C. W. Cryer

Numerische Mathematik II

Westfälische
Wilhelms-Universität
Münster

C. W. Cryer

Numerische Mathematik II

Westfälische
Wilhelms-Universität
Münster

Inhaltsverzeichnis

Kapitel 1

Approximation in normierten Räumen

1.1 Einleitung

In einem vorherigen Kapitel ist die Approximation einer stetigen Funktion $x(t)$ durch ein Polynom $(n-1)$-ten Grades $p_{n-1}(t)$ behandelt worden. Das approximierende Polynom $p_{n-1}(t)$ wurde durch Interpolation definiert. Die Menge aller Polynome $(n-1)$-ten Grades bildet einen n-dimensionalen Vektorraum X_n. Die Funktion x ist Element des Vektorraumes $X = C[a,b]$,

$$C[a,b] = \text{Menge aller stetigen Funktionen auf } [a,b]$$

Interpolation kann deshalb als eine Methode betrachtet werden, eine Approximation aus X_n von $x \in X$ zu bestimmen. In diesem Kapitel werden wir das allgemeine Approximationsproblem untersuchen.

Das Approximationsproblem

Sei X ein reeller normierter Raum. Sei $X_n \subset X$ ein endlich-dimensionaler Teilraum von X (mit Dimension n). Sei $x \in X$. Bestimme eine Bestapproximation von x aus X_n, d.h. ein Element $x^* \in X_n$ mit

$$\|x - x^*\| = d(x, X_n) := \inf_{y \in X_n} \|x - y\|.$$

Dieses Problem kann grafisch dargestellt werden (Abb. 1.1), wobei zu berücksichtigen ist, daß das Bild in $\mathbb{R}^2$ gezeichnet worden ist und deshalb evtl. einige implizite Voraussetzungen beinhaltet.

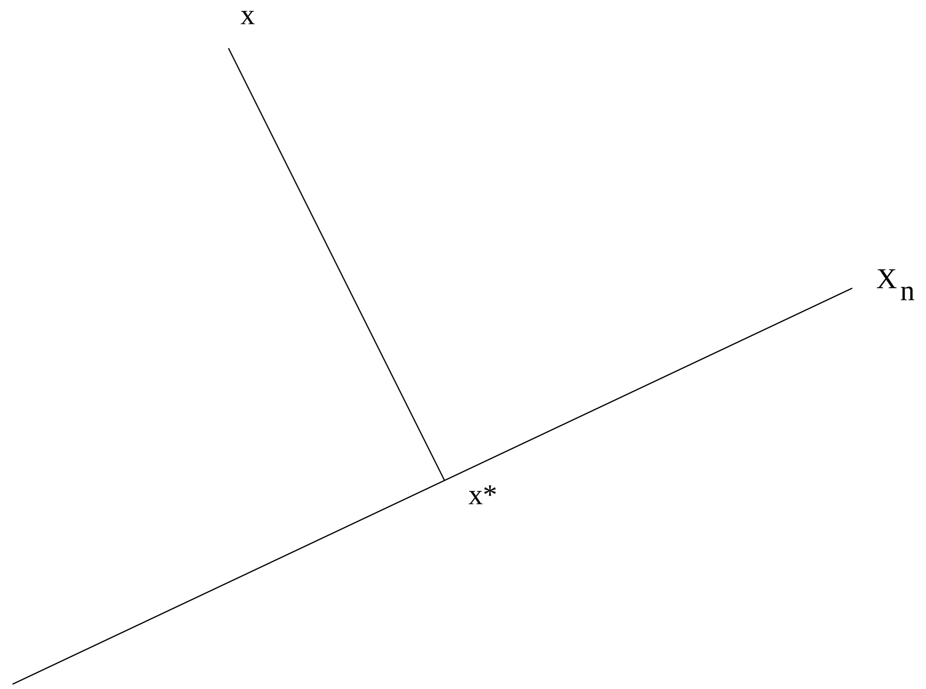


Abbildung 1.1: Eine Bestapproximation x^* von x

1.2 Banachräume

Grundlegende Kenntnisse von Banachräumen werden vorausgesetzt.

1.3 Existenz einer Bestapproximation

Satz 1.1 *Sei X ein reeller normierter Raum und X_n ein endlich-dimensionaler Unterraum von X. Sei $x \in X$. Dann existiert eine Bestapproximation $x^* \in X_n$ des Elementes x, d.h.*

1. $x^* \in X_n$
2. $\|x - x^*\| = \inf_{y \in X_n} \|x - y\|$.

Beweis: Sei $y_k \in X_n$ eine Minimalfolge, d.h.

$$\|x - y_k\| \longrightarrow d = \inf_{y \in X_n} \|x - y\| \quad \text{für } k \longrightarrow \infty.$$

O.E. gilt

$$\|y_k\| \leq \|x\| + 2d.$$

Da die Folge $\{y_k\}$ beschränkt und X_n endlich-dimensional ist, existiert nach Hilfssatz 1.1 eine konvergente Teilfolge $\{\tilde{y}_k\}, \tilde{y}_k \longrightarrow x^*$. Die Ungleichung

$$\|x - x^*\| \leq \|x - \tilde{y}_k\| + \|\tilde{y}_k - x^*\|$$

zusammen mit

$$\begin{aligned} \|x - \tilde{y}_k\| &\longrightarrow d \\ \|\tilde{y}_k - x^*\| &\longrightarrow 0 \end{aligned}$$

zeigt, daß

$$\|x - x^*\| = d \, ,$$

d.h. x^* ist eine Bestapproximation. □

Hilfssatz 1.1 (Erweiterter Satz von Bolzano-Weierstrass) *In jedem endlich-dimensionalen normierten Raum Y enthält jede beschränkte Folge $\{y_k\}$ eine konvergente Teilfolge $\{\tilde{y}_k\}, \tilde{y}_k \longrightarrow y \in Y$.*

Beweis: Ist $y_k = 0$ für unendlich viele k, setze $\tilde{y}_k = 0$, $y = 0$.
Sonst darf man o.E. annehmen, daß $y_k \neq 0$ für alle k. Sei $x_1, \dots, x_n$ eine Basis von Y. Sei

$$y_k = \sum_{i=1}^{n} \alpha_i^k x_i.$$

Setze

$$\begin{aligned} r_k &:= \sum_{i=1}^{n} |\alpha_i^k| \\ u_k &:= y_k / r_k = \sum_{i=1}^{n} \frac{\alpha_i^k}{r_k} x_i \qquad \mu_i^k := \frac{\alpha_i^k}{r_k} \end{aligned}$$

$$\tilde{\mu}_i^k \longrightarrow \mu_i \in \mathbb{R} \, , \quad 1 \leq i \leq n.$$

Sei $u := \sum_{i=1}^{n} \mu_i x_i$. Dann gilt

$$\tilde{u}_k = \sum_{i=1}^{n} \tilde{\mu}_i^k x_i \longrightarrow u \, , \ \text{ für } \ k \longrightarrow \infty.$$

Aus den Definitionen von μ_i^k und r_k sehen wir, daß

$$\sum_{i=1}^{n} |\mu_i| = \lim_{k \to \infty} \sum_{i=1}^{n} |\tilde{\mu}_i^k| = \lim_{k \to \infty} \sum_{i=1}^{n} \frac{|\tilde{\alpha}_i^k|}{\tilde{r}_k} = 1$$

und

$$\sum_{i=1}^{n} |\tilde{\mu}_i^k| = 1.$$

Folglich gilt $\tilde{u}_k \neq 0$, $u \neq 0$ (da $x_1, \dots, x_n$ eine Basis bilden). Es folgt, daß es eine positive Konstante $\delta > 0$ gibt mit

$$\begin{aligned} \|\tilde{u}_k\| &\geq \delta \, , \ \text{ für alle } \ k \, , \\ \|u\| &\geq \delta \, . \end{aligned}$$

□

Als Folgerung erhalten wir:

$$r_k = \|y_k\|/\|u_k\| \leq \left(\frac{1}{\delta}\right) \|y_k\| \leq \left(\frac{1}{\delta}\right) \sup_k \|y_k\| \ ,$$

so daß die Folge $\{r_k\}$ beschränkt ist. Folglich gibt es eine Teilfolge $\{\hat{y}_k\}$ von $\{\tilde{y}_k\}$ und destomehr von $\{y_k\}$ mit

$$\begin{aligned} \hat{r}_k &\longrightarrow r \\ \hat{u}_k &\longrightarrow u \\ \hat{y}_k &\longrightarrow y := ru \ . \end{aligned}$$

1.4 Eindeutigkeit der Bestapproximation

Die Bestapproximation ist nicht immer eindeutig — es gibt mehrere einfache Gegenbeispiele. Es gilt aber:

Satz 1.2 *In einem normierten Raum ist die Menge der Bestapproximationen konvex.*

Es gibt hinreichende Bedingungen dafür, daß die Bestapproximation eindeutig ist. Sei X ein Banachraum. Die Einheitskugel

$$B := \{x \in X : \|x\| \leq 1\}$$

hat mehrere Eigenschaften:

1. B ist abgeschlossen.

2. B ist konvex.

3. B ist absorbierend, d.h. für jedes $x \in X$ gibt es ein $\lambda \in \mathbb{R}$ mit
$$\frac{x}{\lambda} \in B.$$

4. B ist balanciert, d.h.
$$\lambda B \subset B \ \text{ für alle } \ |\lambda| \leq 1 \ .$$

5. B ist beschränkt, d.h., für eine Umgebung U von $0 \in X$ gibt es ein $\lambda_0 \in \mathbb{R}$ mit $\lambda B \subset U$ für alle $|\lambda| < \lambda_0$.

Satz 1.3 (Kolmogorow) *Ein topologischer Vektorraum X ist genau dann normierbar, wenn:*

1. *X ist ein Hausdorffraum (d.h. Elemente sind trennbar).*

2. *Es gibt eine beschränkte konvexe Umgebung von 0.*

Beweis: Siehe z.B. Kantorovich und Akilov, Functional Analysis in Normed Spaces, S. 403. □

Obwohl die Einheitskugel in einem Banachraum konvex ist, gibt es doch qualitative Unterschiede (siehe Abb. 1.2).

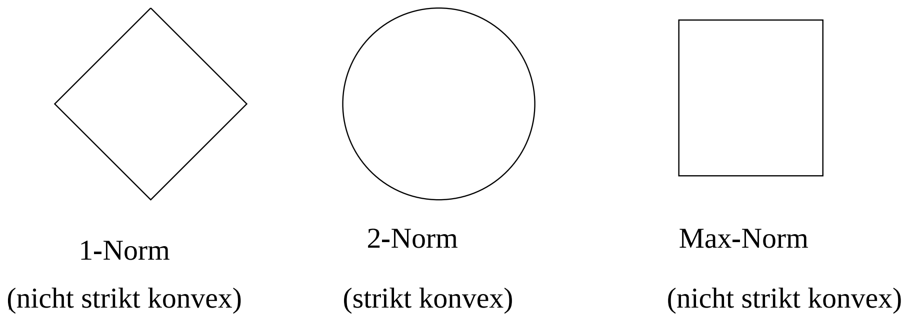


Abbildung 1.2: Einheitskugeln bzgl. dreier Normen in $\mathbb{R}^2$

Definition 1.1 *Eine Norm heißt* strikt konvex, *falls aus* $||x|| = ||y|| = 1, x \neq y$, *stets*

$$\left\| \frac{x+y}{2} \right\| < 1$$

folgt.

Satz 1.4 *In einem strikt konvexen normierten Raum besitzt das Approximationsproblem höchstens eine Lösung.*

In den nächsten Absätzen werden drei Spezialfälle betrachtet: $X = L_2, X = L_\infty$ und $X = L_1$.

1.5 Approximieren in L_2

In diesem Abschnitt betrachten wir das Approximationsproblem in Räumen wie:

$$\begin{aligned}
\ell_2 &= \left\{ x = \{x_i : i \in \mathbb{N}\} : \|x\| = \|x\|_2 = \left[\sum_{i=1}^{\infty} x_i^2 \right]^{1/2} \right\} \\
L_2(a,b) &= \left\{ x(t), a \le t \le b : \|x\| = \|x\|_2 = \left[\int_a^b |x(t)|^2 dt \right]^{1/2} \right\} \\
\mathbb{R}^n &= \left\{ x = (x_1, \ldots, x_n) : \|x\| = \|x\|_2 = \left[\sum_{i=1}^{n} x_i^2 \right]^{1/2} \right\}
\end{aligned}$$

Diese Räume sind reelle Hilberträume mit dem Skalarprodukt (oder Innenprodukt)

$$\begin{aligned}
(x,y) &:= \sum_{i=1}^{\infty} x_i y_i & (\ell_2) \\
(x,y) &:= \int_a^b x(t)y(t)dt & (L_2) \\
(x,y) &:= \sum_{i=1}^{n} x_i y_i & (\mathbb{R}^n)
\end{aligned}$$

In allen Fällen gilt die Parallelogrammgleichung:

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Das Skalarprodukt (x, y) erfüllt die Bedingungen:

1. $(x+y, z) = (x, z) + (y, z)$
2. $(\alpha x, y) = \alpha(x, y),\ \alpha \in \mathbb{R}$
3. $(x, y) = (y, x)$
4. $(x, x) \ge 0$ und $(x, x) = 0$ genau dann, wenn $x = 0$.

Satz 1.5 *Ein reeller normierter Raum X mit Norm $\|\cdot\|$ ist auch ein reeller Innenproduktraum mit einem Skalarprodukt $(\cdot,\cdot)$, wofür $\|x\|^2 = (x, x)$, genau dann, wenn die Parallelogrammgleichung in X gilt.*

Beweis: Siehe Dunford - Schwartz, Linear Operators, S. 393. □

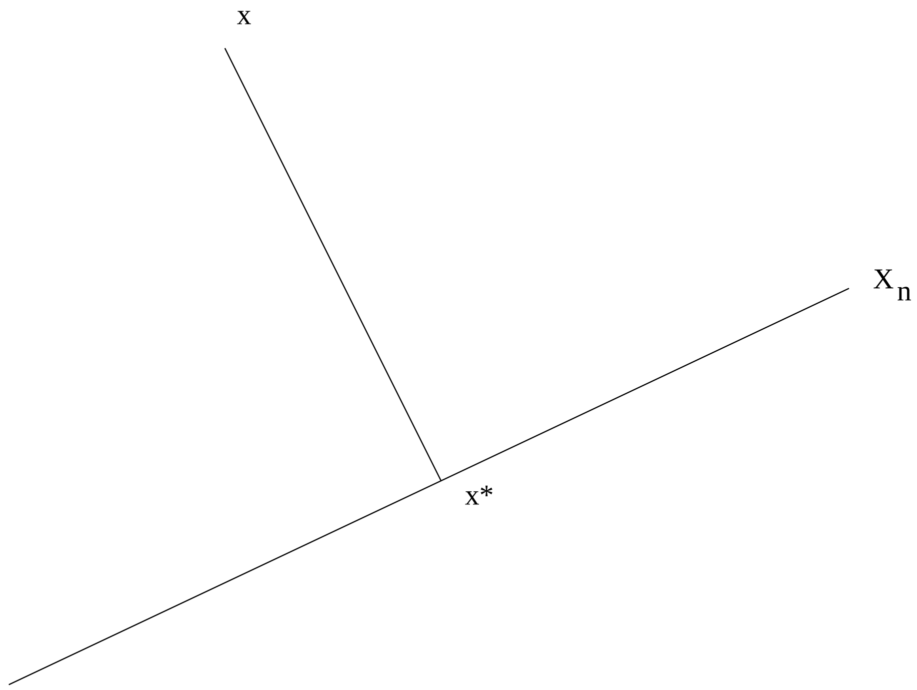


Abbildung 1.3: Approximation in einem Innenproduktraum

Ist X ein Innenproduktraum, dann ist $\|\cdot\|$ strikt konvex (folgt sofort aus der Parallelogrammgleichung), so daß das Approximationsproblem eine eindeutige Lösung besitzt. In diesem Fall stimmt das einfache Bild (siehe Abb. 1.3), d.h., der Fehler $x-x^*$ ist orthogonal zu allen Elementen von X_n.

Satz 1.6 *Sei X ein Innenproduktraum. Sei $x_1, \ldots, x_n$ eine Basis von X_n. Es gilt*

$$x^* = \sum_{i=1}^{n} \alpha_i^* x_i \ ,$$

wobei die Koeffizienten α_i^* eindeutig durch die Normalgleichungen

$$(x^* - x, x_k) = \sum_{i=1}^{n} \alpha_i^* (x_i, x_k) - (x, x_k) = 0$$

für $1 \leq k \leq n$ bestimmt sind.

Beweis: Sei $x \in X$ und $x^* \in X_n$ die Bestapproximation. Sei x_k ein Basiselement. Ist $(x - x^*, x_k) \neq 0$, setze

$$v := x^* + \epsilon x_k \in X_n \ ,$$

wo $\epsilon \in \mathbb{R}$ noch zu bestimmen ist, so daß

$$\begin{aligned} \|x - v\|^2 &= (x - x^* - \epsilon x_k, x - x^* - \epsilon x_k) \\ &= \|x - x^*\|^2 + \epsilon^2 \|x_k\|^2 - 2\epsilon (x - x^*, x_k) \ . \end{aligned}$$

Mit

$$\epsilon := \frac{(x - x^*, x_k)}{\|x_k\|^2} \neq 0$$

folgt

$$\|x - v\|^2 = \|x - x^*\|^2 - \frac{[(x - x^*, x_k)]^2}{\|x_k\|^2} < \|x - x^*\|^2 ,$$

ein Widerspruch. □

Die Koeffizienten α_i^* der Bestapproximation

$$x^* = \sum_{i=1}^{n} \alpha_i^* x_i$$

erfüllen also die Normalgleichungen

$$A\alpha^* = b \tag{1.1}$$

mit

$$\begin{aligned}
A &= (a_{ik}) \in \mathtt{Mat}(n \times n) \\
a_{ik} &= (x_i, x_k), \quad 1 \leq i, k \leq n \\
b &= (b_i) \in \mathbb{R}^n \\
b_i &= (x, x_i), \quad 1 \leq i \leq n \\
\alpha^* &= (\alpha_1^*, \ldots, \alpha_n^*)^T \in \mathbb{R}^n .
\end{aligned}$$

Da die Bestapproximation x^* existiert und eindeutig ist, ist die Gramsche Matrix A regulär und die Gramsche Determinante

$$g(x_1, \ldots, x_n) := \det((x_i, x_k)) \neq 0.$$

Beispiel: Sei $X = L_2(0, 1)$. Bestimme die Bestapproximation der Form

$$x^*(t) = \sum_{i=1}^{n} \alpha_i^* t^{i-1}$$

von einer stetigen Funktion $x(t)$.
Die Koeffizienten α_i^* erfüllen die Normalgleichungen (1.1) mit

$$\begin{aligned}
b &= (b_1, \ldots, b_n)^T \\
b_i &= \int_0^1 x(t) t^{i-1} dt \\
A &= (a_{ik}) \\
a_{ik} &= \int_0^1 t^{i-1} t^{k-1} dt = \frac{1}{i + k - 1} ,
\end{aligned}$$

d.h. die Gramsche Matrix A ist die $n \times n$ Hilbertmatrix.

1.6 Approximation in $C[a,b]$

Sei Ω eine abgeschlossene beschränkte Teilmenge des $\mathbb{R}^n$. $C[\Omega]$ bezeichnet den Raum aller Funktionen $x(t)$, die auf Ω definiert und stetig sind, mit der Maximumnorm:

$$\|x\| = \|x\|_\infty = \max_{t\in\Omega} |x(t)|.$$

Wie leicht festzustellen ist, ist diese Norm nicht strikt konvex, so daß die Eindeutigkeit der Lösung des Approximationsproblems nicht aus der allgemeinen Theorie folgt.

Beispiel:

$$\begin{aligned} \Omega &:= [0,1]\times[0,1] \subset \mathbb{R}^2 \\ x(\mathbf{t}) &:= t_1 t_2\ ,\ \mathbf{t} = (t_1,t_2) \in \Omega \\ X_5 &:= \text{span}\{1, t_1, t_2, t_1^2, t_2^2\} \end{aligned}$$

Es kann gezeigt werden, daß

$$d(x, X_5) = \inf_{y\in X_5} \|x-y\|_\infty = \frac{1}{4}\ .$$

Sei

$$\begin{aligned} x_1^* &:= \frac{1}{2}\left(t_1^2+t_2^2\right) - \frac{1}{4} \\ x_2^* &:= t_1 + t_2 - \frac{1}{2}\left(t_1^2+t_2^2\right) - \frac{1}{4}\ . \end{aligned}$$

Sei nun $\|x^* - x\| = \frac{1}{4}$, $x^* \in X_5$. Dann existiert $\lambda \in [0,1]$, so daß

$$x^* = \lambda x_1^* + (1-\lambda)x_2^*.$$

(Siehe Buck, R.C.: Linear Spaces and Approximation Theory. In: On Numerical Approximation, R.E. Langer (ed.), University of Wisconsin Press, Madison 1959.)

Für den Raum $X = C[a,b]$ und spezielle Unterräume $X_n \subset X$ gibt es allerdings eine schöne Theorie zum Approximationsproblem, die wir jetzt kurz beschreiben. (Siehe z.B. Meinardus, G., Approximation von Funktionen und ihre numerische Behandlung, Springer 1964.)

Im folgenden sei $-\infty < a < b < \infty$,

$$\begin{aligned} X &= C[a,b] \\ \|\cdot\| &= \|\cdot\|_\infty \\ X_n &= \text{span}\{x_1,\ldots,x_n\} \subset C[a,b] \\ x &\in X\ . \end{aligned}$$

Sei $x \in X$. Ein Punkt $t \in [a,b]$ heißt *Extremalpunkt* von x, wenn

$$|x(t)| = \|x\|.$$

Satz 1.7 *Es sei $h_0 \in X_n$ und D die Menge der Extremalpunkte der Funktion $x(t) - h_0(t)$. Gibt es ein $h \in X_n$, so daß*

$$(x(t) - h_0(t))h(t) > 0 \quad \textit{für} \;\; t \in D\,,$$

so gibt es auch ein $h_1 \in X_n$ mit

$$\|x - h_1\| < \|x - h_0\|.$$

Beweis: Die Menge D ist kompakt, da x und h_0 stetige Funktionen sind. Es sei

$$\min_{t \in D} |x(t) - h_0(t)| \cdot |h(t)| = \alpha > 0$$

und

$$\|h\| = \beta > 0.$$

Wähle eine offene Teilmenge U von $[a,b]$, so daß $D \subset U$ und

$$(x(t) - h_0(t)) \cdot h(t) > \frac{\alpha}{2} \quad \text{für} \;\; t \in U.$$

Dann ist

$$\gamma := \|x - h_0\| - \max_{t \in [a,b] \setminus U} |x - h_0| > 0.$$

Sei

$$\begin{aligned} \delta &:= \min\left(\frac{\gamma}{2\beta}, \frac{\alpha}{2\beta^2}\right) \\ h_1(x) &:= h_0(x) + \delta h(x)\,. \end{aligned}$$

Dann gilt für $t \in U$:

$$\begin{aligned} |x - h_1|^2 &= |x - h_0|^2 - 2\delta h(x - h_0) + \delta^2 |h|^2 \\ &\leq |x - h_0|^2 - 2\frac{\delta\alpha}{2} + \delta^2\beta^2 \\ &\leq |x - h_0|^2 - \frac{\alpha\delta}{2} \\ &\leq \|x - h_0\|^2 - \frac{\alpha\delta}{2} \end{aligned}$$

Für $t \in [a,b] \setminus U$ ergibt sich:

$$\begin{aligned} |x - h_1| &\leq |x - h_0| + \delta|h| \\ &\leq \|x - h_0\| - \gamma + \delta\beta \\ &\leq \|x - h_0\| - \gamma/2. \end{aligned}$$

□

Definition 1.2 *Ein Teilraum X_n von $C[a,b]$ (der Dimension n) erfüllt die* Haarsche Bedingung *(ist* unisolvent*), wenn jede Funktion aus X_n, die nicht identisch gleich Null ist, an höchstens $n-1$ Punkten aus $[a,b]$ verschwindet.*

Bemerkung: Sei $\{x_1, \ldots, x_n\}$ eine Basis von X_n. Die folgenden Aussagen sind äquivalent:

a) X_n ist unisolvent

b) Für jede Unterteilung von $[a,b]$, $a \leq t_1 < \cdots < t_n \leq b$ gilt $\det(x_i(t_j)) \neq 0$.

c) Für jede Unterteilung $a \leq t_1 < \cdots < t_n \leq b$ und jede Wahl von y_i, $1 \leq i \leq n$, existiert genau ein $u \in X_n$ mit

$$u(t_i) = y_i \ , \ 1 \leq i \leq n.$$

Beispiele von unisolventen Teilräumen:

a) $X_n = \mathcal{P}_{n-1}$

b) $x_i(t) = e^{\lambda(i-1)t} \ , \ 1 \leq i \leq n$

c) Sei

$$\begin{aligned} [a,b] &:= [-1,+1] \ , \ n := 2 \\ x_1(t) &:= 1 \ , \ x_2(t) := t^2 \end{aligned}$$

Dann ist X_2 nicht unisolvent.

Satz 1.8 *Der lineare Teilraum X_n von $C[a,b]$ sei unisolvent. Dann gibt es zu jedem $x \in C[a,b]$ genau eine Bestapproximation $x^* \in X_n$.*

Beweis: Man darf o.E. annehmen, daß $x \notin X_n$ (sonst wäre der Satz trivial).
Sei h eine Bestapproximation von x. Dann besteht die Menge D der Extremalpunkte von $x(t)-h(t)$ aus mindestens $n+1$ Punkten aus $[a,b]$. Denn gäbe es weniger als $n+1$ solcher Punkte, so existierte wegen der Haarschen Bedingung ein h_1 mit

$$h_1(t_i) = x(t_i) - h(t_i)$$

für alle diese Extremalpunkte x_i. Es würde also

$$(x(t_i) - h(t_i))h_1(t_i) > 0$$

für alle Extremalpunkte gelten. Nach Satz 1.7 wäre dann $h(t)$ keine Bestapproximation.

Mit zwei Bestapproximationen h_1 und h_2 ist auch (wegen der Konvexität der Menge aller Bestapproximationen)

$$h := \frac{1}{2}h_1 + \frac{1}{2}h_2$$

eine Bestapproximation. Aus den Extremalpunkten von $x - h$ greift man $n+1$ Punkte t_i , $a \leq t_1 < t_2 \cdots < t_{n+1} \leq b$, heraus. Man erhält nun aus

$$x(t_i) - h(t_i) = \pm\|x - h\|$$

wegen

$$|x(t_i) - h_j(t_i)| \leq \|x - h\| \ , \ j = 1, 2$$

sofort

$$h_1(t_i) = h_2(t_i) \ , \ 1 \leq i \leq n+1.$$

Dann ist wegen der Haarschen Bedingung $h_1 = h_2$. □

Satz 1.9 *Sei X_n unisolvent, $x \in X$, $u \in X_n$ und $a \leq t_1 < \cdots < t_{n+1} \leq b$. $(x-u)(t_j)$ habe alternierende Vorzeichen, d.h.*

$$\exists \sigma \ , \ |\sigma| = 1 : \mathit{sgn}[(x-u)(t_j)] = (-1)^j \sigma \ , \ 1 \leq j \leq n+1.$$

Dann gilt:

$$\min_{1 \leq j \leq n+1} |(x-u)(t_j)| \leq \|x - x^*\| \leq \|x - u\|.$$

Beweis: Durch Widerspruch. Sei

$$\min_{1 \leq j \leq n+1} |(x-u)(t_j)| > \|x - x^*\|.$$

Dann folgt:

$$|(x-u)(t_j)| > |(x-x^*)(t_j)| \ , \ 1 \leq j \leq n+1$$

und so

$$\begin{aligned} \text{sgn}[u(t_j) - x^*(t_j)] &= (-1)^j \sigma \ , \ 1 \leq j \leq n+1 \\ u(t_j) - x^*(t_j) &\neq 0 \ , \ 1 \leq j \leq n+1 \ . \end{aligned}$$

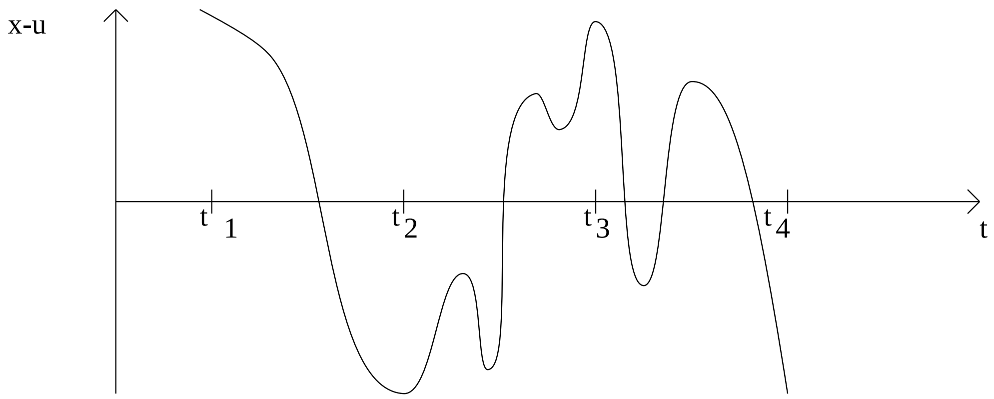
Die Funktion $u - x^* \in X_n$ hat also mindestens n Vorzeichenwechsel, d.h. mindestens n Nullstellen. Da X_n unisolvent ist, folgt $u = x^*$ und damit der gewünschte Widerspruch. □

Bemerkung: Satz 1.9 ermöglicht es, eine untere und obere Grenze für $\|x - x^*\|$ zu bestimmen, wobei allerdings nur die Berechnung der unteren Grenze einfach ist.

Definition 1.3 *$a \le t_1 < \cdots < t_m \le b$ heißt* Alternante *(der Länge m) zu $x-u$, falls $x-u$ an den Stellen t_j mit alternierenden Vorzeichen seinen Maximalwert annimmt, d.h.*

$$\exists \sigma\ ,\ |\sigma| = 1 : (x-u)(t_j) = \sigma(-1)^j \|x-u\|.$$

Beispiel: Eine Alternante der Länge 4.



Satz 1.10 (Tschebyscheff (1899), de la Vallee-Poussin (1919)) *Sei X_n ein n-dimensionaler Teilraum von $C[a,b]$. X_n erfülle die Haarsche Bedingung. Sei $x \in C[a,b]$. Dann ist $u \in X_n$ genau dann eine Bestapproximation von x, wenn es eine Alternante der Länge $n+1$ zu $x-u$ gibt.*

Beweis: „$\Longleftarrow$": Es gebe eine Verteilung $a \le t_1 < \cdots < t_{n+1} \le b$ mit

$$(x-u)(t_j) = \sigma(-1)^j \|x-u\|.$$

Nach Satz 1.9 gilt:

$$\begin{aligned} \|x-u\| &= \min_{1\le j\le n+1} |(x-u)(t_j)| \\ &\le \|x-x^*\| \\ &\le \|x-u\|\ . \end{aligned}$$

„$\Longrightarrow$": (Skizze) Sei x^* die Bestapproximation von x. Wenn es keine Alternante der Länge $n+1$ zu $x-x^*$ gibt, dann existiert $h \in X_n$, so daß

$$(x-x^*)h > 0$$

für alle Extremalpunkte von $x-x^*$. Jetzt benutzt man Satz 1.7 und erhält einen Widerspruch. $\square$

Beispiel: Bestimme die Bestapproximation von $x(t) := t^2\ ,\ x \in C[0,+1]$, aus $X_2 = \mathrm{span}\{1,t\}$.

Lösung: Ist x^* die Bestapproximation, dann existiert eine Alternante der Länge 3, $\{t_1, t_2, t_3\}$, zu $x - x^*$. Da $x - x^*$ stetig differenzierbar ist, ist die Ableitung von $x - x^*$ gleich Null in jedem ⟨⟨ inneren ⟩⟩ Punkt t_j, $0 < t_j < 1$. Sei $x^*(t) = a + bt$, dann folgt:

$$\frac{d}{dt}(x - x^*)(t) = 2t - b\ ,$$

so daß

$$t_1 = 0\ ,\ \ t_2 = +\frac{b}{2}\ ,\ \ t_3 = 1.$$

Weiter gilt:

$$x(t_j) - x^*(t_j) = \sigma d(-1)^j$$

mit

$$d := \|x - x^*\|\ ,\ \ |\sigma| = 1\ ,$$

d.h.

$$x(0) - x^*(0) = 0 - \alpha_1 = -\sigma d \tag{1.2}$$

$$x\left(\frac{b}{2}\right) - x^*\left(\frac{b}{2}\right) = \frac{\alpha_2^2}{4} - \alpha_1 - \frac{\alpha_2^2}{2} = +\sigma d \tag{1.3}$$

$$x(1) - x^*(1) = 1 - \alpha_1 - \alpha_2 = -\sigma d \tag{1.4}$$

(1.2) $\Longrightarrow \sigma d = \alpha_1$
(1.2), (1.4) $\Longrightarrow \alpha_2 = 1$
(1.3) $\Longrightarrow \frac{1}{4} - \alpha_1 - \frac{1}{2} = \alpha_1$,
d.h.

$$\alpha_1 = -\frac{1}{8}.$$

Zusammenfassend: Die Bestapproximation ist

$$x^*(t) = -\frac{1}{8} + t.$$

Siehe Abbildung 1.4.

1.6.1 Iterationsmethoden nach Remez

Sei $x \in C[a, b]$ gegeben. Sei $X_n = \text{span}\{x_1, \ldots, x_n\}$ unisolvent. Die Methoden von Remez berechnen eine Approximation von x mit Hilfe einer Folge $M_0, M_1, \ldots$ von Punktmengen

$$M_k := \left\{t_1^{(k)}, \ldots, t_{n+1}^{(k)}\right\}$$

mit

$$a \le t_1^{(k)} < \cdots < t_{n+1}^{(k)} \le b.$$

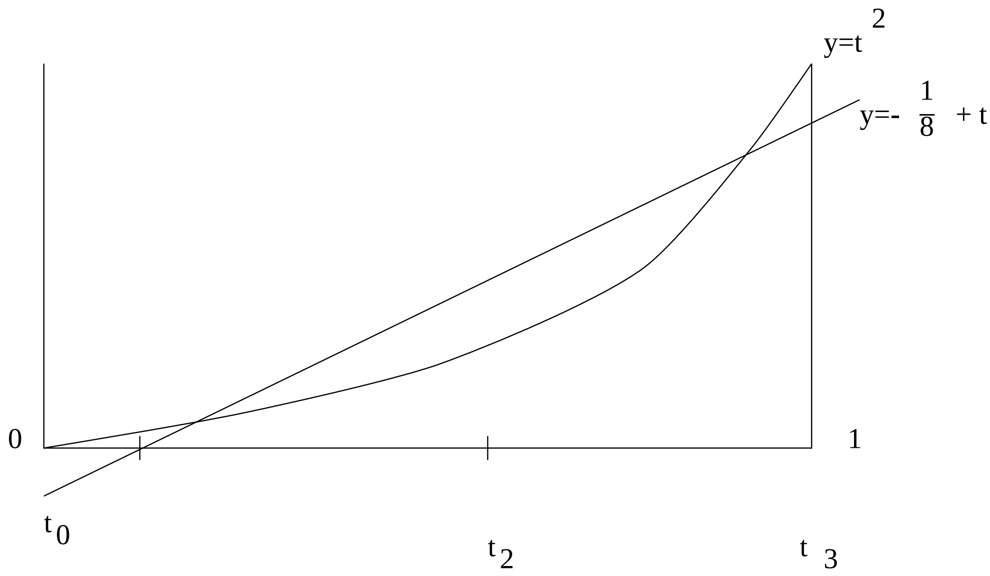


Abbildung 1.4: Die Bestapproximation von t^2 durch lineare Polynome

Sei M_k gegeben. Hierzu konstruiert man $\lambda_k \in \mathbb{R}$ und $h_k \in X_n$, die die Bedingungen

$$h_k\left(t_i^{(k)}\right) + \lambda_k(-1)^i = x\left(t_i^{(k)}\right) \;, \quad 1 \le i \le n+1$$

erfüllen, d.h.

$$A\begin{pmatrix}\alpha^{(k)} \\ \hline \lambda_k\end{pmatrix} = \mathbf{c}^{(k)}$$

mit

$$\begin{aligned}
\alpha^{(k)} &= \left(\alpha_1^{(k)}, \dots, \alpha_n^{(k)}\right)^T \\
\mathbf{c}^k &= \left(x\left(t_1^{(k)}\right), \dots, x\left(t_n^{(k)}\right)\right)^T \\
h_k &= \sum_{i=1}^{n} \alpha_i^{(k)} x_i
\end{aligned}$$

$$A = \begin{pmatrix}
x_1\left(t_1^{(k)}\right) & \cdots & x_n\left(t_1^{(k)}\right) & -1 \\
x_1\left(t_2^{k}\right) & \cdots & x_n\left(t_2^{(k)}\right) & +1 \\
\vdots & & \vdots & \vdots \\
x_1\left(t_{n+1}^{(k)}\right) & \cdots & x_n\left(t_{n+1}^{(k)}\right) & (-1)^n
\end{pmatrix}$$

Daß A regulär ist, folgt aus der Haarschen Bedingung. Folglich existieren α^k, λ_k (und h_k), und die sind auch eindeutig.

Nun besteht eine Alternative. Entweder ist

$$\|h_k - x\| = \lambda_k \;,$$

dann ist (siehe Satz 1.9) h_k die Bestapproximation von x, oder es ist

$$\|h_k - x\| > \lambda_k.$$

Dann gibt es einen Punkt $\xi_k \in [a,b]$ mit $|h_k(\xi_k) - x(\xi_k)| = \|h_k - x\| > \lambda_k$. In diesem Fall wird M_{k+1} gebildet, indem ξ_k zu M_k hingenommen und ein Punkt $t_\mu^{(k)}$ gestrichen wird. Der Punkt $t_\mu^{(k)}$ wird so gewählt, daß die Zahlen $(h_k - x)\left(t_i^{(k+1)}\right)$ alternieren, d.h.

$$\operatorname{sgn}\left[(\varphi_k)\left(t_i^{(k+1)}\right)\right] = \xi \operatorname{sgn}\left[(\varphi_k)\left(t_i^{(k)}\right)\right]$$

mit $\varphi_k := h_k - x$ und $\xi \in \{1, -1\}$. (Siehe Meinardus, S. 100.)

Voraussetzungen			**Streiche**
$a \le \xi_k < t_1^{(k)} \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$\operatorname{sgn} \varphi_k(t_1^{(k)})$	$t_1^{(k)}$
$a \le \xi_k < t_0^{(k)} \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$-\operatorname{sgn} \varphi_k(t_1^{(k)})$	$t_{n+1}^{(k)}$
$t_\nu^{(k)} < \xi_k < t_{\nu+1}^{(k)} \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$\operatorname{sgn}(\varphi_k(t_\nu^{(k)}))$	$t_\nu^{(k)}$
$t_\nu^{(k)} < \xi_k < t_{\nu+1}^{(k)} \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$-\operatorname{sgn}(\varphi_k(t_\nu^{(k)}))$	$t_{\nu+1}^{(k)}$
$t_{n+1}^{(k)} < \xi_k \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$\operatorname{sgn} \varphi_k(t_{n+1}^{(k)})$	$t_{n+1}^{(k)}$
$t_{n+1}^{(k)} < \xi_k \operatorname{sgn} \varphi_k(\xi_k)$	$=$	$-\operatorname{sgn} \varphi_k(t_{n+1}^{(k)})$	$t_1^{(k)}$

Der Konvergenzbeweis dieser *ersten* Methode von Remez, die Beschreibung und der Konvergenzbeweis der *zweiten* Methode von Remez werden in Meinardus 1964, S. 102, Rice 1964, S. 178 und Ralston 1965, S. 301, gegeben.

1.6.2 Approximation durch rationale Funktionen

Überraschenderweise läßt sich die Theorie auf der Approximation von $x \in C[a,b]$ durch rationale Funktionen

$$R(t) = \frac{P_n(t)}{Q_m(t)} = \frac{\sum_{i=0}^{n} \alpha_i t^i}{\sum_{i=0}^{m} \beta_i t^i}$$

übertragen. Sei $V_{n,m}$ die Menge aller Quotienten $P_n(t)/Q_m(t)$.

Satz 1.11 *Sei $x \in C[a,b]$, $n, m \in \mathbb{N}$. Es gibt genau eine Bestapproximation $x^* \in V_{n,m}$ zu x.*

Beweis: Siehe z.B. Meinardus, S. 150. □

Satz 1.12 *Die Funktion*

$$R_{n,m}(t) = \frac{P_n(t)}{Q_m(t)} \in V_{n,m}$$

ist genau dann die Bestapproximation zu $x \in C[a,b]$, wenn eine Alternante zu $x - R_{n,m}$ der Länge

$$\ell(n,m,P_n,Q_m) + 1$$

existiert. Hier ist

$$\ell(n,m,P_n,Q_m) = \begin{cases} n+1, & \textit{falls } P_n \equiv 0 \\ m+n-\delta+1, & \textit{sonst} \end{cases}$$

$$\delta := \min(m - \textit{Grad } P_m, n - \textit{Grad } P_n).$$

Beweis: Siehe z.B. Meinardus, S. 154, Davis, S. 153. □

1.7 Approximation in $L_1[a,b]$

Dieses Approximationsproblem wird nur selten behandelt und wir beschränken uns darauf, einen Satz zu zitieren und einige Literaturhinweise zu geben.

Satz 1.13 *Sei $X = L_1[a,b]$, $x_i \in C[a,b]$, $1 \le i \le n$, $X_n = \textit{span}\{x_1, \ldots, x_n\}$. Sei X_n unisolvent. Sei $x \in X$ eine stetige Funktion.*
Dann existiert genau eine Bestapproximation $x^* \in X_n$ von x.

Beweis: Siehe Watson [1980, S. 135] oder Rice [1964, S. 109]. □

Literatur

Cody, W.J., **Waite**, W.: Software Manual for the Elementary Functions. Prentice Hall, 1980.

Davis, P.J.: Interpolation and Approximation. Blaisdell, 1963.

Fike, C.: Computer Evaluation of Mathematical Functions. Prentice Hall, 1968.

Hastings, C.B.: Approximations for Digital Computers. Princeton University Press, 1955.

Meinardus, G.: Approximation von Funktionen und ihre numerische Behandlung. Springer, 1964.

Ralston, A.: A First Course in Numerical Analysis. McGraw Hill, 1965.

Rice, J.R.: The Approximation of Functions. Vol. I - Linear Theory. Addison-Wesley, 1964.

Watson, G.A.: Approximation Theory and Numerical Methods. Wiley, 1980.

Kapitel 2

Eigenwertprobleme

2.1 Einleitung

Definition 2.1 *Sei* $A \in Mat(n, n, \mathcal{C})$*.* $\lambda \in \mathcal{C}$ *heißt* Eigenwert *von* A *, wenn es* $x \in \mathcal{C}^n$ *gibt mit* $x \neq 0$, $Ax = \lambda x$ *.* x *heißt dann Eigenvektor zu* λ*, und* (λ, x) *heißt* Eigenpaar von A*.*

In diesem Kapitel betrachten wir mehrere numerische Methoden zur numerischen Bestimmung von Eigenwerten und Eigenvektoren.

Es ist oft nützlich, zwischen *algebraischen* und *geometrischen* Eigenwerten zu unterscheiden. Die Nullstellen des charakteristischen Polynoms

$$p(\lambda) := \det(A - \lambda I) = 0$$

sind Eigenwerte; die Vielfachheit der Nullstelle λ_i des Polynoms p heißt ihre *algebraische* Vielfachheit $\sigma(\lambda_i)$. Die Matrix A besitzt dann n (algebraische) Eigenwerte.

Die Eigenvektoren zum Eigenwert λ_i bilden einen linearen Teilraum $L(\lambda_i)$ des $\mathbb{R}^n$. Die Dimension von $L(\lambda_i)$, also die Anzahl linear unabhängiger Eigenvektoren zum Eigenwert λ_i, heißt die *geometrische* Vielfachheit $\rho(\lambda_i)$ von λ_i. Es gilt

$$\begin{aligned} 1 \leq \rho(\lambda_i) &\leq \sigma(\lambda_i) \\ \sum_i \rho(\lambda_i) \leq \sum_i \sigma(\lambda_i) &= n \, . \end{aligned}$$

Literatur

Gourlay, A.R., **Watson**, G.A.: Computational Methods for Matrix Eigenproblems. Chichester, John Wiley, 1973.

Parlett, P.N.: The Symmetric Eigenvalue Problem. Prentice-Hall, 1980.

Saad, Y.: Numerical Methods for Large Eigenvalue Problems. Manchester Univ. Press, 1992.

Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford, Clarendon Press, 1965.

2.2 Anwendungen

Die folgenden Anwendungen treten u.a. auf:

1. Sei $b, u^{(0)} \in \mathcal{C}^n$ gegeben und $u^{(1)}, u^{(2)}, \ldots, u^{(k)}, \ldots$ durch
$$u^{(k+1)} = Au^{(k)} + b \ , \ k \geq 0 \tag{2.1}$$
definiert. Dann gilt $u^{(k)} \longrightarrow b$ als $k \to \infty$ nur, wenn
$$\rho(A) = \max\{|\lambda| : \lambda \quad \text{ein Eigenwert von } A\} < 1 \ .$$
Systeme wie in (2.1) kommen in der Praxis häufig vor, und A ist oft sehr groß ($n \geq 10^6$), aber schwach besetzt.

2. Sei $u(t)$ eine Lösung des Anfangswertproblems
$$\begin{aligned} \dot{u}(t) &= \frac{du(t)}{dt} = Au(t) + b \ , \ t > 0 \\ u(0) &= u_0 \ . \end{aligned}$$
Dann gilt:
$$u(t) = -A^{-1}b + e^{At}(u(0) + A^{-1}b) \ ,$$
und
$$u(t) \longrightarrow -A^{-1}b$$
nur, wenn alle Eigenwerte von A negativen Realteil haben.

3. Schwingungen (Bücher von Collatz, Timoshenko, Frazer, Duncan und Collar), Volkswirtschaft (Samuelson). Flugzeugflügel, Kraftwerke.

Wie die oben erwähnten Beispiele zeigen, entstehen Eigenwertaufgaben oft in Zusammenhang mit Stabilitätsproblemen, und es ist dann nicht nötig, alle Eigenwerte zu berechnen. Es ist oft nur notwendig, die Resolventenmenge $\rho(A)$ zu berechnen. Deshalb ist es sinnvoll, Algorithmen für die folgenden Probleme zu betrachten:

1. Bestimmung des größten oder kleinsten Eigenwertes

2. Bestimmung aller Eigenwerte

3. Bestimmung mehrerer Eigenvektoren.

2.3 Die Gutartigkeit von Eigenwertproblemen

Im allgemeinen kann das Eigenwertproblem schlecht konditioniert sein. Kleine Störungen an der Matrix A können große Änderungen bei den Eigenwerten und Eigenvektoren verursachen.

Beispiel:

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

   (a) A_0 hat *einen* Eigenwert $\lambda = 1$ (von geometrischer Vielfaltigkeit zwei), aber *zwei* Eigenvektoren
$$x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \; x_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

   (b) Ist $\epsilon \neq 0$, hat A_ϵ *einen* Eigenwert $\lambda = 1$ und *einen* Eigenvektor
$$x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$
   Die Anzahl der Eigenvektoren von A_ϵ ändert sich deshalb von 2 (für $\epsilon = 0$) nach 1 (für $\epsilon > 0$) .

2.
$$B_\epsilon := \begin{pmatrix} 0 & \epsilon \\ 1 & 0 \end{pmatrix}.$$

   (a) B_0 hat *einen* Eigenwert ($\lambda = 0$) und *einen* Eigenvektor
$$x_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

   (b) Für $\epsilon > 0$ hat B_ϵ *zwei* Eigenwerte
$$\lambda_1(\epsilon) = +\epsilon^{1/2} , \; \lambda_2(\epsilon) = -\epsilon^{1/2}$$
   und *zwei* reelle Eigenvektoren:
$$x_1(\epsilon) = \begin{pmatrix} \epsilon^{1/2} \\ 1 \end{pmatrix}, \; x_2(\epsilon) = \begin{pmatrix} -\epsilon^{1/2} \\ 1 \end{pmatrix}.$$
   Es ist zu bemerken, daß obwohl $\Delta B = B_\epsilon - B_0 = 0(\epsilon)$ ist, gilt:
$$\begin{aligned} \Delta\lambda_1 &= \lambda_1(\epsilon) - \lambda_1(0) = 0(\epsilon^{1/2}) , \\ \Delta x_1 &= x_1(\epsilon) - x_1(0) = 0(\epsilon^{1/2}) . \end{aligned}$$
   Für ϵ klein ist $|\Delta\lambda|/\|\Delta B\|$ sehr groß.

(c) Für $\epsilon < 0$ hat B_ϵ komplexe Eigenwerte und Eigenvektoren.

Da das Eigenwertproblem für allgemeine Matrizen schlecht konditioniert sein kann, beschränken wir uns i.A. auf zwei Klassen von Matrizen:

1. Allgemeine Matrizen A mit n verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_n$, d.h. $\lambda_i \neq \lambda_j$ für $i \neq j$,

2. Hermitesche Matrizen A , d.h. $A^H = A$, wobei $A^H = (a_{ij}^H)$ mit $a_{ij}^H = \overline{a}_{ji}$.

Für solche Matrizen ist das Eigenwertproblem gut konditioniert.

Die Eigenwerte von A sind die Nullstellen des charakteristischen Polynoms $\det(A - \lambda E)$. Es liegt deshalb nahe, die Koeffizienten des charakteristischen Polynoms auszuwerten und dann die Nullstellen des Polynoms numerisch zu berechnen. Dies ist aber nicht zu empfehlen, weil die Berechnung der Nullstellen eines Polynoms oft schlecht konditioniert ist, wie das Beispiel von Wilkinson zeigt.

Werden in der Praxis die Nullstellen eines Polynoms gewünscht, sollte zunächst gefragt werden, ob das Polynom aus einem Eigenwertproblem entstanden ist und - falls ja - sollte das ursprüngliche Problem gelöst werden.

Im allgemeinen Fall sollte man invariante Unterräume statt Eigenvektoren betrachten. Der Unterraum $M \subset \mathbb{R}^n$ heißt *invariant* bzgl. der Matrix $A \in \mathrm{Mat}(n,n)$, falls $Ax \in M$ für alle $x \in M$.

Die Jordansche Normalform einer Matrix A entspricht einer Zerlegung von $\mathbb{R}^n$ in invariante Unterräume von A :

$$\begin{aligned} \mathbb{R}^n &= \sum_{k=1}^{r} M_k \ , \\ AM_k &\subset M_k \ . \end{aligned}$$

Diese Zerlegung ist gut konditioniert.

2.4 Theoretische Grundlagen

Die benötigten Eigenschaften von Matrizen werden jetzt kurz skizziert:

Satz 2.1 *Sei $A \in$ Mat$(n,n,\mathbb{C})$ mit n verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_n$.*

1. *Es gibt n linear unabhängige rechte Eigenvektoren $x_1, \ldots, x_n$:*

$$Ax_i = \lambda_i x_i \ , \ x_i \neq 0 \ , \ 1 \leq i \leq n$$

und n linear unabhängige linke Eigenvektoren $y_1, \ldots, y_n$:

$$A^T y_i = \lambda_i y_i \ , \ y_i \neq 0 \ , \ 1 \leq i \leq n \ .$$

2. *Jeder Eigenvektor $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ ist bis auf einen Faktor eindeutig bestimmt.*

3.

$$x_i^T y_j = \begin{cases} d_i \neq 0 & , \ \textit{falls} \quad i = j \ , \\ 0 & , \ \textit{falls} \quad i \neq j \ . \end{cases}$$

4. *Sei*

$$X = (x_1, \ldots, x_n) \ , \quad Y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} \ ,$$

$$\Lambda = \textit{diag}(\lambda_i) \ , \ D = \textit{diag}(d_i).$$

Dann gilt:

$$\begin{aligned} AX &= X\Lambda \ , \\ YX &= D \ . \end{aligned}$$

Bemerkung: Wenn A echt komplex ist, ist $A^T \neq A^H$.

Beweis:

1. Da λ_i ein Eigenwert von A ist, existiert x_i mit

$$Ax_i = \lambda_i x_i \ , \ x_i \neq 0 \ .$$

Daraus schließt man, daß

$$\det(A^T - \lambda_i E) = \det(A - \lambda_i E) = 0 \ ,$$

so daß λ_i ein Eigenwert von A^T ist und deshalb y_i existiert mit

$$A^T y_i = \lambda_i y_i \ .$$

Falls $x_1, \ldots, x_n$ nicht linear unabhängig sind, existiert (eventuell nach Umordnung von $x_1, \ldots, x_n$) $r > 1$ und $\alpha_i \neq 0$ für $1 \leq i \leq r$:

$$\sum_{i=1}^{r} \alpha_i x_i = 0.$$

Durch Multiplikation mit $(A - \lambda_r E)$ erhält man:

$$\sum_{i=1}^{r} \alpha_i(\lambda_i - \lambda_r)x_i = \sum_{i=1}^{r-1} \alpha_i(\lambda_i - \lambda_r)x_i = 0 \; .$$

Da $\alpha_i(\lambda_i - \lambda_r) \neq 0$, erreicht man durch Wiederholung des Widerspruchs, daß

$$\alpha_1 x_1 = 0 \; , \; \alpha_1 \neq 0 \; .$$

2. Sei z ein Eigenvektor zum Eigenwert λ_j. Dann gilt:

$$z = \sum_{1}^{n} \alpha_i x_i$$

und durch Multiplikation mit $(A - \lambda_j E)$

$$0 = \sum_{1}^{n} \alpha_i(\lambda_i - \lambda_j)x_i \; ,$$

woraus folgt, daß $\alpha_i = 0$ für $i \neq j$ und deshalb $z = \alpha_j x_j$.

3. Aus der Identität

$$\lambda_i y_j^T x_i = y_j^T A x_i = x_i^T A^T y_j = \lambda_j x_i^T y_j$$

folgt, daß $y_j^T x_i = 0$ falls $i \neq j$. Es existiert α_k^j :

$$y_j = \sum_{k=1}^{n} \alpha_k^j x_k \; .$$

Da $x_k^T y_j = 0$ für $k \neq j$, gilt

$$y_j^T y_j = \alpha_j^j y_j^T x_j \; .$$

Da $y_j \neq 0$, ist $d_j := x_j^T y_j \neq 0$.

4. Folgt aus 1 - 3.

$\square$

Definition 2.2 *Eine Matrix* $A \in Mat(n,n,\mathcal{C})$ *heißt* hermitesch, *wenn* $A = A^H$ *mit* $A^H = \overline{A}^T$ *, d.h.*

$$a_{ij}^H := \overline{a}_{ji} \; , \; 1 \leq i,j \leq n \; .$$

Satz 2.2 *A sei eine $n \times n$ komplexe hermitesche Matrix. Dann hat A n reelle Eigenwerte $\lambda_1, \ldots, \lambda_n$ und dazu n linear unabhängige orthogonale Eigenvektoren $x_1, \ldots, x_n$:*

$$\begin{aligned} Ax_i &= \lambda_i x_i \; , \; 1 \leq i \leq n \; , \\ A^H x_i &= \lambda_i x_i \; , \; 1 \leq i \leq n \; , \end{aligned}$$

$$x_i^H x_j = \begin{cases} 1 & , \textit{ falls} \quad i = j \\ 0 & , \textit{ falls} \quad i \neq j \end{cases} .$$

Sei $U = (x_1, \dots, x_n)$, $\Lambda = diag(\lambda_i)$. *Dann gilt:*

$$\begin{aligned} AU &= U\Lambda \,, \\ U^H U &= E \,, \; (U \quad \textit{ist unitär}) \\ U^H A U &= \Lambda \,. \end{aligned}$$

Falls A *reell ist, ist* U *auch reell und damit orthogonal.*

Beweis: Folgt aus Satz 2.1, falls $\lambda_1, \dots, \lambda_n$ verschieden sind. Sonst durch Induktion. □

Satz 2.3 *Seien* A, B *und* T $n \times n$ *Matrizen mit* T *nicht singulär und* $B = T^{-1}AT$. *Sei* (λ, x) *ein Eigenpaar von* A *, dann ist* $(\lambda, T^{-1}x)$ *ein Eigenpaar von* B. A *und* B *sind* ähnlich.

Satz 2.4 (Gerschgorin) *a) Die Vereinigung aller (Gerschgorin–) Kreisscheiben*

$$K_i := \{\mu \in \mathcal{C} \,|\, |\mu - a_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}|\}$$

enthält alle Eigenwerte der $n \times n$*-Matrix* $A = (a_{ik})$.

b) Ist die Vereinigung

$$M_1 := \bigcup_{i=1}^{k} K_i$$

disjunkt von der Vereinigung M_2 *der übrigen Kreisscheiben, so enthält* M_1 *genau* k *algebraische Eigenwerte von* A *und* M_2 *genau* $n - k$ *algebraische Eigenwerte.*

Beweis:

a) Sei (λ, x) , $x = (x_1, \dots, x_n)^T \in \mathcal{C}^n$ ein Eigenpaar von A. Da $x \neq 0$ gibt es ein k mit

$$|x_k| = \|x\|_\infty = \max |x_i| > 0 \,.$$

Aus der k-ten Gleichung des Systems

$$Ax = \lambda x$$

folgt

$$|a_{kk} - \lambda| \; |x_k| = |\sum_{\substack{i=1 \\ i \neq k}}^{n} a_{ki} x_i| \leq \sum_{\substack{i=1 \\ i \neq k}}^{n} |a_{ki}| \cdot \|x\|_\infty \,,$$

so daß $\lambda \in K_k$.

b) Sei $A = D + B$ mit $D := \text{diag}(A)$, $s \in [0,1]$, und

$$A(s) := D + sB \ .$$

Die entsprechenden Gerschgorin Kreisscheiben sind

$$\begin{aligned} K_i(s) &:= \{\mu \in \mathcal{C} : |\mu - a_{ii}(s)| \le \sum_{\substack{k=1\\k\neq i}}^{n} |a_{ik}(s)|\} \\ &= \{\mu \in \mathcal{C} : |\mu - a_{ii}| \le s \sum_{\substack{k=1\\k\neq i}}^{n} |a_{ik}|\} \ . \end{aligned}$$

Die Gerschgorin Kreisscheiben von $A(s)$ haben also die gleichen Mittelpunkte wie die von A, und ihre Radien sind um den Faktor s kleiner.

Die Eigenwerte von $A(0)$ sind $a_{11}, \dots, a_{nn}$, so daß M_1 und M_2 genau $n-k$ bzw. k Eigenwerte von $A(0)$ erhält.

Die (algebraischen) Eigenwerte von $A(s)$ sind die n Nullstellen des charakteristischen Polynoms

$$p(\lambda, s) = \det(A(s) - \lambda I) = (-1)^n \lambda^n + \text{niedrigere Glieder.}$$

Die algebraischen Eigenwerte ändern sich deshalb stetig mit s. Für jedes $\epsilon > 0$ gibt es folglich $\delta > 0$, so daß für $s, t \in [0,1]$ mit $|s-t| < \delta$ eine Permutation $\sigma = \sigma(s,t)$ existiert mit

$$|\lambda_i(s) - \lambda_{\sigma(i)}(t)| < \epsilon \ , \quad \text{für} \quad 1 \le i \le n \ .$$

Die Behauptung b) wird jetzt durch Widerspruch bewiesen. Sei

$$d := \inf_{\substack{x \in M_1\\y \in M_2}} |x - y|$$

Es gibt $\epsilon > 0$, so daß für $s, t \in [0,1]$ mit $|s-t| < \epsilon$ die algebraischen Eigenwerte von $A(s)$ und $A(t)$ sich so ordnen lassen, daß

$$|\lambda_i(s) - \lambda_i(t)| \le d/2 \ , \ 1 \le i \le n \ .$$

$M_1(0)$ enthält genau k der Eigenwerte $\lambda_i(0)$. Wenn $M_1(1)$ nicht auch genau k der Eigenwerte $\lambda_i(1)$ enthält, gibt es s und t , so daß

1. $|s-t| < \epsilon$
2. $M_1(s)$ und $M_1(t)$ eine unterschiedliche Anzahl der Nullstellen $\lambda_i(s)$ bzw. $\lambda_i(t)$ enthalten. Ohne Einschränkung nehme man an, daß

$$\lambda_i(s) \in \begin{cases} M_1(s) & , \quad i \le k \\ M_2(s) & , \quad \text{sonst} \end{cases}$$

$$\lambda_i(t) \in \begin{cases} M_1(t) & , \quad i \le r \\ M_2(t) & , \quad \text{sonst} \end{cases}$$

mit $k > r$. Es gilt aber, daß

$$|\lambda_{r+1}(t) - \lambda_{r+1}(s)| < \epsilon < \frac{d}{2} \; .$$

Da $\lambda_{r+1}(t) \in M_2$ und $\lambda_{r+1}(s) \in M_1$ und deshalb

$$|\lambda_{r+1}(t) - \lambda_{r+1}(s)| > d \; ,$$

ist ein Widerspruch vorhanden.

□

2.5 Das Jacobi–Verfahren

Das Jacobi–Verfahren ist ein iteratives Verfahren, womit eine gegebene reelle symmetrische Matrix A durch sukzessive orthogonale Transformationen U_k approximativ auf Diagonalform gebracht wird:

$$\begin{aligned} A_0 &:= A \\ A_{k+1} &:= U_k A_k U_k^T \; , \; k \ge 0 \\ A_k &\longrightarrow \Lambda \quad \text{für} \quad k \to \infty \; , \end{aligned} \tag{2.2}$$

wobei Λ eine Diagonalmatrix ist.

Da A_0 und A_k äquivalent sind, sind die Eigenwerte von A durch Λ gegeben. Ist $A_k \doteq \Lambda$ und $U = U_k U_{k-1} \dots U_1$, dann ist

$$AU^T \doteq U^T \Lambda \; ,$$

so daß U^T Approximationen zu den Eigenvektoren von A enthält.

Die einfachsten orthogonalen Matrizen sind die *ebenen Drehungen*

$$T_{pq} = T_{pq}(\varphi) = E - (1-c)(e_p e_p^T + e_q e_q^T) + s(e_p e_q^T - e_q e_p^T) \; , \; 1 \le p < q \le n$$

$$
= \begin{array}{cl}
\left(\begin{array}{cccccccc}
1 & & & & & & & O \\
& 1 & & & & & & \\
& & c & 0 & \dots & & s & \\
& & 0 & 1 & & 0 & & \\
& & & & \ddots & & & \\
& & & & & 1 & & \\
O & & -s & 0 & & & c & \\
& & & & & & & 1
\end{array}\right)
&
\begin{array}{l}
\\ \\ p\text{-te Zeile} \\ \\ \\ \\ q\text{-te Zeile} \\ \\
\end{array}
\\
\begin{array}{cccccccc}
 & & \uparrow & & & & \uparrow & \\
& & p\text{-te} & & & & q\text{-te} & \\
& & \text{Spalte} & & & & \text{Spalte} &
\end{array}
&
\end{array}
$$

wobei $c, s \in \mathbb{R}$,

$$c^2 + s^2 = 1,$$

und φ später definiert wird. (In der Literatur wird oft $A_{k+1} = U_k^T A_k U_k$ geschrieben.)

Sei A eine $n \times n$ reelle symmetrische Matrix und $\tilde{A} = T_{pq} A T_{pq}^T$. Dann ist $\tilde{A}$ auch symmetrisch, und es gilt:

$$
\begin{array}{rcl}
\tilde{a}_{ij} & = & a_{ij}\ , \quad \text{falls} \quad i \neq p, q\ ,\ j \neq p, q. \\
\tilde{a}_{pp} & = & c^2 a_{pp} + 2cs a_{pq} + s^2 a_{qq} \\
\tilde{a}_{qq} & = & s^2 a_{pp} - 2cs a_{pq} + c^2 a_{qq} \\
\tilde{a}_{pq} & = & \tilde{a}_{qp} = +cs(a_{qq} - a_{pp}) + (c^2 - s^2) a_{pq} \\
\tilde{a}_{pj} & = & \tilde{a}_{jp} = c a_{pj} + s a_{qj}\ ,\ j \neq p, q \\
\tilde{a}_{qj} & = & \tilde{a}_{jq} = -s a_{pj} + c a_{qj}\ ,\ j \neq p, q.
\end{array}
\tag{2.3}
$$

Wir untersuchen jetzt, ob es möglich ist, c und s so zu bestimmen, daß $\tilde{a}_{pq} = 0$. Dazu setzen wir

$$c = \cos(\varphi)\ ,\ s = \sin(\varphi)\ ,\ t = \tan\ \varphi\ . \tag{2.4}$$

Dann gilt $\tilde{a}_{pq} = 0$ nur dann, wenn

$$\sin(2\varphi)(a_{pp} - a_{qq}) = 2\ \cos(2\varphi) a_{pq}\ .$$

Es ist immer möglich, diese Gleichung zu lösen:

$$\varphi = \frac{1}{2} \arctan\left(\frac{2a_{pq}}{a_{pp} - a_{qq}}\right)\ . \tag{2.5}$$

Das Jacobi–Verfahren besteht wesentlich aus (2.2), (2.3) - (2.5):

$$\begin{aligned} A_0 &:= A \\ A_{k+1} &:= U_k A_k U_k^T \ , \ k \geq 0 \\ U_k &:= T_{p_k q_k}(\varphi_k) \\ \varphi_k &:= \frac{1}{2} \arctan \left(\frac{2a_{p_k q_k}^{(k)}}{a_{p_k p_k}^{(k)} - a_{q_k q_k}^{(k)}} \right) \ . \end{aligned} \tag{2.6}$$

Die Wahl der Folge $\{(p_k, q_k)\}$ und die Berechnung von φ_k muß noch definiert werden.

Die Auswahl von (p_k, q_k) wird auf folgende Methoden gebracht:

1. **Das klassische Jacobi–Verfahren**

$$|a_{p_k q_k}^{(k)}| = |a_k| := \max_{\substack{1 \leq i \ , \ j \leq n \\ i \neq j}} |a_{ij}^{(k)}| \ .$$

   $a_{p_k q_k}^{(k)}$ sollte also ein betragsgrößtes nichtdiagonales Element von A_k sein. Dies wurde von Jacobi in 1846 vorgeschlagen.

2. **Das allgemeine zyklische Jacobi–Verfahren**
   Auf einem Computer ist es zeitraubend, das betragsgrößte Element von A_k zu finden. Deshalb werden oft die Paare (p_k, q_k) in einer vorgeschriebenen Reihenfolge genommen. Sei

$$N = n(n-1)/2 \ .$$

   Sei

$$m_r = (u_r, v_r) \ , \ 0 \leq r < N \ ,$$

   mit

   (a) $u_r, v_r \in \mathbb{N}$;

   (b) $1 \leq u_r < v_r \leq n$;

   (c) Für jedes $i, j \in \mathbb{N}$ mit $1 \leq i < j \leq n$ gibt es r mit $m_r = (i, j)$.

   Sei $M = (m_1, \ldots, m_N)$. Dann werden die Paare (p_k, q_k) folgendermaßen definiert:

$$(p_k, q_k) = m_r \ , \ r = k(\text{mod } N) \ .$$

   Nach N Drehungen ist jedes nichtdiagonale Element von A genau einmal durch Null ersetzt worden, und der Prozeß wiederholt sich wieder. Deshalb wird eine Folge von N Drehungen als *Zyklus* bezeichnet.

3. **Das (spezielle) zyklische Jacobi–Verfahren**
   In das spezielle zyklische Jacobi–Verfahren wird

$$\begin{aligned} M = M_s \quad := \quad & ((1,2)\ ,\ (1,3),\dots,(1,n)\ ,\ (2,3),\dots,(2,n) \\ ,\dots,\ & (i,i+1),\dots,(i,n),\dots,(n-1,n)) \end{aligned}$$

   gesetzt.

4. **Das allgemeine zyklische Jacobi–Verfahren mit Grenzen**
   Dieses Verfahren unterscheidet sich von dem allgemeinen zyklischen Jacobi–Verfahren nur dadurch, daß eine (p_k, q_k) Drehung nur dann gemacht wird, wenn $|a_{pq}^{(k)}|$ eine vorgeschriebene Grenze $\epsilon_k(A_k)$ überschreitet. Die (p_k, q_k) Drehung wird nur gemacht, wenn

$$|a_{pq}^{(k)}| \geq \epsilon_k(A_k)\ .$$

   Zwei übliche Methoden, um ϵ_k zu bestimmen, sind:

   (a)

$$\epsilon_k(A_k) = \left[\frac{1}{2n^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} (a_{ij}^{(k)})^2\right]^{1/2} \tag{2.7}$$

   (b) $\epsilon_0(A_0 := \max |a_{ij}|/10)$

$$\epsilon_k(A_k) := \begin{cases} \epsilon_{k-1}(A_{k-1}) & , \text{ falls } \quad (i,j) \quad \text{existiert mit} \\ & \ |a_{ij}^k| \geq \epsilon_{k-1}(A_{k-1}) \\ \epsilon_{k-1}(A_{k-1})/10 & , \text{ sonst.} \end{cases} \tag{2.8}$$

Berechnung von φ_k und A_{k+1}

Nach Rutishauser (1966) können die Formeln (2.3) - (2.5) günstiger geschrieben werden:

$$\theta := \cot(2\varphi) = (a_{pp} - a_{qq})/2a_{pq}\ .$$

Dann gilt mit $t := \tan\ \varphi$:

$$t^2 + 2t\theta - 1 = 0\ . \tag{2.9}$$

t wird als betragsmäßig kleinste Nullstelle von (2.9) genommen:

$$t_1 := [|\theta| + (\theta^2 + 1)^{1/2}]^{-1}$$

$$t := \begin{cases} t_1 & , \text{ falls } \quad \theta \geq 0 \\ -t_1 & , \text{ sonst.} \end{cases}$$

$$\left.\begin{array}{rcl} c & := & [1+t^2]^{-1/2} \\ s & := & t \cdot c \\ \tau & := & \tan(\varphi/2) = s/(1+c) \end{array}\right\}$$

$$\begin{array}{rcl} \tilde{a}_{pp} & = & a_{pp} + t a_{pq} \ , \ \tilde{a}_{qq} = a_{qq} - t a_{pq} \\ \tilde{a}_{pj} & = & a_{pj} + s(a_{qj} - \tau a_{pj}) \ , \ j \neq p, q \\ \tilde{a}_{qj} & = & a_{qj} - s(a_{pj} + \tau a_{qj}) \ , \ j \neq p, q \end{array}$$

Mit diesen Formeln sind die Rundungsfehler kleiner als mit den vorherigen Formeln.

Hilfssatz 2.1 *Sei A_k reell und symmetrisch. Sei $T_{pq}(\varphi_k)$ und A_{k+1} durch (2.6) gegeben. Dann gilt:*

$$S(A_{k+1}) := \sum_{1 \le i < j \le n}^{n} [a_{i,j}^{(k+1)}]^2 = S(A_k) - [a_{pq}^{(k)}]^2 \ .$$

Beweis:

Methode 1:

Nur die p-ten und q-ten Zeile von A_k ändern sich. Für $j \neq p, q$ gilt:

$$\begin{pmatrix} a_{pj}^{(k+1)} \\ a_{qj}^{(k+1)} \end{pmatrix} = \hat{T}_{pq} \begin{pmatrix} a_{pj}^{(k)} \\ a_{qj}^{(k)} \end{pmatrix} ,$$

mit

$$\hat{T}_{pq} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} .$$

Da $\hat{T}_{pq}$ orthogonal, ist:

$$\left(a_{pj}^{(k+1)}\right)^2 + \left(a_{qj}^{(k+1)}\right)^2 = \left(a_{pj}^{(k)}\right)^2 + \left(a_{qj}^{(k)}\right)^2 \ . \tag{2.10}$$

Da A_k und A_{k+1} symmetrisch sind, gilt:

$$\begin{array}{rcl} 2S(A_{k+1}) & - & 2S(A_k) = \displaystyle\sum_{\substack{j=1 \\ j \neq p}}^{n} \left[\left(a_{pj}^{(k+1)}\right)^2 - \left(a_{pj}^{(k)}\right)^2 \right] \\ & + & \displaystyle\sum_{\substack{j=1 \\ j \neq q}}^{n} \left[\left(a_{qj}^{(k+1)}\right)^2 - \left(a_{qj}^{(k)}\right)^2 \right] + \sum_{\substack{i=1 \\ i \neq p,q}}^{n} \left[\left(a_{ip}^{(k+1)}\right)^2 - \left(a_{ip}^{(k)}\right)^2 \right] \\ & + & \displaystyle\sum_{\substack{i=1 \\ i \neq p,q}}^{n} \left[\left(a_{iq}^{(k+1)}\right)^2 - \left(a_{iq}^{(k)}\right)^2 \right] \\ & = & \displaystyle\sum_{\substack{j=1 \\ j \neq p,q}}^{n} \left[\left(a_{pj}^{(k+1)}\right)^2 + (a_{qj}^{k+1})^2 - \left(a_{pj}^{(k)}\right)^2 - \left(a_{pj}^{(k)}\right)^2 \right] \\ & + & 2\left[\left(a_{pq}^{(k+1)}\right)^2 - \left(a_{pq}^{(k)}\right)^2 \right] \end{array}$$

Wegen (2.10) gilt:

$$2S(A_{k+1}) - 2S(A_k) = 2\left[\left(a_{pq}^{(k+1)}\right)^2 - \left(a_{pq}^{(k)}\right)^2\right] .$$

Methode 2:

Wir benutzen die folgenden Beobachtungen:

1. Sei U eine orthogonale Matrix und x ein Vektor. Dann gilt:

$$\|Ux\|_2 = \|x\|_2$$

2. Sei A eine quadratische Matrix, U eine orthogonale Matrix und $\tilde{A} = UAU^T$. Dann gilt (wegen 1)

$$\sum_{i,j=1}^{n} a_{ij}^2 = \sum_{i,j=1}^{n} \tilde{a}_{ij}^2 ,$$

3. Sei $\tilde{A} = UAU^T$. Dann gilt

$$2S(A) + \sum_{i=1}^{n} a_{ii}^2 = 2S(\tilde{A}) + \sum_{i=1}^{n} \tilde{a}_{ii}^2$$

4. Sei weiter

$$A_1 = \begin{pmatrix} a_{pp}^{(k)} & a_{pq}^{(k)} \\ a_{qp}^{(k)} & a_{qq}^{(k)} \end{pmatrix}$$

$$\begin{aligned} \tilde{A}_1 &= \hat{T}_{pq} A_1 \hat{T}_{pq}^T , \\ \tilde{a}_{pq} &= 0 . \end{aligned}$$

Dann gilt wegen (2):

$$\left(a_{pp}^{(k)}\right)^2 + \left(a_{qq}^{(k)}\right)^2 + 2\left(a_{pq}^{(k)}\right)^2 = \left(a_{pp}^{(k+1)}\right)^2 + \left(a_{qq}^{(k+1)}\right)^2 .$$

5. $S(\tilde{A}) = S(A) - \left(a_{pq}^{(k)}\right)^2$.

□

Hilfssatz 2.2 *Für die Matrizen* A_k

$$A_{k+1} := U_k A_k U_k^T$$

gilt

$$S(A_k) := \sum_{i<j} \left(a_{ij}^{(k)}\right)^2 \longrightarrow 0 , \quad \textit{für} \quad k \to \infty$$

und $|\varphi_k| \le \frac{\pi}{4}$*. Dann gibt es* $\Lambda = diag(\lambda_i)$ *, so daß*

$$A_k \longrightarrow \Lambda , \quad \textit{als} \quad k \to \infty .$$

Beweis: Wilkinson, Seite 268, und van Kempen, Seite 19. □

Satz 2.5 *Das klassische Jacobi–Verfahren ist (linear) konvergent.*

Beweis: Da $a^{(k)}_{p_k q_k}$ ein betragsmäßig größtes Nichtdiagonalelement von $A^{(k)}$ ist, gilt

$$|a_{p_k q_k}|^2 \geq S(A_k)/N$$

mit $N := n(n-1)/2$ und deshalb, mit Hilfe von Hilfssatz 2.1,

$$\begin{aligned} S(A_{k+1}) &= S(A_k) - \left(a^{(k)}_{pq}\right)^2 \\ &\leq (1 - 1/N) S(A_k) \; . \end{aligned}$$

□

Bemerkung: Das klassische Jacobi-Verfahren ist sogar quadratisch konvergent, d.h.

$$S(A_{k+N}) \leq K(S(A_k))^2$$

(siehe van Kampen [1966]).

Satz 2.6 *Das allgemeine zyklische Jacobi–Verfahren mit Grenzen ist konvergent, falls ϵ_k durch (2.7) oder (2.8) definiert ist.*

Die Konvergenz des allgemeinen Jacobi–Verfahrens kann nur unter Nebenbedingungen gewährleistet sein, wie die folgenden Beispiele zeigen:

Beispiel: (Forsythe und Henrici, 1960, S. 16)

$$A_0 = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 4 \end{pmatrix} , \quad \varphi_k = \pi/2 \; ,$$

$$M = M_s = ((1,2),(1,3),(2,3)) \; .$$

Dann ist $A_{k+6} = A_k$ und $S(A_k) = S(A_0)$ für alle k .

Beispiel: (Forsythe und Henrici, 1960, S. 17)

$$A_0 = \begin{pmatrix} a & \epsilon & 1 \\ \epsilon & a+c & 0 \\ 1 & 0 & a+2c \end{pmatrix} , \quad 0 < \epsilon \leq 1 \; , \; c \geq 4.$$

$M = M_s$ wie in Beispiel 2.5. $\varphi_k \in (\pi/4, \pi/2)$. Dann gilt $S(A_k) \longrightarrow 0$ nicht.

Beispiel: (Hansen, 1963, S. 451)

$$A = \begin{pmatrix} a & 0 & 0 & -b \\ 0 & a & b & 0 \\ 0 & b & a & 0 \\ -b & 0 & 0 & a \end{pmatrix} , \quad a \neq 0 \; , \; b \neq 0 \; ,$$

$$M = ((1,2),(3,4),(1,4),(2,3),(1,3),(2,4)) .$$

$$\varphi_k = \begin{cases} -\pi/4 & , \text{ falls } \quad k = 12r + t \quad \text{mit} \quad r \in \mathbb{N} \quad \text{und} \quad t = 1,9,10 \\ +\pi/4 & , \text{ sonst.} \end{cases}$$

Dann gilt $A_{k+12} = A_k$. Dieses Beispiel ist heutzutage von Interesse, weil Brent und Luk (1982) es als Basis für einen parallelen Algorithmus verwandt haben.

Wie die Beispiele 2.5 und 2.5 zeigen, kann es vorkommen, daß das (spezielle) zyklische Jacobi–Verfahren nicht konvergiert, falls $|\varphi_k| \geq \pi/4$. Es ist aber immer möglich, $|\varphi_k| \leq \pi/4$ zu nehmen, und unter diesen zusätzlichen Voraussetzungen läßt sich die Konvergenz beweisen.

Satz 2.7 *Für das spezielle zyklische Jacobi–Verfahren mit $|\varphi_k| \leq \pi/4$ gilt:*

$$S(A_{(r+1)N}) \leq S(A_{rN})(1 - 2^{-N})^{1/2} .$$

Beweis: Henrici und Zimmermann, 1968. □

Satz 2.8 *Seien die Eigenwerte von A verschieden, dann ist das spezielle zyklische Jacobi–Verfahren quadratisch konvergent. Es gibt also eine Konstante $K = K(A)$:*

$$S(A_{k+N}) \leq K[S(A_k)]^2 .$$

Beweis: Henrici und Zimmermann, 1968, S. 491. □

Die quadratische Konvergenz des speziellen zyklischen Jacobi–Verfahrens ist die theoretische Erklärung dafür, daß in der Praxis nur 5 - 10 Zyklen nötig sind, bevor die nichtdiagonalen Elemente von A^k bis zur Maschinengenauigkeit Null sind.

Parallele Jacobi-Verfahren werden in Golub und van Loan [1989] diskutiert.

Offene Fragen

1. Warum kann Rutishauser annehmen, daß mit seinen Algorithmen alle nichtdiagonalen Elemente Null werden?

2. Unter welchen Bedingungen konvergiert das allgemeine zyklische Jacobi–Verfahren?

Literatur

Dollinger, E.: Ein linear konvergentes zyklisches jacobiähnliches Verfahren für beliebige reelle Matrizen. Numer. Math. 38, 245-253 (1981).

Forsythe, G.E., **Henrici**, P.: The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix. Trans. Amer. Math. Soc. 94, 1-23(1960).

Golub, G.H., **van Loan**, C.F.: Matrix Computations. Second edition. John Hopkins Univ. Press, 1989.

Hansen, E.R.: On Cyclic Jacobi Methods. J. Soc. Indust. Appl. Math., Vol. 11, No. 2, June 1963.

Henrici, P., **Zimmermann**, Katharina: An Estimate for the Norms of Certain Cyclic Jacobi Operators. Linear Algebra and Its Applications 1, 489-501(1968).

Kempen van, H.P.M.: On the Quadratic Convergence of the Special Cyclic Jacobi Method. Numer. Math. 9, 19-22(1966).

Rutishauser, H.: The Jacobi method for real symmetric matrices. Numer. Math. 9, 1-10(1966).

Wilkinson, J.H. und **Reinsch**, C.: Linear Algebra. New York, Springer, 1971.

Wilkinson, J.H.: Rounding Errors in Algebraic Processes. Englewood Cliffs, N.J., Prentice Hall, 1963.

Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford, Clarendon Press, 1965.

2.6 Die Potenzmethode (Vektoriteration)

Sei A eine komplexe $n \times n$-Matrix mit Eigenwerten $\lambda_1, \ldots, \lambda_n$,

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$$

und Eigenvektoren $x_1, \ldots, x_n$.

Ausgehend von einem Startvektor $x^{(0)}$ bildet man der Reihe nach die Vektoren

$$x^{(k+1)} := Ax^{(k)} = A^k x^{(0)} \ , \ k = 0, 1, \ldots \ . \tag{2.11}$$

Die Eigenvektoren x_i bilden eine Basis des $\mathcal{C}^n$, und es gibt einen Vektor

$$c = (c_1, \ldots, c_n)^T \in \mathcal{C}^n$$

mit

$$x^{(0)} = \sum_{i=1}^{n} c_i x_i \tag{2.12}$$

Es folgt aus (2.11) und (2.12)

$$x^{(k)} = \sum_{i=1}^{n} c_i \lambda_i^k x_i = \lambda_1^k (c_1 x_1 + r_k) \tag{2.13}$$

mit

$$r_k := \sum_{i=2}^{n} \left(\frac{\lambda_i}{\lambda_1}\right)^k c_i x_i = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) ,$$

so daß

$$r_k \longrightarrow 0 \quad \text{für} \quad k \longrightarrow \infty .$$

Zur Berechnung von λ_1 wählt man einen komplexen Vektor $d \in \mathcal{C}^n$ und bildet

$$\alpha^{(k)} := (x^{(k)}, d) := d^H x^{(k)} .$$

Es folgt dann aus (2.13)

$$\frac{\alpha^{(k+1)}}{\alpha^{(k)}} = \frac{d^H x^{(k+1)}}{d^H x^{(k)}} = \lambda_1 \frac{c_1 d^H x_1 + d^H r_{k+1}}{c_1 d^H x_1 + d^H r_k} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \longrightarrow \lambda_1 \quad \text{für} \quad k \to \infty.$$

Beispiel:

$$A = \begin{bmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{bmatrix}$$

k	$\alpha^{(k)}/\alpha^{(k-1)}$
2	12,1429
3	10,8118
4	11,0348
5	10,9937
6	11,0011
7	10,9998
8	11,0000

mit

$$\begin{aligned} x^{(0)} &= (0,0,1)^T \\ d &= (1,1,1)^T \end{aligned}$$

Die Konvergenzgeschwindigkeit der Vektoriteration wird durch den Wert von $\left|\frac{\lambda_2}{\lambda_1}\right|$ bestimmt. Es gibt mehrere Methoden, die Konvergenzgeschwindigkeit zu steigern:

1. **Verschiebungen (Shifts)** Die Potenzmethode wird auf die Matrix $B = A - \mu I$ angewandt, wobei die Verschiebung μ so gewählt wird, daß

   $$\frac{\max_{2 \leq i \leq n} |\lambda_i - \mu|}{|\lambda_1 - \mu|}$$

   möglichst klein ist.

2. **Inverse Iteration (Wielandt)** Sei ein guter Näherungswert μ für einen der Eigenwerte $\lambda_1, \dots, \lambda_n$ bekannt. Es soll gelten:

$$|\lambda_j - \mu| << |\lambda_k - \mu| \,, \quad \text{für alle} \quad k \neq j \,.$$

Ausgehend von einem Startvektor $x^{(0)}$ bildet man die Vektoren $x^{(k)}$,

$$x^{(k+1)} = (A - \mu I)^{-1} x^{(k)} \,,\ k \geq 0 \,.$$

Daraus folgt

$$(A - \mu I) x^{(k+1)} = x^{(k)} \,.$$

Die Eigenwerte von $(A - \mu I)^{-1}$ sind

$$\beta_1 := \frac{1}{\lambda_1 - \mu}, \dots, \beta_n := \frac{1}{\lambda_n - \mu} \,.$$

Der betragsmäßig größte Eigenwert ist

$$\beta_j = \frac{1}{\lambda_j - \mu}$$

und die Konvergenzgeschwindigkeit ist recht gut, da

$$|\beta_j| >> |\beta_k| \,, \quad \text{für} \quad k \neq j \,.$$

In der Praxis wird inverse Iteration dafür angewandt, den Eigenvektor x_j zum λ_j zu berechnen.

2.7 Das QR-Verfahren

Das QR-Verfahren ist heutzutage das meistbenutzte Verfahren für Eigenwertprobleme. Es ist eine Verbesserung des LR-Verfahrens von Rutishauser und wurde im Jahre 1961 von Francis veröffentlicht. Das Verfahren besteht aus mehreren Teilen:

1. Zuerst wird die ursprüngliche reelle Matrix durch unitäre Ähnlichkeitstransformation in eine Matrix A von einfachrerer Gestalt transformiert. Falls die ursprüngliche Matrix symmetrisch ist, ist A eine tridiagonale Matrix, d.h. $a_{ij} = 0$ falls $|i - j| > 1$, sonst ist A eine (obere) Hessenberg–Matrix, d.h. $a_{ij} = 0$ falls $j < i - 1$. Der Grund für diesen ersten Schritt ist, daß die folgenden Schritte viel effizienter sind, wenn sie auf Matrizen von einfachrerer Gestalt angewendet werden.

2. Die neue Matrix A wird durch unitäre Ähnlichkeitstransformationen transformiert:

$$\begin{aligned} A_1 &:= A \\ A_k &= Q_k R_k \\ A_{k+1} &:= R_k Q_k \end{aligned}$$

wobei Q_k eine orthogonale Matrix ist und R_k eine rechte Dreiecksmatrix.

3. Der Prozeß wird durch eine *Shiftstrategie* und durch spezielle Algorithmen für komplexe Eigenwerte beschleunigt. Diese sind äußerst wichtig, da hierdurch die Effizienz des Algorithmus sehr stark erhöht wird.

2.7.1 Reduktion auf Hessenberg–Gestalt

Hilfssatz 2.3 *Für jeden Vektor $x \in \mathbb{R}^m$ gibt es eine* Householder Matrix

$$T = E - 2uu^T$$

wobei E die Einheitsmatrix ist und $u \in \mathbb{R}^m$, so daß

1. $\|u\|_2 = 1$,
2. T *ist orthogonal,*
3. $Tx = -\sigma e_1$, $\sigma \in R^1$.

Beweis: Sei $x = (\xi_1, \xi_2, \ldots, \xi_m)$,

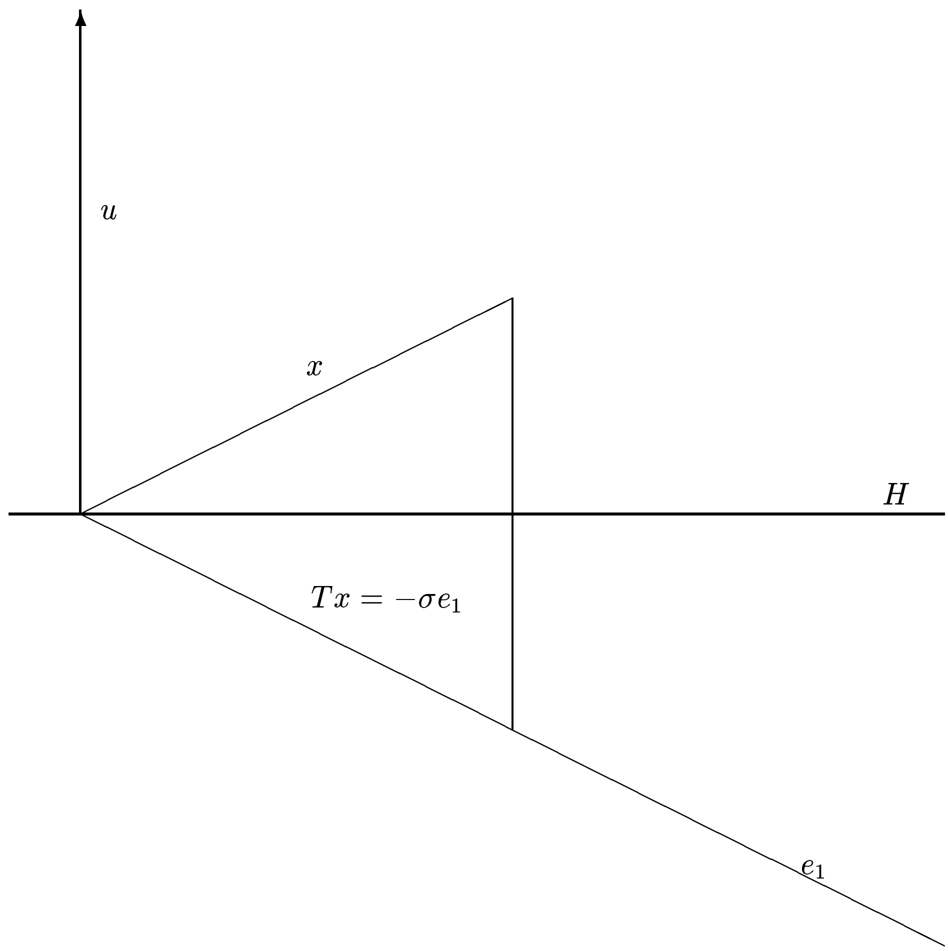
$$\begin{aligned} \sigma &:= \mathrm{sign}(\xi_1)\|x\|_2 \, , \\ v &:= x + \sigma e_1 \, , \\ u &:= v/\|v\| \, . \end{aligned} \tag{2.14}$$

Dann gilt:

$$\begin{aligned} \|v\|^2 &= (x^T + \sigma e_1^T)(x + \sigma e_1) \\ &= \|x\|^2 + \sigma^2 + 2|\xi_1| \; \|x\| \\ &= 2(\|x\|^2 + |\xi_1| \; \|x\|) \\ &= 2\|x\| \; (\|x\| + |\xi_1|). \\ Tx &= (I - 2uu^T)x \\ &= x - 2x\frac{vv^T}{\|v\|^2} \\ &= x - 2(x + \sigma e_1)(x^T + \sigma e_1^T)x/\|v\|^2 \\ &= x - \frac{(x + \sigma e_1)(\|x\|^2 + |\xi_1| \; \|x\|)}{\|x\|^2 + |\xi_1| \; \|x\|} \\ &= -\sigma e_1 \, . \end{aligned}$$

□

Bemerkung: Um Exponentenüber- oder -unterlauf zu vermeiden, betrachtet man $x/\max|\xi_i|$ statt x .



Bemerkung: T ist eine Spiegelung in der Ebene $H = \{x : x^T u = 0\}$.

Beispiel:

$$\begin{aligned} x &= (6,3,2)^T \\ \|x\| &= 7 \\ \sigma &= 7 \\ v &= (13,3,2)^T \\ u &= \frac{1}{(182)^{1/2}}\,(13,3,2)^T\ . \end{aligned}$$

Bemerkung: Für Berechnungen mit Papier und Bleistift ist es oft leichter, (2.14) durch $\sigma = -\text{sign}(\xi_1)\|x\|_2$ zu ersetzen, da die Zahlen dann kleiner sind. Dies ist natürlich nur möglich, wenn $x \neq \xi_1 e_1$ ist. Im oberen Beispiel bekommen wir dann $v = (-1,3,2)^T$ und $u = (-1,3,2)/(14)^{1/2}$.

Sei nun

$$A_r = \begin{pmatrix} H_r & & \vdots & C_r \\ \multicolumn{4}{c}{\dotfill} \\ O & \vdots & b_r & \vdots & B_r \end{pmatrix}$$

eine $n \times n$ reelle Matrix, wobei

C_r sei eine $r \times (n-r)$ Matrix
B_r sei eine $(n-r) \times (n-r)$ Matrix
b_r sei ein $(n-r)$ Vektor
H_r sei eine $r \times r$ obere Hessenberg–Matrix, d.h. $H_r = (h_{ij})$ mit $h_{ij} = 0$ für $j < i - 1$.

Sei P_r die Matrix

$$P_r = \begin{pmatrix} I & \vdots & O \\ \multicolumn{3}{c}{\dotfill} \\ O & \vdots & T_r \end{pmatrix} ,$$

$T_r = E_{n-r} - 2u_r u_r^T$, mit $T_r b_r = -\sigma_r e_{n-r}$. Dann gilt:

$$A_{r+1} := P_r A_r P_r = \begin{pmatrix} H_{r+1} & & \vdots & C_{r+1} \\ \multicolumn{4}{c}{\dotfill} \\ O & \vdots & b_{r+1} & \vdots & B_{r+1} \end{pmatrix}$$

mit

$$H_{r+1} := \begin{pmatrix} H_r & & \vdots & c_r \\ \multicolumn{4}{c}{\dotfill} \\ O & \vdots & -\sigma_r & \vdots & d_r \end{pmatrix} ,$$

Durch Wiederholung erreichen wir

$$A_n = H_n = P_{n-1} \ldots P_1 A_1 P_1 \ldots P_{n-1} = QAQ^T ,$$

wobei A_n eine obere Hessenberg–Matrix und Q eine orthogonale Matrix ist.

Bemerkung: Es ist auch möglich, A_1 zur Hessenberg–Gestalt mit Givens Transformationen (ebene Drehungen) zu bringen. Die benötigten Multiplikationen sind:

Givens: $\sum (n-r-1)[4n + 4(n-r-1)] = \frac{10n^3}{3} + 0(n^2)$

Householder: $5n^3/3 + 0(n^2)$.

Die Givens Transformation wurde zuerst benutzt, später die Householder Transformation, da sie nur halb soviel Operationen benötigt. Noch später zeigte sich, daß die Berechnungen mit Givens Transformation so organisiert werden können, daß der Rechenaufwand mit dem für die Householder Transformation vergleichbar ist.

2.7.2 Konvergenz des QR-Verfahrens

Ist A durch Ähnlichkeitstransformationen zur einfacheren Gestalt gebracht worden, dann wird das QR-Verfahren angewendet.

Das QR-Verfahren ist wie folgt definiert:

$$\begin{aligned} A_1 &:= A && (2.15)\\ A_s &= Q_s R_s && (2.16)\\ A_{s+1} &:= R_s Q_s && (2.17) \end{aligned}$$

wobei Q_s orthogonal und R_s eine rechte Dreiecksmatrix ist.

Die QR-Zerlegung (2.16) beruht auf dem folgenden Hilfssatz:

Hilfssatz 2.4 *a) A sei eine $n \times n$ reelle Matrix. Es gibt eine orthogonale $n \times n$-Matrix Q und eine rechte $n \times n$-Dreiecksmatrix R, damit*

$$A = QR\ .$$

b) Sei A regulär. Dann sind Q und R bis auf die Multiplikation mit einer $n \times n$ orthogonalen Diagonalmatrix D eindeutig bestimmt. Sei also

$$A = Q_1 R_1 = Q_2 R_2$$

dann gibt es eine orthogonale Diagonalmatrix

$$D = diag(\pm 1)$$

mit

$$Q_1 = Q_2 D\ ,\ R_1 = D R_2\ .$$

Beweis:

a) Durch wiederholte Benutzung von Hilfssatz 2.4

$$A_0 := A$$

$$A_k = \begin{pmatrix} U_k & \vdots & c_k & \vdots & C_k \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \vdots & b_k & \vdots & B_k \end{pmatrix}$$

mit U_k eine $k \times k$ rechte Dreiecksmatrix, $c_k \in \mathbb{R}^k$, $b_k \in \mathbb{R}^{n-k}$, $C_k \in \mathbb{R}^{k\times(n-k-1)}$, $B_k \in \mathbb{R}^{(n-k)\times(n-k-1)}$. Nach Hilfssatz 2.4 gibt es eine $(n-k) \times (n-k)$ Householder Matrix H_k :

$$H_k b_k = -\sigma_k (1, 0, 0, \dots, 0)^T\ .$$

Sei

$$Q_k := \begin{pmatrix} I_k & 0 \\ 0 & H_k \end{pmatrix} ,$$

$$\begin{aligned} A_{k+1} &:= Q_k A_k \ , \\ Q &:= Q_0^T \dots Q_{n-1}^T \ , \\ R &:= U_n \ . \end{aligned}$$

Dann gilt:

$$A = QR \ .$$

b) R_1 und R_2 sind regulär. Deshalb folgt aus $Q_1 R_1 = Q_2 R_2$

$$Q_2^T Q_1 = R_1^{-1} R_2 \ .$$

$R_3 := R_1^{-1} R_2$ ist eine orthogonale obere Dreiecksmatrix.

Es folgt (z.B. induktiv), daß R_3 eine orthogonale Diagonalmatrix D ist. Da $D^T D = I$, gilt $D = \mathrm{diag}(\pm 1)$.

□

Hilfssatz 2.5 *Seien A_s, Q_s und R_s durch (2.15) bis (2.17) definiert. Sei*

$$\begin{aligned} P_s &:= Q_1 Q_2 \dots Q_s \ , && (2.18) \\ U_s &:= R_s \dots R_2 R_1 \ . && (2.19) \end{aligned}$$

Dann gilt:

a) A_{s+1} ist ähnlich zu A_s, $A_{s+1} = Q_s^T A_s Q_s$

b) $A_{s+1} = P_s^T A_1 P_s$

c) $A_1^s = P_s U_s$

Beweis: Einfache Manipulationen. □

Der nächste Hilfssatz wird ohne Beweis von Wilkinson (Wilkinson [1965, S. 518]) benutzt. Es kann wohl sein, daß ein besserer Beweis existiert.

Hilfssatz 2.6 *F_s seien $n \times n$ Matrizen mit $F_s \longrightarrow 0$ für $s \longrightarrow \infty$. Es sei weiter*

$$(I + F_s) = Q_s R_s$$

mit Q_s (orthogonal) und R_s (obere Dreiecksmatrix mit positiven Diagonalelementen). Dann gilt:

$$\left.\begin{aligned} Q_s &\longrightarrow I \\ R_s &\longrightarrow I \end{aligned}\right\} \quad \textit{für} \quad s \longrightarrow \infty$$

Beweis: Zuerst bemerken wir, daß

$$\|Q_s\|_2 = 1 \ ,$$

und

$$\|R_s\|_2 = \|Q_s^T \cdot (E + F_s)\|_2 \leq \|Q_s^T\|_2 \cdot (1 + \|F_s\|_2) \ ,$$

so daß o.E. angenommen werden kann, daß

$$\|R_s\|_2 \leq 2 \ .$$

Es gilt auch

$$(E + F_s)^{-1} = E + G_s$$

mit

$$\|G_s\| \longrightarrow 0 \quad \text{für} \quad s \longrightarrow \infty \ .$$

(Begründung: Sei

$$(E + F)^{-1} = E + G.$$

Es folgt

$$G = (E + F)^{-1} - E = (E + F)^{-1}(E - (E + F)) = -(E + F)^{-1} \cdot F \ .$$

Da

$$\|(E + F)^{-1}\| \leq \frac{1}{1 - \|F\|} \ ,$$

folgt:

$$\|G\| \leq \frac{\|F\|}{1 - \|F\|} \ .)$$

Sei nun

$$Q_s = (D_s + L_s + U_s) \ ,$$

wobei D_s, L_s und U_s Diagonal-, linke und rechte Matrizen sind.

Aus

$$Q_s^T = R_s \cdot (E + F_s)^{-1} = R_s \cdot (E + G_s) = R_s + R_s G_s$$

folgt

$$\|U_s^T\|_\infty \leq \|R_s G_s\|_\infty \leq 2\|G_s\|_\infty \ ,$$

so daß

$$U_s \longrightarrow 0 \quad \text{für} \quad s \longrightarrow \infty \ .$$

Es darf weiter o.E. angenommen werden, daß $D_s + L_s$ regulär ist und

$$\|(D_s + L_s)^{-1}\|_2 < 2 \ .$$

Für alle $x \in \mathbb{R}^n$ gilt nämlich

$$\|x\|_2 = \|Q_s x\|_2 = \|(D_s + L_s + U_s)x\|_2 \leq \|(D_s + L_s)x\|_2 + \|U_s\|_2 \|x\|_2 \ ,$$

so daß

$$\|(D_s + L_s)x\|_2 \geq (1 - \|U_s\|_2) \cdot \|x\|_2 \qquad .)$$

Sei nun

$$V_s := (D_s + L_s)^{-1} U_s \ ,$$

so daß

$$Q_s = (D_s + L_s) \cdot (I + V_s)$$

und $V_s \to 0$ für $s \to \infty$. Aus $Q_s^T Q_s = E$ folgt dann

$$(I + V_s^T) \cdot (D_s + L_s^T) \cdot (D_s + L_s) \cdot (I + V_s) = E \ ,$$

und

$$D_s + L_s = (D_s + L_s^T)^{-1} (I + V_s^T)^{-1} (I + V_s)^{-1} \ .$$

Folglich gilt

$$D_s + L_s = (D_s + L_s^T)^{-1} \cdot (I + W_s)$$

mit $W_s \to 0$ für $s \to \infty$.

Wenn die *unteren* Dreiecksmatrizen links und rechts verglichen werden, ergibt sich

$$\begin{aligned} \|L_s\|_\infty &\leq 0 + \|(D_s + L_s^T)^{-1} \cdot W_s\|_\infty \\ &\leq \|(D_s + L_s^T)^{-1}\|_\infty \cdot \|W_s\|_\infty \\ &\longrightarrow 0 \quad \text{für} \quad s \longrightarrow \infty \ . \end{aligned}$$

Da Q_s orthogonal ist und $Q_s = D_s + L_s + U_s$ mit $L_s \longrightarrow 0$ und $U_s \longrightarrow 0$, folgt sofort, daß

$$D_s^2 \longrightarrow E \quad \text{für} \quad s \longrightarrow \infty \ .$$

Da R_s positive Diagonalelemente hat, folgt aus

$$R_s = Q_s^T \cdot (E + F_s)$$

daß

$$R_s \longrightarrow E$$

und zuletzt

$$Q_s \longrightarrow E \ .$$

$\square$

Satz 2.9 *Die $n \times n$ reelle Matrix A habe betragsmäßig verschiedene Eigenwerte:*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| \ .$$

Der Faktor R_s von A_s habe positive Diagonalelemente. Dann gibt es eine Permutation P von den Eigenwerten

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = P \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$$

so daß

$$\begin{aligned} a_{ij}^{(s)} &\longrightarrow 0 \, , \; i > j \\ a_{ii}^{(s)} &\longrightarrow \mu_i \, . \end{aligned}$$

Beweis: (siehe Wilkinson, S. 516) $x_1, \ldots, x_n$ seien die Eigenvektoren von A und $X = (x_1, \ldots, x_n)$. Dann gilt

$$AX = XD \, , \; D = \text{diag}(\lambda_i). \tag{2.20}$$

$$A = XDY \, , \; Y = X^{-1}. \tag{2.21}$$

Sei P, L, U , so daß

$$PY = LU \, , \tag{2.22}$$

wobei P eine Permutationsmatrix ist (im allgemeinen können wir $P = E$ wählen), L eine linke Dreiecksmatrix mit Diagonalelementen 1 und U eine rechte Dreiecksmatrix.

Sei Q (orthogonal) und R (rechte Dreiecksmatrix mit positiven Diagonalelementen), so daß

$$XP^T = QR \, . \tag{2.23}$$

Aus (2.21) bis (2.23) folgt:

$$\begin{aligned} A^s \equiv A_1^s &= XD^sY \\ &= XP^T(PD^sP^T)PY \\ &= QR(PD^sP^T)LU \\ &= QR\tilde{D}^sLU \\ &= QR(\tilde{D}^sL\tilde{D}^{-s})\tilde{D}^sU \end{aligned}$$

mit $\tilde{D} = PDP^T$ eine Diagonalmatrix. Dann gilt:

$$\tilde{D}^sL\tilde{D}^{-s} = E + E_s \, , \quad \text{mit} \quad E_s \longrightarrow 0 \quad \text{für} \quad s \longrightarrow 0 \, .$$

(Wenn $P = E$, ist

$$(D^sLD^{-s})_{ij} = \ell_{ij}(\lambda_i/\lambda_j)^s \, .$$

I.a. sind weitere Überlegungen nötig - siehe Wilkinson, S. 519 .)

Es folgt:

$$\begin{aligned} A_1^s &= Q(E + RE_sR^{-1})R\tilde{D}^sU \\ &= Q(E + F_s)R\tilde{D}^sU \ , \quad \text{mit} \quad F_s \longrightarrow 0 \quad \text{für} \quad s \longrightarrow \infty \ . \end{aligned}$$

Sei $E + F_s = \tilde{Q}_s\tilde{R}_s$, wo $\tilde{R}_s$ positive Diagonalelemente hat. Dann gilt

$$A_1^s = (Q\tilde{Q}_s)(\tilde{R}_sR\tilde{D}^sU) \ .$$

Sei

$$\begin{aligned} \tilde{D} &= |\tilde{D}|D_1 \ , \ D_1^2 = E \\ U &= D_2(D_2^{-1}U) \ , \ D_2^2 = E \ , \end{aligned}$$

wobei $|\tilde{D}|$ und $(D_2^{-1}U)$ positive Diagonalelemente haben. Dann gilt (für Diagonalmatrizen AB gilt $AB = BA$)

$$A_1^s = Q\tilde{Q}_sD_2D_1^s[(D_2D_1^s)^{-1}\tilde{R}_sR(D_2D_1^s)|\tilde{D}|^s(D_2^{-1}U] \ .$$

Aus den Hilfssätzen 2.4 und 2.5 folgt:

$$P_s = Q\tilde{Q}_sD_2D_1^s. \tag{2.24}$$
$$U_s = (D_2D_1^s)^{-1}\tilde{R}_sR(D_2D_1^s)|\tilde{D}|^sD_2^{-1}U \ . \tag{2.25}$$

Es folgt aus (2.18), (2.24) und Hilfssatz 2.6:

$$\begin{aligned} Q_s &= P_{s-1}^{-1}P_s \\ &= D_1^{-s+1}D_2^{-1}\tilde{Q}_{s-1}^{-1}Q^{-1}Q\tilde{Q}_sD_2D_1^s \\ &= D_1 + D_1^{-s+1}D_2^{-1}(\tilde{Q}_{s-1}^{-1}\tilde{Q}_s - E)D_2D_1^s \\ &\longrightarrow D_1 \end{aligned}$$

Auf die gleiche Weise:

$$\begin{aligned} R_s &= U_sU_{s-1}^{-1} \\ &= [(D_2D_1^s)^{-1}\tilde{R}_sR(D_2D_1^s)|\tilde{D}|^sD_2^{-1}U] \cdot \\ &\cdot \quad [U^{-1}D_2]|\tilde{D}|^{-s+1}D_1^{-s+1}D_2^{-1}R^{-1}\tilde{R}_{s-1}^{-1}(D_2D_1^{s-1}) \ , \\ &= (D_2D_1^s)^{-1}\tilde{R}_sRD_1|\tilde{D}|R^{-1}\tilde{R}_{s-1}^{-1}(D_2D_1^{s-1}). \\ \text{diag}(R_s) &= \text{diag}(\tilde{R}_s)\text{diag}(\tilde{R}_{s-1}^{-1})|\tilde{D}| \\ &\longrightarrow |\tilde{D}| \ . \end{aligned}$$

$\square$

2.7.3 Nichtkonvergenz des QR-Verfahrens

Das QR-Verfahren konvergiert nicht immer, wenn die Voraussetzungen von Satz 2.9 nicht erfüllt sind.

Beispiel:

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

Dann gilt: $Q_1 = A_1$, $R_1 = E$, $A_2 = A_1$. Dieses Beispiel zeigt auch, daß - selbst wenn A reell ist - die QR-Folge $\{A_k\}$ nicht notwendig zur Jordan-Gestalt konvergiert.

2.7.4 Beschleunigungsstrategien

1. **Zerlegung von A_s** Falls $a^{(s)}_{r+1,r}$ sehr klein ist, z.B.

   $$\left|a^{(s)}_{r+1,r}\right| \leq \text{eps}\|A\|_\infty ,$$

   dann ist

   $$A_s \doteq \begin{pmatrix} B_s & \vdots & C_s \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D_s \end{pmatrix} ,$$

   wobei B_s eine $r \times r$ Matrix ist. Die Eigenwerte von B_s und D_s werden dann getrennt berechnet.

2. **Shift-Strategie**

   $$\begin{aligned} A_s - \kappa_s E &= Q_s R_s \\ A_{s+1} &= R_s Q_s + \kappa_s E . \end{aligned}$$

   Zwei Methoden werden benutzt, um κ_s zu berechnen:

   (a) $\kappa_s := a^{(s)}_{nn}$

   (b) Seien $\lambda_\pm = a_s \pm i b_s$ die Eigenwerte von

   $$\begin{pmatrix} a^{(s)}_{n-1,n-1} & a^{(s)}_{n-1,n} \\ a^{(s)}_{n,n-1} & a^{(s)}_{n,n} \end{pmatrix} ,$$

   dann ist

   $$\kappa_s := \begin{cases} \lambda_+ & , \text{ falls } \ |\lambda_+ - a^{(s)}_{nn}| \leq |\lambda_- - a^{(s)}_{nn}| \\ \lambda_- & , \text{ sonst.} \end{cases}$$

Kapitel 3

Numerische Integration

3.1 Einleitung

Sei $\Omega \subset \mathbb{R}^n$ und $f : \Omega \longrightarrow \mathbb{R}$. Die Aufgabe der numerischen Integration besteht darin, das Integral

$$I\, f := \int\limits_{\Omega} f(x)dx$$

durch $I_n f$ zu approximieren,

$$I_n f := \sum_{k=1}^{n} A_k f(x_k) \ ,$$

wobei die x_k *Stützstellen* und die Konstanten A_k *Gewichte* heißen.

In diesem Kapitel wird ein einfacher spezieller Fall

$$I\, f := \int\limits_{a}^{b} f(x)dx$$

ausführlich studiert.

Literatur

Davenport, J.H.: On the Integration of Algebraic Functions. Springer.

Davis, P.J., **Rabinowitz**, P.: Methods of Numerical Integration. New York:] Academic Press.

Engels, H.: Numerical Quadrature and Cubature. London, Academic Press, 1980.

Geddes, K.O., **Czapor**, S.R., **Labahn**, G.: Algorithms for Computer Algebra. Kluwer, 1992.

Krylov, V.I.: Approximate Calculation of Integrals. New York:] Macmillan, 1962.

Levin, M., **Girshovich**, J.: Optimal Quadrature Formulas. Teubner, 1979.

Piessens, R., **de Doncker-Kapenga**, E., **Überhuber**, C.W., **Kahaner**, D.K.: Quadpack: A Subroutine Package for Automatic Integration. Berlin, Springer, 1983.

Stroud, A.H.: Approximate Calculation of Multiple Integrals. Englewood Cliffs, N.J., Prentice Hall, 1971.

3.2 Die Formeln von Newton-Cotes

Seien $-\infty < a < b < \infty$, $f \in C[a,b]$ und

$$I\ f = \int_a^b f(x)dx \ .$$

Die Formeln von Newton-Cotes entstehen, wenn gleichverteilte Stützstellen x_k vorgeschrieben werden und $I\ f$ durch

$$\tilde{I}\ f := \sum_k A_k f(x_k)$$

approximiert wird. Dabei werden die Gewichte A_k so gewählt, daß gilt:

$$\tilde{I}\ p = I\ p \ , \ p \in P_{m-1} \ ,$$

wobei m die Anzahl der Gewichte A_k ist.

Die *offenen Newton-Cotes-Formeln* entstehen bei der Wahl:

$$x_k = a + k \cdot \frac{b-a}{n} \ , \ 1 \le k \le n-1 \ , \ m = n-1 \ ,$$

so daß die Endpunkte des Intervalls $[a,b]$ nicht Stützstellen sind.

Die *geschlossenen Newton-Cotes-Formeln* entstehen bei der Wahl:

$$\begin{aligned} x_k &= a + k \cdot \frac{b-a}{n} = a + kh \ , \ 0 \le k \le n \ , \ m = n+1 \ , \ h := (b-a)/n \ , \\ \tilde{I}\ f &:= I_n f = \sum_{k=o}^{n} A_k f(x_k) \ . \end{aligned}$$

Da die geschlossenen Newton-Cotes-Formelen mehrere Anwendungen finden, werden nur diese Formeln weiter betrachtet.

Eine Möglichkeit, die Gewichte A_k zu bestimmen, ist, die Funktion f durch ihr Interpolationspolynom p zu ersetzen, das an ihrer Stelle integriert wird. In der Form von Lagrange lautet p :

$$\begin{aligned} p(x) &= \sum_{k=o}^{n} f(x_k)\, \omega_k(x) \qquad \text{mit} \\ \omega_k(x) &= \prod_{\substack{\ell=0 \\ \ell\neq k}}^{n} \frac{x-x_\ell}{x_k-x_\ell} \end{aligned}$$

Es folgt:

$$\begin{aligned} I_n = \int\limits_a^b p(x)\, dx &= \sum_{k=0}^{n} \int\limits_a^b \omega_k(x)\, f(x_k)\, dx = \sum_{k=0}^{n} A_k f(x_k) \qquad \text{mit} \\ A_k = \int\limits_a^b \omega_k(x)\, dx &= \int\limits_a^b \prod_{\substack{\ell=0 \\ \ell\neq k}}^{n} \frac{x-x_\ell}{x_k-x_\ell}\, dx \end{aligned}$$

Es ist möglich, den Ausdruck für A_k zu vereinfachen. Man setze

$$x = h\,t + a$$

und erhält:

$$A_k = h \int\limits_0^n \prod_{\substack{\ell=0 \\ \ell\neq k}}^{n} \frac{t-\ell}{k-\ell}\, dt =: h\, a_k$$

Beispiel: $n = 1$ die *Trapezregel*:

$$\begin{aligned} a_0 &= \int\limits_0^1 \frac{t-1}{-1}\, dt = \frac{1}{2} \\ a_1 &= \int\limits_0^1 t\, dt = \frac{1}{2} \\ I_1 &= \frac{h}{2}\,(f(a)+f(b)) = \frac{b-a}{2}\,(f(a)+f(b)) \end{aligned}$$

Das Integral wird also durch die Fläche eines Trapezes genähert (siehe Abbildung 3.1):

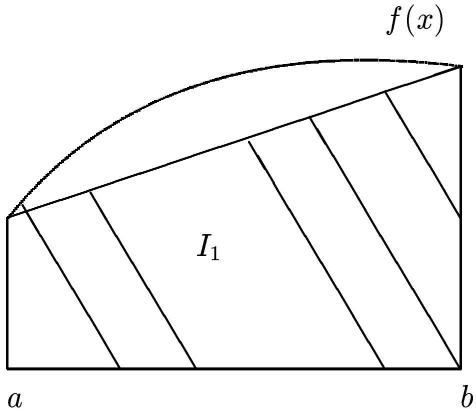


Abbildung 3.1: Graphische Darstellung der Trapezregel

Beispiel: $n = 2$ die Simpsonsche Regel:

$$\begin{aligned} a_0 &= \frac{1}{3} \qquad a_1 = \frac{4}{3} \qquad a_2 = \frac{1}{3} \\ I_2 &= \frac{h}{3}\left(f(a) + 4f\left(\frac{b-a}{2}\right) + f(b)\right) \end{aligned}$$

Die Simpsonsche Regel ist einfach und relativ genau und wird als zusammengesetzte Regel häufig angewandt.

Wir untersuchen jetzt den Fehler.

Satz 3.1 *1. Sei $f \in C^{n+1}[a,b]$. Dann gilt:*

$$\begin{aligned} |I\,f - I_n f| &\leq h^{n+2} c_n \max_{[a,b]} |f^{(n+1)}(x)| \qquad \textit{mit} \\ c_n &\leq \frac{1}{(n+1)!} \int_0^n \prod_{k=0}^n |t-k|\; dt\,. \end{aligned}$$

2. Sei n gerade und $f \in C^{n+2}[a,b]$. Dann folgt

$$\begin{aligned} |I\,f - I_n f| &\leq h^{n+3} c_n^* \max_{[a,b]} |f^{(n+2)}(x)| \qquad \textit{mit} \\ c_n^* &\leq \frac{n}{2}\, c_n \end{aligned}$$

Beweis:

1. Es gilt:

$$I\ f - I_n f = \int\limits_a^b (f-p)(x)\ dx\ .$$

Ferner hat man die folgende Abschätzung für den Interpolationsfehler:

$$\begin{aligned} (f-p)(x) &= \frac{w(x)}{(n+1)!}\, f^{(n+1)}(\xi) \qquad \text{mit} \\ w(x) &= \prod_{k=0}^{n} (x-x_k) \qquad \text{und} \qquad \xi = \xi(x) \in (a,b)\ . \end{aligned}$$

Es folgt sofort:

$$|I\ f - I_n f| \le \frac{1}{(n+1)!} \int\limits_a^b |w(x)|\ \max_{[a,b]} |f^{(n+1)}(x)|\ dx\ .$$

c_n ergibt sich aus der Berechnung von $\int\limits_a^b |w(x)|\ dx$:

$$\int\limits_a^b |w(x)|\ dx = \int\limits_a^b \prod_{k=0}^{n} |x-x_k|\ dx\ , = h^{n+2} \int\limits_0^n \prod_{k=0}^{n} |t-k|\ dt$$

nach der Substitution $x = a + h\ t$.

2. Für gerades n ist w schiefsymmetrisch bezüglich der Intervallmitte $c = \frac{a+b}{2}$, es gilt also

$$\int\limits_a^b w(x)\ dx = 0\ .$$

Damit hat man

$$\begin{aligned} \int\limits_a^b (f-p)(x)\ dx &= \frac{1}{(n+1)!} \int\limits_a^b w(x) f^{(n+1)}(\xi)\ dx \\ &= \frac{1}{(n+1)!} \int\limits_a^b w(x) \left\{ f^{(n+1)}(c) + (\xi - c) f^{(n+2)}(\eta) \right\}\ dx \\ &= \frac{1}{(n+1)!} \int\limits_a^b w(x)(\xi - c) f^{(n+2)}(\eta)\ dx \end{aligned}$$

Wegen $|\xi - c| \leq \frac{b-a}{2} = \frac{nh}{2}$ gilt:

$$
\begin{aligned}
|\int_a^b (f-p)(x)\,dx| &\leq \frac{1}{(n+1)!}\int_a^b |w(x)|\,\frac{nh}{2}\,\max_{[a,b]}\,|f^{(n+2)}(x)|\,dx \\
&= h^{n+2}c_n\frac{nh}{2}\,\max_{[a,b]}\,|f^{(n+2)}(x)| \\
&= h^{n+3}c_n^*\,\max_{[a,b]}\,|f^{(n+2)}(x)|
\end{aligned}
$$

□ Da das Maximum hoher Ableitungen von f sehr schwer zu bestimmen ist, sind diese Formeln zur praktischen Abschätzung des Fehlers unbrauchbar. Ihr Nutzen liegt in der Information, mit welcher Potenz von h der Fehler abfällt, und daß er vom Maximum einer höheren Ableitung abhängt.

In dem Beweis von Satz 3.1 sind einige Abschätzungen gemacht worden. Die Konstanten c_n sind deshalb nicht bestmöglich und können (mit wesentlich mehr Aufwand) verbessert werden. Es gilt:

$$
\begin{aligned}
I_n f &= \frac{h}{\gamma_n}\sum_{k=0}^{n}\sigma_k^{(n)}f(x_k)\ , \\
I\,f - I_n f &= h^{p+1}K_n f^{(p)}(\xi)\ .
\end{aligned}
$$

Die Werte von γ_k , $\sigma_k^{(n)}$, p und K_n sind für $1 \leq n \leq 4$ der folgenden Tabelle zu entnehmen: (siehe Tabelle 3.1)

n	σ_0	σ_1	σ_2	σ_3	σ_4	γ_n	Fehler $I - I_n^{NC}$	Bezeichnung
1	1	1				2	$-h^3\frac{1}{12}f^{(2)}(\xi)$	Trapezregel
2	1	4	1			6	$-h^5\frac{1}{90}f^{(4)}(\xi)$	Simpson-Regel
3	1	3	3	1		8	$-h^5\frac{3}{80}f^{(4)}(\xi)$	Newtonsche $\frac{3}{8}$ - Regel
4	7	32	12	32	7	90	$-h^7\frac{8}{945}f^{(6)}(\xi)$	Milne-Regel
$2r$							$h^{n+3}K_n f^{(n+2)}(\xi)$	n gerade
$2r+1$							$h^{n+2}K_n f^{(n+1)}(\xi)$	n ungerade

Tabelle 3.1: Die abgeschlossenen Newton-Cotes-Formeln für $1 \leq n \leq 4$

Es gibt andere historische Integrationsformeln, z.B. *Weddles Rule*:

$$
\begin{aligned}
\int_{-1}^{+1} f(x)dx &= \frac{1}{10}\left[f(-1) + 5f\left(-\frac{2}{3}\right) + f\left(-\frac{1}{3}\right) + 6f(0) + f\left(\frac{1}{3}\right)\right. \\
&+ \left. 5f\left(\frac{2}{3}\right) + f(1)\right] + \frac{1}{306180}f^{(6)}(0) + \ldots\ ,
\end{aligned}
$$

die wegen der kleinen Koeffizienten beliebt war.

Es ist möglich, die Fehlerabschätzung etwas präziser zu formulieren, und zwar in der Form

$$I\,f - I_n f = \int_a^b K_n(t) f^{(n)}(t)dt$$

wobei $K_n(t)$ der Peano-Kern heißt. Z.B., für die Simpsonsche Methode gilt:

$$I\,f - I_2^{NC} f = \int_a^b K_2(t) f^{(4)}(t)dt$$

mit

$$K_2(t) := \begin{cases} \frac{1}{72}(t-a)^3(3t-a-b), & a \le t \le \frac{a+b}{2} \\ \frac{1}{72}(b-t)^3(b+2a-3t), & \frac{a+b}{2} \le t \le b \end{cases}$$

(Siehe Abbildung 3.2.)

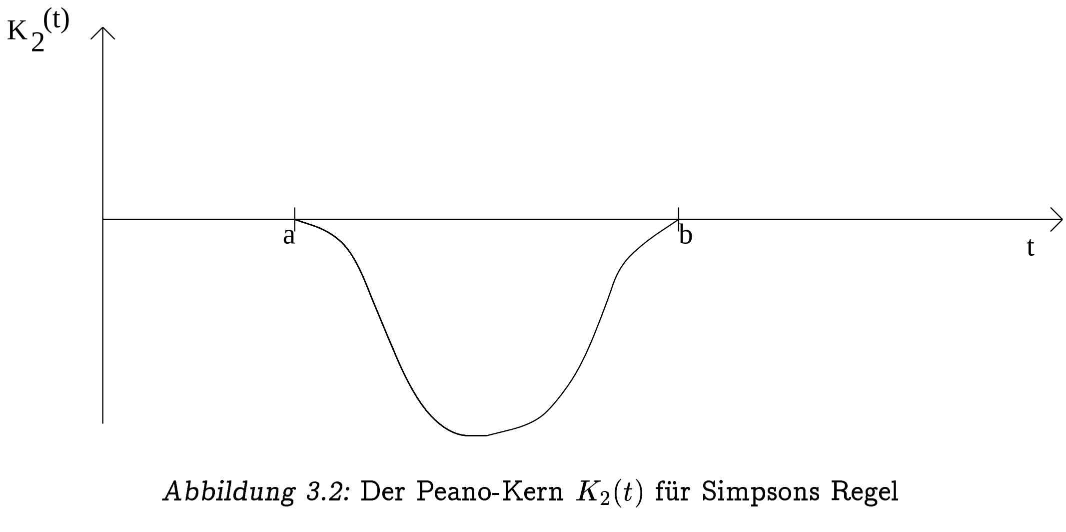


Abbildung 3.2: Der Peano-Kern $K_2(t)$ für Simpsons Regel

Die Fehlerabschätzungen spielen eine wichtige Rolle bei zusammengesetzten Regeln und bei der Lösung von Integralgleichungen, z.B.

$$f(s) + \int_a^b k(s,t) f(t)dt = g(s)\ , \quad a \le s \le b.$$

Beispiel:

$$\begin{aligned} I = \int_0^1 e^x\,dx = e-1 &= 1.7183 \\ I_1 = \frac{1}{2}\,(1+e) &= 1.8591 \\ I_2 = \frac{1}{6}\,(1+4e^{\frac{1}{2}}+e) &= 1.7189 \\ I_3 = \frac{1}{8}\,(1+3e^{\frac{1}{3}}+3e^{\frac{2}{3}}+e) &= 1.7185 \\ I_4 = \frac{1}{90}\,(7+32e^{\frac{1}{4}}+12e^{\frac{2}{4}}+32e^{\frac{3}{4}}+7e) &= 1.7183 \end{aligned}$$

Für $n \geq 8$ können negative Gewichte auftreten, was aus Rundungsfehlergründen nicht gut ist. Wie wir später sehen werden, kann man Formeln höherer Genauigkeit konstruieren, indem man die oben angegebenen Regeln auf Teilintervalle anwendet.

Beispiel: Sei

$$I\ f := \int_{-4}^{+4} \frac{dx}{1+x^2} = 2\arctan 4 \doteq 2,6516.$$

In der Tabelle 3.2 werden die Approximationen I_n^{NC} eingetragen. Es ist klar, daß die Folge $\{I_n^{NC}\}$ divergent ist.

n	I_n^{NC}
2	$5,4902$
4	$2,2776$
6	$3,3288$
8	$1,9411$
10	$3,5956$

Tabelle 3.2: Ein Beispiel dafür, daß die Newton-Cotes Formeln divergieren können

Beispiel: Eine Anwendung der numerischen Integration ist die Lösung von Integralgleichungen.

Man betrachte die Fredholmsche Integralgleichung zweiter Art:

$$f(s) + \int_a^b k(s,t)f(t)dt = g(s)\ ,\quad a \leq s \leq b\ ,$$

wobei die Funktion g und der Kern $k(s,t)$ bekannt sind und die Funktion f zu finden ist.

Als Lösungsansatz kann man die Methode von Nyström benutzen. Seien $x_0, \ldots, x_n$ die Stützstellen zu der Newton-Cotes Integrationsformel I_n^{NC}. Die Integralgleichung wird approximiert durch ein System von $n+1$ linearen Gleichungen für die Werte von f in den $n+1$ Stützstellen:

$$f(x_i) + \sum_{j=0}^{n} A_j k(x_i, x_j) f(x_j) = g(x_i) \ , \ 0 \le i \le n \ ,$$

wobei

$$I_n^{NC} f = \sum_{j=0}^{n} A_j f(x_j).$$

3.3 Die direkte Konstruktion von Integrationsformeln

Im vorigen Abschnitt wurden die Newton-Cotes-Integrationsformeln mit Hilfe des Interpolationspolynoms konstruiert. Im nächsten Abschnitt werden die Gaußschen Integrationsformeln ähnlich definiert. Es ist aber möglich, solche Integrationsformeln ab initio zu konstruieren, wie wir jetzt durch ein Beispiel erklären.

Beispiel:

$$\begin{aligned} I\,f &:= \int_{-1}^{+1} f(x)\,dx \ , \\ G\,f &:= \sum_{k=1}^{2} A_k f(x_k) \ . \end{aligned}$$

Bestimme A_1, A_2, x_1 und x_2 derart, daß

$$G\,f = I\,f \ , \qquad \text{für} \quad f \in P_3 \ . \tag{3.1}$$

Die Bedingung (3.1) ist dazu äquivalent, daß

$$\begin{aligned} G\,q_j &= I\,q_j \ , \qquad 0 \le j \le 3 \\ q_j &\in C[-1,+1] \ , \\ q_j(x) &:= x^j \end{aligned}$$

Es gilt also

$$\begin{array}{lrcl} I & A_1 + A_2 & = & 2 \\ II & A_1 x_1 + A_2 x_2 & = & 0 \\ III & A_1 x_1^2 + A_2 x_2^2 & = & 2/3 \\ IV & A_1 x_1^3 + A_2 x_2^3 & = & 0 \end{array} \tag{3.2}$$

Sei

$$p(x) := (x - x_1)(x - x_2) \ ,$$

so daß es Konstanten α und β gibt mit

$$p(x) = x^2 + \alpha\, x + \beta \ .$$

Durch gewichtete Kombinationen der Gleichungen (3.2) erhalten wir:

$$\begin{array}{rcl} I + \alpha\, II + \beta\, III & : & 2 + 2/3\beta = A_1 p(x_1) + A_2 p(x_2) = 0 \\ II + \alpha\, III + \beta\, IV & : & 2/3\alpha = A_1 x_1 p(x_1) + A_2 x_2 p(x_2) = 0 \end{array}$$

woraus folgt:

$$\begin{array}{rcl} \alpha & = & 0 \ , \\ \beta & = & -1/3 \ , \\ p(x) & = & x^2 - 1/3 \ , \\ x_1 & = & -1/\sqrt{3} \ , \\ x_2 & = & +1/\sqrt{3} \ . \end{array}$$

Die Gleichungen I und II werden:

$$\begin{array}{rcl} A_1 + A_2 & = & 2 \\ A_1 \cdot \left(\dfrac{-1}{\sqrt{3}}\right) + A_2 \cdot \left(\dfrac{+1}{\sqrt{3}}\right) & = & 0 \ , \end{array}$$

woraus folgt:

$$A_1 = A_2 = 1 \ .$$

Die gewünschte Integrationsformel lautet:

$$G\ f = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{+1}{\sqrt{3}}\right) \ .$$

3.4 Die Formeln von Gauß

Sei
$$-\infty \leq a < b \leq +\infty \ .$$

Sei $w \in C[a,b]$ mit:

1. $w(x) \geq 0 \ , \qquad a \leq x \leq b \ .$
2. $w(x) > 0$ für fast alle $x \in (a,b)$, so daß aus
   (a) $f \in C[a,b]$
   (b) $f(x) \geq 0$
   (c) $\int\limits_a^b w(x)f(x)dx = 0$ folgt: $f \equiv 0$.
3. $\int\limits_a^b w(x)x^n dx < \infty$, für $n \in \mathbb{N}$.

Wir setzen
$$(f,g) := \int\limits_a^b w(x)f(x)g(x)dx$$

vorausgesetzt, daß das Integral existiert. Ist $(f,g) = 0$, dann sagen wir, daß f und g orthogonal sind und schreiben $f \perp g$.

Seien
$$a \leq x_1 < x_2 \cdots < x_n \leq b \ .$$

Eine Gaußsche Integrationsformel G_n hat die Gestalt:
$$G_n f = \sum_{k=1}^{n} A_k f(x_k)$$

(notabene: n Stützstellen) und erfüllt die Bedingungen:
$$G_n p = Ip := \int\limits_a^b w(x)p(x)dx \ , \qquad \text{für alle} \qquad p \in P_{2n-1} \ .$$

Damit ist eine Gaußsche Integrationsformel wesentlich genauer als eine Newton-Cotes-Integrationsformel mit der gleichen Anzahl Stützstellen, was sie der geeigneten Wahl der Stützstellen x_k zu verdanken hat.

Der folgende Satz zeigt, daß die Gaußsche Integrationsformel optimal ist:

Satz 3.2 *Es gibt keine Formel G_n , die in $\mathcal{P}_{2n}$ exakt ist.*

Beweis: Die Annahme, daß

$$G_n f = \int_a^b w\ f\ dx \qquad \forall\ f \in \mathcal{P}_{2n}$$

führt mit

$$f := \prod_{j=1}^{n} (x - x_j)^2 \in \mathcal{P}_{2n}$$

zu einem Widerspruch:

$$G_n f = 0 \neq I\ f = \int_a^b w\ f\ dx > 0\ .$$

□

Um G_n zu konstruieren, wird eine Basis $\{p_k\}$ von orthogonalen Polynomen mit folgenden Eigenschaften benutzt:

1. $p_k \in P_k \qquad k \geq 0$.
2. $p_k(x) = x^k + \sum_{j=0}^{k-1} a_j^{(k)} x^j$.
3. $p_n \perp p_m$, d.h. $(p_n, p_m) = 0$ für $n \neq m$.
4. $p_{-1}(x) \equiv 0$.
5. p_n hat n unterschiedliche reelle Nullstellen $a \leq x_1 < x_2 < \cdots < x_n \leq b$.

Es wird später bewiesen, daß eine solche Basis tatsächlich existiert. Wir geben zuerst ein Beispiel.

Beispiel: Die Tschebyscheff Polynome erster Art Diese Polynome werden wie folgt definiert:

$$T_n(x) = \cos(n\ \arccos\ x)\ ,\ -1 \leq x \leq 1.$$

Sie haben folgende Eigenschaften:

Orthogonalität

$$\int_{-1}^{+1} \frac{T_n(x) T_m(x)}{\sqrt{1 - x^2}}\, dx = \begin{cases} 0 & \text{für} \quad m \neq n \\ \pi/2 & \text{für} \quad m = n > 0 \\ \pi & \text{für} \quad m = n = 0 \end{cases}$$

Rekursion

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \ , \ n \geq 1$$

Nullstellen

$$T_n(x_k) = 0 \quad \text{für} \quad x_k = \cos\left(\frac{2k+1}{2n}\,\pi\right) \ , \ 0 \leq k \leq n-1$$

Extremstellen

$$\begin{aligned} |T_n(x)| \leq 1 \quad &\text{für} \quad -1 \leq x \leq 1 \\ T_n(x_i) = (-1)^i \quad &\text{für} \quad x_i = \cos\left(\frac{\pi\, i}{n}\right) \ , \ 0 \leq i \leq n \ . \end{aligned}$$

Formel von Rodrigues

$$T_n(x) = \frac{(-1)^n}{2^n n!}\sqrt{1-x^2}\frac{d^n}{dx^n}\left\{\frac{(1-x^2)^n}{\sqrt{1-x^2}}\right\}$$

Reihenentwicklungen

Satz 3.3 *Sei* $f \in C[-1,+1]$ *Lipschitz stetig und*

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int\limits_{-1}^{+1} \frac{f(x)T_0(x)}{\sqrt{1-x^2}}\,dx \\ a_n &= \frac{2}{\pi}\int\limits_{-1}^{+1} \frac{f(x)T_n(x)}{\sqrt{1-x^2}}\,dx \ , \ n \geq 1 \end{aligned}$$

Dann konvergiert die Folge

$$\sum_{n=0}^{\infty} a_n T_n(x)$$

gleichmäßig gegen die Funktion $f(x)$ *im Intervall* $[-1,+1]$.

Beweis: Siehe z.B. Rivlin [1974, S. 135]. □

Um eine gute Annäherung für die Entwicklung von f in einer Folge von Tschebyscheff Polynomen zu erhalten, kann man folgendermaßen vorgehen:

1. Bestimme eine konvergente Taylorentwicklung von f,
$$f(x) = \sum_{k=0}^{\infty} b_k x^k \ , \ -1 \le x \le 1.$$
2. Wähle $N \in \mathbb{N}$ und setze
$$S_N(x) = \sum_{k=0}^{N} b_k x^k.$$
3. Ersetze die Potenzen von x durch Tschebyscheff Polynome.

Beispiel: $f(x) = e^x$

1. $$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$
2. $$S_4(x) := \sum_{k=0}^{4} \frac{x^k}{k!}$$
3. Aus
$$\begin{aligned} 1 &= T_0 \\ x &= T_1 \\ x^2 &= \frac{1}{2}\,(T_0 + T_2) \\ x^3 &= \frac{1}{4}\,(3T_1 + T_3) \\ x^4 &= \frac{1}{8}(3T_0 + 4T_2 + T_4) \end{aligned}$$
ergibt sich
$$S_4(x) = 1,2656T_0 + 1,1250T_1 + 0,2708T_2 + 0,0417T_3 + 0,0052T_4$$

Satz 3.4 *Seien $x_1, \dots, x_n$ die Nullstellen von p_n. Seien A_k die Gewichte der Newton-Cotes-Formel zu den Stützstellen $x_1, \dots, x_n$ und der Funktion $\omega(x)$:*
$$A_k := \int_a^b \omega(x) \prod_{\substack{j=1 \\ j \ne k}}^{n} \frac{(x - x_j)}{(x_k - x_j)} dx.$$
Sei
$$G_n f := \sum_{k=1}^{n} A_k f(x_k) \ .$$
Dann gilt:

1.

$$G_n f = I\ f\ , \qquad f \in P_{2n-1}$$

2.

$$A_k = \int_a^b w(x) \prod_{\substack{j=1 \\ j \neq k}}^{n} \left(\frac{x - x_j}{x_k - x_j} \right)^2 dx$$

3. Sei $f \in C^{(2n)}[a, b]$. Dann gibt es $\xi \in (a, b)$ mit:

$$I\ f - G_n f = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b w(x)(p_n(x))^2 dx\ .$$

Beweis:

1. G_n ist die Newton-Cotes-Formel zu den Nullstellen $x_1, \ldots, x_n$ von p_n , also

$$G_n(f) = \sum_{j=1}^{n} \int_a^b w(x) \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} f(x_j) dx\ .$$

G_n ist offenbar exakt in $\mathcal{P}_{n-1}$, da ja gerade das Interpolationspolynom vom Grad $n-1$ integriert wird. Ist $f \in \mathcal{P}_{2n-1}$, so gilt $f = q\ p_n + r$ mit $q, r \in \mathcal{P}_{n-1}$. Es folgt

$$\begin{aligned} I\ f = \int_a^b w\ f\ dx &= \underbrace{\int_a^b w\ q\ p_n dx}_{=0\ ,\quad \text{da} \quad q \in \mathcal{P}_{n-1}} + \int_a^b w\ r\ dx\ , \\ &= G_n(r) \quad , \quad \text{da} \quad r \in \mathcal{P}_{n-1} \\ &= G_n(r) + \underbrace{G_n(q\ p_n)}_{=0\ , \quad \text{da} \quad p_n(x_j)=0\ ,\ j=1,\ldots,n} \\ &= G_n(r + q\ p_n) \\ &= G_n(f) \end{aligned}$$

2. Mit

$$w_j(x) := \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}$$

ist $w_j^2 \in \mathcal{P}_{2n-2}$, also

$$\int_a^b w\ w_j^2\ dx = G_n(w_j^2) = \sum_{k=1}^{n} A_k w_j^2(x_k) = \sum_{k=1}^{n} A_k \delta_{jk} = A_j$$

3. Ist $f \in C^{(2n)}[a,b]$, können wir f in den Stützpunkten $x_1, x_1, \ldots, x_n, x_n$ interpolieren und erhalten bei Anwendung von dividierten Differenzen

$$\begin{aligned} f(x) &= [f(x_1) + (x-x_1)[x_1,x_1]f + (x-x_1)^2[x_1,x_1,x_2]f + \cdots \\ &+ \left(\prod_{j=1}^{n-1}(x-x_j)^2\right)(x-x_n)[x_1,x_1,\ldots,x_{n-1},x_{n-1},x_n,x_n]f \\ &+ \left(\prod_{j=1}^{n}(x-x_j)^2\right)[x_1,x_1,\ldots,x_n,x_n,x]f \\ &= q(x) + r(x)p_n(x)^2\,, \quad \text{mit} \quad q \in P_{2n-1} \end{aligned}$$

Es folgt:

$$I\,f - G_n f = (I\,q - G_n q) + (I\,rp_n^2 - G_n rp_n^2) = 0 + (I\,rp_n^2 - 0) = I\,rp_n^2.$$

Aus dem Mittelwertsatz der Integralrechnung folgt:

$$I\,f - G_n f = r(\xi)\int_a^b w(x)(p_n(x))^2 dx = \frac{f^{(2n)}(\eta)}{(2n)!}\,(p_n,p_n)\,,$$

wobei im letzten Schritt Hilfssatz 3.1 benutzt worden ist.

□

Hilfssatz 3.1 *Sei* $x_0, \ldots, x \in \mathbb{R}$. *Sei*

$$\begin{aligned} \alpha &= \min\{x_0,\ldots,x_m\} \\ \beta &= \max\{x_0,\ldots,x_m\} \end{aligned}$$

$f \in C^m[\alpha,\beta]$. *Es gibt* $\xi \in [\alpha,\beta]$ *mit*

$$[x_0,x_1,\ldots,x_m]f = \frac{f^m(\xi)}{m!}.$$

Beweis: Siehe (z.B.) E. Isaacson und H.B. Keller: Analysis of Numerical Methods, S. 252. □

Beispiel: $[a,b] = [-1,+1]$, $\quad w = 1$

Die x_j sind die Nullstellen der Legendre Polynome P_n.

n	x_1	x_2	x_3	A_1	A_2	A_3
1	0			2		
2	$-\sqrt{\frac{1}{3}}$	$+\sqrt{\frac{1}{3}}$		1	1	
3	$-\sqrt{\frac{3}{5}}$	0	$\sqrt{\frac{3}{5}}$	$\frac{5}{9}$	$\frac{8}{9}$	$\frac{5}{9}$

$$I = \int_{-1}^{1} e^x dx = 2.350402$$

Die Simpson-Regel liefert: $I_2 = 2.362054$. Dagegen ist mit gleich vielen Funktionswertungen $G_3 = 2.350337$. Es folgt aus der Nichtnegativität der Gewichte A_k für die Formel von Gauß, daß

$$\lim_{n\to\infty} G_n f = If \quad \text{für} \quad f \in C[a,b].$$

Um dies zu beweisen, benutzt man den folgenden Satz:

Satz 3.5 (Stekloff und Polya) *Sei*

$$\begin{aligned} If &= \int_a^b f(x)\,dx \\ I_n f &= \sum_{k=1}^{n} A_k^{(n)} f(x_k^{(n)}) \;. \end{aligned}$$

Es gilt:

$$\lim_{n\to\infty} I_n f = If \quad \textit{für alle} \quad f \in C[a,b]$$

genau dann, wenn

1. *$I_n p \to Ip$ für jedes Polynom p .*
2. *Eine Konstante K existiert derart, daß*

$$\sum_{k=1}^{n} |A_k^{(n)}| \le K \quad \textit{für} \quad n \in \mathbb{N}.$$

Beweis: Siehe (z.B.) I.P. Natanson: Konstruktive Funktionentheorie. S. 451. □

3.5 Existenz von Orthogonalpolynomen

Hierfür benutzen wir das Schmidtsche Verfahren zur Konstruktion einer orthogonalen Basis, ausgehend von einer beliebigen Basis.

Seien

$$q_n \in \mathcal{P}_n \,, \qquad q_n(x) := x^n \,, \qquad n \in \mathbb{N}_0 \;.$$

Dann ist $\{q_0, \dots, q_n\}$ eine Basis für $\mathcal{P}_n$. Seien

$$\begin{aligned} p_0 &:= q_0 \,, \\ p_n &:= q_n - \sum_{i=0}^{n-1} \frac{(q_n, p_i)}{(p_i, p_i)}\, p_i \,, \quad n \in \mathbb{N} \;. \end{aligned}$$

Satz 3.6 *Seien $\{p_n\}$ wie oben definiert. Dann gilt:*

1. *p_n ist wohldefiniert, $p_n \in \mathcal{P}_n$, und*
$$p_n(x) = x^n + r_{n-1} \ , \quad r_{n-1} \in \mathcal{P}_{n-1} \ .$$

2. *$p_n \perp p_s$ für $s < n$, d.h.*
$$(p_n, p_s) = 0 \ , \qquad \textit{für} \quad s < n \ .$$

3. *$(p_n, p_n) > 0$.*

4. *$\{p_0, p_1, \ldots, p_n\}$ ist eine Basis für $\mathcal{P}_n$.*

5. *$p_n \perp \mathcal{P}_{n-1}$, d.h.*
$$(p_n, p) = 0 \quad \textit{für alle} \quad p \in \mathcal{P}_{n-1} \ .$$

6. *Sei $n \geq 1$. Die Polynome p_n genügen der Rekursionsformel*
$$p_n(x) = (x - \delta_n) p_{n-1}(x) - \gamma_n^2 p_{n-2}(x) \ ,$$
wobei
$$p_{-1}(x) :\equiv 0 \ , \qquad \mathcal{P}_{-1} = \{p_{-1}\} \ , \qquad \delta_n := \frac{(x p_{n-1}, p_{n-1})}{(p_{n-1}, p_{n-1})} \ , \qquad n \in \mathbb{N}$$
$$\gamma_n^2 := \begin{cases} 0 & , \ \textit{für} \quad n = 1 \ , \\ \frac{(p_{n-1}, p_{n-1})}{(p_{n-2}, p_{n-2})} & , \ \textit{für} \quad n > 1 \end{cases}$$

Beweis: Der Beweis erfolgt durch Induktion über die Behauptung:

H_r : Die Bedingungen 1) bis 6) gelten für $0 \leq n \leq r$.

Die Behauptung H_0 ist offensichtlich wahr.
Sei H_{n-1} wahr. Da $(p_i, p_i) > 0$ für $0 \leq i < n$, ist p_n definiert und 1) und 2) folgen sofort.
Da $p_n(x) = x^n + r_{n-1}$, ist $p_n \not\equiv 0$ und 3) folgt.

Sei $q \in \mathcal{P}_n$, $q = \alpha x^n + q_{n-1}$ mit $q_{n-1} \in \mathcal{P}_{n-1}$. Dann ist
$$q = \alpha p_n + w_{n-1} \ , \qquad w_{n-1} \in \mathcal{P}_{n-1} \ .$$

4) und 5) folgen sofort.

Um 6) zu beweisen, muß man zwischen dem Fall $n = 1$ und dem Fall $n > 1$ unterscheiden.
Ist $n = 1$, dann kann man 6) durch direkte Berechnung bestätigen.
Ist $n > 1$, dann setzen wir
$$q(x) := x \ p_{n-1}(x) = x^n + s_{n-1}, s_{n-1} \in \mathcal{P}_{n-1} \ .$$

Aus 4) folgt:

$$q = p_n + \sum_{i=0}^{n-1} c_i p_i \ , \qquad c_i \in \mathbb{R} \ .$$

Sei $0 \le r < n-2$. Dann gilt:

$$(q, p_r) = (p_n + \sum_{i=0}^{n-1} c_i p_i, p_r) = c_r(p_r, p_r) \ .$$

Andererseits gilt:

$$(q, p_r) = (p_{n-1}, x p_r) = 0 \ ,$$

da $p_{n-1} \perp \mathcal{P}_{n-2}$, d.h. $c_r = 0$ für $0 \le r < n-2$.

Ebenso folgt:

$$\delta_n = c_{n-1} = \frac{(x\, p_{n-1}, p_{n-1})}{(p_{n-1}, p_{n-1})}$$

und

$$\gamma_n^2 = c_{n-2} = \frac{(x\, p_{n-2}, p_{n-1})}{(p_{n-2}, p_{n-2})} = \frac{(p_{n-1}, p_{n-1})}{(p_{n-2}, p_{n-2})}$$

da

$$x\, p_{n-2} = p_{n-1} + \delta_{n-1} p_{n-1} + \gamma_{n-1}^2 p_{n-2}$$

□

Satz 3.7 *Sei $\{p_n\}$ eine Basis von Orthogonalpolynomen. Dann gilt:*

1. *p_n hat n unterschiedliche Nullstellen in (a, b) .*
2. *Die Nullstellen von p_{n-1} trennen die Nullstellen von p_n .*

Beweis:

1. Wenn sich das Vorzeichen von p_n in (a, b) nicht ändert, setzen wir $r = 0$ und $q(x) \equiv 1$. Sonst seien $x_1 < x_2 < \cdots < x_r$ jene Punkte in (a, b) , wo das Vorzeichen von $p_n(x)$ sich ändert. ($x_1, \ldots, x_r$ sind die Nullstellen von p_n im Intervall (a, b) , deren Vielfachheit ungerade ist.) Sei

$$q := \prod_{i=1}^{r} (x - x_i) \ .$$

Ist $r < n$, dann ist $q \in \mathcal{P}_{n-1}$ und $q \perp p_n$. Aber das Vorzeichen von $q\, p_n$ ändert sich in (a, b) nicht, so daß $(q, p_n) > 0$. Es folgt, daß $r = n$.

2. Der Beweis erfolgt durch Induktion.
   Sei $n = 2$, dann gilt
   $$p_2(x) = (x - \delta_2)p_1(x) - \gamma_2^2 p_0(x) .$$
   Sei $x_1 \in (a, b)$ die Nullstelle von p_1. Da $p_0(x) = 1$, folgt, daß
   $$p_2(x_1) = -\gamma_2^2 p_0(x_1) < 0 .$$
   Da $p_2(x) = x^2 + r_1(x)$ mit $r_1 \in \mathcal{P}_1$, ist $p_2(x)$ positiv für $|x|$ groß. Deshalb hat p_2 mindestens eine Nullstelle $y_1 \in (-\infty, x_1)$ und eine Nullstelle $y_2 \in (x_1, +\infty)$. Es gilt:
   $$a < y_1 < x_1 < y_2 < b .$$
   Seien $n > 2$,
   $$x_1 < \cdots < x_{n-2}$$
   die Nullstellen von p_{n-2} und
   $$y_1 < \cdots < y_{n-1}$$
   die Nullstellen von p_{n-1}. Wegen der Induktionshypothese gilt:
   $$a < y_1 < x_1 < y_2 < \cdots < x_{n-2} < y_{n-1} < b .$$
   Weiter gilt:
   $$p_n(y_i) = -\gamma_n^2 p_{n-2}(y_i) .$$
   Zwischen den beiden nacheinanderfolgenden Nullstellen y_i liegt genau eine Nullstelle von p_{n-2} , so daß
   $$\text{sign } (p_{n-2}(y_i)) = (-1)^{i-1} \text{ sign } p_{n-2}(y_1) , \quad 1 \le i \le n .$$
   Folglich gilt:
   $$\text{sign } (p_n(y_i)) = -\text{sign } (p_n(y_{i+1})) , \quad 1 \le i \le n-1 ,$$
   so daß $p_n(x)$ eine Nullstelle z_{i+1} zwischen y_i und y_{i+1} hat, d.h.
   $$a < y_1 < z_2 < y_2 < \cdots < z_{n-1} < y_{n-1} < b .$$
   Weiter gilt:
   $$\begin{aligned} \text{sign } (p_n(y_1)) &= -\text{sign } (p_{n-2}(y_1)) \\ \text{sign } (p_n(y_n)) &= -\text{sign } (p_{n-2}(y_n)) \\ \text{sign } (p_n(x)) &= \text{sign } (p_{n-2}(x)) , \quad \text{für} \quad |x| \longrightarrow \infty . \end{aligned}$$
   Es folgt, daß p_n zwei weitere Nullstellen z_1 und z_n besitzt mit
   $$a < z_1 < y_1 \qquad \text{und} \qquad y_n < z_n < b .$$
   Damit ist die Induktionshypothese bestätigt:
   $$a < z_1 < y_1 < \cdots < y_{n-1} < z_n < b .$$

□

Beispiele von Orthogonalpolynomen

a	b	$w(x)$	Bezeichnung
-1	1	1	Legendre
0	∞	e^{-x}	Laguerre
$-\infty$	$+\infty$	e^{-x^2}	Hermite

3.6 Zusammengesetzte Regeln

Kennzeichnend für die moderne Numerik ist, daß man lokale Approximationen niedriger Ordnung über globale Approximationen höherer Ordnung stellt. Dementsprechend wird ein Integral

$$I(\Omega)f := \int\limits_{\Omega} f(x)dx$$

oft dadurch approximiert, daß das Gebiet Ω in mehrere Teilgebiete Ω_k aufgeteilt

$$\Omega = \bigcup_{k=1}^{m} \Omega_k \ , \qquad \Omega_k \cap \Omega_\ell = \phi \qquad \text{für} \qquad k \neq \ell \ ,$$

und jedes der entsprechenden Integrale gesondert approximiert wird,

$$I(\Omega)f = \sum_{k=1}^{m} \int\limits_{\Omega_k} f(x)dx \ , = \sum_{k=1}^{m} I(\Omega_k)f \ , \doteq \sum_{k=1}^{m} \tilde{I}(\Omega_k)f \ ,$$

wobei $\tilde{I}(\Omega_k)f$ eine Approximation zu $I(\Omega)f$ ist. Die Vorteile dieser Vorgehensweise sind:

1. Die Teilgebiete Ω_k können so klein gewählt werden, daß es genügt, $\tilde{I}(\Omega_k)f$ mit Hilfe von Integrationsformeln niedriger Ordnung zu berechnen.

2. Stünden mehrere Prozessoren zur Verfügung, könnte die Berechnung verschiedener Approximationswerte $\tilde{I}(\Omega_k)f$ auf verschiedene Prozessoren verteilt werden.

3. Falls der Integrand f in einem Teilgebiet Ω_k besonders stark variiert, kann der Berechnung von $\tilde{I}(\Omega_k)f$ besonders viel Sorgfalt gewidmet werden.

Wir werden diese Ideen zuerst in einem einfachen, aber sehr wichtigen Spezialfall anwenden:

$$\begin{aligned}
\Omega &= [a,b] \subset \mathbb{R}^1 \ , \\
\Omega_k &= [a + \frac{k-1}{m}(b-a) \ , \ a + \frac{k}{m}(b-a)] \ , \\
\tilde{I}(\Omega_k)f &= I_n^{NC}(\Omega_k)f \ ,
\end{aligned}$$

wobei $I_n^{NC}(\Omega_k)f$ durch eine geschlossene Newton-Cotes-Formel mit $(n+1)$ Stützpunkten berechnet wird,

$$I_n^{NC,m}(\Omega)f := \sum_{k=1}^{m} \left\{ h \sum_{i=0}^{n} a_i^{(n)} f(x_{(k-1)n+i}) \right\}$$

mit

$$\begin{aligned} x_j &:= a + j\,h\,, \qquad 0 \le j \le m \cdot n\,, \\ h &:= \frac{b-a}{m \cdot n}\,, \end{aligned}$$

wo $a_i^{(n)}$ die Gewichte der geschlossenen Newton-Cotes Integrationsformel mit n Intervallen sind. Diese Integrationsformel heißt die *wiederholte Newton-Cotes Integrationsformel* oder die *zusammengesetzte Newton-Cotes Integrationsformel.*

Beispiel:$[n = 1]$

$$\begin{aligned} I_1^{NC,m}([a,b])f &= \sum_{k=1}^{m} \{h[\frac{1}{2}f(a+(k-1)h) + \frac{1}{2}f(a+kh)]\} \\ &= h\left[\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{k=1}^{m-1} f(a+kh)\right] \end{aligned}$$

mit

$$h = \frac{b-a}{m}\,.$$

Diese Formel, die wiederholte Trapezregel, kommt in mehreren Anwendungen vor. Statt $I_1^{NC,m}([a,b])f$ schreibt man oft $T(h)$, wenn der Integrand f und das Intervall $[a,b]$ festgelegt worden sind.
Es gilt:

$$T(h) = \frac{h}{2}[f(x_0) + 2f(x_1) + \cdots + 2f(x_{m-1}) + f(x_m)]$$

mit

$$x_k := a + kh\,.$$

Diese Formulierung der Integrationsformel wird oft benutzt, obwohl sie rechnerisch ungünstig ist, da mehrere unnötige Multiplikationen mit dem Faktor 2 vorliegen.

Beispiel:$[n = 2]$

$$I_n^{NC,m}([a,b])f = \sum_{k=1}^{m}\{h[\frac{1}{3}f(x_{2k-2}) + \frac{4}{3}f(x_{2k-1}) + \frac{1}{3}f(x_{2k})]\}$$

mit

$$h := \frac{b-a}{2m}\,.$$

Diese Integrationsformel kann auch folgendermaßen umformiert werden:

$$\begin{aligned} I_n^{NC,m}([a,b])f &= \frac{h}{3}[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots \\ &+ 2f(x_{2m-2}) + 4f(x_{2m-1}) + f(x_{2m})] \,. \end{aligned}$$

Diese Formel, die wiederholte Simpsonregel, ist vielleicht jene Formel, die am meisten benutzt wird, wenn keine numerische Software vorhanden ist. Sie ist einerseits sehr einfach und andererseits ziemlich genau. Statt $I_2^{NC,m}([a,b])f$ schreibt man oft $S(h)$, wenn der Integrand f und das Intervall $[a,b]$ festgelegt worden sind.

Ein Vorteil der zusammengesetzten Integrationsformel ist die einfache Fehlerabschätzung, wie das aus dem folgenden Satz hervorgeht, den wir der Einfachheit halber nur für die zusammengesetzten Newton-Cotes Integrationsformeln formulieren:

Satz 3.8 *Seien* $-\infty < a < b < +\infty$ *und* $I_n^{NC,m}([a,b])f$ *eine zusammengesetzte Newton-Cotes Integrationsformel. Dann gilt:*

1. *Sei*

$$r = \begin{cases} n+1 & , \textit{ für } n \textit{ ungerade} \,, \\ n+2 & , \textit{ für } n \textit{ gerade} \,. \end{cases}$$

Sei $f \in C^{(r)}([a,b])$. *Dann gibt es* $\xi \in (a,b)$ *und* $C_n \in \mathbb{R}$ *mit*

$$I\ f - I_n^{NC,m}([a,b])f = (b-a)C_n h^r f^{(r)}(\xi) \,.$$

2. *Sei* $f \in C[a,b]$. *Dann gilt:*

$$\lim_{m\to\infty} I_n^{NC,m}([a,b])f = I\ f \,.$$

Beweis:

1. Es gibt (siehe Satz 3.1 und Tabelle 3.1) ein $\xi_k \in (x_{(k-1)n}, x_{kn})$ mit

$$\int\limits_{x_{(k-1)n}}^{x_{kn}} f\ dx - \sum_{i=0}^{n} h\ a_i^{(n)} f(x_{(k-1)n+i}) = c_n\ h^{r+1} f^{(r)}(\xi_k) \,.$$

Durch diese Abschätzung, von $k = 1$ bis $k = m$ zu summieren, erhält man:

$$E_n^m f := I\ f - I_n^{NC,m}([a,b])f = c_n h^{r+1} \sum_{k=1}^{m} f^{(r)}(\xi_k) \,.$$

Sei

$$\begin{aligned} g_r &:= \min_{x\in[a,b]} f^{(r)}(x) \\ G_r &:= \max_{x\in[a,b]} f^{(r)}(x) \end{aligned}$$

Dann ist

$$g_r \leq \frac{1}{m} \sum_{k=1}^{m} f^{(r)}(\xi_k) \leq G_r \; .$$

Folglich gibt es ein $\xi \in [a, b]$ mit

$$\frac{1}{m} \sum_{k=1}^{m} f^{(r)}(\xi_k) = f^{(r)}(\xi) \; ,$$

d.h.

$$E_n^m f = \frac{c_n}{n} (b-a) h^r f^{(r)}(\xi) \; .$$

2. Wir überprüfen die Bedingungen des Satzes von Steklow und Polya:

   (a) Sei f ein Polynom. Sei r wie in 1). Dann folgt:

   $$|I\ f - I_n^{NC,m} f| \leq \max_{a \leq x \leq b} |f^{(r)}(x)| \cdot C_n \cdot b - a \cdot \frac{\cdot (b-a)^r}{(mn)^r} \; ,$$

   so daß

   $$\lim_{m \to \infty} I_n^{NC,m} f = I\ f \; .$$

   Damit ist die erste Bedingung des Satzes von Steklow und Polya erfüllt.

   (b) Die Integrationsformel $I_n^{NC,m}$ läßt sich folgendermaßen schreiben:

   $$\sum_{j=0}^{mn} A_j^{m,n} f(x_j) =: I_n^{NC,m} f \; .$$

   Es folgt:

   $$\sum_{j=0}^{mn} |A_j^{m,n}| \leq \sum_{k=1}^{m} h \sum_{i=0}^{n} |a_i^{(n)}| \leq K := \frac{b-a}{n} \sum_{i=0}^{n} |a_i^{(n)}| \; .$$

   Damit ist die zweite Bedingung des Satzes von Steklow und Polya bewiesen.

□

Die Fehlerabschätzungen für einige zusammengesetzte Newton-Cotes Formeln werden in Tabelle 3.3 zusammengefaßt.

Beispiel:

1. $I = \int\limits_0^1 e^x dx$

h	T_h	$I - T_h$	Quotient aufeinanderfolgender Fehler
1	1.8591	-0.1409	–
1/2	1.7539	-0.0357	3.95
1/4	1.7272	-0.0089	4.01

n	Bezeichnung	Fehlerabschätzung
1	Trapezregel	$h^2 \cdot \frac{b-a}{12} \cdot f^{(2)}(\xi)$
2	Simpson-Regel	$h^4 \cdot \frac{b-a}{180} \cdot f^{(4)}(\xi)$
3	3/8-Regel	$h^4 \cdot \frac{b-a}{80} \cdot f^{(4)}(\xi)$
4	Milne-Regel	$h^6 \cdot \frac{2}{945} \cdot b - a \cdot f^{(6)}(\xi)$

Tabelle 3.3: Fehlerabschätzungen für zusammengesetzte Newton-Cotes Regeln

Der Fehlerquotient zeigt den Abfall mit h^2 : Bei Halbierung von h geht der Fehler auf ungefähr ein Viertel zurück. Eine Überprüfung dieses Abfalls ist eine gute Kontrolle auf Rechenfehler.

2. $I = \int\limits_0^1 \sqrt{x}\, dx$

h	T_h	$I - T_h$	Quotient aufeinanderfolgender Fehler
1	0.5000	0.1667	–
1/2	0.6036	0.0631	2.67
1/4	0.6433	0.0234	2.70

Hier fällt der Fehler nicht mit h^2 ab. Unsere Abschätzung ist nicht anwendbar, da $\sqrt{x} \notin C^2[0,1]$.

3.7 Praktische Anwendungen

Wenn gewünscht wird, ein Integral $I = \int\limits_\Omega f\, dx$ mit einem Fehler von höchstens ϵ zu berechnen, können die folgenden Hinweise nützlich sein:

1. Überprüfe, ob das Integral existiert, d.h. ob die Voraussetzungen für die Existenz vorhanden sind. Wenn das Integral durch mehrere Transformationen und Manipulationen hergeleitet worden ist, ist es durchaus möglich, daß sich Fehler einschleichen, so daß z.B. das Integral divergiert.

2. Überprüfe, ob der Integrand sich in einigen Teilgebieten besonders schnell ändert, d.h. ob die Ableitungen von f in einigen Teilgebieten betragsmäßig groß werden. Wenn ja, sollten solche Gebiete gesondert behandelt werden:

$$I = I_g + I_{ng} = \int\limits_{\Omega_{\text{glatt}}} f\, dx + \int\limits_{\Omega_{\text{nicht-glatt}}} f\, dx .$$

Hierdurch kann eine wesentliche Effizienzsteigerung erzielt werden, da die Berechnung von I_{ng} eine viel dichtere Verteilung von Stützpunkten erfordert.

3. Überprüfe, ob I analytisch ausgewertet werden kann. Wenn nicht, überprüfe, ob es eine Funktion g gibt, so daß das Integral $\int_\Omega g\, dx$ analytisch ausgewertet werden kann und $(f - g)$ entweder sehr klein oder sehr glatt und deshalb leichter als f zu berechnen ist.

4. Wenn der gewünschte Fehler ϵ sehr klein ist, müssen Rundungsfehler eventuell berücksichtigt werden.

5. Sollte das Integral öfter berechnet werden (z.B. wenn f von einem Parameter abhängig ist), muß die Effizienz besonders berücksichtigt werden.

6. Wenn eine exakte Fehlerabschätzung oder Erfahrung mit ähnlichen Integralen nicht vorliegt, muß der Fehler empirisch abgeschätzt werden. Die kanonische Methode ist, die Berechnungen mit drei verschiedenen Intervallen h durchzuführen:

$$h_1 < h_2 < h_3 \ ,$$
$$h_2 = \alpha\, h_1 \ , \quad h_3 = \alpha\, h_2 \ , \quad \alpha \in (0,1)$$

Man setzt voraus, daß

$$\begin{aligned} I_1 := I(h_1) &\doteq I + c\, h_1^s \\ I_2 := I(h_2) &\doteq I + c\, h_2^s \\ I_3 := I(h_3) &\doteq I + c\, h_3^s \\ I_1 - I_2 &= c(h_1^s - h_2^s) = c\, h_1^s(1 - \alpha^s) \\ I_2 - I_3 &= c(h_2^s - h_3^s) = c\, h_2^s(1 - \alpha^s) \end{aligned}$$

Es folgt

$$\begin{aligned} s &= \ell n \left(\frac{I_2 - I_3}{I_1 - I_2} \right) / \ell n\ \alpha \ , \\ I &= I_3 - c\, h_3^s = I_3 - c\, h_2^s \cdot \left(\frac{h_3}{h_2} \right) = I_3 - (I_2 - I_3) \frac{\alpha^s}{1 - \alpha^s} \end{aligned}$$

z.B., wenn $\alpha = 1/2$ und $s = 2$, gilt:

$$I = I_3 - \frac{(I_2 - I_3)}{3} = \frac{4\, I_3 - I_2}{3} \ .$$

Man kann diese Formel benutzen, um eine bessere Approximation zu I zu bekommen. Wichtig ist aber, daß der berechnete Wert von α mit dem theoretischen Wert übereinstimmt.

3.8 Die Euler-Maclaurin Formel

Sei $h = \frac{b-a}{n}$, $n \in \mathbb{N}$ und $T(h)$ die Trapezsumme

$$T(h) := h\left[\frac{1}{2}f(a) + f(a+h) + \cdots + f(b-h) + \frac{1}{2}f(b)\right]$$

Wir möchten eine asymptotische Entwicklung für $T(h)$ herleiten. Es ist zunächst erforderlich, die Bernoulli-Polynome $B_n(x)$ und Bernoulli-Zahlen B_n zu betrachten. Die Polynome $B_n(x)$ werden mit Hilfe einer Erzeugungsfunktion definiert:

$$e^{xt}\frac{t}{e^t-1} = \sum_{n=0}^{\infty}\frac{B_n(x)}{n!}t^n , \tag{3.3}$$

so daß z.B.

$$\begin{aligned}
B_0(x) &\equiv 1\\
B_1(x) &= x - \frac{1}{2}\\
B_2(x) &= x^2 - x + \frac{1}{6}\\
B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x\\
B_4(x) &= x^4 - 2x^3 + x^2 - 1/30\\
B_5(x) &= x^5 - \frac{5}{2}x^4 + \frac{5}{3}x^3 - \frac{1}{6}x\\
B_6(x) &= x^6 - 3x^5 + \frac{5}{2}x^4 - \frac{1}{2}x^2 + \frac{1}{42}\\
B_7(x) &= x^7 - \frac{7}{2}x^6 + \frac{7}{2}x^5 - \frac{7}{6}x^3 + \frac{1}{6}x\\
B_8(x) &= x^8 - 4x^7 + \frac{14}{3}x^6 - \frac{7}{3}x^4 + \frac{2}{3}x^2 - \frac{1}{30}
\end{aligned}$$

Die Bernoulli-Zahlen werden durch $B_n := B_n(0)$ definiert, so daß aus (3.3)

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

folgt, so daß z.B.

$$\begin{aligned}
B_0 &= 1\\
B_1 &= -\frac{1}{2}\\
B_2 &= \frac{1}{6}\\
B_3 &= 0\\
B_4 &= -1/30\\
B_5 &= 0\\
B_6 &= 1/42\\
B_7 &= 0\\
B_8 &= -1/30\\
B_9 &= 0\\
B_{10} &= 5/66
\end{aligned}$$

Bemerkung: Wir benutzen die Notation von Abramowitz und Stegun, Handbook of Mathematical Functions. Leider werden in der Literatur verschiedene Definitionen benutzt, z.B. Stoer, Einführung in die Numerische Mathematik, setzt

$$B_{2k}^{Stoer} = (-1)^k B_k$$

und Atkinson, An Introduction to Numerical Analysis, setzt

$$\sum_{n=1}^{\infty} \frac{B_n^{Atkinson}(x)}{n!} t^n = \frac{t(e^{xt}-1)}{e^t-1}.$$

Wird die Gleichung (3.3) nach x differenziert, so folgt

$$\sum_{n=0}^{\infty} \frac{B_n'(x)}{n!} t^n = te^{xt}\frac{t}{e^t-1} = t\sum_{n=0}^{\infty} \frac{B_n(x)}{n!} t^n \ .$$

Der Vergleich der Koeffizienten von t^n ergibt:

$$B_n'(x) = nB_{n-1}(x)\ , \quad n \geq 0$$

und

$$B_n(x) = B_n + n\int_0^x B_{n-1}(t)dt \tag{3.4}$$

mit $B_{-1}(x)$ und $B_{-1} = 0$.

Weitere Eigenschaften sind:

Satz 3.9 *1.*

$$B_{2n+1} = 0 \ \textit{für} \ n \geq 1 \ .$$

2.

$$B_n(0) = B_n(1) \ \textit{für} \ n = 0 \ \textit{und} \ n \geq 2 \ . \tag{3.5}$$

3.

$$B_n(1-x) = (-1)^n B_n(x) \ .$$

4. $B_{2k}(x) - B_{2k}$ *hat das gleiche Vorzeichen für* $x \in [0,1]$ *und erreicht sein Extremum für* $x = \frac{1}{2}$*.*

5.

$$B_n\left(\frac{1}{2}\right) = -(1 - 2^{-n+1}) B_n$$

6.

$$(-1)^{n+1} B_{2n} > \frac{2 \cdot (2n)!}{(2\pi)^{2n}} \ \textit{für} \ n \geq 1 \ ,$$

d.h. die Bernoulli-Zahlen werden sehr groß.

Beweis:

1. Die Funktion
$$\frac{t}{e^t - 1} - B_1 t = t\left(\frac{e^t - 1}{e^t + 1}\right)$$
ist symmetrisch um $t = 0$.

2.
$$\sum_{n=0}^{\infty} \frac{B_n(0)}{n!} t^n - \sum_{n=0}^{\infty} \frac{B_n(1)}{n!} t^n = \frac{t}{e^t - 1} - \frac{te^t}{e^t - 1} = -t.$$

3.
$$\sum_{n=0}^{\infty} \frac{B_n(1-x)}{n!} t^n - \sum_{n=0}^{\infty} \frac{B_n(x)}{n!} (-t)^n = \frac{te^{(1-x)t}}{e^t - 1} - \frac{(-t)e^{-xt}}{e^{-t} - 1} = 0.$$

4. Man betrachtet zuerst die Nullstellen von $B_{2k-1}(x)$. Es gilt:

$$B_{2k-1}(0) = B_{2k-1}(1) = B_{2k-1}\left(\frac{1}{2}\right) = 0 \text{ für } k \geq 2.$$

Ist $\overline{x} \in [0,1]$ eine weitere Nullstelle von $B_{2k-1}(x)$, so ist $1-\overline{x}$ auch eine Nullstelle. Also hat $B_{2k-1}(x)$ mindestens fünf Nullstellen in $[0,1]$. Nach dem Satz von Rolle hat

$$(2k-1)\,B_{2k-2}(x) = B'_{2k-1}(x)$$

dann mindestens vier Nullstellen in $(0,1)$. Dazu kommen die Nullstellen $x=0$ und $x=1$, so daß $B_{2k-3}(x)$ mindestens fünf Nullstellen hat. Durch Induktion folgt, daß $B_3(x)$ mindestens fünf Nullstellen hat, ein Widerspruch.
Damit ist bewiesen, daß für $k \geq 2$ $B_{2k-1}(x)$ nur die drei Nullstellen $0, \frac{1}{2}, 1$ in $[0,1]$ hat. Die Funktion $B_{2k}(x) - B_{2k}$ ist symmetrisch um $x = \frac{1}{2}$ und ihre Ableitung

$$[B_{2k}(x) - B_{2k}]' = 2k\,B_{2k-1}(x)$$

ist Null für $x = \frac{1}{2}$ und $x = 0$, aber sonst ungleich Null. Es folgt, daß $B_{2k}(x)$ in $[0,1]$ das gleiche Vorzeichen besitzt und daß

$$\max_{0\leq x\leq 1} |B_{2k}(x) - B_{2k}| = B_{2k}\left(\frac{1}{2}\right).$$

5.
$$\frac{t}{e^{t/2} - e^{-t/2}} = \sum_{n=0}^{\infty} B_n\left(\frac{1}{2}\right)\frac{t^n}{n!}$$

und

$$\frac{t}{e^{t/2} - e^{-t/2}} = \sum_{n=0}^{\infty} [2^{-n+1} - 1] B_n \frac{t^n}{n!}.$$

□

Die durch

$$B_n^*(x) := B_n(x-i)\ ,\quad i \leq x < i+1\ ,\quad i \in \mathbb{N}$$

definierte Funktion ist eine periodische Fortsetzung der Funktion $B_n(x)$ mit Periode 1. Für $n \geq 2$ ist $B_n^*(x)$ stetig (siehe (3.5)); $B_0^* \equiv 1$ und $B_1^*(x)$ ist stückweise linear (siehe Abbildung 3.3).

Aus (3.4) und (3.5) folgt:

$$B_n^*(x) = B_n^*(0) + n\int_0^x B_{n-1}^*(t)dt \text{ für } n = 0 \text{ und } n \geq 2\ . \tag{3.6}$$

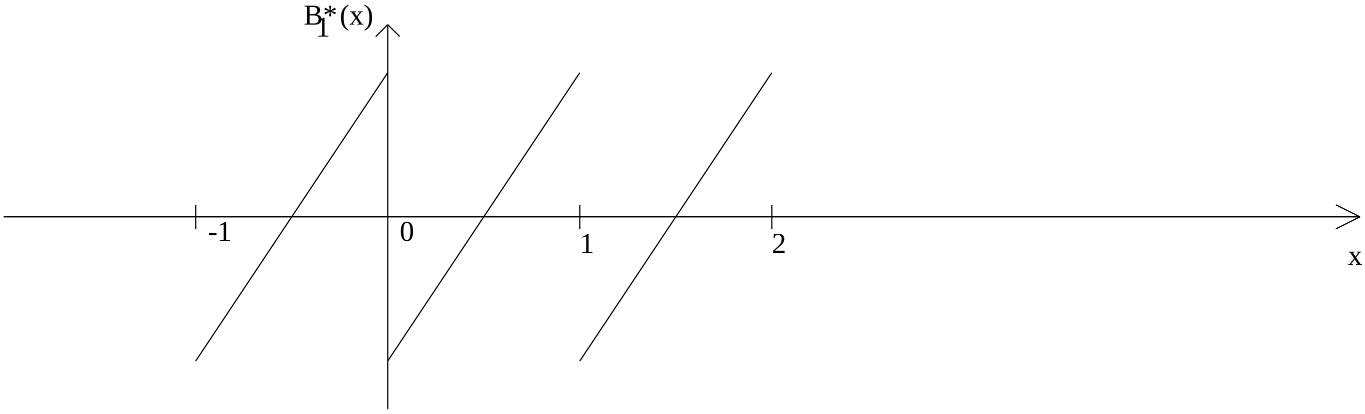


Abbildung 3.3: Die Funktion $B_1^*(x)$

Satz 3.10 (Euler-MacLaurin-Formel) *Sei $m \geq 0$, $n \geq 1$, $h = \frac{b-a}{n}$, $x_j = a + jh$. Sei $f \in C^{2m+2}[a,b]$. Dann gilt:*

$$\begin{aligned} T(h) &= \int_a^b f(x)dx + \sum_{i=1}^m h^{2i} \frac{B_{2i}}{(2i)!}[f^{(2i-1)}(b) - f^{(2i-1)}(a)] \\ &- \frac{h^{2m+2}}{(2m+2)!} \int_a^b \left[B_{2m+2}^* \left(\frac{x-a}{h} \right) - B_{2m+2} \right] f^{(2m+2)}(x)dx \; . \end{aligned} \tag{3.7}$$

Beweis: Statt f betrachten wir die Funktion $g(t)$, die durch

$$g(t) := f(a+th)$$

definiert ist. Dann gilt:

$$\int_a^b f(x)dx = h \int_0^n g(t)dt.$$

Sei $0 \leq i < n$. Dann folgt durch partielle Integration

$$\begin{aligned} \int_i^{i+1} g(t)dt &= \int_i^{i+1} g(t) \frac{d}{dt}(B_1^*(t))dt \\ &= \int_0^1 g(t+i) \frac{d}{dt}(B_1(t))dt \\ &= [g(t+i)B_1(t)]_0^1 - \int_0^1 g'(t+i)B_1(t)dt \\ &= \frac{1}{2}[g(i+1) + g(i)] - \int_0^1 g'(t+i)B_1(t)dt \end{aligned}$$

und nach Summation über i:

$$T(h) = \int_a^b f(x)dx + h\int_0^n g'(t)B_1^*(t)dt \tag{3.8}$$

$$= \int_a^b f(x)dx + h\int_0^n g'(t)\frac{d}{dt}\left[\frac{1}{2}B_2^*(t)\right]dt \tag{3.9}$$

Sei $u, v \in C[0,n]$, $u \in C^1[0,n]$, v stetig differenzierbar mit Ausnahme von endlich vielen Stellen t_j und existiere

$$\lim_{t\to t_j+0} v'(t) = v'(t_j+0)$$
$$\lim_{t\to t_j-0} v'(t) = v'(t_j-0)\ ,$$

dann folgt

$$\int_0^n uv'dt = [uv]_0^n - \int_0^n u'v\ dt\ . \tag{3.10}$$

Die Funktionen $B_k^*(t)$ sind stetig für $k \geq 2$ (siehe (3.5)) und $\frac{d}{dt}B_k^* = k\ B_{k-1}^*$ (siehe (3.6)). Durch wiederholte partielle Integration folgt aus (3.8), (3.9) und (3.10):

$$\int_0^n g'(t)B_1^*(t)dt = \sum_{k=1}^{2m}\frac{(-1)^{k+1}}{(k+1)!}\left[g^{(k)}(t)B_{k+1}^*(t)\right]_{t=0}^{t=n} + \frac{(-1)^{2m+2}}{(2m+1)!}\int_0^n g^{(2m+1)}(t)B_{2m+1}^*(t)dt\ .$$

Da $B_j^*(0) = B_j^*(1) = 0$ für j ungerade, entfällt die Hälfte der Terme in der Summe. Weiter gilt:

$$g^{(k)}(t) = h^k f^{(k)}(a+th)\ . \tag{3.11}$$

Das Integral kann noch einmal durch partielle Integration transformiert werden, wobei wir diesmal statt der Identität

$$B_{2m+1}^*(t) = \frac{1}{2m+2}\frac{d}{dt}\left(B_{2m+2}^*(t)\right)$$

die Identität

$$B_{2m+1}^*(t) = \frac{1}{2m+2}\frac{d}{dt}\left[B_{2m+2}^*(t) - B_{2m+2}^*(0)\right]$$

benutzen. Zusammenfassend kann die Gleichung (3.11) wie folgt geschrieben werden:

$$\begin{aligned}\int_0^n g'(t)B_1^*(t)dt &= \sum_{i=1}^m \frac{h^{2i-1}B_{2i}}{(2i)!}\left[f^{(2i-1)}(b) - f^{(2i-1)}(a)\right] \\ &- \frac{1}{(2m+2)!}\int_0^n g^{(2m+2)}(t)\left[B_{2m+2}^*(t) - B_{2m+2}(0)\right]dt\end{aligned}$$

Der Satz folgt aus (3.8), (3.11) und (3.12). □

Es ist manchmal möglich, den Fehlerterm $E_m(f)$

$$E_m(f) := -\frac{h^{2m+2}}{(2m+2)!}\int_a^b \left[B_{2m+2}^*\left(\frac{x-a}{h}\right) - B_{2m+2}\right] f^{2m+2}(x)dx$$

umzuformen.

Satz 3.11 *Sei* $f \in C^{2m+2}[a,b]$. f *habe das gleiche Vorzeichen auf* $[a,b]$. *Dann gilt: Es gibt* $\theta \in [0,1]$ *mit*

$$E_m(f) = +\theta(2 - 2^{-2m-1})\frac{h^{2m+2}}{(2m+2)!}B_{2m+2}\left[f^{(2m+1)}(b) - f^{(2m+2)}(a)\right] \ ,$$

d.h. der Fehler hat das gleiche Vorzeichen wie der erste vernachlässigte Term der Euler-MacLaurin Entwicklung und ist höchstens zweimal so groß.

Beweis: Siehe Krylov [1962, S. 217]. □

Die Euler-MacLaurin Formel (3.7) hat u.a. folgende Anwendung:

Die Summation von langsam konvergierenden Folgen, wie z.B.

$$\sum_{n=1}^\infty \frac{1}{n^2}.$$

Sei $f(x) = 1/x^2$, $a = 10$, $b = \infty$, $h = 1$. Es folgt:

$$T(h) = \frac{1}{2}f(10) + \sum_{n=1}^\infty \frac{1}{(10+n)^2} = \int_{10}^\infty \frac{1}{x^2}dx + \sum_{i=1}^m \frac{B_{2i}}{10^{2i+1}} + E_m \ ,$$

so daß

$$\sum_{n=1}^\infty \frac{1}{n^2} = \sum_{n=1}^9 \frac{1}{n^2} + \frac{\frac{1}{2}}{10^2} + \frac{1}{10} + \sum_{i=1}^m \frac{B_{2i}}{10^{2i+1}} + E_m.$$

Wähle z.B. $m = 2$, dann folgt:

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{n^2} &= 1,5397677311\ldots + \frac{1}{200} + \frac{1}{10} + \frac{1}{6}\frac{1}{10^3} - \frac{1}{30}\frac{1}{10^5} + E_2 \\ &= 1,6449340644 + E_2\end{aligned}$$

mit

$$|E_2| \leq 2 \cdot \frac{1}{42} \cdot 10^{-7}.$$

3.9 Romberg-Integration

Eine wichtige Anwendung der Euler-MacLaurin Formel ist die Romberg- Integration. Sei $f \in C^{2m+2}[a,b]$. Nach Satz 3.10 gilt:

$$T(h) = \tau_0 + \tau_1 h^2 + \tau_2 h^4 + \cdots + \tau_m h^{2m} + h^{2m+2}\alpha_{m+1}(h)$$

mit

$$\begin{aligned}\tau_0 &:= I\,f = \int_a^b f(x)dx \\ \tau_i &= \frac{B_{2i}}{(2i)!}\left[f^{(2i-1)}(b) - f^{(2i-1)}(a)\right] \;,\quad 1 \leq i \leq m \\ \alpha_{m+1}(h) &= -\frac{h^{2m+2}}{(2m+2)!}\int_a^b \left[B^*_{2m+2}\left(\frac{x-a}{h}\right) - B_{2m+2}\right] f^{(2m+2)}(x)dx\end{aligned}$$

Wichtig dabei ist, daß $\tau_0, \ldots, \tau_m$ Konstanten sind.
Sei nun

$$\begin{aligned}h_0 &:= (b-a) \\ h_i &:= \frac{h_0}{n_i} \;,\quad n_i \in \mathbb{N}\;,\end{aligned}$$

z.B. die *Romberg-Folge*

$$h_i := \frac{b-a}{2^i} \;,\quad i \geq 0.$$

Man berechnet

$$T_{i0} := T(h_i) \;,\quad 0 \leq i \leq m$$

und

$$T_{ik} := T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1} \;,\quad 1 \leq k \leq i \leq m.$$

$$
\begin{array}{lllll}
T_{00} & & & & \\
T_{10} & T_{11} & & & \\
T_{20} & T_{21} & T_{22} & & \\
\vdots & \vdots & & \ddots & \\
T_{m0} & T_{m1} & T_{m2} & \cdots & T_{mm}
\end{array}
$$

Tabelle 3.4: Das Romberg-Tableau

Für die Romberg-Folge gilt:

$$T_{ik} := T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{4^k - 1}\ , \quad 1 \leq k \leq i \leq m.$$

Die Zahlen T_{ik} können als Tableau dargestellt werden (siehe Tabelle 3.4).

Die Zahl T_{ik} ist gleich $\tilde{T}_{ik}(0)$, wo $\tilde{T}_{ik}$ dasjenige Polynom in h^2 vom Grade $2k$,

$$\tilde{T}_{ik}(h) := a_0 + a_1 h^2 + \cdots + a_k h^{2k}\ ,$$

ist, für das gilt:

$$\tilde{T}_{ik}(h_s) = T(h_s)\ , \quad i - k \leq s \leq i.$$

Die Berechnung von T_{ik} erfolgt nach dem Schema von Neville für Polynominterpolation. Es gilt:

$$\tilde{T}_{ik}(h) = \frac{\left(h^2 - h_{i-k}^2\right) \tilde{T}_{i,k-1}(h) + \left(h_i^2 - h^2\right) \tilde{T}_{i-1,k-1}}{h_i^2 - h_{i-k}^2}$$

(siehe Tabelle 3.5), so daß

$$\tilde{T}_{ik}(0) = T_{ik}.$$

$$
\begin{array}{cccccc}
h_{i-k}^2 & T_{0,i-k} & & & & \\
h_{i-k+1}^2 & T_{0,i-k+1} & T_{1,i-k+1} & & & \\
\vdots & \vdots & & \ddots & & \\
h_{i-1}^2 & T_{0,i-1} & T_{1,i-1} & & T_{i-1,k-1} & \\
h_i^2 & T_{0,i} & T_{1,i} & & T_{i,k-1} & T_{i,k}
\end{array}
$$

Tabelle 3.5: Das Neville-Schema

Beispiel: Für $I = \int\limits_1^2 \frac{1}{x} dx = \ln\ 2 = 0,693147180\ldots$ ergeben sich die Daten in Tabelle 3.6 (Bauer et al. [1963]):

Es kann bewiesen werden, daß

.750 000 000				
.708 333 333	.694 444 444			
.697 023 809	.693 253 967	.693 174 603		
.694 121 851	.693 154 532	.693 147 901	.693 147 479	
.693 391 202	.693 147 652	.693 147 193	.693 147 182	.693 147 181.

Tabelle 3.6:

1.
$$T_{ik} - \int_a^b f(x)dx = \frac{4^{-(i-k)(k+1)} B_{2k+2}}{2^{k(k+1)}(2k+2)!} f^{(2k+2)}(\xi) ,$$
so daß T_{ik} etwa um einen Faktor 4^{k+1} genauer ist als $T_{i-1,k}$ (siehe Bauer et al. [1963, S. 210]).

2. Sei $f \in C[a,b]$. Dann gilt
$$\lim_{k\to\infty} T_{kk} = \int_a^b f(x)dx.$$
(Siehe Bauer et al. [1963, Satz 2].)

3.10 Mehrdimensionale Integration

Die Berechnung von mehrdimensionalen Integralen für allgemeine Gebiete $\Omega \subset \mathbb{R}^n$ ist nicht leicht. Wir geben nur zwei einfache Beispiele:

Beispiel: Ist $\Omega \subset \mathbb{R}^2$ das Rechteck $[a_1, a_2] \times [b_1, b_2]$, dann kann wiederholte eindimensionale Integration benutzt werden:

$$I(\Omega)f = \int_{a_1}^{a_2} dx \int_{b_1}^{b_2} dy\ f(x,y) \doteq \sum_{i=0}^{n} A_i \sum_{j=0}^{m} B_j f(x_i, y_j) ,$$

wobei

$$\begin{aligned} I_n^{NC}[a_1,a_2]f &:= \sum_{i=0}^{n} A_i f(x_i) \\ I_m^{NC}[b_1,b_2]g &:= \sum_{j=0}^{m} B_j g(y_j) \end{aligned}$$

Beispiel: Ist $\Omega \subset \mathbb{R}^2$ ein spezielles Gebiet, so existieren manchmal spezielle Formeln, z.B. ist $\Omega := \{(x,y) : x^2 + y^2 \le h^2\}$, so gilt:

$$\int_\Omega f(x,y)dx\,dy = \sum_{i=1}^{n} w_i f(x_i, y_i) + R$$

mit $R = O(h^6)$ und

(x_i, y_i)	w_i
$(0,0)$	$1/2$
$(\pm h, 0)$	$1/24$
$(0, \pm h)$	$1/24$
$\left(\pm\frac{h}{2}, \pm\frac{h}{2}\right)$	$1/6$

Tabelle 3.7: Eine spezielle Integrationsformel für den Kreis mit Radius h

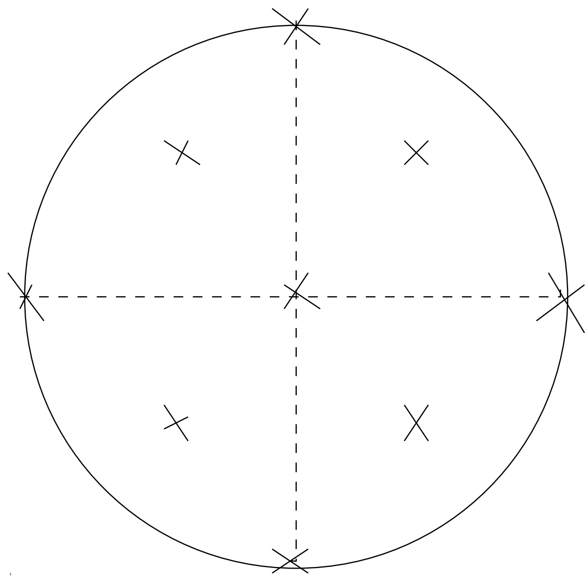


Abbildung 3.4: Die Stützstellen für eine spezielle Integrationsformel für einen Kreis

Literatur

Bauer, F.L., **Rutishauser**, H. und **Stiefel**, E.: New aspects in numerical quadrature. In Experimental Arithmetic, High Speed Computing and Mathematics. Amer. Math. Soc., 1963.

Davis, P.J., **Rabinowitz**, P.: Methods of Numerical Integration. Second Edition. New York:] Academic Press, 1984

Engels, H.: Numerical Quadrature and Cubature. London, Academic Press, 1980.

Ghizzetti, A., **Ossicini**, A.: Quadrature Formulae. New York:] Academic Press, 1970.

Krylov, V.I.: Approximate Calculation of Integrals. New York:] Macmillan, 1962.

Levin, M., **Girshovich**, J.: Optimal Quadrature Formulas. Teubner, 1979.

Piessens, R., **de Doncker-Kapenga**, E., **Überhuber**, C.W., **Kahaner**, D.K.: Quadpack:] A Subroutine Package for Automatic Integration. Berlin:] Springer, 1983.

Rivlin, T.J.: The Chebyshev Polynomials. New York:] Wiley, 1974.

Stroud, A.H.: Approximate Calculation of Multiple Integrals. Englewood Cliffs, N.J.:] Prentice Hall, 1971.

Kapitel 4

Gewöhnliche Differentialgleichungen

4.1 Einleitung

Sei $m \in \mathbb{N}$, $r_k \in \mathbb{N}$ für $1 \leq k \leq m$, $I \subset \mathbb{R}$ ein offenes Intervall und $y_1, \ldots, y_m$ Funktionen,

$$y_j : I \longrightarrow \mathbb{R} \, , \; 1 \leq i \leq m \, .$$

Seien $F_1, \ldots, F_m$ vorgeschriebene Funktionen,

$$F_k : I \times \prod_{k=1}^{m} \mathbb{R}^{r_k+1} \longrightarrow \mathbb{R} \, , \; 1 \leq k \leq m \, .$$

Die Funktionen y_j sind Lösungen des Systems von Differentialgleichungen

$$F_k = 0 \, , \; 1 \leq k \leq m \, ,$$

falls

$$\begin{aligned} F_k(t, y_1(t), Dy_1(t) \quad & , \ldots , \quad D^{r_1} y_1(t) \; , \; y_2(t), \ldots , \\ D^{r_2} y_2(t) \quad & , \ldots , \quad y_m(t), \ldots , D^{r_m} y_m(t)) = 0 \, , \quad \text{für} \quad t \in I \, , \; 1 \leq k \leq m \, , \end{aligned} \tag{4.1}$$

mit

$$D^j y(t) := \frac{d^j y(t)}{dt^j}$$

Die Variable t heißt *unabhängige Variable*, $r := \max \; r_k$ heißt *Ordnung* des Systems. Sind alle Funktionen F_k in $y_1, \ldots, D^{r_n} y_m$ linear, so heißt das System ein *lineares* System. Unter einem expliziten System versteht man ein System, das nach den Ableitungen der höchsten Ordnungen aufgelöst ist:

$$F_k(t, y_1, \ldots, D^{r_n} y_n) = D^{r_k} y_k + G_k \, ,$$

wobei G_k von der Ableitung $D^{r_k}y_k$ unabhängig ist.

Beispiel:

$$y' + 2ty = 0 \ , \ 0 < t < 1 \ .$$

Dies ist eine lineare Differentialgleichung erster Ordnung. Mit den Bezeichnungen von (4.1):

$$\begin{array}{rcl} m = 1 \ , \ r = 1 & , & I = (0,1) \ , \\ F_1(t,u,v) & = & v + 2tu \ . \end{array}$$

Beispiel:

$$\begin{array}{rcl} \ddot{x} & = & x + 2\dot{y} - \mu' \dfrac{x+\mu}{[(x+\mu)^2 + y^2]^{3/2}} - \mu \dfrac{x-\mu'}{[(x-\mu')^2 + y^2]^{3/2}} \ , \ t > 0 \\ \ddot{y} & = & y - 2\dot{x} - \mu' \dfrac{y}{[(x+\mu)^2 + y^2]^{3/2}} - \mu \dfrac{y}{[(x-\mu')^2 + y^2]^{3/2}} \ , \ t > 0 \end{array}$$

ist ein explizites System von zwei Differentialgleichungen zweiter Ordnung, wobei μ und μ' Konstanten sind. Dieses System beschreibt die Bewegung eines Satelliten, der sich im Erd-Mond-Kräftefeld bewegt. Die Koordinaten des Satelliten zum Zeitpunkt t sind $(x(t), y(t))$.

4.2 Modellierung

Differentialgleichungen sind wichtig, weil sie viele Anwendungen haben. Aus der großen Vielfalt wählen wir einige aus, die hoffentlich ohne physikalische Kenntnisse leicht verständlich sind.

4.2.1 Beispiel 1: Der freie Fall

Ein zunächst festgehaltener Körper wird zum Zeitpunkt $t = 0$ mit Geschwindigkeit v_0 losgelassen. Der weitere Verlauf dieses Vorgangs wird mathematisch durch das Newtonsche Gesetz beschrieben:

Masse $\cdot$ Beschleunigung $=$ Kraft
$m \cdot \ddot{s} = -m \cdot g$

mit

$$\begin{aligned} m &:= \text{Masse des Körpers} \\ g &:= \text{Erdbeschleunigung} = 981\text{cm}/sek^2 \\ s &:= \text{Höhe des Körpers über einem festgelegten Niveau} \end{aligned}$$

und Anfangsbedingungen

$$\begin{aligned} s(0) &= s_0 = \text{Anfangshöhe} \\ \dot{s} &= v_0 = \text{Anfangsgeschwindigkeit} \end{aligned}$$

4.2.2 Beispiel 2: Räuber-Beute Systeme

In der Natur gibt es mehrere Räuber-Beute Systeme, z.B. das Hase-Luchs System. Solche Systeme werden oft durch Differentialgleichungen modelliert, z.B.

$$\begin{aligned} \frac{db}{dt} &= \alpha b - \beta b r \\ \frac{dr}{dt} &= -\gamma r + \delta b r \end{aligned}$$

mit positiven Konstanten $\alpha, \beta, \gamma, \delta$. Dieses System kann sowohl quantitativ als auch qualitativ behandelt werden (siehe Abbildungen 4.1, 4.2 und 4.3).

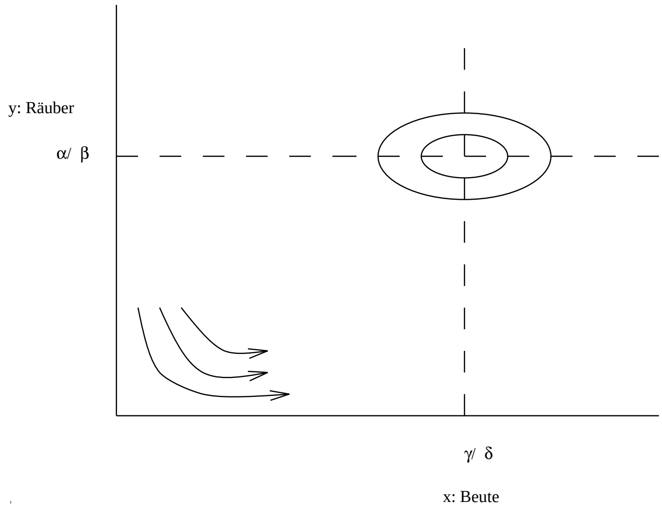


Abbildung 4.1: Ruhepunkte

4.2.3 Beispiel 3: Lineare elektrische Netzwerke

Eine elektrische Schaltung oder ein Netzwerk werden durch eine Menge K von *Knotenpunkten*, eine Menge Z von *Zweigen*, die die einzelnen Knotenpunkte verbinden, und die zugehörigen Induktivitäten L, Widerstände R und Kapazitäten C definiert.

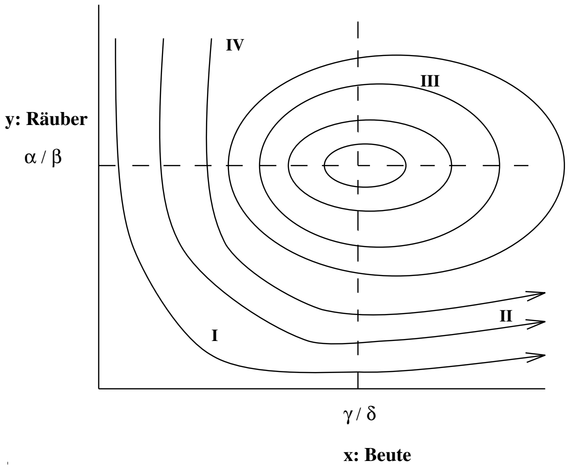


Abbildung 4.2: Geschlossene Bahnen

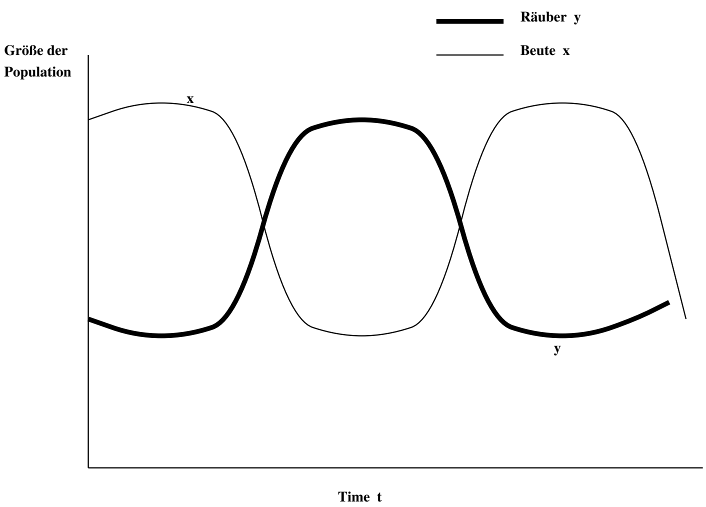


Abbildung 4.3: Zeitabhaengigkeit

Es gelten folgende Gesetze:

Erstes Kirchhoffsches Gesetz: Die algebraische Summe aller einem Knotenpunkt zufließenden Ströme ist gleich Null.

Zweites Kirchhoffsches Gesetz: In jedem (beliebig herausgegriffenen) in sich geschlossenen Kreis ist die Summe der Teilspannungen gleich der Summe der Stromquellen.

Zur Berechnung der Teilspannungen gilt folgendes:

1. **Widerstände** Sei I die Stromstärke, U die Spannung und R der Widerstand eines Leiters. Dann gilt(Ohmsches Gesetz):

$$U = R\,I$$

2. **Induktivität** Sei I die Stromstärke, U die Spannung und L die Induktivität eines

Leiters. Dann gilt:

$$U = L\frac{dI}{dt} = L\,\dot{I}$$

3. **Kapazität** Sei Q die Ladung, I die Stromstärke, U die Spannung und C die Kapazität eines Leiters. Es gilt:

$$\begin{aligned} U &= \frac{Q}{C} \\ \frac{dQ}{dt} &= \dot{Q} = I \end{aligned}$$

Wenn ein Netzwerk gegeben ist, kann die entsprechende Differentialgleichung mit Hilfe der beiden Kirchhoffschen Gesetze aufgestellt werden. Die folgende Methode führt aber zu wenigen Gleichungen.

Zuerst konstruiert man einen vollständigen Baum:

Definition 4.1 *Ein vollständiger Baum eines Netzwerkes ist ein Baum, wobei jeder Knotenpunkt nur einmal vorkommt. (Siehe Abbildung 4.4)*

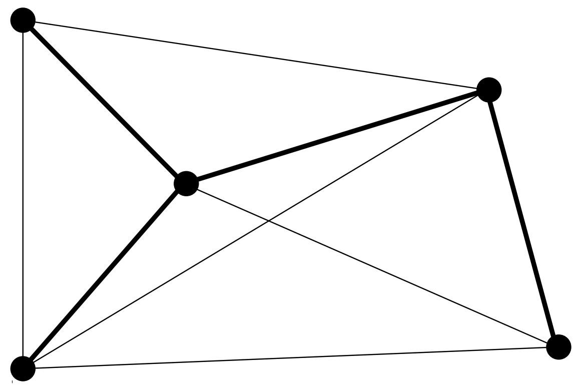
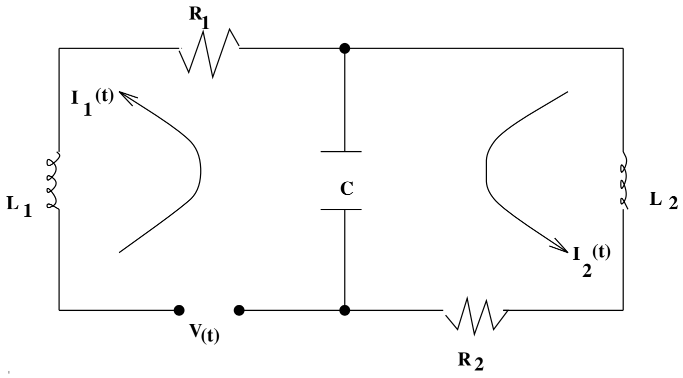


Abbildung 4.4: Jedem Zweig, der nicht im Baum vorkommt, ordnet man eine Stromstärke zu. Für jeden Zweig, der nicht im Baum vorkommt, gibt es genau einen Kreis, der aus diesem Zweig und Bestandteilen des Baumes besteht.

Beispiel

$$\begin{aligned} L_2\frac{dI_2}{dt} + R_2 I_2 + \frac{Q_1+Q_2}{C} &= O \\ L_1\frac{dI_1}{dt} + R_1 I_1 + \frac{Q_1+Q_2}{C} &= V(t) \\ \frac{dQ_1}{dt} &= I_1 \\ \frac{dQ_2}{dt} &= I_2 \end{aligned}$$



Literatur

Hort, W.: Die Differentialgleichungen des Ingenieurs. Julius Springer, Berlin, 1925.

Pipes, L.A.: Applied Mathematics for Engineers and Physicists. McGraw-Hill, New York, 1958.

Promberger, M.: Anwendung von Matrizen und Tensoren in der Theoretischen Elektrotechnik. Akademie-Verlag, Berlin, 1960.

(veraltet, aber in der math. Bibliothek)

4.3 Einige analytische Lösungsverfahren

4.3.1 Transformation in ein System von Differentialgleichungen erster Ordnung

Jedes System

$$F_k(t, y_1, \ldots, D^{r_m} y_m) = 0 \ , \ 1 \leq k \leq m \ , \tag{4.2}$$

kann in ein äquivalentes System

$$F(t, u, \dot{u}) = 0 \tag{4.3}$$

transformiert werden mit

$$\begin{aligned} u &: I \longrightarrow \mathbb{R}^N \\ N &= \sum_{k=1}^{m} r_k \\ F &: I \times \mathbb{R}^N \times \mathbb{R}^N \longrightarrow \mathbb{R}^N \ , \end{aligned}$$

indem man N Hilfsfunktionen

$$\begin{aligned} u_1 &:= y_1, \ldots, u_{r_1} := D^{r_1-1}y_1 \\ u_{r_1+1} &:= y_2, \ldots, u_{r_1+r_2} := D^{r_1-1}y_2 \\ u_{N-r_n+1} &:= y_n, \ldots, u_N := D^{r_n-1}y_n \end{aligned}$$

Durch die Einführung einer weiteren Variablen kann das System (4.3) in ein *autonomes* System

$$\begin{aligned} v &: I \longrightarrow \mathbb{R}^{N+1}, \\ G(v,\dot{v}) &= 0, \\ G &: \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \longrightarrow \mathbb{R}^{N+1} \end{aligned} \tag{4.4}$$

transformiert werden. Ist das System (4.2) explizit, so nehmen die Gleichungen (4.3) und (4.4) die folgende Gestalt an:

$$\begin{aligned} \dot{u} &= f(t,u), \\ \dot{v} &= g(v), \qquad \text{mit} \\ f &: I \times \mathbb{R}^N \longrightarrow \mathbb{R}^N, \\ g &: \mathbb{R}^{N+1} \longrightarrow \mathbb{R}^{N+1}. \end{aligned} \tag{4.5}$$

Diese Transformationen sind wichtig, da sie einheitliche numerische Verfahren erlauben: die meisten numerischen Verfahren sind der Gleichung (4.5) angepaßt.

4.3.2 Differentialgleichungen mit getrennten Veränderlichen

Die Differentialgleichung

$$\dot{y} = f(x)g(y)$$

kann durch folgende Schritte gelöst werden:

1. $$\frac{dy}{dx} = f(x)g(y)$$

2. $$\frac{dy}{g(y)} = f(x)dx$$

3. $$\int \frac{dy}{g(y)} = \int f(x)dx + C$$

   (C ist eine Integrationskonstante.)

4.

$$G(y) = F(x) + c$$

(G und F seien Stammfunktionen von $\frac{1}{g}$ bzw. f.)

5.

$$y(x) = \phi(F(x) + c)$$

wobei ϕ die Umkehrfunktion von G ist:

$$\phi(G(\alpha)) = \alpha \ . \tag{4.6}$$

Diese Berechnungen können wie folgt begründet werden, wobei vorausgesetzt wird, daß die auftretenden Integrale eigentliche Integrale sind:

$$\begin{aligned} y(x) &= \phi(F(x) + c) \\ \frac{dy}{dx} &= \left.\frac{d\phi(\alpha)}{d\alpha}\right|_{\alpha=F(x)+c} \cdot \frac{dF(x)}{dx} = \left.\frac{d\phi(\alpha)}{d\alpha}\right|_{\alpha=F(x)+c} \cdot f(x) \ . \end{aligned}$$

Aus der Identität (4.6) folgt nach Differentiation nach α

$$\left.\frac{d\phi}{d\xi}\right|_{\xi=G(\alpha)} \cdot \frac{dG(\alpha)}{d\alpha} = 1$$

oder

$$\left.\frac{d\phi}{d\xi}\right|_{\xi=G(\alpha)} \cdot \frac{1}{g(\alpha)} = 1 \ ,$$

so daß

$$\left.\frac{d\phi}{d\xi}\right|_{\xi=F(x)+c} = \left.\frac{d\phi}{d\xi}\right|_{\xi=G(y)} = g(y) \ .$$

Zusammenfassend:

$$\frac{dy}{dx} = \left.\frac{d\phi}{d\xi}\right|_{\xi=F(x)+c} \cdot f(x) = g(y) \cdot f(x)$$

wie erwünscht.

Beispiel (Ince, S. 5):

$$\frac{dy}{dx} = \frac{x(y^2-1)}{y(x^2-1)} = \left(\frac{x}{x^2-1}\right) \cdot \left(\frac{y^2-1}{y}\right) \ .$$

Die Lösung ergibt sich aus folgenden Schritten:

1.

$$dy \cdot \frac{y}{y^2-1} = dx \cdot \frac{x}{x^2-1}$$

2.
$$\int \frac{y}{y^2-1}\,dy = \int \frac{x}{x^2-1}\,dx$$

3.
$$\begin{aligned} \ell n|y^2-1| &= \ell n|x^2-1| + C \\ y^2-1 &= c(x^2-1)\ ,\ c = \pm e^C \end{aligned}$$

4.
$$y = \sqrt{1+c(x^2-1)}\ .$$

4.3.3 Homogene lineare Differentialgleichungen erster Ordnung mit konstanten Koeffizienten

Wir betrachten die Gleichung
$$\dot{y} = Ay$$
mit $A \in \mathrm{Mat}(n,n,\mathcal{C})$. Sei C eine reguläre Matrix, $z := C^{-1}y$ und $y = Cz$. Es folgt:
$$\dot{z} = C^{-1}\dot{y} = C^{-1}Ay = C^{-1}ACz = Bz \tag{4.7}$$
mit $B := C^{-1}AC$. Die Matrix C wird so gewählt, daß B die Jordansche Normalform hat:
$$B = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix}\ ,\quad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & O & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}\ ,$$
mit $J_i \in \mathrm{Mat}(n_i,n_i)$. Der Vektor $z(t)$ wird entsprechend partitioniert:
$$z(t) = (z^{(1)}(t),\dots,z^{(k)}(t))^T$$
mit
$$z^{(i)}(t) \in \mathbb{R}^{n_i}\ .$$
Das Gleichungssystem (4.7) nimmt dann eine vereinfachte Gestalt an
$$\dot{z}^{(i)}(t) = J_i z^{(i)}(t)\ ,\ 1 \le i \le k\ .$$
Es ist leicht festzustellen, daß das Gleichungssystem
$$\dot{w} = Jw\ ,$$
$$J := \begin{pmatrix} \lambda & 1 & O \\ & \ddots & 1 \\ O & & \lambda \end{pmatrix} \in \mathrm{Mat}(s,s)\ ,$$

$$\begin{aligned}
\dot{w}_1 &= \lambda w_1 + w_2 \\
&\vdots \\
\dot{w}_{s-1} &= \lambda w_{s-1} + w_s \\
\dot{w}_s &= \lambda w_s
\end{aligned}$$

die allgemeine Lösung

$$\begin{aligned}
w_s &= \alpha_s e^{\lambda t} \\
w_{s-1} &= (\alpha_{s-1} + \alpha_s t)e^{\lambda t} \\
\cdots \\
w_1 &= (\alpha_1 + \alpha_2 t + \cdots + \frac{\alpha_s t^{s-1}}{(s-1)!})e^{\lambda t}
\end{aligned} \tag{4.8}$$

hat. Die allgemeine Gestalt von $z^{(i)}$ ist aus (4.8) zu entnehmen.

Die Rechenarbeit wird oft erleichtert, wenn man den Ansatz

$$y = p(t)e^{\lambda t},$$

λ ein Eigenwert von A, macht.

Beispiel (Walter, S. 125):

$$\begin{aligned}
\dot{y}_1 &= y_1 - y_2 \ , \\
\dot{y}_2 &= 4y_1 - 3y_2 \ , \\
\dot{y} &= Ay \ ,
\end{aligned}$$

$$A = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix} .$$

Das charakteristische Polynom ist

$$p(\lambda) = \det(A - \lambda I) = \lambda^2 + 2\lambda + 1 = (\lambda + 1)^2$$

Das Gleichungssystem

$$Au = -u$$

hat die Lösung $u = (1,2)^T$, so daß

$$y(t) = \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-t}$$

eine Lösung der Differentialgleichung ist. Um eine zweite Lösung zu finden, wird der Ansatz

$$y = (a + bt)e^{-t}$$

gemacht. Es folgt

$$\begin{aligned} Ab &= -b \, , \\ Aa &= b - a \, . \end{aligned}$$

Es ergibt sich:

$$y = \begin{pmatrix} t \\ -1 + 2t \end{pmatrix} e^{-t} \, .$$

Die allgemeine Lösung ist:

$$y(t) = \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-t} + \beta \begin{pmatrix} 1 \\ -1 + 2t \end{pmatrix} e^{-t} \, ,$$

wobei α und β Konstanten sind.

4.3.4 Lineare inhomogene Differentialgleichungen erster Ordnung

$$\dot{y} + g(x)y = h(x) \, .$$

Schritt 1: Berechnung der allgemeinen Lösung der homogenen Gleichung
Die Gleichung

$$\dot{y} + g(x)y = 0$$

ist eine Gleichung mit getrennten Veränderlichen. Es folgt, daß

$$u(x) = e^{-G(x)} \, , \quad \text{mit} \quad G(x) = \int g(t)dt$$

eine Lösung ist.

Schritt 2: Methode der Variation der Konstanten
Man macht den Ansatz:

$$y(x) = c(x)e^{-G(x)} \, .$$

Es folgt

$$\dot{c}(x) = h(x)e^{G(x)}$$

und

$$c(x) = \int h(t)e^{G(t)}dt + c_0 \, .$$

Schritt 3: Zusammenfassung
Die allgemeine Lösung ist:

$$y(x) = ce^{-G(x)} + c(x)e^{-G(x)} \, .$$

Beispiel (Ince, S. 20):

$$x\dot{y} - (x+1)y = x^2 - x^3 \ .$$

Es folgt:

$$\dot{y} - \frac{x+1}{x}\, y = x - x^2 \ .$$

Schritt 1:

$$\begin{aligned} \frac{du}{u} &= \frac{x+1}{x}\, dx \\ \ell n\ u &= x + \ell n\ x \\ u &= xe^x \end{aligned}$$

Schritt 2:

$$\begin{aligned} y &= c(x)u(x) \\ \dot{y} &= \dot{c}(x)u(x) + c(x)\dot{u}(x) = \dot{c}(x)u(x) + c(x) \cdot \frac{x+1}{x} \cdot u(x) \ , \end{aligned}$$

so daß

$$\dot{y} - \frac{x+1}{x}\, y = \dot{c}(x)u(x) = x - x^2 \ .$$

Dies ist gleichbedeutend mit

$$\dot{c}(x) = \frac{(x-x^2)}{u(x)} = (1-x)e^{-x}$$

und

$$c(x) = xe^{-x} + c_0 \ .$$

Schritt 3:

$$y(x) = cu(x) + c(x)u(x) = cxe^x + x^2 \ .$$

4.3.5 Existenz und Eindeutigkeit des Anfangswertproblems

Der Hauptsatz lautet:

Satz 4.1 *Sei $a, b \in \mathbb{R}$ endlich und $n \in \mathbb{N}$. Sei*

$$S := I \times \mathbb{R}^n \ , \ I = [a,b] \subset \mathbb{R} \ .$$

Die Funktion $f : S \longrightarrow \mathbb{R}^n$ sei definiert und stetig. Weiter gebe es eine Konstante L, so daß

$$\|f(x,u) - f(x,v)\| \leq L\|u - v\|$$

für alle $x \in I$ und $u, v \in \mathbb{R}^n$.(f erfüllt also eine Lipschitzbedingung mit Konstante L.) Dann existiert zu jedem $x_0 \in [a,b]$ und jedem $y_0 \in \mathbb{R}^n$ genau eine Funktion $y(x)$ mit:

1. $y(x)$ ***ist stetig für*** $x \in [a, b]$ *,*

2. $y(x)$ ***ist stetig differenzierbar für*** $x \in (a, b)$ *,*

3.

$$\left.\frac{dy}{dx}\right|_{a+0} = \lim_{x \to a} \dot{y}(x) \quad \textbf{\textit{und}} \quad \left.\frac{dy}{dx}\right|_{b-0} = \lim_{x \to b} \dot{y}(x)$$

4. $\dot{y}(x) = f(x, y(x))$ ***für*** $x \in [a, b]$

5. $y(x_0) = y_0$ *.*

4.3.6 Randwertaufgaben

Sei $y(x)$ eine Lösung der Gleichung

$$\dot{y}(x) = f(x, y(x)) \; .$$

Es ist oft physikalisch sinnvoll, folgende Randwertaufgabe zu betrachten:

Randwertaufgabe: Sei $a, b \in \mathbb{R}^n$, $a < b$. Sei $A, B \in \mathrm{Mat}(n, 2n, \mathcal{C})$ und $c \in \mathbb{R}^n$ gegeben. Bestimme $y(x)$, so daß:

1. $\dot{y}(x) = f(x, y(x))$, $a < x < b$
2. $A\binom{y(a)}{y'(a)} + B\binom{y(b)}{y'(b)} = c$.

Beispiel: (Freier Fall)

$$\begin{aligned} \ddot{y}(t) &= -g \\ y(0) &= 0 \\ \dot{y}(1) &= 1 \end{aligned}$$

Lösung

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -g \end{aligned}$$

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \quad B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} ,$$

$$c = (0, 1)^T$$

Die allgemeine Lösung der Gleichung ist:

$$y(t) = \alpha + \beta t - \frac{1}{2}\, g t^2$$

Die Bedingungen

$$y(0) = 0 \ , \ \dot{y}(1) = 1$$

werden erfüllt, wenn

$$\begin{aligned} y(0) &= \alpha = 0 \ , \\ \dot{y}(1) &= \beta - g = 1 \ . \end{aligned}$$

Es folgt $\alpha = 0$, $\beta = 1 + g$ und

$$y(t) = (1+g)t - \frac{1}{2}\, gt^2 \ .$$

Die Existenz- und Eindeutigkeitstheorie für Randwertaufgaben ist verwickelter als die für Anfangswertaufgaben, und wir verzichten auf Einzelheiten.

Die numerische Lösung von Randwertproblemen führt zur Lösung großer linearer Gleichungssysteme.

Kapitel 5

Numerische Methoden für Anfangswertaufgaben

5.1 Einleitung

Wir betrachten das Anfangswertproblem

$$\begin{aligned} \dot{y} &= f(x,y) \ , \\ y(x_0) &= y_0 \end{aligned}$$

mit

$$\begin{aligned} y(t) &\in \mathbb{R}^n \ , \ y_0 \in \mathbb{R}^n \ , \\ f &: I \times \mathbb{R}^n \longrightarrow \mathbb{R}^n \ , \\ x_0 &\in I \end{aligned}$$

Die Existenz und Eindeutigkeit der Lösung wird vorausgesetzt.

Es gibt zwei allgemeine Klassen von Lösungsverfahren: Einschritt- und lineare Mehrschrittverfahren, die zuerst kurz und anschließend ausführlicher beschrieben werden.

Beide Klassen von Verfahren gehen davon aus, daß die exakte Lösung $y(x)$ in einer Folge $\{x_i\}$

$$x_0 < x_1 < x_2 < \cdots$$

von Stützstellen approximiert wird, wobei die Approximation zu $y(x_i)$ gelegentlich durch $\tilde{y}(x_i)$, $y^h(x_i)$, y_i^h , y_i usw. bezeichnet wird.

Grundlegende Fragen sind:

1. Wie wird $\tilde{y}(x_i)$ definiert und berechnet?

2. Wie verhält sich der Fehler $e(x_i) := y(x_i) - \tilde{y}(x_i)$?

Einschritt- und lineare Mehrschrittverfahren unterscheiden sich darin, wie ihre Bezeichnungen verdeutlichen, daß bei Einschrittverfahren $\tilde{y}(x_{i+1})$ nur von einem Wert $\tilde{y}(x_i)$ abhängig ist, während bei einem r-Schritt linearer Mehrschrittverfahren $\tilde{y}(x_{i+r})$ von den r Werten $\tilde{y}(x_{i+r-1})$ bis $\tilde{y}(x_i)$ abhängt.

Einschrittverfahren: Ein Einschrittverfahren hat die Gestalt:

$$\tilde{y}(x_{i+1}) = \tilde{y}(x_i) + h_i \Phi(x_i, \tilde{y}(x_i), h_i; f)$$

mit

$$h_i := x_{i+1} - x_i \ .$$

Lineare Mehrschrittverfahren: Ein lineares Mehrschrittverfahren hat die Gestalt:

$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}) = h \sum_{k=0}^{r} b_k f(x_{i+k}, \tilde{y}(x_{i+k})) \ ,$$

wobei a_k, b_k vorgeschriebene Konstanten sind und

$$x_{i+k} := x_i + kh$$

und h die Schrittweite heißt.

5.2 Lineare Mehrschrittverfahren

Zur Lösung des Anfangswertproblems

$$\begin{aligned} \dot{y} &= f(x, y) \\ y(x_0) &= y_0 \end{aligned}$$

wird folgender Ansatz gemacht:

1. Eine Schrittweite $h > 0$ wird gewählt und die Stützpunkte

   $$x_i := x_i(h) := x_0 + ih \ , \ i \in \mathbb{N}$$

   entsprechend definiert.

2. Die Approximation $\tilde{y}(x_i, h)$, $0 \leq i \leq r-1$ wird berechnet. Es liegt nahe,

   $$\tilde{y}(x_0, h) = y_0$$

   zu setzen. Die übrigen Werte $\tilde{y}(x_i, h)$, $1 \leq i \leq r-1$ können z.B. mit Hilfe eines Einschritt- oder Mehrschrittverfahrens niedriger Ordnung berechnet werden.

3. Die Approximationen $\tilde{y}(x_{i+r}, h)$, $i \geq 0$ werden berechnet:
$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}, h) = h \sum_{k=0}^{r} b_k f(x_{i+k}, \tilde{y}(x_{i+k}, h)) \ , \ i \geq 0 \ .$$

Ein lineares Mehrschrittverfahren wird eindeutig durch die Konstanten a_k, b_k bestimmt. Es ist oft sehr nützlich, die Polynome

$$\begin{aligned} \rho(z) &:= \sum_{k=0}^{r} a_k z^k \\ \sigma(z) &:= \sum_{k=0}^{r} b_k z^k \end{aligned}$$

einzuführen, die ebenfalls das lineare Mehrschrittverfahren eindeutig bestimmen.

5.2.1 Herleitung von Linearen Mehrschrittverfahren durch Integration

Für nichtnegative ganze Zahlen m, k, q ist

$$y(x_{p+k}) - y(x_{p-m}) = \int_{x_{p-m}}^{x_{p+k}} f(x, y(x))dx \doteq \int_{x_{p-m}}^{x_{p+k}} P_q(x)dx = h \sum_{i=0}^{q} \beta_{q,i} \, f_{p-i}$$

wobei $P_q(x)$ das interpolierende Polynom ist:

1. Grad $P_q \leq q$
2. $P_q(x_{p-i}) = f_{p-i} := f(x_{p-i}, y(x_{p-i}))$, $0 \leq i \leq q$.

Unter Benutzung der Lagrangeschen Interpolationsformel

$$P_q(x) = \sum_{i=0}^{q} f_{p-i} L_i(x) \ , \ L_i(x) = \prod_{\substack{t=0 \\ t \neq i}}^{q} \frac{x - x_{p-t}}{x_{p-i} - x_{p-t}}$$

erhält man:

$$\begin{aligned} \beta_{q,i} &= \frac{1}{h} \int_{x_{p-m}}^{x_{p+k}} L_i(x)dx \\ &= \frac{1}{h} \int_{x_{p-m}}^{x_{p+k}} \prod_{\substack{t=0 \\ t \neq i}}^{q} \frac{x - x_{p-t}}{x_{p-i} - x_{p-t}} dx \ , \\ &= \int_{-m}^{+k} \prod_{\substack{t=0 \\ t \neq i}}^{q} \left(\frac{s+t}{-i+t} \right) ds. \ \ (x = x_p + sh) \end{aligned}$$

Das entsprechende lineare Mehrschrittverfahren ist:

$$\tilde{y}(x_{p+k}) - \tilde{y}(x_{p-m}) = h \sum_{i=0}^{q} \beta_{q,i} \, \tilde{f}_{p-i}$$

mit

$$\tilde{f}_{p-i} := f(x_{p-i}, \tilde{y}(x_{p-i})) \; .$$

Dies ist ein r-Schritt lineares Mehrschrittverfahren, wobei

$$r := k + \max(m, q) \; .$$

Beispiel: Die Adams-Bashforth Formeln: $k = 1, m = 0, q = 0, 1, 2, \ldots$. Z.B., falls $q = 2$,

$$y(x_{p+1}) - y(x_p) = \int\limits_{x_p}^{x_{p+1}} f(x, y(x))dx \; .$$

$P \equiv P_2$ muß die Bedingungen erfüllen:

$$\begin{aligned} P(x_p) &= f_p \; , \\ P(x_{p-1}) &= f_{p-1} \; , \\ P(x_{p-2}) &= f_{p-2} \; . \end{aligned}$$

Mit $x = x_p + sh$ sei

$$\begin{aligned} L_0(x) &= \frac{(x - x_{p-1})(x - x_{p-2})}{(x_p - x_{p-1})(x_p - x_{p-2})} = \frac{(s+1)(s+2)}{2} =: \ell_0(s) \\ L_1(x) &= \frac{(x - x_p)(x - x_{p-2})}{(x_{p-1} - x_p)(x_{p-1} - x_{p-2})} = \frac{s(s+2)}{(-1)} =: \ell_1(s) \\ L_2(x) &= \frac{(x - x_p)(x - x_{p-1})}{(x_{p-2} - x_p)(x_{p-2} - x_{p-1})} = \frac{s(s+1)}{2} =: \ell_2(s) \end{aligned}$$

Es folgt

$$P(x) = \sum_{i=0}^{2} f_{p-i} L_i(x)$$

und

$$y(x_{p+1}) - y(x_p) = \int\limits_{x_p}^{x_{p+1}} f(x, y(x))dx \doteq \int\limits_{x_p}^{x_{p+1}} P(x)dx = h \int\limits_0^1 \sum_{i=0}^{2} f_{p-i}\ell_i(s)ds \; .$$

Da

$$\begin{aligned}
\int_0^1 \ell_0(s)ds &= \int_0^1 \frac{(s+1)(s+2)}{2}\, ds = \frac{23}{12} \\
\int_0^1 \ell_1(s)ds &= \int_0^1 \frac{s(s+2)}{(-1)}\, ds = -\frac{16}{12} \\
\int_0^1 \ell_2(s)ds &= \int_0^1 \frac{s(s+1)}{2}\, ds = \frac{5}{12}
\end{aligned}$$

ergibt sich

$$y(x_{p+1}) - y(x_p) \doteq \frac{h}{12}\left(23 f_p - 16 f_{p-1} + 5 f_{p-2}\right) .$$

Das entsprechende lineare Mehrschrittverfahren ist:

$$\tilde{y}_{p+3} - \tilde{y}_{p+2} = \frac{h}{12}\left(23\tilde{f}_{p+2} - 16\tilde{f}_{p+1} + 5\tilde{f}_p\right) .$$

Beispiel: Die Adams-Moulton Formeln: $k = 0, m = 1, q = 0, 1, 2, \ldots$. Z.B. für $q = 2$:

$$\tilde{y}_{p+3} - \tilde{y}_{p+2} = \frac{h}{12}\left(5\tilde{f}_{p+3} + 8\tilde{f}_{p+2} - \tilde{f}_p\right) .$$

5.2.2 Ein numerisches Beispiel

Wird später beigefügt.

5.2.3 Theorie der linearen Mehrschrittverfahren

Definition 5.1 *Das lineare Mehrschrittverfahren*

$$\sum_{k=0}^{r} a_k \tilde{y}_{i+k} = h \sum_{k=0}^{r} b_k \tilde{f}_{i+k}$$

heißt konvergent*, falls:*
Für jedes *Anfangswertproblem*

$$\begin{aligned}
\dot{y}(x) &= f(x, y(x)) \ , \ a \le x \le b \\
y(a) &= y_0 \ ,
\end{aligned}$$

mit $f : [a,b] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ *stetig und Lipschitz stetig bezüglich* y, *gilt:*
Sei $\tilde{y}(x_i, h)$, $0 \leq i < r$ *vorgeschrieben mit*

$$\lim_{h \to 0} \tilde{y}(x_i, h) = y_0 \ , \ 0 \leq i < r \ .$$

Sei $\epsilon(x_i, h)$ *gegeben mit*

$$|\epsilon(x, h)| \leq \psi(h)$$

wo

$$\lim_{h \to 0} \psi(h) = 0 \ .$$

Sei

$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}, h) = h \sum_{k=0}^{r} b_k f(x_{i+k}, \tilde{y}(x_{i+k}, h)) + h\epsilon(x_{i+k}, h) \ .$$

Dann gilt:

$$\lim_{\substack{h \to 0 \\ x_n = x}} \tilde{y}(x_n, h) = y(x) \ , \quad \textit{für} \quad a \leq x \leq b \ .$$

Bemerkung:

1. Der Term $h\epsilon(x_{i+k}, h)$ stellt den Rundungsfehler dar.
2. Die Anfangswerte $\tilde{y}(x_i, h)$, $0 \leq i < r$ sind i.a. nicht exakt, da die exakte Lösung $y(x)$ i.a. nicht bekannt ist.

Definition 5.2 *Das lineare Mehrschrittverfahren* (ρ, σ) *erfüllt die* Stabilitätsbedingung, *falls:*

1. *Für jede Nullstelle* λ *von* $\rho(\lambda)$ *gilt* $|\lambda| \leq 1$.
2. *Aus* $|\lambda| = 1$ *und* $\rho(\lambda) = 0$ *folgt, daß* λ *nur einfache Nullstelle von* ρ *ist.*

Beispiel: Für die r-Schritt Adams-Bashforth und Adams-Moulton Methoden ist

$$\rho(z) = z^r - z^{r-1} = z^{r-1}(z-1) \ ,$$

so daß diese Methoden stabil sind.

Satz 5.1 *Wenn das lineare Mehrschrittverfahren* (ρ, σ) *konvergiert, so ist die Stabilitätsbedingung erfüllt.*

Beweis: Der Beweis erfolgt durch Widerspruch. Sei (ρ, σ) ein r-Schritt lineares Mehrschrittverfahren, das die Stabilitätsbedingung nicht erfüllt. Dann gibt es ein $z_1 \in \mathbb{C}$ mit $\rho(z_1) = 0$ und entweder

$$\text{a)} \qquad |z_1| > 1$$

oder

$$\text{b)} \qquad |z_1| = 1 \quad \text{und} \quad \rho'(z_1) = 0 \; .$$

z_1 ist also eine mehrfache Nullstelle von ρ. □

Wir betrachten nun das Anfangswertproblem

$$\begin{aligned} y'(x) &= 0 \; , \; 0 \le x \le 1 \\ y(0) = y_0 &= 0 \end{aligned}$$

mit der Lösung $y(x) \equiv 0$. Zur Lösung dieses Problems wird das lineare Mehrschrittverfahren (ρ, σ) eingesetzt mit

$$\begin{aligned} h &:= 1/n \; , \\ \epsilon(x,h) &:= 0 \; , \\ \tilde{y}(x_i,h) &:= hu_i \; , \; 0 \le i < r \; , \end{aligned}$$

wobei die Konstanten u_i noch festzulegen sind. Es gilt

$$\begin{aligned} \psi(h) &= 0 \; , \\ \lim_{h\to 0} \tilde{y}(x_i,h) &= y_0 \; . \end{aligned}$$

Die Folge $\{\tilde{y}(x_i,h)\}$ erfüllt die Differenzengleichung

$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k},h) = h \sum_{k=0}^{r} b_k \tilde{f}_{i+k} = 0$$

Sei $z_1, \ldots, z_m$ die Nullstelle von $\rho(z)$ mit den Vielfachheiten $\sigma_1, \ldots, \sigma_m$. Aus der Theorie der Differenzengleichungen bilden die Folgen

$$\begin{aligned} v_i^{(k,s)} &:= i(i-1)\ldots(i-s+1) z_k^{i-s} \; , \\ 0 \le s &< \sigma_k \; , \\ 1 \le k &\le m \end{aligned}$$

ein System von r linear unabhängigen Lösungen.

Fall a: Sei

$$u_i := v_i^{(1,0)} = z_1^i \; , \; 0 \le i < r \; .$$

Dann ist

$$\tilde{y}(x_i,h) = h \; z_1^i$$

und

$$\tilde{y}(1,h) = h \; z_1^n = \frac{1}{n} \; z_1^n \; .$$

Fall b: Sei

$$u_i := v_i^{(1,1)} = i \; z_1^{i-1} \; , \; 0 \le i < r \; .$$

Dann ist

$$\tilde{y}(x_i, h) = hi\ z_1^{i-1}$$

und

$$\tilde{y}(1, h) = z_1^{n-1} .$$

In beiden Fällen wird die Bedingung

$$\lim_{h \to 0} \tilde{y}(1, h) = y(1) = 0$$

nicht erfüllt.

Die bisherigen Überlegungen haben sich nur auf das Polynom $\rho(z)$ bezogen, was für die Konvergenz nicht hinreichend sein kann.

Sei (ρ, σ) ein lineares Mehrschrittverfahren. Sei

$$L_h : C^\infty(\mathbb{R}) \longrightarrow C^\infty(\mathbb{R})$$

mit

$$(L_h y)(x) := \sum_{k=0}^{r} a_k y(x + kh) - h \sum_{k=0}^{r} b_k \dot{y}(x + kh) .$$

Die Funktionen $y(x+kh)$, $\dot{y}(x+kh)$ haben Taylor-Reihen mit Entwicklungspunkt $h = 0$:

$$\begin{aligned} y(x + kh) &= \sum_{j=0}^{\infty} y^{(j)}(x) \cdot \frac{(kh)^j}{j!} \\ \dot{y}(x + kh) &= \sum_{j=0}^{\infty} y^{(j+1)}(x) \cdot \frac{(kh)^j}{j!} \end{aligned}$$

Es folgt:

$$(L_h y)(x) = \sum_{j=0}^{\infty} C_j y^j(x) h^j$$

mit

$$\begin{aligned} C_0 &= \sum_{k=0}^{r} a_k = \rho(1) \\ C_1 &= \sum_{k=0}^{r} k a_k - \sum_{k=0}^{r} b_k = \dot{\rho}(1) - \sigma(1) \\ C_j &= \frac{1}{j!} \cdot \sum_{k=0}^{r} a_k k^j - \frac{1}{(j-1)!} \cdot \sum_{k=0}^{r} b_k k^{j-1} \ , \ j \geq 1 . \end{aligned}$$

Definition 5.3 *Das lineare Mehrschrittverfahren* (ρ, σ) *ist von der Ordnung* p *, wenn für den entsprechenden Operator* L *gilt:*

$$\begin{aligned} C_j &= 0 \, , \; 0 \le j \le p \, , \\ C_{p+1} &\neq 0 \, . \end{aligned}$$

Das Verfahren (ρ, σ) *heißt* konsistent, *wenn das Verfahren mindestens die Ordnung 1 hat.*

Beispiel: Für das Adams-Bashforth 3-Schritt Verfahren

$$\tilde{y}_{p+3} - \tilde{y}_{p+2} = \frac{h}{12} \left(23 \tilde{f}_{p+2} - 16 \tilde{f}_{p+1} + 5 \tilde{f}_p \right)$$

gilt:

$$\begin{aligned} \rho(z) &= z^3 - z^2 \\ \sigma(z) &= \frac{1}{12} \left(23 z^2 - 16 z + 5 \right) . \end{aligned}$$

Es gilt:

$$\begin{aligned} C_0 &= (1-1) = 0 \\ 1! C_1 &= (3-2) - \frac{1}{12} (23 - 16 + 5) = 0 \\ 2! C_2 &= (3^2 - 2^2) - \frac{2}{12} (23.2 - 16.1 + 5.0) = 0 \\ 3! C_3 &= (3^3 - 2^3) - \frac{3}{12} (23.2^2 - 16.1^2 + 5.0^2) = 0 \\ 4! C_4 &= (3^4 - 2^4) - \frac{4}{12} (23.2^3 - 16.1^3 + 5.0^3) = 9 \, . \end{aligned}$$

Die Methode hat deshalb die Ordnung $p = 3$ und $C_4 = \frac{3}{8} \neq 0$.

Satz 5.2 *Wenn das lineare Mehrschrittverfahren* (ρ, σ) *konvergiert, so ist es konsistent:*

$$\begin{aligned} &a) \quad \rho(1) = 0 \\ &b) \quad \dot{\rho}(1) = \sigma(1) \, . \end{aligned}$$

Beweis:

a) Wir betrachten das Anfangswertproblem

$$\begin{aligned} \dot{y}(x) &= 0 \, , \; 0 \le x \le 1 \\ y(0) &= 1 \, . \end{aligned}$$

Als Approximation nehmen wir:

$$\begin{aligned} \tilde{\eta}(x_i, h) &= 1 \ , \ 0 \le i < r \\ \sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}, h) &= 0 \ , \ i \ge 0 \ . \end{aligned}$$

Da die Methode konvergent ist, gilt:

$$\lim_{\substack{h \to 0 \\ x_n = nh = x}} \tilde{\eta}(x_n, h) = y(x) = 1 \quad \text{für} \quad x \ge 0 \ .$$

Sei $\{u_i\}$ die Lösung des Problems

$$\begin{aligned} u_i &= 1 \ , \ 0 \le i < r \\ \sum_{k=0}^{r} a_k u_{i+k} &= 0 \ , \ i \ge 0 \ . \end{aligned}$$

Dann gilt:

$$\tilde{\eta}(x_{i+k}, h) = u_{i+k} \ .$$

Insbesondere

$$u_n = \tilde{\eta}(1, 1/n) \ , \ n > r \ .$$

Da

$$\tilde{\eta}(1, 1/n) \longrightarrow y(1) = 1 \quad \text{für} \quad n \to \infty \ ,$$

folgt

$$u_n \longrightarrow 1 \quad \text{für} \quad n \to \infty \ .$$

Es folgt

$$0 = \lim_{i \to \infty} \sum_{k=0}^{r} a_k u_{k+i} = \sum_{k=0}^{r} a_k \lim_{i \to \infty} u_{k+i} = \sum_{k=0}^{r} a_k = \rho(1) \ .$$

b) Wir betrachten das Anfangswertproblem

$$\begin{aligned} \dot{y}(x) &= 1 \ , \ x \ge 0 \\ y(0) &= 0 \end{aligned}$$

mit der exakten Lösung $y(x) = x$ und die Differenzengleichung

$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}, h) = h \sum_{k=0}^{r} b_k \tag{5.1}$$

Zuerst wird eine Lösung der Differenzengleichung der Gestalt

$$\tilde{y}(x_i, h) = Khi$$

gesucht. Es gilt:

$$\sum_{k=0}^{r} a_k Kh(i+k) = Kh \sum_{k=0}^{r} (a_k k + a_k i) = Kh(\dot\rho(1) + i\rho(1))$$

und

$$h \sum_{k=0}^{r} b_k = h\sigma(1) \; .$$

Wie schon in Teil a) bewiesen, gilt $\rho(1) = 0$. Aus Satz 5.2 folgt, daß $\dot\rho(1) \neq 0$. Wir können folgern, daß

$$\tilde y(x_i, h) := Khi$$

eine Lösung der Gleichung (5.1) ist, wenn

$$K = \frac{\sigma(1)}{\dot\rho(1)} \; .$$

Wird $\epsilon(x_i, h) = 0$ und

$$\tilde y(x_i, h) = Khi \; , \; 0 \le i < r$$

gesetzt, dann gilt

$$\lim_{h\to 0} \tilde y(x_i, h) = 0 = y(0) \; ,$$

so daß die Voraussetzungen in der Definition von Konvergenz erfüllt sind. Es folgt:

$$\lim_{h\to 0} \tilde y(1, h) = K = y(1) = 1 \; .$$

□

Satz 5.3 *Das lineare Mehrschrittverfahren (ρ, σ) ist genau dann konvergent, wenn es konsistent und stabil ist.*

Beweis: Siehe z.B. Henrici. □

Obwohl jedes konvergente lineare Mehrschrittverfahren eine beliebig genaue Lösung erzeugen kann, ist es wichtig, Methoden höherer Ordnung zu benutzen, wie der folgende Satz verdeutlicht:

Satz 5.4 *Sei (ρ, σ) ein konvergentes lineares Mehrschrittverfahren der Ordnung p. Sei $f : I \times \mathbb{R}^n \to \mathbb{R}^n$ $(p+1)$-mal stetig differenzierbar. Sei $y(x)$ die Lösung des Anfangswertproblems*

$$\begin{aligned} \dot y(x) &= f(x, y(x)) \; , \; x \in I \; , \\ y(a) &= y_0 \; . \end{aligned}$$

Sei $\tilde{\eta}(x_i, h)$ die berechnete Approximation

$$\sum_{k=0}^{r} a_k \tilde{y}(x_{i+k}, h) = h \sum_{k=0}^{r} b_k f(x_{i+k}, \tilde{y}(x_{i+k}, h))$$

mit

$$\tilde{y}(x_i, h) - y(x_i) = O(h^q)\ ,\ 0 \leq i < r\ .$$

Sei

$$s := \min(p, q)\ .$$

Dann gibt es eine stetige Funktion $e(x)$, so daß

$$\tilde{y}(x, h) = y(x) + h^s e(x) + O(h^{s+1})\ .$$

Beweis: Siehe z.B. Henrici, S. 249. □

Bemerkung: Globale Ordnung $=$ Lokale Ordnung $= 1$.

Es folgt aus Satz 5.4, daß es praktisch wichtig ist, lineare Mehrschrittverfahren höherer Ordnung zu finden.

Satz 5.5 *Das lineare Mehrschrittverfahren (ρ, σ) ist genau dann ein Verfahren p-ter Ordnung, wenn die Funktion*

$$\psi(\mu) = \frac{\rho(\mu)}{\ln(\mu)} - \sigma(\mu)$$

die Zahl $\mu = 1$ als p-fache Nullstelle besitzt.

Beweis: Wenn das Verfahren (ρ, σ) die Ordnung p hat, dann gilt

$$\begin{aligned}(L_h y)(x) &= \sum_{k=0}^{r} a_k y(x + kh) - h \sum_{k=0}^{r} b_k \dot{y}(x + kh) \\ &= C_{p+1} h^{p+1} y^{(p+1)}(x) + O(h^{p+2}) \qquad (5.2)\end{aligned}$$

mit

$$C_{p+1} \neq 0 \quad \text{für alle} \quad y \in C^\infty\ .$$

Für $y(x) = e^x$ gilt

$$(L_h y)(x) = [\rho(e^h) - h\sigma(e^h)]e^x\ . \qquad (5.3)$$

Es folgt aus (5.2) und (5.3), daß das Verfahren (ρ, σ) die Ordnung p mit Konstante C_{p+1} hat, genau dann wenn

$$\rho(e^h) - h\sigma(e^h) = C_{p+1} h^{p+1} + O(h^{p+2})$$

mit $C_{p+1} \neq 0$.
Der Satz folgt mit Hilfe der Substitution

$$\mu := e^h ,$$

da für μ in einer kleinen Umgebung von 1

$$\ln\ \mu = \ln\ 1 + (\mu - 1) + O(\mu - 1)^2 = (\mu - 1) + O(\mu - 1)^2$$

□

Beispiel: Für das Euler-Verfahren gilt:

$$\rho(z) = z - 1, \quad \sigma(z) = 1, \quad \psi(\mu) = \frac{\rho(\mu)}{\ln\ \mu} - \sigma(\mu) = \frac{\mu - 1}{\ln\ \mu} - 1\ .$$

Die Taylor-Reihe von $\ln\ \mu$ im Punkte $\mu = 1$ ist:

$$\ln\ \mu = \frac{(\mu - 1)}{1} - \frac{(\mu - 1)^2}{2} + \frac{(\mu - 1)^3}{3} + \cdots$$

Es folgt:

$$\begin{aligned}
\frac{\mu - 1}{\ln\ \mu} &= \frac{\mu - 1}{(\mu - 1)\left(1 - \left[\frac{(\mu-1)}{2} - \frac{(\mu-1)^2}{3} \cdots\right]\right)} \\
&= 1 + \left[\frac{\mu - 1}{2} - \frac{(\mu - 1)^2}{3} \cdots\right] - [\ ]^2 + [\ ]^2 \cdots \\
&= 1 + \left(\frac{\mu - 1}{2} - \frac{(\mu - 1)^2}{3}\right) - \left(\frac{\mu - 1}{2}\right)^2 + O((\mu - 1)^3)
\end{aligned}$$

und

$$\begin{aligned}
\psi(\mu) &= 1 + \frac{\mu - 1}{2} - \frac{7}{12}(\mu - 1)^2 - 1 + O((\mu - 1)^3) \\
&= \frac{\mu - 1}{2} + O((\mu - 1)^2)\ .
\end{aligned}$$

Das Euler-Verfahren hat deshalb die Ordnung $p = 1$ mit der Konstante $C_2 = \frac{1}{2}$.

Beispiel:

$$\rho(z) = (z - 1)(z - \lambda).$$

Sei

$$s := \mu - 1$$

Dann folgt:

$$\begin{aligned}
\frac{\rho(\mu)}{\ln\ \mu} &= \frac{s(s + 1 - \lambda)}{\ln(1 + s)} = \frac{s(s + 1 - \lambda)}{s\left(1 - \frac{s}{2} + \frac{s^2}{3} - \frac{s^3}{4} \cdots\right)} \\
&= (1 - \lambda) + \frac{3 - \lambda}{2}\, s + \frac{5 + \lambda}{12}\, s^2 - \frac{1 + \lambda}{24}\, s^3 + O(s^4)
\end{aligned}$$

wobei benutzt worden ist, daß

$$\frac{1}{\ln(1+s)} = \frac{1}{s}\left(1+\frac{s}{2}-\frac{s^2}{12}+\frac{s^3}{24}-\frac{19s^4}{720}\cdots\right).$$

Die maximale Ordnung 3 wird erreicht, falls

$$\begin{aligned}\sigma(1+s) &= (1-\lambda)+\frac{3-\lambda}{2}\,s+\frac{5+\lambda}{12}\,s^2\,, \quad \text{d.h.}\\ \sigma(\mu) &= 1-\lambda+\frac{3-\lambda}{2}\,(\mu-1)+\frac{5+\lambda}{12}\,(\mu-1)^2\,.\end{aligned}$$

Satz 5.6 *(Dahlquist) Ein lineares r-Schritt-Mehrschrittverfahren, welches die Stabilitätsbedingung erfüllt, besitzt eine Ordnung p mit*

$$p \leq \begin{cases} r+1 & \text{falls } r \text{ ungerade} \\ r+2 & \text{falls } r \text{ gerade}\end{cases}$$

Falls $p = r+2$ mit r gerade ist, liegt jede Nullstelle von ρ am Einheitskreis.

Beweis: Siehe Henrici S. 231 und S. 232. □

Literatur

Brenan, K.E., **Campbell**, S.L., **Pethhold**, L.R.: Numerical Solution of Initial-Value Problems for Differential- Algebraic Equations. North-Holland, 1989.

Butcher, J.C.: The numerical Analysis of Ordinary Differential Equations. Wiley, 1987.

Fatunia, S.O.: Numerical Methods for Initial Value Problems in Ordinary Differential Equations. Academic Press, 1987.

Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen, 1. Einschrittverfahren. Teubner, 1972.

Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen, 2. Mehrschrittverfahren. Teubner, 1977.

Hairer, E., **Norsett**, S.P., **Wanner**, G.: Solving Ordinary Differential Equations I. Nonstiff Problems. Springer, 1987.

Hairer, E. and **Wanner**, G.: Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Equations. Springer, 1991.

Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. Wiley, 1962.

Lambert, J.D.: Computational Methods in Ordinary Differential Equations. Wiley, 1973.

5.2.4 Grundbegriffe

Wir betrachten das Anfangswertproblem

$$\dot{y}(x) = f(x, y(x)) \ , \ x > x_0; \qquad y(x_0) = y_0 \tag{5.4}$$

Näherungswerte $\tilde{y}(x_i)$ erhält man durch

$$\tilde{y}(x_{i+1}; h) = \tilde{y}(x_i; h) + h_i \Phi(x_i, \tilde{y}(x_i; h), h_i; f); \qquad \tilde{y}(x_0) = y_0 \tag{5.5}$$

Beispiel: Das Euler-Verfahren:

$$\Phi(x, y, h; f) := f(x, y) \ ,$$

so daß

$$\begin{aligned} \tilde{y}(x_0) &= y_0 \\ \tilde{y}(x_{i+1}) &= \tilde{y}(x_i) + h_i f(x_i, \tilde{y}(x_i)) \ . \end{aligned}$$

Sei nun x und z_0 fest gewählt und $z(t)$ die exakte Lösung des Anfangswertproblems

$$\dot{z}(t) = f(t, z(t)) \ , \ t > x; \qquad z(x) = z_0$$

Sei

$$\Delta(x, z_0, h; f) := \begin{cases} \frac{z(x+h) - z_0}{h} \ , & h > 0 \\ f(x, z_0) \ , & h = 0 \end{cases}$$

Es gilt:

$$z(x + h) = z_0 + h\Delta(x, z_0, h; f) \ .$$

Seien $y(x)$ und $\tilde{y}(x_i)$ durch (5.4) bzw. (5.5) definiert. Es folgt:

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + h_i \Delta(x_i, y(x_i), h_i; f) \\ \tilde{y}(x_{i+1}; h) &= \tilde{y}(x_i; h) + h_i \Phi(x_i, \tilde{y}(x_i; h), h_i; f) \end{aligned}$$

wobei die letzte Gleichung manchmal der Einfachheit halber als

$$\tilde{y}(x_{i+1}) = \tilde{y}(x_i) + h_i \Phi(x_i, \tilde{y}(x_i), h_i; f)$$

oder sogar als

$$\tilde{y}_{i+1} = \tilde{y}_i + h_i \Phi(x_i, \tilde{y}_i, h_i; f)$$

geschrieben wird, wenn dadurch keine Mißverständnisse entstehen können.

Definition 5.4

$$e(x,h) := \tilde{y}(x,h) - y(x)$$

heißt globaler Diskretisierungsfehler *und*

$$\tau(x,z,h) := \Phi(x,z,h;f) - \Delta(x,z,h;f)$$

heißt relativer lokaler Diskretisierungsfehler.

Definition 5.5 *Ein Einschrittverfahren hat Ordnung p, falls $p \in \mathbb{N}$ die größtmögliche Zahl ist, wofür gilt:*

$$\tau(x,z,h) = O(h^p)\ ,$$

für alle $x \in [a,b]$, $z \in \mathbb{R}^n$, und $f \in C^\infty$.

5.2.5 Konstruktion von Einschrittverfahren

Die grundlegende Idee ist es, die Funktionen Δ und Φ in Taylor-Reihen zu entwickeln. Die Taylor-Reihe von Δ ist bekannt:

$$\Delta(x,z,h) = \frac{z(x+h)-z(x)}{h} = \sum_{k=0}^{\infty} \frac{h^k D^{k+1} z(x)}{(k+1)!} = \sum_{k=0}^{\infty} \frac{h^k}{(k+1)!}\, f^{(k)}(x,z)$$

wobei

$$f^{(k)}(x,z) := \frac{d^k}{dx^k}\, f(x,z(x))\ ,$$

so daß

$$\begin{aligned}
f^{(1)}(x,z) &= \frac{d}{dx}\, f(x,z(x)) = f_x + f_y \cdot \frac{dz}{dx} = f_x + f_y f \\
f^{(2)}(x,z) &= (f_{xx} + f_{xy}\cdot f) + (f_{yx}\cdot f + f_{yy}\cdot f \cdot f + f_y \cdot f_x + f_y \cdot f_y \cdot f) \\
&= f_{xx} + 2 f_{xy}\cdot f + f_{yy}\cdot f^2 + f_y\cdot f_x + f_y^2 \cdot f \quad \text{usw.}
\end{aligned}$$

Zusammenfassend:

$$\begin{aligned}
\Delta(x,z,h) &= f + \frac{h}{2}(f_x + f_y\cdot f) \\
&+ \frac{h^2}{6}(f_{xx} + 2 f_{xy}\cdot f + f_{yy}\cdot f^2 + f_y\cdot f_x + f_y^2\cdot f) \\
&+ \cdots
\end{aligned} \tag{5.6}$$

Die Taylor-Reihe von Φ kann erst bestimmt werden, wenn Φ vorgeschrieben worden ist.

Beispiel:

$$\Phi(x,z,h) = a_1 f(x,z) + a_2 f(x + p_1 h, z + p_2 h f(x,z))$$

Die Taylor-Reihe-Entwicklung für Φ ergibt sich aus:

$$\begin{aligned} f(x+p_1h, z+p_2hf(x,z)) &= f+(p_1hf_x+p_2hf_y\cdot f)+ \\ &+ \frac{1}{2!}((p_1h)^2f_{xx}+2p_1hp_2hf_{xy}\cdot f+(p_2h)^2f_{yy}\cdot f^2) \\ &+ \cdots \end{aligned}$$

Es folgt

$$\begin{aligned} \Phi(x,z,h) &= (a_1+a_2)f+h(p_1f_x+p_2f_y\cdot f)+ \\ &+ \frac{h^2}{2!}(p_1^2f_{xx}+2p_1p_2f_{xy}\cdot f+p_2^2f_{yy}\cdot f^2) \\ &+ \cdots \end{aligned} \tag{5.7}$$

Durch den Vergleich der Entwicklungen (5.6) und (5.7) folgt:

$$\begin{aligned} h^0f: \quad & 1 = a_1+a_2 \\ h^1f_x: \quad & \tfrac{1}{2} = p_1a_2 \\ h^1f_y\cdot f: \quad & \tfrac{1}{2} = p_2a_2 \end{aligned}$$

Es gibt somit vier Parameter a_1, a_2, p_1 und p_2 und drei algebraische Gleichungen (5.7). Die allgemeine Lösung ist:

$$\begin{aligned} a_1 &= 1-a_2 \\ p_1 &= \frac{1}{2a_2} \\ p_2 &= \frac{1}{2a_2} \end{aligned}$$

für $a_2 \in \mathbb{R}\backslash\{0\}$.
Für die Wahl $a_2 = \frac{1}{2}$ erhält man

$$\Phi(x,z,h) = \frac{1}{2}f(x,z)+\frac{1}{2}f(x+h, z+hf(x,z))$$

was der Methode von Heun (1900) entspricht:

$$\tilde{y}(x+h) = \tilde{y}(x)+\frac{1}{2}\,hf(x,\tilde{y}(x))+\frac{1}{2}\,hf(x+h,\tilde{y}(x)+hf(x,\tilde{y}(x))).$$

Ein weiteres Beispiel ist die „klassische" Runge-Kutta Methode der Ordnung 4:

$$\begin{aligned} k_1 &= f(x,z) \\ k_2 &= f(x+\frac{h}{2}, z+\frac{h}{2}\,k_1) \\ k_3 &= f(x+\frac{h}{2}, z+\frac{h}{2}\,k_2) \\ k_4 &= f(x+h, z+hk_3) \\ \Phi &= \frac{1}{6}(k_1+2k_2+2k_3+k_4) \end{aligned}$$

Diese Methode erreicht die Ordnung 4 mit Hilfe der 4 Auswertungen von f.

5.2.6 Konvergenz

Satz 5.7 *Es sei* $\Phi(x, y; h)$ *stetig für* $[a, b] \times (-\infty, +\infty) \times (0, h_0)$ *und Lipschitz-stetig bezüglich* y*:*

$$|\Phi(x, u, h) - \Phi(x, v, h)| \leq L|u - v|$$

Dann ist die Konsistenzbedingung

$$\Phi(x, y, 0) = f(x, y)$$

nötig und hinreichend dafür, daß $\tilde{y}(x, h) \longrightarrow y(x)$ *konvergiert.*

Satz 5.8 *Es sei* Φ *stetig auf*

$$G := [a, b] \times \mathbb{R}^n \times [0, h_0]$$

und Lipschitz-stetig, d.h.

$$\|\Phi(x, u; h) - \Phi(x, v; h)\| \leq M\|u - v\|.$$

Es sei weiter

$$|\tau(x, y(x); h)| = |\Delta(x, y(x), h) - \Phi(x, y(x), h)| \leq N h^p.$$

Dann gibt es eine Konstante $K = K(|b - a|, M)$*, so daß*

$$|e(x; h)| \leq h^p K\ N$$

für $x \in [a, b],\ h \leq h_0$.

Beweis: Durch Subtraktion der Gleichungen

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + h\Delta(x_i, y(x_i), h\ ;\ )f \\ \tilde{y}(x_{i+1}; h) &= \tilde{y}(x_i; h) + h\Phi(x_i,\ \tilde{y}(x_i; h), h; f) \end{aligned}$$

erhält man für den Fehler

$$\tilde{i} := ||\tilde{y}(x_i; h) - y(x_i)||$$

die Rekursionsformel

$$\tilde{e}_{i+1} \leq \tilde{e}_i + h||\Phi(x_i, \tilde{y}(x_i\ ;\ h), h\ ;\ f) - \Phi(x_i, y(x_i), h\ ;\ f)|| + h||\Phi(x_i, y(x_i), h\ ;\ f)||.$$

Mit Hilfe der Lipschitzbedingung für Φ und der gegebenen Abschätzung für den relativen lokalen Diskretisierungsfehler τ, folgt

$$\tilde{e}_{i+1} \leq \tilde{e}_i + h||\tilde{y}(x_i; h) - y(x_i)|| + h\ N\ h^p = \tilde{e}_i + h\ \tilde{e}_i + N\ h^{p+1}.$$

Hilfssatz 5.1 ergibt wegen $\tilde{e}_0 = \tilde{y}(x_0; h) - y(x_0) = 0$,

$$|\tilde{e}_i| \le N\ h^p\ \frac{e^{ihM} - 1}{M} \le N\ h^p\ \frac{e^{nhM} - 1}{M} = h^p N\ K$$

mit

$$\begin{aligned} n &:= \frac{b-a}{h}, \\ K &:= \frac{e^{M(b-a)} - 1}{M}. \end{aligned}$$

□

Hilfssatz 5.1 *Genügen die Zahlen ξ_1 einer Abschätzung der Form*

$$|\ \xi_{i+1}| \le (1+\delta)|\xi_1| + B, \quad B \le 0, \quad i = 0, 1, 2, \dots ,$$

so gilt

$$|\ \xi_n| \le e^{n\delta}|\xi_0| + \frac{e^{n\delta} - 1}{\delta} B.$$

5.2.7 Asymptotische Entwicklungen

Satz 5.9 *Es sei $f(x,y) \in C^\infty(a,b)$ und $\tilde{y}(x;h)$ die von einem Einschrittverfahren der Ordnung p, $p \le N$, gelieferte Näherungslösung $y(x)$ des Anfangswertproblems*

$$y^{'} = f(x,y)\ , \quad y(x_0)\ , \quad x_0 \in [a,b].$$

Dann besitzt $\tilde{y}(x;h)$ eine asymptotische Entwicklung der Form

$$\tilde{y}(x;h) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \ldots + h^N e_N(x) + h^{N+1} E_{N+1}(x;h)\ ,$$

mit $e_p(x_0) = 0$, die für alle $x \in [a,b]$ und alle $h = h_n = \frac{x-x_0}{n}$, $n = 1, 2, \dots$, gilt. Dabei sind die Funktionen $e_i(x)$ von h unabhängig und das Restglied $E_{N+1}(x;h)$ ist bei festem x für alle $h = h_n = \frac{x-x_0}{n}$, $n = 1, 2, \dots$, beschränkt.

Beispiel:

$$y^{'} = x \quad y(0) = 1$$

Wir benutzten die Eulersche Methode:

$$\eta_{i+1} = \eta_i + h x_i = \eta_i + h^2 i.$$

Wie leicht geprüft werden kann:

$$\begin{aligned} \eta_i &= \frac{i(i-1)}{2}\ h^2 \\ \eta(x,h) &= \frac{x^2}{2} - \frac{x}{2}\ h \end{aligned}$$

5.3 Differenzengleichungen

Differenzengleichungen spielen eine wichtige Rolle in vielen Teilen der numerischen Mathematik.

5.3.1 Einführung

Definition 5.6 *Sei $k \in \mathbb{N}_+$ und I eine Menge von aufeinanderfolgenden ganzen Zahlen. Seien S und T Teilmengen von $\mathbb{R}$ oder $\not\subset$. Sei $\{F_n\}$ eine Folge von Abbildungen*

$$F_n : S^{k+1} \to T\,,\ n \in I\,.$$

Die Folge $\{u_n\}$ heißt Lösung der Differenzengleichung

$$F_n(u_n, u_{n+1}, \dots, u_{n+k}) = t_n\,, \tag{5.8}$$

falls die folgenden Bedingungen erfüllt sind:

1. *u_j ist definiert und $u_j \in S$ für alle $j = n + \ell$, mit $n \in I$ und $0 \le \ell \le k$.*
2. *$t_n \in T$ für $n \in I$.*
3. *$F_n(u_n, u_{n+1}, \dots, u_{n+k}) = t_n$ für $n \in I$.*

Die Gleichung (5.8) heißt eine *Differenzengleichung der Ordnung k.*

Definition 5.7 *Die Differenzengleichung (5.8) heißt* linear, *wenn für alle $n \in I$ die Funktion F_n eine lineare Funktion ist:*

$$F_n(y_0, \dots, y_k) = \sum_{j=0}^{k} a_{jn}\, y_j\,.$$

Ein wichtiger spezieller Fall entsteht, wenn die Koeffizienten a_{jn} konstant sind, d.h.

$$a_{jn} = a_j \quad , \quad n \in I\,.$$

Die Differenzengleichung (5.8) heißt homogen, *wenn $t_n = 0$ für $n \in I$.*

1. **Beispiel**
   $I = \mathbb{N}, \qquad k = 1\,, \qquad S = T = \mathbb{R},$

   $$\begin{aligned} F_n(y_0, y_1) &:= y_1 - n\, y_0\,, \\ t_n &:= 0\,. \end{aligned} \tag{5.9}$$

   Die Gleichung (5.9) ist eine lineare homogene Differenzengleichung erster Ordnung.

   Die Folge $\{u_n\}$ mit

   $$u_n = \Gamma(n) = (n-1)!$$

   ist eine Lösung.

2. **Beispiel**
$I = \mathbb{Z}\,, \qquad k = 2\,, \qquad S = T = \mathbb{R}\,,$

$$\begin{aligned} F_n(y_0, y_1, y_2) &:= y_0 - 2y_1 \cos\varphi + y_2, \\ t_n &:= 0 \end{aligned} \tag{5.10}$$

mit $\varphi \in \mathbb{R}$. Die Gleichung (5.10) ist eine lineare homogene Differenzengleichung zweiter Ordnung. Die Folge $\{u_n\}$,

$$u_n := \cos n\varphi$$

ist eine Lösung.

3. **Beispiel**
Die Differenzengleichung

$$u_{n+1} - [u_n + \alpha u_n(1 - u_n)] = 0$$

ist eine Differenzengleichung erster Ordnung. Obwohl sie sehr einfach ist, zeigt sie „chaotische" Eigenschaften für $1,8 < \alpha < 3$.

5.3.2 Lineare Differenzengleichungen k-ter Ordnung

Seien $a_0, \ldots, a_k$ beliebige reelle oder komplexe Konstanten mit $a_0 a_k \neq 0$. Sei

$$F_n(y_0, \ldots, y_k) = \sum_{i=0}^{k} a_i y_i\,.$$

Wir betrachten die homogene lineare Differenzengleichung

$$(Lu)_n := \sum_{i=0}^{k} a_i u_{n+i} = 0i \tag{5.11}$$

als auch die inhomogene Gleichung

$$(Lu)_n := \sum_{i=0}^{k} a_i u_{n+i} = t_n\,. \tag{5.12}$$

Der Operator L wird als Abbildung der Folge $u := \{u_n\}$ auf der Folge $Lu := \{(Lu)_n\}$ definiert, und die Gleichungen (5.11) und (5.12) können deshalb als

$$Lu = 0 \quad \text{bzw.} \quad Lu = t$$

geschrieben werden.

Für lineare Differenzengleichungen mit konstanten Koeffizienten wird $I = \mathbb{Z}$ gesetzt, falls nicht anders festgelegt.

Definition 5.8 *Sei $Lu = 0$ eine homogene lineare Differenzengleichung*

$$(Lu)_n = \sum_{j=0}^{k} a_j u_{n+j} \,.$$

Das Polynom

$$p(z) := \sum_{j=0}^{k} a_j z^j$$

heißt das charakteristische Polynom *von L.*

Definition 5.9 *Die Folgen*

$$u^{(1)} = \{u_n^{(1)}, \ldots, u^{(k)} = \{u_n^{(k)}\}$$

seien k Lösungen von $Lu = 0$. Ihre Wronski-Determinante *an der Stelle n wird definiert durch*

$$w_n(u^{(1)}, \ldots, u^{(k)}) = w_n := \begin{vmatrix} u_{n+k}^{(1)} & \ldots & u_{n+k}^{(k)} \\ u_{n+k}^{(1)} & \ldots & \\ u_{n+1}^{(1)} & \ldots & u_{n+1}^{(k)} \end{vmatrix}$$

Satz 5.10

1. *Es seien $u^{(1)}, \ldots, u^{(k)}$ k Lösungen von $Lu = 0$. Dann läßt sich jede Lösung von $Lu = 0$ genau dann als Linearkombination dieser Lösungen darstellen, wenn ihre Wronski-Determinante für alle n von Null verschieden ist.*

2. *Es sei v eine spezielle (partikuläre) Lösung von $Luu = t$ und $u^{(1)}, \ldots, u^{(k)}$ ein Sysytem von k linear unabhängigen Lösungen der homogenen Gleichung $Lu = 0$.*
   Dann läßt sich jede Lösung ω der Gleichung $L\omega = t$ in der Form

   $$\omega = \sum_{j=1}^{k} c_j u^{(j)} + v$$

   mit passend gewählten Konstanten $c_1, \ldots, c_k$ darstellen.

3. *Das charakteristische Polynom p der homogen linearen Differenzengleichung k-ter Ordnung*

   $$(Lu)_n = \sum_{j=0}^{k} a_j u_{n+j}$$

   habe die verschiedenen Nullstellen

   $$z_1, z_2, \ldots, z_m \,,$$

mit den Vielfachheiten σ_r , $1 \le r \le m$. Dann bilden die k Folgen

$$\begin{aligned} &u_n^{(r,s)} := n(n-1)\cdots(n-s+1)z_r^{n-s}\,,\\ &0 \le s < \sigma_r\,,\\ &1 \le r \le m \end{aligned}$$

ein System von k linear unabhängigen Lösungen von $Lu = 0$.

Beweis:

1. $u_n^{(r,s)}$, $0 \le s < \sigma_r$ ist eine Lösung von $Lu = 0$. Da z_r eine Nullstelle des Polynoms p der Vielfachheit σ_r ist, gilt:

   $$p(z_r) = p'(z_r) = \cdots = p^{(s)}(z_r) = 0\,.$$

   Es folgt

   $$\begin{aligned} (Lu^{(r,s)})_n &= \sum_{j=0}^{m} a_j(n+j)(n+j-1)\cdots(n+j-s+1)z_r^{n+j-s}\,,\\ &= \left(\frac{dz}{d}\right)^s (p(z)z^n)|_{z=z_r} = 0 \quad \text{(Leibnizregel)}\,. \end{aligned}$$

2. $u_n^{(r,s)}$ sind linear unabhängig, falls $z_1, \ldots, z_k$ verschieden sind. Dies ergibt sich aus

   $$w_{-1} = \begin{vmatrix} z_1^{k-1} & \ldots & z_k^{k-1} \\ \vdots & & \vdots \\ z_1^0 & & z_k^0 \end{vmatrix} = \prod_{1\le s<t\le k} (z_s - z_t)\,.$$

   (Vandermonde Determinante.)

3. $u_n^{(r,s)}$ sind linear unabhängig im allgemeinen Fall. Dies ergibt sich aus

   $$w_{-1} = \prod_{1\le s<t\le m} (z_s - z_t)^{\sigma_s\sigma_t} \prod_{r=1}^{m} (\sigma_r - 1)\,!!$$

   mit

   $$0\,!! = 1\,, \qquad q\,!! := \prod_{j=1}^{q} j\,!$$

   (Siehe Henricii, *Discrete Variable Methods in Ordinary Differential Equations*, S. 214, wobei es einige Druckfehler gibt.)

□

Aus Satz 5.10.2 folgt, daß sich die allgemeine Lösung von $L\omega = t$ als die Summe $u + v$ darstellen läßt, wobei $Lu = 0$ und $Lv = t$. In Satz 5.10.3 wird die allgemeine Lösung von $Lu = 0$ konstruiert. Es bleibt die Frage, wie bestimmt man eine partikuläre Lösung v ?

Zur Konstruktion von v gibt es mehrere systematische Methoden. Für die Beispiele hier kann v durch Geschicklichkeit bestimmt werden, da t nur einfache Formen annimmt.

5.3.3 Stabilität der Lösungen von Differenzengleichungen

In vielen Anwendungen ist es wichtig festzustellen, ob alle Lösungen der homogenen Gleichung $Luu = 0$ beschränkt sind, d.h. ob alle Nullstellen des charakteristischen Polynoms $p(z)$ innerhalb des Einheitskreises liegen.

Es gibt mehrere Methoden, Aussagen über die Nullstellen eines Polynoms p zu gewinnen (siehe z.B. die Bücher von Marsden und Obreschkoff), wobei die Sätze von Routh (1877) und Hurwitz (1889) besonders zu erwähnen sind. Hier wird nur der Satz von Hurwitz zitiert:

Satz 5.11 (Hurwitz) *Sei p ein Polynom n-ter Grad mit reellen Koeffizienten,*

$$p(z) = \sum_{j=0}^{n} a_j z^j .$$

Sei $a_0 > 0$. Sei $a_j := 0$ für $j > n$ und $j < 0$. Für $1 \leq j \leq n$ sei $\triangle_j$ die $j \times j$ Determinante

$$\triangle_1 := |a_1| ,$$

$$\triangle_2 := \begin{vmatrix} a_1 & a_0 \\ a_3 & a_2 \end{vmatrix}$$

$$\cdots$$

$$\triangle_j := \begin{vmatrix} a_1 & a_0 & a_{-1} & \cdot & a_{2-j} \\ a_3 & a_2 & a_1 & \cdot & a_{4-j} \\ a_5 & a_4 & a_3 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{2j-1} & a_{2j-2} & a_{2j-3} & \cdot & a_j \end{vmatrix}$$

Das Polynom p hat genau dann nur Nullstellen mit negativem Realteil, wenn alle n Determinanten $\triangle_j$ streng positiv sind.

Bemerkung: Alle Nullstellen des Polynoms $q(\zeta)$ der Ordnung n liegen innerhalb des Einheitskreises genau dann, wenn alle Nullstellen des Polynoms

$$p(z) = (1-z)^n q\left(\frac{1+z}{1-z}\right)$$

negativen Realteil haben, da die gebrochene lineare Funktion

$$z = \frac{\zeta - 1}{\zeta + 1} \qquad , \qquad \zeta = \frac{1+z}{1-z}$$

den Einheitskreis in der komplexen ζ-Ebene eineindeutig konform auf die linke Halbebene in der z-Ebene abbildet.

Literatur

Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. John Wiley, 1962.

Henrici, P.: Elemente der numerischen Analysis I. Bibliographisches Institut, Mannheim, 1972.

Marden, M.: Geometry of Polynomials. AMS, 1966.

Nörlund, N.E.: Vorlesungen über Differenzenrechnung. 1923.

Obreschkoff, N.: Verteilung und Berechnung der Nullstellen reeller Polynome. VEB Deutscher Verlag der Wissenschaften, Berlin 1963.

Kapitel 6

Iterationsverfahren zur Lösung großer linearer Gleichungssysteme

6.1 Einleitung

Zur Lösung von Differentialgleichungen und verschiedenen technischen Problemen ist es oft erforderlich, sehr große lineare Gleichungssysteme

$$Ax = b\ ,\ x \in \mathbb{R}^n\ ,\ A \in \mathtt{Mat}(n,n)\ ,\ n \geq 10^3$$

zu lösen. Bei solchen Problemen hat die Matrix A fast immer eine spezielle Struktur, die den Einsatz von effizienten Iterationsverfahren ermöglicht.

Die folgenden Themen werden behandelt:

Das Jacobi Verfahren
Das Gauß-Seidel Verfahren
Verfahren der konjugierten Gradienten

6.2 Hilfsmittel

Die folgenden Begriffe und Sätze werden später benötigt.

Definition 6.1 *Zwei Matrizen $A, B \in M(n,n)$ heißen* ähnlich, *wenn es eine reguläre Matrix $S \in M(n,n)$ gibt mit*

$$B = SAS^{-1}\ .$$

Definition 6.2 *Sei $A \in M(n,n)$. Ein $\lambda \in \mathcal{C}$ heißt* Eigenwert *von A, wenn es ein v gibt mit $v \neq 0$, so daß gilt:*

$$Av = \lambda v\ .$$

Jedes vom Nullvektor verschiedene v *mit*

$$Av = \lambda v$$

heißt Eigenvektor *von* A *(zum Eigenwert* λ*). Man spricht auch vom* Eigenpaar (λ, v) .

Definition 6.3 *Sind* λ_i , $1 \leq i \leq \ell$, *die Eigenwerte von* $A \in M(n,n)$, *so nennt man*

$$\rho(A) := \max_{1 \leq i \leq \ell} |\lambda_i|$$

den Spektralradius *von* A .

Definition 6.4 *Eine Matrix* $J \in M(r,r)$ *heißt* Jordanmatrix *zum Eigenwert* λ, *wenn*

$$J = \begin{pmatrix} \lambda & 1 & & O \\ & \lambda & 1 & \\ & & \ddots & 1 \\ O & & & \lambda \end{pmatrix}$$

Satz 6.1 (Jordansche Normalform) *Sei* $A \in M(n,n)$. *Es gibt eine reguläre Matrix* $S \in M(n,n)$ *, so daß gilt*

$$SAS^{-1} = J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_\ell \end{pmatrix}$$

wobei für $1 \leq r \leq \ell$ $\quad J_r \in Mat(n_r, n_r)$ *eine Jordanmatrix ist,*

$$J_r = \begin{pmatrix} \lambda_r & 1 & & O \\ & \lambda_r & 1 & \\ & & \ddots & 1 \\ O & & & \lambda_r \end{pmatrix}$$

Beweis: Siehe Fischer, Lineare Algebra, Anhang B. □

Die charakteristischen Polynome $p_r(\lambda) = \det(J_r - \lambda I) = (\lambda_r - \lambda)^{n_r}$ heißen die *Elementarteiler* von A.

Satz 6.2 *Für alle Eigenwerte* λ *von* A *gilt*

$$|\lambda| \leq \boldsymbol{lub}(A) := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

für jede Vektornorm $\|\cdot\|$.

Beweis: Sei (λ, v) ein Eigenpaar von A, d.h. $Av = \lambda v$ mit $v \neq 0$. Dann folgt

$$|\lambda| \, \|v\| = \|Av\| \leq \|A\| \cdot \|v\| \, .$$

$\square$

Satz 6.3 *a) Zu jeder Matrix A und jedem $\epsilon > 0$ existiert eine Vektornorm $\|\cdot\|_\wedge$ mit*

$$\boldsymbol{lub}(A)_\wedge = \|A\|_\wedge \leq \rho(A) + \epsilon \, .$$

b) Hat jeder Eigenwert λ von A mit der Eigenschaft $|\lambda| = \rho(A)$ nur lineare Elementarteiler, so existiert sogar eine Vektornorm $\|\cdot\|_\wedge$ mit

$$\|A\|_\wedge := \boldsymbol{lub}(A)_\wedge = \rho(A) \, .$$

Beweis:

Teil a): Sei $SAS^{-1} = J$, wobei J in Jordannormalform ist, $J = \operatorname{diag}(J_1, \ldots, J_\ell)$,

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_\ell \end{pmatrix} , \quad J_r = \begin{pmatrix} \lambda_r & 1 & & O \\ & \lambda_r & 1 & \\ & & \ddots & 1 \\ O & & & \lambda_r \end{pmatrix}$$

Sei $C(\epsilon) := C := \operatorname{diag}(C_1, \ldots, C_\ell)$ mit

$$C_r = C_r(\epsilon) = \operatorname{diag}(1, \epsilon, \ldots, \epsilon^{n_r - 1}) \in \operatorname{Mat}(n_r, n_r) \, .$$

Es folgt

$$C^{-1}JC = \operatorname{diag}(C_1^{-1}J_1C_1, \ldots, C_\ell^{-1}J_\ell C_\ell) \, ,$$

mit

$$C_r^{-1}J_rC_r = \begin{pmatrix} \lambda_r & \epsilon & & O \\ & \lambda_r & \epsilon & \\ & & \ddots & \epsilon \\ O & & & \lambda_r \end{pmatrix}$$

Es folgt sofort, daß

$$\|C^{-1}JC\|_\infty \leq \rho(A) + \epsilon \, ,$$

so daß gilt

$$\|TAT^{-1}\|_\infty \leq \rho(A) + \epsilon$$

mit

$$T := C^{-1}S \, .$$

Daraus folgt mit Hilfssatz 6.1

$$\|A\|_\wedge := \|TAT^{-1}\|_\infty \leq \rho(A) + \epsilon\ .$$

Teil b): Für

$$\epsilon < \rho(A) - \max_{|\lambda_r| < \rho(A)} |\lambda_r|$$

setze

$$D_r = \begin{cases} C_r(\epsilon) & , \quad |\lambda_r| < \rho(A) \\ I & , \quad |\lambda_r| = \rho(A) \end{cases}$$

$$D = \text{diag}(D_1, \ldots, D_\ell)\ .$$

Dann gilt:

$$\|D^{-1}JD\|_\infty = \rho(A)\ .$$

□

Hilfssatz 6.1 *Sei $\|\cdot\|$ eine Vektornorm und $T \in Mat(n,n)$ eine reguläre Matrix. Dann ist $\|\cdot\|_\wedge$,*

$$\|x\|_\wedge := \|Tx\|$$

eine Vektornorm und

$$lub(A)_\wedge = \|TAT^{-1}\|$$

für alle $A \in Mat(n,n)$.

Beweis: $\|\cdot\|_\wedge$ ist eine Vektornorm, da die drei notwendigen Voraussetzungen offensichtlich erfüllt sind.
Es gilt für $A \in \text{Mat}(n,n)$

$$\text{lub}(A)_\wedge = \sup_{x \neq 0} \frac{\|Ax\|_\wedge}{\|x\|_\wedge} = \sup_{y \neq 0} \frac{\|AT^{-1}y\|_\wedge}{\|T^{-1}y\|_\wedge} = \sup_{y \neq 0} \frac{\|TAT^{-1}y\|}{\|y\|} = \|TAT^{-1}\|.$$

□

Satz 6.4 *Sei $A \in Mat(n,n)$. Der Limes*

$$R := \lim_{k \to \infty} \|A^k\|^{1/k}$$

existiert und $R = \rho(A)$.

Beweis: Sei

$$R_k := \|A^k\|^{1/k}$$

und (λ, x) ein Eigenpaar von A. Dann gilt:

$$R_k \geq \left[\frac{\|A^k x\|}{\|x\|}\right]^{1/k} = |\lambda| \ ,$$

so daß

$$R_k \geq \rho(A) \ . \tag{6.1}$$

Sei nun $\epsilon > 0$ beliebig. Es folgt aus Satz 6.3, daß es eine Norm $\| \ \|_\epsilon$ gibt mit

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

Alle Normen auf dem $\mathbb{R}^n$ sind äquivalent, und es gibt deshalb eine Konstante $m = m(\epsilon)$ mit

$$\frac{1}{m} \ \|x\| \leq \|x\|_\epsilon \leq m\|x\| \ .$$

Es folgt

$$\begin{aligned} R_k &= \left[\sup_{x \neq 0} \frac{\|A^k x\|}{\|x\|}\right]^{1/k} \leq \left[\sup_{x \neq 0} \frac{m^2 \|A^k x\|_\epsilon}{\|x\|_\epsilon}\right]^{1/k} \\ &= m^{2/k} \cdot [\|A^k\|_\epsilon]^{1/k} \leq m^{2/k} \cdot \|A\|_\epsilon \leq m^{2/k} \cdot (\rho(A) + \epsilon) \end{aligned} \tag{6.2}$$

Für alle $m > 0$ gilt

$$\lim_{k \to \infty} \ m^{2/k} = 1 \ .$$

Der Satz folgt deshalb sofort aus den Ungleichungen (6.1) und (6.2). □

Bemerkung:

a) $\rho(A) \leq \|A\|$

b) $\rho(A) < 1$ und $\rho(B) < 1$ impliziert i.a. nicht $\rho(AB) < 1$.

Beispiel:

a)

$$A = \begin{pmatrix} 1/2 & 100 \\ 0 & 1/2 \end{pmatrix} \ , \quad A^k = \begin{pmatrix} (\frac{1}{2})^k & 100k(\frac{1}{2})^{k-1} \\ 0 & (\frac{1}{2})^k \end{pmatrix}$$

b)

$$A = \begin{pmatrix} 1/2 & 100 \\ 0 & 1/2 \end{pmatrix} \ , \quad B = \begin{pmatrix} 1/2 & 0 \\ 100 & 1/2 \end{pmatrix} \ , \quad AB = \begin{pmatrix} 10^4 + 1/4 & 50 \\ 50 & 1/4 \end{pmatrix}$$

$$\rho(A) = \rho(B) = 1/2 \ , \quad \rho(AB) > 1 \ .$$

Genauere Information über die Potenzen A^m einer Matrix A erhält man aus der Jordanschen Normalform.

Satz 6.5 *A sei eine $n \times n$-Matrix mit $\rho(A) > 0$. Dann gilt:*

$$\|A^m\| \sim \nu \binom{m}{p-1} [\rho(A)]^{m-(p-1)} \ , \ m \longrightarrow \infty \ ,$$

wobei p die größte Ordnung aller Jordanscher Blockmatrizen J von A mit $\rho(J) = \rho(A)$ und ν eine positive Konstante ist.

Beweis: Siehe Varga, Matrix Iterative Analysis, S. 65. □

Beispiel:

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, \quad A^m = \begin{pmatrix} \lambda^m & m\lambda^{m-1} \\ 0 & \lambda^m \end{pmatrix}.$$

6.3 Das Jacobi-, Gauß-Seidel- und SOR-Verfahren — eine Einleitung

In diesem Abschnitt werden drei Iterationsverfahren zur Lösung des Gleichungssystems

$$Ax = c \tag{6.3}$$

vorgestellt. Im nächsten Abschnitt wird die Konvergenz dieser Verfahren untersucht. Als Beispiel nehmen wir

$$A = \begin{pmatrix} 4 & 0 & -1 & -1 \\ 0 & 4 & -1 & -1 \\ -1 & -1 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{pmatrix}, \quad c = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{6.4}$$

Bei allen drei Verfahren wird A als Differenz von zwei Matrizen M und N dargestellt,

$$A = M - N \ ,$$

wobei gilt:

1. M sei regulär
2. M sei „leicht invertierbar", d.h. die Gleichung $My = c$ sei für beliebiges c leicht zu lösen.

Das Gleichungssystem (6.3) kann dann in der Gestalt

$$Mx = b + Nx$$

oder

$$x = M^{-1}b + M^{-1}Nx$$

geschrieben werden. Es liegt deshalb nahe, das Iterationsverfahren

$$x^{(k+1)} = Bx^{(k)} + d\ ,\ \ k \geq 0$$

zu benutzen mit

$$\begin{aligned} B &:= M^{-1}N\ , \\ d &:= M^{-1}b\ . \end{aligned}$$

Es ist manchmal sinnvoll, die folgende Zerlegung von A zu benutzen:

$$\begin{aligned} A &= D - L - R \quad \text{mit} \\ D &= \text{diag}(A) \\ L &= \text{linke Dreiecksmatrix} \\ R &= \text{rechte Dreiecksmatrix} \end{aligned}$$

1. **Das Jacobi[1] -Verfahren oder Gesamtschrittverfahren**

   Zur Lösung von

   $$Ax = b$$

   wird die Zerlegung

   $$\begin{aligned} A &= M - N\ , \\ M &:= D = \text{diag}(A)\ , \\ N &:= D - A = L + R \end{aligned}$$

   benutzt. Es folgt

   $$x^{(k+1)} = Bx^{(k)} + d$$

   mit

   $$\begin{aligned} d &= D^{-1}b \\ B &= D^{-1}(D - A) = D^{-1}(L + U) = I - D^{-1}A\ . \end{aligned}$$

   Es folgt weiter:

   $$Dx^{(k+1)} + (A - D)x^{(k)} = b\ .$$

[1] Carl Gustav Jacobi (1804-1855)

Dies ist die Gleichung, der die Bezeichnung *Gesamtschrittverfahren* zugrunde liegt.

Die Matrix $B = I - D^{-1}A$ wird oft als *Jacobi-Matrix* bezeichnet.

Beispiel: A, b und x seien wie in (6.4). Dann gilt für das Jacobi-Verfahren

$$B = \begin{pmatrix} 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 \end{pmatrix}.$$

Mit

$$\begin{aligned} x^{(0)} &= (0,0,0,0)^T \quad \text{gilt:} \\ x^{(1)} &= (0,5,0,5,0,5,0,5)^T \,, \\ x^{(2)} &= (0,75,0,75,0,75,0,75)^T \\ x^{(10)} &= (0,999023,0,999023,0,999023,0,999023)^T \\ &\vdots \\ x^{(20)} &= (0,9999,0,9999,0,9999,0,9999)^T \end{aligned}$$

2. **Das Gauß-Seidel-(Einzelschritt-) Verfahren**

$$\begin{aligned} M &= D - L \,, \\ N &= R \,, \\ B &= G := (D-L)^{-1}R \,, \\ (D-L)x^{k+1} &- Rx^{(k)} = c \end{aligned}$$

Beispiel: Mit

$$\begin{aligned} x^{(0)} &= (0,0,0,0)^T \quad \text{gilt:} \\ x^{(1)} &= (0,5,0,5,0,75,0,75)^T \\ x^{(2)} &= (0,875,0,875,0,9375,0,9375)^T \\ x^{(10)} &= (0,99998,0,99998,0,99999,0,99999)^T \end{aligned}$$

3. **Das S.O.R.-Verfahren**

$$
\begin{aligned}
\omega &\in (0,2)\ ,\\
M &= \frac{1}{\omega}\,D - L\ ,\\
N &= \frac{1-\omega}{\omega}\,D + R\ ,\\
B(\omega) &= L_\omega := (D-\omega L)^{-1}((1-\omega)D+\omega R)\ ,\\
(D-\omega L)x^{(k+1)} &- ((1-\omega)D+\omega R)x^{(k)} = c
\end{aligned}
$$

Beispiel: Mit

$$
\begin{aligned}
x^{(0)} &= (0,0,0,0)^T\\
\omega &= 1,5 \quad \text{gilt:}\\
x^{(1)} &= (0,75, 0,75, 1,3125, 1,3125)^T\ ,\\
x^{(10)} &= (0,99882, 0,99882, 1,00002, 1,00002)^T\ .
\end{aligned}
$$

Mit

$$
\begin{aligned}
x^{(0)} &= (0,0,0,0)^T\\
\omega &= 1,0718 \quad \text{gilt:}\\
x^{(1)} &= (0,535898, 0,535898, 0,823085, 0,823085)^T\\
x^{(7)} &= (1,0000, 1,0000, 1,0000, 1,0000)^T
\end{aligned}
$$

6.4 Konvergenzbetrachtungen

Definition 6.5 *Das Iterationsverfahren*

$$
\begin{aligned}
x^{k+1} &= Bx^k + d\ ,\\
B &= M^{-1}N\ ,\ d = M^{-1}c
\end{aligned}
$$

zur Lösung des Gleichungssystems

$$
Ax \equiv (M-N)x = c
$$

heißt konvergent*, falls für alle Startvektoren* $x^{(0)}$ *und Vektoren* d *die Folge* $\{x^{(k)}\}$ *gegen die exakte Lösung* $x = A^{-1}c$ *konvergiert.*

Wir können jetzt notwendige und hinreichende Bedingungen für Konvergenz herleiten. Der folgende Satz ist sowohl eine Erweiterung (da eine hinreichende Bedingung gegeben wird) als auch eine Spezialisierung (da nur lineare Abbildungen betrachtet werden) des Banachschen Fixpunktsatzes.

Satz 6.6 *Das Iterationsverfahren*

$$x^{(k+1)} = Bx^{(k)} + d \tag{6.5}$$

ist genau dann konvergent, wenn

$$\rho(B) < 1 \ .$$

Beweis:

a) Sei (6.5) konvergent. Für jedes d konvergiert die Folge $x^{(0)} := 0$, $x^{(k+1)} = Bx^{(k)} + d$. Sei

$$x := \lim_{k \to \infty} x^{(k)} \ .$$

Es gilt $x = Bx + d$, so daß die Gleichung

$$y = By + d$$

für jedes d eine Lösung hat. Die Abbildung $F := I - B$,

$$F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

ist also surjektiv und deshalb auch bijektiv (Fischer, S. 86).

Für den Fehler $f^{(k)} := x - x^{(k)}$ gilt:

$$f^{(k+1)} = Bf^{(k)} \ .$$

Wähle $x^{(0)}$ so, daß $f^{(0)} = v$, wobei $\{\lambda, v\}$ ein Eigenpaar für B ist:

$$\begin{aligned} x^{(0)} &:= 0 \ , \\ d &:= (I - B)v \end{aligned}$$

Es folgt:

$$f^{(k)} = \lambda^k v \ , \ k \in \mathbb{N} \ ,$$

und daher $|\lambda| < 1$.

b) Sei umgekehrt $\rho(B) < 1$. Es folgt aus Satz 6.3, daß eine Norm existiert mit $\|B\| < 1$. Es folgt, daß die Abbildung $g(u) := Bu + d$ kontrahierend ist. Man benutzt jetzt den Banachschen Fixpunktsatz.

□

Eine einfache aber oft nützliche notwendige Bedingung für die Konvergenz des Gesamtschrittverfahrens wird jetzt hergeleitet.

Definition 6.6 *Die Matrix* $A \in Mat(n,n)$ *erfüllt das* starke Zeilensummenkriterium *falls*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ , \ 1 \leq i \leq n \ .$$

Satz 6.7 *Die Matrix A erfülle das starke Zeilensummenkriterium. Dann konvergiert das Gesamtschrittverfahren.*

Beweis: Es gilt für die Jacobi-Matrix $B = (b_{ij})$,

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|/|a_{ii}| < 1 \ , \ 1 \leq i \leq n, \qquad b_{ii} = 0 \ , \ 1 \leq i \leq n. \tag{6.6}$$

Es gibt jetzt zwei einfache Methoden, um zu zeigen, daß $\rho(B) < 1$.

Methode 1: Wegen (6.6) gilt:

$$\rho(B) < \|B\|_\infty < 1 \ .$$

Methode 2: Sei $\{\lambda, v\}$ ein Eigenpaar von B. Wähle k , so daß

$$|v_k| = \|v\|_\infty = \max_j \ |v_j|$$

und betrachte die k-te Gleichung des Gleichungssystems $\lambda v = Bv$:

$$|\lambda v_k| = \sum_{j=1}^{n} |b_{ij}| \ |v_j| \leq \|v\|_\infty \sum_{j=1}^{n} |b_{ij}| < \|v\|_\infty$$

Es folgt, daß $|\lambda| < 1$. □

Satz 6.8 *Sei $A \in Mat(n,n)$ und $L\omega$ die zugehörige SOR- Iterationsmatrix. Dann gilt:*

$$\rho(L_\omega) \geq |\omega - 1| \ .$$

Beweis:

$$L_\omega = (D - \omega L)^{-1}((1-\omega)D + \omega R) \ .$$

Es folgt:

$$(D - \omega L)L_\omega = (1-\omega)D + \omega R \ .$$

Weiter gilt:

$$\begin{aligned}
\det(D) \cdot \det(L_\omega) &= \det(D - \omega L) \cdot \det(L_\omega) \\
&= \det((D - \omega L)L_\omega) \ , \\
&= \det((1-\omega)D + \omega R) \ , \\
&= (1-\omega)^n \cdot \det(D) \ , \quad \text{so daß} \\
\det(L_\omega) &= (1-\omega)^n \ .
\end{aligned}$$

Das Produkt der evtl. vielfältigen Eigenwerte von L_ω ist gleich dem konstanten Term von $\det(\lambda I - L_\omega)$:

$$\prod_{i=1}^{n} \lambda_i = \det(-L_\omega) = (\omega - 1)^n \ .$$

□

Bemerkung: Es folgt aus Satz 6.8, daß das SOR-Verfahren nur für $\omega \in (0,2)$ konvergent ist.

Satz 6.9 (Ostrowski-Reich) *Sei $A = D - E - E^*$ eine $n \times n$ hermitesche Matrix, D eine positiv definite hermitesche Matrix, $D - \omega E$ nichtsingulär für $0 \leq \omega \leq 2$. Sei*

$$L_\omega = (D - \omega E)^{-1}(1-\omega)D + \omega E^*) \ .$$

Dann ist $\rho(L_\omega) < 1$ genau dann, wenn A positiv definit ist und $\omega \in (0,2)$.

Beweis: Zuerst wird eine Hilfsgleichung hergeleitet. Die Fehler $e^m := x - x^m$ erfüllen die Gleichungen:

$$\begin{aligned} e^{m+1} &= L_\omega e^m \\ (D - \omega E)e^{m+1} &= (\omega E^* + (1-\omega)D)e^m \end{aligned} \tag{6.7}$$

Sei

$$\delta^m := e^m - e^{m+1} \ , \ m \geq 0 \ .$$

Aus (6.7) folgt

$$(D - \omega E)\delta^m = [(D - \omega E) - (\omega E^* + (1-\omega)D)]e^m = \omega A e^m \tag{6.8}$$

und

$$\begin{aligned} \omega A e^{m+1} &= \omega(D - E - E^*)e^{m+1} \\ &= [(\omega - 1)D - \omega E^*]e^{m+1} + (D - \omega E)e^{m+1} \\ &= [(1-\omega)D + \omega E^*](e^m - e^{m+1}) \\ &= (1-\omega)D\delta^m + \omega E^* \delta^m \ . \end{aligned} \tag{6.9}$$

Werden Gleichungen (6.8) und (6.9) mit ϵ_m^* bzw. e_{m+1}^* multipliziert, dann folgt, daß

$$S_1 := e^{m*}[D - \omega E]\delta^m = \omega e^{m*} A e^m \quad \text{und} \tag{6.10}$$

$$S_2 := e^{m+1*}[(1-\omega)D + \omega E^*]\delta^m = \omega e^{m+1*} A e^{m+1} \tag{6.11}$$

Sei

$$S := S_1 - S_2 \ . \tag{6.12}$$

Es gilt:

$$
\begin{aligned}
S &= e^{m*}[D-\omega E]\delta^m - e^{m+1*}[(1-\omega)D+\omega E^*]\delta^m \\
&= e^{m*}[D-\omega E]\delta^m - (e^m-\delta^m)^*[(1-\omega)D+\omega E^*]\delta^m \\
&= \delta^{m*}[(1-\omega)D+\omega E^*]\delta^m + e^{m*}[\omega D-\omega E-\omega E^*]\delta^m \\
&= \delta^{m*}[(1-\omega)D+\omega E^*]\delta^m + e^{m*}\omega A\delta^m \\
&= \delta^{m*}[(1-\omega)D+\omega E^*]\delta^m + \delta^{m*}\omega A e^m .
\end{aligned}
$$

Mit Hilfe von (6.8) folgt:

$$S = \delta^{m*}\{[(1-\omega)D+\omega E^*]+[D-\omega E]\}\delta^m = (2-\omega)\delta^{m*}D\delta^m. \tag{6.13}$$

Aus (6.10), (6.11), (6.12) und (6.13) folgt:

$$(2-\omega)\delta^{m*}D\delta^m = \omega\{e^{m*}Ae^m - e^{m+1*}Ae^{m+1}\}\ . \tag{6.14}$$

Diese Gleichung ist der Schlüssel zum Konvergenzbeweis, da daraus folgt, daß unter bestimmten Voraussetzungen die Zielfunktion $e^{m*}Ae^m$ monoton fallend ist.

Wähle ein Eigenpaar (λ, v) von L_ω und setze $e^0 = v$. Dann folgt aus (6.14):

$$
\begin{aligned}
e^1 &= L_\omega e^0 = \lambda e^0 \\
\delta^0 &= (1-\lambda)e^0 && (6.15)\\
(2-\omega)|1-\lambda|^2 e^{0*}De^0 &= \omega(1-|\lambda|^2)e^{0*}Ae^0. && (6.16)
\end{aligned}
$$

Wir können jetzt den Beweis durchführen.

a) Sei nun A positiv definit und $\omega \in (0,2)$. Die Matrix D ist positiv definit. Wegen $e^0 = v \neq 0$ gilt $e^{0*}De^0 > 0$. Es folgt aus (6.16), daß entweder $\lambda = 1$ oder $1-|\lambda|^2 > 0$.

   Wäre $\lambda = 1$, dann folgte aus (6.15), daß $\delta^0 = 0$ und folglich aus (6.8), daß $e^{0*}Ae^0 = 0$, was der Annahme $e^0 = v \neq 0$ widerspricht. Es gilt deshalb

$$|\lambda| < 1\ .$$

   Da λ einen beliebigen Eigenwert darstellt, ist $\rho(B) < 1$.

b) Sei nun $\rho(L_\omega) < 1$. Dann konvergiert die Folge $\{x^k\}$ für jedes x^0, so daß $e^m \to 0$ für $m \to \infty$.

   Es ist bekannt aus Satz 6.8, daß

$$|\omega - 1| \leq \rho(L_\omega) < 1\ ,$$

   d.h. $\omega \in (0,2)$.

Es gilt, da $\omega \neq 0$,

$$A = \frac{1}{\omega}\,(D - \omega E)(I - L_\omega)\ ,$$

so daß A regulär ist.

Wäre A nicht positiv definit, hätte A mindestens einen negativen Eigenwert λ mit Eigenvektor v. Setze $e^0 := v$. Es gelte:

$$e^{0*}Ae^0 < 0\ .$$

Es würde dann aus (6.14) folgen, daß

$$e^{m+1*}Ae^{m+1} \leq e^{m*}Ae^m \leq \cdots \leq e^{0*}Ae^0 < 0\ ,$$

was der Konvergenz der Folge $\{e^m\}$ gegen Null widerspricht.

Zusammenfassend muß gelten:

$$\begin{array}{lll} \omega & \in & (0,2)\ , \\ A & & \text{positiv definit}\ . \end{array}$$

□

Bemerkung: Die Behauptung im Teil b) des Beweises, daß $\omega \in (0,2)$, ist nicht vollständig bewiesen und möglicherweise sogar falsch.

Literatur

Berman, A., **Plemmons**, R.J.: Nonnegative Matrices in the Mathematical Sciences. Academic Press, 1979.

Golub, G.H., **van Loan**, C.F.: Matrix Computations. North Oxford Academic, Oxford, 1983.

Hageman, L.A., **Young**, D.M.: Applied Iterative Methods. Academic Press, 1981.

Hackbusch, W.: Multigrid Methods and Applications. Springer, 1985.

Varga, R.S.: Matrix Iterative Analysis. 1962.

Kapitel 7

Komplexität von Algorithmen

7.1 Einleitung

Wir haben jetzt mehrere Algorithmen für lineare Algebra kennengelernt. Wenn mehrere Algorithmen vorhanden sind, ist es wünschenswert, die Algorithmen in bezug auf verschiedene Kriterien zu betrachten, damit der „beste" Algorithmus ausgesucht werden kann. Mögliche Kriterien sind:

1. Rechenaufwand
2. Speicherbedarf
3. Stabilität oder Gutartigkeit
4. Zusätzliche Vor- und Nachteile

Nachdem mehrere Kriterien vorhanden sind und die einzelnen Kriterien sich nicht immer genau beurteilen lassen, ist es nicht immer möglich, den „besten" Algorithmus zu finden. Dagegen ist es oft möglich, „schlechte" Algorithmen zu erkennen.

Die Kriterien *Rechenaufwand* und *Speicherbedarf*, die seit eh und je von Numerikern untersucht worden sind, werden heutzutage unter der Rubrik *Komplexität* zusammengefaßt. Die Theorie der Komplexität von Algorithmen ist ein sehr interessantes und lebhaftes Teilgebiet der Informatik. Hier befassen wir uns nur mit der Frage des Rechenaufwandes für einige Algorithmen der linearen Algebra. Die Anzahl Multiplikationen, Additionen (Subtraktionen) und Divisionen wird mit $M(n)$, $A(n)$ und $D(n)$ bezeichnet, wobei der Parameter n die Größe des Problems beschreibt.

Es gibt mehrere Möglichkeiten:

a) Die Anzahl Operationen ist gering, aber das Problem sollte sehr oft gelöst werden (z.B. dreidimensionale Grafik).

b) Es gibt bekannte Konstanten K und α , so daß:

$$M(n) \sim Kn^{\alpha} \ ,$$

also

$$M(n) = Kn^{\alpha} + O(n^{\alpha})$$

oder

$$(M(n)/Kn^{\alpha}) \longrightarrow 1 \quad \text{für} \quad n \longrightarrow \infty \ .$$

c) Es gibt eine Konstante α mit

$$M(n) = O(n^{\alpha}) \ .$$

Es gibt also eine Konstante K , womöglich unbekannt, mit

$$M(n) \leq Kn^{\alpha} \ .$$

Wir betrachten nur Fälle b) und c).

7.2 Lineare Gleichungssysteme

Die Komplexität von einigen Algorithmen zur Lösung linearer Gleichungssysteme wird in Tabelle 7.1 zusammengefaßt:

Algorithmus	$M(n)$	Matrix A
Gauß-Elimination	$\sim \frac{1}{3}\, n^3$	allgemein
Gauß-Jordan	$\sim \frac{1}{2}\, n^3$	allgemein
Choleski	$\sim \frac{1}{6}\, n^3$	symmetrisch
Strassen (1969)	$\sim 4.7\, n^{2.81}$	allgemein
Coppersmith und Winograd (1982)	$0(n^{2.496})$	allgemein

Tabelle 7.1: Algorithmen zur Lösung der Gleichung $Ax = b$

Die Ergebnisse für die Gauß-, Gauß-Jordan- und Choleski-Algorithmen sind seit langem bekannt und lassen sich aus einfachen Überlegungen bestimmen, wobei die Formeln

$$\begin{aligned}
\sum_{k=1}^{n} 1 &= n \\
\sum_{k=1}^{n} k &= \frac{n(n+1)}{2} \\
\sum_{k=1}^{n} k^2 &= \frac{n(n+1)(2n+1)}{6}
\end{aligned}$$

nützlich sind.

Es wurde vermutet, daß $M(n) = O(n^3)$ für alle Methoden zur Lösung des Gleichungssystems $Ax = b$ mit vollbesetzter Matrix A. Diese Vermutung wurde durch die Erfindung des Strassen Algorithmus widerlegt, dem der nächste Abschnitt gewidmet ist.

7.2.1 Der Strassen Algorithmus

Der Strassen Algorithmus ist ein Algorithmus zur Berechnung des Matrixprodukts $C = A \cdot B$, wo A, B und C quadratische Matrizen sind. Der Algorithmus wird induktiv definiert.

Seien A und B $2n \times 2n$ Matrizen, die folgendermaßen partitioniert sind:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

Sei $C = A \cdot B$,

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

Sei

$$\begin{aligned} p_1 &:= (a_{11} + a_{22})(b_{11} + b_{22}) \\ p_2 &:= (a_{21} + a_{22})b_{11} \\ p_3 &:= a_{11}(b_{12} - b_{22}) \\ p_4 &:= a_{22}(-b_{11} + b_{21}) \\ p_5 &:= (a_{11} + a_{12})b_{22} \\ p_6 &:= (-a_{11} + a_{21})(b_{11} + b_{12}) \\ p_7 &:= (a_{12} - a_{22})(b_{21} + b_{22}) \end{aligned} \tag{7.1}$$

Es gilt:

$$\begin{aligned} c_{11} &= p_1 + p_4 - p_5 + p_7 \\ c_{12} &= p_3 + p_5 \\ c_{21} &= p_2 + p_4 \\ c_{21} &= p_1 + p_3 - p_2 + p_6 \end{aligned} \tag{7.2}$$

Sei nun $M(n)$ und $A(n)$ die Anzahl elementarer Multiplikationen bzw. Additionen, die bei dem Strassen Algorithmus erforderlich sind. Es folgt aus (7.1) und (7.2):

$$\begin{aligned} M(2n) &= 7M(n) \\ A(2n) &= 7A(n) + 18n^2 \end{aligned}$$

da bei der Berechnung der p_i 7 Multiplikationen von $n \times n$- Matrizen erforderlich sind und außerdem 18 $n \times n$-Matrizen addiert werden müssen.

Sei

$$\begin{aligned} u_k &= M(2^k) \ , \\ v_k &= A(2^k) \ , \ k \geq 0 \ . \end{aligned}$$

Dann gilt:

$$\begin{aligned} u_{k+1} &= 7u_k \ , \\ v_{k+1} &= 7v_k + 18 \cdot 4^k \ , \\ u_0 &= 1 \ , \\ v_0 &= 0 \ . \end{aligned}$$

u_k und v_k sind also Lösungen einer homogenen bzw. inhomogenen Differenzengleichung erster Ordnung.

Es folgt sofort, daß

$$\begin{aligned} u_k &= 7^k \ , \\ v_k &= \alpha \, 7^k + w_k \end{aligned}$$

wobei α ein Parameter und w_k eine Lösung der inhomogenen Gleichung ist. Der Ansatz

$$w_k = \beta \, 4^k$$

ergibt

$$4\beta = 7\beta + 18$$

oder

$$\beta = -6 \ .$$

Mit $\alpha = 6$ wird die Bedingung $v_0 = 0$ erfüllt, so daß

$$v_k = 6 \cdot 7^k - 6 \cdot 4^k \leq 6 \cdot 7^k$$

Für $M(n)$ und $A(n)$ ist:

$$\begin{aligned} n &= 2^k \ , \\ k &= \log_2 n \ , \\ M(2^k) &= u_k = 7^k \ , \\ A(2^k) &= v_k \leq 6 \cdot 7^k \ , \\ M(n) &= 7^{\log_2 n} = 7^{\log_7 n \cdot \log_2 7} = n^{\log_2 7} \ , \\ A(n) &\leq 6 \cdot n^{\log_2 7} \ , \end{aligned}$$

wobei $\log_2 7 \doteq 2,81$. Es gilt also:

$$\begin{aligned} A(n) &= O(n^{2,81}) \\ M(n) &= O(n^{2,81}) \end{aligned}$$

Zur Berechnung von A^{-1} sind folgende Überlegungen nötig: Sei

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

dann gilt

$$A^{-1} = \begin{pmatrix} I & -a_{11}^{-1}a_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} a_{11}^{-1} & 0 \\ 0 & d^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -a_{21}a_{11}^{-1} & I \end{pmatrix}$$

mit $d := a_{22} - a_{21}a_{11}^{-1}a_{12}$. Es werden

$$I(n) = 2I(n/2) + 6M(n/2) + 6A(n/2) + 2.n^2/4$$

Operationen erforderlich, also

$$I(n) = 2I(n/2) + 6T(n/2) + n^2/2 \leq 3 \sum_{i=1}^{\log_2 n} 2^i T(n/2^i) + O(n^2) = O(n^{\log_2 7}) \ ,$$

wobei $T(n)$ die Anzahl Operationen für Matrixmultiplikation ist. Damit d nicht singulär ist, wird die Gleichung

$$Ax = b$$

mit

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

ersetzt.

7.3 Eigenwertprobleme

Die Komplexität von einigen Algorithmen zur Bestimmung der Eigenwerte einer Matrix A wird in Tabelle 7.2 zusammengefaßt:

7.3.1 Das QR-Verfahren

Wir betrachten den Fall, wenn A_s eine Hessenbergsche Matrix ist. Es ist möglich, A_s zur diagonalen Gestalt zu bringen mit $n-1$ ebenen Drehungen.

$$\begin{aligned} F_1 &:= A_s \\ F_{k+1} &:= T_{k,k+1}(\varphi_k)F_k \ , \ k = 1, \dots, n-1 \end{aligned}$$

Algorithmus	Transformation zu Einfacherer Gestalt	Multiplikationen pro Zyklus	Konvergenzgeschwindigkeit
Jacobi	—	$2n^3$	quadratisch ($\alpha = 2$)
QR	$5/3n^3$	$4n^2$	kubisch ($\alpha = 3$)
Schnelle QR (Gentleman)	$5/3n^3$ (weniger für schwachbesetzte A)	$2n^2$	kubisch ($\alpha = 3$)
Potenzmethode	—	n^2	linear ($\alpha = 1$)

Tabelle 7.2: Algorithmen zur Berechnung von Eigenwerten

$$T_{k,k+1}(\varphi) = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & c & & & s & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & -s & & & c & & & \\ & & & & & & & 1 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 1 \end{pmatrix}$$

$$c = \cos(\varphi)\ ,\ s = \sin(\varphi)$$

$$F_k = \begin{pmatrix} R_k & & \vdots & C_k \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \vdots\ b_k & \vdots & B_k \end{pmatrix} = (f_{ij}^{(k)})$$

R_k sei eine $k \times k$ rechte Dreiecksmatrix, $b_k^T = (f_{k+1,k}^{(k)}, 0, \ldots, 0)$

$$\begin{aligned} d_k &:= \left[\left[f_{k,k}^{(k)}\right]^2 + \left[f_{k+1,k}^{(k)}\right]^2\right]^{1/2} , \\ c_k &:= f_{k,k}^{(k)}/d_k , \\ s_k &:= f_{k+1,k}^{(k)}/d_k , \\ f_{k,j}^{(k+1)} &:= c_k f_{k,j}^{(k)} + s_k f_{k+1,j}^{(k)} ,\ k+1 \le j \le n , \\ f_{k+1,j}^{(k+1)} &:= -s_k f_{k,j}^{(k)} + c_k f_{k+1,j}^{(k)} ,\ k+1 \le j \le n . \\ f_{k,k}^{(k+1)} &:= d_k , \\ f_{k+1,k}^{(k+1)} &:= 0 . \end{aligned}$$

Insgesamt

$$\sum_{k=1}^{n-1}[4(n-k)+2] = \frac{n-1}{2}\,[4n-2+6] = 2(n^2-1) \quad \text{Multiplikationen} .$$

Sei

$$Q^T := T_{n-1,n}\ldots T_{1,2} \ ,$$

dann gilt:

$$\begin{aligned} Q^T A_s &= B_n \qquad \text{oder} \\ A_s &= QB_n \\ A_{s+1} &= B_nQ = B_nT_{1,2}\ldots T_{n-1,n} \ . \end{aligned}$$

Die Berechnung von B_nQ erfordert

$$\sum_{k=1}^{n-1} 4k+2 = \frac{n-1}{2}\,(4n-2+6) = 2(n^2-1) \quad \text{Multiplikationen} .$$

Insgesamt erfordert die Berechnung von A_{s+1} $4n^2 + 0(n)$ Multiplikationen, und A_{s+1} ist wieder eine Hessenberg Matrix.

Bemerkung: Im speziellen Fall, wenn A_s eine tridiagonale Matrix ist, ist A_{s+1} auch tridiagonal, und die Berechnung von A_{s+1} erfordert nur $0(n)$ Multiplikationen.

7.3.2 Schnelle ebene Drehungen

Sei A eine $n \times n$ Matrix mit

$$A = DBD \ ,$$

wobei D eine Diagonalmatrix ist. Dann gilt:

$$T_{pq}AT_{pq}^T = T_{pq}DBDT_{pq}^T = \overline{D}\,\overline{B}\,\overline{D} \ ,$$

wobei $\overline{D}$ und $\overline{B}$ noch nicht bestimmt sind. $\overline{D}$ sollte eine Diagonalmatrix sein mit $\overline{d}_i = d_i$ für $i \neq p, q$. Z.B.:

$$\begin{aligned} \overline{b}_{p\ell} &= \frac{d_pc}{\overline{d}_p}\,b_{p\ell} + \frac{d_qs}{\overline{d}_q}\,b_{q\ell} \ , \ \ell \neq p,q \\ \overline{b}_{q\ell} &= \frac{d_{p(-s)}}{\overline{d}_p}\,b_{p\ell} + \frac{d_qc}{\overline{d}_q}\,b_{q\ell} \ , \ \ell \neq p,q \ . \end{aligned}$$

Durch geeignete Wahl von $\overline{d}_p$ und $\overline{d}_q$ ist es möglich, die Anzahl der Multiplikationen zu halbieren. Diese Methode, die sogenannte FGR (Fast Givens Rotation) wird von Gentleman (1973, 1975) und Rath (1982) untersucht.

Literatur

Borodin, A., **Munro**, I.: The Computational Complexity of Algebraic and Numeric Problems. American Elsevier Publishing Co., Inc. 1975.

Cook, Stephen A.: An Overview of Computational Complexity. Commun. ACM, 26, No. 6, June 1983.

Hartmanis, J. (Hrsg.): Computational Complexity Theory. AMS, 1989.

Kronsjö, L.: Algorithms. Their Complexity and Efficiency. Wiley, 1987.

Mehlhorn, K.: Effiziente Algorithmen. Teubner, 1977.