L)\,[\mu I - B(\omega)] &= \mu(I + \omega A_D^{-1} A_L) - A_x D^{-1}\,[(1-\omega)A_D - \omega A_U] \\ &= (\mu + \omega - 1) I + \sqrt{\mu}\omega \left(\sqrt{\mu} A_D^{-1} A_L + \frac{1}{\sqrt{\mu}} A_D^{-1} A_U \right). \end{aligned}$$

In Verbindung mit der Regularität von $I + \omega A_D^{-1} A_L$ folgern wir hieraus, daß $\mu \neq 0$ genau dann Eigenwert von $B(\omega)$ ist, wenn

$$\lambda = \frac{\mu + \omega - 1}{\sqrt{\mu}\omega} \tag{6.27}$$

Eigenwert von

$$-\sqrt{\mu} A_D^{-1} A_L - \frac{1}{\sqrt{\mu}} A_D^{-1} A_U$$

ist. Da A konsistent geordnet sein soll, ist damit $\mu \neq 0$ Eigenwert von $B(\omega)$ genau dann, wenn λ Eigenwert der Matrix $-A_D^{-1}(A_L + A_U)$ ist.

Die quadratische Gleichung

$$\mu + \omega - 1 = \sqrt{\mu}\omega\lambda$$

hat die Lösungen

$$\mu = \left(\frac{\omega\lambda}{2} \pm \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega} \right)^2 .$$

Durch Wahl von $\alpha = -1$ in der obigen Definition erkennen wir, daß für konsistent geordnete Matrizen mit λ stets auch $-\lambda$ Eigenwert von $-A_D^{-1}(A_L + A_U)$ ist. Zur Berechnung von $\rho[B(\omega)]$ untersuchen wir

$$\mu = \left(\frac{\omega|\lambda|}{2} + \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega} \right)^2 .$$

Von den Nullstellen von

$$\omega^2\lambda^2 - 4\omega + 4 = 0, \quad \text{d.h. } \omega = \frac{2}{1 \pm \sqrt{1-\lambda^2}},$$

liegt nur eine im Intervall $(0, 2)$, und zwar

$$\omega_0(\lambda) = \frac{2}{1 + \sqrt{1-\lambda^2}} > 1.$$

Damit gilt

$$|\mu(\omega)| = \left(\frac{\omega|\lambda|}{2} + \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega} \right)^2, \quad 0 < \omega < \omega_0(\lambda). \tag{6.28}$$

Für $\omega_0 < \omega < 2$ werden die Eigenwerte komplex mit

$$|\mu(\omega)| = \omega - 1, \quad \omega_0(\lambda) < \omega < 2. \tag{6.29}$$

Aus den Gleichungen (6.28) und (6.29) sieht man, daß $|\mu(\omega)|$ monoton nichtfallend in $|\lambda|$ ist, also

$$\rho\left[B(\omega)\right] = \begin{cases} \left(\frac{\omega\Lambda}{2} + \sqrt{\frac{\omega^2\Lambda^2}{4} + 1 - \omega} \right)^2, & 0 < \omega < \omega_0(\Lambda), \\ \omega - 1, & \omega_0(\Lambda) < \omega < 2. \end{cases} \tag{6.30}$$

Wir untersuchen die Funktion

$$f(\omega) := \frac{\omega\Lambda}{2} + \sqrt{\frac{\omega^2\Lambda^2}{4} + 1 - \omega}.$$

Neben $f(0) = 1$ gilt

$$f'(\omega) = \frac{\Lambda}{2} + \frac{\Lambda^2\omega - 2}{2\sqrt{\omega^2\Lambda^2 + 4 - 4\omega}} < 0.$$

Die letzte Ungleichung ergibt sich wegen

$$\Lambda^2(4 - 4\omega + \omega^2\Lambda^2) < 4 - 4\Lambda^2\omega + \omega^2\Lambda^4 = (2 - \omega\Lambda^2)^2.$$

Damit ist der durch Gleichung (6.30) beschriebene Spektralradius für $0 < \omega < \omega_0$ monoton fallend und für $\omega_0 < \omega < 2$ monoton wachsend. An den Intervallendpunkten $\omega = 0$ und $\omega = 2$ erhält man jeweils den Wert 1 als Spektralradius, also $\rho\left[B(\omega)\right] < 1$ für alle Werte $0 < \omega < 2$. Der minimale Spektralradius ergibt sich gerade für $\omega = \omega_0$ zu $\rho\left[B(\omega)\right] = \omega_0 - 1$. □

Auf der Basis von Satz 6.17 können wir noch Aufwandsabschätzungen für Relaxationsverfahren angeben. Im Spezialfall ohne Relaxation ($\omega = 1$) haben wir folgendes Resultat.

Beispiel 6.18. *Unter den Voraussetzungen von Satz 6.17 konvergiert das Einzelschritt-Verfahren etwa doppelt so schnell wie das Gesamtschritt-Verfahren (beide Verfahren mit $\omega = 1$).*

Beweis. Wegen Gleichung (6.27) gilt $\mu = \lambda^2$ für $\omega = 1$, also für die Spektralradien

$$\rho(B_{ESV}) = [\rho(B_{GSV})]^2 .$$

Die im Anschluß von Satz 5.14 gegebene a-priori Abschätzung zeigt, daß die Anzahl der für eine vorgegebene Toleranz erforderlichen Iterationsschritte indirekt proportional zum Logarithmus des Spektralradius ist. Damit folgt

$$\frac{\text{Iterationszahl(ESV)}}{\text{Iterationszahl(GSV)}} \sim \frac{ln\rho(B_{GSV})}{ln\rho(B_{ESV})} \sim \frac{1}{2}. \qquad \square$$

Das nachfolgende Beispiel zeigt den deutlichen Einfluß eines optimal gewählten Relaxationsparameters auf die erforderliche Iterationszahl im Fall spezieller Tridiagonalmatrizen.

Beispiel 6.19. *Im Falle der Tridiagonalmatrix A aus Beispiel 6.3 (bzw. 1.3) gilt bei optimalem Wert des SOR-Relaxationsparameters für den Verfahrensaufwand*

$$\frac{\text{Iterationszahl}(SOR^{opt})}{\text{Iterationszahl}(GSV)} \approx \frac{\pi}{4(m+1)}.$$

Beweis. Einsetzen in die Eigenwertgleichung $Bx = \lambda x$ zeigt, daß die GSV-Iterationsmatrix

$$B := -A_D^{-1}(A_L + A_U) = \text{tridiag}(\frac{1}{2},\ 0,\ \frac{1}{2})$$

die folgenden Eigenwerte λ_j und Eigenvektoren $x_j = (x_{j,1}, \ldots, x_{j,m})^T$ hat

$$\lambda_j = \cos\frac{\pi j}{m+1}; \qquad x_{j,k} = \sin\frac{\pi jk}{m+1},\ \ k = 1, ..., m, \quad j = 1, ..., m.$$

Dazu verwendet man das Additionstheorem

$$\frac{1}{2}\sin\frac{\pi j(k-1)}{m+1} + \frac{1}{2}\sin\frac{\pi j(k+1)}{m+1} = \cos\frac{\pi j}{m+1}\sin\frac{\pi jk}{m+1}.$$

Der größte Eigenwert der GSV-Iterationsmatrix ist damit

$$\Lambda = \cos\frac{\pi}{m+1} \approx 1 - \frac{\pi^2}{2(m+1)^2}$$

und damit

$$-\ln\rho(B) \approx \frac{\pi^2}{2(m+1)^2}.$$

Satz 6.17 ergibt den optimalen Parameterwert

$$\omega_0 = \frac{2}{1 + \sin\frac{\pi}{m+1}}$$

sowie

$$\rho\left[B(\omega_0)\right] = \frac{1 - \sin\frac{\pi}{m+1}}{1 + \sin\frac{\pi}{m+1}} \approx 1 - \frac{2\pi}{m+1},$$

also

$$-\ln \rho\,[B(\omega_0)] \approx \frac{2\pi}{m+1}.$$

Daraus folgt

$$\frac{\ln \rho(B)}{\ln \rho\,[B(\omega_0)]} \approx \frac{\pi}{4(m+1)}$$

und mit der schon in Beispiel 6.18 benutzten a-prori Abschätzung daraus die Behauptung. □

Beispiel 6.20. Wir wollen anhand der Poisson-Matrix aus Beispiel 1.4 (bzw. auch Beispiel 2.19) die Leistungsfähigkeit verschiedener iterativer Verfahren testen. (vgl. auch Programmieraufgabe). Zunächst wird in Abbildung 6.1. (angefertigt von J. Löwe) die Entwicklung des Residuums $\|y - Ax\|_2$ in Abhängigkeit von der Iterationszahl für das Gesamt-, Einzelschritt- und (optimierte) SOR-Verfahren sowie die Fälle $n^2 = 400$ bzw. $n^2 = 1.600$ gezeigt. Der Gewinn durch das SOR-Verfahren ist offensichtlich.

Ferner wird für das SOR-Verfahren die Abhängigkeit des optimalen Relaxationsparameters ω in Abhängigkeit von der Dimension n^2 des Lösungsvektors gezeigt. Schließlich wird noch die Entwicklung des Residuums $\|y - Ax\|_2$ in Abhängigkeit von der Iterationszahl für das (optimierte) SOR-Verfahren in Abhängigkeit von der Dimension $n^2 \in \{30^2, 40^2, 50^2, 60^2, 70^2, 80^2, 90^2, 100^2\}$ demonstriert. □

Bemerkung 6.21. Die scharfen Abschätzungen für den Spektralradius des Gesamt- bzw. (relaxierten) Einzelschrittverfahrens aus Beispiel 6.19 zeigen, daß sich die Konvergenzeigenschaften der Basis-Iterationsverfahren bei wachsender Dimension m ständig verschlechtern. Dies wird auch in Abbildung 6.1 für wachsende Dimension n^2 beim optimierten SOR-Verfahren deutlich. Das erfordert eine grundsätzliche Verbesserung des Verfahrensansatzes (vgl. Abschnitt 6.1). Im Kurs *Numerische Mathematik II* wird ein derartiger Ansatz mit *Krylov-Unterraumverfahren* studiert. Die hier untersuchten elementaren Iterationsverfahren verwendet man zur *Vorkonditionierung*. □

6.6 Verfahren der Nachiteration

Abschließend zeigen wir eine Anwendung von Iterationsverfahren zur Verbesserung der Näherungslösung linearer Gleichungssysteme $Ax = y$. Sei x_0 eine irgendwie ermittelte Näherungslösung des Systems, z.B. die rundungsfehlerbehaftete Lösung mit einem direkten Verfahren. Dabei gelte im allgemeinen Fall nur $x_0 = A_{appr}^{-1} y$ mit einer möglicherweise fehlerbehafteten inversen Matrix A_{appr}^{-1}. Dann verschwindet in der Regel nicht das sogenannte *Residuum* oder *Defekt*

$$r_0 := y - Ax_0.$$

Für eine (nun gesuchte) verbesserte Lösung $x_1 = x_0 + \delta_0$ ist die Gleichung $Ax_1 = y$ äquivalent zur *Defektkorrektur-Gleichung*

$$A\delta_0 = r_0.$$

Die Näherungslösung dieser Gleichung mit dem für x_0 verwendeten Verfahren liefert $\delta_0 = A_{appr}^{-1} r_0$ sowie

$$x_1 = x_0 + A_{appr}^{-1}\{y - Ax_0\} = (I - A_{appr}^{-1}A)x_0 + A_{appr}^{-1} y.$$

Eventuell wird diese Prozedur iterativ wiederholt. Unter der Annahme $A_{appr}^{-1}A \approx I$ ist $\|I - A_{appr}^{-1}A\| \ll 1$. Man benötigt dann nur wenige derartige Schritte der Nachiteration. Diese Überlegungen motivieren die

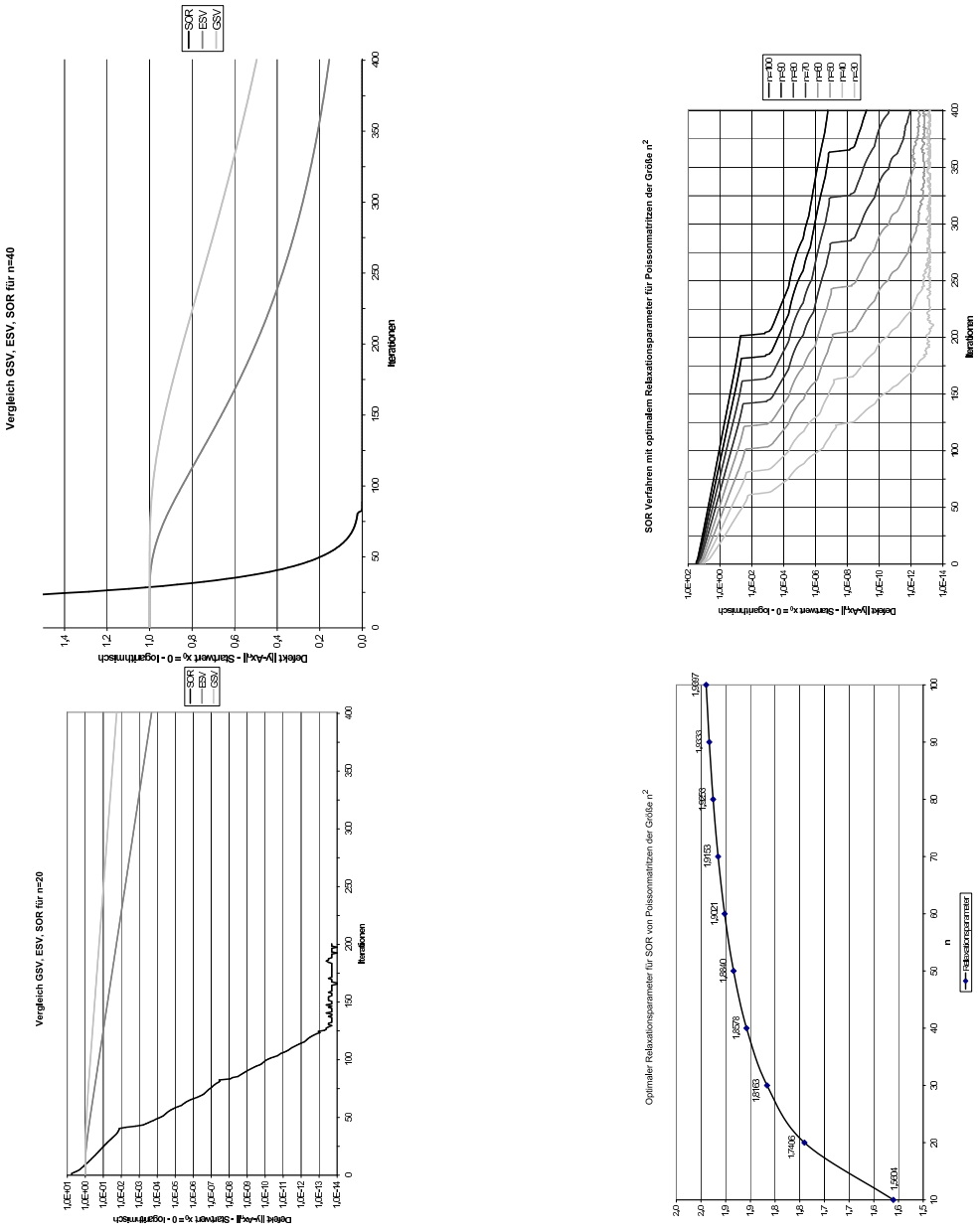

Abbildung 6.1: Vergleich iterativer Verfahren für die Poisson-Matrix

Definition 6.22. *Die Verbesserung einer Näherungslösung x_0 des linearen Gleichungssystems $Ax = y$ nach dem Defektkorrektur-Prinzip*

$$x_{n+1} := x_n + \delta_n, \quad \delta_n = A_{appr}^{-1} r_n, \qquad n = 0, 1, ...,$$

mit dem Defekt $r_n := y - Ax_n$ heißt Verfahren der Nachiteration.

Wir merken noch an, daß dieses Prinzip vor allem Anwendung bei den sogenannten *Mehrgitter-Verfahren* zur effizienten Lösung von Gleichungssystemen, die bei der Diskretisierung von partiellen Differentialgleichungen entstehen, findet.

Kapitel 7

Iterationsverfahren für skalare nichtlineare Gleichungen

Die beiden folgenden Kapiteln behandeln Näherungsverfahren für nichtlineare Gleichungen

$$\text{Finde } (x_1, ..., x_m) \in \mathbb{R}^m : \qquad f_i(x_1, ..., x_m) = 0, \qquad i = 1, ..., m \tag{7.1}$$

bzw. in Kurzform mit $x = (x_1, ..., x_m)^T$ und $f = (f_1, ..., f_m)^T$

$$\text{Finde } x \in \mathbb{R}^m : \qquad f(x) = 0. \tag{7.2}$$

Die bislang untersuchten linearen Systeme sind natürlich ein Spezialfall mit $f(x) := Ax - b$.

Zunächst betrachten wir den Fall $m = 1$, der z.B. das Problem der Nullstellenbestimmung von Polynomen einschließt. Hier konzentrieren wir uns auf die Entwicklung der grundlegenden Ideen. Das folgende Kapitel zum Fall $m \geq 1$ enthält die wesentlichen mathematischen Resultate.

7.1 Bisektionsverfahren

Wir untersuchen zuerst das einfache *Bisektionsverfahren* für Funktionen $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$. Sei

$$f \in C[a, b] \quad \text{und} \quad f(a) \cdot f(b) < 0,$$

d.h. f wechselt das Vorzeichen. Nach dem Zwischenwertsatz existiert dann mindestens eine reelle Nullstelle $x^* \in (a, b) =: (a_0, b_0)$.

Man halbiert dann das Intervall mit $x_0 = \frac{1}{2}(a_0 + b_0)$ und untersucht das Vorzeichen von $f(x_0)$. Bei $f(a_0) \cdot f(x_0) < 0$ setzen wir die Nullstellensuche auf (a_0, x_0) fort, bei $f(x_0) \cdot f(b_0) < 0$ auf (x_0, b_0). Dann wird das Verfahren sukzessiv wiederholt.

Das Verfahren konvergiert für jede stetige Funktion f. Fordert man eine Genauigkeit δ der Nullstellenermittlung, so ergibt sich die benötigte Schrittanzahl aus der Forderung $2^{-n}|b-a| \leq \delta$ zu

$$n := 1 + \left[-\log_2 \frac{\delta}{b-a} \right] = 1 + \left[-\frac{\ln \frac{\delta}{b-a}}{\ln 2} \right].$$

Dabei ist $[x]$ die größte ganze Zahl, die kleiner oder gleich x ist. Das Verfahren konvergiert natürlich nur langsam. Es ist geeignet zur groben Lokalisierung von Nullstellen, die dann mit schneller konvergierenden Verfahren approximiert werden.

7.2 Einfache Iteration

Bessere Konvergenz ist zu erwarten, wenn man neben dem Vorzeichen der Funktion (wie beim Bisektionsverfahren) auch die Funktionswerte in die Berechnung einbezieht. Dazu schreiben wir Gleichung $f(x) = 0$ in Fixpunktform

$$x = g(x) \tag{7.3}$$

und verwenden das Verfahren der sukzessiven Approximation (oder *einfache Iteration*)

$$x_{n+1} = g(x_n), \quad n = 0, 1, 2, ... \tag{7.4}$$

Eine hinreichende Konvergenzaussage gibt der

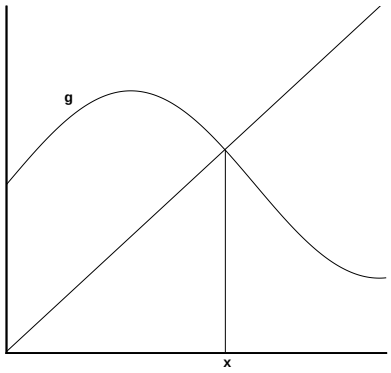

Abbildung 7.1: Fixpunktgleichung $x = g(x)$

Satz 7.1. *Seien $G \subset \mathbb{R}$ ein abgeschlossenes Intervall und $g : G \to G$ eine stetig differenzierbare Funktion mit*

$$q := \sup_{x \in G} |g'(x)| < 1. \tag{7.5}$$

Dann konvergiert das Verfahren (7.4) bei beliebigem Startwert $x_0 \in G$ gegen die eindeutig bestimmte Lösung $x^* \in G$ der Gleichung (7.3). Es gelten die a-priori bzw. a-posteriori Fehlerabschätzung

$$|x_n - x^*| \leq \frac{q^n}{1-q} |x_1 - x_0|, \quad n \in \mathbb{N} \tag{7.6}$$

bzw.

$$|x_n - x^*| \leq \frac{q}{1-q} |x_n - x_{n-1}|, \quad n \in \mathbb{N}. \tag{7.7}$$

Beweis. Der Raum $(\mathbb{R}, \|\cdot\|)$ mit $\|\cdot\| = |\cdot|$ ist vollständig. Der Mittelwertsatz der Differentialrechnung zeigt die Existenz von $\xi \in (x, y)$ mit $g(x) - g(y) = g'(\xi)(x - y)$. Daher gilt

$$|g(x) - g(y)| \leq \sup_{\xi \in G} |g'(\xi)||x - y| = q|x - y|, \ \ \forall x, y \in G, \tag{7.8}$$

und g ist kontrahierend. Die Sätze 5.13 und 5.14 ergeben die Behauptung. □

Die Abbildungen zeigen die Konvergenz des Verfahrens der sukzessiven Approximation in den ersten beiden Fällen mit $|g'(x)| \leq q < 1$. Die dritte Skizze verdeutlicht die Divergenz des Verfahrens für $|g'(x)| \geq \tilde{q} > 1$.

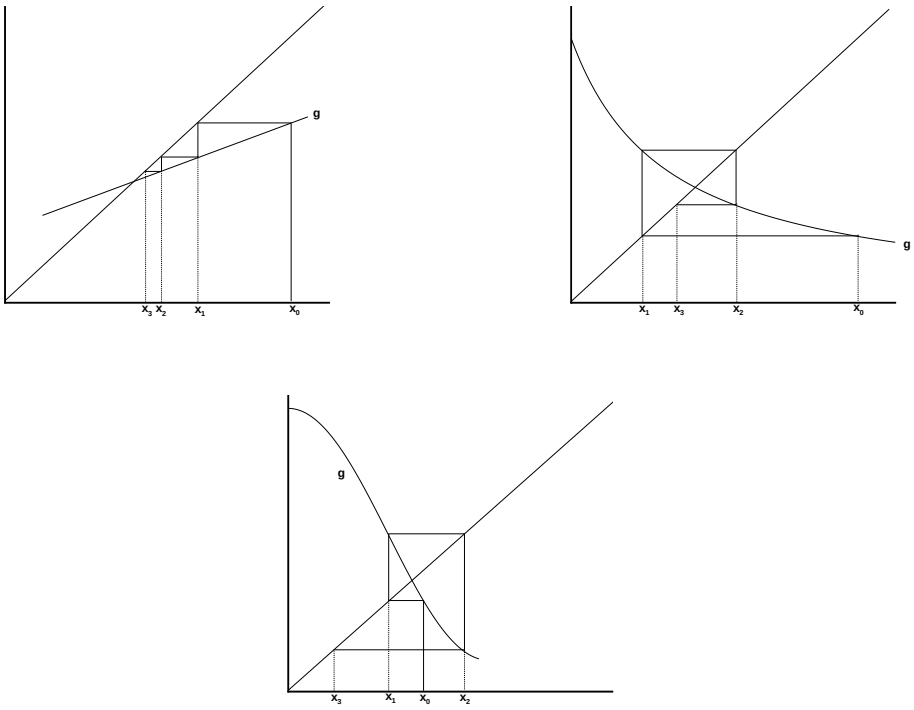

Abbildung 7.2: Konvergenz bzw. Divergenz der einfachen Iteration

Der lokale Charakter von Konvergenzaussagen für das beschriebene Iterationsverfahren bei nichtlinearen Gleichungen wird durch den folgenden Satz verdeutlicht.

Satz 7.2. *Sei x^* Lösung der Gleichung $g(x) = x$ für eine stetig differenzierbare Funktion g mit $|g'(x^*)| < 1$. Dann findet man eine Umgebung U von x^* so, daß das Verfahren der sukzessiven Approximation $x_{n+1} := g(x_n)$ für jeden Startwert $x_0 \in U$ gegen den Fixpunkt x^* konvergiert.*

Beweis. Wegen der stetigen Differenzierbarkeit von g existieren Zahlen $q < 1$ und $\delta > 0$ derart, daß $|g'(x)| \leq q$ für alle $x \in U := [x^* - \delta, x^* + \delta]$. Daraus folgt

$$|g(x) - x^*| = |g(x) - g(x^*)| \leq q|x - x^*| \leq |x - x^*| \leq \delta,$$

d.h. $g : U \to U$. Dann ist der Satz 7.1 anwendbar. □

Eine wesentliche Schwierigkeit des Verfahrens der sukzessiven Approximation besteht in der geschickten Wahl der Fixpunktgleichung, d.h. in der Wahl von g. Dies soll an verschiedenen Beispielen demonstriert werden.

Beispiel 7.3. $x = g(x) := \cos x$

Für die Funktion $g(x) := \cos x$ gilt offenbar $g : [0, 1] \to [0, 1]$ sowie

$$q = \sup_{0 \leq x \leq 1} |g'(x)| = \sin 1 < 1.$$

Nach Satz 7.1 konvergiert das Verfahren der sukzessiven Approximation für jeden Startwert $x_0 \in [0, 1]$ gegen die Lösung x^* von $x = \cos x$.

n	x_n	n	x_n
0	1.00000000	⋮	⋮
1	0.54030231	45	0.73908513
2	0.85755322	46	0.73908514
3	0.65428979	47	0.73908513
4	0.79348036	48	0.73908513
5	0.70136877		
6	0.76395968		
7	0.72210243		

Deutlich wird das oszillierende Verhalten der Lösungsfolge in Umgebung der Nullstelle x^*. Die Konvergenzgeschwindigkeit ist unbefriedigend. □

Beispiel 7.4. $f(x) := x + \ln x$

Eine grobe Skizze zeigt, daß f eine Nullstelle $x^* \in (0,1)$ besitzt. Wir wählen zunächst naiv $g(x) := -\ln x$ und lösen $g(x) = x$. Offenbar ist aber $|g'(x)| = \frac{1}{x} > 1$ im Intervall $(0,1)$. Das Verfahren der sukzessiven Approximation divergiert dann.

Wir gehen nun zur inversen Funktion $\tilde{g}(x) := e^{-x}$ über und lösen die zum Ausgangsproblem äquivalente Gleichung $\tilde{g}(x) = x$. Es gilt im Fixpunkt

$$|\tilde{g}'(x^*)| = e^{-x^*} < 1.$$

Damit ist der Satz 7.2 anwendbar. Ferner bildet $\tilde{g}$ das Intervall $[a,1]$ für jedes $a \in (0, 1/e)$ in sich ab. Wegen

$$q = \sup_{a \le x \le 1} |\tilde{g}'(x)| = e^{-a} < 1$$

konvergiert dann das Verfahren der einfachen Iteration sogar für beliebige Startwerte $x_0 > 0$ gegen die Lösung x^* von $\tilde{g}(x) = x$.

n	x_n	n	x_n
0	1.00000000	9	0.56487935
1	0.36787944	10	0.56842873
2	0.69220063	⋮	⋮
3	0.50047350	⋮	⋮
4	0.60624354	31	0.56714328
5	0.54539579	32	0.56714330
6	0.57961234	33	0.56714329
7	0.56011546	34	0.56714329
8	0.57114312	35	0.56714329

Erneut oszilliert die Lösungsfolge um die Nullstelle x^*. Die Konvergenzgeschwindigkeit ist nicht zufriedenstellend. □

Beispiel 7.5. *Division durch Iteration*

Zur Berechnung von $x = \frac{1}{a}, a > 0$ betrachten wir die quadratische Funktion $g : \mathbb{R} \to \mathbb{R}$ mit

$$g(x) := 2x - ax^2.$$

Die Gleichung $x = g(x)$ hat die Fixpunkte $x^* = 0$ und $x^* = \frac{1}{a}$. Man findet sofort, daß die Abbildung g das Intervall $(0, 2/a)$ in $(0, 1/a)$ abbildet. Ferner ist

$$g'(x) = 2(1 - ax).$$

Wegen $g'(1/a) = 0$ besitzt der Fixpunkt $x^* = 1/a$ eine Umgebung, so daß g dort eine kontrahierende Abbildung ist. Wegen der Ungleichung $x < g(x) >< 1/a$ im Intervall $(0, 1/a)$ ist die Folge $(x_n)_n$ der einfachen Iteration monoton wachsend und beschränkt. Nach Satz 7.2 konvergiert das Verfahren der sukzessiven Approximation dann für jeden Startwert $x_0 \in (0, 2/a)$ gegen den Fixpunkt $x^* = 1/a$. Die Tabelle veranschaulicht die Konvergenz des Verfahrens für $a = 2$ und die Startwerte $x = 0.3$ bzw. $x = 0, 4$.

n	x_n	x_n
0	0.30000000	0.40000000
1	0.42000000	0.48000000
2	0.48720000	0.49920000
3	0.49967232	0.49999872

Die Konvergenzgeschwindigkeit ist im Unterschied zu den beiden vorherigen Beispielen sehr gut. Man beachte u.a., daß $g'(x) \to 0$ für $x \to x^* = 1/a$ gilt. . □

Beispiel 7.6. *Babylonisches Wurzelziehen*
Zur Lösung der Gleichung $f(x) = 0$ mit $f(x) := x^2 - a$ mit $a > 0$ untersuchen wir die Funktion $g : [0, \infty) \to (0, \infty)$ mit

$$g(x) := \frac{1}{2}\left(x + \frac{a}{x}\right).$$

Diese quadratische Funktion hat den Fixpunkt $x^* = \sqrt{a}$. Nach der Ungleichung vom geometrisch–arithmetischen Mittel ist $g(x) > \sqrt{a}$ für $x > 0$, d.h. es gilt $g : (0, \infty) \to [\sqrt{a}, \infty)$. Wegen

$$g'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)$$

ist

$$q := \sup_{\sqrt{a} \le x < \infty} |g'(x)| = \frac{1}{2}.$$

Damit ist der Satz 7.1 anwendbar, und das Verfahren der sukzessiven Approximation konvergiert für jeden Startwert $x_0 > 0$ gegen den Fixpunkt $x^* = \sqrt{a}$ mit der a-posteriori Fehleraussage

$$|x_n - \sqrt{a}| \le |x_n - x_{n-1}|.$$

Die Tabelle verdeutlicht wie in Beispiel 7.5 den raschen Konvergenzprozeß, hier für den Wert $a = 2$.

n	x_n
0	5.00000000
1	2.70000000
2	1.72037037
3	1.44145537
4	1.41447098
5	1.41421359
6	1.41421356

□

7.3 Newton-Verfahren

Bei den bisher betrachteten Verfahren zur Nullstellenbestimmung wurde im Algorithmus nicht die Differenzierbarkeit der betrachteten Funktion benutzt. Lediglich im Konvergenzbeweis wurde die stetige Differenzierbarkeit auf einem kompakten Intervall als hinreichende Konvergenzbedingung benutzt. Wir verwenden nun auch im Algorithmus Informationen über die Ableitung.

Sei x_0 eine Näherung an eine Nullstelle der Funktion f. Dann beschreibt

$$G(x) := f(x_0) + f'(x_0)(x - x_0) \tag{7.9}$$

die Tangente an die durch f beschriebene Kurve im Punkt $(x_0, f(x_0))$. Der Schnittpunkt der Tangente mit der $x-$Achse wird als neue Näherung x_1 an die Nullstelle von f gewählt. Sie ergibt sich aus der Gleichung

$$f(x_0) + f'(x_0)(x_1 - x_0) = 0 \tag{7.10}$$

zu

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{7.11}$$

Eine andere Motivation ergibt sich aus der Taylorschen Formel durch Abbruch nach dem linearen Glied

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) =: G(x). \tag{7.12}$$

Neuer Näherungswert für die Nullstelle von f ist die Nullstelle x_1 der linearen Funktion G, d.h.

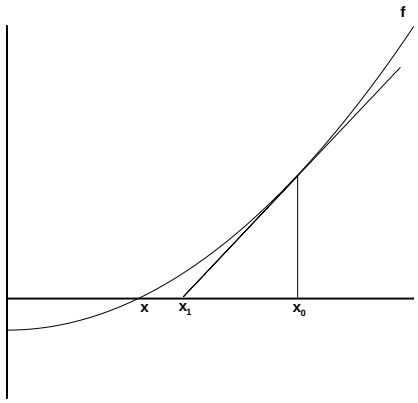

Abbildung 7.3: Newton-Verfahren

erneut (7.11). Sukzessive Wiederholung dieser Vorgehensweise ergbt das Newton- Verfahren.

Definition 7.7. *Seien $G \subset \mathbb{R}$ offen und $f : G \to \mathbb{R}$ eine stetig differenzierbare Funktion mit $f'(x) \neq 0$ für alle $x \in G$. Das Iterationsverfahren*

$$x_{n+1} := x_n - (f'(x_n))^{-1} f(x_n), \quad n = 0, 1, 2, ... \tag{7.13}$$

mit Startwert $x_0 \in G$ heißt Newton-Verfahren *zur Lösung der Aufgabe $f(x) = 0$.*

Offenbar ist das Newton-Verfahren ein Spezialfall der einfachen Iteration mit

$$x_{n+1} = g(x_n), \qquad g(x) := x - (f'(x))^{-1} f(x).$$

Wir wenden nun das Newton-Verfahren auf die Beispiele 7.3 - 7.6 an.

Beispiel 7.8. $f(x) := x - \cos x$

Für die Funktion $f(x) := x - \cos x$ ergibt sich als Vorschrift für das Newton-Verfahren

$$x_{n+1} = x_n - \frac{x_n - \cos x_n}{1 + \sin x_n}.$$

Im Unterschied zur sukzessiven Approximation in Beispiel 7.3 finden wir jetzt eine sehr gute Konvergenz der Lösungsfolge.

n	x_n
0	1.00000000
1	0.75036387
2	0.73911289
3	0.73908513
4	0.73908513

□

Beispiel 7.9 $f(x) := x - e^{-x}$

Für die bereits im Beispiel 7.4 betrachtete Funktion $f(x) := x - e^{-x}$ lautet die Iterationsvorschrift des Newton-Verfahrens

$$x_{n+1} = x_n - \frac{x_n - e^{-x_n}}{1 + e^{-x_n}}.$$

Es ergibt sich im Unterschied zur sukzessiven Approximation im Beispiel 7.4 ein sehr günstiges Konvergenzverhalten.

n	x_n
0	1.00000000
1	0.53788284
2	0.56698699
3	0.56714329
4	0.56714329

□

Beispiel 7.10. *Division durch Iteration*

Sei $f(x) := \frac{1}{x} - a, \;\; a > 0$. Das Newton-Verfahren liefert die Vorschrift

$$x_{n+1} := x_n - \frac{\frac{1}{x_n} - a}{-\frac{1}{(x_n)^2}} = x_n(2 - ax_n).$$

Dies ist genau die im Beispiel 7.5 gewählte Vorschrift für die Iteration. Es gelten also die bereits dort gefundenen Resultate. □

Beispiel 7.11. *Babylonisches Wurzelziehen*

Sei $f(x) := x^2 - a, \;\; a > 0$. Das Newton-Verfahren lautet

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Dies ist exakt die bereits im Beispiel 7.6 benutzte Iterationsvorschrift. □

Die Beispiele 7.8 und 7.9 zeigen eine sehr viel schnellere Konvergenz als das Verfahren der sukzessiven Approximation. (Bei den Beispielen 7.10 und 7.11 hatten wir durch geeignete Wahl der Fixpunktgleichung bereits das Newton-Verfahren erzeugt.) Durch Anwendung von Satz 7.2 erhalten wir bereits die folgende Konvergenzaussage.

Satz 7.12. *Sei x^* einfache Nullstelle der in Umgebung von x^* zweimal stetig differenzierbaren Funktion f, d.h. es gilt $f'(x^*) \neq 0$. Dann konvergiert das Newton-Verfahren in einer hinreichend kleinen Umgebung von x^**

Beweis: Für die Verfahrensfunktion $g(x) := x - f(x)/f'(x)$ gilt

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f''(x)}{[f'(x)]^2} f(x),$$

also $g'(x^*) = 0$. Dann ergibt Satz 7.2 die Behauptung. □

Wegen $g'(x^*) = 0$ ist eine besonders schnelle Konvergenz des Newton-Verfahrens (als Spezialfall der einfachen Iteration) zu erwarten. Eine vertiefende Konvergenzanalyse führen wir im folgenden Kapitel durch. Eine Modifikation des Newton-Verfahrens für mehrfache Nullstellen behandeln wir in einer Übungsaufgabe.

7.4 Newton-artige Verfahren

Bei der praktischen Durchführung des Newton-Verfahrens erweist sich (insbesondere im Fall nichtlinearer Gleichungssysteme mit $m > 1$) die Berechnung der Ableitungen $f'(x_n)$ als aufwendig. Man versucht daher, die Ableitung in geeigneter Weise zu approximieren.

Beim *vereinfachten* oder *modifizierten Newton–Verfahren*

$$x_{n+1} := x_n - (f'(x_0))^{-1} f(x_n), \quad n = 0, 1, ... \tag{7.14}$$

ersetzt man die Tangente an die Kurve f im Näherungswert $(x_n, f(x_n))$ durch Parallelen zur Tangente an f im Punkt $(x_0, f(x_0))$. Eine Verbesserung dieses Verfahrens kann erwartet werden, wenn man die Ableitung $f'(x_0)$ nach N Schritten abändert zu $f'(x_N)$. Diese Aufdatierung der Ableitung kann zyklisch wiederholt werden.

Bei der *Sekanten-Methode* geht man von zwei Näherungswerten x_0 und (dem etwa mit einem Newton-Schritt berechneten) x_1 aus. Wir ersetzen nun die Tangente an den Punkt $(x_0, f(x_0))$ durch die Sekante durch die Kurvenpunkte $(x_0, f(x_0))$ sowie $(x_1, f(x_1))$, d.h.

$$g(x) := f(x_1) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_1).$$

Der Schnittpunkt der Sekante mit der $x-$Achse wird als neue Näherung x_2 der Nullstelle von f gewählt. Wir erhalten aus $g(x_2) = 0$, daß

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1).$$

Sukzessive Wiederholung ergibt die Vorschrift

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n), \quad n = 0, 1, ..., \tag{7.15}$$

d.h. die Ableitung $f'(x_n)$ wird durch einen einseitigen Differenzenquotienten approximiert.

Die *Regula falsi* kombiniert das Sekanten-Verfahren mit dem Bisektionsverfahren. Seien x_0 und $\tilde{x}_0$ Näherungswerte an die Nullstelle x^*, und es gelte $f(x_0)f(\tilde{x}_0) < 0$. Nach dem Zwischenwertsatz wird die Nullstelle dann von x_0 und $\tilde{x}_0$ eingeschlossen. Dann berechnen wir die Nullstelle ξ der Sekante durch die Kurvenpunkte $(x_0, f(x_0))$ und $(\tilde{x}_0, f(\tilde{x}_0))$, d.h.

$$\xi = \frac{x_0 f(\tilde{x}_0) - \tilde{x}_0 f(x_0)}{f(\tilde{x}_0) - f(x_0)}. \tag{7.16}$$

Nun wählen wir als neue Näherungen zur Einschließung der Nullstelle die Werte $x_1 = \xi, \tilde{x}_1 = \tilde{x}_0$, falls $f(x_0)f(\xi) > 0$, bzw. $x_1 = x_0, \tilde{x}_1 = \xi$, falls $f(\tilde{x}_0)f(\xi) > 0$. Dann wird die Nullstelleneinschließung in der beschriebenen Weise sukzessiv fortgesetzt.

7.5 Nullstellenbestimmung von Polynomen

Wir wollen nun das Newton-Verfahren zur Nullstellenbestimmung von Polynomen

$$P(x) = a_0 x^m + a_1 x^{m-1} + \cdots + a_{m-1} x + a_m \tag{7.17}$$

verwenden. Zur praktischen Durchführung der Iterationsvorschrift

$$x_{n+1} := x_n - \frac{P(x_n)}{P'(x_n)}, \quad n = 0, 1, 2, ... \tag{7.18}$$

müssen die Größen $P(x_n)$ und $P'(x_n)$ effektiv berechnet werden. Ein sehr ökonomisches Verfahren ist das *Horner-Schema* (Horner 1819), das auf einer geeigneten Klammersetzung im Polynom basiert:

$$P(z) = (\cdots((a_0 z + a_1)z + a_2)z + \cdots + a_{m-1})z + a_m. \tag{7.19}$$

Durch rekursive Berechnung der Werte

$$\begin{aligned} b_0 &:= a_0 \\ b_1 &:= b_0 z + a_1 \\ b_2 &:= b_1 z + a_2 \\ &\vdots \qquad \vdots \\ b_m &:= b_{m-1} z + a_m \end{aligned}$$

erhalten wir über m Multiplikationen und Additionen den Wert

$$P(z) = b_m. \tag{7.20}$$

Dies ist deutlich effizienter als die Funktionswertberechnung in der Monombasis $\{x^i\}_{i=0}^m$. Bei letzterer Variante sind im Extremfall $O(m^2)$ wesentliche Rechenoperationen erforderlich. Es sei angemerkt, daß der Funktionswertaufruf eine sich sehr oft wiederholende Grundoperation in vielen komplexeren Algorithmen darstellt. Sehr häufig ersetzt man auch aufwendig zu berechnende Funktionen (wenigstens lokal) durch leichter zu berechnende Polynome (s. dazu ab Kapitel 9 der Vorlesung).

Die effiziente Funktionswertberechnung läßt sich auch auf Ableitungen erweitern. Für das Polynom

$$P_1(x) = b_0 x^{m-1} + b_1 x^{m-2} + \cdots + b_{m-2}x + b_{m-1} = \sum_{k=0}^{m-1} b_k x^{m-1-k} \tag{7.21}$$

gilt nach kurzer Rechnung

$$\begin{aligned} P_1(x)(x-z) + b_m &= \sum_{k=0}^{m-1} b_k x^{m-1-k}(x-z) + b_m \\ &= \sum_{k=1}^{m} \underbrace{(b_k - b_{k-1}z)}_{=a_k} x^{m-k} + \underbrace{b_0}_{=a_0} x^m = P(x). \end{aligned}$$

Somit erzeugt das Horner-Schema für die Nullstelle z von $P(\cdot)$ das durch Division des Linearfaktors $(x-z)$ entstehende Polynom $P_1(\cdot)$. Nach Differentiation folgt

$$P'(x) = P_1'(x)(x-z) + P_1(x)$$

und damit in effektiver Weise die gesuchte Größe

$$P'(z) = P_1(z). \tag{7.22}$$

Induktive Fortsetzung der Differentiation ergibt

$$P^{(j)}(x) = P_1^{(j)}(x)(x-z) + jP_1^{(j-1)}(x)$$

also

$$P^{(j)}(z) = jP_1^{(j-1)}(z), \qquad j = 1, ..., m. \tag{7.23}$$

Nun erzeugt man rekursiv Polynome P_j vom Grade $m-j$ durch Anwendung des Horner-Schemas auf das Vorgängerpolynom P_{j-1} und erhält

$$P^{(j)}(z) = j!P_j(z), \qquad j = 1, ..., m. \tag{7.24}$$

Wir fassen das Verfahren, das auch im komplexwertigen Fall benutzt werden kann, in dem folgenden Satz zusammmen.

Satz 7.13. *Sei $P(x) = a_0x^m + a_1x^{m-1} + \cdots + a_{m-1}x + a_m$ Polynom vom Grade m. Für jedes $z \in \mathbb{C}$ stehen im vollständigen Horner-Schema*

	a_0	a_1	a_2	$\cdots$	$\cdots$	a_{m-1}	a_m
z	b_0	b_1	b_2	$\cdots$	$\cdots$	b_{m-1}	b_m
z	b_0'	b_1'	b_2'	$\cdots$	$\cdots$	b_{m-1}'	
z	b_0''	b_1''	b_2''	$\cdots$	b_{m-2}''		
$\vdots$	$\vdots$	$\vdots$					
z	$b_0^{(m-1)}$	$b_1^{(m-1)}$					
z	$b_0^{(m)}$						

die Ableitungen

$$b_{m-j}^{(j)} = \frac{P^{(j)}(z)}{j!}, \quad j = 0, 1, ..., m \tag{7.25}$$

von P an der Stelle z. Die Bildungsvorschrift ist rekursiv definiert durch

$$b_0^{(j)} := b_0^{(j-1)}, \quad b_k^{(j)} := z b_{k-1}^{(j)} + b_k^{(j-1)}, \quad k = 1, ..., m-j \tag{7.26}$$

für $j = 0, ..., m$ und beginnend mit

$$b_k^{(-1)} := a_k, \quad k = 1, ..., m. \tag{7.27}$$

Zur Illustration betrachten wir das folgende

Beispiel 7.14. *Laguerre-Polynom 4.Grades*
Untersucht wird das Laguerre-Polynom 4. Grades, das zur Vereinfachung mit dem Faktor 24 multipliziert wird:

$$P_4(x) = 24\left(\frac{1}{24}x^4 - \frac{2}{3}x^3 + 3x^2 - 4x + 1\right).$$

Das vollständige Horner-Schema an der Stelle $x = 0.4$ ergibt sich zu

z	1	-16	72	-96	24
0.4	1	-15.6	65.76	-69.696	-3.8784
0.4	1	-15.2	59.68	-45.824	
0.4	1	-14.8	53.76		
0.4	1	-14.4			
0.4	1				

Damit erhalten wir die Entwicklung des Polynoms an der Stelle $x = 0.4$ mit

$$P_4(x) = (x-0.4)^4 - 14.4(x-0.4)^3 + 53.76(x-0.4)^2 - 45.824(x-0.4) - 3.8784.$$

Weiterhin erhalten wir die Werte der Ableitungen $P_4(0.4) = -3.8784$, $P_4'(0.4) = -45.824$, $P_4^{(2)}(0.4) = 2! \cdot 53.76 = 107.52$, $P_4^{(3)}(0.4) = 3! \cdot (-14.4) = -86.4$ und $P_4^{(4)}(0.4) = 4! \cdot 1 = 24$. □

Man kann nun folgenden Plan zur Bestimmung der reellen Nullstellen eines Polynoms entwerfen. Wir betrachten zunächst den Fall, daß alle Nullstellen einfach sind.

1. Grobe Lokalisierung der Nullstellen $z_m < z_{m-1} < ... < z_2 < z_1$ (z.B. mittels Bisektionsverfahren)

2. Näherungsberechnung von z_1 über Newton-Verfahren mit Startwert $x_0 > z_1$ (Konvergenz aus Monotoniegründen)

3. Abspaltung des Linearfaktors $(x - z_1)$ mittels Horner-Schema und Wiederholung des 2. Schritts für reduziertes Polynom

4. Sukzessive Näherungsberechnung aller Nullstellen nach 2. und 3. Schritt

5. Wiederholung der Newton-Iteration für das vollständige Polynom für alle Nullstellen z_i mit näherungsweise berechneten Nullstellen aus 3. Schritt als Startwert (Vermeidung von Rundungsfehlern !).

Schließlich gehen wir noch kurz auf den Fall *mehrfacher* Nullstellen des Polynoms ein. Exemplarisch gelte die Darstellung

$$P(x) = (x - z)^l Q(x)$$

bei einer Nullstelle z der Ordnung $l \in \mathbb{N},\ l \geq 2$ und mit einem Polynom Q vom Grad $m - l$ und $Q(z) \neq 0$. Bei Anwendung des Newton-Verfahrens gemäß (7.18) erhält man bei mehrfacher Anwendung der Regel von l'Hospital

$$|g'(z)| = \left|1 - \frac{1}{l}\right| < 1$$

mit $g(x) := x - f(x)/f'(x)$. Nach Satz 7.2 konvergiert dann das Newton-Verfahren lokal. In diesem Sinne kann der oben besprochene Plan zur Berechnung der reellen Nullstellen eines Polynoms auf den Fall mehrfacher Nullstellen erweitert werden. (Man erkennt, daß sich die Kontraktionskonstante des Newton-Verfahrens mit wachsender Ordnung l der Nullstelle verschlechtert. In einer Übungsaufgabe behandeln wir, wie das Konvergenzverhalten wesentlich verbessert werden kann.)

Nachfolgender Satz und Beispiel unterstreichen die Notwendigkeit des letzten Schritts des oben genannten Planes zur Nullstellenberechnung.

Satz 7.15. *Seien P_1 und P_2 Polynome, z sei einfache Nullstelle von P_1. Dann hat das gestörte Polynom $P_\epsilon := P_1 + \epsilon P_2$ mit $0 < \epsilon \ll 1$ in erster Näherung die Nullstelle*

$$z_\epsilon = z - \epsilon \frac{P_2(z)}{P_1'(z)}.$$

Beweis. Wir setzen $z_\epsilon = z + \alpha\epsilon$ und erhalten unter Beachtung der Taylor-Entwicklung von P_ϵ an der Stelle z

$$\begin{aligned} 0 &= P_\epsilon(z_\epsilon) = P_1(z + \alpha\epsilon) + \epsilon P_2(z + \alpha\epsilon) \\ &= P_1(z) + P_1'(z)\alpha\epsilon + P_2(z)\epsilon + 0(\epsilon)^2 \end{aligned}$$

sowie unter Berücksichtigung von $P_1(z) = 0$, daß in erster Näherung $\alpha P_1'(z) + P_2(z) = 0$ gilt und damit

$$\alpha = -\frac{P_2(z)}{P_1'(z)}. \qquad \square$$

Beispiel 7.16. *Schlecht konditioniertes Problem*
Das Polynom 10. Grades

$$P_1(x) = \prod_{i=1}^{10} (x - i) = x^{10} - 55x^9 + ... + 10!$$

hat die Nullstellen $z_i = i,\ \ i = 1, ..., 10$. Stört man nun P_1 mit dem Polynom $\epsilon P_2(x) = \epsilon 55x^9$, so erhalten wir wegen $P_1'(10) = 9!$ in erster Näherung eine Störung der Nullstelle $z = 10$ von

$$z_\epsilon = 10 - \epsilon \frac{55 \cdot 10^9}{9!} \approx 10(1 - 1.5 \cdot 10^4 \epsilon). \qquad \square$$

Im Zusammenhang mit der Lösung von Eigenwertaufgaben wird im Kurs *Numerische Mathematik II* eine stabilere Methode zur Nullstellenbestimmung über die QR-Methode behandelt.

Kapitel 8

Iterationsverfahren für nichtlineare Gleichungssysteme

Nach der Behandlung des skalaren Falles in Kapitel 7 untersuchen wir nun allgemeiner Iterationsverfahren zur Lösung von Systemen nichtlinearer Gleichungen

$$\text{Finde } \; x \in \mathbb{R}^m : \qquad f(x) = 0 \tag{8.1}$$

mit $f = (f_1, ..., f_m)^T, \;\; x = (x_1, ..., x_m)^T, \;\; m \geq 1.$

8.1 Einfache Iteration (Sukzessive Approximation)

Sei das System (8.1) überführt in die Fixpunktform

$$x = g(x) \qquad \text{bzw.} \qquad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} g_1(x_1, ..., x_m) \\ g_2(x_1, ..., x_m) \\ \vdots \\ g_m(x_1, ..., x_m) \end{pmatrix}. \tag{8.2}$$

Dann ergibt sich das Verfahren der *einfachen Iteration* (bzw. sukzessiven Approximation)

$$x^{(n+1)} = g(x^{(n)}) \qquad \text{bzw.} \qquad \begin{pmatrix} x_1^{(n+1)} \\ x_2^{(n+1)} \\ \vdots \\ x_m^{(n+1)} \end{pmatrix} = \begin{pmatrix} g_1(x_1^{(n)}, ..., x_m^{(n)}) \\ g_2(x_1^{(n)}, ..., x_m^{(n)}) \\ \vdots \\ g_m(x_1^{(n)}, ..., x_m^{(n)}) \end{pmatrix}. \tag{8.3}$$

Wir formulieren zunächst zur Vorbereitung der Konvergenzanalyse von (8.3) den Mittelwertsatz für stetig differenzierbare Funktionen $g : G \subset \mathbb{R}^m \to \mathbb{R}^m$.

Definition 8.1. *$G \subset \mathbb{R}^m$ heißt* konvex, *wenn gilt:* $\lambda x + (1-\lambda)y \in G, \; \forall x, y \in G, \; \forall \lambda \in (0,1)$.

Satz 8.2. *Sei $G \subset \mathbb{R}^m$ offen und konvex sowie $g : G \to \mathbb{R}^m$ eine Abbildung*

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix} = \begin{pmatrix} g_1(x_1, ..., x_m) \\ \vdots \\ g_m(x_1, ..., x_m) \end{pmatrix}$$

mit auf G partiell stetig differenzierbaren Funktionen $g_i, i = 1, ..., m$; d.h. $g_i \in C^1(G)$. Dann gilt in jeder Norm $\|\cdot\|$ auf $\mathbb{R}^m$ der Mittelwertsatz

$$\|g(x) - g(y)\| \leq \max_{0\leq\lambda\leq 1} \|g'(\lambda x + (1-\lambda)y)\|\|x - y\| \tag{8.4}$$

für alle $x, y \in G$. Dabei ist

$$g'(x) = \left(\frac{\partial g_i}{\partial x_k}(x)\right)_{i,k=1}^{m} \tag{8.5}$$

die Jacobische Funktionalmatrix *von g.*

Beweis: (i) In einem vorbereitenden Schritt betrachten wir eine stetige vektorwertige Funktion $h : [0, 1] \to \mathbb{R}^m$ und zeigen

$$\left\| \int_0^1 h(\lambda) d\lambda \right\| \leq \int_0^1 \|h(\lambda)\| d\lambda. \tag{8.6}$$

Die linke Seite von (8.6) ist mit komponentenweiser Definition des Integrals zu verstehen. Die Abbildung $\lambda \mapsto \|h(\lambda)\|$ stetig und somit die rechte Seite in (8.6) wohldefiniert. Bei Zerlegung des Intervalls $[0, 1]$ mit $\lambda_j = j/n, j = 0, 1, ..., n,\ \ n \in \mathbb{N}$ konvergieren die Riemann-Summen

$$\sum_{j=0}^{n} \|h(\lambda_j)\|(\lambda_j - \lambda_{j-1}) \to \int_0^1 \|h(\lambda)\| d\lambda, \quad n \to \infty$$

$$\sum_{j=0}^{n} h(\lambda_j)(\lambda_j - \lambda_{j-1}) \to \int_0^1 h(\lambda) d\lambda, \quad n \to \infty.$$

Die zweite Dreiecksungleichung (vgl. Satz 4.3) liefert dann

$$\begin{aligned} 0 \quad &\leq \quad \left| \left\| \sum_{j=0}^{n} h(\lambda_j)(\lambda_j - \lambda_{j-1}) \right\| - \left\| \int_0^1 h(\lambda) d\lambda \right\| \right| \\ &\leq \quad \left\| \sum_{j=0}^{n} h(\lambda_j)(\lambda_j - \lambda_{j-1}) - \int_0^1 h(\lambda) d\lambda \right\| \to 0, \quad n \to \infty, \end{aligned}$$

daher

$$\left\| \sum_{j=0}^{n} h(\lambda_j)(\lambda_j - \lambda_{j-1}) \right\| \to \left\| \int_0^1 h(\lambda) d\lambda \right\|, \ \ n \to \infty.$$

Führen wir nun in der Dreiecksungleichung

$$\left\| \sum_{j=0}^{n} h(\lambda_j)(\lambda_j - \lambda_{j-1}) \right\| \leq \sum_{j=0}^{n} \|h(\lambda_j)\|(\lambda_j - \lambda_{j-1})$$

den Grenzübergang $n \to \infty$ aus, so ergibt sich die Ungleichung (8.6).

(ii) Komponentenweise (d.h. für $i = 1, ..., m$) gilt für fixierte, jedoch beliebige Werte $x, y \in G$ der Hauptsatz der Differential- und Integralrechnung in der Form

$$g_i(x) - g_i(y) = \int_0^1 \frac{dg_i}{d\lambda}(\lambda x + (1-\lambda)y) d\lambda$$

bzw. mit Kettenregel

$$g_i(x) - g_i(y) = \int_0^1 \sum_{k=1}^m \frac{\partial g_i}{\partial x_k}(\lambda x + (1-\lambda)y)(x_k - y_k)d\lambda.$$

Hierbei wurde implizit die Konvexität von G benutzt. Damit folgt in Vektorform

$$g(x) - g(y) = \int_0^1 g'(\lambda x + (1-\lambda)y)(x-y)d\lambda. \tag{8.7}$$

(Beachte: Matrix-Vektor-Multiplikation unter dem Integral !)

(iii) Ungleichung (8.6) ergibt bei Anwendung auf Gleichung (8.7) die Behauptung wegen

$$\begin{aligned} \|g(x) - g(y)\| &\leq \int_0^1 \|g'(\lambda x + (1-\lambda)y)(x-y)\|d\lambda \\ &\leq \int_0^1 \|g'(\lambda x + (1-\lambda)y)\|\|x-y\|d\lambda \\ &\leq \sup_{0\leq\lambda\leq 1} \|g'(\lambda x + (1-\lambda)y)\|\|x-y\|. \qquad \square \end{aligned}$$

Wir zeigen nun einen Konvergenzsatz für das Verfahren der einfachen Iteration.

Satz 8.3. *Sei $G \subset \mathbb{R}^m$ abgeschlossen und konvex sowie $g : G \to G$ eine Abbildung mit Funktionalmatrix $g' \in C(\overline{G})$ (d.h. jedes Matrixelement von g' ist stetig in G und stetig auf den Rand ∂G fortsetzbar). Schließlich gelte eine der folgenden Bedingungen:*

$$(i) \qquad q_\infty := \sup_{x\in G} \max_{i=1,\ldots,m} \sum_{k=1}^m \left|\frac{\partial g_i}{\partial x_k}(x)\right| < 1 \tag{8.8}$$

$$(ii) \qquad q_1 := \sup_{x\in G} \max_{k=1,\ldots,m} \sum_{i=1}^m \left|\frac{\partial g_i}{\partial x_k}(x)\right| < 1 \tag{8.9}$$

$$(iii) \qquad q_2 := \sup_{x\in G} \left\{ \sum_{i,k=1}^m \left|\frac{\partial g_i}{\partial x_k}(x)\right|^2 \right\}^{1/2} < 1. \tag{8.10}$$

Dann konvergiert das Verfahren der einfachen Iteration

$$x^{(n+1)} := g(x^{(n)}), \quad n = 0, 1, 2, \ldots \tag{8.11}$$

bei beliebigem Startvektor $x^{(0)} \in G$ gegen die eindeutig bestimmte Lösung x^* des nichtlinearen Gleichungssystems $g(x) = x$. Für $\mu = 1, 2, \infty$ gelten für $q_\mu < 1$ die a-priori bzw. die a-posteriori Fehleraussage

$$\|x^{(n)} - x^*\|_\mu \leq \frac{q_\mu^n}{1-q_\mu}\|x^{(1)} - x^{(0)}\|_\mu, \ \forall n \in \mathbb{N} \tag{8.12}$$

bzw.

$$\|x^{(n)} - x^*\|_\mu \leq \frac{q_\mu}{1-q_\mu}\|x^{(n)} - x^{(n-1)}\|_\mu, \ \forall n \in \mathbb{N}. \tag{8.13}$$

Beweis. Satz 4.18 über die Darstellung spezieller Matrixnormen $\|\cdot\|_\mu$ zeigt zunächst

$$\sup_{x\in G} \|g'(x)\|_\mu = q_\mu, \qquad \mu = 1, \infty; \quad \sup_{x\in G} \|g'(x)\|_2 \leq q_2.$$

Bei $q_\mu < 1$ ist nach dem Mittelwertsatz 8.1 die Abbildung $g : G \to G$ kontrahierend. Der Banachsche Fixpunktsatz 5.13 und Satz 5.14 ergeben die Behauptung. □

Ein zum lokalen Konvergenzsatz 8.3 analoges Resultat gibt der

Satz 8.4. *Sei x^* Fixpunkt der Gleichung $g(x) = x$ bei stetig differenzierbarer Funktion g. Es gelte $\|g'(x^*)\| < 1$ bezüglich einer (beliebigen) Norm $\|\cdot\|$ auf $\mathbb{R}^m$. Dann existiert eine Umgebung U von x^*, so daß das Verfahren der einfachen Iteration $x^{(n+1)} := g(x^{(n)})$ bei beliebigem Startvektor $x^{(0)} \in U$ gegen x^* konvergiert.*

Beweis: Übungsaufgabe (analog zu Beweis von Satz 8.3.) □

8.2 Gesamt- und Einzelschrittverfahren

Das Verfahren der einfachen Iteration

$$x_i^{(n+1)} = g_i\left(x_1^{(n)}, ..., x_m^{(n)},\right), \quad i = 1, ..., m, \;\; n = 0, 1, ... \tag{8.14}$$

wird in Analogie zum linearen Fall (vgl. Abschnitt 5) auch als *nichtlineares Gesamtschrittverfahren (Jacobi-Iteration)* bezeichnet.

Beispiel 8.5.: (Erweiterung von Beispiel 1.3)
Die Diskretisierung des Randwertproblems $-u''(t) = f(t, u(t)),\; 0 < t < 1$ mit $u(0) = u(1) = 0$ (d.h. eine nichtlineare Erweiterung des Beispiels 1.3) führt auf das nichtlineare System

$$Ax = F(x)$$

mit

$$A = \text{tridiag}(-1, 2, -1), \quad x = (x_1, ..., x_m)^T, \quad F(x) = \frac{1}{m^2}(f(t_1, x_1), ..., f(t_m, x_m))^T.$$

Dabei ist $x_i, i = 1, \ldots, m$ eine Approximation an $u(t_i), i = 1, \ldots, m$. Das Gesamtschrittverfahren (GSV) geht aus von der Fixpunktform

$$x = g(x) := -A_D^{-1}(A_L + A_U)x + A_D^{-1}F(x), \;\; A = A_D + A_L + A_U.$$

Dann gilt

$$g'(x) = -A_D^{-1}(A_L + A_U) + A_D^{-1}F'(x)$$

mit

$$F'(x) = \frac{1}{m^2}\text{diag}\left(\frac{\partial f}{\partial x}(t_1, x_1), ..., \frac{\partial f}{\partial x}(t_m, x_m)\right).$$

Nach Beispiel 6.19 und Satz 4.23 (über den Spektralradius) ist

$$\| - A_D^{-1}(A_L + A_U)\|_2 = \rho(-A_D^{-1}(A_L + A_U)) = \cos\left(\frac{\pi}{m+1}\right).$$

Ferner gilt

$$\|F'(x)\|_2 \le \frac{c}{m^2}, \quad c := \sup_{t\in[0,1], x\in\mathbb{R}} \left|\frac{\partial f}{\partial x}(t, x)\right|,$$

damit

$$\|g'(x)\|_2 \le \cos\left(\frac{\pi}{m+1}\right) + \frac{c}{m^2}.$$

Für hinreichend großes m (d.h. hinreichend feine Diskretisierung des Randwertproblems mit $h = 1/(m+1)$) und $c < \pi^2$ findet man unter Beachtung des Taylorschen Satzes, daß $\|g'(x)\|_2 < 1$. Damit ist der Satz 8.3 anwendbar, d.h. das nichtlineare GSV konvergiert. □

Statt des nichtlinearen GSV kann man auch das *nichtlineare Einzelschrittverfahren (Gauß-Seidel Iteration)* betrachten:

$$\begin{aligned} x_1^{(n+1)} &:= g_1\left(x_1^{(n)}, x_2^{(n)}, ..., x_m^{(n)}\right) \\ x_2^{(n+1)} &:= g_2\left(x_1^{(n+1)}, x_2^{(n)}, ..., x_m^{(n)}\right) \\ &\vdots \quad \vdots \quad \vdots \\ x_m^{(n+1)} &:= g_m\left(x_1^{(n+1)}, x_2^{(n+1)}, ..., x_{m-1}^{(n+1)}, x_m^{(n)}\right). \end{aligned}$$

Schließlich können zur Konvergenzbeschleunigung wieder Relaxationsvarianten der genannten Verfahren verwendet werden. Hinsichtlich weiterer Details und hinreichender Konvergenzaussagen sei verwiesen auf

- J.M. Ortega, W.C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press 1970,
- G. Maeß: *Vorlesungen über Numerische Mathematik II*, Abschn. 5.3.1, Akademie-Verlag, Berlin 1988 bzw.
- H. Schwetlick: *Numerische Lösung nichtlinearer Gleichungen*, Verlag der Wissenschaften, Berlin 1979.

8.3 Newton-Verfahren

Für nichtlineare Gleichungssysteme

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix} = \begin{pmatrix} f_1(x_1, ..., x_m) \\ \vdots \\ f_m(x_1, ..., x_m) \end{pmatrix} = 0 \tag{8.15}$$

mit stetig differenzierbarer Funktion f betrachten wir die Näherung

$$f(x) \approx f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)}) =: g(x) \tag{8.16}$$

mit

$$f'(x) = \left(\frac{\partial f_i}{\partial x_k}(x)\right)_{i,k=1}^{m}.$$

Sei $x^{(1)}$ Lösung von $g(x) = 0$ und somit auch Lösung des linearen (!) Systems

$$f(x^{(0)}) + f'(x^{(0)})(x^{(1)} - x^{(0)}) = 0$$

bzw.

$$x^{(1)} = x^{(0)} - \left[f'(x^{(0)})\right]^{-1} f(x^{(0)}).$$

Sukzessive Wiederholung führt auf das Newton-Verfahren.

Definition 8.6. *Seien $G \subset \mathbb{R}^m$ offen und $f : G \to \mathbb{R}^m$ eine stetig differenzierbare Funktion mit einer für alle $x \in G$ nichtsingulären Jacobischen Funktionalmatrix $f'(x)$. Dann heißt das Iterationsverfahren*

$$x^{(n+1)} := x^{(n)} - \left[f'(x^{(n)})\right]^{-1} f(x^{(n)}), \;\; n = 0, 1, 2, ... \tag{8.17}$$

mit Startvektor $x^{(0)} \in G$ Newton-Verfahren *zur Lösung von $f(x) = 0$.*

In jedem Schritt ist also ein lineares Gleichungssystem

$$f'(x^{(n)})w^{(n)} = -f(x^{(n)}), \;\; n = 0, 1, 2, ... \tag{8.18}$$

mit Aufdatierung

$$x^{(n+1)} = x^{(n)} + w^{(n)}, \;\; n = 0, 1, 2, ... \tag{8.19}$$

zu lösen. Die Berechnung der aktuellen Jacobischen Funktionalmatrix $f'(x^{(n)})$ ist natürlich sehr aufwendig bei großen Werten von m.

Wir beweisen nun einen Satz zur lokalen Konvergenz des Newton-Verfahrens.

Satz 8.7. *Seien $G \subset \mathbb{R}^m$ offen und konvex und $f : G \to \mathbb{R}^m$ stetig differenzierbar. Für $x^{(0)} \in G$ genüge f in einer (beliebigen) Norm $\|\cdot\|$ auf $\mathbb{R}^m$ folgenden Bedingungen:*

(i) *Es existiert eine Zahl $\gamma > 0$ mit* $\|f'(x) - f'(y)\| \le \gamma\|x - y\|, \qquad \forall x, y \in G.$

(ii) *Für alle $x \in G$ existieren $[f'(x)]^{-1}$ und eine Zahl $\beta > 0$, so daß*

$$\| \left[f'(x)\right]^{-1} \| \le \beta, \qquad \forall x \in G.$$

(iii) *Mit* $\alpha := \| \left[f'(x^{(0)})\right]^{-1} f(x^{(0)})\|$ *gilt:* $\rho := \alpha\beta\gamma < \frac{1}{2}$.

(iv) *Mit* $r := 2\alpha$ *gelte:* $B[x^{(0)}; r] := \{x : \|x - x^{(0)}\| \le r\} \subset G.$

Dann gelten folgende Aussagen:

1. *Das Newton-Verfahren*

$$x^{(n+1)} := x^{(n)} - \left[f'(x^{(n)})\right]^{-1} f(x^{(n)}), \qquad n \in \mathbb{N}_0$$

   ist mit dem Startvektor $x^{(0)}$ wohldefiniert.

2. *Die Lösungsfolge $(x^{(n)})$ konvergiert gegen eine Nullstelle x^* von f mit*

$$\|x^{(n)} - x^*\| \le 2\alpha\rho^{2^n - 1}, \qquad n \in \mathbb{N}_0.$$

   In der Kugel $B(x^{(0)}; r)$ ist x^* die einzige Nullstelle von f.

Beweis. a) *Vorbereitender Schritt:* Wir beginnen mit einer Anwendung des Mittelwertsatzes (vgl. Satz 8.2). Aus dessen Beweis ergab sich

$$f(y) - f(x) = \int_0^1 f'(\lambda x + (1 - \lambda)y)(y - x)d\lambda.$$

Daraus ergibt sich mittels Nullergänzung

$$f(y) - f(x) - f'(z)(y-x) = \int_0^1 \{f'(\lambda x + (1-\lambda)y) - f'(z)\}(y-x)d\lambda$$

und durch (8.6) (vgl. Beweis von Satz 8.2) sowie Voraussetzung (i) und Integration

$$\begin{aligned} \|f(y) - f(x) - f'(z)(y-x)\| &\leq \gamma\|y-x\| \int_0^1 \|\lambda(x-z) + (1-\lambda)(y-z)\|d\lambda \\ &\leq \frac{1}{2}\gamma\|y-x\| \left(\|x-z\| + \|y-z\|\right). \end{aligned}$$

Mit $z = x$ ergibt sich

$$\|f(y) - f(x) - f'(x)(y-x)\| \leq \frac{\gamma}{2}\|y-x\|^2, \quad \forall x, y \in G. \tag{8.20}$$

Im Beweisschritt e) benötigen wir folgende Abschätzung, die mit der Wahl $z = x^{(0)}$ folgt

$$\|f(y) - f(x) - f'(x^{(0)})(y-x)\| \leq r\gamma\|y-x\|, \quad \forall x, y \in B[x^{(0)}; r]. \tag{8.21}$$

b) *Wohldefiniertheit des Verfahrens:* Wir zeigen hierzu und in Vorbereitung des Beweises der Cauchy-Konvergenz der Lösungsfolge mittels vollständiger Induktion, daß für die Lösungsfolge $(x^{(n)})$ gilt

$$\|x^{(n)} - x^{(0)}\| < r, \quad \|x^{(n)} - x^{(n-1)}\| \leq \alpha\rho^{2^{n-1}-1}, \ \ n \in \mathbb{N}. \tag{8.22}$$

Induktionsanfang: Für $n = 1$ gilt wegen Voraussetzung (iii)

$$\|x^{(1)} - x^{(0)}\| = \left\| \left[f'(x^{(0)})\right]^{-1} f(x^{(0)}) \right\| = \alpha = \frac{r}{2} < r.$$

Induktionsbeweis: Sei die Induktionsbehauptung (8.22) bis zu einer Zahl $n \in \mathbb{N}$ richtig. Wegen Voraussetzung (ii) und $x^{(n)} \in B(x^{(0)}, r) \subset G$ ist das nächste Folgenglied $x^{(n+1)}$ wohldefiniert. Unter Beachtung von Voraussetzung (ii), (8.20), der Induktionsannahme, von Voraussetzung (iii) sowie der Definition von ρ schließen wir

$$\begin{aligned} \|x^{(n+1)} - x^{(n)}\| &= \left\| \left[f'(x^{(n)})\right]^{-1} f(x^{(n)}) \right\| \ \leq \ \beta\|f(x^{(n)})\| \\ &= \beta \left\| f(x^{(n)}) - f(x^{(n-1)}) - f'(x^{(n-1)})(x^{(n)} - x^{(n-1)}) \right\| \\ &\leq \frac{1}{2}\beta\gamma\|x^{(n)} - x^{(n-1)}\|^2 \ \leq \ \frac{1}{2}\beta\gamma \left[\alpha\rho^{2^{n-1}-1}\right]^2 \\ &= \frac{1}{2}\alpha\rho^{2^n-1} \ < \ \alpha\rho^{2^n-1}. \end{aligned}$$

Dreiecksungleichung, die gerade gezeigte Abschätzung und die Definition von r zeigen nun

$$\begin{aligned} \|x^{(n+1)} - x^{(0)}\| &\leq \|x^{(n+1)} - x^{(n)}\| + ... + \|x^{(1)} - x^{(0)}\| \\ &\leq \alpha(1 + \rho + \rho^3 + \rho^7 + ... + \rho^{2^n-1}) \\ &< \frac{\alpha}{1-\rho} \ \leq \ 2\alpha \ = \ r. \end{aligned}$$

Damit ist der Induktionsbeweis für (8.22) erbracht.

c) *Existenz des Grenzwertes und Fehlerabschätzung:* Für $k \in \mathbb{N}$ folgt über die Dreiecksungleichung und (8.22) sowie wegen $\rho < \frac{1}{2}$, daß

$$\begin{aligned} \|x^{(n)} - x^{(n+k)}\| &\leq \|x^{(n)} - x^{(n+1)}\| + ... + \|x^{(n+k-1)} - x^{(n+k)}\| \\ &\leq \alpha(\rho^{2^n-1} + \rho^{2^{n+1}-1} + ... + \rho^{2^{n+k-1}-1}) \\ &= \alpha\rho^{2^n-1}(1 + \rho^{2^n} + ... + (\rho^{2^n})^{2^{k-1}}) \\ &< 2\alpha\rho^{2^n-1} \;\to\; 0, \quad n \to \infty. \end{aligned} \tag{8.23}$$

Damit ist $(x^{(n)})$ Cauchy-Folge. Satz 5.2 zeigte die Vollständigkeit des $\mathbb{R}^m$, damit existiert

$$x^* = \lim_{n\to\infty} x^{(n)}.$$

Grenzübergang $n \to \infty$ in (8.22) ergibt $\|x^* - x^{(0)}\| \leq r$, somit $x^* \in B[x^{(0)}; r]$. Schließlich liefert der Grenzübergang $k \to \infty$ in (8.23) die zu zeigende Fehlerabschätzung.

d) *Nachweis, daß x^* Nullstelle von f ist*: Nach Definition des Newton-Verfahrens und Nullergänzung sowie Anwendung der Dreiecksungleichung in Verbindung mit Voraussetzung (i) folgern wir

$$\begin{aligned} \|f(x^{(n)})\| &= \left\| f'(x^{(n)})(x^{(n+1)} - x^{(n)}) \right\| \\ &\leq \left\| f'(x^{(n)}) - f'(x^{(0)}) + f'(x^{(0)}) \right\| \; \|x^{(n+1)} - x^{(n)}\| \\ &\leq \left\{ \gamma\|x^{(n)} - x^{(0)}\| + \|f'(x^{(0)})\| \right\} \; \|x^{(n+1)} - x^{(n)}\| \\ &\to 0, \quad n \to \infty, \end{aligned}$$

damit $\lim\limits_{n\to\infty} f(x^{(n)}) = 0$. Wegen der Stetigkeit von f gilt somit auch $f(x^*) = 0$.

e) *Eindeutigkeit der Nullstelle in B*: Wir betrachten hierzu die Funktion

$$g : B[x^{(0)}; r] \to \mathbb{R}^m, \qquad g(x) := x - \left[f'(x^{(0)}) \right]^{-1} f(x).$$

Ausgehend von der Identität

$$g(x) - g(y) = \left[f'(x^{(0)}) \right]^{-1} \Big(f(y) - f(x) - f'(x^{(0)})(y - x) \Big)$$

ergeben die Voraussetzungen (ii), (iii) sowie Aussage (8.21)

$$\|g(x) - g(y)\| \leq \beta\gamma r\|y - x\| \leq 2\rho\|y - x\|, \quad \forall y, x \in B[x^{(0)}; r].$$

Somit ist g wegen $\rho < \frac{1}{2}$ kontraktiv. Nach dem Fixpunktsatz von Banach hat dann g auf $B[x^{(0)}; r]$ höchstens einen Fixpunkt. Die zu zeigende Eindeutigkeit der Nullstelle von f folgt dann wegen der Äquivalenz der Fixpunktgleichung $x = g(x)$ zu $f(x) = 0$. □

Der folgende Satz zeigt den lokalen Konvergenzcharakter des Newton-Verfahrens.

Satz 8.8. *Sei $G \subseteq \mathbb{R}^m$ offen, $f : G \to \mathbb{R}^m$ zweifach stetig differenzierbar und x^* Nullstelle von*

f mit $\det f'(x^*) \neq 0$. *Dann gibt es ein* $\delta > 0$ *so, daß das Newton-Verfahren für jeden Startvektor* $x^{(0)}$ *mit* $\|x^{(0)} - x^*\| < \delta$ *gegen* x^* *konvergiert.*

Beweis: Wegen der Stetigkeit der zweiten partiellen Ableitungen kann der Mittelwertsatz 8.2 auf die Komponenten von f' angewendet werden. Dann existiert eine Zahl $\gamma > 0$ so, daß

$$\left\|f'(x) - f'(y)\right\| \leq \gamma\|x - y\|, \qquad \forall x, y \in B[x^*; R]. \tag{8.24}$$

in einer geeigneten abgeschlossenen Kugelumgebung $B[x^*; R] := \{x : \|x - x^*\| \leq R\} \subset G$ gilt. Wir gehen nun aus von der Identität

$$f'(x) = f'(x^*)\left\{I + \left[f'(x^*)\right]^{-1}\left[f'(x) - f'(x^*)\right]\right\} \equiv f'(x^*)(I - B).$$

Nach Abschätzung (8.24) erhalten wir

$$\begin{aligned} \|B\| &= \left\|[f'(x^*)]^{-1}(f'(x) - f'(x^*)\right\| \\ &\leq \left\|[f'(x^*)]^{-1}\right\|\gamma\|x - x^*\| \ \leq \ \left\|[f'(x^*)]^{-1}\right\|\gamma R. \end{aligned}$$

Durch geeignete Wahl von R folgt $\|B\| < 1$. Nach Satz 5.15 ist $I - B$ und damit $f'(x)$ invertierbar. Ferner gilt

$$\left\|\left[f'(x)\right]^{-1}\right\| \leq \beta, \qquad \forall x \in B[x^*; R]$$

mit geeigneter Konstante $\beta > 0$.
Wegen der Stetigkeit von f und $f(x^*) = 0$ findet man eine Zahl $\delta < \frac{1}{2}R$ derart, daß

$$\|f(x^{(0)})\| < \min\left\{\frac{R}{4\beta}; \frac{1}{2\beta^2\gamma}\right\}, \qquad \forall x^{(0)} \text{ mit } \|x^{(0)} - x^*\| < \delta.$$

Mit der Festlegung $\alpha := \|[f'(x^{(0)}]^{-1}f(x^{(0)})\|$ erhält man

$$\alpha\beta\gamma \leq \|f(x^{(0)})\|\beta^2\gamma < \frac{1}{2}; \qquad 2\alpha \leq 2\beta\|f(x^{(0)})\| < \frac{1}{2}R.$$

Für die offene und konvexe Kugel $G := B(x^*; R)$ und alle $x^{(0)}$ mit $\|x^{(0)} - x^*\| < \delta$ sind dann die Voraussetzungen von Satz 8.7 erfüllt. □

Eine einfache Anwendung von Satz 8.8 reproduziert nochmals das Ergebnis von Satz 7.12 für den skalaren Fall ($m = 1$).

Satz 8.9. *Sei* $f : (a, b) \to \mathbb{R}$ *zweimal stetig differenzierbar und* x^* *einfache Nullstelle von* f. *Dann existiert ein* $\delta > 0$ *so, daß das Newton-Verfahren bei beliebigem Startvektor* $x^{(0)}$ *mit* $|x^{(0)} - x^*| \leq \delta$ *gegen* x^* *konvergiert.*

Beweis. Für einfache Nullstellen ist $f'(x^*) \neq 0$ und damit Satz 8.8 anwendbar. □

Abschließend bestimmen wir die Konvergenzordnung des Newton-Verfahrens für nichtlineare Gleichungssysteme.

Definition 8.10. *Die Folge* $(x^{(n)})$ *auf dem normierten Raum* X konvergiert von der Ordnung $p \geq 1$ gegen x, *falls eine Zahl* C *existiert (für* $p = 1$ *mit* $C \in [0, 1)$*) mit*

$$\|x^{(n+1)} - x\| \leq C\|x^{(n)} - x\|^p, \quad n = 1, 2, ... \tag{8.25}$$

Satz 8.11. *Unter den Voraussetzungen von Satz 8.7 konvergiert das Newton-Verfahren von 2. Ordnung.*

Beweis: *Übungsaufgabe !* □

Anhand der Beispiele 7.5 und 7.6 prüft man nach, daß für das Newton-Verfahren tatsächlich jeweils quadratische Konvergenz vorliegt.

8.4 Newton-ähnliche Verfahren

Die Berechnung der Jacobi-Matrix in jedem Schritt des Newton-Verfahrens ist im mehrdimensionalen Fall (insbesondere bei $m \gg 1$ viel zu aufwendig. Man sucht daher wie im skalaren Fall ($m = 1$) nach Vereinfachungen.

Für das *vereinfachte Newton-Verfahren* (vgl. auch Abschnitt 7.4)

$$x^{(n+1)} := x^{(n)} - \left[f'(x^{(0)})\right]^{-1} f(x^{(n)}), \qquad n \in \mathbb{N}_0$$

kann man beweisen, daß es unter den Voraussetzungen von Satz 8.7 nur linear gegen die (lokal eindeutig bestimmte) Nullstelle. Dies wird dem Leser als Übungsaufgabe überlassen.

Auch für das Sekanten-Verfahren findet man geeignete Verallgemeinerungen im mehrdimensionalen Fall, vgl. z.B. *Ortega/Rheinboldt*). Man kann jedoch wiederum nur lineare Konvergenz erwarten.

Bei modifizierten Newton-Verfahren bestimmt man Näherungen an die inverse Jacobi-Matrix $[f'(x^{(n)}]^{-1}$ derart, daß überlineare Konvergenz

$$\|x^{(n+1)} - x^*\| \leq C_n \|x^{(n)} - x^*\|, \quad n \in \mathbb{N}_0$$

mit $C_n \to 0, n \to \infty$ bei geringeren Kosten als für das vollständige Newton-Verfahren erzielt wird. Eine wichtige Klasse bilden die Broyden-Verfahren, vgl. z.B. *Ortega/Rheinboldt*).

Kapitel 9

Polynomiale Interpolation

Im Teil II der Vorlesung behandeln wir die Probleme der numerischen Interpolation und Integration im univariaten bzw. eindimensionalen Fall. (Betreffs der Erweiterung auf den multivariaten bzw. mehrdimensionalen Fall wird auf weiterführende Veranstaltungen verwiesen.)

Eindimensionale Interpolationsaufgaben behandeln folgende Fragestellung:

Gegeben sind $n+1$ *Paare* (x_i, y_i), $i = 0, 1, ..., n$ *reeller Zahlen mit* $x_i \neq x_k, i \neq k$ *und eine Klasse reellwertiger Funktionen* $\Phi(\cdot; c_0, ..., c_n)$, *die von* $n+1$ *Parametern* $c_0, ..., c_n$ *abhängen. Gesucht sind die Parameterwerte* $c_0, ..., c_n$ *mit der (Interpolations-) Eigenschaft*

$$\Phi(x_i; c_0, ..., c_n) = y_i, \quad i = 0, 1, ..., n. \tag{9.1}$$

Dies entspricht der Lösung eines *nichtlinearen* Gleichungssystems. Speziell erhält man ein *lineares* System, wenn die Parameter $c_0, ..., c_n$ nur linear in Φ eingehen. Mit einer geeigneten Basis $\{\phi_0, \ldots, \phi_n\}$ gelte dann

$$\Phi(x; c_0, ..., c_n) := \sum_{i=0}^{n} c_i \phi_i(x). \tag{9.2}$$

Wir gehen hier nur auf diesen *linearen* Fall ein. Aus historischer Sicht ist die *polynomiale Interpolation* der wichtigste Spezialfall. In der Monom-Basis mit $\phi_i(x) = x^i$ ergibt sich zum Beispiel

$$\Phi(x; c_0, ..., c_n) := \sum_{i=0}^{n} c_i x^i.$$

Zunächst wurde diese Interpolationsform vorwiegend benutzt, um in Tafelwerken angegebene diskrete Funktionswerte zu interpolieren, vor allem durch lineare und quadratische Funktionen. Im Zeitalter elektronischer Rechentechnik hat das nur noch geringe Bedeutung, da die Funktionswerte durch schnelle Routinen verfügbar sind. Hingegen ist die polynomiale Interpolation nach wie vor wesentlich zur Begründung anderer numerischer Verfahren (z.B. numerische Integration, vgl. Abschn. 13-14, Differentiation und Näherungslösung von Differentialgleichungen, vgl. *Numerische Mathematik* II, oder allgemeiner von Operatorgleichungen).

Bei der Interpolation von *Meßreihen* $(x_i, y_i), i = 0, ..., n$ ist die polynomiale Interpolation vor allem bei größeren Werten von n wenig geeignet, da das interpolierende Polynom oft an den Intervallenden stark oszilliert. Abhilfe schafft hier die *Interpolation durch Splines*, d.h. die Annäherung durch stückweise polynomiale Funktionen (vgl. Kap. 11). *Trigonometrische Polynome* spielen bei der Interpolation periodischer Daten eine zentrale Rolle (vgl. Kap. 10).

Bei den einzelnen Interpolationsansätzen (vgl. Kap. 9-12) behandeln wir Existenz- und Eindeutigkeitsaussagen, dann die praktische Darstellung und Berechnung und teilweise Fehler- und Konvergenzaussagen.

9.1 Lagrangesche Interpolation

Sei Π_n der lineare Raum der Polynome vom Grad kleiner oder gleich n als Unterraum des linearen Raumes $C[a,b]$ der stetigen Funktionen auf $[a,b]$ mit $a < b$.

Wir werden mehrfach Gebrauch von folgender Eigenschaft algebraischer Polynome machen.

Satz 9.1. *(i) Jedes Polynom aus Π_n, $n \in \mathbb{N}_0$, das mehr als n (komplexe) Nullstellen (bei Berücksichtigung ihrer Vielfachheit) besitzt, verschwindet identisch.*
(ii) Die Monome $\phi_i(x) = x^i, i = 0, ..., n$ bilden eine Basis des Raumes Π_n, d.h.

$$\dim(\Pi_n) = n + 1, \qquad \Pi_n = \text{span}\{1, x, ..., x^n\}.$$

Beweis: Aussage (i) beweist man, ausgehend vom Fundamentalsatz der Algebra, am einfachsten per Induktion über n. Aussage (ii) ist dann Folgerung von Aussage (i). *(Übungsaufgabe !)*

Die *Lagrangesche Interpolationsaufgabe* lautet: Bestimme ein Polynom $P_n \in \Pi_n$ zu gegebenen Stützstellen $x_0, ..., x_n \in \mathbb{R}$ mit $x_i \neq x_k, i \neq k$ und Funktionswerten $y_0, ..., y_n$, so daß

$$P_n(x_i) = y_i, \quad i = 0, ..., n. \tag{9.3}$$

Satz 9.2. *Die Lagrangesche Interpolationsaufgabe ist eindeutig lösbar. P_n hat die* Lagrange-Darstellung

$$P_n(x) = \sum_{k=0}^{n} y_k l_k(x) \tag{9.4}$$

mit den Lagrange-Polynomen

$$l_k(x) := \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}, \quad k = 0, ..., n. \tag{9.5}$$

Beweis: Per Ansatz gilt $P_n \in \Pi_n$ sowie

$$l_k(x_i) = \delta_{ik} = \begin{cases} 0, & k \neq i \\ 1, & k = i \end{cases}.$$

Damit ist die Interpolationsbedingung erfüllt.

Zum Beweis der Eindeutigkeit seien $p_1, p_2 \in \Pi_n$ zwei interpolierende Polynome. Für deren Differenz $p := p_1 - p_2$ ist $p(x_i) = 0, \ i = 0, ..., n$, d.h. das Polynom $p \in \Pi_n$ hat $n + 1$ Nullstellen und verschwindet daher identisch nach Satz 9.1 (i). □

Der Ansatz von Lagrange liefert einen wesentlich eleganteren Beweis als der naive Ansatz

$$P_n(x) := \sum_{k=0}^{n} c_k x^k$$

in der Monom-Basis. Dieser führt auf das inhomogene lineare System

$$\sum_{k=0}^{n} c_k x_j^k = y_j, \quad j = 0, ..., n$$

mit nichtsingulärer Koeffizientenmatrix $(x_j^k)_{0\le j,k\le n}$ *(Übungsaufgabe !).*

Die Lagrange-Darstellung von Interpolationspolynomen eignet sich wegen des einfachen symmetrischen Aufbaus gut für theoretische Zwecke. Für die praktische Berechnung ist sie wenig geeignet, da z.B. die Hinzunahme weiterer Stützstellen eine völlige Neuberechnung erfordert. Ein weiteres Problem ist, daß die Lagrange-Polynome für große Werte von n stark anwachsen und oszillieren können. In diesem Sinne ist das Problem der Lagrange-Interpolation schlecht konditioniert.

Beispiel 9.3. MAPLE stellt die Funktion `interp` zur polynomialen Interpolation bereit. Wir benutzen sie, um an den Stellen datax $:= [1,2,3,4,5,6,7,8,9,10]$ mit den zufällig erzeugten Werten datay $:= [1,8,3,3,5,3,9,7,7,9]$ ein Polynom 9. Grades zu erzeugen. Man erhält prompt

$$\begin{aligned} P(x) &= -\frac{17}{10368}x^9 + \frac{51}{640}\,x^8 - \frac{99979}{60480}\,x^7 + \frac{18353}{960}\,x^6 - \frac{2333359}{17280}\,x^5 \\ &\quad + \frac{1151497}{1920}\,x^4 - \frac{1344587}{810}\,x^3 + \frac{1310431}{480}\,x^2 - \frac{6016069}{2520}\,x + 836 \end{aligned}$$

Abb. 9.1 zeigt, daß das Polynom an den Intervallenden zum "Überschwingen" neigt.

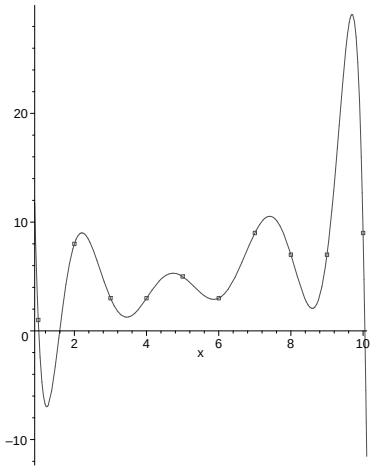

Abbildung 9.1: Interpolation eines Polynoms 9. Grades mit Zufallsdaten

9.2 Newtonsche Interpolation

Die Idee der *Newton-Darstellung* des Interpolationspolynoms besteht darin, die Funktionen $\{n_0, ..., n_n\}$ mit

$$n_k(x) := \prod_{i=0}^{k-1}(x - x_i) \in \Pi_k \tag{9.6}$$

als Basis des Π_n zu nutzen. Das führt über den Ansatz

$$N_n(x) := \sum_{k=0}^{n} c_k n_k(x).$$

auf das lineare Gleichungssytem

$$\sum_{k=0}^{n} c_k n_k(x_j) = y_j, \quad j = 0, ..., n. \tag{9.7}$$

Wegen $n_k(x_j) = 0$ für $k > j$ ist die Koeffizientenmatrix eine untere Dreiecksmatrix, die ferner nichtverschwindende Hauptdiagonalelemente hat. Damit können die gesuchten Werte $c_0, ..., c_n$ sukzessiv berechnet werden.

Eine systematische Vorgehensweise erhält man über die dividierten Differenzen.

Definition 9.4. *Seien $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 0, ..., n$ gegeben mit $x_i \neq x_k, i \neq k$. Dann sind die* dividierten Differenzen D_i^k der Ordnung k an der Stelle x_i *rekursiv definiert durch*

$$D_i^0 := y_i, \qquad i = 0, ..., n, \tag{9.8}$$

$$D_i^k := \frac{D_{i+1}^{k-1} - D_i^{k-1}}{x_{i+k} - x_i}, \quad i = 0, ..., n-k, \;\; k = 1, ..., n. \tag{9.9}$$

Sinnvoll ist die Anordnung der dividierten Differenzen in folgendem Differenzenschema (hier $n = 3$) mit einem Aufwand von $0(n^2)$ wesentlichen Rechenoperationen:

$$\begin{array}{c|l} x_0 & y_0 = D_0^0 \\ & \qquad D_0^1 \\ x_1 & y_1 = D_1^0 \qquad D_0^2 \\ & \qquad D_1^1 \qquad D_0^3 \\ x_2 & y_2 = D_2^0 \qquad D_1^2 \\ & \qquad D_2^1 \\ x_3 & y_3 = D_3^0 \end{array} \tag{9.10}$$

Schema der dividierten Differenzen im Fall $n = 3$

Beispiel 9.5. Gegeben seien die Paare $(0, 1), (1, 3), (3, 2)$. Dann lautet das Differenzenschema

$$\begin{array}{c|l} 0 & 1 \\ & \qquad 2 \\ 1 & 3 \qquad -\frac{5}{6} \\ & \qquad -\frac{1}{2} \\ 3 & 2 \end{array}$$

. □

Lemma 9.6. *Für die dividierten Differenzen gilt*

$$D_i^k = \sum_{j=i}^{i+k} y_j \left(\prod_{\substack{r=i \\ r \neq j}}^{i+k} \frac{1}{x_j - x_r} \right), \quad i = 0, ..., n-k, \quad k = 0, ..., n. \tag{9.11}$$

Beweis: Wir führen den Induktionsbeweis nach der Ordnung k durch. Die Behauptung ist per Definition richtig für $k = 0$.

Sei nun die Aussage für $k - 1$ bewiesen. Über die Definition dividierter Differenzen und die Induktionsannahme ergibt sich

$$D_i^k = \frac{1}{x_{i+k} - x_i} \left\{ \sum_{j=i+1}^{i+k} y_j \left(\prod_{\substack{r=i+1 \\ r \neq j}}^{i+k} \frac{1}{x_j - x_r} \right) - \sum_{j=i}^{i+k-1} y_j \left(\prod_{\substack{r=i \\ r \neq j}}^{i+k-1} \frac{1}{x_j - x_r} \right) \right\}$$

$$
\begin{aligned}
&= \frac{1}{x_{i+k}-x_i}\sum_{j=i+1}^{i+k-1} y_j\left\{\frac{1}{x_j-x_{i+k}}-\frac{1}{x_j-x_i}\right\}\prod_{\substack{r=i+1\\ r\neq j}}^{i+k-1}\frac{1}{x_j-x_r}\\
&\quad +\frac{1}{x_{i+k}-x_i}y_{i+k}\prod_{r=i+1}^{i+k}\frac{1}{x_{i+k}-x_r}-\frac{1}{x_{i+k}-x_i}y_i\prod_{r=i+1}^{i+k-1}\frac{1}{x_i-x_r}\\
&= \sum_{j=i+1}^{i+k-1} y_j\frac{1}{(x_j-x_{i+k})(x_j-x_i)}\prod_{\substack{r=i+1\\ r\neq j}}^{i+k-1}\frac{1}{x_j-x_r}\\
&\quad +y_{i+k}\prod_{r=i}^{i+k}\frac{1}{x_{i+k}-x_r}+y_i\prod_{r=i+1}^{i+k}\frac{1}{x_i-x_r}\\
&= \sum_{j=i}^{i+k} y_j\left(\prod_{\substack{r=i\\ r\neq j}}^{i+k}\frac{1}{x_j-x_r}\right). \qquad \square
\end{aligned}
$$

Satz 9.7. *Das nach Satz 9.2 eindeutig bestimmte Interpolationspolynom lautet in Newtonscher Darstellung*

$$N_n(x) = y_0 + \sum_{k=1}^{n} D_0^k n_k(x) \;\equiv\; y_0 + \sum_{k=1}^{n} D_0^k\left[\prod_{r=0}^{k-1}(x-x_r)\right]. \tag{9.12}$$

Beweis: Wir führen den Induktionsbeweis nach der Polynomordnung n durch und verwenden die Eindeutigkeit der Lagrange-Darstellung nach Satz 9.2.

Offenbar ist die Aussage für $n = 1$ richtig. Sei nun die Behauptung richtig für $n-1$ mit $n \geq 2$. Dann betrachten wir die Differenz $d_n := P_n - N_n$. Wegen des rekursiven Aufbaus des Newton-Polynoms gilt

$$d_n(x) = P_n(x) - N_{n-1}(x) - D_0^n\prod_{r=0}^{n-1}(x-x_r).$$

Nach Lemma 9.6 und Satz 9.2 gilt

$$
\begin{aligned}
D_0^n\prod_{r=0}^{n-1}(x-x_r) &= \left(\sum_{j=0}^{n} y_j\prod_{\substack{r=0\\ r\neq j}}^{n}\frac{1}{x_j-x_r}\right)\prod_{r=0}^{n-1}(x-x_r)\\
&= y_n\prod_{r=0}^{n-1}\frac{x-x_r}{x_n-x_r}+s_{n-1}(x) \;=\; y_n l_n(x)+s_{n-1}(x)
\end{aligned}
$$

mit geeignetem $s_{n-1} \in \Pi_{n-1}$. Daher verschwindet der Koeffizient vor x^n im Polynom d_n, d.h. $d_n \in \Pi_{n-1}$. Nach Induktionsannahme ist

$$N_{n-1}(x_i) = y_i = P_n(x_i), \quad i = 0, ..., n-1,$$

daher

$$d_n(x_i) = 0, \quad i = 0, ..., n-1.$$

Damit muß $d_n \in \Pi_{n-1}$ nach Satz 9.1 identisch verschwinden, also gilt $N_n = P_n$. $\square$

Beispiel 9.8. Das Newtonsche Interpolationspolynom zum Beispiel 9.5 lautet

$$N_2(x) = 1 + 2x - \frac{5}{6}x(x-1). \qquad \square$$

Die Newton-Darstellung des Interpolationspolynoms hat gegenüber der Darstellung von Lagrange den wesentlichen Vorteil, daß bei Hinzunahme eines neuen Punktepaares (x_{n+1}, y_{n+1}) die vorherige Rechnung nicht überflüssig wird.

Wir betrachten nun den *Aufwand der Funktionswertberechnung* mittels des Newtonschen Interpolationspolynoms. Setzt man wie beim Horner-Schema in geeigneter Weise Klammern

$$\begin{aligned} N_n(x) &= c_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) + ... + c_1(x-x_0) + c_0 \\ &= (...(c_n(x-x_{n-1}) + c_{n-1})(x-x_{n-2}) + ... + c_1)(x-x_0) + c_0, \end{aligned}$$

so erhält man einen Funktionswert durch n Multplikationen und $2n$ Additionen. Die Berechnung der Koeffizienten $c_0, ..., c_n$ erfordert selbst $\frac{1}{2}n(n+1)$ Divisionen und $n(n+1)$ Additionen.

Man kann nun die Berechnung eines *einzelnen* Wertes des Interpolationspolynoms sogar ohne explizite Ermittlung der Polynomkoeffizienten vornehmen. Das entstehende *Schema von Neville* ist dem der Berechnung der dividierten Differenzen sehr ähnlich.

Satz 9.9. *Seien $n+1$ Paare $(x_i, y_i), i = 0, ..., n$ mit $x_i \neq x_k, i \neq k$ und $y_i \in \mathbb{R}$ gegeben. Dann gilt für die eindeutig bestimmten Interpolationspolynome*

$$P_i^k \in \Pi_k, \;\; i = 0, ..., n-k, \;\; k = 0, ..., n$$

mit der Interpolationseigenschaft

$$P_i^k(x_j) = y_j, \quad j = i, ..., i+k$$

die Rekursionsbeziehung

$$P_i^0(x) = y_i, \tag{9.13}$$

$$P_i^k(x) = \frac{(x-x_i)P_{i+1}^{k-1}(x) - (x-x_{i+k})P_i^{k-1}(x)}{x_{i+k} - x_i}, \quad k = 1, ..., n. \tag{9.14}$$

Beweis: Wir führen den Induktionsbeweis nach der Polynomordnung k durch. Für $k = 1$ ist die Aussage richtig. Die Induktionsbehauptung sei nun bewiesen für $k-1$ mit $k \geq 2$. Das durch die zweite Rekursionsbeziehung im Satz definierte Polynom p_k gehört zu Π_k. Weiter gilt nach Induktionsannahme

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

und

$$p_k(x_i) = y_i, \quad p_k(x_{i+k}) = y_{i+k}.$$

Daraus folgt die Behauptung. Der gesuchte Interpolationswert ist $p_n(x) = P_0^n(x)$. $\square$

9.3 Interpolationsfehlerabschätzungen

Wir betrachten die Interpolation einer gegebenen („komplizierten") stetigen Funktion $f : [a, b] \to \mathbb{R}$ durch das („einfachere") Interpolationspolynom $(L_n f)(\cdot) \in \Pi_n$ mit der Interpolationseigenschaft

$$(L_n f)(x_i) = f(x_i), \quad i = 0, ..., n. \tag{9.15}$$

Der Operator $L_n : C[a, b] \to \Pi_n$ mit $f \mapsto L_n f$ heißt auch *Lagrange-Interpolationsoperator.*

Verschiedentlich (z.B. bei der numerischen Integration) benötigt man Aussagen über den *Interpolationsfehler* $f - L_n f$.

Satz 9.10. *Sei $f : [a, b] \to \mathbb{R}$ $(n + 1)-$mal stetig differenzierbar. Dann hat das Restglied $R_n f := f - L_n f$ bei der Polynominterpolation an $n + 1$ paarweise verschiedenen Stützstellen $x_0, ..., x_n \in [a, b]$ die Lagrangesche Darstellung*

$$(R_n f)(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!} \prod_{k=0}^{n} (x - x_k), \quad x \in [a, b] \tag{9.16}$$

mit einer von der Entwicklungsstelle x abhängigen Stelle $\xi \in [a, b]$.

Beweis: Ist x selbst Stützstelle, so ist die Aussage des Satzes erfüllt. Sei nun

$$\omega_{n+1}(x) := \prod_{k=0}^{n} (x - x_k) \tag{9.17}$$

und für festes, aber beliebiges $x \in [a, b], x \neq x_k,\ k = 0, ..., n$ eine Hilfsfunktion $g : [a, b] \to \mathbb{R}$ definiert durch

$$g(y) := f(y) - (L_n f)(y) - \omega_{n+1}(y) \frac{f(x) - (L_n f)(x)}{\omega_{n+1}(x)},\ \ y \in [a, b]. \tag{9.18}$$

Nach den Voraussetzungen an f ist g $(n + 1)-$mal stetig differenzierbar. Ferner hat g die $n + 2$ Nullstellen $x, x_0, ..., x_n$. Der *Satz von Rolle* besagt nun, daß die Ableitung g' zwischen zwei Nullstellen von g wieder eine Nullstelle besitzt. Damit hat g' $n + 1$ paarweise verschiedene Nullstellen auf $[a, b]$. Sukzessive Wiederholung dieses Arguments ergibt, daß $g^{(n+1)}$ eine Nullstelle ξ in $[a, b]$ hat. Man berechnet

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n + 1)! \frac{(R_n f)(x)}{\omega_{n+1}(x)}$$

und erhält die Behauptung. □

Da die Zwischenstelle ξ im allgemeinen Fall nicht explizit berechnet werden kann, schätzt man oft den Interpolationsfehler ab durch

Korollar 9.11. *Unter den Voraussetzungen von Satz 9.10 ist*

$$\|f - L_n f\|_\infty \leq \frac{1}{(n + 1)!} \|\omega_{n+1}\|_\infty \|f^{(n+1)}\|_\infty \tag{9.19}$$

mit $\omega_{n+1}(x) := \prod_{i=0}^{n} (x - x_i)$ und $\|g\|_\infty := \max_{x \in [a,b]} |g(x)|$.

Beispiel 9.12. Sei $f(x) = \sin x$ auf dem Intervall $[0, 1]$. Mit der *Maple*-Funktion `interp`

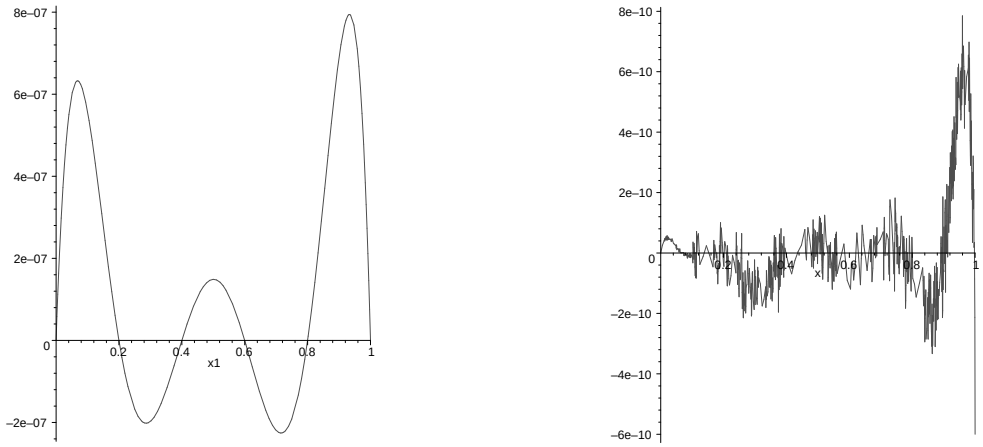

Abbildung 9.2: Interpolationsfehler $f - L_n f$ für $f(x) = \sin x$ und a) $n = 5$ sowie b) $n = 10$

berechnen wir die Interpolationspolynome $L_5 f(x)$ und $L_{10} f(x)$ zu den äqudistanten Stützstellen $x_i = i/5,\ i = 0, ..., 5$ bzw. $x_i = i/10,\ i = 0, ..., 10$.

Die Abbildung 9.2 zeigt jeweils den Interpolationsfehler $f - L_5 f$ bzw. $f - L_{10} f$. Offenbar ist $f \in C^\infty[0, 1]$ mit $\|f^{(n)}\|_\infty \leq 1$. Eine sehr grobe Abschätzung ist $\|\omega_{n+1}\|_\infty \leq 1$. Damit folgt nach Satz 9.10

$$|f(x) - (L_n f)(x)| \leq \frac{1}{(n+1)!}.$$

Abb. 9.2 zeigt sogar kleinere Fehlerwerte. Mit wachsender Stützstellenzahl n wird er sogar "exponentiell klein". □

Bemerkung 9.13. Eine grobe Abschätzung in Korollar 9.11 ergibt $\|\omega_{n+1}\|_\infty \leq (b-a)^{n+1}$, mit $h := b - a$ gilt dann

$$\|f - L_n f\|_\infty \leq \frac{h^{n+1}}{(n+1)\ !} \|f^{n+1}\|_\infty.$$

Korollar 9.11 verdeutlicht in Verbindung mit dem Beispiel, daß bei hinreichend oft differenzierbaren ("glatten") Funktionen die Interpolation mit Polynomen hoher Ordnung sinnvoll sein kann. Eventuell zerlegt man dazu das Intervall $[a, b]$ in kleinere Intervalle und interpoliert f bei stetiger Verheftung an den Intervallenden stückweise durch Lagrange-Polynome. Das ist auch die Grundidee der in Anwendungen sehr wichtigen *Finite-Element-Methoden* höherer Ordnung.

Zu beachten ist jedoch, daß auch die Größen $\|f^{(n+1)}\|_\infty$ nicht beliebig stark bezüglich n anwachsen sollten. RUNGE untersuchte das Beispiel der auf $[a, b] = [-1, 1]$ beliebig oft differenzierbaren Funktion $f(x) = \frac{1}{1+25x^2}$. Man beobachtet bereits für moderate Werte von n die Divergenz der Interpolationspolynome an den Intervallenden. □

Bemerkung 9.14. Korollar 9.11 legt nahe, bei Wahlmöglichkeit der Stützstellen diese so festzulegen, daß

$$\max_{x \in [a,b]} \left| \prod_{k=0}^{n} (x - x_k) \right| \to Min.! \tag{9.20}$$

In der Approximationstheorie (vgl. *Numerische Mathematik* II) wird gezeigt, daß dabei die Stützstellen

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2(n-i)+1}{2(n+1)} \pi \right), \quad i = 0, ..., n \tag{9.21}$$

entstehen (Nullstellen der *Tschebychev-Polynome*). Bei dieser Wahl erreicht man auch eine wesentlich bessere Aproximation für das Beispiel von RUNGE. □

9.4 Konvergenz von Interpolationspolynomen

Wir betrachten nun die Konvergenz von Interpolationspolynomen bei wachsender Stützstellenzahl. Offenbar gilt

Satz 9.15. *Seien $f \in C^\infty[a,b]$ und $\|f^{(n)}\|_\infty \leq M$ für alle $n = 0, 1, ...$ Dann konvergiert der Interpolationsfehler $R_n f$ für $n \to \infty$ gleichmäßig auf $[a,b]$ gegen Null.*

Beweis: Die Behauptung folgt aus $|(R_n f)(x)| \leq M \frac{(b-a)^{n+1}}{(n+1)!} \to 0, \ n \to \infty.$ □

Dieser Satz ist zwar für viele Standardfunktionen, jedoch in der Regel nicht auf praktisch interessierende Funktionen anwendbar. Insbesondere ist die starke Glattheitsforderung an die zu interpolierende Funktion stark einschränkend.

Das folgende Beispiel zeigt, daß im Fall lediglich *stetiger* Funktionen der Fehler bei wachsender Stützstellenzahl sogar divergieren kann.

Beispiel 9.16. Sei

$$f(x) := \begin{cases} x \sin \frac{\pi}{x}, & x \in (0,1], \\ 0, & x = 0. \end{cases}$$

Mit $x_k = \frac{1}{k+1}, k = 0, ..., n$ ist wegen $f(x_k) = 0, k = 0, ..., n$ offenbar $(L_n f)(x) = 0$ für alle $n \in \mathbb{N}$. Die Folge der Interpolationspolynome konvergiert bei dieser Stützstellenwahl nur an den Stellen $x_k, k \in \mathbb{N}_0$ gegen die Funktion f. □

Zur Verdeutlichung führen wir noch ohne Beweis folgende Sätze über die Konvergenz von Interpolationspolynomen an.

Satz 9.17. *(Marcinkiewicz)*
Für jede Funktion $f \in C[a,b]$ gibt es eine Folge von Stützstellen $(x_k^{(n)}), k = 0, ..., n$ und $n \in \mathbb{N}_0$ derart, daß die entsprechende Folge $(L_n f)$ von Interpolationspolynomen $L_n f \in \Pi_n$ mit $(L_n f)(x_k^{(n)}) = f(x_k^{(n)}), \ k = 0, ..., n$ auf $[a,b]$ gleichmäßig gegen f konvergiert.

Satz 9.18. *(Faber)*
Für jede Folge $(x_k^{(n)})$ existiert eine Funktion $f \in C[a,b]$ so, daß die zugehörige Folge $(L_n f)$ von Interpolationspolynomen auf $[a,b]$ nicht gleichmäßig gegen f konvergiert.

Bei geeigneter Wahl der Stützstellen gilt jedoch wenigstens

Satz 9.19. *Sei $f \in C^1[a,b]$. Dann konvergiert die Folge der mit den Tschebychev-Stützstellen (vgl. Bemerkung 9.14) gebildeten Interpolationspolynome gleichmäßig auf $[a,b]$ gegen f.*

Insgesamt ergibt sich aus den Überlegungen dieses Abschnitts, daß die Interpolation mit Polynomen hohen Grades im allgemeinen Fall (insbesondere bei geringer Glattheit der zu interpolierenden Funktionen) nicht sinnvoll ist. Wir werden im Kapitel 11 über die Spline-Interpolation sehen, wie durch stückweise polynomiale Interpolation unter relativ geringen Glattheitsforderungen an die zu interpolierende Funktion gleichmäßige Konvergenz erzielt werden kann.

9.5 Verallgemeinerung

Wir betrachten die Interpolation mittels allgemeinerer Funktionensysteme.

Definition 9.20. *Ein $m-$dimensionaler Unterraum $U \subset C[a,b]$ heißt* unisolvent *bezüglich der paarweise verschiedenen Stützstellen $x_1, ..., x_m \in [a,b]$, wenn jede Funktion $u \in U$ mit den Nullstellen $u(x_i) = 0,\ i = 1, ...m$ identisch verschwindet.*

Im vorliegenden Kapitel hatten wir den Fall

$$U = \Pi_n[a,b] \subset C[a,b]$$

mit verschiedenen Basisfunktionen (Monom-Basis, Basis mit Lagrange- bzw. Newton-Polynomen) und $m = n + 1$ besprochen.

Satz 9.21. *Seien der $m-$dimensionale Unterraum $U \subset C[a,b]$ bezüglich der paarweise verschiedenen Stützstellen $x_1, ..., x_m \in [a,b]$ unisolvent und m Werte $y_1, ..., y_m \in \mathbf{R}$ gegeben. Dann existiert genau eine Funktion $u \in U$ mit der Interpolationseigenschaft $u(x_i) = y_i,\ \ i = 1, ..., m$.*

Beweis: Sei $U = \text{span}\{\phi_1, ..., \phi_m\}$. Mit dem Ansatz

$$u = \sum_{k=1}^{m} c_k \phi_k$$

führt die Interpolationsaufgabe auf das lineare Gleichungssystem

$$\sum_{k=1}^{m} c_k \phi_k(x_i) = y_i, \quad i = 1, ..., m. \tag{9.22}$$

Wegen der Bedingung der Unisolvenz an U hat das zugehörige homogene System nur die triviale Lösung. Folglich ist das Gleichungssystem eindeutig lösbar. □

Kapitel 10

Trigonometrische Interpolation

Bei Anwendungen in der Signal- und Steuerungstechnik treten in der Regel periodische Daten bzw. periodische Funktionen mit der Eigenschaft

$$\exists T > 0: \quad f(x+T) = f(x), \quad \forall x \in \mathbb{R} \tag{10.1}$$

auf. Unter dem Aspekt der modernen Kommunikationstechnik kommt es hier vor allem auf eine sehr effiziente Auswertung derartiger Daten an.

Algebraische Polynome (vgl. Kapitel 9) sind nicht zur Interpolation derartiger Daten geeignet, da sie nicht periodisch sind. Wir wenden uns daher *trigonometrischen Polynomen* zu. Bei äquidistanter Verteilung der Interpolationsstützstellen spricht man auch von *diskreter Fourier-Transformation.*

10.1 Trigonometrische Polynome

Wir werden ohne Beschränkung der Allgemeinheit annehmen, daß für die Periodenlänge gilt $T = 2\pi$. Dann ist die Betrachtung der folgenden Funktionensysteme sinnvoll.

Definition 10.1. *Die Elemente des Raumes*

$$T_n := \text{span}\{1, \cos x, \sin x, ..., \cos(nx), \sin(nx)\} \tag{10.2}$$

heißen trigonometrische oder Fourier-Polynome *vom Grad kleiner oder gleich* n.

Die Begriffsbildung wird durch folgende Überlegung gerechtfertigt: Über die Additionstheoreme für trigonometrische Funktionen folgt, daß mit $p_1 \in T_{n_1}$ und $p_2 \in T_{n_2}$ für das Produkt gilt $p_1 p_2 \in T_{n_1+n_2}$. (*Übungsaufgabe !*)

Nachfolgend wollen wir eine funktionalanalytische Charakterisierung des Raumes T_n vornehmen und einen Zusamenhang zwischen Fourier- und algebraischen Polynomen herstellen. Dazu beweisen wir zunächst das folgende Resultat.

Lemma 10.2 *(i) Hat ein trigonometrisches Polynom aus* T_n *mehr als* $2n$ *paarweise verschiedene Nullstellen im Periodizitätsintervall* $[0, 2\pi)$*, so verschwindet es dort identisch.*

(ii) Die Funktionen $\cos(kx),\ k = 0, \ldots, n$ *und* $\sin(kx),\ k = 1, \ldots, n$ *sind linear unabhängig auf dem Raum* $C[0, 2\pi]$.

Beweis: (i) Wir betrachten ein trigonometrisches Polynom $p_n \in T_n$ der Form

$$p_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)] \tag{10.3}$$

und gehen nun zu einer Formulierung in der Menge $\mathbb{C}$ über. Dabei ist i die *imaginäre Einheit.* Unter Verwendung von

$$c_k = \frac{1}{2}(a_k - ib_k), \quad c_{-k} = \frac{1}{2}(a_k + ib_k), \quad k = 0, ..., n$$

und mit $b_0 = 0$ folgt aus der Eulerschen Formel

$$e^{it} = \cos t + i \sin t$$

die zu (10.3) äquivalente komplexe Formulierung

$$p_n(x) = c_0 + \sum_{k=1}^{n} [(c_k + c_{-k}) \cos(kx) + i(c_k - c_{-k}) \sin(kx)] = \sum_{k=-n}^{n} c_k e^{ikx}. \tag{10.4}$$

Mit der Substitution $z = e^{ix}$ und der Bezeichnung

$$q_{2n}(z) = \sum_{k=-n}^{n} c_k z^{n+k}$$

erhalten wir

$$p_n(x) = z^{-n} q_{2n}(z).$$

Wir nehmen an, daß das trigonometrische Polynom p_n mehr als $2n$ paarweise verschiedene Nullstellen in $[0, 2\pi)$ hat. Dann hat das algebraische Polynom $q_{2n} \in \Pi_{2n}$ mehr als $2n$ paarweise verschiedene Nullstellen auf dem Einheitskreis in $\mathbb{C}$, denn die Funktion $t \mapsto e^{it}$ bildet das Intervall $[0, 2\pi)$ bijektiv auf den Einheitskreis in $\mathbb{C}$ ab. Nach Satz 9.1 muß das algebraische Polynom q_{2n} identisch verschwinden. Wegen (10.4) muß auch das trigonometrische Polynom p_n identisch verschwinden.

(ii) Wir nehmen an, daß

$$\frac{1}{2}a_0 + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)] = 0, \qquad x \in [0, 2\pi].$$

Das trigonometrische Polynom mit den Koeffzienten $a_0, a_1, b_1, \ldots, a_n, b_n$ hat offenbar mehr als $2n$ paarweise verschiedene Nullstellen in $[0, 2\pi)$. Nach Aussage (i) müssen dann aber die Koeffizienten sämtlich verschwinden. □

Als Folgerung aus Lemma 10.2 erhalten wir folgendes wesentliche Ergebnis.

Satz 10.3. *Der lineare Raum T_n der trigonometrischen Polynome vom Grad kleiner oder gleich n hat die Dimension $2n + 1$ und ist unisolvent bezüglich $2n + 1$ paarweise verschiedener Stützstellen aus dem Intervall $[0, 2\pi)$.*

10.2 Trigonometrische Interpolation

Definition 10.4. *Seien $2n+1$ paarweise verschiedene Stützstellen $x_0, ..., x_{2n}$ im Intervall $[0, 2\pi)$ sowie $2n + 1$ Werte $y_0, ..., y_{2n} \in \mathbb{R}$ vorgegeben. Bei der* trigonometrischen Interpolation *sucht man ein trigonometrisches Polynom $p_n \in T_n$ mit der Interpolationseigenschaft*

$$p_n(x_j) = y_j, \quad j = 0, ..., 2n.$$

Nachfolgender Satz zeigt, daß diese Aufgabe wohlgestellt ist.

Satz 10.5. *Die Aufgabe der trigonometrischen Interpolation ist eindeutig lösbar. Mit Hilfe der Lagrange-Polynome*

$$l_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{2n} \frac{\sin\left(\frac{1}{2}(x - x_j)\right)}{\sin\left(\frac{1}{2}(x_k - x_j)\right)}, \quad k = 0, ..., 2n \tag{10.5}$$

lautet die Lagrange-Darstellung des Interpolationspolynoms

$$p_n(x) = \sum_{k=0}^{2n} y_k l_k(x). \tag{10.6}$$

Beweis: Zunächst stellen wir die Existenz und Eindeutigkeit der Interpolationsfunktion fest, da wegen Satz 10.3 der Satz 9.21 anwendbar ist.

Die konkrete Gestalt von p_n folgt wegen $l_k(x_j) = \delta_{kj}$ und $l_k \in T_n, k = 0, ..., 2n$. Die Aussage $l_k \in T_n$ ergibt sich dabei aus dem Additionstheorem

$$\sin\frac{x - x_0}{2} \sin\frac{x - x_1}{2} = \frac{1}{2}\cos\frac{x_1 - x_0}{2} - \frac{1}{2}\cos\left(x - \frac{x_1 + x_0}{2}\right),$$

d.h. l_k ist ein Produkt aus n trigonometrischen Polynomen vom Grad 1. □

Wir hatten bereits im Fall der Interpolation mit algebraischen Polynomen gesehen, daß die Lagrange-Basis für praktische Zwecke wenig geeignet ist. Dies gilt natürlich auch im Fall der trigonometrischen Interpolation. Daher gehen wir künftig von dem Ansatz (10.3) bzw. (10.4) aus. Das Problem besteht somit in der effizienten Bestimmung der Koeffizienten a_k, b_k bzw. c_k.

Zur Vereinfachung der Darstellung betrachen wir nun den wichtigen Spezialfall *äquidistanter* Stützstellen. Sei zunächst eine *ungerade* Anzahl von Stützstellen angenommen mit

$$x_j = j\frac{2\pi}{2n + 1}, \quad j = 0, ..., 2n. \tag{10.7}$$

Dann gilt die Summenformel

$$\sum_{j=0}^{2n} e^{ikx_j} = \begin{cases} 2n + 1, & k = 0, \\ 0, & k = \pm 1, ..., \pm 2n \end{cases}, \tag{10.8}$$

denn für $e^{ix_k} \neq 1$ ergeben die geometrische Summe und die Eulersche Formel

$$\sum_{j=0}^{2n} e^{ikx_j} = \sum_{j=0}^{2n} e^{ijx_k} = \frac{1 - e^{i(2n+1)x_k}}{1 - e^{ix_k}} = 0.$$

Zur Vereinfachung der Berechnung setzen wir das trigonometrische Interpolationspolynom in komplexer Form an

$$p_n(x) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

Aus der Interpolationsforderung

$$p_n(x_j) = y_j, \quad j = 0, ..., 2n$$

schließen wir auf das eindeutig lösbare lineare Gleichungssystem

$$\sum_{k=-n}^{n} c_k e^{ikx_j} = y_j, \quad j = 0, ..., 2n.$$

Zur Lösung des Systems multiplizieren wir mit e^{-irx_j}, $r = -n, \ldots, n$, summieren über j, vertauschen die Summationsfolge und benutzen Beziehung (10.8)

$$\sum_{j=0}^{2n} y_j e^{-irx_j} = \sum_{k=-n}^{n} c_k \sum_{j=0}^{2n} e^{i(k-r)x_j} = (2n+1)c_r$$

und daher

$$c_k = \frac{1}{2n+1} \sum_{j=0}^{2n} y_j e^{-ikx_j}, \quad k = -n, ..., n. \tag{10.9}$$

Die Formeln (10.3) und (10.4) zeigen den Zusammenhang zwischen reeller und komplexer Formulierung des trigonometrischen Polynoms, insbesondere ist

$$a_k = c_k + c_{-k}, \qquad b_k = i(c_k - c_{-k}), \qquad k = 0, \ldots, n.$$

Daraus ergibt sich folgendes Resultat.

Satz 10.6. *Es existiert genau ein trigonometrisches Polynom*

$$p_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)]$$

mit der Interpolationseigenschaft

$$p_n\left(j\frac{2\pi}{2n+1}\right) = y_j, \quad j = 0, ..., 2n.$$

Für die Koeffizienten gelten die Formeln

$$a_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \cos\left(\frac{2\pi}{2n+1} jk\right), \quad k = 0, ..., n \tag{10.10}$$

$$b_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \sin\left(\frac{2\pi}{2n+1} jk\right), \quad k = 1, ..., n. \tag{10.11}$$

Im Falle einer *geraden* Anzahl der Stützstellen gibt es im äquidistanten Fall mit

$$x_j = j\frac{\pi}{n}. \quad j = 0, ..., 2n-1 \tag{10.12}$$

zu den $2n+1$ Freiheitsgraden eines trigonometrischen Polynoms aus dem Raum T_n zunächst nur $2n$ Interpolationsbedingungen. Man kann jedoch die Basisfunktion $\sin nx$ weglassen, da sie in den Stützstellen nur Nullstellen hat. Dann gilt

Satz 10.7. *Es existiert genau ein trigonometrisches Polynom*

$$p_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n-1} [a_k \cos(kx) + b_k \sin(kx)] + \frac{1}{2}a_n \cos(nx)$$

mit der Interpolationseigenschaft

$$p_n\left(j\frac{\pi}{n}\right) = y_j, \quad j = 0, ..., 2n-1.$$

Die Koeffizienten sind

$$a_k = \frac{1}{n}\sum_{j=0}^{2n-1} y_j \cos\left(\frac{\pi}{n}jk\right), \qquad k = 0, ..., n, \tag{10.13}$$

$$b_k = \frac{1}{n}\sum_{j=0}^{2n-1} y_j \sin\left(\frac{\pi}{n}jk\right), \qquad k = 1, ..., n-1. \tag{10.14}$$

Beweis: Der lineare Raum $\tilde{T}_n$ der trigonometrischen Polynome aus T_n mit verschwindendem Koeffizienten $b_n = 0$ vor $\sin(nx)$ hat offenbar die Dimension $2n$. Wir betrachten nun ein Element $p \in \tilde{T}_n$ mit den Nullstellen $p(x_j) = 0$, $j = 0, ..., 2n-1$. Ferner wird der Punkt x^* mit $x^* \neq x_j, j = 0, ..., 2n-1$ gewählt und

$$Q(x) := p(x) - p(x^*)\frac{\sin(nx)}{\sin(nx^*)}$$

gesetzt. Dann ist $Q \in T_n$ und hat die $2n+1$ Nullstellen $x_j, j = 0, ...2n-1$ sowie x^*. Daher muß Q verschwinden und p ist ein Vielfaches von $\sin(nx)$. Nach Definition verschwindet dann auch $p \in \tilde{T}_n$. Daher ist $\tilde{T}_n$ unisolvent bezüglich der äquidistanten Stützstellen $x_j, j = 0, ..., 2n-1$.

Existenz und Eindeutigkeit der Interpolationsfunktion ergeben sich dann aus Satz 9.21. Die gesuchte Darstellung gewinnt man analog zum Beweis von Satz 10.6. □

Man spricht im Fall äquidistanter Stützstellenverteilung von der *diskreten Fourier-Transformation*. Wir wollen diesen Punkt etwas genauer betrachten:

Man kann zeigen (vgl. Approximationstheorie im Kurs *Numerische Mathematik II*), daß eine gegebene Funktion $f \in C[0, 2\pi]$ in der Klasse der trigonometrischen Polynome vom maximalen Grad n bezüglich der $L^2(0, 2\pi)$-Norm am besten durch das Polynom

$$(P_n f)(x) = \frac{1}{2}\tilde{a}_0 + \sum_{k=1}^{n}\Big[\tilde{a}_k \cos(kx) + \tilde{b}_k \sin(kx)\Big]$$

mit den Fourier-Koeffizienten (für $k = 0, \ldots, n$)

$$\tilde{a}_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)\ dx, \qquad \tilde{b}_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(kx)\ dx \tag{10.15}$$

approximiert wird. Man erhält diese Formeln auch, wenn man in den Formeln (10.10), (10.11) für a_k bzw. b_k formal den Grenzübergang $n \to \infty$ ausführt.

Andererseits erhält man die Formeln (10.10), (10.11) aus den Formeln (10.15) bei Anwendung der zusammengesetzten numerischen Integration mittels der sogenannten Trapezregel, die wir im Kapitel 13 behandeln werden. Der aufgezeigte Zusammenhang verdeutlicht die Begriffsbildung "diskrete Fourier-Transformation".

10.3 Berechnung der Fourier-Koeffizienten

Für Anwendungen ist eine schnelle, effiziente Berechnung von Ausdrücken der Form

$$\sum_{k=1}^{n} \alpha_k \cos(kx) \quad \text{bzw.} \quad \sum_{k=1}^{n} \beta_k \sin(kx) \tag{10.16}$$

erforderlich. Das betrifft sowohl die Berechnung der Fourier-Koeffizienten a_k und b_k in den Sätzen 10.5 und 10.6 (*Fourier-Analyse*) als auch die Berechnung der Werte der trigonometrischen Interpolationspolynome $p_n(x)$ (*Fourier-Synthese*).

Eine erste Idee besteht in der Verwendung des Horner-Schemas (d.h. geeignete Klammersetzung) zur Berechnung des komplexen algebraischen Polynoms

$$p(z) = \sum_{k=0}^{n} a_k z^k.$$

Mittels Horner-Schema folgt

$$\begin{aligned} B_n &= a_n, \\ B_{k-1} &= B_k z + a_{k-1}, \qquad k = n-1, ..., 1 \end{aligned}$$

und damit $p(z) = B_0$. Durch Substitution $z = e^{ix}$ und Aufspaltung in Real- und Imaginärteil $B_k = u_k + iv_k$ ergibt sich bei (vereinfachend angenommenen) reellen Werten a_k

$$\begin{aligned} u_n &= a_n, & v_n &= 0, \\ u_{k-1} &= u_k \cos x - v_k \sin x + a_{k-1}, & v_{k-1} &= u_k \sin x + v_k \cos x. \end{aligned}$$

Der numerische Aufwand zur Auswertung eines trigonometrischen Polynoms besteht also in der Berechnung von zwei Funktionswerten der Sinus- bzw. Kosinus-Funktion sowie $0(n)$ Multiplikationen und Additionen. Zur Berechnung aller Koeffizienten a_k und b_k eines trigonometrischen Polynoms werden damit insgesamt $0(n^2)$ Multiplikationen und Additionen benötigt.

Bei großen Werten von n möchte man diesen Aufwand weiter reduzieren. Eine dafür geeignete Methode ist die *schnelle Fourier-Transformation (FFT - fast Fourier transformation)* nach Cooley/ Tukey (1965). Eine geschickte Ausnutzung von Symmetrien der Einheitswurzeln in $\mathbf{C}$ im Falle von $n = 2^s$ ist dabei der Schlüssel. Eine gute Gesamtdarstellung findet man z.B. in dem Artikel von H.R. Schwarz: *Elementare Darstellung der schnellen Fourier-Transformation* in Computing 18 (1977), 107-116. Wir betrachten hier nur die tragende Idee zur Berechnung der komplexwertigen Fourier-Koeffizienten

$$c_k^{(n)} = \frac{1}{n} \sum_{j=0}^{n-1} y_j \exp\left[-\frac{2\pi i}{n} kj\right], \quad k = 0, ..., n-1. \tag{10.17}$$

Gelte $n = pq$. Später betrachten wir den Spezialfall $p = 2, q = 2^{s-1}$. Setzt man in der obigen Summe für den Laufindex $j = rp + l$, so ergibt sich für $l = 0, ..., p-1, \;\; k = 0, ..., n-1$

$$\begin{aligned} c_k^{(n)} &= \frac{1}{p} \sum_{l=0}^{p-1} \frac{1}{q} \sum_{r=0}^{q-1} y_{l+rp} \exp\left(-\frac{2\pi i}{n} k(l+rp)\right) \\ &= \frac{1}{p} \sum_{l=0}^{p-1} \left\{ \frac{1}{q} \sum_{r=0}^{q-1} y_{l+rp} \exp\left(-\frac{2\pi i}{q} kr\right) \right\} \exp\left(\frac{-2\pi i}{n} kl\right) \\ &= \frac{1}{p} \sum_{l=0}^{p-1} c_{k,l}^{(q)} \exp\left(\frac{-2\pi i}{n} kl\right). \end{aligned} \tag{10.18}$$

Dabei stellen die Größen $c_{k,l}^{(q)}$ Koeffizienten für die kleinere Zahl q von Stützstellen bei festem Index l dar. Insbesondere spaltet man bei $p = 2$ die Summe in die Terme mit geradem und ungeradem Index auf.

Bei der FFT wendet man diese Prozedur rekursiv an. Im Fall $p = 2$ halbiert sich also jeweils die Zahl der pro Fourier-Koeffizient zu berücksichtigenden Stützstellen. Hinweise zur effektiven Programmierung der FFT findet man in der o.a. Arbeit von Schwarz.

Nachfolgend geben wir eine Abschätzung des für die FFT erforderlichen Rechenaufwandes. Sei dazu M_n die Anzahl der Multiplikationen zur Berechnung aller Koeffizienten $c_k^{(n)}$ bei n Stützstellen. Aus Gleichung (10.18) erhalten wir

$$M_n = (p-1)n + pM_q,$$

denn bei Kenntnis von $c_{k,l}^{(q)}$ sind noch $p-1$ Multiplikationen auszuführen. Ferner sind zuvor die Koeffizienten $c_{k,l}^{(q)}$ für $l = 0, ..., p-1$ zu ermitteln. Mit der Wahl $n = 2^s = pq = 2 \cdot 2^{s-1}$ folgt damit die Funktionalgleichung

$$M_{2^s} = 2^s + 2M_{2^{s-1}}$$

mit der Lösung $M_{2^s} = s \cdot 2^s$ bzw. $M_n = n \cdot \log_2 n$.

Beispiel 10.8. Die folgende Überlegung verdeutlicht den erheblich reduzierten Aufwand bei der FFT. Sei $n = 2^8 = 256$. Dann ist $n^2 = 65.536$ und $n \cdot log_2 n = 2.048$, d.h. der Aufwand der FFT liegt um den Faktor 32 niedriger als bei der Anwendung des Horner-Schemas. □

10.4 Konvergenz trigonometrischer Polynome

Konvergenzuntersuchungen für die Folge $(p_n)_n$ der trigonometrischen Interpolationspolynome an eine gegebene periodische Funktion f gestalten sich komplizierter als entsprechende Untersuchungen bei algebraischen Interpolationsproblemen (vgl. Kap. 9). Es gilt jedoch das folgende Resultat.

Satz 10.9. *Sei $f \in C(\mathbb{R})$ eine gegebene $2\pi-$periodische Funktion. Dann konvergiert die Folge der trigonometrischen Interpolationspolynome p_n an f in der L^2-Norm gegen f, d.h.*

$$\lim_{n\to\infty} \|p_n - f\|_2 = 0.$$

Ist darüber hinaus sogar $f \in C^1(\mathbb{R})$, so konvergiert die Folge in der Maximumnorm, d.h.

$$\lim_{n\to\infty} \|p_n - f\|_\infty = 0$$

bzw. $f(x) = \lim_{n\to\infty} p_n(x)$ für alle $x \in [0, 2\pi]$.

Beim Beweis benutzt man maßgeblich den Approximationssatz von Weierstraß für trigonometrische Polynome. Wir gehen darauf im Rahmen der Approximationstheorie (vgl. Numerische Mathematik II) ein.

Insbesondere ist das Konvergenzresultat in der L^2-Norm im Zusammenhang mit der Minimaleigenschaft von Fourier-Reihen (vgl. Abschnitt 10.2) als Grenzfall für $n \to \infty$ zu sehen. Man kann die Anwendung der Fourier-Approximation auch auf solche Funktionen f erweitert werden, für die noch $\|f\|_{L^2(0,2\pi)}$ endlich ist. Diese Funktionen bilden den Lebesgue-Raum $L^2(0, 2\pi)$, der durch "Vervollständigung" von $C[0, 2\pi]$ bezüglich der Norm $\|\cdot\|_{L^2(0,2\pi)}$ ensteht.

Das folgende Beispiel verdeutlicht, daß die (punktweise) Konvergenz der Fourier-Polynome für $n \to \infty$ nicht sehr schnell ist, wenn die gegebene Kurve nicht mehr stetig stetig ist.

Beispiel 10.10. Wir zeigen als Beispiel für die "Sägezahn"-Impulskurve

$$f(x) = x, \quad 0 \le x < 1; \qquad f(x+1) = f(x), \quad x \in \mathbb{R}$$

die mit *Maple* berechneten Fourier-Ansätze mit $n = 15$ und $n = 100$. Abb. 10.1 verdeutlicht, daß sich die Approximation mit wachsendem n nur langsam der vorgegebenen Funktion nähert, insbesondere in deren Unstetigkeitspunkten. Bei den lokalen Oszillationen in deren Umgebung spricht auch vom "Gibbs-Phänomen". □

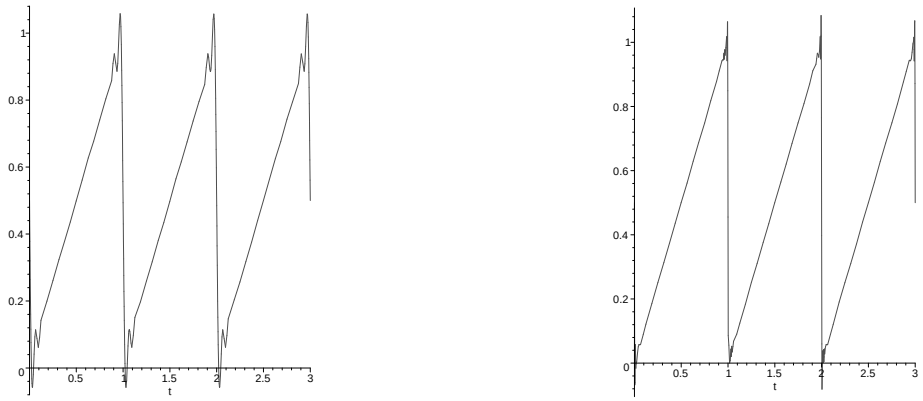

Abbildung 10.1: Fourier-Approximation der "Sägezahnkurve" mit $n = 15$ und $n = 100$

Für den Fall hinreichender glatter periodischer Funktionen zitieren wir ohne Beweis noch folgendes Resultat.

Satz 10.11. *Sei $f \in C^k(\mathbb{R})$ $2\pi-$periodisch mit $k \in \mathbb{N}$ und sei p_n das trigonometrische Interpolationspolynom nach Satz 10.5 bzw. Satz 10.6. Dann gibt es eine nur von k abhängige Konstante C_k mit*

$$\|f - p_n\|_{L^2} \le C_k n^{-k} \left(\int_0^{2\pi} \left[|f(x)|^2 + |f^{(k)}(x)|^2 \right] dx \right)^{\frac{1}{2}}, \quad \forall n \in \mathbf{N}.$$

Dabei ist

$$\|g\|_{L^2} := \left(\int_0^{2\pi} |g(x)|^2 \, dx \right)^{\frac{1}{2}}.$$

Insbesondere konvergiert für $f \in C^\infty(\mathbb{R})$ der Fehler $\|f - p_n\|_{L^2}$ für $n \to \infty$ schneller als jede Potenz n^{-k} bei festem $k \in \mathbb{N}$ gegen Null.

Kapitel 11

Spline-Interpolation

Im Kapitel 9 hatten wir festgestellt, daß Interpolationspolynome bei Vergrößerung der Zahl n der Stützstellen, d.h. genauer im Grenzprozeß $n \to \infty$, nicht notwendig gegen die zu interpolierende Funktion f konvergieren. Einen Ausweg bietet die stückweise polynomiale Interpolation durch *Splines*. Sie bilden auch eine wichtige Grundlage für die numerische Integration (vgl. Kapitel 13) sowie für die Diskretisierung von Differentialgleichungen (vgl. Numerische Mathematik II).

11.1 Räume von Spline-Funktionen

Definition 11.1. *Sei $a = x_0 < x_1 < ... < x_n = b$ eine Unterteilung des Intervalls $[a, b]$. Dann heißt eine Funktion $s : [a, b] \to \mathbb{R}$* Spline m-ten Grades, *falls*

$$(i) \quad s \in C^{m-1}[a, b],$$

$$(ii) \quad s\Big|_{[x_{j-1},x_j]} \in \Pi_m[x_{j-1}, x_j], \quad j = 1, ..., n.$$

Die Menge der Splines $m-$ter Ordnung zu einer gegebenen Unterteilung von $[a, b]$ in n Teilintervalle bezeichnen wir mit $S_n^m[a, b]$.

Beispiel 11.2. *Lineare Splines ($m = 1$)*

Splines 1. Ordnung sind (global) stetige und stückweise lineare Funktionen. Bei gegebener Funktion $f \in C[a, b]$ lauten die Interpolationforderungen

$$s(x_j) = f(x_j), \qquad j = 0, ..., n,$$

d.h. die Spline-Funktion entsteht einfach durch stetige Verheftung der auf den Teilintervallen stückweise linearen Interpolationspolynome in den Punkten $(x_j, f(x_j))$. Auf jedem Teilintervall $[x_{j-1}, x_j]$ gilt nach Satz 9.10 die Fehlerabschätzung

$$|(s - f)(x)| \le \frac{1}{8}|f''(\xi)|h_j^2, \quad \xi \in [x_{j-1}, x_j], h_j := x_j - x_{j-1}.$$

Damit konvergiert die Spline-Funktion 1. Ordnung auf $[a, b]$ zumindest für $f \in C^2[a, b]$ gleichmäßig gegen f, wenn $h := \max_{j=1,...,n} h_j \to 0$. □

Bei manchen Anwendungen (z.B. Design-Problemen) sind Splines 1. Ordnung nachteilig, da ein "glatterer" Übergang an den Teilpunkten erforderlich wäre. Einen Kompromiß zwischen Glattheit und Rechenaufwand bieten *kubische Splines.*

Beispiel 11.3. *Kubische Splines ($m = 3$)*

Nach Definition gilt $s \in C^2[a,b]$ und $s \in \Pi_3[x_{j-1}, x_j]$ für $j = 1, ..., n$. Dem englischen Wort *spline* entspricht etwa das deutsche Wort *Strak*. Das ist ein Werkzeug aus dem Schiffbau zur Führung glatter Kurven durch vorgegebene Punkte, d.h. der Strak wird dort durch Lager fixiert. In der Mechanik leitet man für die Balkenbiegung $u(x)$ die Gleichung $u^{(4)}(x) = 0$ ab. Zwischen den Lagerpunkten wird der Strak also durch ein Polynom 3. Grades beschrieben. Eine mathematische Darstellung der Biegelinie $u = u(x)$ erhalten wir dann unter Beachtung der Stetigkeit der Auslenkung u, der Biegung u' sowie des Biegemomentes u'' in den Lagerpunkten.

In Abb. 11.1 (i) vergleichen wir für den in Abschn. 9.1 betrachteten Datensatz (vgl. Abb. 9.1) die mit der MAPLE-Funktion `spline` für $m = 1$ erzeugte Spline-Funktion mit dem globalen Interpolationspolynom 9. Ordnung. Die Spline-Funktion ist zwar an den Kopplungspunkten weniger glatt, vermeidet aber die Randoszillationen des globalen Polynoms. Ferner betrachten wir

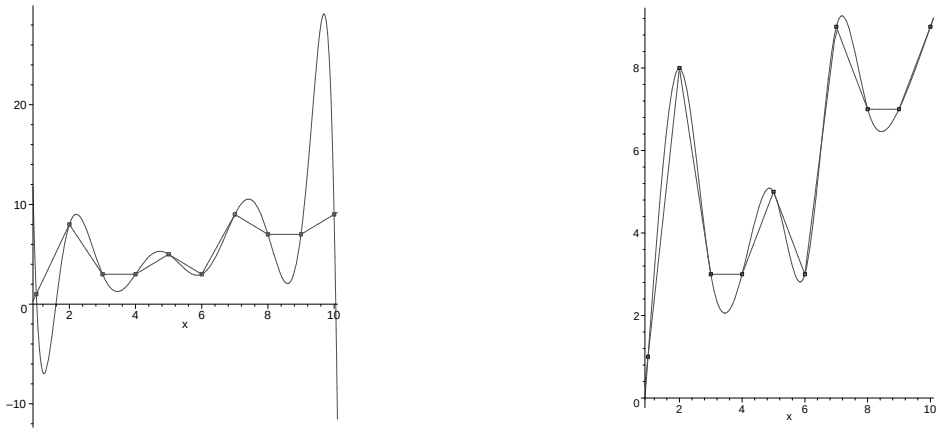

Abbildung 11.1: Interpolation von Zufallsdaten (i) Splines 1. Ordnung und Lagrange-Interpolation 9. Grades, (ii) Splines 1. bzw. 3. Ordnung

in Abb. 11.1 (ii) die Interpolation durch Splines 1. und 3. Ordnung. Der "glattere" Verlauf für kubische Splines ist offensichtlich. Vergleicht man ferner die Splines mit $m = 3$ und $m = 9$ (hier nicht gezeigt), so ist die Spline-Funktion mit $m = 9$ zwar "glatter", sie tendiert aber ebenso wie das globale Polynom 9. Grades in (i) zu Randoszillationen. Auch dies spricht für die Wahl $m = 3$. □

Zur Ermittlung einer Basis des Raumes $S_n^m[a,b]$ der Splines der Ordnung m definieren wir über die Funktion

$$x_+^m := \begin{cases} x^m, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

das System der *Kardinalsplines*

$$\begin{cases} \phi_k(x) := (x - x_0)^k, & k = 0, \ldots, m \\ \psi_j(x) := (x - x_j)_+^m, & j = 1, \ldots, n-1. \end{cases} \tag{11.1}$$

Lemma 11.4. *Die $n + m$ Funktionen des Systems (11.1) sind linear unabhängig.*

Beweis: Sei

$$\sum_{k=0}^{m} a_k (x - x_0)^k + \sum_{j=1}^{n-1} b_j (x - x_j)_+^m = 0, \quad x \in [a,b].$$

Nach Definition ist

$$\sum_{k=0}^{m} a_k(x-x_0)^k = 0, \qquad x \in [x_0, x_1],$$

damit $a_k = 0,\ k = 0, \ldots, m$. Weiter gilt $b_1(x-x_1)_+^m = 0,\ x \in [x_1, x_2]$ und daher $b_1 = 0$. Sukzessiv schließen wir auf $b_j = 0,\ j = 1, ..., n-1$. Daraus folgt die lineare Unabhängigkeit der Funktionen des Systems (11.1). □

Der nachfolgende Satz zeigt, daß die Funktionen des Systems (11.1) eine Basis des $S_n^m[a,b]$ bilden. Jede Funktion $s \in S_n^m[a,b]$ hat dann die Kardinalspline-Darstellung

$$s(x) = \sum_{k=0}^{m} a_k(x-x_0)^k + \sum_{j=1}^{n-1} b_j(x-x_j)_+^m, \quad x \in [a,b]. \tag{11.2}$$

Satz 11.5. *Der Raum $S_n^m[a,b]$ der Splines ist ein linearer Raum der Dimension $m+n$. Die Funktionen des Systems (11.1) bilden eine Basis von $S_n^m[a,b]$.*

Beweis: Wir zeigen die Gültigkeit der Darstellung (11.2) durch vollständige Induktion über die Teilintervalle, d.h.

$$s(x) = \sum_{k=0}^{m} a_k(x-x_0)^k + \sum_{j=1}^{i-1} b_j(x-x_j)_+^m, \quad x \in [x_0, x_i],\ \ i = 1, ..., n-1.$$

Der Induktionsanfang für $i = 1$ ist offenbar richtig wegen $s \in \Pi_m[x_0, x_1]$. Sei nun die Gültigkeit der Darstellung bewiesen für eine Zahl $i \in \{1, \ldots, n-1\}$. Dann verschwindet die Differenz

$$d(x) := s(x) - \sum_{k=0}^{m} a_k(x-x_0)^k - \sum_{j=1}^{i-1} b_j(x-x_j)_+^m$$

auf dem Intervall $[x_0, x_i]$. Ferner ist $d \in \Pi_m[x_i, x_{i+1}]$. Weiter gilt dann für d wegen der Spline-Definition $d^{(j)}(x_i) = 0,\ j = 0, \ldots, m-1$. Dies ergibt die Gestalt

$$d(x) = b_i(x-x_i)_+^m, \quad x \in [x_i, x_{i+1}].$$

Wegen $(x-x_i)_+^m = 0$ auf $[x_0, x_i]$ ist die Behauptung auch für $i+1$ richtig.
Die Linearität des Raumes $S_n^m[a,b]$ ist offensichtlich. Schließlich ergibt sich die Dimension aus Lemma 11.4 und der Analyse aus dem ersten Teil des Beweises. □

11.2 Interpolation in Spline-Räumen

Bei Interpolation einer Funktion f durch Splines $m-$ter Ordnung an den $n+1$ Stützstellen $x_0, ..., x_n$, d.h.

$$s(x_j) = f(x_j), \qquad j = 0, ..., n,$$

werden $m-1$ Freiheitsgrade des Raumes $S_n^m[a,b]$ nicht genutzt. Eine Ausnahme bildet hier der Fall stückweise linearer Splines ($m = 1$, vgl. Beispiel 11.2). Die Menge $S_n^1[a,b]$ ist unisolvent bezüglich der Stützstellen $x_0, \ldots, x_n$. (Der Nachweis sei als *Übungsaufgabe* überlassen.)
Für Splines mit $m \geq 2$ kann man zusätzlich, zum Beispiel am Rand des Intervalls $[a,b]$, Bedingungen stellen. Aus Symmetriegründen betrachten wir nur ungerade Zahlen $m = 2l-1$ mit $l \in \mathbf{N},\ l \geq 2$ für eine der *Randbedingungen*

$$s^{(j)}(a) = f^{(j)}(a), \quad s^{(j)}(b) = f^{(j)}(b), \qquad j = 1, \ldots, l-1, \tag{11.3}$$

bzw.

$$s^{(l+j)}(a) = s^{(l+j)}(b) = 0, \qquad j = 0, \ldots, l-2, \tag{11.4}$$

oder für periodische Funktionen f mit Periode $b-a$ die *Periodizitätsbedingungen*

$$s^{(j)}(a) = f^{(j)}(a) = s^{(j)}(b) = f^{(j)}(b), \qquad j = 1, \ldots, l-1, \tag{11.5}$$

Im Fall (11.3) spricht man auch von *Hermite-Randbedingungen*. Die sogenannten *natürlichen Randbedingungen* (11.4) sind jedoch nicht von praktischer Bedeutung.

Zunächst zeigen wir eine Extremaleigenschaft für Splines.

Lemma 11.6. *Sei $f \in C^l[a,b]$ für $l \in \mathbb{N},\ l \geq 2$ und $s \in S_n^m[a,b]$ mit $m = 2l-1$ die interpolierende Spline-Funktion, d.h. $s(x_j) = f(x_j),\ j = 0, ..., n$. Ferner gelte eine der Randbedingungen (11.3), (11.4) bzw. (11.5) . Dann gilt*

$$\int_a^b \left[f^{(l)}(x) - s^{(l)}(x)\right]^2 dx = \int_a^b \left[f^{(l)}(x)\right]^2 dx - \int_a^b \left[s^{(l)}(x)\right]^2 dx.$$

Beweis: Ausmultiplikation ergibt

$$\int_a^b \left[f^{(l)}(x) - s^{(l)}(x)\right]^2 dx = \int_a^b \left[f^{(l)}(x)\right]^2 dx - \int_a^b \left[s^{(l)}(x)\right]^2 dx - 2S$$

mit

$$S := \int_a^b \left[f^{(l)}(x) - s^{(l)}(x)\right] s^{(l)}(x) dx.$$

Auf S wenden wir $(l-1)$-fach die Regel der partiellen Integration an und berücksichtigen jeweils einer der Bedingungen (11.3), (11.4) bzw. (11.5). Dann erhalten wir

$$S = (-1)^{l-1} \int_a^b \left[f'(x) - s'(x)\right] s^{(m)}(x) dx,$$

da $f \in C^l[a,b]$ und für $s \in C^{m-1}[a,b]$ die Ableitung $s^{(m)}$ stückweise existiert. Erneute partielle Integration liefert mit den Interpolationsbedingungen sowie wegen $s \in \Pi_m[x_{j-1}, x_j]$, daß

$$\begin{aligned} S &= (-1)^{l-1} \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} \left[f'(x) - s'(x)\right] s^{(m)}(x) dx \\ &= (-1)^{l-1} \sum_{j=1}^{n} \left(\left[f(x) - s(x)\right] s^{(m)}(x) \Big|_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} \left[f(x) - s(x)\right] s^{(m+1)}(x)\ dx \right) \ = \ 0. \end{aligned}$$

Daraus folgt die Behauptung. □

Lemma 11.6 ergibt die Ungleichung

$$\int_a^b \left[s^{(l)}(x)\right]^2 dx \leq \int_a^b \left[f^{(l)}(x)\right]^2 dx.$$

Dies erlaubt für den Fall kubischer Splines, d.h. $m = 3$, eine geometrische Interpretation. Für die Krümmung einer durch $y = g(x)$ beschriebenen Kurve gilt

$$k(x) = \frac{g''(x)}{(1 + [g'(x)]^2)^{3/2}},$$

d.h. im Falle $|g'(x)| \ll 1$ ist näherungsweise $k(x) \approx g''(x)$ bzw.

$$\|k\|_2^2 := \int_a^b [k(x)]^2 \, dx \approx \|g\|_2^2.$$

Die abgeleitete Ungleichung bedeutet nun, daß der interpolierende kubische Spline unter allen Funktionen $g \in C^2[a,b]$, die die gleichen Interpolationsforderungen erfüllen, näherungsweise die mittlere Krümmung $\|k\|_2$ minimiert.

Als für uns wesentlichste Folgerung aus Lemma 11.6 zeigen wir jetzt, daß die Bestimmung der interpolierenden Spline-Funktion $s \in S_n^m[a,b]$ mit $m = 2l-1,\ l \geq 1$ an eine gegebene Funktion f unter einer der Randbedingungen (11.3), (11.4) oder (11.5) wohldefiniert ist. Speziell muß man sich nicht explizit um die Stetigkeitsbedingungen in den Ableitungen $s^{(j)}$ für $j = 1, \ldots, 2l-2$ der Spline-Funktion in den inneren Stützstellen $x_1, \ldots, x_{n-1}$ kümmern. Genauer gilt

Satz 11.7. *Unter den Voraussetzungen von Lemma 11.6 existiert genau ein Spline $s \in S_n^m[a,b]$, der die Funktion f an den Stützstellen $x_0, ..., x_n$ interpoliert und eine der Randbedingungen (11.3), (11.4) oder (11.5) erfüllt.*

Beweis: Unter Verwendung der Darstellung (11.2) durch Kardinalsplines ergibt sich aus den Interpolationsforderungen und den Randbedingungen (11.3), (11.4) bzw. (11.5) ein Gleichungssystem für die Koeffizienten a_k und b_k. Nach dem Beweis zum Satz 9.21 ist lediglich zu zeigen, daß der interpolierende Spline s zu $f \equiv 0$ auf $[a,b]$ identisch verschwindet.

Lemma 11.6 mit $f \equiv 0$ liefert $\int_a^b \left[s^{(l)}(x)\right]^2 dx = 0$, damit $s^{(l)}(x) = 0$ bzw. $s \in \Pi_{l-1}[a,b]$. Wegen der Randbedingungen für s folgt aber $s(x) \equiv 0$. □

11.3 B-Splines

Eine naheliegende Variante zur Berechnung der Koeffizienten a_k, b_k im Ansatz (11.2) wäre, das entstehende Gleichungssystem aus den Interpolationsforderungen

$$s(x_j) = f(x_j), \qquad j = 0, \ldots, n$$

und einer der Randbedingungen (11.3) oder (11.5) zu lösen. Dieser Weg ist aber bei großer Zahl $n \gg 1$ von Stützstellen nicht geeignet, da die entsprechende Koeffizientenmatrix sehr stark besetzt und schlecht konditioniert ist. Dies liegt vor allem daran, daß ein Teil der Kardinalsplines das gesamte Intervall $[a,b]$ als Träger hat.

Einen Ausweg stellt die Verwendung von Basisfunktionen in $S_n^m[a,b]$ mit kleinem Träger dar, d.h. sie sind nur auf einem kleinem Teilgebiet des Intervalls $[a,b]$ von Null verschieden. Auf Schönberg gehen die sogenannten *B-Splines* (basic spline curves) zurück, die wir hier vereinfachend für den Fall einer äquidistanten Zerlegung mit $x_k = a + kh,\ h = \frac{b-a}{n}$ beschreiben.

Definition 11.8. *Ausgehend von*

$$B_0(x) := \begin{cases} 1, & |x| \leq \frac{1}{2}, \\ 0, & |x| > \frac{1}{2} \end{cases}$$

definiert man B-Splines rekursiv durch

$$B_{m+1}(x) := \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} B_m(t) \, dt, \quad x \in \mathbb{R}, \qquad m = 0, 1, \ldots. \tag{11.6}$$

Durch vollständige Induktion nach m gewinnt man folgendes Resultat.

Lemma 11.9. *Die durch (11.6) definierten B-Splines gehören zum Raum $C^{m-1}(\mathbb{R})$, sind nichtnegativ und verschwinden außerhalb des Intervalls $[-\frac{m}{2}-\frac{1}{2}, \frac{m}{2}+\frac{1}{2}]$. Sie sind stückweise polynomial vom Grad m auf den Intervallen $[i, i+1]$ für ungerade Zahlen m bzw. $[i-\frac{1}{2}, i+\frac{1}{2}]$ für gerade Zahlen m für beliebige ganze Zahlen i.*

Beweis: *(Übungsaufgabe !)* □

Beispiel 11.10. *(B-Spline der Ordnung $m = 1, 2, 3$)*

Für $m = 1$ erhalten wir

$$B_1(x) = \begin{cases} 1 - |x|, & |x| \le 1, \\ \\ 0, & |x| \ge 1. \end{cases}$$

Diese wegen ihrer Form *Hutfunktionen* (hat functions) genannten Splines spielen bei der Näherungslösung von Differential- und Integralgleichungen eine wichtige Rolle (vgl. Numerische Mathematik II). Sie stellen das einfachste Beispiel für sogenannte *finite Elemente* dar.

Man kann diese stückweise linearen, jedoch global stetigen Funktionen leicht auf den mehrdimensionalen Fall verallgemeinern, wenn man sie über simplizialen Gebieten im $\mathbb{R}^n$ definiert. Das sind im Fall $n = 1$ gerade Intervalle, für $n = 2$ Dreiecke und für $n = 3$ Tetraeder.

Eine weitere Modifikation im Fall $m = 1$ ist, daß man global stetige, jedoch stückweise polynomiale Funktionen höherer Ordnung über den Intervallen $[x_{j-1}, x_j]$ konstruiert. Wir werden diese Idee bei der numerischen Integration in Kapitel 13 benutzen.

Für $m = 2$ erhält man

$$B_2(x) = \frac{1}{2} \begin{cases} 2 - (|x| - \frac{1}{2})^2 - (|x| + \frac{1}{2})^2, & |x| \le \frac{1}{2}, \\ \\ (|x| - \frac{3}{2})^2, & \frac{1}{2} \le |x| \le \frac{3}{2}, \\ \\ 0, & |x| \ge \frac{3}{2}. \end{cases}$$

Schließlich gewinnt man für $m = 3$ die Darstellung

$$B_3(x) = \frac{1}{6} \begin{cases} (2 - |x|)^3 - 4(1 - |x|)^3, & |x| \le 1, \\ \\ (2 - |x|)^3, & 1 \le |x| \le 2, \\ \\ 0, & |x| \ge 2. \end{cases}$$

Abbildung 11.2 zeigt die B-Splines der Ordnungen $m \in \{0, 1, 2\}$. □

Wir wollen jetzt zeigen, daß bei *äquidistanter* Zerlegung eines Intervalles $[a, b]$ durch B-Splines eine Basis des Raumes $S_n^m[a, b]$ erzeugt wird.

Satz 11.11. *Sei durch $x_k = a + hk$ mit $k = 0, \ldots, n$ und $n \ge 2$ eine äquidistante Zerlegung des Intervalls $[a, b]$ mit der Schrittweite $h = \frac{1}{n}(b - a)$ gegeben. Gelte ferner $m = 2l - 1$ mit $l \in \mathbb{N}$. Dann erhält man eine Basis des Raumes $S_n^m[a, b]$ durch die transformierten B-Splines*

$$B_{m,k}(x) := B_m\left(\frac{x - x_k}{h}\right), \quad k = -l + 1, \ldots, n + l - 1, \qquad x \in [a, b]. \tag{11.7}$$

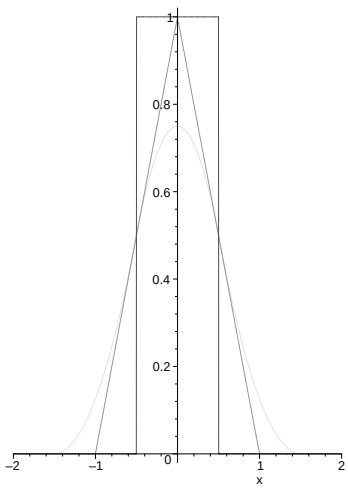

Abbildung 11.2: B-Splines der Ordnung $m \in \{0, 1, 2\}$

Beweis: (i) Wir beweisen zunächst durch Induktion nach $m \in \mathbb{N}_0$, daß die verschobenen B-Splines

$$B_m(\cdot - k), \qquad k = 0, \ldots, m$$

auf dem Intervall $I_m = \left[\frac{1}{2}(m-1), \frac{1}{2}(m+1)\right]$ linear unabhängig sind. Für $m = 0$ ist die Behauptung nach Definition 11.8 offenbar erfüllt.
Sie sei auch für eine Zahl $m - 1$ mit $m \in \mathbb{N}$ richtig. Gelte nun

$$\sum_{k=0}^{m} \gamma_k B_m(x - k) = 0, \qquad x \in I_m. \tag{11.8}$$

Differentiation in (11.8) ergibt unter Beachtung der Definition von B-Splines

$$\sum_{k=0}^{m} \gamma_k \left[B_{m-1}\left(x - k + \frac{1}{2}\right) - B_{m-1}\left(x - k - \frac{1}{2}\right)\right] = 0, \qquad x \in I_m.$$

Die Funktionen $B_{m-1}(\cdot + \frac{1}{2})$ und $B_{m-1}(\cdot - m - \frac{1}{2})$ verschwinden auf I_m nach Satz 11.9, daher

$$\sum_{k=1}^{m} (\gamma_k - \gamma_{k-1})\, B_{m-1}\left(x - k + \frac{1}{2}\right) = 0, \qquad x \in I_m.$$

Damit folgt nach Induktionsvoraussetzung, daß $\gamma_k = \gamma_{k-1}$ für $k = 1, \ldots, m$ bzw. $\gamma_k = \gamma$ für $k = 0, \ldots, m$. Somit können wir Formel (11.8) auch schreiben als

$$\gamma \sum_{k=0}^{m} B_m(x - k) = 0, \qquad x \in I_m.$$

Nach Integration über das Intervall I_m ergibt sich

$$\gamma \sum_{k=0}^{m} \int_{I_m} B_m(x - k)\, dx = \gamma \sum_{k=0}^{m} \int_{\frac{m-1}{2}-k}^{\frac{m+1}{2}-k} B_m(t)\, dt = \gamma \int_{-\frac{1}{2}(m+1)}^{\frac{1}{2}(m+1)} B_m(t)\, dt = 0.$$

Wegen der Nichtnegativität der Splines B_m nach Lemma 11.9 ist $\gamma = 0$. Daraus folgt die lineare Unabhängigkeit auch für die Zahl m und damit die Induktionsaussage.

(ii) Man rechnet unmittelbar nach, daß die nach (11.7) definierten Splines

$$B_{m,k}(x) = B_m\left(\frac{x-x_k}{h}\right) \equiv B_m\left(\frac{x-a}{h} - k\right), \qquad k = -l+1, \ldots, n+l-1$$

unter Beachtung von Lemma 11.9 zum Raum $S_n^m[a,b]$ gehören. Durch sinngemäße Anwendung der in Teil (i) des Beweises gezeigten Aussage folgt, daß diese $m+n$ Funktionen linear unabhängig sind. Die Behauptung folgt dann aus Satz 11.5. □

Man kann die *Spline-Interpolation* mittels B-Spline-Basis wesentlich günstiger als mit Kardinal-Splines berechnen. Dies liegt wesentlich daran, daß die B-Splines nur einen kleinen Träger haben. Im Ergebnis erhält man eine im allgemeinen Fall große, jedoch schwachbesetzte Matrix.

Wir demonstrieren dies am Beispiel kubischer Splines, d.h. für $m = 3$. Nach Beispiel 11.10 verschwindet B_3 außerhalb des Intervalls $(-2, 2)$. Weiter ergibt sich

$$B_3(0) = \frac{2}{3}, \qquad B_3(\pm 1) = \frac{1}{6}, \qquad B_3'(0) = 0, \qquad B_3'(\pm 1) = \mp\frac{1}{2}.$$

Gesucht wird der kubische Spline

$$s(x) = \sum_{j=-1}^{n+1} c_j B_3\left(\frac{x-x_j}{h}\right), \qquad x \in [a,b] \tag{11.9}$$

mit den Interpolationsbedingungen $s(x_j) = y_j := f(x_j)$ mit $j = 0, \ldots, n$ sowie den Randbedingungen

$$s'(a) = a_1, \qquad s'(b) = b_1.$$

Die Randbedingungen werden dabei approximiert mittels

$$s'(a) \approx \frac{c_1 - c_{-1}}{2h} = a_1, \qquad s'(b) \approx \frac{c_{n+1} - c_{n-1}}{2h} = b_1.$$

Dann genügen die gesuchten Koeffizienten $c_{-1}, \ldots, c_{n+1}$ dem folgenden linearen Gleichungssystem

$$AC = F, \qquad C = (c_{-1}, ..., c_{n+1})^T$$

mit $F = (ha_1,\ y_1,\ \ldots, y_n,\ hb_1)^T \in \mathbb{R}^{n+3}$ und der schwach besetzten Matrix

$$A = \frac{1}{6}\begin{pmatrix} -3 & 0 & 3 & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & -3 & 0 & 3 \end{pmatrix} \in \mathbb{R}^{(n+3)\times(n+3)}.$$

Die Matrix A erfüllt das schwache Zeilensummenkriterium und ist unzerlegbar. Dann existiert die inverse Matrix A^{-1}. Ferner sind die grundlegenden Iterationsverfahren (vgl. Kapitel 6) zur Lösung des Gleichungssystems anwendbar.

11.4 Fehlerabschätzungen für Splines

Für den Fall stückweise linearer Splines aus $S_n^1[a,b]$ hatten wir im Beispiel 11.2 bereits Fehlerabschätzungen hergeleitet. Entsprechende Untersuchungen für Spline-Räume $S_n^m[a,b]$ mit $m>1$ sind wesentlich aufwendiger.

Wir zeigen hier exemplarisch für den Fall kubischer Splines Fehlerabschätzungen, die geeignet verallgemeinert werden können. Zunächst betrachten wir Funktionen $f \in C^2[a,b]$.

Satz 11.12. *Seien $f \in C^2[a,b]$ und $s \in S_n^3[a,b]$ der interpolierende kubische Spline. Ferner gelte eine der Randbedingungen (11.3) bzw. (11.5). Dann genügt der Interpolationsfehler den Beziehungen*

$$(i)\ \ \|f-s\|_\infty \le \frac{1}{2}h^{3/2}\|f''\|_2, \qquad (ii)\ \ \|f'-s'\|_\infty \le h^{1/2}\|f''\|_2$$

mit $h := \max_{j=1,\dots,n}|x_j - x_{j-1}|$ und $\|g\|_2 := (\int_a^b |g(t)|^2\ dt)^{\frac{1}{2}}$.

Beweis: (i) Die Fehlerfunktion $r = f - s$ hat $n+1$ Nullstellen $x_0,\dots,x_n$. Der maximale Abstand benachbarter Nullstellen ist kleiner oder gleich h. Nach dem Satz von Rolle gibt es dann n Nullstellen von r'. Der Maximalabstand benachbarter Nullstellen von r' ist höchstens $2h$.

Sei nun z so gewählt, daß $|r'(z)| = \|r'\|_\infty$. Für die zu z nächstliegende Nullstelle ξ von r' gilt offenbar $|\xi - z| \le h$. Unter Benutzung der Schwarzschen Ungleichung und von Lemma 11.6 folgt dann

$$\|r'\|_\infty^2 = \left|\int_\xi^z r''(y)dy\right|^2 \le \left(\int_\xi^z (r'')^2 dy\right)\cdot\int_\xi^z 1^2 dy \le h\int_a^b (r'')^2 dy \le h\|f''\|_2^2.$$

(ii) Wir wählen nun den Wert x so, daß $|r(x)| = \|r\|_\infty$. Mit der zu x nächstliegenden Nullstelle ζ von r gilt $|\zeta - x| \le h/2$ und folglich über (i)

$$\|r\|_\infty = \left|\int_\zeta^x r'(y)dy\right| \le \frac{h}{2}\|r'\|_\infty \le \frac{1}{2}h^{3/2}\|f''\|_2.$$

Damit ist die gesuchte Aussage bewiesen. □

Wir stellen jetzt höhere Glattheitsanforderungen an die zu interpolierende Funktion f, um eine verbesserte Fehlerabschätzung zu gewinnen. Zur Vorbereitung beweisen wir eine Abschätzung der zweiten Ableitungen für kubische Splines im Fall der Randbedingung (11.3). Dazu sei

$$c_j := s''(x_j), \quad j = 0,\dots,n.$$

Wegen $s'' \in \Pi_1[x_{j-1},x_j],\ j = 1,\dots n$ gilt

$$s''(x) = \frac{1}{h}\{c_{j-1}(x_j - x) + c_j(x - x_{j-1})\}, \quad x_{j-1} \le x \le x_j.$$

Zweimalige Integration und Elimination der Integrationskonstanten über die Interpolationsbedingungen $s(x_{j-1}) = f(x_{j-1}), s(x_j) = f(x_j)$ ergibt

$$\begin{aligned} s(x) &= \frac{1}{6h}[c_{j-1}(x_j - x)^3 + c_j(x - x_{j-1})^3 \\ &\quad + \{6f(x_{j-1}) - c_{j-1}h^2\}(x_j - x) + \{6f(x_j) - c_j h^2\}(x - x_j)], \\ s'(x) &= \frac{1}{6h}[-3c_{j-1}(x_j - x)^2 + 3c_j(x - x_{j-1})^2 \\ &\quad + 6\{f(x_j) - f(x_{j-1})\} - (c_j - c_{j-1})h^2]. \end{aligned}$$

Aus den Übergangsbedingungen $s'(x_j - 0) = s'(x_j + 0),\ \ j = 1, ..., n-1$ finden wir

$$c_{j-1} + 4c_j + c_{j+1} = \frac{6}{h^2}\{f(x_{j-1}) - 2f(x_j) + f(x_{j+1})\}, \quad j = 1, ..., n-1.$$

Aus den Randbedingungen $s'(a) = f'(a), s'(b) = f'(b)$ folgt

$$\begin{aligned} 4c_0 + 2c_1 &= \frac{12}{h^2}\{-f(x_0) + f(x_1) - hf'(x_0)\}, \\ 2c_{n-1} + 4c_n &= \frac{12}{h^2}\{f(x_{n-1}) - f(x_n) + hf'(x_n)\}. \end{aligned}$$

Somit erhalten wir das Gleichungssystem $AC = F$ für $C = (c_0, ..., c_n)^T$ mit

$$A = \begin{pmatrix} 4 & 2 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & 2 & 4 \end{pmatrix},$$

und

$$F = \frac{6}{h^2}\begin{pmatrix} 2[-f(x_0) + f(x_1) - hf'(x_0)] \\ f(x_0) - 2f(x_1) + f(x_2) \\ \vdots \\ f(x_{n-2}) - 2f(x_{n-1}) + f(x_n) \\ 2[f(x_{n-1}) - f(x_n) + hf'(x_n)] \end{pmatrix}.$$

Die unzerlegbare Tridiagonalmatrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ erfüllt das schwache Zeilensummenkriterium, insbesondere existiert A^{-1}. Jetzt geben wir eine Abschätzung für die Lösung des Systems an.

Lemma 11.13. *Unter der Voraussetzung $f \in C^2[a, b]$ gilt*

$$\|s''\|_\infty \leq 3\|f''\|_\infty.$$

Beweis: Wir wählen den Index j so, daß $|c_j| = \|c\|_\infty$. Dann folgt

$$4\|c\|_\infty = 4|c_j| \leq |F_j| - 2\|c\|_\infty \leq \|F\|_\infty - 2\|c\|_\infty,$$

also $2\|c\|_\infty \leq \|F\|_\infty$ bzw. $\|A^{-1}\|_\infty \leq \frac{1}{2}$.

Für $f \in C^2[a, b]$ ergibt die Taylor-Entwicklung

$$2[-f(x_0) + f(x_1) - hf'(x_0)] = h^2 f''(\xi_0), \quad \xi_0 \in [x_0, x_1]$$

$$2[f(x_{n-1}) - f(x_n) + hf'(x_n)] = h^2 f''(\xi_n), \quad \xi_n \in [x_{n-1}, x_n].$$

Die Restgliedabschätzung nach Satz 9.10 bei linearer Interpolation von $f(x_j)$ zeigt

$$f(x_{j-1}) - 2f(x_j) + f(x_{j+1}) = h^2 f''(\xi_j), \quad \xi_j \in [x_{j-1}, x_{j+1}]$$

und damit $\|f\|_\infty \leq 6\|f''\|_\infty$. Zusammenfassend ergibt sich

$$\|c\|_\infty \leq \|A^{-1}\|_\infty \|f\|_\infty \leq 3\|f''\|_\infty.$$

Damit ist die Behauptung bewiesen. □

Wir beweisen nun die gesuchte verbesserte Fehlerabschätzung bei erhöhten Glattheitsforderungen an die zu interpolierende Funktion..

Satz 11.14. *Seien $f \in C^4[a,b]$ und $s \in S_n^3[a,b]$ der interpolierende Spline bei äquidistanter Unterteilung von $[a,b]$ und der Randbedingung (11.3). Dann gilt*

$$\|f - s\|_\infty \leq \frac{1}{16} h^4 \|f^{(4)}\|_\infty.$$

Beweis: Sei $L_1 g$ der Polygonzug aus linearen Splines zu $g : [a,b] \to \mathbb{R}$. Nach Beispiel 11.2 gilt für $g \in C^2[x_{j-1}, x_j]$

$$\|g - L_1 g\|_\infty \leq \frac{1}{8} h^2 \|g''\|_\infty, \tag{11.10}$$

daher für $r = f - s$ und mit $L_1 r \equiv 0$

$$\|r\|_\infty = \|r - L_1 r\|_\infty \leq \frac{1}{8} h^2 \|r''\|_\infty. \tag{11.11}$$

Sei w Lösung der Differentialgleichung $w'' = L_1 f''$. Unter Anwendung der Dreiecksungleichung, von Lemma 11.13 auf $s - w \in S_n^3[a,b]$ sowie von (11.10) folgt

$$\begin{aligned} \|f'' - s''\|_\infty &\leq \|f'' - L_1 f''\|_\infty + \|L_1 f'' - s''\|_\infty \\ &\leq 4\|f'' - L_1 f''\|_\infty \\ &\leq \frac{1}{2} h^2 \|f^{(4)}\|_\infty. \end{aligned}$$

Daraus finden wir mit (11.11)

$$\|r\|_\infty = \|f - s\|_\infty \leq \frac{1}{8} h^2 \|f'' - s''\|_\infty \leq \frac{1}{16} h^4 \|f^{(4)}\|_\infty.$$

Damit ist das gesuchte Resultat bewiesen. □

Bemerkung 11.15. Die Fehlerabschätzungen in der Maximum-Norm für $m = 1$ (vgl. Beispiel 11.2)

$$\|f - s_1\|_\infty \leq C \|f''\|_\infty h^2, \qquad \text{falls} \quad f \in C^2[a,b]$$

und für $m = 3$ (vgl. Satz 11.14)

$$\|f - s_3\|_\infty \leq C \|f^{(4)}\|_\infty h^4, \qquad \text{falls} \quad f \in C^4[a,b]$$

sind optimal, d.h. nicht zu verbessern. Wir wollen das exemplarisch für die Spline-Approximation

Tabelle 11.1: Fehler in Maximum-Norm $\|f - s_1\|_\infty := \|f - s_1\|_{C[0,1]}$ für $f(x) = \sin x$

N	5	10	20	40	80
$\|f - s_1\|_\infty$	0.0035	0.0010	0.00026	0.000064	0.000016

1. Grades ($m = 1$) an die Funktion $f(x) = \sin x$ auf dem Intervall $[0,1]$ prüfen. Mittels Maple-Funktion `spline` ermitteln wir bei äquidistanter Aufteilung $x_i = i/N,\ i = 0..., N$ folgende Werte für den Fehler in der Maximum-Norm. Man erkennt in der Tat quadratische Konvergenz, denn

bei Halbierung der Schrittweite verkleinert sich der Fehler etwa um den Faktor $\frac{1}{4}$.

Generell hat man folgendes Resultat bei der Spline-Approximation: Interpoliert man eine Funktion $f \in C^{m+1}[a,b]$ auf einer Zerlegung $x_0, \ldots, x_n$ von $[a,b]$ (ggf. unter Zunahme von hinreichend vielen Zusatzbedingungen) durch eine Spline-Funktion $s_m \in S_n^m[a,b]$, erhält man in der Maximum-Norm bestenfalls die Approximationsordnung h^{m+1}.

Erfüllt die Funktion f nur geringere Glattheitsbedingungen, verschlechtert sich die Approximationsordnung. Im Satz 11.12 hatten wir für $m = 3$ derartige Abschwächungen diskutiert. □

Kapitel 12

Bezier-Kurven

Gegenstand dieses Kapitels ist eine elementare Einführung in die praxisrelevante Problematik des *Computer Aided Geometric Design (CAGD).* Hierbei geht es um die computergestützte Approximation und Darstellung von Kurven und Flächen. Wir beschränken uns hier auf den Fall von Kurven, d.h. Teilmengen $\Gamma \subset \mathbb{R}^m$, die sich durch eine Abbildung

$$x: \ I = [t_0, t_1] \subset \mathbb{R} \to \mathbb{R}^m, \quad x(\cdot) \in C[t_0, t_1]$$

beschreiben lassen.

Ziel ist eine solche Darstellung der Kurve im Computer, die das schnelle Zeichnen und leichte interaktive Manipulieren mit ihr erlauben. Dies wird dadurch gewährleistet, daß (im Unterschied z.B. zur Darstellung von Polynomen in der Monom-Basis) die Kurvenparameter eine geometrische Interpretation erlauben. Der Name *Bezier-Kurve* geht auf Bezier zurück, der seit 1962 bei dem Autohersteller Renault im CAGD-Bereich tätig war.

12.1 Bernstein-Polynome

Die mathematische Grundlage zur Beschreibung von Bezier-Kurven geben die sogenannten *Bernstein-Polynome.*

Definition 12.1. *Die Funktionen*

$$B_i^n(t) := \binom{n}{i} t^i (1-t)^{n-i}, \quad i = 0, ..., n, \quad n \in \mathbb{N} \tag{12.1}$$

heißen Bernstein-Polynome vom Grad n *auf* $[0, 1]$.

Eine Verallgemeinerung auf beliebige Intervalle $[a, b]$ mit $a < b$ erhalten wir mittels affin linearer Transformation

$$t \mapsto \lambda = \lambda(t) := \frac{t-a}{b-a}$$

von $[a, b]$ auf $[0, 1]$.

Definition 12.2. *Die Funktionen*

$$B_i^n(t; a, b) := B_i^n\left(\frac{t-a}{b-a}\right) = \frac{1}{(b-a)^n}\binom{n}{i}(t-a)^i(b-t)^{n-i}, \quad i = 0, ..., n \tag{12.2}$$

heißen Bernstein-Polynome vom Grad n *auf* $[a, b]$.

Bernstein-Polynome spielen auch eine wesentliche Rolle beim Beweis des wichtigen Approximationssatzes von Weierstraß. (Dabei geht es um die Annäherung von stetigen Funktionen durch Polynome, vgl. Approximationstheorie im Kurs *Numerische Mathematik I*).

Nachfolgend untersuchen wir Eigenschaften der Bernstein-Polynome. Aus der Definition ergeben sich folgende Aussagen.

Lemma 12.3. *Es gilt*

(i) $B_i^n \in \Pi_n,\ i = 0, ..., n;$ Grad $(B_i^n) = n$

(ii) $B_i^n(t) \geq 0, \quad t \in [0, 1]$

(iii) $\sum_{i=0}^{n} B_i^n(t) \equiv 1,\ \forall t \in \mathbb{R}$ (d.h. $B_i^n, i = 0, ..., n$ bilden eine Zerlegung der Einheit)

(iv) $t = 0$ ist $i-$fache Nullstelle, $t = 1$ ist $(n-i)-$fache Nullstelle von B_i^n.

(v) B_i^n hat in $[0, 1]$ genau ein Maximum bei $t = i/n$.

Beweis: Die Aussagen (i),(ii) und (iv) sind offensichtlich gültig.

(iii) Die Aussage folgt aus dem binomischen Satz

$$1 \equiv [t + (1 - t)]^n = \sum_{i=0}^{n} \binom{n}{i} t^i (1 - t)^{n-i}.$$

(v) Die erste Ableitung

$$\frac{d}{dt} B_i^n(t) = \binom{n}{i} (1 - t)^{n-i-1} t^{i-1} (i - nt), \quad i = 0, ..., n$$

verschwindet in (0,1) genau für $t_i = i/n$. Ferner wechselt die Ableitung in diesem Punkt das Vorzeichen. □

Weiterhin benötigen wir Rekursionsbeziehungen bzw. Zusammenhänge zwischen den Bernstein-Polynomen.

Lemma 12.4. *Es gilt*

(i) $B_i^n(t) = (1 - t)B_i^{n-1}(t) + tB_{i-1}^{n-1}(t), \quad t \in \mathbb{R},\ i = 1, ..., n$

(ii) $B_i^n(t) = B_{n-i}^n(1 - t), \quad i = 0, ..., n$

(iii) $(1 - t)B_0^n(t) = B_0^{n+1}(t); \quad tB_n^n(t) = B_{n+1}^{n+1}(t)$

(iv) $\frac{d}{dt} B_i^n = \begin{cases} -nB_0^{n-1}, & i = 0, \\ n(B_{i-1}^{n-1} - B_i^{n-1}), & i = 1, ..., n-1, \\ nB_{n-1}^{n-1}, & i = n. \end{cases}$

Beweis: (i)-(iii) Folgerung aus der Definition.
(iv) Differentiation ergibt

$$\frac{d}{dt}B_i^n(t) = \binom{n}{i}\{i(1-t)^{n-i}t^{i-1} - (n-i)(1-t)^{n-i-1}t^i\}.$$

Die Behauptung folgt nach Einsetzen von $i = 0, n$ bzw. $i = 1, ..., n-1$ und unter Beachtung der Definition der B_i^n. □

Wesentlich für die Darstellung von Bezier-Kurven ist der folgende

Satz 12.5. *$B_0^n, ..., B_n^n$ bilden eine Basis von Π_n.*

Beweis: Wir beachten zunächst die Aussage von Lemma 12.3 (i). Es bleibt dann zu zeigen, daß die Funktionen $B_0^n, ..., B_n^n$ linear unabhängig sind. Aus der Darstellung

$$\sum_{i=0}^{n} c_i B_i^n(t) = 0, \quad t \in [0,1].$$

folgt durch Differentiation

$$\sum_{i=0}^{n} c_i \frac{d^j}{dt^j} B_i^n(t) = 0, \quad t \in [0,1], \;\; j = 1, ..., n.$$

Lemma 12.3 (iv) ergibt dann für $t = 0$

$$\sum_{i=j}^{n} c_i \frac{d^j}{dt^j} B_i^n(0) = 0, \quad j = 0, ..., n,$$

denn $t = 0$ ist $i-$fache Nullstelle des Polynoms B_i^n. Induktiv folgt dann $c_n = \ldots = c_0 = 0$ und daraus die lineare Unabhängigkeit der Bernstein-Polynome. □

Bemerkung 12.6. Die Aussagen der Lemmata 12.3 und 12.4 sowie von Satz 12.5 übertragen sich auf die Bernstein-Polynome $B_i^n(t; a, b)$. □

12.2 Bezier-Polygone und Bezier-Kurven

Wir verallgemeinern zuerst den Begriff reellwertiger Polynome.

Definition 12.7. *Die Abbildung*

$$p : \mathbb{R} \to \mathbb{R}^m, \quad p(t) = \sum_{i=0}^{n} b_i t^i, \quad b_i \in \mathbb{R}^m, \quad i = 0, ..., n \tag{12.3}$$

mit $b_n \neq 0$ heißt polynomiale Kurve vom Grad n *in $\mathbb{R}^m$. Π_n^m bezeichnet den linearen Raum der polynomialen Kurven vom Grad n im $\mathbb{R}^m$.*

Neben der Monombasisdarstellung (12.3) wollen wir eine Basisdarstellung im Raum Π_n^m mit Hilfe der *Bernstein-Basis* bilden, d.h. bezüglich der Funktionen

$$p_j(t) = B_j^n(t; a, b), \quad j = 0, ..., n. \tag{12.4}$$

Wir betrachten somit polynomiale Kurven $p \in \Pi_n^m$ mit

$$p(t) = \sum_{i=0}^{n} b_i B_i^n(t; a, b), \quad t \in [a, b]. \tag{12.5}$$

Definition 12.8. *Die Koeffizienten $b_i \in \mathbb{R}^m$, $i = 0, ..., n$ in (12.5) heißen* Kontroll- *oder* Bezier-Punkte.

Der Streckenzug durch die Punkte $b_i, i = 0, ..., n$ heißt Bezier-Polygon.

Die durch Gleichung (12.5) beschriebene polynomiale Kurve wird als Bezier-Kurve *bezeichnet.*

Den Zusammenhang zwischen Bezier-Kurve und -Polygon beschreibt das

Lemma 12.9. *Es gilt*

(i) *Anfangs- und Endpunkte von Bezier-Kurve und -Polygon stimmen überein, d.h. $p(a) = b_0$, $p(b) = b_n$.*

(ii) *Die Bezier-Kurve liegt in der konvexen Hülle der Bezier-Punkte*

$$\text{con } \{b_0, ..., b_n\} := \left\{ \sum_{i=0}^{n} \alpha_i b_i : \ \alpha_i \geq 0, \quad \sum_{i=0}^{n} \alpha_i = 1 \right\}.$$

Beweis: (i) Folgerung aus der Definition.

(ii) Folgerung aus Lemma 12.3, (ii) und (iii) mit $\alpha_i = B_i^n(t : a, b), i = 1, \ldots, n$. □

Für den Berechnungsalgorithmus von Bezier-Kurven benötigen wir Aussagen über die Ableitungen des Bezier-Polynoms.

Satz 12.10. *Für das Bezier-Polynom $p(t) = \sum_{i=0}^{n} b_i B_i^n(t), t \in [0, 1]$ gilt*

$$p^{(j)}(t) = \frac{n!}{(n-j)!} \sum_{i=0}^{n-j} \Delta^j b_i \ B_i^{n-j}(t), \qquad j = 1, ..., n \tag{12.6}$$

mit den Vorwärtsdifferenzen $\Delta^j b_i$, die rekursiv definiert sind durch

$$\Delta^0 b_i := b_i, \quad \Delta^1 b_i := b_{i+1} - b_i, \quad \Delta^j b_i := \Delta^{j-1} b_{i+1} - \Delta^{j-1} b_i. \tag{12.7}$$

Beweis: Wir führen den Beweis mittels Induktion nach der Differentiationsordnung j durch. Der Induktionsanfang für $j = 0$ gilt wegen der Definition des Bezier-Polynoms.

Sei nun die Behauptung richtig für $j \in \{0, ..., n-1\}$. Die Induktionsbehauptung für $j+1$ folgt dann unter Benutzung von Lemma 12.4 (iv), durch Indexverschiebung und unter Beachtung der Definition der Vorwärtsdifferenzen aus

$$\begin{aligned}
p^{(j+1)}(t) &= \frac{n!}{(n-j)!} \sum_{i=0}^{n-j} \Delta^j b_i \frac{d}{dt} B_i^{n-j}(t) \\
&= \frac{n!}{(n-j-1)!} \left\{ \sum_{i=1}^{n-j} \Delta^j b_i \ B_{i-1}^{n-j-1}(t) - \sum_{i=0}^{n-j-1} \Delta^j b_i \ B_i^{n-j-1}(t) \right\} \\
&= \frac{n!}{(n-j-1)!} \sum_{i=0}^{n-j-1} \left\{ \Delta^j b_{i+1} - \Delta^j b_i \right\} B_i^{n-j-1}(t) \\
&= \frac{n!}{(n-(j+1))!} \sum_{i=0}^{n-(j+1)} \Delta^{j+1} b_i \ B_i^{n-(j+1)}(t).
\end{aligned}$$

Damit ist der Induktionsbeweis geführt. □

Eine wichtige Schlußfolgerung aus Satz 12.10 ist

Korollar 12.11. *Für das Bezier-Polynom aus Satz 12.10 gilt*

$$p^{(j)}(0) = \frac{n!}{(n-j)!}\Delta^j b_0, \quad p^{(j)}(1) = \frac{n!}{(n-j)!}\Delta^j b_{n-j}. \tag{12.8}$$

Damit hängen die Werte $p^{(j)}(0)$ und $p^{(j)}(1)$ nur von den Bezier-Punkten $b_0, ..., b_j$ bzw. $b_{n-j}, ..., b_n$ ab. Speziell hat die Bezier-Kurve in den Endpunkten die gleiche Tangente wie das Bezier-Polygon, d.h.

$$p'(0) = n(b_1 - b_0), \qquad p'(1) = n(b_n - b_{n-1}). \tag{12.9}$$

Beispiel 12.12. ($n = 0, 1, 2$ und $m = 2$)

Für $n = 0$ besteht die Bezier-Kurve natürlich nur aus dem Anfangspunkt b_0. Im Fall $n = 1$ stimmt die Bezier-Kurve mit dem Bezier-Polygon, d.h. der Geraden zwischen b_0 und b_1, überein. Somit entsteht für $n = 2$ der erste wirklich interessante Fall.

Die quadratische Kurve liegt offenbar im durch die Punkte b_0, b_1 und b_2 gebildeten Dreieck, dies ist auch deren konvexe Hülle. Sind die drei Punkte nicht kollinear, so ist b_1 genau der Schnittpunkt der Tangenten an die Bezier-Kurve in den Punkten b_0 und b_1.

Wir wählen die Anfangs- und Endpunkt b_0 bzw. b_2 fest und ändern den Zwischenpunkt b_1. Die Abbildung 12.1 zeigt die Kurvenänderung durch Manipulation am Zwischenpunkt. □

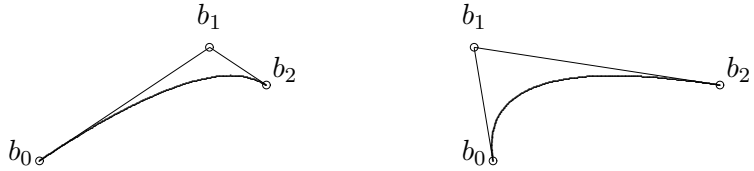

Abbildung 12.1: Bezier-Polynome der Ordnung 2

Beispiel 12.13. ($n = 2$ und $m = 2$)

Wir demonstrieren das Verbinden von Bezier-Kurven für den Fall $n = 2$.

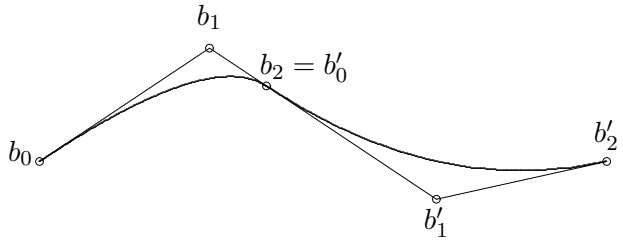

Abbildung 12.2: Bezier-Splines der Ordnung 2

In Abbildung 12.2 werden zwei Bezier-Kurven gleichen Grades zusammengefügt im Punkt $b_2 = b'_0$. Der stetige Übergang in der 1. Ableitung im Koppelpunkt ist nach (12.9) gewährleistet. □

12.3 Algorithmus von de Casteljau

Ziel ist ein stabiles und schnelles Verfahren zur

- Berechnung der Werte $p(t)$ einer Bezier-Kurve durch fortgesetzte Konvexkombination und
- Zerlegung einer Bezier-Kurve in Teilkurven vom gleichen Grad.

Bei Wiederholung der letzteren Prozedur (z.B. durch fortlaufende Halbierung des Intervalles) soll die Ausgangskurve approximiert werden.

Ausgangspunkt ist das Bezier-Polynom auf $[0,1]$ mit

$$p(t) = \sum_{i=0}^{n} b_i B_i^n(t), \quad t \in [0,1].$$

Dann erklären wir *Teilpolynome* $b_i^k \in \Pi_k^m$ unter Beachtung von Lemma 12.4 (ii) durch

$$\begin{aligned} b_i^k(t) &:= \sum_{j=0}^{k} b_{i+j} B_j^k(t) \\ &= \sum_{j=0}^{k} b_{i+j} B_{k-j}^k(1-t), \quad i = 0,...,n-k, \;\; k = 0,...,n. \end{aligned} \tag{12.10}$$

Ein derartiges Teilpolynom stellt also das durch die Kontrollpunkte $b_i, ..., b_{i+k}$ definierte Bezier-Polynom vom Grad k dar. Speziell ist $b_0^n = p$ das Ausgangspolynom.

Man kann nun nach dieser Idee eine dem Neville-Schema verwandte Rekursionsformel bestimmen, die die Grundlage für den *Algorithmus von de Casteljau* bildet. Dabei entsteht das Teilpolynom vom Grad k durch Konvexkombination von zwei Teilpolynomen vom Grad $k-1$.

Satz 12.14. *Die Teilpolynome b_i^k des Bezier-Polynoms p erfüllen die Rekursionsbeziehung*

$$b_i^k(t) = (1-t) b_i^{k-1}(t) + t b_{i+1}^{k-1}(t), \quad i = 0,...,n-k, \;\; k = 1,...,n. \tag{12.11}$$

Beweis: Nach Definition (12.10) sowie nach Lemma 12.4 (i),(iii) finden wir aus der Rechnung

$$\begin{aligned} b_i^k(t) &= b_i B_0^k(t) + \sum_{j=1}^{k-1} b_{i+j} B_j^k(t) + b_{i+k} B_k^k(t) \\ &= b_i (1-t) B_0^{k-1}(t) + \sum_{j=1}^{k-1} b_{i+j} \{(1-t) B_j^{k-1}(t) + t B_{j-1}^{k-1}(t)\} + b_{i+k} t B_{k-1}^{k-1}(t) \\ &= \sum_{j=0}^{k-1} b_{i+j} (1-t) B_j^{k-1}(t) + \sum_{j=1}^{k} b_{i+j} t B_{j-1}^{k-1}(t) \\ &= (1-t) b_i^{k-1}(t) + t b_{i+1}^{k-1}(t) \end{aligned}$$

die gesuchte Aussage. □

Wir können nun das gewünschte Rekursionsschema zur Funktionswertberechnung von $p(t)$ herleiten. Unter Beachtung von $b_i^0(t) = b_i$ und $b_0^n(t) = p(t)$ läßt sich $p(t)$ nach Satz 12.14 durch sukzessive Konvexkombination aus den Werten $b_0, ..., b_n$ analog zum Schema der dividierten Differenzen (vgl. Kap. 9) ermitteln.

Wir notieren das entstehende Schema von DE CASTELJAU speziell für $n = 3$ in Abb. 12.3.

$$\begin{array}{llll} b_0 = b_0^0 & & & \\ & b_0^1 & & \\ b_1 = b_1^0 & & b_0^2 & \\ & b_1^1 & & b_0^3 = p(t) \\ b_2 = b_2^0 & & b_1^2 & \\ & b_2^1 & & \\ b_3 = b_3^0 & & & \end{array}$$

Abbildung 12.3: Schema von de Casteljau für $n = 3$

Dieses Schema kann jetzt auch zur Zerlegung einer Bezier-Kurve in Teilkurven gleichen Grades verwendet werden. Genauer zerlegen wir $p(t)$ in zwei Bezier-Polynome auf $[0, t]$ bzw. $[t, 1]$, die mit dem Ausgangspolynom übereinstimmen.

Satz 12.15. *Die beiden Bezier-Polynome*

$$p_1(t) := \sum_{i=0}^{n} b_0^i(\lambda) B_i^n(t; 0, \lambda), \qquad p_2(t) := \sum_{i=0}^{n} b_i^{n-i}(\lambda) B_i^n(t; \lambda, 1),$$

die aus den Koeffizienten der beiden Randdiagonalen des Schemas von de Casteljau gebildet werden, genügen für jedes $\lambda \in (0, 1)$ der Beziehung

$$p(t) = p_1(t) = p_2(t), \quad t \in \mathbb{R}.$$

Beweis: Wir zeigen die Beziehung nur für p_1. Der entsprechende Nachweis von $p = p_2$ erfolgt völlig analog.

Nach Definition (12.10) und nach Umsummieren finden wir

$$p_1(t) = \sum_{i=0}^{n} \left(\sum_{j=0}^{i} b_j B_j^i(\lambda) \right) B_i^n(t; 0, \lambda) = \sum_{j=0}^{n} b_j \left(\sum_{i=j}^{n} B_j^i(\lambda) B_i^n(t; 0, \lambda) \right).$$

Zum Beweis bleibt zu zeigen, daß

$$B_j^n(t) = \sum_{i=j}^{n} B_j^i(\lambda) B_i^n(t; 0, \lambda), \ \ \forall t \in \mathbb{R}.$$

Unter Anwendung der Definition der Bernstein-Polynome, durch Indexverschiebung und unter Benutzung des binomischen Satzes ergibt sich

$$\begin{aligned} \sum_{i=j}^{n} B_j^i(\lambda) B_i^n(t; 0, \lambda) &= \sum_{i=j}^{n} \left[\binom{i}{j} (1-\lambda)^{i-j} \lambda^j \right] \left[\lambda^{-n} \binom{n}{i} t^i (\lambda - t)^{n-i} \right] \\ &= \binom{n}{j} \lambda^{j-n} \sum_{i=j}^{n} \binom{n-j}{i-j} (1-\lambda)^{i-j} t^i (\lambda - t)^{n-i} \\ &= \binom{n}{j} \lambda^{j-n} t^j \sum_{i=0}^{n-j} \binom{n-j}{i} (1-\lambda)^i t^i (\lambda - t)^{n-j-i} \end{aligned}$$

$$
\begin{aligned}
&= \binom{n}{j}\lambda^{j-n}t^j\ [(1-\lambda)t+\lambda-t]^{n-j} \\
&= \binom{n}{j}(1-t)^{n-j}t^j \\
&= B_j^n(t).
\end{aligned}
$$

Daraus folgt die Behauptung des Satzes. □

In der Praxis wendet man dann Satz 12.15 mit $\lambda = \frac{1}{2}$ an, d.h. die durch die Bezier-Punkte $b_0, \ldots, b_n$ definierte Bezier-Kurve wird halbiert. Wiederholte Durchführung dieses Verfahrens führt auf eine Folge von glatt zusammengesetzten Bezier-Kurven gleicher Ordnung. Diese Folge konvergiert hinreichend schnell gegen die ursprüngliche Bezier-Kurve. Damit ist das Verfahren zur effizienten Visualisierung im Rechner geeignet, wenn man bei hinreichend starker Verfeinerung stückweise Bezier-Polygone zeichnet. Es wurde auch bei den Abbildungen 12.1 und 12.2 benutzt.

Kapitel 13

Numerische Integration nach Newton-Cotes

Ziel der beiden folgenden Kapitel ist die Näherungsberechnung linearer Funtionale, die sich in Form reellwertiger Integrale

$$I(f) := \int_a^b f(x)dx.$$

darstellen lassen. Eine derartige Aufgabenstellung tritt z.B. auf, wenn

- das Integral nicht geschlossen auswertbar ist,
- der Integrand nur punktweise (etwa durch Messungen) bekannt ist oder
- die Lösung von Differential- oder Integralgleichungen numerisch ermittelt werden soll.

Bei den letztgenannten Anwendungen ist die Integralauswertung ein kleines Teilproblem, das aber sehr oft und damit sehr effizient ausgeführt werden muß.

13.1 Interpolationsquadraturen

Betrachtet werden nachfolgend *Quadraturformeln* der Form

$$\int_a^b f(x)dx \approx Q_n(f) \ := \ \sum_{i=0}^{n} a_i f(x_i) \tag{13.1}$$

mit den paarweise verschiedenen *Stützstellen* $x_0, ..., x_n$ aus dem Intervall $[a, b]$ und *Gewichten* $a_0, ..., a_n \in \mathbb{R}$. Es liegt auf der Hand, statt der Funktion f ein Interpolationspolynom zu wählen und das entstehende Integral exakt auszuwerten.

Definition 13.1. *Eine Quadraturformel der Gestalt (13.1) heißt* Interpolationsquadratur der Ordnung n, *falls*

$$\sum_{i=0}^{n} a_i f(x_i) = \int_a^b (L_n f)(x)dx, \quad \forall f \in C[a, b]. \tag{13.2}$$

Dabei sei $L_n f \in \Pi_n[a, b]$ das (eindeutig bestimmte) Interpolationspolynom zu f mit den Stützstellen $x_0, ..., x_n$.

Die nachfolgende Charakterisierung von Interpolationsquadraturen der Ordnung n wird sich als

bequem für die weiteren Überlegungen erweisen.

Satz 13.2. *Eine Quadraturformel ist eine Interpolationsquadratur der Ordnung n genau dann, wenn alle Polynome $p \in \Pi_n[a,b]$ exakt integriert werden, also*

$$\sum_{i=0}^{n} a_i p(x_i) = \int_a^b p(x)dx, \quad \forall p \in \Pi_n[a,b]. \tag{13.3}$$

Beweis: $\Rightarrow$) Wegen $L_n p \equiv p$, für alle $p \in \Pi_n[a,b]$ folgt (13.3) aus (13.2).
$\Leftarrow$) Andererseits folgt (13.2) aus (13.3) wegen

$$\int_a^b (L_n f)(x)dx = \sum_{i=0}^{n} a_i (L_n f)(x_i) = \sum_{i=0}^{n} a_i f(x_i), \quad \forall f \in C[a,b]. \qquad \square$$

Die Existenz und Eindeutigkeit von Interpolationsquadraturen zeigt der

Satz 13.3. *(i) Es existiert genau eine Interpolationsquadratur der Ordnung n zu paarweise verschiedenen Stützstellen $x_0, ..., x_n$.*

(ii) Die Gewichte ergeben sich mit $\omega_{n+1}(x) := \prod_{i=0}^{n}(x - x_i)$ aus

$$a_i = \frac{1}{\omega'_{n+1}(x_i)} \int_a^b \frac{\omega_{n+1}(x)}{x - x_i} dx, \quad i = 0, ..., n.$$

Beweis: (i) Existenz und Eindeutigkeit der Quadraturformel ergeben sich aus der Existenz und Eindeutigkeit des Interpolationspolynoms $L_n f$ zur Funktion f, vgl. dazu auch Satz 9.1.
(ii) Mit Hilfe der Lagrange-Darstellung des Interpolationspolynoms finden wir

$$\int_a^b (L_n f)(x)dx = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x)dx \ ,$$

also

$$a_i = \int_a^b l_i(x)dx = \int_a^b \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} dx = \frac{1}{\omega'_{n+1}(x_i)} \int_a^b \frac{\omega_{n+1}(x)}{x - x_i} dx. \qquad \square$$

Zur Vereinfachung der Formeln beschränken wir uns auf den Fall äquidistanter Stützstellen.

Definition 13.4. *Die Interpolationsquadratur der Ordnung n zu den Stützstellen $x_i = a + ih$, $i = 0, ..., n$ mit der Schrittweite $h = (b-a)/n$ heißt* Newton-Cotes Formel der Ordnung n.

Lemma 13.5. *Die Gewichte der Newton-Cotes Formel der Ordnung n ergeben sich aus*

$$a_i = hA_i, \quad A_i = A_{n-i} = \frac{(-1)^{n-i}}{i!(n-i)!} \int_0^n \prod_{\substack{k=0 \\ k \neq i}}^{n} (z - k)dz, \quad i = 0, ...n. \tag{13.4}$$

Beweis: Wir substituieren $x - x_0 = hz$ in der Formel für die Gewichte a_i im Satz 13.3. Dabei benötigen wir für die Herleitung von (13.4) speziell

$$\omega_{n+1}(x) = h^{n+1} \prod_{k=0}^{n}(z - k), \qquad \omega'_{n+1}(x_i) = (-1)^{n-i}\ i!\ (n-i)!\ h^n.$$

Die Symmetriebeziehung $A_i = A_{n-i}$ ergibt sich nach Substitution $z = n - y$. □

Einfacher als die Berechnung der Gewichte nach Lemma 13.5 ist ihre Ermittlung über die Lösung eines linearen Gleichungssystems, das (unter Beachtung von Satz 13.2) bei exakter Integration der Monome x^i bis zur Ordnung n entsteht. Wir betrachten einige *Spezialfälle:*

Trapez-Regel ($n = 1$:) Wir setzen o.B.d.A. $a = -1, b = 1$, also $h = 2$. Dann ergibt sich das System

$$\begin{aligned} \int_{-1}^{1} x^0 dx = 2 &= a_0 + a_1 = 2A_0 + 2A_1, \\ \int_{-1}^{1} x^1 dx = 0 &= -a_0 + a_1 = -2A_0 + 2A_1 \end{aligned}$$

mit der Lösung $a_0 = a_1 = 1$ bzw. $A_0 = A_1 = 1/2$. Es ergibt sich die *Trapez-Regel*

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\{f(a) + f(b)\} = \frac{h}{2}\{f(x_0) + f(x_1)\}.$$

Simpson-Regel ($n = 2$): Mit $h = 1$ und $a_i = A_i,\ i = 0, 1, 2$ gilt

$$\begin{aligned} \int_{-1}^{1} x^0 dx = 2 &= a_0 + a_1 + a_2 = A_0 + A_1 + A_2, \\ \int_{-1}^{1} x^1 dx = 0 &= -a_0 + a_2 = -A_0 + A_2, \\ \int_{-1}^{1} x^2 dx = \frac{2}{3} &= a_0 + a_2 = A_0 + A_2 \end{aligned}$$

mit der Lösung $A_0 = A_2 = \frac{1}{3}, A_1 = \frac{4}{3}$. Daraus ergibt sich die *Simpson-Regel*

$$\begin{aligned} \int_a^b f(x)dx &\approx \frac{b-a}{6}\{f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\} \\ &= \frac{h}{3}\{f(x_0) + 4f(x_1) + f(x_2)\}. \end{aligned}$$

In Tabelle 13.1 stellen wir die Gewichte der Newton-Cotes Formeln bis zur Ordnung $n = 5$ zusammen. Offenbar sind die modifizierten Gewichte A_i nur von der Zahl der Stützstellen, jedoch nicht von a, b und h abhängig. Die ursprünglichen Gewichte erhält man durch $a_i = hA_i$ mit dem gemeinsamen Faktor $h = (b - a)/n$.

Bemerkung 13.6. Ab $n \geq 8$ wird ein Teil der Gewichte negativ. Das bewirkt eventuell Auslöschungseffekte; ggf. liefert das negative Näherungswerte bei positiven Integranden. □

13.2 Fehlerabschätzungen

Wir leiten nun Fehlerabschätzungen

$$R_n(f) := I(f) - Q_n(f) = \int_a^b f(x)\ dx - \sum_{i=0}^{n} a_i f(x_i)$$

für Interpolationsquadraturen der Ordnung n zur Berechnung des Integrals $I(f)$ ab. Exemplarisch betrachten wir zunächst den Fall der Trapez-Regel.

h	A_0	A_1	A_2	A_3	A_4	A_5	Bezeichnung
1	$\frac{1}{2}$	$\frac{1}{2}$					Trapez-Regel
2	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$				Simpson-Regel
3	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{9}{8}$	$\frac{3}{8}$			Newton-3/8-Regel
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{14}{45}$		1. Milne-Regel
5	$\frac{95}{288}$	$\frac{375}{288}$	$\frac{250}{288}$	$\frac{250}{288}$	$\frac{375}{288}$	$\frac{95}{288}$	2. Milne-Regel

Tabelle 13.1: Gewichte einiger Newton-Cotes-Formeln

Satz 13.7. *Unter der Voraussetzung $f \in C^2[a,b]$ gilt für den Fehler der Trapez-Regel*

$$R_1(f) = \int_a^b f(x)dx - \frac{h}{2}[f(a)+f(b)] = -\frac{h^3}{12}f''(\xi)$$

mit $h = b-a$ und einer Zwischenstelle $\xi \in (a,b)$.

Beweis: Mit der linearen Interpolierenden $L_1 f$ von f zu den Stützstellen $x_0 = a$ und $x_1 = b$ gilt per Konstruktion

$$R_1(f) := \int_a^b [f(x) - (L_1 f)(x)]dx = \int_a^b (x-a)(x-b)\frac{f(x)-(L_1f)(x)}{(x-a)(x-b)}dx.$$

Der Term $(x-a)(x-b)$ wechselt in $[a,b]$ nicht das Vorzeichen und ist dort stetig. Mittels Regel von l'Hospital und unter Beachtung der Interpolationseigenschaft sieht man, daß der Bruch im Integranden zu $C[a,b]$ gehört. Dann zeigt der erweiterte Mittelwertsatz der Integralrechnung

$$R_1(f) = \frac{f(\eta)-(L_1f)(\eta)}{(\eta-a)(\eta-b)}\int_a^b (x-a)(x-b)dx, \quad \eta \in [a,b].$$

Die Restglieddarstellung nach Satz 9.10 für den Fall linearer Interpolation und die Nebenrechnung

$$\int_a^b (x-a)(x-b)dx = -\frac{(b-a)^3}{6}$$

führen auf die Behauptung. □

Wir betrachten jetzt eine Verallgemeinerung dieser Fehleraussage für beliebige $n \in \mathbb{N}_0$. Wesentliches Hilfsmittel für den Beweis ist die folgende Fehlerdarstellung mit Hilfe des *Peano-Kerns*.

Lemma 13.8. *Für eine Funktion $f \in C^{m+1}[a,b]$ mit $m \in \mathbb{N}_0$ und $0 \le m \le n$ gilt die Fehlerdarstellung*

$$R_m(f) := I(f) - Q_n(f) \;=\; \int_a^b f^{(m+1)}(t)K_m(t)\,dt \tag{13.5}$$

mit dem Peano-Kern

$$K_m(t) := \frac{1}{m!}R_x[(x-t)_+^m], \qquad (x-t)_+^m := \begin{cases} (x-t)^m & \text{für} \quad x \ge t \\ 0 & \text{für} \quad x < t \end{cases}. \tag{13.6}$$

Dabei heißt $R_x[(x-t)_+^m]$, daß R_m auf das Argument $(\cdot - t)_+^m$ als Funktion in x anzuwenden ist.

Beweis: *Übungsaufgabe !* □

Die gesuchte Fehlerabschätzung der Interpolationsquadratur liefert folgender Satz.

Satz 13.9. *Sei $f \in C^{(n+1)}[a,b]$ für $n \in \mathbb{N}$. Ferner ändere der Peano-Kern K_n auf dem Intervall $[a,b]$ nicht das Vorzeichen. Dann gibt es eine Zahl $\xi \in (a,b)$, so daß*

$$R_n(f) = \frac{1}{n!} f^{(n+1)}(\xi)\ R_n(x^{n+1}). \tag{13.7}$$

Beweis: *Übungsaufgabe !* □

Bei Anwendung dieses Resultates für die Trapez-Regel, d.h. $n = 1$, erhält man wieder das Ergebnis von Satz 13.7 (vgl. Übungsaufgabe).

Eine genauere Analyse zeigt, daß die mit Satz 13.9 zu gewinnende Fehlerabschätzung für gerade Zahlen n nicht optimal sein muß. Im folgenden Satz finden wir für die Simpson-Regel, d.h. $n = 2$, eine noch bessere Fehlerabschätzung. Beim Beweis des Resultates verwenden wir die sogenannte *Hermitesche Interpolierende*.

Satz 13.10. *Unter der Voraussetzung $f \in C^4[a,b]$ gilt für den Fehler der Simpson-Regel*

$$R_2(f) = \int_a^b f(x)dx - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] = -\frac{h^5}{90} f^{(4)}(\xi)$$

mit $h = (b-a)/2$ und einer Zwischenstelle $\xi \in [a,b]$.

Beweis. Sei $L_2 f$ die quadratische Interpolierende von f zu den Stützstellen $x_0 = a, x_1 = (a+b)/2, x_2 = b$. Für das Polynom $q \in \Pi_3$

$$q(x) := (L_2 f)(x) + \frac{4}{(b-a)^2}[(L_2 f)'(x_1) - f'(x_1)]\omega_2(x) \tag{13.8}$$

mit $\omega_2(x) := (x-x_0)(x-x_1)(x-x_2)$ gelten die Interpolationseigenschaften

$$q(x_i) = f(x_i),\ \ i = 0,1,2; \qquad q'(x_1) = f'(x_1).$$

Aus Symmetriegründen ist $\int_a^b \omega_2(x)dx = 0$. Dann gilt

$$\begin{aligned} R_2(f) &:= \int_a^b [f(x) - (L_2 f)(x)]\ dx \\ &= \int_a^b [f(x) - q(x)]\ dx \\ &= \int_a^b \omega_2(x)(x-x_1)\frac{f(x)-q(x)}{\omega_2(x)(x-x_1)}\ dx. \end{aligned}$$

Offenbar ist $\omega_2(x)(x-x_1) \le 0,\ \ x \in [a,b]$. Der im Bruch stehende Term im Integranden ist stetig auf $[a,b]$, wie man nach zweifacher Anwendung der Regel von l'Hospital und bei Beachtung der Interpolationseigenschaft sieht. Dann ist der erweiterte Mittelwertsatz der Integralrechnung anwendbar. Es ergibt sich hiermit

$$R_2(f) = \frac{f(\eta) - q(\eta)}{\omega_2(\eta)(\eta - x_1)} \int_a^b \omega_2(x)(x-x_1)dx, \quad \eta \in [a,b].$$

Analog zu Satz 9.10 ermittelt man für die kubische Interpolierende q folgende Fehlerabschätzung

$$(f-q)(\eta) = \frac{1}{4!} f^{(4)}(\xi)\omega_2(\xi)(\xi - x_1), \quad \xi \in [a,b].$$

Unter Benutzung von

$$\int_a^b \omega_2(x)(x-x_1)dx = -\frac{(b-a)^5}{120}$$

folgt die Behauptung des Satzes. □

Beispiel 13.11. *Approximation von* $\int_0^1 \frac{1}{x+1}dx = \ln 2$

Die Trapez-Regel ergibt den Näherungswert 0.75, damit $\ln 2 - 0.75 = -0.056....$ Nach Satz 13.7 erhält man mit $|R_1(f)| \leq 1/6$ eine Überschätzung des tatsächlichen Fehlers.

Bei Anwendung der Simpson-Regel findet man den Näherungswert $\frac{25}{36} \approx 0.694444$, also $\ln 2 - 0.694444 = 0.001297....$ Nach Satz 13.10 finden wir die Fehlerschranke $|R_2(f)| \leq 0.0084$, damit wird der Fehler wieder überschätzt. □

Beispiel 13.12. *Approximation von* $\int_{-4}^4 \frac{dx}{1+x^2} = 2\arctan 4 \approx 2.651...$

Bei Anwendung der Newton-Cotes Formeln findet man die folgenden unakzeptablen Näherungswerte

n	1	2	3	4	6	8
$Q_n(f)$	0.471	5.490	2.277	2.278	3.329	1.941

Ursache sind die aus dem Kapitel über polynomiale Interpolation bekannten möglicherweise (und hier tatsächlich) auftretenden Oszillationen der Interpolatiomspolynome bei wachsendem Grad n. □

13.3 Zusammengesetzte Newton-Cotes Formeln

In Analogie zur Spline-Interpolation zerlegt man bei den *zusammengesetzten Newton-Cotes Formeln* das Integrationsintervall in m Teilintervalle und wendet dort jeweils eine Quadraturformel niedriger Ordnung an. Es zeigt sich, daß die Konvergenz der zusammengesetzten Integrationsformeln bei relativ geringen Glätteanforderungen an den Integranden für $m \to \infty$ erzielt wird.

Wendet man bei äquidistanter Zerlegung von $[a,b]$ in Teilintervalle der Breite $h = (b-a)/m$ in jedem Teilintervall die Trapezregel an, so entsteht die *zusammengesetzte Trapez-Regel*

$$\int_a^b f(x)dx \approx T_h(f) := h\left\{\frac{1}{2}f(x_0) + \sum_{i=1}^{m-1} f(x_i) + \frac{1}{2}f(x_m)\right\}, \quad x_i := a + ih, \;\; i = 0,...,m.$$

Satz 13.13. *Sei* $f \in C^2[a,b]$. *Dann gilt für den Fehler der zusammengesetzten Trapez-Regel*

$$\int_a^b f(x)dx - T_h(f) = -\frac{b-a}{12}h^2 f''(\xi), \;\; \xi \in [a,b].$$

Beweis: Satz 13.7 kann intervallweise angewendet werden. Dabei werden die Integrale für die Restglieder vor Anwendung des Mittelwertsatzes zu einem Integral zusammengefaßt. □

Wir wenden nun bei äquidistanter Zerlegung in eine gerade Anzahl von Teilintervallen jeweils die Simpson-Regel an. Für die entstehende *zusammengesetzte Simpson-Regel*

$$\int_a^b f(x)dx \approx S_h(f) := \frac{h}{3}\{f(x_0) + 4f(x_1) + 2f(x_2) + ... + 2f(x_{m-2}) + 4f(x_{m-1} + f(x_m)\}$$

erhalten wir dann den

Satz 13.14. *Für den Fehler der zusammengesetzten Simpson-Regel mit den Stützstellen $x_i = a + ih, i = 0, ..., m$ und der Schrittweite $h = (b-a)/m$ und gerader Zahl der Stützstellen gilt*

$$\int_a^b f(x)dx - S_h(f) = -\frac{b-a}{180}h^4 f^{(4)}(\xi)$$

mit $\xi \in [a, b]$, sofern $f \in C^4[a, b]$.

Beweis: Analog zu Satz 13.13 unter Benutzung von Satz 13.10. □

Wir untersuchen den Effekt der zusammengesetzten Integrationsformeln anhand der Integrale aus den Beispielen 13.11 und 13.12.

Beispiel 13.15. *Approximation von $\int_0^1 \frac{dx}{x+1}$*

Die Tabelle zeigt jeweils den Fehler für die zusammengesetzte Trapez-Regel (zTR) bzw. Simpson-Regel (zSR) bei unterschiedlicher Anzahl n von Teilintervallen.

m	1	2	4	8	16
zTR	-0.5685282	-0.01518615	-0.00387663	-0.00097467	-0.00024402
zSR		-0.00129726	-0.00010679	-0.00000735	-0.00000047

Die theoretisch ermittelten Konvergenzordnungen 2 bzw. 4 nach den Sätzen 13.13 und 13.14 sind aus der Tabelle tatsächlich zu erkennen. □

Beispiel 13.16. *Approximation von $\int_{-4}^4 \frac{dx}{1+x^2} \approx 2.651$*

Die Tabelle zeigt die Näherungswerte bei Anwendung der zusammengesetzten Trapez-Regel bei unterschiedlicher Anzahl n von Teilintervallen.

m	1	2	4	8
zTR	0.4706	4.2350	2.9176	2.6588

Dieses Beispiel zeigt im Vergleich zu Beispiel 13.12 sehr überzeugend den Vorteil zusammengesetzter Newton-Cotes Formeln. □

13.4 Konvergenz von Quadraturformeln (Exkurs)

Wir untersuchen nun, ob die Quadraturformeln für Integranden $f \in C[a, b]$ für $n \to \infty$ konvergieren. Der folgende Satz von SZEGÖ gibt eine hinreichende Bedingung für die Existenz des Integrals.

Satz 13.17. *Für die Folge von Quadraturformeln $Q_n(f) = \sum_{i=0}^{n} a_i^{(n)} f(x_i^{(n)}),\ n = 1, 2, ...$ gelte:*

(i) *Die Quadraturformeln konvergieren für alle Polynome, d.h. für alle Polynome p gilt*

$$\lim_{n\to\infty} Q_n(p) = \int_a^b p(x)dx.$$

(ii) *Für die Koeffizienten der Quadraturformel existiert $C > 0$ mit:* $\sum_{i=0}^{n} |a_i^{(n)}| \le C, \ \forall n \in \mathbb{N}$.

Dann konvergieren die Quadraturformeln für alle Funktionen $f \in C[a,b]$, d.h. es gilt

$$\lim_{n\to\infty} Q_n(f) = \int_a^b f(x)dx.$$

Beweis: Nach dem Approximationssatz von Weierstraß (vgl. Approximationstheorie im Kurs *Numerische Mathematik II*) gibt es für $f \in C[a,b]$ und $\epsilon > 0$ stets ein Polynom p, so daß

$$\|f - p\|_\infty \le \frac{\epsilon}{2(C + b - a)}.$$

Wegen (i) findet man zu jedem Polynom p einen Index $N(\epsilon) \in \mathbb{N}$ derart, daß

$$|Q_n(p) - \int_a^b p(x)dx| \le \frac{\epsilon}{2}, \quad \forall n \ge N(\epsilon).$$

Mittels Dreiecksungleichung folgt dann für alle $n \ge N(\epsilon)$

$$\begin{aligned}
&\left|Q_n(f) - \int_a^b f(x)dx\right| \\
\le\ & |Q_n(f) - Q_n(p)| + \left|Q_n(p) - \int_a^b p(x)dx\right| + \left|\int_a^b p(x)dx - \int_a^b f(x)dx\right| \\
\le\ & \sum_{i=0}^{n} |a_i^{(n)}||f(x_i^{(n)}) - p(x_i^{(n)})| + \left|Q_n(p) - \int_a^b p(x)dx\right| + \int_a^b |p(x) - f(x)|dx \\
\le\ & \frac{C\epsilon}{2(C + b - a)} + \frac{\epsilon}{2} + \frac{(b-a)\epsilon}{2(C + b - a)} = \epsilon. \qquad \square
\end{aligned}$$

Bemerkung 13.18. Es kann mittels funktionalanalytischer Methoden (Satz von Banach-Steinhaus) gezeigt werden, daß die Voraussetzungen des Satzes 13.17 auch notwendig für die Konvergenz der Quadraturformeln für beliebige Funktionen $f \in C[a,b]$ sind. $\square$

Eine unmittelbare Anwendung von Satz 13.17 ist das folgende Resultat von Steklov:

Satz 13.19. *Für eine Folge von Quadraturformeln* $Q_n(f) = \sum_{i=0}^{n} a_i^{(n)} f(x_i^{(n)}), \quad n = 1, 2, \ldots$ *gelte:*

(i) Die Quadraturformeln konvergieren für alle Polynome.
(ii) Für die Gewichte gilt $a_i^{(n)} \ge 0$.

Dann konvergieren die Quadraturformeln für alle Funktionen $f \in C[a,b]$.

Beweis: Wegen der Konvergenzaussage

$$\sum_{i=0}^{n} |a_i^{(n)}| = \sum_{i=0}^{n} a_i^{(n)} = Q_n(1) \to \int_a^b dx = b - a, \ \ n \to \infty$$

folgt, daß auch Voraussetzung (ii) des Satzes 13.17 erfüllt ist. Damit ergibt sich die Behauptung aus dem genannten Satz. □

Bemerkungen 13.20.

(i) Die zusammengesetzte Trapez- bzw. Simpson-Regel haben positive Gewichte. Die Sätze 13.13 und 13.14 zeigten die Konvergenz dieser Quadraturformeln für alle Polynome. Satz 13.19 ergibt dann Konvergenz für alle stetigen Funktionen, falls für die Zahl der Teilintervalle gilt $m \to \infty$.

(ii) Die Newton-Cotes Formeln konvergieren nach einem Resultat von KUSMIN nicht für alle stetigen Funktionen. □

Kapitel 14

Gaußsche Integrationsformeln

Bei Interpolationsquadraturen der Ordnung n zu paarweise verschiedenen Stützstellen $x_0, ..., x_n$ werden die Gewichte $a_0, ..., a_n$ derart bestimmt, daß Polynome vom Grad n exakt integriert werden. Für die *Gaußschen Integrationsformeln* sollen nun neben den Gewichten auch die Stützstellen gewählt werden mit dem Ziel, Polynome vom Grad $2n+1$ exakt zu integrieren. Dies führt für die $2n+2$ Unbekannten $x_0, ..., x_n \in [a, b]$ und $a_0, ..., a_n \in \mathbb{R}$ auf das nichtlineare Gleichungssystem

$$\sum_{i=0}^{n} a_i x_i^k = \int_a^b x^k dx, \quad k = 0, ..., 2n + 1. \tag{14.1}$$

14.1 Problemstellung. Orthogonale Polynome

In verschiedenen Anwendungen ist es günstig, den allgemeineren Fall von Quadraturformeln für *gewichtete* Integrale

$$\int_a^b w(x) f(x) dx \tag{14.2}$$

mit einer *Gewichtsfunktion* w zu betrachten. Dabei sei w eine auf dem Intervall (a, b) stetige und positive Funktion, für die die Integrale $\int_a^b w(x) x^k dx$ für alle $k \in \mathbb{N}_0$ existieren. Speziell soll $\int_a^b w(x) dx$ positiv sein. Typische Beispiele zeigt Tabelle 14.1.

$w(x) = 1$	$[a, b]$ kompakt	Gauß-Legendre
$w(x) = \frac{1}{\sqrt{1-x^2}}$	$[a, b] = [-1, 1]$	Gauß-Tschebyscheff 1. Art
$w(x) = \sqrt{1 - x^2}$	$[a, b] = [-1, 1]$	Gauß-Tschebyscheff 2. Art
$w(x) = e^{-x}$	$[a, b] = [0, \infty)$	Gauß-Laguerre
$w(x) = e^{-x^2}$	$[a, b] = (-\infty, \infty)$	Gauß-Hermite.

Tabelle 14.1: Beispiele von Gewichtsfunktionen

Bemerkung 14.1. Man erhält eine Interpolationsquadratur zu (14.2) wie in Kapitel 13 (dort mit $w \equiv 1$) durch Ersetzung von f durch das entsprechende Interpolationspolynom $L_n f$. Es ist

generell besser, das Integral $\int_a^b (wL_n f)(x)dx$ statt $\int_a^b L_n(wf)(x)dx$ zu betrachten. □

Definition 14.2. *Als* Gaußsche Quadraturformel der Ordnung n *wird eine Interpolationsquadratur*

$$\int_a^b w(x)f(x)dx \approx Q_n(f) := \sum_{i=0}^{n} a_i f(x_i) \tag{14.3}$$

bezeichnet, wenn sie alle Polynome $p \in \Pi_{2n+1}[a,b]$ *exakt integriert, d.h.*

$$Q_n(p) = \int_a^b w(x)p(x)dx, \quad \forall p \in \Pi_{2n+1}[a,b]. \tag{14.4}$$

Nachfolgend wollen wir die Stützstellen und Gewichte der Gaußschen Integralformeln möglichst ohne direkte Lösung des nichtlinearen Gleichungssystems (14.1) bestimmen. Die Idee besteht darin, die Stützstellen als Nullstellen gewisser *orthogonaler Polynome* zu charakterisieren.

Satz 14.3. *Sei w eine Gewichtsfunktion auf $[a,b]$ mit den o.g. Bedingungen. Dann gilt:*
(i) Es existiert ein Orthonormalsystem von Polynomen $p_n \in \Pi_n[a,b], n \in \mathbb{N}_0$ mit

$$\int_a^b w(x)p_n(x)p_m(x)dx = \delta_{nm}, \quad \forall\ n,m \in \mathbb{N}_0.$$

(ii) Die Nullstellen von p_n sind sämtlich reell, einfach und liegen in (a,b).

Beweis: zu (i): Zur Vereinfachung der Schreibweise benutzen wir das gewichtete Skalarprodukt

$$(f,g)_w := \int_a^b w(x)f(x)g(x)dx, \quad f,g \in C[a,b] \tag{14.5}$$

ein. Es ist hier ausreichend zu wissen, daß das Integral in (14.5) für beliebige Polynome f und g nach Voraussetzung an die Gewichtsfunktion w existiert.
Wir konstruieren die Polynome p_i rekursiv durch Orthonormierung der Monome $\{1, x, x^2, ...\}$ bezüglich $(\cdot,\cdot)_w$ mit Hilfe des Orthogonalisierungsverfahrens von E. Schmidt. Sei

$$p_0(x) = \frac{1}{(1,1)_w^{1/2}} = \frac{1}{(\int_a^b w(x)dx)^{1/2}},$$

d.h. $(p_0,p_0)_w = 1$. Wir nehmen jetzt an, daß bereits Polynome $p_i \in \Pi_i[a,b],\ i = 0,\dots,n-1$ definiert wurden mit

$$(p_i,p_j)_w = \delta_{ij}, \quad i,j = 0,...,n-1.$$

Dann konstruieren wir $p_n \in \Pi_n[a,b]$ aus

$$p_n(x) = \gamma_n \left\{ x^n - \sum_{i=0}^{n-1} (x^n,p_i)_w p_i(x) \right\}. \tag{14.6}$$

Man prüft unmittelbar über die Induktionsvoraussetzung nach, daß dann $(p_n,p_m)_w = 0,\ m = 0,...,n-1$. Schließlich wird die Konstante γ_n so gewählt, daß $(p_n,p_n)_w = 1$ gilt. Damit bildet $\{p_i\}$ ein System orthogonaler Polynome bezüglich des Skalarproduktes $(\cdot,\cdot)_w$.

zu (ii): Seien $x_1,...,x_m$ die Nullstellen von p_n in (a,b) mit ungerader Vielfachheit, d.h. die Nullstellen mit Vorzeichenwechsel von p_n. Wir setzen nun

$$q_m(x) = \prod_{i=1}^{m}(x-x_i) \quad \text{mit} \quad \prod_{i=1}^{0}(x-x_i) := 1.$$

Es bleibt zu zeigen, daß $m = n$. Dazu nehmen wir die Ungleichung $m < n$ an. Offenbar gilt $q_m \in \Pi_m[a,b] \subseteq \Pi_{n-1}[a,b]$. Ferner wechselt der Ausdruck $p_n q_m$ auf (a,b) nicht das Vorzeichen, d.h. es folgt

$$(p_n, q_m)_w \neq 0.$$

Andererseits ist per Konstruktion

$$\Pi_m[a,b] = \text{span}\{p_0, ..., p_m\} = \text{span}\{1, x, ..., x^m\}, \quad \forall m \in \mathbb{N}_0.$$

Daraus folgt die Darstellung

$$q_m = \sum_{i=0}^{m} \lambda_i p_i$$

und nach Konstruktion $(p_n, q_m)_w = 0$. Dadurch entsteht ein Widerspruch, d.h. die Annahme $m < n$ ist falsch. Folglich ist $m = n$. □

14.2 Existenz und Konvergenz der Gauß-Formeln

Man wählt als Stützstellen $x_0 < x_1 < \ldots < x_n$ der Gauß-Quadraturformel $Q_n(f)$ die Nullstellen des orthogonalen Polynoms $p_{n+1} \in \Pi_{n+1}[a,b]$. Im zweiten Schritt geben wir dann die Gewichte der Gaußschen Integrationsformeln unter Verwendung der Lagrange-Basisfunktionen an.

Satz 14.4. *Seien $m \in \mathbb{N}$, w eine Gewichtsfunktion auf (a,b) sowie p_{n+1} das zugehörige orthogonale Polynom nach Satz 14.3 mit den Nullstellen $x_0 < x_1 < ... < x_n \in (a,b)$. Für die Gewichte der Quadraturformel*

$$Q_n(f) = \sum_{i=0}^{n} a_i f(x_i)$$

gelte

$$a_i := \int_a^b w(x) l_i(x) dx, \quad l_i(x) := \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}. \tag{14.7}$$

Dann ist $Q_n(f)$ Gaußsche Quadraturformel, d.h.

$$Q_n(p) = \int_a^b w(x) p(x) dx, \quad \forall p \in \Pi_{2n+1}[a,b].$$

Beweis: Sei $p \in \Pi_{2n+1}[a,b]$. Division durch $p_{n+1} \in \Pi_{n+1}[a,b]$ ergibt eine Darstellung der Form

$$p(x) = q(x) p_{n+1}(x) + r(x) \quad \text{mit } q, r \in \Pi_n[a,b].$$

Wegen $p_{n+1}(x_j) = 0, j = 0, ..., n$ ist $p(x_j) = r(x_j), j = 0, ..., n$. Damit folgt

$$\begin{aligned} Q_n(p) &= \sum_{i=0}^{n} a_i p(x_i) = \sum_{i=0}^{n} a_i r(x_i) \\ &= \int_a^b w(x) \left(\sum_{i=0}^{n} r(x_i) l_i(x) \right) dx = \int_a^b w(x) r(x) dx \\ &= \int_a^b w(x) p(x) dx, \end{aligned}$$

denn per Konstruktion der orthogonalen Polynome ist $(q, p_{n+1})_w = 0$. □

Für Konvergenzaussagen benötigen wir das

Lemma 14.5. *Alle Gewichte a_i der Gaußschen Quadraturformeln sind positiv.*

Beweis: Seien $x_0, ..., x_n$ die Stützstellen von $Q_n(\cdot)$, d.h. die Nullstellen von p_{n+1}. Wir definieren nun

$$\omega(x) := \prod_{i=0}^{n}(x - x_i), \quad f_i(x) = \left[\frac{\omega(x)}{x - x_i}\right]^2, \ \ i = 0, ..., n.$$

Damit ist $f_i \in \Pi_{2n}[a, b]$, folglich nach Satz 14.4

$$0 < \int_a^b w(x) f_i(x) dx = Q_n(f_i) = \sum_{j=0}^{n} a_j f_i(x_j) = a_i f_i(x_i).$$

Wegen $f_i(x_i) > 0$ ist $a_i > 0$. □

Wir beweisen nun den folgenden Konvergenzsatz.

Satz 14.6. *Für jede Funktion $f \in C[a, b]$ konvergiert die Folge der Gaußschen Quadraturformeln für $n \to \infty$ gegen $\int_a^b w(x) f(x)\ dx$.*

Beweis: Für $p \in \Pi_m[a, b]$ ist per Konstruktion

$$Q_n(p) = \int_a^b w(x) p(x) dx, \quad \text{falls } \ 2n + 1 \geq m.$$

Der Satz 13.17 (Steklov) gilt auch für den Fall mit Gewichtsfunktion w. Unter Beachtung von Lemma 14.5 ergibt sich dann daraus die Behauptung. □

Bei hinreichend glatten Funktionen f finden wir auch eine Fehlerabschätzung.

Satz 14.7. *Sei $f \in C^{2n+2}[a, b]$. Dann gilt*

$$\int_a^b w(x) f(x) dx - Q_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n + 2)!} \int_a^b w(x) [p_{n+1}(x)]^2 dx \tag{14.8}$$

mit einer Zwischenstelle $\xi \in [a, b]$.

Beweis: Zum Beweis nutzen wir (wie bereits beim Beweis von Satz 13.10 zur Simpson-Regel) die Hermite-Interpolation. Sei $H_n f \in \Pi_{2n+1}[a, b]$ das Hermitesche Interpolationspolynom zu f, d.h.

$$(H_n f)(x_i) = f(x_i), \quad (H_n f)'(x_i) = f'(x_i), \ \ i = 0, ..., n.$$

Es hat die Darstellung

$$(H_n f)(x) = \sum_{i=0}^{n} \{f(x_i) l_i^0(x) + f'(x_i) l_i^1(x)\}$$

mit

$$l_i^0(x) := \{1 - 2 l_i'(x_i)(x - x_i)\}[l_i(x)]^2 \quad l_i^1(x) := (x - x_i)[l_i(x)]^2$$

und den Lagrange-Polynomen $l_i(x)$ (vgl. Übungsaufgaben). Für den Verfahrensfehler gilt

$$\begin{aligned}\int_a^b w(x)f(x)dx - Q_n(f) &= \int_a^b w(x)\{f(x) - (H_nf)(x)\}dx \\ &= \int_a^b w(x)[p_{n+1}(x)]^2 \frac{f(x) - (H_nf)(x)}{[p_{n+1}(x)]^2}dx.\end{aligned}$$

Der Term $w(x)[p_{n+1}(x)]^2$ ist nichtnegativ. Über die Regel von l'Hospital können wir erneut mit Hilfe der Interpolationseigenschaften von H_nf auf die Stetigkeit des im Bruch im Integranden stehenden Ausdrucks auf $[a,b]$ schließen. Aus dem erweiterten Mittelwertsatz der Integralrechnung folgern wir dann

$$\int_a^b w(x)f(x)dx - Q_n(f) = \frac{f(\eta) - (H_nf)(\eta)}{[p_{n+1}(\eta)]^2} \int_a^b w(x)[p_{n+1}(x)]^2 dx$$

mit $\eta \in [a,b]$. Für das Restglied der Hermite-Interpolation findet man analog zu Satz 9.10

$$f(x) - (H_nf)(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \left(\prod_{i=0}^{n}(x - x_i)\right)^2, \quad x \in [a,b],\ \xi \in (a,b).$$

Daraus folgt die Behauptung des Satzes. □

14.3 Legendre-Polynome

Wir untersuchen die Gaußschen Quadraturformeln zur Gewichtsfunktion $w(x) = 1, x \in [-1,1]$.

Lemma 14.8. *Die* Legendre-Polynome L_n mit

$$L_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \qquad n \in \mathbb{N}_0 \tag{14.9}$$

sind ein System orthogonaler Polynome auf $[-1,1]$ *zur Gewichtsfunktion* $w(x) = 1$.

Beweis: Man sieht aus der Definition, daß gilt $L_n \in \Pi_n[-1,1]$. Für $m < n$ rechnet man durch mehrfache partielle Integration nach, daß

$$\int_{-1}^{1} x^m \frac{d^n}{dx^n}(x^2-1)^n dx = 0.$$

Dabei wird berücksichtigt, daß die Stellen $x = \pm 1$ jeweils $n-$fache Nullstelle von $(x^2-1)^n$ sind. Daraus ergibt sich die Orthogonalitätsrelation

$$\int_{-1}^{1} L_n(x)L_m(x)dx = 0, \quad n \neq m. \qquad \square$$

Leider kann man im allgemeinen Fall die Nullstellen der Legendreschen Polynome und damit die Stützstellen und Gewichte der entsprechenden Gaußschen Quadraturformeln nicht mehr geschlossen angeben.

Wir beschränken uns hier auf die Fälle $n = 0$ und $n = 1$. Nach Definition gilt (ohne Orthonormierung)

$$L_0(x) = 1, \qquad L_1(x) = x, \qquad L_2(x) = x^2 - \frac{1}{3}.$$

Die Stützstelle der Gauß-Formel $Q_0(f)$ ist damit $x_0 = 0$. Das zugehörige Gewicht folgt aus der Exaktheitsbedingung

$$a_0 = \int_{-1}^{1} dx = 2.$$

Daraus ergibt sich die Gauß-Formel $Q_0(f) := 2f(0)$. Der Fehler ist nach Satz 14.7

$$\int_{-1}^{1} f(x)dx - Q_0(f) = \frac{1}{3}f''(\xi).$$

Die Stützstellen der Gauß-Formel $Q_1(\cdot)$ sind $x_{0,1} = \pm 1/\sqrt{3}$. Die Gewichte findet man aus den Exaktheitsbedingungen

$$a_0 + a_1 = \int_{-1}^{1} dx = 2, \qquad a_0x_0 + a_1x_1 = \int_{-1}^{1} xdx \; = \; 0$$

zu $a_0 = a_0 = 1$. Damit lautet die Gaußsche Formel für $n = 1$

$$\int_{-1}^{1} f(x)dx \approx Q_1(f) \; := \; f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}.\right)$$

Satz 14.7 ergibt

$$\int_{-1}^{1} f(x)dx - Q_1(f) = \frac{1}{135}f^{(4)}(\xi).$$

14.4 Tschebyscheff-Polynome

Wir betrachten nun den Fall der Gewichtsfunktion

$$w(x) = \frac{1}{\sqrt{1-x^2}}, \quad x \in [-1, 1].$$

Nachfolgend werden wir sehen, daß die zugehörigen orthogonalen Polynome gerade die durch

$$T_n(x) := \cos(n \arccos x), \quad x \in [-1, 1]; \qquad n \in \mathbb{N}_0 \tag{14.10}$$

definierten *Tschebyscheff-Polynome* sind. Es gilt (ohne Normierung)

$$T_0(x) = 1, \qquad T_1(x) = x.$$

Das Additionstheorem

$$\cos(n+1)t + \cos(n-1)t = 2\cos t \; \cos nt \tag{14.11}$$

ergibt die Rekursion

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x).$$

Damit folgt $T_n \in \Pi_n$ und die Struktur $T_n(x) = 2^{n-1}x^n + ...$

Lemma 14.9. *Die Tschebyscheff-Polynome genügen folgenden Aussagen:*
(i) Orthogonalität:

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}} = \begin{cases} \pi, & n = m = 0 \\ \pi/2, & n = m > 0 \\ 0, & n \neq m. \end{cases}$$

(ii) Die Nullstellen sind : $x_i = \cos\left(\frac{2i+1}{2n}\pi\right), \quad i = 0, ..., n-1.$

Beweis: (i): Die Substitution $x = \cos t$ führt auf

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}} dx = \int_0^{\pi} \cos nt \cos mt \; dt$$

und damit auf die Fallunterscheidung unter (i), speziell die Orthogonalität für $n = m$.

(ii) Aus der Definition von T_n folgt für die Nullstellen $x_i = \cos t_i,\ t_i = \frac{2i+1}{2n}\pi,\ i = 0, \ldots, n-1$. □

Als Ergebnis von Lemma 14.9 formulieren wir vorläufig die *Gauß-Tschebyscheff Quadraturformeln* der Ordnung $n-1$ für $n = 1, 2, ...$

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} \approx \; Q_{n-1}(f) \; := \; \sum_{i=0}^{n-1} a_i f\left(\cos\frac{2i+1}{2n}\pi\right).$$

Dabei wurden die Nullstellen des Polynoms $T_n(\cdot)$ als Stützstellen verwendet. Es verbleibt die Bestimmung der Gewichte a_i. Dazu zeigen wir

Lemma 14.10 *Für die Gewichte der Gauß-Tschebyscheff Quadraturformeln der Ordnung $n-1$ erhält man:* $a_i = \frac{\pi}{n}, \quad i = 0, ..., n-1$.

Beweis: Für eine Gauß-Quadraturformel der Ordnung $n-1$ müssen Polynome bis zum Grad $n-1$ exakt integriert werden, speziell auch die Polynome $T_m \in \Pi_m[-1,1],\ m = 0, \ldots, n-1$, d.h.

$$\sum_{j=0}^{n-1} a_j T_m(x_j) = \int_{-1}^{1} \frac{T_m(x)}{\sqrt{1-x^2}}\, dx, \qquad m = 0, \ldots, n-1.$$

Lemma 14.9 ergibt

$$\sum_{j=0}^{n-1} a_j \cos\frac{(2j+1)m\pi}{2n} = \begin{cases} \pi, & m = 0 \\ 0, & n \neq m. \end{cases}$$

Dies ist ein lineares Gleichungssystem für die Gewichte $a_0, \ldots, a_{n-1}$. Die explizite Lösung dieses System kann man mittels der Summenformel (10.8) aus dem Kapitel zur trigonometrischen Interpolation umgehen. Ein Vergleich ergibt, daß mit $a_j = \frac{\pi}{n}$ für $j = 0, \ldots, n-1$ eine Lösung des Gleichungssystems gefunden ist. Diese ist auch eindeutig. □

Damit lauten die *Gauß-Tschebyscheff Quadraturformeln* der Ordnung $n-1$

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} \approx \; Q_{n-1}(f) \; := \; \frac{\pi}{n} \sum_{i=0}^{n-1} f\left(\cos\frac{2i+1}{2n}\pi\right), \qquad n \in \mathbb{N}.$$

Der Nachweis der Exaktheit für Polynome bis zum Grad $2n-1$ sei als Übungsaufgabe gestellt. Aus dem Satz 14.7 haben wir schließlich noch die Fehlerdarstellung

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} - Q_{n-1}(f) = \frac{\pi f^{(2n)}(\xi)}{2^{2n-1}(2n)!}.$$

14.5 Zusammengesetzte Gauß-Formeln

Abschließend untersuchen wir *zusammengesetzte Gaußsche Quadraturformeln* (vorzugsweise mit niedriger Ordnung n). Sei auf dem Intervall $[-1,1]$ eine Gauß-Quadraturformel $n-$ter Ordnung

$$Q_n(g) := \sum_{i=0}^{n} a_i g(t_i) \;\approx\; \int_{-1}^{1} g(t)dt$$

gegeben mit dem Fehler

$$\int_{-1}^{1} g(t)dt - \sum_{i=0}^{n} a_i g(t_i) = C_n g^{(2n+2)}(\xi), \quad \xi \in (-1,1)$$

Die Transformation $x = \frac{a+b}{2} + \frac{b-a}{2}t$ und $f(x) = g(t)$ liefert eine Gauß-Formel $n-$ter Ordnung

$$Q_n(f) := \frac{b-a}{2} \sum_{i=0}^{n} a_i f\left(\frac{a+b}{2} + \frac{b-a}{2} x_i\right) \approx \int_a^b f(x)dx$$

auf einem beliebigen Intervall $[a,b]$. Es ergibt sich für den Fehler

$$\int_a^b f(x)dx - Q_n(f) = C_n \left(\frac{b-a}{2}\right)^{2n+2} f^{(2n+2)}(\zeta), \qquad \zeta \in (a,b).$$

Man unterteilt nun das Intervall $[a,b]$ in m äquidistante Teilintervalle mit der Schrittweite $h = (b-a)/m$ und wendet auf jedem Teilintervall die Gaußsche Quadraturformel der Ordnung n an. Das liefert die zusammengesetzte Gaußsche Quadraturformel $n-$ter Ordnung

$$\int_a^b f(x)dx \approx Q_n^z(f) := \frac{h}{2} \sum_{j=0}^{m-1} \sum_{i=0}^{n} a_i f\left(a + jh + \frac{h}{2} + \frac{b-a}{2} x_i\right).$$

mit einem Fehler der Ordnung $\mathcal{O}(h^{2(n+1)})$.

Beispiel 14.11. Die Tabelle 14.2 zeigt den Fehler zwischen dem Wert des Integrals $\int_0^1 \frac{dx}{1+x} = \ln 2$ und der Näherung bei Anwendung der zusammengesetzten Gaußschen Quadraturformeln mit $n = 0,1$ bei unterschiedlicher Zahl m von Teilintervallen. Bei Verdopplung der Zahl der Quadraturpunkte bzw. beim Halbierung von h reduziert sich der Fehler (entsprechend der Fehlerabschätzung) um den Faktor $\frac{1}{4}$ für $n = 0$ bzw. $\frac{1}{16}$ bei $n = 1$.

m	$n = 0$	$n = 1$
1	0.02648051	0.00083949
2	0.00743289	0.00007054
4	0.00192729	0.00000489
8	0.00048663	0.00000031
16	0.00012197	0.00000002

Tabelle 14.2: Approximation von $\int_0^1 \frac{dx}{1+x} = \ln 2$

Die Konvergenzresultate verdeutlichen die sehr gute Eignung der zusammengesetzten Gauß-Formeln für praktische Anwendungen. Ein kritischer Vergleich mit Beispiel 13.15 für die zusammengesetzten Newton-Cotes Formeln vom Grad $n = 1,2$ zeigt, daß gleichartige (sogar bessere) Ergebnisse für zusammengesetzte Gauß-Formeln der Ordnung $n = 0,1$ erzielt werden.

Wir vermerken noch, daß eine weitere Konvergenzbeschleunigung mit dem Extrapolationsverfahren von Romberg erreicht werden kann. □