Technische Universität Graz

Numerische Mathematik 1

O. Steinbach

Berichte aus dem
Institut für Mathematik D
Numerik und Partielle Differentialgleichungen

Vorlesungsskript 2005/1

Technische Universität Graz

Numerische Mathematik 1

O. Steinbach

Berichte aus dem
Institut für Mathematik
Numerik und Partielle Differentialgleichungen

Vorlesungsskript 2005/1

Technische Universität Graz
Institut für Mathematik D (Numerik und Partielle Differentialgleichungen)
Steyrergasse 30
A 8010 Graz

WWW: `http://www.numerik.math.tu-graz.at`

Vorwort

Dieses Vorlesungsskript basiert auf der Vorlesung **Numerische Mathematik 1** für Studierende der Technischen Mathematik an der Technischen Universität Graz, die ich erstmals im Wintersemester 2004/2005 gehalten habe. Der Umfang umfaßt 3 SWS Vorlesung und 1 SWS Übung. Parallel dazu habe ich die Vorlesung **Technische Numerik** für Studierende der Mechatronik im Maschinenbau (2V, 1U) gehalten.

Die Numerische Mathematik beinhaltet die Entwicklung, Analysis, Anwendung und Implementierung von Algorithmen zur näherungsweisen Lösung von komplexen Aufgabenstellungen aus verschiedenen Bereichen der Natur– und Ingenieurwissenschaften, verstärkt auch aus den Wirtschaftswissenschaften. Für Existenz– und Eindeutigkeitsanalysen bzw. für Konvergenzuntersuchen werden funktionalanalytische Methoden benutzt, während Methoden der Numerischen Linearen Algebra bei der Lösung linearer und nichtlinearer Gleichungssysteme sowie von Eigenwertproblemen zum Einsatz kommen. Darüber hinaus erfordert die Implementierung der Algorithmen entsprechende Kenntnisse aus der Informatik.

Die Lösung eines Anwendungsproblems führt von der Beschreibung des realen Prozesses auf ein analytisches mathematisches Modell, das oft durch gewöhnliche und partielle Differentialgleichungen beschrieben wird. Da dieses in der Regel nicht geschlossen lösbar ist, müssen geeignete mathematische Näherungsverfahren zur Bestimmung eines endlichen Ersatzproblems hergeleitet werden. Daraus abgeleitet wird ein numerischer Algorithmus zur Bestimmung einer Näherungslösung. Bei der Untersuchung der Stabilität und Konvergenz der Folge von Näherungslösungen sind mögliche Fehlerquellen wie z.B. der Modellierungsfehler, der Verfahrensfehler sowie Eingabe– und Rundungsfehler zu berücksichtigen. Letzteres ist bei der Fehlerfortpflanzung zu beachten.
Die Vorlesung **Numerische Mathematik 1** umfaßt zunächst die Grundlagen der Numerischen Mathematik,

– Approximation und Interpolation von Funktionen,

– Numerische Integration,

– Lösungsverfahren für Lineare Gleichungssysteme und nichtlineare Gleichungen,

– Rand– und Anfangswertprobleme für gewöhnliche Differentialgleichungen.

In der anschließenden Vorlesung **Numerische Mathematik 2** steht dann die Numerik partieller Differentialgleichungen, insbesondere die finite Element Methode, im Vordergrund. In weiterführenden Vorlesungen werden dann weitere Aspekte der Numerischen Mathematik vertieft.

Mein Dank für das Erstellen großer Teile des Manuskripts in Latex gilt Frau B. Pöltl. Ich hoffe, daß das Skript gern zur Hand genommen wird und eine gute Basis für die Beschäftigung mit der Numerischen Mathematik darstellen wird.

Graz, März 2005 O. Steinbach

Inhaltsverzeichnis

Kapitel 1

Approximation

Die polynomiale Approximation von Funktionen dient einerseits der Verarbeitung von experimentellen Daten, d.h. für eine gegebene Punktmenge $\{(x_i, f_i)\}_{i=0}^n$ ist eine geeignete funktionale Darstellung $f_n(x)$ zu finden, andererseits ermöglicht die Approximation einer gegebenen Funktion $f(x)$ durch ein Polynom eine einfache Realisierung von Differentiation und Integration. Später werden diese Konzepte zur Approximation von Funktionen auch zur Lösung von Anfangs– und Randwertproblemen gewöhnlicher und partieller Differentialgleichungen eingesetzt.

Gesucht ist eine allgemeine Darstellung der Form

$$f_n(x) = \sum_{k=0}^{n} a_k \varphi_k(x)$$

mit linear unabhängigen **Basisfunktionen** $\{\varphi_k\}_{k=0}^n$ und zu bestimmenden **Zerlegungskoeffizienten** $a_0, \ldots, a_n$. Diese sind durch eine geeignete Approximationsmethode aus den gegebenen Daten f_i bzw. aus der gegebenen Funktion $f(x)$ zu bestimmen.

1.1 Interpolation

Gegeben seien $n+1$ Paare $(x_i, f_i), i = 0, \ldots, n$, mit paarweise verschiedenen **Stützstellen** $x_i \neq x_j$ für $i \neq j$.

Gesucht ist eine Funktion

$$f_n(x) = \sum_{k=0}^{n} a_k \varphi_k(x),$$

die in den Stützstellen x_i die **Interpolationsgleichungen**

$$f_n(x_i) = \sum_{k=0}^{n} a_k \varphi_k(x_i) = f_i = f(x_i) \quad \text{für} \quad i = 0, \ldots, n$$

erfüllt.

Je nach der Wahl der Basisfunktionen φ_k unterscheidet man dabei in

- polynomiale Interpolation,
- trigonometrische Interpolation,
- stückweise polynomiale Interpolation (Splines).

Zu bestimmen sind die $n+1$ Zerlegungskoeffizienten a_k aus den $n+1$ Interpolationsgleichungen

$$\sum_{k=0}^{n} a_k \varphi_k(x_i) = f_i \quad \text{für} \quad i = 0, \dots, n.$$

Diese entsprechen dem **linearen Gleichungssystem**

$$\begin{pmatrix} \varphi_0(x_0) & \dots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \dots & \varphi_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

bzw. $A\underline{a} = \underline{f}$ mit der **Systemmatrix** $A \in \mathbb{R}^{(n+1)\times(n+1)}$ und den Einträgen

$$A[i,k] = \varphi_k(x_i) \quad \text{für} \quad i,k = 0, \dots, n.$$

Die Zerlegungskoeffizienten a_k sind genau dann eindeutig bestimmt, wenn das lineare Gleichungssystem $A\underline{a} = \underline{f}$ eindeutig lösbar ist. Zu untersuchen sind deshalb die Eigenschaften, insbesondere die Invertierbarkeit, der Systemmatrix A in Abhängigkeit der konkret gewählten Basisfunktionen $\{\varphi_k\}_{k=0}^n$.

1.1.1 Monome

Betrachtet werden als Basisfunktionen zunächst die **Monome**

$$\varphi_k(x) = x^k \quad \text{für} \quad k = 0, \dots, n$$

bzw., siehe auch Abbildung 1.1,

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x, \quad \varphi_2(x) = x^2, \quad \dots, \quad \varphi_n(x) = x^n.$$

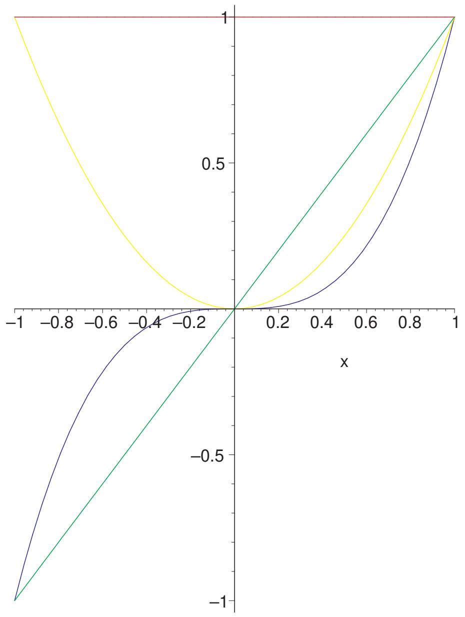

Abbildung 1.1: Monome $\varphi_k(x) = x^k$ für $k = 0, 1, 2, 3$.

Für die Einträge der Systemmatrix A folgt dann

$$A[i,k] = \varphi_k(x_i) = x_i^k \quad \text{für} \quad i,k = 0,\ldots,n$$

bzw. ist

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}$$

eine **Vandermonde**–Matrix. Für diese gilt (vgl. Übungsaufgabe 1.1.)

$$\det A = \prod_{i<j}(x_j - x_i).$$

Für paarweise verschiedene Stützstellen $x_i \neq x_j$ für alle $i \neq j$ folgt somit $\det A \neq 0$ und damit die Invertierbarkeit der Systemmatrix A. Damit ist das lineare Gleichungssystem $A\underline{a} = \underline{f}$ und somit die Interpolationsaufgabe **eindeutig** lösbar. Es zeigt sich jedoch, dass die Systemmatrix A schlecht konditioniert ist. Zu klären ist dabei zunächst der Begriff und die Bedeutung der **Kondition** einer Matrix.

Bei einer schlecht konditionierten Matrix A führt bereits eine kleine Störung der rechten Seite $\underline{f}$ zu einer großen Störung des Ergebnisvektors $\underline{a} = A^{-1}\underline{f}$. Diese Instabilität kann durch die in der Gleitkomma–Arithmetik auftretenden Rundungsfehler noch verstärkt werden.

Beispiel 1.1 *Die Lösung des linearen Gleichungssystems*

$$\begin{pmatrix} 1+\varepsilon & 1-\varepsilon \\ 1-\varepsilon & 1+\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

mit $\varepsilon > 0$ *ist gegeben durch*

$$x = \frac{1}{4\varepsilon}\Big[(1+\varepsilon)a - (1-\varepsilon)b\Big], \quad y = \frac{1}{4\varepsilon}\Big[(1+\varepsilon)b - (1+\varepsilon)a\Big].$$

Für $a = b = 2$ *folgt*

$$x = y = 1,$$

während für $a = 2$ *und* $b = 2 + \delta$

$$x = 1 + \delta\frac{1+\varepsilon}{4\varepsilon}, \quad y = 1 - \delta\frac{1-\varepsilon}{4\varepsilon}$$

gilt. Eine Störung der rechten Seite um δ *zieht also eine Störung der Lösung um* $\mathcal{O}(\frac{\delta}{\varepsilon})$ *nach sich.*

Die **Kondition** einer gegebenen Matrix A kann durch die **spektrale Konditionszahl** charakterisiert werden. Hierzu wird eine beliebige **Vektornorm** in $\mathbb{R}^{n+1}$ betrachtet, z.B. die Euklidische Vektornorm

$$||\underline{u}||_2 = \Big(\sum_{i=0}^{n} u_i^2\Big)^{\frac{1}{2}}.$$

Die durch diese Vektornorm induzierte **Matrixnorm** ist

$$||A||_2 = \sup_{0 \neq \underline{u} \in \mathbb{R}^{n+1}} \frac{||A\underline{u}||_2}{||\underline{u}||_2}.$$

Wegen

$$||A\underline{u}||_2^2 = (A\underline{u}, A\underline{u}) = (A^T A\underline{u}, \underline{u}) \leq \lambda_{max}(A^T A)(\underline{u}, \underline{\mathrm{u}}) = \lambda_{max}(A^T A)||\underline{u}||_2^2$$

und

$$\begin{aligned} ||A\underline{u}_{max}||_2^2 &= (A^T A\underline{u}_{max}, \underline{u}_{max}) = \lambda_{max}(A^T A)(\underline{u}_{max}, \underline{u}_{max}) \\ &= \lambda_{max}(A^T A)\, ||\underline{u}_{max}||_2^2 \end{aligned}$$

folgt

$$||A||_2 = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A^T A),$$

wobei $\sigma_{max}(A^T A)$ den größten **Singulärwert** von A bezeichnet. Die spektrale Konditionszahl der Matrix A ist dann definiert durch

$$\kappa_2(A) = ||A||_2 \cdot ||A^{-1}||_2.$$

Für eine symmetrische und positiv definierte Matrix A folgt

$$\kappa_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}\ .$$

Beispiel 1.2 *Die Eigenwerte der in Beispiel* 1.1 *betrachteten Matrix*

$$A = \begin{pmatrix} 1+\varepsilon & 1-\varepsilon \\ 1-\varepsilon & 1+\varepsilon \end{pmatrix}$$

sind durch

$$\lambda_1(A) = 2, \quad \lambda_2(A) = 2\varepsilon$$

gegeben. Damit ergibt sich für die spektrale Konditionszahl

$$\kappa_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = \frac{1}{\varepsilon}.$$

1.1.2 Lagrange–Polynome

Die Monome $\varphi_k(x) = x^k$ bilden den linearen Raum der Polynome vom Grad n,

$$\Pi_n = \operatorname{span}\{\varphi_k\}_{k=0}^n.$$

Jedes Polynom $f_n \in \Pi_n$ mit maximalen Grad n kann also als Linearkombination der Basisfunktionen φ_k dargestellt werden,

$$f_n(x) = \sum_{k=0}^n a_k \varphi_k(x) = \sum_{k=0}^n a_k x^k.$$

Der Übergang zu einer anderen Basis,

$$\Pi_n = \operatorname{span}\{\psi_k\}_{k=0}^n,$$

ermöglicht für das Interpolationspolynom den Ansatz

$$f_n(x) = \sum_{k=0}^n b_k \psi_k(x)$$

mit zu bestimmenden Zerlegungskoeffizienten $b_k, k = 0, \ldots, n$. Die zugehörigen Interpolationsgleichungen lauten dann

$$f_n(x_i) = \sum_{k=0}^n b_k \psi_k(x_i) = f_i \quad \text{für} \quad i = 0, \ldots, n.$$

Die Basisfunktionen ψ_k sollen nun derart gewählt werden, so daß das resultierende lineare Gleichungssystem $A\underline{b} = \underline{f}$ besonders einfach zu lösen ist. Aus der Forderung

$$\psi_k(x_i) = \begin{cases} 1 & \text{für } i = k, \\ 0 & \text{für } i \neq k \end{cases}$$

folgt dann

$$b_k = f_k \quad \text{für alle } k = 0, \ldots, n$$

und somit

$$f_n(x) = \sum_{k=0}^{n} f_k \psi_k(x).$$

Dies motiviert die Definition der **Lagrange–Polynome**, vgl. Abbildung 1.2,

$$L_k(x) = \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \quad \text{für} \quad k = 0, \ldots, n.$$

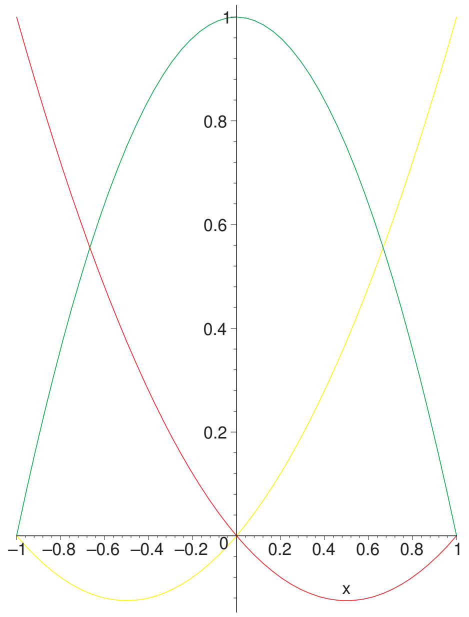

Abbildung 1.2: Lagrange–Polynome $L_k(x)$ für $x_0 = -1$, $x_1 = 0$, $x_2 = 1$.

Die Lagrange–Polynome $\{L_k\}_{k=0}^{n}$ bilden eine Basis im Raum Π_n der Polynome vom Grad n (vgl. Übungsaufgabe 1.2.).

1.1.3 Abschätzung des Interpolationsfehlers

Für eine gegebene Funktion f bezeichne $f_n \in \Pi_n$ das Interpolationspolynom mit

$$f(x_i) = f_n(x_i) \quad \text{für } i = 0, \ldots, n.$$

Dabei seien die $n + 1$ Stützstellen $x_i \in [a, b]$ in einem beschränkten Intervall $[a, b]$ gegeben. Abzuschätzen bleibt der **Fehler**

$$e_n(x) := f(x) - f_n(x) \quad \text{für } x \in [a, b].$$

In den Stützstellen $x = x_i$ folgt aus den Interpolationsgleichungen

$$e_n(x_i) \;=\; f(x_i) - f_n(x_i) \;=\; 0 \quad \text{für } i = 0, \ldots, n.$$

Für beliebiges $\bar{x} \in [a,b]$ mit $\bar{x} \neq x_i$ für $i = 0, \ldots, n$ sei

$$g_\alpha(x) \;=\; e_n(x) - \alpha \prod_{j=0}^{n} (x - x_j)$$

eine von einem skalaren Parameter $\alpha \in I\!R$ abhängige Funktion. Für diese gilt nach Konstruktion

$$g_\alpha(x_i) = 0 \quad \text{für } i = 0, \ldots, n.$$

Wegen $\bar{x} \neq x_i$ für $i = 0, \ldots, n$ ist

$$\bar{\alpha} = e_n(\bar{x}) \left[\prod_{j=0}^{n} (\bar{x} - x_j) \right]^{-1}$$

wohldefiniert und es folgt

$$g_{\bar{\alpha}}(\bar{x}) = 0.$$

Damit hat die Funktion $g_{\bar{\alpha}}(x)$ im Intervall $[a,b]$ $n+2$ voneinander verschiedene Nullstellen $x_0, \ldots, x_n$ und $\bar{x}$. Nach dem Satz von Rolle besitzt dann $g'_{\bar{\alpha}}(x)$ in $[a,b]$ $n+1$ Nullstellen und durch rekursives Anwenden folgt, daß $g_{\bar{\alpha}}^{(n+1)}(x)$ in $[a,b]$ eine Nullstelle $\xi(\bar{x}) \in [a,b]$ besitzt. Für diese ist

$$\begin{aligned}
0 &= g_{\bar{\alpha}}^{(n+1)}\Big(\xi(\bar{x})\Big) \\
&= \frac{d^{n+1}}{dx^{n+1}} \Big[f(x) - f_n(x) - \bar{\alpha} \prod_{i=0}^{n} (x - x_i) \Big]_{x=\xi(\bar{x})} \\
&= f^{(n+1)}\Big(\xi(\bar{x})\Big) - \bar{\alpha}\,(n+1)!
\end{aligned}$$

und somit

$$\bar{\alpha} = \frac{1}{(n+1)!} f^{(n+1)}\Big(\xi(\bar{x})\Big).$$

Damit gilt für den Interpolationsfehler die Darstellung

$$e_n(\bar{x}) = \frac{1}{(n+1)!} f^{(n+1)}\Big(\xi(\bar{x})\Big) \prod_{j=0}^{n} (\bar{x} - x_j), \quad \text{für } \bar{x} \in [a,b],\ \bar{x} \neq x_j.$$

Wegen $e_n(x_i) = 0$ bleibt dies offenbar für alle $x \in [a,b]$ richtig. Daraus folgt für den Interpolationsfehler die Abschätzung

$$\begin{aligned}
\max_{x \in [a,b]} |f(x) - f_n(x)| &= \max_{x \in [a,b]} \left| \frac{1}{(n+1)!} f^{(n+1)}\Big(\xi(x)\Big) \prod_{j=0}^{n} (x - x_j) \right| \\
&\leq \frac{1}{(n+1)!} \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right| \max_{x \in [a,b]} \left| \prod_{j=0}^{n} (x - x_j) \right|.
\end{aligned}$$

Diese Fehlerabschätzung zeigt, dass die Wahl der Stützstellen x_i wesentlich für die Güte der Approximation f_n ist, siehe das folgende Beispiel von Runge.

Beispiel 1.3 *Die Interpolation der Funktion*

$$f(x) = \frac{1}{1+x^2} \quad \textit{für} \quad x \in [-5, +5]$$

in den gleichmäßig verteilten Stützstellen

$$x_i = -5 + \frac{10i}{n} \quad \textit{für } i = 0, \ldots, n$$

ergibt die in Abbildung 1.3 *für* $n = 5$ *und* $n = 10$ *dargestellten Interpolationspolynome mit den in der Nähe der Randpunkte* ± 5 *auftretenden Oszillationen.*

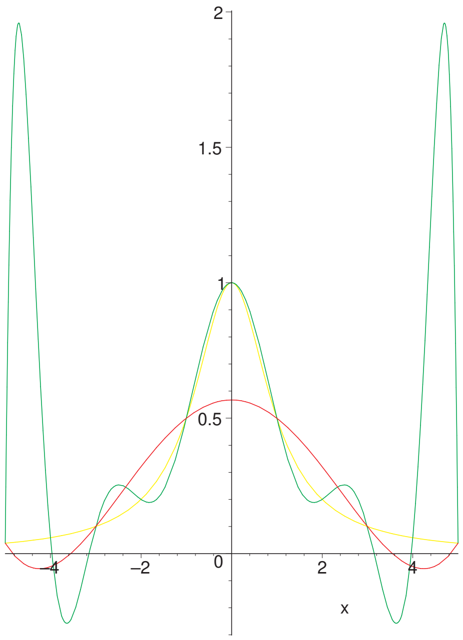

Abbildung 1.3: Interpolation der Funktion $f(x) = \dfrac{1}{1+x^2}$.

Dies motiviert die Wahl der Stützstellen x_i in einer solchen Weise, so daß

$$\max_{x \in [a,b]} \left| \prod_{j=0}^{n} (x - x_j) \right|$$

minimal wird. Die Lösung dieser Minimierungsaufgabe beruht auf den im folgenden Abschnitt behandelten Tschebyscheff–Polynome.

1.1.4 Tschebyscheff–Polynome

Der Raum Π_n der Polynome vom Grad n kann neben der Beschreibung durch die Monome bzw. durch die Lagrange–Polynome auch durch die **Tschebyscheff–Polynome** $T_k(x)$ charakterisiert werden. Diese werden rekursiv definiert durch

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \quad \text{für } k = 1, 2, \ldots \, . \end{aligned}$$

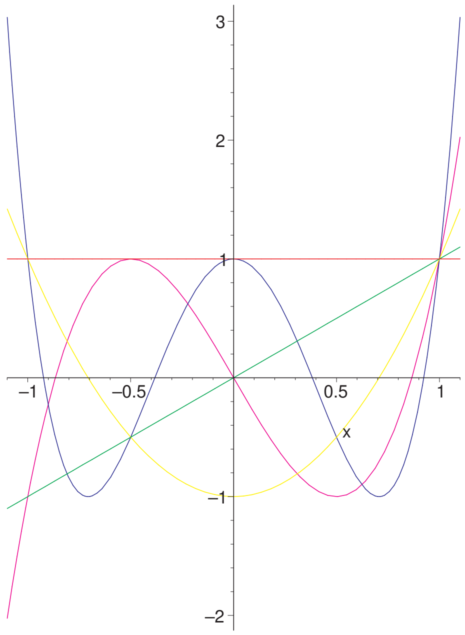

Abbildung 1.4: Tschebyscheff–Polynome $T_k(x)$ für $k = 0, \ldots, 4$.

Lemma 1.1 *Für $x \in [-1, +1]$ und $k = 0, 1, 2, \ldots$ gilt die alternative Darstellung*

$$T_k(x) = \cos(k \arccos x) \ .$$

Entsprechend gilt für $|x| > 1$

$$T_k(x) = \cosh(k \, \mathrm{arccosh} x).$$

Beweis: Übungsaufgabe 1.3. ∎

Aus der Darstellung in Lemma 1.1 lassen sich nun einige wichtige Eigenschaften der Tschebyscheff–Polynome $T_k(x)$ ablesen. Es gilt

$$\max_{x \in [-1,+1]} |T_k(x)| = 1$$

sowie

$$T_k(x_i^{(k)}) = (-1)^i \quad \text{für } x_i^{(k)} = \cos \frac{i\pi}{k}, \ i = 0, \ldots, k.$$

Die **Nullstellen** der Tschebyscheff–Polynome ergeben sich aus der Forderung

$$T_k(x) = \cos(k \ \arccos x) = 0$$

bzw. aus

$$k \ \arccos x = \frac{\pi}{2} + i\pi, \quad i \in I\!N.$$

Daraus folgt

$$\bar{x}_i^{(k)} = \cos \frac{(2i+1)\pi}{2k} \quad \text{für } i = 0, \ldots, k-1.$$

Neben der hier durch die Rekursionsvorschrift erfolgten Definition der Tschebyscheff–Polynome und der dazu äquivalenten Darstellung durch trigonometrische Polynome ermöglichen die Tschebyscheff–Polynome eine dritte Darstellung, die für die Funktionsauswertung von $T_k(x)$ für $x > 1$ wesentlich sein wird.

Lemma 1.2 *Für $k = 0, 1, 2, \ldots$ gilt*

$$\begin{aligned} T_k(x) &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\Big] \\ &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^k + (x+\sqrt{x^2-1})^{-k}\Big]. \end{aligned}$$

Beweis: Für $k = 0, 1$ ist die Behauptung offensichtlich. Für $k > 1$ erfolgt der Nachweis durch vollständige Induktion. Als Induktionsvoraussetzung gelte also

$$\begin{aligned} T_{k-1}(x) &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^{k-1} + (x-\sqrt{x^2-1})^{k-1}\Big], \\ T_k(x) &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\Big]. \end{aligned}$$

Nach Definition der Tschebyscheff–Polynome folgt

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \\ &= x\Big[(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\Big] \\ &\qquad -\frac{1}{2}\Big[(x+\sqrt{x^2-1})^{k-1} + (x-\sqrt{x^2-1})^{k-1}\Big] \\ &= \frac{1}{2}(x+\sqrt{x^2-1})^{k-1}\Big[2x\,(x+\sqrt{x^2-1})\Big] \\ &\qquad +\frac{1}{2}(x-\sqrt{x^2-1})^{k-1}\Big[2x\,(x+\sqrt{x^2-1})-1\Big]. \end{aligned}$$

Die erste Behauptung folgt nun aus

$$2x(x \pm \sqrt{x^2-1}) - 1 = x^2 + 2x\sqrt{x^2-1} + x^2 - 1 = (x \pm \sqrt{x^2-1})^2.$$

Die zweite Behauptung folgt unmittelbar aus

$$x - \sqrt{x^2-1} = \frac{[x-\sqrt{x^2-1}\;][x+\sqrt{x^2-1}\;]}{x+\sqrt{x^2-1}} = \frac{1}{x+\sqrt{x^2-1}}.$$

∎

Die erste Darstellung in Lemma 1.1 gilt nur für Argumente $x \in [-1, +1]$. Sei $[a, b]$ ein beliebiges Intervall mit $0 < a < b$. Für $t \in [a, b]$ ermöglicht die Transformation

$$x := \frac{b+a-2t}{b-a} \in [-1, +1]$$

die Definition der skalierten Tschebyscheff–Polynome

$$\widetilde{T}_k(t) := \frac{T_k\left(\frac{b+a-2t}{b-a}\right)}{T_k\left(\frac{b+a}{b-a}\right)}$$

mit $\widetilde{T}_k(0) = 1$. Ist Π_n der Raum aller Polynome maximalen Grades n, so umfasst Π_n^1 alle Polynome $f_n \in \Pi_n$ mit $f_n(0) = 1$, d.h. $\widetilde{T}_k \in \Pi_n^1$. Die modizierten Tschebyscheff–Polynome $\widetilde{T}_k(t)$ sind die Polynome vom Polynomgrad k mit dem kleinsten Maximum im Intervall $[a, b]$:

Satz 1.1 *Für $0 < a < b$ sind die modifizierten Tschebyscheff–Polynome $\widetilde{T}_k(t)$ Lösung der Minimierungsaufgabe*

$$\min_{p_k \in \Pi_k^1} \max_{t \in [a,b]} |p_k(t)| \;=\; \max_{t \in [a,b]} |\widetilde{T}_k(t)| \;=\; \frac{2q^k}{1+q^{2k}}, \quad q \;=\; \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}}\,.$$

Beweis: Der Beweis erfolgt indirekt durch die Annahme, es existiere ein Polynom $q_k \in \Pi_k^1$ mit

$$\max_{t\in[a,b]} |q_k(t)| \;<\; \max_{t\in[a,b]} |\widetilde{T}_k(t)|\,.$$

Für

$$t_i^{(k)} \;:=\; \frac{1}{2}\left[(b+a)-(b-a)x_i^{(k)}\right], \quad x_i^{(k)} = \cos\frac{i\pi}{k}, \quad i=0,\ldots,k$$

wird durch

$$\widetilde{T}_k(t_i^{(k)}) \;=\; \frac{T_k(x_i^{(k)})}{T_k(\frac{b+a}{b-a})} \;=\; \frac{(-1)^i}{T_k(\frac{b+a}{b-a})}$$

das Maximum bzw. das Minimum des modifizierten Tschebyscheff–Polynoms $\widetilde{T}_k(t)$ angenommen. Dann gilt

$$|q_k(t_i^{(t_k)})| \;\le\; \max_{t\in[a,b]}|q_k(t)| \;<\; \frac{1}{T_k(\frac{b+a}{b-a})} \quad \text{für } i=0,\ldots,k.$$

Insbesondere für $i=2j$ ist $\widetilde{T}_k(t_{2j}^{(k)})>0$ und somit ist

$$-\widetilde{T}_k(t_{2j}^{(k)}) < q_k(t_{2j}^{(k)}) < \widetilde{T}_k(t_{2j}^{(k)}).$$

Entsprechend ergibt sich für $i=2j+1$ $\widetilde{T}_k(t_{2j+1}^{(k)})<0$ und somit

$$\widetilde{T}_k(t_{2j+}^{(k)}) < q_k(t_{2j+1}^{(k)}) < -\widetilde{T}_k(t_{2j+}^{(k)}).$$

Für das Polynom $r_k := \widetilde{T}_k - q_k \in \Pi_k$ folgt dann

$$r_k(t_{2j}^{(k)}) \;=\; \widetilde{T}_k(t_{2j}^{(k)}) - q_k(t_{2j}^{(k)}) \;>\; 0$$

und

$$r_k(t_{2j+1}^{(k)}) \;=\; \widetilde{T}_k(t_{2j+1}^{(k)}) - q_k(t_{2j+1}^{(k)}) \;<\; 0\,.$$

Zwischen den $k+1$ Stellen $t_i^{(k)}$ finden also k Vorzeichenwechsel statt, d.h. das Polynom $r_k(t)$ besitzt im Intervall $[a,b]$ mindestens k Nullstellen. Wegen $\widetilde{T}_k \in \Pi_k^1$ und $q_k \in \Pi_k^1$ ist

$$r_k(0) \;=\; \widetilde{T}_k(0) - q_k(0) \;=\; 1-1 \;=\; 0$$

und somit ist Null eine weitere Nullstelle von $r_k(t)$. Damit besitzt das Polynom $r_k(t)$ vom Polynomgrad k auf der reellen Achse mindestens $k+1$ Nullstellen. Daraus folgt $r_k(t)\equiv 0$ für alle $t\in\mathbb{R}$ und somit $q_k = \widetilde{T}_k$ im Widerspruch zur Annahme.
Zu bestimmen bleibt der maximale Wert von

$$\max_{t\in[a,b]}|\widetilde{T}_k(t)| \;=\; \frac{1}{T_k(\frac{b+a}{b-a})}\,.$$

Mit Lemma 1.2 und

$$\begin{aligned} q \;&=\; \frac{b+a}{b-a} + \sqrt{\left(\frac{b+a}{b-a}\right)^2 - 1} \;=\; \frac{1}{b-a}\left[(b+a)+\sqrt{(b+a)^2-(b-a)^2}\right] \\ &=\; \frac{1}{b-a}\left[b+a+2\sqrt{ab}\right] \;=\; \frac{(\sqrt{b}+\sqrt{a})^2}{(\sqrt{b}+\sqrt{a})(\sqrt{b}-\sqrt{a})} \;=\; \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}} \end{aligned}$$

ist schließlich

$$T_k(\frac{b+a}{b-a}) \;=\; \frac{1}{2}\left[q^k+q^{-k}\right] \;=\; \frac{q^{2k}+1}{2q^k}\,.$$

■

Für das Tschebyscheff–Polynom $T_{k+1} \in \Pi_{k+1}$ gilt die Darstellung

$$T_{k+1}(x) = \alpha \prod_{i=0}^{k} \left(x - \bar{x}_i^{(k+1)} \right)$$

mit den Nullstellen $\bar{x}_i^{(k+1)}$ von T_{k+1}. Andererseits ergibt sich aus der rekursiven Definition der Tschebyscheff–Polynome

$$T_{k+1}(x) = 2^k x^{k+1} + g_k(x)$$

mit einem Polynom $g_k \in \Pi_k$ vom Polynomgrad k. Durch Vergleich der führenden Koeffizienten folgt somit

$$T_{k+1}(x) = 2^k \prod_{i=0}^{k} \left(x - \bar{x}_i^{(k+1)} \right).$$

Damit ergibt sich für die Minimierungsaufgabe zur Bestimmung der optimalen Stützstellen für die polynomiale Interpolationsaufgabe im Intervall $[-1, +1]$,

$$\min_{x_j \in [-1,+1]} \max_{x \in [-1,+1]} \left| \prod_{j=0}^{n} (x - x_j) \right| = \left| \prod_{i=0}^{n} \left(x - \bar{x}_i^{(n+1)} \right) \right| = 2^{-n} \max_{x \in [-1,+1]} |T_{n+1}(x)| = 2^{-n},$$

und somit die Fehlerabschätzung

$$\max_{x \in [-1,+1]} |f(x) - f_n(x)| \leq \frac{2^{-n}}{(n+1)!} \max_{x \in [-1,+1]} \left| f^{(n+1)}(x) \right|,$$

falls das Interpolationspolynom $f_n(x)$ die Interpolationsgleichungen

$$f_n\left(\bar{x}_i^{(n+1)}\right) = f\left(\bar{x}_i^{(n+1)}\right) \quad \text{für} \quad i = 0, \ldots, n$$

in den Nullstellen $\bar{x}_i^{(n+1)}$ von $T_{n+1}(x)$ erfüllt. Durch eine geeignete Transformation können die Stützstellen der Interpolationsaufgabe auf ein beliebiges Intervall $[a, b]$ übertragen werden.
Zu bestimmen bleiben die Zerlegungskoeffizienten von f_n als Lösung des zugehörigen linearen Gleichungssystems. Der Ansatz

$$f_n(x) = \sum_{k=0}^{n} a_k T_k(x)$$

mit Tschebyscheff–Polynomen $T_k(x)$ führt dann auf die Interpolationsgleichungen

$$f_n\left(\bar{x}_i^{(n+1)}\right) = \sum_{k=0}^{n} a_k T_k\left(\bar{x}_i^{(n+1)}\right) = f_i \quad \text{für } i = 0, \ldots, n.$$

Zur Bestimmung der Zerlegungskoeffizienten a_k werden die Interpolationsgleichungen mit $T_\ell(\bar{x}_i^{(n+1)})$ multipliziert und über $i = 0, \ldots, n$ summiert,

$$\sum_{i=0}^{n} \sum_{k=0}^{n} a_k T_k\left(\bar{x}_i^{(n+1)}\right) T_\ell\left(\bar{x}_i^{(n+1)}\right) = \sum_{i=0}^{n} f_i T_\ell\left(\bar{x}_i^{(n+1)}\right) .$$

Es gilt

$$\sum_{i=0}^{n} T_k\left(\bar{x}_i^{(n+1)}\right) T_\ell\left(\bar{x}_i^{(n+1)}\right) = \begin{cases} 0 & \text{für } k \neq \ell, \\ \frac{1}{2}(n+1) & \text{für } k = \ell \neq 0, \\ n+1 & \text{für } k = \ell = 0 \end{cases}$$

und somit

$$a_0 = \frac{1}{(n+1)} \sum_{i=0}^{n} f_i \quad \text{für } k = 0$$

bzw.

$$a_k = \frac{2}{n+1} \sum_{i=0}^{n} f_i T_k \left(\bar{x}_i^{(n+1)} \right) \quad \text{für } k = 1, \ldots, n.$$

Dies ist gleichbedeutend mit

$$a_k = \frac{2}{n+1} \sum_{i=0}^{n} f_i \cos k \frac{(2i+1)\pi}{2(n+1)} \quad \text{für } k = 1, \ldots, n.$$

Eine effiziente Berechnung der Zerlegungskoeffizienten a_k kann schließlich durch eine **schnelle Fouriertransformation** realisiert werden, siehe zum Beispiel [14].

Die bisher verwendeten Ansatzfunktionen zur Bestimmung des Interpolationspolynom sind **global**, d.h. sie sind stets im gesamten Intervall $[a, b]$ auszuwerten. Die Anwendung der Fehlerabschätzung für ein Interpolationspolynom n–ten Grades erfordert darüberhinaus die Stetigkeit der $(n+1)$–ten Ableitung der zu interpolierenden Funktion. Für viele Anwendungen ist dies aber eine zu starke Restriktion. Deshalb sollen im folgenden Approximationsmethoden betrachtet werden, die neben **lokalen Ansatzfunktionen** auch Fehlerabschätzungen für Funktionen mit geringerer Regularität ermöglichen.

1.1.5 Stückweise lineare Ansatzfunktionen

Gegeben seien im Intervall $[a, b]$ $n+1$ gleichmäßig verteilte Stützstellen x_i mit

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b.$$

Offenbar gilt

$$x_i = a + i\,\frac{b-a}{n} = a + ih \quad \text{für } i = 0, \ldots, n$$

mit der **Schrittweite**

$$h = \frac{b-a}{n} \to 0 \quad \text{für } n \to \infty.$$

In den Stützstellen x_i seien zugehörige Funktionswerte $f_i = f(x_i)$ gegeben. Die Approximation $f_n(x)$ ist dann wie in Abbildung 1.5 definiert als stückweise lineare Funktion

$$f_n(x) = f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} [f_i - f_{i-1}] \quad \text{für } x \in [x_{i-1}, x_i],\ i = 1, \ldots, n.$$

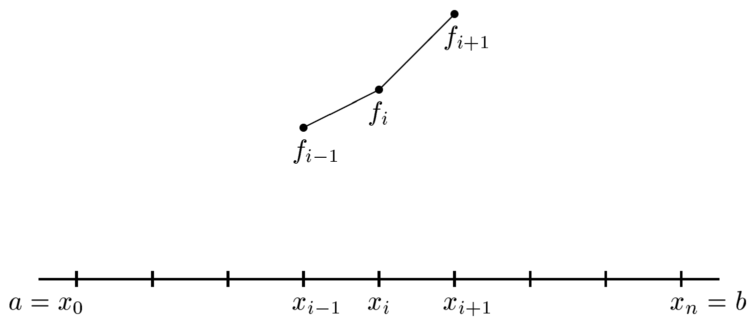

Abbildung 1.5: Stückweise lineare Interpolation.

Für die Abschätzung des Interpolationsfehlers $f(x) - f_n(x)$ führt die Taylor-Entwicklung von $f(x)$ für $x \in (x_{i-1}, x_i)$ um x_{i-1} auf

$$f(x) = f(x_{i-1}) + (x - x_{i-1}) f'(x_{i-1}) + \frac{1}{2} f''(\xi)(x - x_{i-1})^2$$

mit einer Zwischenwertstelle $\xi \in (x_{i-1}, x)$. Entsprechend gilt für $x = x_i$

$$f_i = f(x_i) = f(x_{i-1}) + (x_i - x_{i-1})f'(x_{i-1}) + \frac{1}{2}f''(\bar{\xi})(x_i - x_{i-1})^2$$

mit $\bar{\xi} \in (x_{i-1}, x_i)$. Für $x \in (x_{i-1}, x_i)$ folgt dann, unter Ausnutzung der Interpolationsbedingung $f_{i-1} = f(x_{i-1})$,

$$\begin{aligned}
f(x) - f_n(x) &= f(x_{i-1}) + (x - x_{i-1})f'(x_{i-1}) + \frac{1}{2}f''(\xi)(x - x_{i-1})^2 \\
&\qquad -\Big[f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}}(f_i - f_{i-1})\Big] \\
&= (x - x_{i-1})f'(x_{i-1}) + \frac{1}{2}f''(\xi)(x - x_{i-1})^2 \\
&\qquad -\frac{x - x_{i-1}}{x_i - x_{i-1}}\Big[(x_i - x_{i-1})f'(x_{i-1}) + \frac{1}{2}f''(\bar{\xi})(x_i - x_{i-1})^2\Big] \\
&= \frac{1}{2}f''(\xi)(x - x_{i-1})^2 - \frac{1}{2}f''(\bar{\xi})(x - x_{i-1})(x_i - x_{i-1})
\end{aligned}$$

und somit

$$\begin{aligned}
|f(x) - f_n(x)| &\le \frac{1}{2}|f''(\xi)|(x - x_{i-1})^2 + \frac{1}{2}|f''(\bar{\xi})||x - x_{i-1}||x_i - x_{i-1}| \\
&\le \frac{1}{2}h^2\Big[|f''(\xi)| + |f''(\bar{\xi})|\Big] \le h^2 \max_{\xi \in (x_{i-1}, x_i)} |f''(\xi)|.
\end{aligned}$$

Damit ergibt sich als Fehlerabschätzung in der Maximum–Norm

$$\max_{x \in (x_{i-1}, x_i)} |f(x) - f_n(x)| \le h^2 \max_{\xi \in (x_{i-1}, x_i)} |f''(\xi)| \quad \text{für } i = 1, \ldots, n,$$

bzw.

$$\max_{x \in [a,b]} |f(x) - f_n(x)| \le h^2 \max_{\xi \in [a,b]} |f''(\xi)|.$$

Im folgenden soll eine Abschätzung des Interpolationsfehlers in der L_2–Norm

$$\int_a^b \Big[f(x) - f_n(x)\Big]^2 dx$$

gewonnen werden. Für $x \in (x_{i-1}, x_i)$ ergibt sich zunächst aus den Interpolationsgleichungen $f(x_{i-1}) = f_n(x_{i-1})$

$$f(x) - f_n(x) = f(x) - f(x_{i-1}) - f_n(x) + f_n(x_{i-1}) = \int_{x_{i-1}}^{x} [f'(\xi) - f_n'(\xi)]d\xi.$$

Dann gilt, unter Verwendung der Cauchy–Schwarz–Ungleichung,

$$\begin{aligned}
|f(x) - f_n(x)|^2 &= \Big|\int_{x_{i-1}}^{x} [f'(\xi) - f_n'(\xi)]d\xi\Big|^2 \\
&\le \left[\int_{x_{i-1}}^{x} 1 \cdot |f'(\xi) - f_n'(\xi)|d\xi\right]^2 \\
&\le \int_{x_{i-1}}^{x} 1^2 d\xi \cdot \int_{x_{i-1}}^{x} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi \\
&\le (x - x_{i-1}) \cdot \int_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi \quad \text{für } x \in (x_{i-1}, x_i).
\end{aligned}$$

Integration nach $x \in (x_{i-1}, x_i)$ liefert

$$\int\limits_{x_{i-1}}^{x_i} \Big[f(x) - f_n(x)\Big]^2 dx \leq \frac{1}{2} h^2 \int\limits_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi.$$

Wegen

$$\int\limits_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big] d\xi \; = \; f(x_i) - f(x_{i-1}) - f_n(x_i) + f_n(x_{i-1}) = 0$$

ergibt sich mit $e_h(\xi) = f(\xi) - f_n(\xi)$

$$\begin{aligned}
\int\limits_{x_{i-1}}^{x_i} \Big[e_h'(\xi)\Big]^2 d\xi \; &= \; \int\limits_{x_{i-1}}^{x_i} \Big[e_h'(\xi) - \frac{1}{h} \int\limits_{x_{i-1}}^{x_i} [e_h'(s)] ds\Big]^2 d\xi \\
&= \; \int\limits_{x_{i-1}}^{x_i} \Big[\frac{1}{h} \int\limits_{x_{i-1}}^{x_i} [e_h'(\xi) - e_h'(s)] ds\Big]^2 d\xi \\
&= \; \frac{1}{h^2} \int\limits_{x_{i-1}}^{x_i} \Big[\int\limits_{x_{i-1}}^{x_i} 1 \cdot \int\limits_{s}^{\xi} e_h''(t) dt \; ds\Big]^2 d\xi \\
&\leq \; \frac{1}{h^2} \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-1}}^{x_i} 1^2 ds \int\limits_{x_{i-1}}^{x_i} \Big[\int\limits_{s}^{\xi} e_h''(t) dt\Big]^2 ds \; d\xi \\
&= \; \frac{1}{h} \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-1}}^{x_i} \Big[\int\limits_{s}^{\xi} f''(t) dt\Big]^2 ds \; d\xi
\end{aligned}$$

wegen $e_h''(\xi) = f''(\xi)$ für die lineare Approximation $f_n(x)$, $x \in (x_{i-1}, x_i)$.
Durch wiederholte Anwendung der Cauchy–Schwarz–Ungleichung ist weiterhin

$$\begin{aligned}
\int\limits_{x_{i-1}}^{x_i} \Big[e_h'(\xi)\Big]^2 d\xi \; &\leq \; \frac{1}{h} \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-1}}^{x_i} \Big[\int\limits_{s}^{\xi} 1 \cdot f''(t) dt\Big]^2 ds \; d\xi \\
&\leq \; \frac{1}{h} \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-1}}^{x_i} \Big|\int\limits_{s}^{\xi} 1^2 dt\Big| \Big|\int\limits_{s}^{\xi} |f''(t)|^2 dt\Big| ds \; d\xi \\
&\leq \; \frac{1}{h} \int\limits_{x_{i-1}}^{x_i} \int\limits_{x_{i-1}}^{x_i} |\xi - s| \int\limits_{x_{i-1}}^{x_i} |f''(t)|^2 dt \; ds \; d\xi \\
&\leq \; h^2 \int\limits_{x_{i-1}}^{x_i} |f''(t)|^2 dt.
\end{aligned}$$

Insgesamt gilt also

$$\int\limits_{x_{i-1}}^{x_i} \Big[f(x) - f_n(x)\Big]^2 dx \; \leq \; \frac{1}{2} \, h^4 \int\limits_{x_{i-1}}^{x_i} [f''(x)]^2 dx$$

und durch Summation über alle Elemente (x_{i-1}, x_i) folgt die gewünschte Fehlerabschätzung

$$\int\limits_{a}^{b} \Big[f(x) - f_n(x)\Big]^2 dx \; \leq \; \frac{1}{2} \, h^4 \int\limits_{a}^{b} [f''(x)]^2 dx$$

bzw.

$$\|f - f_n\|_{L_2[a,b]} \leq \frac{1}{\sqrt{2}}\, h^2 \, \|f''\|_{L_2[a,b]}.$$

Für eine globale Darstellung der stückweise linear Interpolierenden

$$f_n(x) = \sum_{k=0}^{n} a_k \varphi_k(x)$$

können Basisfunktionen durch die Forderung

$$\varphi_k(x) = \begin{cases} 1 & \text{für } x = x_k, \\ 0 & \text{für } x = x_\ell \neq x_k, \\ \text{stückweise linear} & \text{sonst} \end{cases}$$

definiert werden, vergleiche hierzu auch Abbildung 1.6.

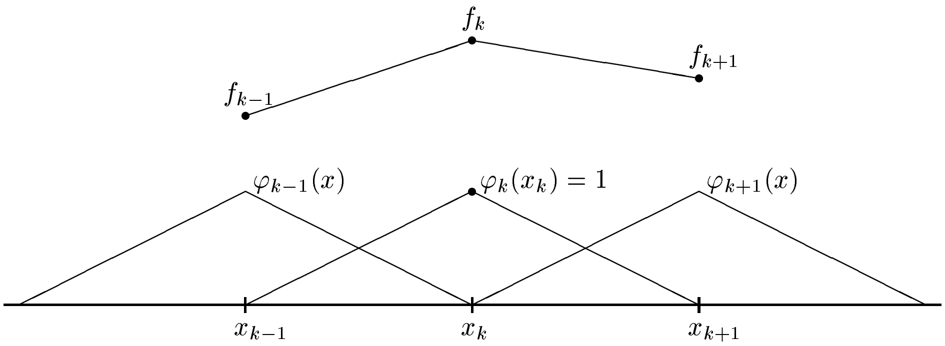

Abbildung 1.6: Ansatzfunktionen $\varphi_k(x)$ sowie $\varphi_{k\pm1}(x)$.

Für die Basisfunktionen $\varphi_k(x)$ ergibt sich daraus die funktionale Darstellung

$$\varphi_k(x) = \begin{cases} \dfrac{x - x_{k-1}}{x_k - x_{k-1}} & \text{für } x \in [x_{k-1}, x_k] \\[2ex] \dfrac{x_{k+1} - x}{x_{k+1} - x_k} & \text{für } x \in [x_k, x_{k+1}] \\[2ex] 0 & \text{sonst} \end{cases}$$

für $k = 0, \ldots, n$. Analog können auf diese Weise auch stückweise konstante Basisfunktionen

$$\psi_k(x) = \begin{cases} 1 & \text{für } x \in (x_{k-1}, x_k), \\ 0 & \text{sonst} \end{cases}$$

für $k = 1, \ldots, n$ erklärt werden.

Die Verwendung lokaler Basisfunktionen, z.B. stückweise linearer Ansatzfunktionen, ermöglicht eine einfache Auswertung des Interpolationspolynoms. Jedoch verlangt die Interpolationsaufgabe die Stetigkeit der zu approximierenden Funktion. Diese Voraussetzung kann durch die Verwendung von Projektionsmethoden vermieden werden.

1.2 Projektionsmethoden

Gesucht ist jetzt eine Approximation

$$f_n(x) = \sum_{k=0}^{n} a_k \varphi_k(x)$$

mit zunächst beliebigen Ansatzfunktionen $\varphi_k(x)$, die den zugehörigen Fehler

$$\int_a^b [f(x) - f_n(x)]^2 dx$$

minimiert. Zu lösen ist also das Minimierungsproblem

$$F(\underline{a}) = \min_{\underline{b} \in \mathbb{R}^{n+1}} F(\underline{b})$$

mit dem Funktional

$$\begin{aligned} F(\underline{b}) &= \int_a^b \Big[f(x) - \sum_{k=0}^n b_k \varphi_k(x)\Big]^2 dx \\ &= \int_a^b [f(x)]^2 dx - 2\sum_{k=0}^n b_k \int_a^b f(x)\varphi_k(x)dx + \sum_{k=0}^n \sum_{\ell=0}^n b_k b_\ell \int_a^b \varphi_k(x)\varphi_\ell(x)dx. \end{aligned}$$

Aus der notwendigen Minimierungsbedingung

$$\frac{\partial}{\partial b_i} F(\underline{b}) = 0 \quad \text{für alle } i = 0, \ldots, n$$

für den Lösungsvektor $\underline{a} \in \mathbb{R}^{n+1}$ folgt

$$\begin{aligned} 0 &= \frac{\partial}{\partial b_i}\left[\int_a^b [f(x)]^2 dx - 2\sum_{k=0}^n b_k \int_a^b f(x)\varphi_k(x)ds + \sum_{k=0}^n \sum_{\ell=0}^n b_k b_\ell \int_a^b \varphi_k(x)\varphi_\ell(x)\right] \\ &= -2\int_a^b f(x)\varphi_i(x)ds + \frac{\partial}{\partial b_i}\left[b_i^2 \int_a^b \varphi_i(x)\varphi_i(x)dx\right. \\ &\qquad \left. + \sum_{\substack{k=0\\k\neq i}}^n \sum_{\ell=0}^n b_k b_\ell \int_a^b \varphi_k(x)\varphi_\ell(x)dx + \sum_{\substack{\ell=0\\\ell\neq i}}^n b_i b_\ell \int_a^b \varphi_i(x)\varphi_\ell(x)\right] \\ &= -2\int_a^b f(x)\varphi_i(x)dx + 2b_i \int_a^b \varphi_i(x)\varphi_i(x)dx \\ &\qquad + \sum_{\substack{k=0\\k\neq i}}^n b_k \int_a^b \varphi_k(x)\varphi_i(x)dx + \sum_{\substack{\ell=0\\\ell\neq i}}^n b_\ell \int_a^b \varphi_i(x)\varphi_\ell(x)dx \\ &= -2\int_a^b f(x)\varphi_i(x)dx + 2\sum_{k=0}^n b_k \int_a^b \varphi_k(x)\varphi_i(x)dx. \end{aligned}$$

Dies ist gleichbedeutend mit

$$\sum_{k=0}^n a_k \int_a^b \varphi_k(x)\varphi_i(x)dx = \int_a^b f(x)\varphi_i(x)dx \quad \text{für } i = 0, \ldots, n,$$

bzw. mit dem linearen Gleichungssystem

$$M_h \underline{a} = \underline{f}_h$$

mit der **Massematrix**

$$M_h[i,k] = \int\limits_a^b \varphi_k(x)\varphi_i(x)dx$$

und dem Vektor der rechten Seite,

$$f_i = \int\limits_a^b f(x)\varphi_i(x)dx,$$

für $i, k = 0, \ldots, n$.
Die voneinander linear unabhängigen Basisfunktionen $\{\varphi_k\}_{k=0}^n$ bilden einen linearen Raum

$$S_h = \mathrm{span}\{\varphi_k\}_{k=0}^n.$$

Die Approximation $f_h \in S_h$ ist also Lösung des **Variationsproblems**

$$\int\limits_a^b f_h(x)\varphi_i(x)dx = \int\limits_a^b f(x)\varphi_i(x)dx \quad \text{für } i = 0, \ldots, n$$

bzw. von

$$\int\limits_a^b f_h(x)g_h(x)dx = \int\limits_a^b f(x)g_h(x)dx \quad \text{für alle } g_h \in S_h.$$

Offenbar gilt die **Galerkin–Orthogonalität**

$$\int\limits_a^b [f(x) - f_h(x)]g_h(x)dx = 0 \quad \text{für alle } g_h \in S_h.$$

Die Approximation $f_h = Q_h f \in S_h$ wird als $\boldsymbol{L_2}$**–Projektion** von f bezeichnet.
Für den Fehler $f(x) - Q_h f(x)$ folgt in der L_2–Norm mit der Galerkin–Orthogonalität

$$\begin{aligned}
||f - Q_h f||^2_{L_2[a,b]} &= \int\limits_a^b [f(x) - Q_h f(x)][f(x) - Q_h f(x)]dx \\
&= \int\limits_a^b [f(x) - Q_h f(x)][f(x) - g_h(x)]dx + \int\limits_a^b [f(x) - Q_h f(x)][g_h(x) - f_h(x)]dx \\
&= \int\limits_a^b [f(x) - Q_h f(x)][f(x) - g_h(x)]dx \\
&\leq ||f - Q_h f||_{L_2[a,b]} ||f - g_h||_{L_2[a,b]}
\end{aligned}$$

für alle $g_h \in S_h$ und somit, in Übereinstimmung mit der ursprünglichen Minimierungsaufgabe $F(\underline{a}) \leq F(\underline{b})$ für alle $\underline{b} \in \mathbb{R}^{n+1}$,

$$||f - Q_h f||_{L_2[a,b]} \leq ||f - g_h||_{L_2[a,b]} \quad \text{für alle } g_h \in S_h$$

bzw.

$$||f - Q_h f||_{L_2[a,b]} = \min_{g_h \in S_h} ||f - g_h||_{L_2[a,b]}.$$

Für stückweise lineare Ansatzfunktionen ergibt sich für die Einträge der Massematrix

$$\begin{aligned}
M_h[i,k] &= \int\limits_a^b \varphi_k(x)\varphi_i(x)dx = 0 \quad \text{für } k \neq i, i \pm 1, \\
M_h[k \pm 1, k] &= \int\limits_{x_k}^{x_{k+1}} \frac{x_{k+1}-x}{x_{k+1}-x_k}\frac{x-x_k}{x_{k+1}-x_k}dx = \frac{1}{h_{k+1}^2}\int\limits_0^{h_{k+1}}(h_{k+1}-t)tdt = \frac{1}{6}h_{k+1}, \\
M_h[k,k] &= \int\limits_a^b [\varphi_k(x)]^2 dx = \int\limits_{x_{k-1}}^{x_k}\left[\frac{x-x_{k-1}}{x_k-x_{k-1}}\right]^2 dx + \int\limits_{x_k}^{x_{k+1}}\left[\frac{x_{k+1}-x}{x_{k+1}-x_k}\right]^2 dx \\
&= \frac{1}{h_k^2}\int\limits_0^{h_k} t^2 dt + \frac{1}{h_{k+1}^2}\int\limits_0^{h_k+1}(h_{k+1}-t)^2 dt = \frac{1}{3}h_k + \frac{1}{3}h_{k+1}.
\end{aligned}$$

Für eine gleichmäßige Unterteilung mit $h_k = h$ für alle $k = 1, \ldots, n$ folgt somit

$$M_h = \frac{h}{6}\begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & \ddots & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix}.$$

Die Massematrix M_h ist

- symmetrisch und positiv definit;
- schwach besetzt, d.h. M_h besitzt $2+3(n-1)+2$ **Nichtnulleinträge**;
- Tridiagonalmatrix;
- von hierarchischer Struktur.

Damit können für die Lösung des linearen Gleichungssystems $M_h\underline{a} = \underline{f}$ effiziente Lösungsverfahren verwendet werden. Darauf soll an dieser Stelle jedoch nicht weiter eingegangen werden [14].
Aus der Abschätzung des Interpolationsfehlers für stückweise lineare Ansatzfunktionen folgt schließlich die Fehlerabschätzung

$$||f - Q_h f||_{L_2[a,b]} \leq ||f - I_h f||_{L_2[a,b]} \leq \frac{1}{\sqrt{2}}\, h^2\, ||f''||_{L_2[a,b]}.$$

Diese erfordert die Quadratintegrierbarkeit der zweiten Ableitung $f''(x)$ der zu approximierenden Funktion $f(x)$. Am Beispiel der L_2–Projektion mit stückweise konstanten Ansatzfunktionen sollen nun die Fälle untersucht werden, wenn die zu approximierende Funktion eine geringere Regularität aufweist.
Mit den stückweise konstanten Ansatzfunktionen

$$\psi_k(x) = \begin{cases} 1 & \text{für } x \in (x_{k-1}, x_k) \\ 0 & \text{sonst} \end{cases}$$

für $k = 1, \ldots, n$ führt der Ansatz

$$(Q_h f)(x) = \sum_{k=1}^n a_k \psi_k(x)$$

auf das Variationsproblem

$$\sum_{k=1}^{n} a_k \int_a^b \psi_k(x)\psi_i(x)dx \;=\; \int_a^b f(x)\psi_i(x)dx \quad \text{für } i=1,\ldots,n.$$

Wegen

$$\int_a^b \psi_k(x)\psi_i(x)dx \;=\; \begin{cases} h_k & \text{für } k=i, \\ 0 & \text{sonst} \end{cases}$$

folgt

$$a_k = \frac{1}{h_k}\int_a^b f(x)\psi_k(x)dx = \frac{1}{h_k}\int_{x_{k-1}}^{x_k} f(x)dx.$$

Für $x \in (x_{k-1}, x_k)$ ist $(Q_h f)(x) = a_k\psi_k(x) = a_k$ und somit

$$\begin{aligned} f(x) - Q_h f(x) &= f(x) - \frac{1}{h_k}\int_{x_{k-1}}^{x_k} f(y)dy \\ &= \frac{1}{h_k}\int_{x_{k-1}}^{x_k} \Big[f(x) - f(y)\Big]dy \;=\; \frac{1}{h_k}\int_{x_{k-1}}^{x_k}\int_y^x f'(s)ds\,dy. \end{aligned}$$

Die wiederholte Anwendung der Cauchy–Schwarz–Ungleichung liefert

$$\begin{aligned} \Big[f(x) - Q_h f(x)\Big]^2 &= \frac{1}{h_k^2}\left[\int_{x_{k-1}}^{x_k} 1\cdot\int_y^x f'(s)ds\,dy\right]^2 \\ &\le \frac{1}{h_k^2}\int_{x_{k-1}}^{x_k} 1^2 dy \int_{x_{k-1}}^{x_k}\left[\int_y^x f'(s)ds\right]^2 dy \;=\; \frac{1}{h_k}\int_{x_{k-1}}^{x_k}\left[\int_y^x 1\cdot f'(s)ds\right]^2 dy \\ &\le \frac{1}{h_k}\int_{x_{k-1}}^{x_k}\left|\int_y^x 1^2 ds\right|\cdot\left|\int_y^x [f'(s)]^2 ds\right| dy \;\le\; \frac{1}{h_k}\int_{x_{k-1}}^{x_k}|x-y|\int_{x_{k-1}}^{x_k}[f'(s)]^2 ds\,dy \\ &\le h_k\int_{x_{k-1}}^{x_k}[f'(s)]^2 ds. \end{aligned}$$

Integration bezüglich $x \in (x_{k-1}, x_k)$ ergibt

$$\int_{x_{k-1}}^{x_k}\Big[f(x) - Q_h f(x)\Big]^2 dx \;\le\; h_k^2\int_{x_{k-1}}^{x_k}[f'(s)]^2 ds$$

und somit

$$||f - Q_h f||^2_{L_2[a,b]} \;\le\; \sum_{k=1}^{n} h_k^2\int_{x_{k-1}}^{x_k}[f'(s)]^2 ds$$

bzw.

$$||f - Q_h f||_{L_2[a,b]} \;\le\; h\,||f'||_{L_2[a,b]}$$

mit $h = \max_{k=1,\ldots,n} h_k$. Die Voraussetzung der Quadratintegrierbarkeit der ersten Ableitung $f'(x)$ gewährleistet also ein lineares Konvergenzverhalten für den Fehler $f(x) - f_h(x)$ in der L_2–Norm.

Ausgangspunkt für eine noch schwächere Fehlerabschätzung ist für $s \in (0,1)$

$$f(x) - Q_h f(x) = \frac{1}{h_k} \int\limits_{x_{k-1}}^{x_k} \Big[f(x) - f(y)\Big] dy = \frac{1}{h_k} \int\limits_{x_{k-1}}^{x_k} \frac{f(x) - f(y)}{|x-y|^{\frac{1}{2}+s}} |x-y|^{\frac{1}{2}+s} dy.$$

Mit der Cauchy–Schwarz–Ungleichung folgt daraus

$$\begin{aligned}
\Big[f(x) - Q_h f(x)\Big]^2 &= \frac{1}{h_k^2} \left[\int\limits_{x_{k-1}}^{x_k} \frac{f(x) - f(y)}{|x-y|^{\frac{1}{2}+s}} |x-y|^{\frac{1}{2}+s} dy \right]^2 \\
&\leq \frac{1}{h_k^2} \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy \int\limits_{x_{k-1}}^{x_k} |x-y|^{1+2s} dy \\
&\leq \frac{1}{h_k^2} \, h_k^{1+2s} \, h_k \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy \\
&= h_k^{2s} \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy
\end{aligned}$$

und Integration bezüglich $x \in (x_{k-1}, x_k)$ liefert

$$\int\limits_{x_{k-1}}^{x_k} \Big[f(x) - Q_h f(x)\Big]^2 dx \;\leq\; h_k^{2s} \int\limits_{x_{k-1}}^{x_k} \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy dx$$

und somit

$$\int\limits_a^b \Big[f(x) - Q_h f(x)\Big]^2 dx \;\leq\; \sum_{k=1}^n h_k^{2s} \int\limits_{x_{k-1}}^{x_k} \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy dx$$

bzw.

$$\begin{aligned}
||f - Q_h f||_{L_2[a,b]} &\leq h^s \left[\sum_{k=1}^n \int\limits_{x_{k-1}}^{x_k} \int\limits_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy dx \right]^{\frac{1}{2}} \\
&\leq h^s \left[\int\limits_a^b \int\limits_a^b \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dy dx \right]^{\frac{1}{2}}.
\end{aligned}$$

Aus der Galerkin–Orthogonalität

$$\int\limits_a^b \Big[f(x) - Q_h f(x)\Big] g_h(x) dx = 0 \quad \text{für alle } g_h \in S_h$$

und der Cauchy–Schwarz–Ungleichung folgt andererseits

$$\begin{aligned}
||f - Q_h f||^2_{L_2[a,b]} &= \int_a^b \Big[f(x) - Q_h f(x)\Big]\Big[f(x) - Q_h f(x)\Big] dx \\
&= \int_a^b \Big[f(x) - Q_h f(x)\Big] f(x) dx \\
&\le \left[\int_a^b [f(x) - Q_h f(x)]^2 dx\right]^{\frac{1}{2}} \left[\int_a^b [f(x)]^2 dx\right]^{\frac{1}{2}} \\
&= ||f - Q_h f||_{L_2[a,b]} ||f||_{L_2[a,b]}
\end{aligned}$$

und somit

$$||f - Q_h f||_{L_2[a,b]} \le ||f||_{L_2[a,b]}.$$

Je nach Voraussetzung an die zu approximierende Funktion $f(x)$ gelten also die Fehlerabschätzungen

$$\begin{aligned}
||f - Q_h f||_{L_2[a,b]} &\le ||f||_{L_2[a,b]}, \\
||f - Q_h f||_{L_2[a,b]} &\le h^s \left[\int_a^b \int_a^b \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}} dx dx\right]^{\frac{1}{2}} \quad \text{für } s \in (0,1), \\
||f - Q_h f||_{L_2[a,b]} &\le h\, ||f''||_{L_2[a,b]}.
\end{aligned}$$

Auf analoge Betrachtungen für die L_2–Projektion mit stückweise linearen Ansatzfunktionen soll hier verzichtet werden.

Übungsaufgaben

1.1. Für die Vandermonde–Matrix

$$V_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix}$$

zeige man

$$\det V_n = \prod_{i<j} (x_j - x_i).$$

1.2. Gegeben seien $n+1$ paarweise verschiedene Stützstellen x_i mit $x_i \neq x_j$ für $i \neq j$. Man zeige, daß die zugehörigen Lagrange–Polynome

$$L_i(x) = \sum_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad \text{für } i = 0, \dots, n$$

eine Basis im Raum Π_n der Polynome vom Grad n bilden.

1.3. Für $x \in [-1,+1]$ und $k = 0,1,2,\dots$ zeige man die alternative Darstellung

$$T_k(x) = \cos(k \text{ arccos } x)$$

der Tschebyscheff–Polynome $T_k(x)$.

1.4. Man beweise die Orthogonalität der Tschebyscheff–Polynome,

$$\int_{-1}^{1} \frac{T_k(x)T_\ell(x)}{\sqrt{1-x^2}}dx \;=\; \begin{cases} 0 & \text{für } k \neq \ell, \\ \frac{\pi}{2} & \text{für } k = \ell \neq 0, \\ \pi & \text{für } k = \ell = 0. \end{cases}$$

1.5. Gegeben seien die vier Stützstellen $x_i = i$ für $i = 0, 1, 2, 3$. Man bestimme eine stetig differenzierbare Ansatzfunktion $\varphi(x)$, welche in den Intervallen (x_{i-1}, x_i) für $i = 1, 2, 3$ durch ein quadratisches Polynom beschrieben werden kann, und für welche

$$\varphi(x_0) \;=\; \varphi'(x_0) \;=\; \varphi(x_3) \;=\; \varphi'(x_3) \;=\; 0$$

gilt.

Kapitel 2

Numerische Integration

Für eine gegebene Funktion $f(x)$ ist das bestimmte Integral

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

durch eine geeignete Näherungsformel

$$I_n = \sum_{i=0}^{n} \omega_i f(x_i)$$

mit **Stützstellen** x_i und **Gewichten** ω_i zu berechnen. Ein numerisches Integrationsverfahren heißt von der **Ordnung** $\boldsymbol{p}$, falls p die größte ganze Zahl ist, für die das Verfahren alle Polynome kleineren Grades als p exakt integriert,

$$\int_a^b g(x)dx = \sum_{i=0}^{n} \omega_i g(x_i) \quad \text{für alle } g \in \Pi_{p-1}.$$

Eine erste Idee für die Herleitung numerischer Integrationsformeln besteht im Ersetzen der Funktion $f(x)$ durch das Interpolationspolynom $f_n \in \Pi_n$ mit

$$f_n(x_i) = f(x_i) \quad \text{für } i = 0, \ldots, n.$$

Dies führt zu

$$I_n = \int_a^b f_n(x)dx,$$

und aus der Darstellung des Interpolationsfehlers

$$f(x) - f_n(x) \; = \; \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{i=0}^{n} (x - x_i)$$

mit einer Zwischenwertstelle $\xi = \xi(x) \in (a, b)$ folgt für den Fehler der numerischen Integrationsformel

$$I - I_n \; = \; \int_a^b [f(x) - f_n(x)]dx \; = \; \frac{1}{(n+1)!} \int_a^b f^{(n+1)}\Big(\xi(x)\Big) \prod_{i=0}^{n} (x - x_i)dx.$$

Für eine **äquidistante** Unterteilung der Stützstellen,

$$x_i = a + i\frac{b-a}{n} = a + ih \quad \text{für } i = 0, \ldots, n,$$

und für die Lagrange–Darstellung des Interpolationspolynoms,

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

folgt

$$I_n = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x) dx = \sum_{i=0}^{n} \omega_i f(x_i)$$

mit den Integrationsgewichten

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

Mit den Substitutionen

$$x_i = a + ih \quad \text{für } i = 0, \ldots, n, \quad x = a + th \quad \text{für } t \in [0, n], \quad dx = h\, dt$$

ist

$$\omega_i = h \int_0^n \prod_{j\neq i} \frac{(a + th) - (a + jh)}{(a + ih) - (a + jh)} dt = h \int_0^n \prod_{j\neq i} \frac{t - j}{i - j} dt = \frac{b - a}{n} \widetilde{\omega}_i$$

mit

$$\widetilde{\omega}_i = \int_0^n \prod_{j=0, j\neq i}^{n} \frac{t - j}{i - j} dt \quad \text{für } i = 0, \ldots, n.$$

Die resultierenden numerischen Integrationsformeln sind die **Newton–Cotes–Formeln**

$$I_n = \frac{b - a}{n} \sum_{i=0}^{n} \widetilde{\omega}_i f(x_i).$$

Ist die Integrationsformel exakt für konstante Funktionen, dann folgt für $f(x) = 1$

$$I = \int_a^b dx \, = \, b - a \, = \, I_n \, = \, \frac{b - a}{n} \sum_{i=0}^{n} \widetilde{\omega}_i$$

und somit

$$\frac{1}{n} \sum_{i=0}^{n} \widetilde{\omega}_i = 1.$$

Für eine stabile numerische Auswertung ist weiterhin die Positivität $\widetilde{\omega}_i > 0$ der Integrationsgewichte zu fordern.

Beispiel 2.1 *Für $n = 1$ lauten die Stützstellen*

$$x_0 = a, \quad x_1 = b$$

und für die Integrationsgewichte ergibt sich

$$\begin{aligned} \widetilde{\omega}_0 &= \int_0^1 \frac{t - 1}{0 - 1} dt = \int_0^1 (1 - t)\, dt = \frac{1}{2}, \\ \widetilde{\omega}_1 &= \int_0^1 \frac{t - 0}{1 - 0} dt = \int_0^1 t\, dt = \frac{1}{2}. \end{aligned}$$

Damit ist

$$I_1 \;=\; (b-a)\Big[\frac{1}{2}f(a)+\frac{1}{2}f(b)\Big] \;=\; \frac{b-a}{2}\Big[f(a)+f(b)\Big]$$

die **Trapezregel**. *Der Fehler ist*

$$I - I_1 \;=\; \int\limits_a^b \frac{1}{2!}f''\Big(\xi(x)\Big)(x-a)(x-b)dx \;=\; -\frac{1}{2}\int\limits_a^b f''\Big(\xi(x)\Big)\underbrace{(x-a)(b-x)}_{>0 \text{ für } x\in(a,b)}dx.$$

Die Substitution

$$s(x) \;=\; \int (x-a)(b-x)dx \;=\; -\frac{1}{3}x^3+\frac{1}{2}(a+b)x^2-abx$$

ergibt eine für $x\in(a,b)$ streng monoton steigende Funktion, für die die Umkehrfunktion $x=x(s)$ existiert. Mit der Transformation der Integrationsgrenzen,

$$s_a = \frac{1}{6}a^3-\frac{1}{2}a^2b \quad \text{für } x=a, \qquad s_b=\frac{1}{6}b^3-\frac{1}{2}ab^2 \quad \text{für } x=b,$$

ergibt sich aus dem Mittelwertsatz der Integralrechnung

$$\begin{aligned} I - I_1 \;&=\; -\frac{1}{2}\int\limits_{s_a}^{s_b} f''\Big(\xi(x(s))\Big)ds \;=\; -\frac{1}{2}[s_b-s_a]f''\Big(\xi(x(\bar{s}))\Big) \\ &=\; -\frac{1}{2}f''(\eta)\Big[\frac{1}{6}b^3-\frac{1}{2}ab^2+\frac{1}{2}a^2b-\frac{1}{6}a^3\Big] \;=\; -\frac{1}{12}f''(\eta)(b-a)^3. \end{aligned}$$

Damit ist die Trapezregel ein Verfahren 2. Ordnung, d.h. lineare Funktionen werden exakt integriert.

Beispiel 2.2 *Für $n=2$ sind die Stützstellen durch*

$$x_0=a, \quad x_1=\frac{1}{2}(a+b), \quad x_2=b$$

gegeben und für die Integrationsgewichte ergibt sich

$$\widetilde{w}_0=\frac{1}{3}, \quad \widetilde{w}_1=\frac{4}{3}, \quad \widetilde{w}_2=\frac{1}{3}.$$

Die resultierende Integrationsformel

$$I_2 \;=\; \frac{1}{6}(b-a)\Big[f(a)+4f(\frac{a+b}{2})+f(b)\Big]$$

ist die **Simpson–Regel** *mit dem Fehler*

$$I - I_2 = -\frac{(b-a)^5}{2880}f^{(4)}(\eta),$$

d.h. kubische Funktionen werden exakt integriert. Der Beweis erfolgt analog zur Mittelpunktformel (vgl. Übungsaufgabe 2.1)

$$\int\limits_a^b f(x)dx = (b-a)f\Big(\frac{a+b}{2}\Big)+\frac{1}{24}f''(\eta)(b-a)^3$$

mittels Taylorentwicklung um $\frac{a+b}{2}$ und Anwendung des Mittelwertsatzes.

Als notwendiges Kriterium für die Konvergenz der bisherigen numerischen Integrationsformeln ist

$$|b-a| < 1$$

vorauszusetzen. Der allgemeine Fall kann durch zusammengesetzte Integrationsformeln

$$I = \int_a^b f(x)dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)dx$$

mit Stützstellen

$$x_k = a + k\frac{b-a}{n} \quad \text{für } k = 0, \ldots, n$$

und numerischer Integration der verbleibenden Integrale behandelt werden. Mit der Simpson–Regel folgt zum Beispiel

$$\begin{aligned} I_n &= \sum_{k=1}^{n} \frac{1}{6}(x_k - x_{k-1})\Big[f(x_{k-1}) + 4f(\frac{x_{k-1}+x_k}{2} + f(x_k)\Big] \\ &= \frac{b-a}{6n}\sum_{k=0}^{n}\Big[f(x_{k-1}) + 4f(\frac{x_{k-1}+x_k}{2}) + f(x_k)\Big]. \end{aligned}$$

Für eine Fehleranalysis von zusammengesetzten Integrationsformeln siehe Übungsaufgabe 2.2.

Bei den bisherigen Betrachtungen wurden die Integrationspunkte x_k als gegeben vorausgesetzt. Allgemein enthält die Integrationsformel

$$I_n = \sum_{i=0}^{n} \omega_i f(x_i)$$

$2(n+1)$ frei wählbare Parameter (ω_i, x_i). Diese können aus der Forderung der exakten Integration von Polynomen $f(x) = x^\alpha$ für $\alpha = 0, \ldots, 2n+1$ gewonnen werden,

$$\int_a^b x^\alpha \, dx = \sum_{i=0}^{n} \omega_i x_i^\alpha.$$

Beispiel 2.3 *Für $n = 2$ und das Integrationsintervall $[a, b] = [0, 1]$ ergibt sich das nichtlineare Gleichungssystem*

$$\int_a^b x^\alpha \, dx = \frac{1}{\alpha+1} = \sum_{i=0}^{2} \omega_i x_i^\alpha \quad \text{für } \alpha = 0, \ldots, 5.$$

Aus Symmetriegründen ist

$$x_0 = t, \quad x_1 = \frac{1}{2}, \quad x_2 = 1-t$$

und

$$\omega_0 = \omega_2 = \omega$$

zu wählen. Aus der Gleichung für $\alpha = 0$,

$$\omega_1 + \omega_2 + \omega_3 = 1,$$

folgt dann

$$\omega_1 = 1 - 2\omega.$$

Man prüft leicht nach, daß dann auch die Gleichung für $\alpha = 1$ erfüllt ist,

$$\frac{1}{2} = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = \omega t + (1-2\omega)\frac{1}{2} + \omega(1-t) = \frac{1}{2}.$$

Für $\alpha = 2$ bzw. für $\alpha = 3$ ergibt sich

$$\frac{1}{12} = \omega\Big[2t^2 - 2t + \frac{1}{2}\Big],$$

während für $\alpha = 4$ bzw. für $\alpha = 5$

$$\frac{11}{80} = \omega\Big[2t^4 - 4t^3 + 6t^2 - 4t + \frac{7}{8}\Big]$$

folgt. Gleichsetzen liefert

$$40t^4 - 80t^3 + 54t^2 - 14t + 1 = 0$$

mit den Lösungen

$$t_{1/2} = \frac{1}{2} \pm \frac{\sqrt{15}}{10}, \quad t_{3/4} = \frac{1}{2}.$$

Für

$$t = \frac{1}{2} - \frac{\sqrt{15}}{10} \quad \textit{folgt} \quad \omega = \frac{5}{18}$$

und somit lauten die Stützstellen

$$x_0 = \frac{1}{2} - \frac{\sqrt{15}}{10}, \quad x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2} + \frac{\sqrt{15}}{10}$$

und die zugehörigen Integrationsgewichte

$$w_0 = \frac{5}{18}, \quad w_1 = \frac{8}{18}, \quad w_2 = \frac{5}{18}.$$

Es stellt sich die Frage, wie dieser Zugang und insbesondere die Lösung des nichtlinearen Gleichungssystems verallgemeinert werden kann.
Zur Berechnung des Integrals

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

wird eine numerische Integrationsfomel

$$I_n = \sum_{i=0}^{n} \omega_i f(x_i)$$

betrachtet, welche Polynome $f_m(x)$ von möglichst maximalen Polynomgrad $m > n$ exakt integriert.
Allgemein bezeichnet

$$f_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x)$$

das Interpolationspolynom mit Lagrange–Funktionen $L_i(x)$. Für $f_m \in \Pi_m$ gilt dann

$$f_m(x) = \underbrace{\sum_{i=0}^{n} f_m(x_i) L_i(x)}_{f_n \in \Pi_n} + \underbrace{\prod_{i=0}^{n} (x - x_i)}_{p_{n+1} \in \Pi_{n+1}} \underbrace{g_{m-(n+1)}(x)}_{\in \Pi_{m-(n+1)}}$$

sowie

$$f_m(x_i) \; = \; f_n(x_i) \quad \text{für } i = 0, \ldots, n.$$

Einsetzen in die Integrationsformel ergibt

$$\int_a^b f_m(x)dx = \int_a^b f_n(x)dx + \int_a^b g_{m-(n+1)}(x) p_{n+1}(x)dx = \sum_{i=0}^{n} f_m(x_i) \int_a^b L_i(x)dx,$$

falls

$$\int_a^b g_{m-(n+1)}(x) p_{n+1}(x) dx \;=\; 0$$

erfüllt ist. Mit

$$g_{m-(n+1)}(x) \;=\; \sum_{j=0}^{m-(n+1)} a_j p_j(x)$$

mit noch zu bestimmenden linear unabhängigen Polynomen $p_j(x)$ folgt dies aus der Orthogonalität

$$\int_a^b p_j(x) p_{n+1}(x) dx \;=\; 0 \quad \text{für } j = 0, \ldots, n.$$

Wegen $j = 0, \ldots, m - (n+1)$ ist dies gleichbedeutend mit

$$m \leq 2n + 1,$$

d.h. Polynome maximalen Grades $2n + 1$ werden exakt integriert.
Benötigt wird also eine Folge von zueinander orthogonalen Polynomen $\{p_k\}_{k=0}^{n+1}$ mit

$$\int_a^b p_k(x) p_\ell(x) dx \;=\; 0 \quad \text{für } k \neq \ell.$$

Die Stützstellen der numerischen Integrationsformel ergeben sich dann aus den Nullstellen von $p_{n+1}(x)$.
Ausgehend von der Basis $\{x^k\}_{k=0}^{n+1}$ der Monome x^k kann durch Anwendung des **Orthogonalisierungsverfahrens nach Gram–Schmidt** ein System orthogonaler Polynome konstruiert werden.

Beispiel 2.4 *Für $n = 2$ und $[a, b] = [0, 1]$ ist zunächst*

$$p_0(x) = 1.$$

Mit dem Ansatz

$$p_1(x) = x - \alpha_0 p_0(x) = x - \alpha_0$$

und der Bedingung

$$0 = \int_0^1 p_1(x) p_0(x) dx = \int_0^1 [x - \alpha_0] dx = \frac{1}{2} - \alpha_0$$

ergibt sich

$$p_1(x) = x - \frac{1}{2}.$$

Für

$$p_2(x) = x^2 - \alpha_1 p_1(x) - \alpha_0 p_0(x)$$

ist analog

$$\begin{aligned} 0 &= \int_0^1 p_2(x) p_0(x) dx = \int_0^1 [x^2 - \alpha_0] dx = \frac{1}{3} - \alpha_0, \\ 0 &= \int_0^1 p_2(x) p_1(x) dx = \int_0^1 \Big[x^2 - \alpha_1 (x - \frac{1}{2})\Big](x - \frac{1}{2}) dx \;=\; \frac{1}{12} - \alpha_1 \frac{1}{12} \end{aligned}$$

und somit

$$p_2(x) = x^2 - \frac{1}{3} - (x - \frac{1}{2}) = x^2 - x + \frac{1}{6}.$$

Entsprechend folgt

$$p_3(x) = x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}.$$

Zu bestimmen sind die Nullstellen von $p_3(x)$ durch Lösen der kubischen Gleichung

$$20x^3 - 30x^2 + 12x - 1 = 0$$

mit den Lösungen

$$x_0 = \frac{1}{2} - \frac{\sqrt{15}}{10}, \quad x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2} + \frac{\sqrt{15}}{10}.$$

Für die Integrationsgewichte ist zunächst

$$\omega_0 = \int_0^1 \frac{x - x_1}{x_0 - x_1} \frac{x - x_2}{x_0 - x_2} dx = \frac{5}{18}$$

und die Werte für $\omega_1 = \frac{4}{9}$ und $\omega_2 = \frac{5}{18}$ ergeben sich analog.

Allgemein sind die bezüglich des Intervalles $[-1, +1]$ orthogonalen Polynome durch die **Legendre–Polynome**

$$P_k(x) = \frac{1}{2^k k!} \frac{d^k}{dx^k}(x^2 - 1)^k \quad \text{für } k = 0, 1, 2, \ldots$$

gegeben:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x).$$

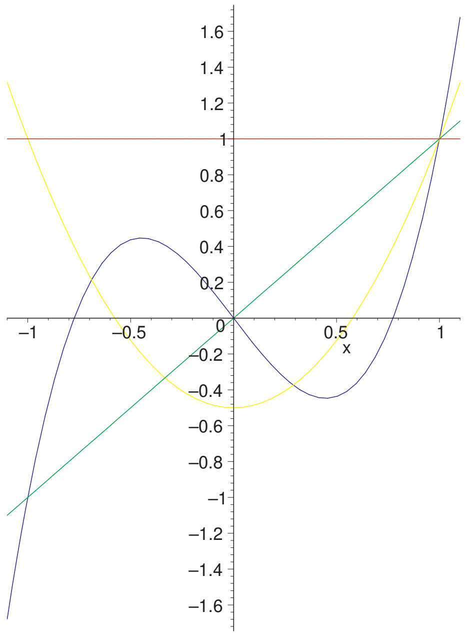

Abbildung 2.1: Legendre–Polynome $P_k(x) = x^k$ für $k = 0, 1, 2, 3$.

Für die Legendre–Polynome gilt die Orthogonalität

$$\int_{-1}^{1} P_k(x)P_\ell(x)dx = \frac{2}{2k+1} \cdot \begin{cases} 1 & \text{für } k = \ell, \\ 0 & \text{für } k \neq \ell. \end{cases}$$

Andererseits genügen die Legendre–Polynome der Rekursion

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ (k+1)P_{k+1}(x) &= (2k+1)xP_k(x) - kP_{k-1}(x) \quad \text{für } k = 1, 2, \ldots. \end{aligned}$$

Mit den Nullstellen x_k des Legendre–Polynoms $P_{n+1}(x)$ als Stützstellen ergeben sich somit die **Gauss–Legendre–Integrationsformeln**

$$\int_{-1}^{1} f(x)dx = \sum_{i=0}^{n} \omega_i f(x_i) + \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \Big[\prod_{i=0}^{n}(x - x_i)\Big]^2 dx$$

mit den Integrationsgewichten

$$\omega_i = \int_{-1}^{1} L_i(x)dx = \frac{2(1 - x_i^2)}{(n+1)^2[P_n(x_i)]^2} \quad \text{für } i = 0, \ldots, n.$$

Abschließend sollen geeignete numerische Integrationsformeln für gewichtete Integrale der Form

$$I = \int_{-1}^{1} \frac{f(x)}{\sqrt{1 - x^2}} dx$$

betrachtet werden. Polynome $f_m \in \Pi_m$ mit $m > n$ erlauben die Darstellung

$$f_m(x) = \sum_{i=0}^{n} f_m(x_i)L_i(x) + \prod_{i=0}^{n}(x - x_i)g_{m-(n+1)}(x).$$

Einsetzen in die Integrationsformel ergibt

$$\begin{aligned} \int_a^b \frac{f_m(x)}{\sqrt{1-x^2}}dx &= \int_a^b \frac{f_n(x)}{\sqrt{1-x^2}}dx + \int_a^b \frac{g_{m-(n+1)}(x)p_{n+1}(x)}{\sqrt{1-x^2}}dx \\ &= \sum_{i=0}^{n} f_m(x_i) \int_a^b \frac{L_i(x)}{\sqrt{1-x^2}}dx, \end{aligned}$$

falls

$$\int_a^b \frac{g_{m-(n+1)}(x)p_{n+1}(x)}{\sqrt{1-x^2}}dx = 0$$

erfüllt ist. Diese Orthogonalität ist aber gerade für die Tschebyscheff–Polynome $T_k(x)$ erfüllt (siehe Übungsaufgabe 1.4.),

$$\int_{-1}^{1} \frac{T_k(x)T_\ell(x)}{\sqrt{1-x^2}}dx = 0 \quad \text{für } k \neq \ell.$$

Werden als Stützstellen $x_i = \bar{x}_i^{(n+1)}$ der numerischen Integrationsformel die Nullstellen des Tschebyscheff–Polynoms $T_{n+1}(x)$ gewählt, so ist die Integrationsformel

$$I_n = \sum_{i=0}^{n} \omega_i f_m\Big(\bar{x}_i^{(n+1)}\Big), \quad \omega_i = \int_{-1}^{1} \frac{L_i(x)}{\sqrt{1-x^2}} dx$$

exakt für alle Polynome f_m vom Polynomgrad $m \leq 2n+1$. Diese Eigenschaft ist wesentlich für die Berechnung der Koeffizienten a_k bei der Interpolation mit Tschebyscheff–Polynomen (vergleiche Abschnitt 1.1.4).

Übungsaufgaben

2.1. Man beweise die Fehlerabschätzung für die Mittelpunktformel,

$$\int_a^b f(x)dx = (b-a)f(\frac{a+b}{2}) + \frac{1}{24} f''(\eta)(b-a)^3, \quad \eta \in (a,b).$$

2.2. Für $n \in I\!N$ werde die zusammengesetzte Integrationsformel

$$\int_0^1 f(x)dx = \frac{1}{n} \sum_{k=1}^{n} f(\frac{2k-1}{2n}) + R_n$$

betrachtet. Man gebe eine Abschätzung für das Restglied R_n an.

2.3. Man bestimme die Stützstellen x_k und die Integrationsgewichte ω_k der Gauss–Legendre–Integrationsformel

$$\int_0^1 f(x)dx = \omega_0 f(x_0) + \omega_2 f(x_1) + R_1 .$$

2.4. Man zeige

$$\sum_{i=0}^{n} T_k\Big(\bar{x}_i^{(n+1)}\Big) T_\ell\Big(\bar{x}_i^{(n+1)}\Big) = \begin{cases} 0 & \text{für } k \neq \ell, \\ \frac{1}{2}(n+1) & \text{für } k = \ell \neq 0, \\ n+1 & \text{für } k = \ell = 0. \end{cases}$$

2.5. Für das Referenz–Dreieck $\tau = \{(x,y) \in I\!R^2 : 0 < x < 1,\ 0 < y < 1-x\}$ ist eine numerische Integrationsformel

$$\int_\tau f(x,y)dxdy = \frac{1}{2} \sum_{k=1}^{3} \omega_k f(x_k, y_k) + R_3$$

herzuleiten, die Polynome

$$p_2(x,y) = \sum_{\alpha_1+\alpha_2 \leq 2} a_\alpha x^{\alpha_1} y^{\alpha_2}$$

exakt integriert.

Kapitel 3

Lineare Gleichungssysteme

Zu bestimmen ist der Lösungsvektor $\underline{x} \in \mathbb{R}^n$ des linearen Gleichungssystems

$$A\underline{x} = \underline{f}$$

mit einer regulären, d.h. invertierbaren, Matrix $A \in \mathbb{R}^{n \times n}$. Die Dimension n widerspiegelt in der Regel einen bei der Approximation auftretenden Diskretisierungsparameter, so daß der Fall $n \to \infty$ von Interesse ist.

Die Lösungsverfahren können in zwei Klassen eingeteilt werden. Die **direkten** Verfahren wie zum Beispiel das Gauß'sche Eliminationsverfahren oder die LU–Zerlegung liefern bis auf Rundungsfehler die exakte Lösung des linearen Gleichungssystems. Der Aufwand hierfür beträgt im allgemeinen jedoch $\mathcal{O}(n^3)$ wesentliche Operationen: eine Verdoppelung der Freiheitsgrade verachtfacht die notwendige Rechenzeit zur Lösung.

In vielen Anwendungen resultiert das lineare Gleichungssystem aus einer Approximationsmethode zur näherungsweisen Lösung eines mathematischen Modells. Der Lösungsvektor beschreibt also eine bereits fehlerbehaftete Näherungslösung. Damit ist es in der Regel ausreichend, auch die Lösung des linearen Gleichungssystems mittels eines **Iterationsverfahren** mit einer hinreichenden Genauigkeit zu berechnen. Zu den **klassischen** Iterationsverfahren gehören das **Einschrittverfahren** (**Jacobi**) und das **Gesamtschrittverfahren** (**Gauss–Seidel**) sowie die daraus resultierenden Relaxationsverfahren. Für den Fall einer symmetrischen und positiv definiten Systemmatrix A soll hier das **Verfahren konjugierter Gradienten** behandelt werden, während für den allgemeinen Fall einer regulären Matrix A das **Verfahren des verallgemeinerten minimalen Residuums** (**GMRES**) betrachtet wird.

Für weiterführende Betrachtungen zu effizienten Lösungsverfahren für lineare Gleichungssysteme sei hier auf entsprechende weiterführende Vorlesungen bzw. Literatur verwiesen, siehe zum Beispiel [4, 9, 14].

3.1 Das Verfahren konjugierter Gradienten

Sei zunächst $A = A^\top > 0$ symmetrisch und positiv definit. Ein System von linear unabhängigen Vektoren $\{\underline{p}^k\}_{k=0}^{n-1}$ heißt ***A*–orthogonal** oder **konjugiert**, falls

$$(A\underline{p}^k, \underline{p}^\ell) = 0 \quad \text{für } k, \ell = 0, \ldots, n-1 \text{ und } k \neq \ell$$

sowie

$$(A\underline{p}^k, \underline{p}^k) > 0 \quad \text{für } k = 0, \ldots, n-1$$

erfüllt ist. Die letzte Forderung folgt dabei unmittelbar aus der positiven Definitheit von A.

Für ein zunächst beliebig gegebenes System linear unabhängiger Vektoren $\{\underline{w}^k\}_{k=0}^{n-1}$ kann mittels des **Gram–Schmidtschen–Orthogonalisierungsverfahrens** ein System orthogonaler Vektoren $\{\underline{p}^k\}_{k=0}^{n-1}$ konstruiert werden:

Setze
$\underline{p}^0 := \underline{w}^0$
Für $k = 0, \ldots, n-2$ berechne
$\underline{p}^{k+1} := \underline{w}^{k+1} - \sum_{\ell=0}^{k} \beta_{k,\ell}\underline{p}^\ell, \quad \beta_{k,\ell} = \frac{(A\underline{w}^{k+1}, \underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)}$

Algorithmus 3.1: Konstruktion A–orthogonaler Vektoren.

Einsetzen des Lösungsansatzes

$$\underline{x} = \underline{x}^0 - \sum_{\ell=0}^{n-1} \alpha_\ell \underline{p}^\ell$$

in das lineare Gleichungssystem $A\underline{x} = \underline{f}$ ergibt

$$A\underline{x} = A\underline{x}^0 - \sum_{\ell=0}^{n-1} \alpha_\ell A\underline{p}^\ell = \underline{f}.$$

Das Bilden des Euklidischen Skalarproduktes mit $\underline{p}^k$ liefert wegen der A–Orthogonalität des Vektorsystems $\{\underline{p}^k\}_{k=0}^{n-1}$

$$(A\underline{x}^0, \underline{p}^k) - \sum_{\ell=0}^{n-1} \alpha_\ell \underbrace{(A\underline{p}^\ell, \underline{p}^k)}_{= 0 \text{ für } \ell \neq k} = (\underline{f}, \underline{p}^k)$$

und somit

$$\alpha_k = \frac{(A\underline{x}^0 - \underline{f}, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)} \quad \text{für } k = 0, \ldots, n.$$

Für die **Näherungslösung**

$$\underline{x}^{k+1} = \underline{x}^0 - \sum_{\ell=0}^{k} \alpha_\ell \underline{p}^\ell = \underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell \underline{p}^\ell - \alpha_k \underline{p}^k = \underline{x}^k - \alpha_k \underline{p}^k$$

ist das zugehörige **Residuum** gegeben durch

$$\underline{r}^{k+1} = A\underline{x}^{k+1} - \underline{f} = A\underline{x}^0 - \sum_{\ell=0}^{k} \alpha_\ell A\underline{p}^\ell - \underline{f} = A\underline{x}^k - \underline{f} - \alpha_k A\underline{p}^k = \underline{r}^k - \alpha_k A\underline{p}^k$$

und wegen

$$(A\underline{p}^\ell, \underline{p}^k) = 0 \quad \text{für } k \neq \ell$$

folgt

$$(A\underline{x}^0 - \underline{f}, \underline{p}^k) \; = \; (A\underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell A\underline{p}^\ell - \underline{f}, \underline{p}^k) \; = \; (\underline{r}^k, \underline{p}^k)$$

und somit

$$\alpha_k = \frac{(\underline{r}^k, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)}.$$

Damit ergibt sich für $k = 0, \ldots, n-2$ die Iterationsvorschrift

$$\underline{x}^{k+1} = \underline{x}^k - \alpha_k \underline{p}^k, \quad \underline{r}^{k+1} = \underline{r}^k - \alpha_k A\underline{p}^k, \quad \alpha_k = \frac{(\underline{r}^k, \underline{p}^k)}{(A\underline{p}^k, \underline{p}^k)}.$$

Nach Konstruktion gilt

$$(\underline{r}^{k+1}, \underline{p}^k) = (\underline{r}^k - \alpha_k A\underline{p}^k, \underline{p}^k) = 0$$

für alle $k = 0, \ldots, n-2$ und somit für $k = 1$ die Induktionsvoraussetzung

$$(\underline{r}^k, \underline{p}^\ell) = (\underline{r}^1, \underline{p}^0) = 0 \quad \text{für} \quad \ell = 0, \ldots, k-1.$$

Durch vollständige Induktion folgt, daß dann auch

$$(\underline{r}^{k+1}, \underline{p}^\ell) = 0 \quad \text{für alle } \ell = 0, \ldots, k$$

für alle $k = 2, \ldots, n-1$ gilt. Zunächst ist

$$(\underline{r}^{k+1}, \underline{p}^k) = 0$$

und für $\ell < k$ folgt aus der Induktionsvoraussetzung und der A–Orthogonalität

$$(\underline{r}^{k+1}, \underline{p}^\ell) = (\underline{r}^k - \alpha_k A\underline{p}^k, \underline{p}^\ell) = (\underline{r}^k, \underline{p}^\ell) - \alpha_k (A\underline{p}^k, \underline{p}^\ell) = 0.$$

Insgesamt gilt also

$$(\underline{r}^{k+1}, \underline{p}^\ell) = 0 \quad \text{für } \ell = 0, \ldots, k \text{ und } k = 0, \ldots, n-2.$$

Aus der Konstruktion der Suchrichtungen,

$$\underline{p}^\ell = \underline{w}^\ell - \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} \underline{p}^j \quad \text{bzw.} \quad \underline{w}^\ell = \underline{p}^\ell + \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} \underline{p}^j,$$

folgt weiterhin

$$(\underline{r}^{k+1}, \underline{w}^\ell) = (\underline{r}^{k+1}, \underline{p}^\ell) + \sum_{j=0}^{\ell-1} \beta_{\ell-1,j} (\underline{r}^{k+1}, \underline{p}^j) \;=\; 0 \quad \text{für } \ell = 0, \ldots, k.$$

Damit ist das Residuum $\underline{r}^{k+1}$ orthogonal zu allen Basisvektoren $\underline{w}^\ell$, $\ell = 0, \ldots, k$. Das Vektorsystem

$$\{\underline{w}^0, \underline{w}^1, \ldots, \underline{w}^k, \underline{r}^{k+1}\}$$

ist also **linear unabhängig**, so daß die neue Suchrichtung als

$$\underline{w}^{k+1} \;=\; \underline{r}^{k+1} \quad \text{bzw.} \quad \underline{w}^\ell \;=\; \underline{r}^\ell \quad \text{für } \ell = 0, \ldots, n-1$$

gewählt werden kann. Damit folgt

$$(\underline{r}^{k+1}, \underline{p}^\ell) = (\underline{r}^{k+1}, \underline{r}^\ell) = 0 \quad \text{für } \ell = 0, \ldots, k \text{ und } k = 0, \ldots, n-2.$$

Für das Orthogonalisierungsverfahren nach Gram–Schmidt ergibt sich dann

$$\underline{p}^0 = \underline{w}^0 = \underline{r}^0, \quad \underline{p}^{k+1} = \underline{r}^{k+1} - \sum_{\ell=0}^{k} \beta_{k\ell} \underline{p}^\ell \quad \text{für } k = 0, \ldots, n-2$$

mit

$$\beta_{k\ell} = \frac{(A\underline{w}^{k+1}, \underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)} = \frac{(\underline{r}^{k+1}, A\underline{p}^\ell)}{(A\underline{p}^\ell, \underline{p}^\ell)}.$$

Ohne Einschränkung der Allgemeinheit sei $\alpha_\ell \neq 0$. Andernfalls würde aus der Rekursion

$$\underline{r}^{\ell+1} = \underline{r}^\ell - \alpha_\ell A\underline{p}^\ell$$

und der Orthogonalität

$$(\underline{r}^{\ell+1}, \underline{r}^{\ell}) = 0$$

die Gleichheit

$$0 = (\underline{r}^{\ell+1}, \underline{r}^{\ell}) = (\underline{r}^{\ell}, \underline{r}^{\ell})$$

folgen und somit die Forderung $\underline{r}^{\ell} = 0$ nach sich ziehen, d.h. $\underline{x}^{\ell} = \underline{x}$ ist die exakte Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$.
Aus

$$\underline{r}^{\ell+1} = \underline{r}^{\ell} - \alpha_\ell A\underline{p}^{\ell}$$

folgt dann

$$A\underline{p}^{\ell} = \frac{1}{\alpha_\ell}[\underline{r}^{\ell} - \underline{r}^{\ell+1}]$$

und somit ergibt sich für den Zähler von $\beta_{k\ell}$

$$(\underline{r}^{k+1}, A\underline{p}^{\ell}) = \frac{1}{\alpha_\ell}(\underline{r}^{k+1}, \underline{r}^{\ell} - \underline{r}^{\ell+1}) = 0 \quad \text{für} \quad \ell = 0, \ldots, k-1$$

und daher

$$\beta_{k\ell} = 0 \quad \text{für} \quad \ell = 0, \ldots, k-1.$$

Weiterhin ist

$$(\underline{r}^{k+1}, A\underline{p}^{k}) = \frac{1}{\alpha_\ell}(\underline{r}^{k+1}, \underline{r}^{k+1}) \quad \text{für} \quad \ell = k.$$

Damit ergibt sich

$$\underline{p}^{k+1} = \underline{r}^{k+1} - \beta_k \underline{p}^{k}$$

mit

$$\beta_k = \frac{(\underline{r}^{k+1}, A\underline{p}^{k})}{(A\underline{p}^{k}, \underline{p}^{k})} = -\frac{1}{\alpha_k}\,\frac{(\underline{r}^{k+1}, \underline{r}^{k+1})}{(A\underline{p}^{k}, \underline{p}^{k})}.$$

Aus

$$\underline{r}^{k+1} = \underline{r}^{k} - \alpha_k A\underline{p}^{k}$$

folgt

$$\alpha_k A\underline{p}^{k} = \underline{r}^{k} - \underline{r}^{k+1}$$

und somit

$$\alpha_k(A\underline{p}^{k}, \underline{p}^{k}) = (\underline{r}^{k} - \underline{r}^{k+1}, \underline{p}^{k}) = (\underline{r}^{k}, \underline{p}^{k}) = (\underline{r}^{k}, \underline{r}^{k} - \beta_{k-1}\underline{p}^{k-1}) = (\underline{r}^{k}, \underline{r}^{k}) = \rho_k.$$

Damit ist schließlich

$$\beta_k = -\frac{\rho_{k+1}}{\rho_k}$$

bzw.

$$\alpha_k = \frac{(\underline{r}^{k}, \underline{p}^{k})}{(A\underline{p}^{k}, \underline{p}^{k})} = \frac{\rho_k}{(A\underline{p}^{k}, \underline{p}^{k})}.$$

Die resultierende Methode ist das **Verfahren konjugierter Gradienten (CG)**, welches auf Hestenes und Stiefel [6] zurückgeht.

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in I\!R^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$. Setze $\underline{p}^0 := \underline{r}^0$ und berechne $\varrho_0 = (\underline{r}^0, \underline{r}^0)$. Stoppe, falls $\varrho_0 < \varepsilon^2$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist.

Berechne für $k = 0, 1, \ldots, n-2$:

$$\underline{s}^k = A\underline{p}^k, \ \sigma_k = (\underline{s}^k, \underline{p}^k), \ \alpha_k = \frac{\varrho_k}{\sigma_k}$$

$$\underline{x}^{k+1} := \underline{x}^k - \alpha_k \underline{p}^k$$

$$\underline{r}^{k+1} := \underline{r}^k - \alpha_k \underline{s}^k$$

$$\varrho_{k+1} := (\underline{r}^{k+1}, \underline{r}^{k+1})$$

Stoppe, falls $\varrho_{k+1} < \varepsilon^2 \varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Berechne andernfalls die neue Suchrichtung

$$\underline{p}^{k+1} := \underline{r}^{k+1} + \beta_k \underline{p}^k, \ \beta_k := \frac{\varrho_{k+1}}{\varrho_k}$$

Algorithmus 3.2: Iterationsvorschrift des konjugierten Gradientenverfahrens.

Aus den Induktionsvoraussetzungen

$$\underline{r}^0 \ \in \ \mathrm{span}\{\underline{r}^0\}, \quad \underline{p}^0 \ = \ \underline{r}^0 \in \mathrm{span}\{\underline{r}^0\}$$

folgt wegen

$$\underline{r}^{\ell+1} \ = \ \underline{r}^\ell - \alpha_\ell A \underline{p}^\ell, \quad \underline{p}^{\ell+1} \ = \ \underline{r}^{\ell+1} + \beta_\ell \underline{p}^\ell$$

durch vollständige Induktion nach $\ell = 0, \ldots, k-1$

$$\underline{p}^k \in \mathrm{span}\{\underline{r}^0, A\underline{r}^0, \ldots, A^k \underline{r}^0\} \ =: \ S_k(A, \underline{r}^0).$$

Hierbei bezeichnet $S_k(A, \underline{r}^0)$ den k–ten **Krylov–Raum** der Matrix A zum Anfangsresiduum $\underline{r}^0$. Nach Konstruktion ist durch

$$S_k(A, \underline{r}^0) \ = \ \mathrm{span}\{\underline{p}^0, \underline{p}^1, \ldots, \underline{p}^k\}$$

eine A–orthogonale Basis von $S_k(A, \underline{r}^0)$ gegeben.

Satz 3.1 *Für eine symmetrische und positiv definite Matrix $A = A^\top > 0$ konvergiert das konjugierte Gradientenverfahren mit der Konvergenzabschätzung*

$$||\underline{x}^k - \underline{x}||_A \ \leq \ \frac{2q^k}{1 + q^{2k}} \, ||\underline{e}^0||_A$$

mit

$$q \ = \ \frac{\sqrt{\kappa_2(A)} + 1}{\sqrt{\kappa_2(A)} - 1}, \quad \kappa_2(A) \ = \ \|A\|_2 \|A^{-1}\|_2 \ = \ \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Beweis: Für die durch das konjugierte Gradientenverfahren konstruierte Näherungslösung

$$\underline{x}^k \ = \ \underline{x}^0 - \sum_{\ell=0}^{k-1} \alpha_\ell \underline{p}^\ell$$

folgt aus der A–Orthogonalität der Suchrichtungen $\underline{p}^\ell$ für die durch die symmetrische und positiv definite Matrix A induzierte Norm des Fehlers

$$\|\underline{x}^k - \underline{x}\|_A^2 \ = \ \|\sum_{\ell=k}^{n-1} \alpha_\ell \underline{p}^\ell\|_A^2 \ = \ \sum_{\ell=k}^{n-1} \sum_{j=k}^{n-1} \alpha_\ell \alpha_j (A\underline{p}^\ell, \underline{p}^j) \ = \ \sum_{\ell=k}^{n-1} \alpha_\ell^2 \|\underline{p}^\ell\|_A^2.$$

Für eine beliebige Linearkombination

$$\underline{w} = \sum_{\ell=0}^{k-1} w_\ell \underline{p}^\ell$$

mit beliebigen Koeffizienten $w_0, \ldots, w_{k-1}$ folgt analog

$$\|\underline{x}^0 - \underline{w} - \underline{x}\|_A^2 = \sum_{\ell=0}^{k-1} [w_\ell - \alpha_\ell]^2 \, \|\underline{p}^\ell\|_A^2 + \sum_{\ell=k}^{n-1} \alpha_\ell^2 \, \|\underline{p}^\ell\|_A^2,$$

und somit

$$\|\underline{x}^k - \underline{x}\|_A \le \|\underline{x}^0 - \underline{w} - \underline{x}\|_A \quad \text{für alle } \underline{w} \in S_{k-1}(A, \underline{r}^0).$$

Die Näherungslösung $\underline{x}^k$ ist also Lösung des Minimierungsproblems

$$\|\underline{x}^k - \underline{x}\|_A = \min_{\underline{w} \in S_{k-1}(A, \underline{r}^0)} \|\underline{x}^0 - \underline{w} - \underline{x}\|_A.$$

Mit

$$\underline{r}^0 = A\underline{x}^0 - \underline{f} = A(\underline{x}^0 - \underline{x}) = A\underline{e}^0, \quad \underline{e}^0 = \underline{x}^0 - \underline{x}$$

ergibt sich

$$\underline{x}^0 - \underline{w} - \underline{x} = \underline{e}^0 - \sum_{\ell=0}^{k-1} w_\ell A^\ell \underline{r}^0 = \underline{e}^0 - \sum_{\ell=0}^{k-1} w_\ell A^{\ell+1} \underline{e}^0 = \sum_{\ell=0}^{k} \underline{\widetilde{w}}_\ell A^\ell \underline{e}^0$$

mit $\widetilde{w}_0 = 1$ und $\widetilde{w}_\ell = w_{\ell-1}$ für $\ell = 1, \ldots, k$. Damit gilt

$$\underline{x}^0 - \underline{w} - \underline{x} = p_k(A)\underline{e}^0$$

mit einem Matrix–Polynom $p_k \in \Pi_k^1$, d.h. $\underline{x}^k$ ist Lösung des Minimierungsproblems

$$\|\underline{x}^k - \underline{x}\|_A = \min_{p_k \in \Pi_k^1} \|p_k(A)\underline{e}^0\|_A.$$

Die Eigenvektoren $\{\underline{v}^j\}_{j=1}^n$ der symmetrischen und positiv definiten Matrix A bilden ein Orthonormalsystem. Die zugehörigen Eigenwerte von A seien $\lambda_j(A)$. Dann ergibt sich für den Anfangsfehler

$$\underline{e}^0 = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)\underline{v}^j$$

und in der Folge

$$p_k(A)\underline{e}^0 = p_k(A) \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)\underline{v}^j = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(A)\underline{v}^j = \sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A))\underline{v}^j.$$

Aus der Orthonormalität der Eigenvektoren $\underline{v}^j$ ergibt sich weiterhin

$$\begin{aligned}
\|p_k(A)\underline{e}^0\|_A^2 &= (Ap_k(A)\underline{e}^0, p_k(A)\underline{e}^0) \\
&= (A\sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A))\underline{v}^j, \sum_{i=1}^n (\underline{e}^0, \underline{v}^i) p_k(\lambda_i(A))\underline{v}^i) \\
&= (\sum_{j=1}^n (\underline{e}^0, \underline{v}^j) p_k(\lambda_j(A))\lambda_j(A)\underline{v}^j, \sum_{i=1}^n (\underline{e}^0, \underline{v}^i) p_k(\lambda_i(A))\underline{v}^i) \\
&= \sum_{i=1}^n \sum_{j=1}^n (\underline{e}^0, \underline{v}^i)(\underline{e}^0, \underline{v}^j) p_k(\lambda_i(A)) p_k(\lambda_j(A))\lambda_j(A)(\underline{v}^j, \underline{v}^i) \\
&= \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)^2 [p_k(\lambda_j(A))]^2 \lambda_j(A) \\
&\le \max_{j=1,\ldots,n} [p_k(\lambda_j(A))]^2 \sum_{j=1}^n (\underline{e}^0, \underline{v}^j)^2 \lambda_j(A) = \max_{j=1,\ldots,n} [p_k(\lambda_j(A))]^2 \, \|\underline{e}^0\|_A^2
\end{aligned}$$

und somit

$$\|\underline{x}^k - \underline{x}\|_A \leq \min_{p_k \in \Pi_k^1} \max_{j=1,\ldots,n} |p_k(\lambda_j(A))| \, \|\underline{e}^0\|_A \leq \min_{p_k \in \Pi_k^1} \max_{\lambda \in [\lambda_{\min}(A), \lambda_{\max}(A)]} |p_k(\lambda)| \, \|\underline{e}^0\|_A \, .$$

Mit Satz 1.1 ist schließlich

$$\min_{p_k \in \Pi_k^1} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p_k(\lambda)| \; = \; \frac{2q^k}{1+q^{2k}}, \quad q \; = \; \frac{\sqrt{\lambda_{\max}(A)} + \sqrt{\lambda_{\min}(A)}}{\sqrt{\lambda_{\max}(A)} - \sqrt{\lambda_{\min}(A)}} \; = \; \frac{\sqrt{\kappa_2(A)}+1}{\sqrt{\kappa_2(A)}-1}.$$

■

Die Konvergenzgeschwindigkeit des konjugierten Gradientenverfahrens wird maßgeblich durch die extremalen Eigenwerte der Matrix A bestimmt. Aus der Charakterisierung der extremalen Eigenwerte mittels des Rayleigh–Quotienten,

$$\lambda_{\min}(A) \; = \; \min_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})} \; \leq \; \max_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})} \; = \; \lambda_{\max}(A),$$

folgt

$$\lambda_{\min}(A)\,(\underline{x},\underline{x}) \; \leq \; (A\underline{x},\underline{x}) \; \leq \; \lambda_{\max}(A)\,(\underline{x},\underline{x}) \quad \text{für alle } \underline{x} \in \mathbb{R}^n.$$

Aus den **Spektraläquivalenzungleichungen**

$$c_1^A\,(\underline{x},\underline{x}) \; \leq \; (A\underline{x},\underline{x}) \; \leq \; c_2^A\,(\underline{x},\underline{x}) \quad \text{für alle } \underline{x} \in \mathbb{R}^n$$

folgt deshalb für die Abschätzung der spektralen Konditionszahl

$$\kappa_2(A) \; = \; \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \; \leq \; \frac{c_2^A}{c_1^A} \, .$$

Abhängig von der jeweiligen Anwendung strebt $\kappa_2(A) \to \infty$ für $n \to \infty$. Ziel ist deshalb die Herleitung eines Verfahrens, welches eine beschränkte Anzahl notwendiger Iterationsschritte zum Erreichen einer vorgegebenen relativen Genauigkeit ε unabhängig von der Dimension n gewährleistet.

Eine symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n \times n}$ gestattet die Faktorisierung

$$B \; = \; V \mathrm{diag}(\lambda_k(B)) V^\top$$

mir der durch die orthonormalen Eigenvektoren von B gebildeten orthonormalen Matrix V. Die Positivität der Eigenwerte $\lambda_k(B)$ ermöglicht die Definition der symmetrischen und positiv definiten Matrix

$$B^{1/2} \; = \; V \mathrm{diag}(\sqrt{\lambda_k(B)}) V^\top$$

mit $B^{1/2}B^{1/2} = B$ und der inversen Matrix $B^{-1/2} = (B^{1/2})^{-1}$.

Anstelle des linearen Gleichungssystems $A\underline{x} = \underline{f}$ wird jetzt das transformierte System

$$B^{-1/2}AB^{-1/2}B^{1/2}\underline{x} \; = \; B^{-1/2}\underline{f}$$

bzw. $\widetilde{A}\widetilde{\underline{x}} = \widetilde{\underline{f}}$ mit

$$\widetilde{A} = B^{-1/2}AB^{-1/2}, \quad \widetilde{\underline{x}} = B^{1/2}\underline{x}, \quad \widetilde{\underline{f}} = B^{-1/2}\underline{f}$$

betrachtet. Für die symmetrisch und positiv definite Matrix $\widetilde{A}$ kann nun das in Algorithmus 3.2 angegebene konjugierte Gradientenverfahren angewendet werden. Dessen Konvergenzgeschwindigkeit ergibt sich aus der Abschätzung der spektralen Konditionszahl $\kappa_2(\widetilde{A})$, welche unmittelbar aus den Spektraläquivalenzungleichungen

$$c_1^{\widetilde{A}}\,(\widetilde{\underline{x}},\widetilde{\underline{x}}) \; \leq \; (\widetilde{A}\widetilde{\underline{x}},\widetilde{\underline{x}}) \; \leq \; c_2^{\widetilde{A}}\,(\widetilde{\underline{x}},\widetilde{\underline{x}}) \quad \text{für alle } \widetilde{\underline{x}} \in \mathbb{R}^n$$

folgt. Einsetzen der Transformationen ergibt die dazu äquivalenten Spektraläquivalenzungleichungen

$$c_1^{\widetilde{A}}\,(B\underline{x},\underline{x}) \,\leq\, (A\underline{x},\underline{x}) \,\leq\, c_2^{\widetilde{A}}\,(B\underline{x},\underline{x}) \quad \text{für alle } \underline{x} \in I\!R^n .$$

Für die Lösung des linearen Gleichungssystems $\widetilde{A}\underline{\widetilde{x}} = \underline{\widetilde{f}}$ mit der symmetrisch und positiv definiten Matrix $\widetilde{A}$ kann die in Algorithmus 3.2 angegebene Iterationsvorschrift angewendet werden. Mit den zugehörigen Transformationen ergibt sich für die Näherungslösung $\underline{\widetilde{x}}^k$ sowie für das zugehörige Residuum $\underline{\widetilde{r}}^k$

$$\underline{\widetilde{x}}^k \,=\, B^{1/2}\underline{x}^k, \quad \underline{\widetilde{r}}^k \,=\, \widetilde{A}\underline{\widetilde{x}}^k - \underline{\widetilde{f}} \,=\, B^{-1/2}(A\underline{x}^k - \underline{f}) \,=\, B^{-1/2}\underline{r}^k .$$

Aus dem Ansatz

$$\underline{\widetilde{x}} \,=\, \underline{\widetilde{x}}^0 - \sum_{\ell=0}^{n-1} \widetilde{\alpha}_\ell \underline{\widetilde{p}}^\ell, \quad \widetilde{\alpha}_\ell \,=\, \frac{(\underline{\widetilde{r}}^k,\underline{\widetilde{r}}^k)}{(\widetilde{A}\underline{\widetilde{p}}^k,\underline{\widetilde{p}}^k)}$$

für die Lösung $\underline{\widetilde{x}}$ des transformierten linearen Gleichungssystems $\widetilde{A}\underline{\widetilde{x}} = \underline{\widetilde{f}}$ folgt durch Multiplikation mit $B^{-1/2}$

$$\underline{x} \,=\, \underline{x}^0 - \sum_{\ell=0}^{n-1} \widetilde{\alpha}_\ell B^{-1/2}\underline{\widetilde{p}}^\ell \,=\, \underline{x}^0 - \sum_{\ell=0}^{n-1} \widetilde{\alpha}_\ell \underline{p}^\ell$$

mit $\underline{p}^\ell = B^{-1/2}\underline{\widetilde{p}}^\ell$ bzw. $\underline{\widetilde{p}}^\ell = B^{1/2}\underline{p}^\ell$. Weiter ist

$$\widetilde{\varrho}_k \,=\, (\underline{\widetilde{r}}^k,\underline{\widetilde{r}}^k) \,=\, (B^{-1}\underline{r}^k,\underline{r}^k)$$

sowie

$$\widetilde{\sigma}_k \,=\, (\widetilde{A}\underline{\widetilde{p}}^k,\underline{\widetilde{p}}^k) \,=\, (A\underline{p}^k,\underline{p}^k).$$

Für die Konstruktion der transformierten Suchrichtungen $\underline{\widetilde{p}}^{k+1}$ ergibt sich schließlich

$$\underline{\widetilde{p}}^{k+1} \,=\, \underline{\widetilde{r}}^{k+1} + \widetilde{\beta}_k \underline{\widetilde{p}}^k$$

bzw. durch Multiplikation mit $B^{-1/2}$

$$\underline{p}^{k+1} \,=\, B^{-1}\underline{r}^{k+1} + \widetilde{\beta}_k \underline{p}^k .$$

Die resultierende Iterationsvorschrift des vorkonditionierten konjugierten Gradientenverfahrens ist in Algorithmus 3.3 angegeben.

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in I\!R^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$.
Berechne $\underline{v}^0 = B^{-1}\underline{r}^0,\ \underline{p}^0 := \underline{v}^0,\ \varrho_0 = (\underline{v}^0,\underline{r}^0)$. Stoppe, falls $\varrho_0 < \varepsilon^2$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist.

Berechne für $k = 0, 1, \ldots, n-2$:

$\underline{s}^k = A\underline{p}^k,\ \sigma_k = (\underline{s}^k,\underline{p}^k),\ \alpha_k = \dfrac{\varrho_k}{\sigma_k}$

$\underline{x}^{k+1} := \underline{x}^k - \alpha_k \underline{p}^k$

$\underline{r}^{k+1} := \underline{r}^k - \alpha_k \underline{s}^k$

$\underline{v}^{k+1} = B^{-1}\underline{r}^{k+1}$

$\varrho_{k+1} := (\underline{v}^{k+1},\underline{r}^{k+1})$

Stoppe, falls $\varrho_{k+1} < \varepsilon^2\varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Berechne andernfalls die neue Suchrichtung

$\underline{p}^{k+1} := \underline{v}^{k+1} + \beta_k \underline{p}^k,\ \beta_k := \dfrac{\varrho_{k+1}}{\varrho_k}$

Algorithmus 3.3: Konjugiertes Gradientenverfahrens mit Vorkonditionierung.

Neben einer Matrix–Vektor–Multiplikation mit A ist in Algorithmus 3.3 pro Iterationsschritt jeweils eine Anwendung der Vorkonditionierung $\underline{v} = B^{-1}\underline{r}$ zu realisieren. Damit muß die Matrix B neben den entsprechenden **Spektraläquivalenzungleichungen** auch eine **effiziente Anwendung** von B^{-1} ermöglichen. Sind beide Bedingungen efüllt, so wird B als **Vorkonditionierung** zu A bezeichnet. Diese ist in der Regel **problemabhängig** zu konstruieren und stellt in vielen Anwendungen eine Herausforderung dar.

3.2 Verfahren des minimalen Residuums

Betrachtet wird jetzt das lineare Gleichungsystem $A\underline{x} = \underline{f}$ mit einer invertierbaren Matrix A, d.h. es wird weder die Symmetrie noch die positive Definitheit der Systemmatrix A vorausgesetzt. Wie beim CG–Verfahren wird für eine Anfangsnäherung $\underline{x}^0$ das zugehörige Residuum $\underline{r}^0 = A\underline{x}^0 - \underline{f}$ und der dadurch induzierte Krylov–Raum

$$S_k(A, \underline{r}^0) \;=\; \text{span}\left\{\underline{r}^0, A\underline{r}^0, \ldots, A^k\underline{r}^0\right\}$$

eingeführt. Für diesen soll eine Basis $\{\underline{v}^\ell\}_{\ell=0}^{n-1}$ **orthonormaler** Vektoren mit

$$(\underline{v}^k, \underline{v}^\ell) \;=\; \delta_{k\ell}$$

konstruiert werden. Die Anwendung des Gram–Schmidt–Orthogonalisierungsverfahrens führt bei geeigneter Wahl der Ausgangsvektoren auf die **Methode von Arnoldi**:

Sei mit $\underline{x}^0 \in I\!R^n$ eine beliebige Startnäherung gegeben.
Berechne $\underline{r}^0 := A\underline{x}^0 - \underline{f}$ und setze

$$\underline{v}^0 = \frac{\underline{r}^0}{\|\underline{r}^0\|_2}.$$

Berechne für $k = 0, 1, \ldots, n-2$:

$$\hat{\underline{v}}^{k+1} = A\underline{v}^k - \sum_{\ell=0}^{k} \beta_{k\ell}\underline{v}^\ell$$

mit

$$\beta_{k\ell} = (A\underline{v}^k, \underline{v}^\ell).$$

Abbruch für $\|\hat{\underline{v}}^k\|_2 = 0,$ setze andernfalls

$$\underline{v}^{k+1} = \frac{\hat{\underline{v}}^{k+1}}{\|\hat{\underline{v}}^{k+1}\|_2}.$$

Algorithmus 3.4: Methode von Arnoldi.

Der Ansatz der Näherungslösung

$$\underline{x}^{k+1} \;=\; \underline{x}^0 - \sum_{\ell=0}^{k} \alpha_\ell \underline{v}^\ell$$

ergibt für das zugehörige Residuum

$$\underline{r}^{k+1} \;=\; A\underline{x}^{k+1} - \underline{f} \;=\; \underline{r}^0 - \sum_{\ell=0}^{k} \alpha_\ell A\underline{v}^\ell, \quad \underline{r}^0 = A\underline{x}^0 - \underline{f}.$$

Zu bestimmen bleiben die Zerlegungskoeffizienten α_ℓ durch Minimierung des Residuums $\underline{r}^{k+1}$ in der Euklidischen Vektornorm,

$$\|\underline{r}^{k+1}\|_2 \;=\; \|\underline{r}^0 - \sum_{\ell=0}^{k} \alpha_\ell A\underline{v}^\ell\|_2 \to \min_{\alpha_0,\ldots,\alpha_k} .$$

Für eine alternative Darstellung des Residuenvektors $\underline{r}^{k+1}$ folgt aus der Methode von Arnoldi zunächst

$$A\underline{v}^\ell \;=\; \underline{\hat{v}}^{\ell+1} + \sum_{j=0}^{\ell} \beta_{\ell,j}\underline{v}^j \;=\; \|\underline{\hat{v}}^{\ell+1}\|_2 \underline{v}^{\ell+1} + \sum_{j=0}^{\ell} \beta_{\ell,j}\underline{v}^j \;=\; \sum_{j=0}^{\ell+1} \beta_{\ell,j}\underline{v}^j$$

mit

$$\beta_{\ell,j} \;=\; \begin{cases} (A\underline{v}^\ell, \underline{v}^j) & \text{für } j = 0, \ldots, \ell, \\ \|\underline{\hat{v}}^{\ell+1}\|_2 & \text{für } j = \ell + 1. \end{cases}$$

Dann ist

$$\underline{r}^{k+1} \;=\; \underline{r}^0 - \sum_{\ell=0}^{k} \alpha_\ell A \underline{v}^\ell \;=\; \underline{r}^0 - \sum_{\ell=0}^{k} \alpha_\ell \sum_{j=0}^{\ell+1} \beta_{\ell,j}\underline{v}^j \;=\; \underline{r}^0 - V_{k+1} H_k \underline{\alpha}$$

mit der durch die orthonormalen Vektoren $\underline{v}^j$ gebildeten Matrix

$$V_{k+1} \;=\; \left(\underline{v}^0, \underline{v}^1, \ldots, \underline{v}^{k+1}\right) \in I\!R^{n \times (k+2)}$$

und mit der durch

$$H_k[j,\ell] \;=\; \begin{cases} \beta_{\ell,j} & \text{für } j \le \ell + 1, \\ 0 & \text{für } j > \ell + 1 \end{cases}$$

definierten oberen **Hessenberg–Matrix** $H_k \in I\!R^{(k+2)\times(k+1)}$. Nach Konstruktion folgt weiterhin

$$\underline{r}^0 \;=\; \|\underline{r}^0\|_2 \underline{v}^0 \;=\; \|\underline{r}^0\|_2 \, V_{k+1} \underline{e}^0$$

mit dem ersten Einheitsvektor $\underline{e}^0 = (1, 0, \ldots, 0)^\top \in I\!R^{k+2}$. Mit der Invarianz der Euklidischen Vektornorm bezüglich orthonormalen Matrizen ergibt sich

$$\|\underline{r}^{k+1}\|_2 = \|\underline{r}^0 - V_{k+1} H_k \underline{\alpha}\|_2 = \|V_{k+1}\left(\|\underline{r}^0\|_2 \underline{e}^0 - H_k\underline{\alpha}\right)\|_2 = \|\|\underline{r}^0\|_2 \underline{e}^0 - H_k \underline{\alpha}\|_2 .$$

Wegen $H_k \in I\!R^{(k+2)\times(k+1)}$ und $\underline{\alpha} \in I\!R^{k+1}$ sowie $\underline{e}^0 \in I\!R^{k+2}$ entspricht die Forderung $H_k\underline{\alpha} = \|\underline{r}^0\|_2 \underline{e}^0$ einem überbestimmten linearen Gleichungssystem mit $k+2$ Gleichungen für $k+1$ Unbekannte. Zu bestimmen bleibt jener Lösungsvektor $\underline{\alpha} \in I\!R^{k+1}$, der das verbleibende Residuum minimiert.

Sei $Q_k \in I\!R^{(k+2)\times(k+2)}$ eine orthonormale Matrix mit $Q_k^\top Q_k = I_{k+2}$, so daß $Q_k H_k \in I\!R^{(k+2)\times(k+1)}$ obere Dreiecksgestalt besitzt. Dann gilt

$$\begin{aligned} \|\underline{r}^{k+1}\|_2 \;&=\; \| \, \|\underline{r}^0\|_2 \underline{e}^0 - H_k \underline{\alpha}\|_2 \;=\; \| \, \|\underline{r}^0\|_2 Q_k \underline{e}^0 - Q_k H_k \underline{\alpha}\|_2 \\ &=\; \| \, \|\underline{r}^0\|_2 Q_k \underline{e}^0 - R_k \underline{\alpha}\|_2 \;=\; \|\underline{r}^0\|_2 \, |(Q_k \underline{e}^0)_{k+1}| , \end{aligned}$$

falls

$$(R_k \underline{\alpha})_\ell \;=\; \|\underline{r}^0\|_2 \, (Q_k \underline{e}^0)_\ell \quad \text{für } \ell = 0, \ldots, k$$

erfüllt ist.

Zu bestimmen bleibt eine orthonormale Matrix $Q_k \in I\!R^{(k+2)\times(k+2)}$, welche die obere Hessenberg–Matrix

$$H_k \;=\; \begin{pmatrix} \beta_{0,0} & \beta_{1,0} & \cdots & \beta_{k,0} \\ \beta_{0,1} & \beta_{1,1} & & \vdots \\ 0 & \beta_{1,2} & \ddots & \vdots \\ & 0 & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \in I\!R^{(k+2)\times(k+1)}$$

in eine obere Dreiecks–Matrix

$$R_k \;=\; Q_k H_k \;=\; \begin{pmatrix} r_{0,0} & r_{0,1} & \dots & r_{0,k} \\ 0 & r_{1,1} & & \vdots \\ 0 & 0 & \ddots & \vdots \\ & & \ddots & r_{k,k} \\ & & & 0 \end{pmatrix} \in I\!R^{(k+2)\times(k+1)}$$

transformiert. Ausgehend von dem Spaltenvektor

$$\underline{h}^j \;=\; (\beta_{j,0}, \dots, \beta_{j,j-1}, \beta_{j,j}, \beta_{j,j+1}, 0, \dots, 0)^\top \in I\!R^{k+2}$$

ist zunächst jene orthonormale Matrix G_j gesucht, so daß

$$G_j \underline{h}^j \;=\; \Big(\beta_{j,0}, \dots, \beta_{j,j-1}, \widetilde{\beta}_{j,j}, 0, 0, \dots, 0\Big)^\top .$$

Offenbar ist die Betrachtung einer orthonormalen Transformationsmatrix $\widetilde{G}_j \in I\!R^{2\times 2}$ mit

$$\widetilde{G}_j \begin{pmatrix} \beta_{j,j} \\ \beta_{j,j+1} \end{pmatrix} \;=\; \begin{pmatrix} \widetilde{\beta}_{j,j} \\ 0 \end{pmatrix}$$

mit der allgemeinen Darstellung

$$\widetilde{G}_j \;=\; \begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix}, \quad a_j^2 + b_j^2 \;=\; 1,$$

ausreichend. Die Koeffizienten a_j und b_j ergeben sich aus der Forderung

$$-b_j \beta_{j,j} + a_j \beta_{j,j+1} \;=\; 0\,.$$

Unter Beachtung der Normierung $a_j^2 + b_j^2 = 1$ folgt daraus

$$a_j \;=\; \frac{\beta_{j,j}}{\sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2}}, \quad b_j \;=\; \frac{\beta_{j,j+1}}{\sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2}}$$

und daher

$$\widetilde{\beta}_{j,j} \;=\; a_j \beta_{j,j} + b_j \beta_{j,j+1} \;=\; \sqrt{\beta_{j,j}^2 + \beta_{j,j+1}^2} \;>\; 0,$$

wenn $\beta_{j,j+1} > 0$ vorausgesetzt wird. Für $j = 0, \dots, k$ sind die resultierenden Transformationsmatrizen die **Givens–Rotationen**

$$G_j \;=\; \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & a_j & b_j & & & \\ & & & -b_j & a_j & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix} \in I\!R^{(k+2)\times(k+2)}$$

mit $G_j[j,j] = G_j[j+1,j+1] = a_j$.

Deren rekursive Anwendung liefert mit

$$
\begin{aligned}
G_kG_{k-1}\dots G_2G_1G_0H_k &= G_kG_{k-1}\dots G_2G_1G_0 \begin{pmatrix} \beta_{0,0} & \beta_{1,0} & \dots & \beta_{k,0} \\ \beta_{0,1} & \beta_{1,1} & \dots & \beta_{k,1} \\ 0 & \beta_{1,2} & \ddots & \vdots \\ & 0 & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= G_kG_{k-1}\dots G_2G_1 \begin{pmatrix} \widetilde{\beta}_{0,0} & \widetilde{\beta}_{1,0} & \dots & \widetilde{\beta}_{k,0} \\ 0 & \bar{\beta}_{1,1} & \dots & \bar{\beta}_{k,1} \\ & \beta_{1,2} & \ddots & \vdots \\ & & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= G_kG_{k-1}\dots G_2 \begin{pmatrix} \widetilde{\beta}_{0,0} & \widetilde{\beta}_{1,0} & \dots & \widetilde{\beta}_{k,0} \\ 0 & \widetilde{\beta}_{1,1} & \dots & \widetilde{\beta}_{k,1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \beta_{k,k} \\ & & & \beta_{k,k+1} \end{pmatrix} \\
&= \begin{pmatrix} \widetilde{\beta}_{0,0} & \widetilde{\beta}_{1,0} & \dots & \widetilde{\beta}_{k,0} \\ 0 & \widetilde{\beta}_{1,1} & \dots & \widetilde{\beta}_{k,1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \widetilde{\beta}_{k,k} \\ & & & 0 \end{pmatrix} = R_k
\end{aligned}
$$

die gewünschte obere Dreiecksmatrix, deren Invertierbarkeit aus der Positivität der Diagonaleinträge $\widetilde{\beta}_{j,j}$ folgt. Beim Übergang von H_k zu H_{k+1}, d.h. bei Hinzunahme eines weiteren Spaltenvektors $\underline{h}^{k+1}$ bzw. einer weiteren Suchrichtung $\underline{v}^{k+2}$, und vor der Anwendung der zugehörigen orthonormalen Matrix G_{k+1} sind **alle** vorherigen Transformationen $G_k, \dots, G_0$ auf $\underline{h}^{k+1}$ anzuwenden.
Nach Konstruktion ist

$$
Q_k = G_kG_{k-1}\dots G_1G_0 \in I\!R^{(k+2)\times(k+2)}
$$

orthonormal und zu untersuchen bleibt die Auswertung von

$$
\begin{aligned}
Q_k\underline{e}^0 = G_k\dots G_0 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} &= G_k\dots G_1 \begin{pmatrix} a_0 \\ -b_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = G_k\dots G_2 \begin{pmatrix} a_0 \\ a_1(-b_0) \\ (-b_0)(-b_1) \\ \vdots \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} a_0 \\ a_1(-b_0) \\ (-b_0)(-b_1) \\ \vdots \\ a_k(-b_0)\cdots(-b_{k-1}) \\ (-b_0)\cdots(-b_k) \end{pmatrix} \in I\!R^{k+2}.
\end{aligned}
$$

Während die ersten $k+1$ Komponenten von $Q_k\underline{e}^0$ die rechte Seite des linearen Gleichungssystems zur Bestimmung des Koeffizientenvektors $\underline{\alpha} \in I\!R^{k+1}$ bilden, beschreibt die letzte Komponente das verbleibende Residuum

$$\varrho_{k+1} \;=\; \|\underline{e}^0\|_2\,|(Q_k\underline{e}^0)_{k+1}| \;=\; \|\underline{e}^0\|_2 \prod_{j=0}^{k} b_j\,.$$

Wegen

$$b_j \;=\; \frac{\beta_{j,j+1}}{\sqrt{\beta_{j,j+1}^2+\beta_{j,j}^2}} \;=\; \frac{\|\hat{\underline{v}}^j\|_2}{\sqrt{\|\hat{\underline{v}}^j\|_2^2+(A\underline{v}^j,\underline{v}^j)^2}}$$

folgt für $(A\underline{v}^j,\underline{v}^j) \neq 0$ wegen $\beta_j < 1$ ein monotones Abklingverhalten des Fehlers. Insbesondere für die Abbruch–Situation der Methode von Arnoldi, $\|\hat{\underline{v}}^k\|_2 = 0$, folgt $b_k = 0$ und somit

$$\varrho_{k+1} \;=\; \|\underline{r}^{k+1}\|_2 \;=\; 0,$$

d.h. $\underline{x}^{k+1} = \underline{x}$ ist die exakte Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$.
Das resultierende Iterationsverfahren ist das in Algorithmus 3.5 angegebene **verallgemeinerte Verfahren des minimalen Residuums (GMRES)** [12].

Für eine beliebig gegebene Startnäherung $\underline{x}^0 \in \mathbb{R}^n$ sei $\underline{r}^0 = A\underline{x}^0 - \underline{f}$.
Berechne $\varrho_0 = \|\underline{r}^0\|_2$. Stoppe, falls $\varrho_0 < \varepsilon$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist. Setze andernfalls $\underline{v}^0 = \dfrac{1}{\varrho_0}\underline{r}^0$, $p_0 = \varrho_0$.
Berechne für $k = 0, 1, \ldots, n-2$:

$$\underline{w}^k = A\underline{v}^k$$

$$\widetilde{\underline{v}}^{k+1} = \underline{w}^k - \sum_{\ell=0}^{k} \beta_{k\ell}\underline{v}^\ell, \quad \beta_{k\ell} = (\underline{w}^k, \underline{v}^\ell), \qquad \beta_{kk+1} \;=\; \|\widetilde{\underline{v}}^{k+1}\|_2$$

Falls $\beta_{kk+1} = 0$, stoppe und berechne Näherungslösung $\underline{x}^{k+1}$.

$$\underline{v}^{k+1} \;=\; \frac{1}{\beta_{kk+1}}\widetilde{\underline{v}}^{k+1}$$

Berechne für $\ell = 0, \ldots, k-1$:

$$\widetilde{\beta}_{k\ell} = a_\ell\beta_{k\ell} + b_\ell\beta_{k\ell+1}$$

$$\widetilde{\beta}_{k\ell+1} = -b_\ell\beta_{k\ell} + a_\ell\beta_{k\ell+1}$$

$$a_k \;=\; \frac{\beta_{kk}}{\sqrt{\beta_{kk}^2+\beta_{kk+1}^2}}, \quad b_k \;=\; \frac{\beta_{kk+1}}{\sqrt{\beta_{kk}^2+\beta_{kk+1}^2}}, \quad \widetilde{\beta}_{kk} \;=\; \sqrt{\beta_{kk}^2+\beta_{kk+1}^2}$$

$$p_{k+1} \;=\; -b_k p_k, \quad p_k \;=\; a_k p_k, \quad \varrho_{k+1} \;=\; |p_{k+1}|$$

Stoppe, falls $\varrho_{k+1} < \varepsilon\varrho_0$ mit einer vorgegebenen Fehlergenauigkeit ε erreicht ist und berechne die Näherungslösung:

Berechne für $\ell = k, k-1, \ldots, 0$:

$$\alpha_\ell \;=\; \frac{1}{\beta_{\ell\ell}}\left[p_\ell - \sum_{j=\ell+1}^{k}\beta_{\ell j}\alpha_j\right]$$

$$\underline{x}^{k+1} \;=\; \underline{x}^0 - \sum_{\ell=0}^{k}\alpha_\ell\underline{v}^\ell$$

Algorithmus 3.5: Iterationsvorschrift GMRES Verfahren.

Übungsaufgaben

3.1. Gegeben sei die Matrix $A \in \mathbb{R}^{n \times n}$ durch

$$A = I + \underline{e}\,\underline{e}^\top \quad \text{mit } \underline{e} = (1, \ldots, 1)^\top .$$

Man bestimme alle Eigenwerte der Matrix A.

3.2. Für die in Aufgabe **3.1** gegebene Matrix A bestimme man die Anzahl der zur näherungsweisen Lösung des linearen Gleichungssystems $A\underline{x} = \underline{f}$ notwendigen CG–Iterationen, um für die Startnäherung $\underline{x}^0 = \underline{0}$ eine relative Fehlerreduktion von ε zu gewährleisten.

3.3. Betrachtet werde die Massematrix M_h der stückweise linearen Ansatzfunktionen bezüglich einer gleichmässigen Unterteilung des Intervalles $[0, 1]$ in n Elemente der Länge $h = 1/n$. Man finde positive Konstanten c_1 und c_2 derart,so daß die Spektraläquivalenzungleichungen

$$c_1\, h\, \|\underline{u}\|_2^2 \;\leq\; (M_h \underline{u}, \underline{u}) \;\leq\; c_2\, h\, \|\underline{u}\|_2^2$$

für alle Vektoren $\underline{u} \in \mathbb{R}^{n+1}$ erfüllt sind.

Hinweis: Man drücke die Bilinearform $(M_h \underline{u}, \underline{u})$ als Integral aus und betrachte die lokalen Beiträge in Matrixform der Dimension 2.

Kapitel 4

Nichtlineare Gleichungen

Für die näherungsweise Bestimmung von **Nullstellen** einer gegebenen Funktion $f : [a,b] \to I\!R$, d.h. zur näherungsweisen Lösung der Gleichung

$$f(x) = 0,$$

werden in diesem Kapitel **Iterationsverfahren** betrachtet, welche eine Folge von Näherungslösungen $\{x_k\}_{k \in I\!N}$ mit

$$\lim_{k \to \infty} x_k \;=\; \bar{x}, \quad f(\bar{x}) \;=\; 0$$

erzeugen. Die Folge $\{x_k\}_{k \in I\!N}$ konvergiert gegen den Grenzwert $\bar{x}$ mit der Ordnung $p \geq 1$, falls

$$|x_{k+1} - \bar{x}| \;\leq\; c\, |x_k - \bar{x}|^p \quad \text{für alle } k \geq k_0$$

mit einem gewissen $k_0 \in I\!N$ und einer positiven Konstanten $c \in I\!R$ gilt. Für $p = 1$ muß $c < 1$ vorausgesetzt werden. Das Konvergenzverhalten ist im allgemeinen abhängig von der Wahl der Anfangsnäherung x_0. Muß x_0 in einer geeigneten Umgebung von $\bar{x}$ vorausgesetzt werden, so spricht man von **lokaler Konvergenz**. Gilt die Konvergenz für eine beliebige Startnäherung $x_0 \in [a,b]$, so folgt **globale Konvergenz**.

4.1 Bisektionsverfahren

Sei $f : [a,b] \to I\!R$ eine stetige Funktion und es gelte

$$f(a)\, f(b) \;<\; 0\,.$$

Daraus folgt die Existenz wenigstens einer Nullstelle $\bar{x} \in (a,b)$ mit $f(\bar{x}) = 0$. In Verbindung mit einer rekursiven Unterteilung des Intervalles $[a,b]$ kann daraus das **Bisektionsverfahren** zur näherungsweisen Bestimmung der Nullstelle $\bar{x} \in (a,b)$ abgeleitet werden, siehe Algorithmus 4.1.

Setze
$a_0 = a$, $b_0 = b$ mit $f(a_0) f(b_0) < 0$.
Definiere für $k = 0, 1, 2, \ldots$
$$x_k = \frac{1}{2}(a_k + b_k)$$
und setze
$$a_{k+1} = a_k,\; b_{k+1} = x_k \quad \text{falls } f(x_k) f(a_k) < 0,$$
$$a_{k+1} = x_k,\; b_{k+1} = b_k \quad \text{falls } f(x_k) f(a_k) > 0.$$

Algorithmus 4.1: Bisektionsverfahren.

Im Fall $f(x_k) = 0$ ist $x_k = \bar{x}$ die exakte Nullstelle. Nach Konstruktion gilt für die Nullstelle $\bar{x} \in (a_k, b_k)$. Für die Näherungslösung $x_k = a_k$ (bzw. $x_k = b_k$) ist also

$$|x_k - \bar{x}| \le |b_k - a_k| .$$

Die rekursive Bisektion des Intervalles $[a, b]$ liefert andererseits

$$|b_k - a_k| = 2^{-k}\,|b - a|,$$

so daß insgesamt die **Fehlerabschätzung**

$$|x_k - \bar{x}| \le \frac{1}{2^k}\,|b - a|$$

folgt. Damit ergibt sich die Konvergenz für jede beliebige Startnäherung, d.h. es gilt globale Konvergenz.

Beispiel 4.1 *Zur Bestimmung der Nullstelle* $\bar{x} = 2$ *von* $f(x) = x^2 - 4$ *im Intervall* $[a, b] = [1, 4]$ *wird das in Algorithmus* 5.1 *angegebene Bisektionsverfahren angewendet:*

k	a_k	b_k	x_k	$f(a_k)$	$f(b_k)$	$f(x_k)$	$\|x_k - \bar{x}\|$
0	1	4	2.5	-3	12	2.25	0.5
1	1	2.5	1.75	-3	2.25	-0.9375	0.25
2	1.75	2.5	2.125	-0.9375	2.25	0.515625	0.125

Die Nachteile des Bisektionsverfahrens sind ein langsames Konvergenzverhalten und einer unter Umständen nicht monotone Fehlerreduktion, wie das nächste Beispiel zeigt.

Beispiel 4.2 *Zur Bestimmung der Nullstelle* $\bar{x} \approx 0.9062$ *von*

$$f(x) = \frac{x}{8}(63x^4 - 70x^2 + 15)$$

im Intervall $[a, b] = [0.8, 1]$ *ergibt sich*

k	a_k	b_k	x_k	$f(a_k)$	$f(b_k)$	$f(x_k)$	$\|x_k - \bar{x}\|$
0	0.8	1	0.9	-0.399	1	-0.04	0.0062
1	0.9	1	0.95	-0.04	1	0.3727	0.038

und somit

$$|x_1 - \bar{x}| > |x_0 - \bar{x}| .$$

4.2 Methode der sukzessiven Approximation

Für die Herleitung eines allgemeinen Iterationsverfahrens zur Lösung der nichtlinearen Gleichung

$$f(x) = 0 \quad \text{in } [a, b]$$

werde eine dazu äquivalente **Fixpunktgleichung**

$$x = \Phi(x) \quad \text{in } [a, b]$$

betrachtet. Daraus ergibt sich, ausgehend von einer Startnäherung $x_0 \in [a, b]$, die **Methode der sukzessiven Approximation**,

$$x_{k+1} = \Phi(x_k) \quad \text{für } k = 0, 1, 2, \ldots .$$

Zu untersuchen bleibt die **Konstruktion** von $\Phi(x)$, der Nachweis der **Konvergenz** der Näherungslösungen x_k und die Abschätzung der **Konvergenzgeschwindigkeit**.

Beispiel 4.3 *Zur Lösung der Gleichung $f(x) = x^2 - 4 = 0$ werden zunächst die folgenden Äquivalenzumformungen betrachtet:*

$$x^2 - 4 = 0 \quad \Leftrightarrow \quad x^2 = 4 \quad \Leftrightarrow \quad 2x^2 = x^2 + 4 \quad \Leftrightarrow \quad x = \frac{1}{2}\Big(x + \frac{4}{x}\Big).$$

Daraus ergibt sich, ausgehend von einer Anfangsnäherung $x_0 \in [a, b]$, die Iterationsvorschrift des **Babylonischen Wurzelziehens**,

$$x_{k+1} \; = \; \frac{1}{2}\Big(x_k + \frac{4}{x_k}\Big) \quad \text{für } k =, 1, 2, \ldots .$$

Für die Anfangsnäherung $x_0 = 4$ folgt:

k	x_k	$\lvert x_k - \bar{x}\rvert$
0	4	2
1	2.5	0.5
2	2.05	0.05
3	2.00061	0.00061

Für $a > 1$ wird nun für die Lösung der Gleichung $f(x) = x^2 - a = 0$ die Iterationsvorschrift

$$x_{k+1} \; = \; \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big) \; = \; \Phi(x_k) \quad \text{für } k = 0, 1, 2, \ldots, \quad x_0 \; = \; a,$$

betrachtet.

Lemma 4.1 *Es gilt*

$$x_k^2 \; \geq \; x_{k+1}^2 \; \geq \; a$$

sowie

$$x_k \; \geq \; x_{k+1} \; \geq \; 1.$$

Beweis: Aus $(\alpha - \beta)^2 \geq 0$ folgt $2\alpha\beta \leq \alpha^2 + \beta^2$ und somit

$$\alpha\beta \; \leq \; \frac{1}{4}(\alpha + \beta)^2.$$

Für $\alpha = x_k$ und $\beta = a/x_k$ gilt also

$$1 \; \leq \; a \; = \; x_k \frac{a}{x_k} \; \leq \; \frac{1}{4}\Big(x_k + \frac{a}{x_k}\Big)^2 \; = \; x_{k+1}^2$$

und somit $x_{k+1}^2 \geq a$ bzw. $x_{k+1} \geq 1$ für alle $k \in \mathbb{N}_0$. Insbesondere gilt also auch $x_k^2 \geq a$. Dann folgt

$$1 \; \leq \; x_{k+1} \; = \; \frac{1}{2}\Big(x_k + \frac{a}{x_k}\Big) \; \leq \; \frac{1}{2}\Big(x_k + \frac{x_k^2}{x_k}\Big) \; = \; x_k \,.$$

■

Folgerung 4.1 *Die Folge $\{x_k\}_{k \in \mathbb{N}}$ von Näherungslösungen ist monoton fallend und nach unten beschränkt. Daraus folgt die Konvergenz der Näherungslösungen x_k gegen die Lösung $\bar{x}$ der Fixpunktgleichung $\bar{x} = \Phi(\bar{x})$.*

Zur Abschätzung der Konvergenzgeschwindigkeit wird zunächst die Differenz der Rekursionsvorschrift

$$\Phi(x) \; = \; \frac{1}{2}\Big(x + \frac{a}{x}\Big)$$

für zwei Argumente $x, y \geq a$ abgeschätzt:

$$\begin{aligned} |\Phi(x) - \Phi(y)| &= \left|\frac{1}{2}\Big(x + \frac{a}{x}\Big) - \frac{1}{2}\Big(y + \frac{a}{y}\Big)\right| = \frac{1}{2}\left|x - y + a\Big(\frac{1}{x} - \frac{1}{y}\Big)\right| \\ &= \frac{1}{2}\left|x - y + a\,\frac{y-x}{xy}\right| = \frac{1}{2}\left|1 - \frac{a}{xy}\right|\,|x-y|\,. \end{aligned}$$

Für $x = x_{k+1} = \Phi(x_k)$ und $y = \bar{x} = \Phi(\bar{x})$ folgt mit $a = \bar{x}^2$

$$|x_{k+1} - \bar{x}| = |\Phi(x_k) - \Phi(\bar{x})| \leq \frac{1}{2}\left|1 - \frac{a}{x_k\bar{x}}\right|\,|x_k - \bar{x}| = \frac{1}{2}\frac{1}{x_k}|x_k - \bar{x}|^2$$

und wegen $x_k \geq 1$ ergibt sich die **quadratische** Konvergenzabschätzung

$$|x_{k+1} - \bar{x}| \leq \frac{1}{2}\,|x_k - \bar{x}|^2\,.$$

4.3 Banachscher Fixpunktsatz

Die Vorgehensweise des Babylonischen Wurzelziehens zur Lösung der Gleichung $f(x) = x^2 - a = 0$ soll nun zur Lösung allgemeiner nichtlinearen Gleichungen

$$f(x) = 0 \quad \text{in } [a,b]$$

bzw. der dazu äquivalenten Fixpunktgleichungen

$$x = \Phi(x) \quad \text{in } [a,b]$$

übertragen werden. Wie oben wird die **Methode der sukzessiven Approximation**

$$x_{k+1} = \Phi(x_k) \quad \text{für } k = 0, 1, 2, \ldots$$

betrachtet. Die Funktion Φ erfüllt in $D = [a,b]$ eine **Lipschitz–Bedingung** mit einer positiven **Lipschitz–Konstanten** c_L, falls

$$|\Phi(x) - \Phi(y)| \leq c_L\,|x - y|$$

für alle $x, y \in D = [a,b]$ erfüllt ist. Gilt $c_L = q < 1$, dann heißt Φ **Kontraktion** auf $D = [a,b]$.

Satz 4.1 (Banachscher Fixpunktsatz) *Sei $D = [a,b]$ abgeschlossen und $\Phi : D \to D$ eine Kontraktion auf D mit $q < 1$. Dann hat die Fixpunktgleichung $x = \Phi(x)$ genau eine Lösung $x_0 \in D$. Bildet man für $x_0 \in D := [a,b]$ die sukzessiven Approximationen $x_{k+1} = \Phi(x_k)$, so gelten die* **a priori** *Fehlerabschätzung*

$$|x_k - x| \leq \frac{q^k}{1-q}\,|x_1 - x_0|$$

sowie die **a posteriori** *Fehlerabschätzung*

$$|x_k - x| \leq \frac{q}{1-q}\,|x_k - x_{k-1}|\,.$$

Beweis: Wegen $\Phi : D \to D$ folgt für $x_0 \in D$ durch vollständige Induktion $x_{k+1} = \Phi(x_k) \in D$ für alle $k \in \mathbb{N}_0$. Durch rekursive Anwendung der Kontraktionsabschätzung folgt dann

$$|x_{k+1} - x_k| = |\Phi(x_k) - \Phi(x_{k-1})| \leq q\,|x_k - x_{k-1}| \leq q^k\,|x_1 - x_0|\,.$$

Für $p \in I\!N_0$ ist

$$\begin{aligned} |x_{k+p} - x_k| &= \left|\sum_{i=0}^{p-1}(x_{k+i+1} - x_{k+i})\right| \le \sum_{i=0}^{p-1} |x_{k+i+1} - x_{k+i}| \le \sum_{i=0}^{p-1} q^{k+i}|x_1 - x_0| \\ &= q^k\,|x_1 - x_0| \sum_{i=0}^{p-1} q^i = q^k \frac{1-q^p}{1-q}\,|x_1 - x_0| \le \frac{q^k}{1-q}\,|x_1 - x_0|\,. \end{aligned}$$

Daraus folgt

$$\lim_{k\to\infty} |x_{k+p} - x_k| \le \lim_{k\to\infty} \frac{q^k}{1-q}\,|x_1 - x_0| = 0$$

für alle $p \in I\!N_0$. $\{x_k\}_{k\in I\!N}$ ist also **Cauchy–Folge** und somit existiert der Grenzwert

$$\bar{x} = \lim_{k\to\infty} x_k \in D = \overline{D}$$

bzw.

$$\bar{x} = \lim_{i\to\infty} x_i = \lim_{p\to\infty} x_{k+p}$$

für ein fest gewähltes $k \in I\!N_0$. Damit folgt die a priori Fehlerabschätzung

$$|\bar{x} - x_k| = \left|\lim_{p\to\infty}(x_{k+p} - x_k)\right| \le \frac{q^k}{1-q}\,|x_1 - x_0|.$$

Zu zeigen bleibt, daß der Grenzwert $\bar{x} = \lim_{k\to\infty} x_k$ Lösung der Fixpunktgleichung $x = \Phi(x)$ ist. Es gilt

$$\begin{aligned} |\bar{x} - \Phi(\bar{x})| &= |\bar{x} - \Phi(x_k) + \Phi(x_k) - \Phi(\bar{x})| \le |\bar{x} - \Phi(x_k)| + |\Phi(x_k) - \Phi(\bar{x})| \\ &= |\bar{x} - x_{k+1}| + |\Phi(x_k) - \Phi(\bar{x})| \le |\bar{x} - x_{k+1}| + q\,|x_k - \bar{x}| \\ &\le \frac{q^{k+1}}{1-q}\,|x_1 - x_0| + q\,\frac{q^k}{1-q}\,|x_1 - x_0| = 2\,\frac{q^{k+1}}{1-q}\,|x_1 - x_0| \end{aligned}$$

für alle $k \in I\!N$. Daraus folgt

$$|\bar{x} - \Phi(\bar{x})| \le 2\,|x_1 - x_0| \lim_{k\to\infty} \frac{q^{k+1}}{1-q} = 0$$

und somit $\bar{x} = \Phi(\bar{x})$.
Für zwei Fixpunkte $\bar{x} = \Phi(\bar{x})$ und $\widetilde{x} = \Phi(\widetilde{x})$ ist

$$|\bar{x} - \widetilde{x}| = |\Phi(\bar{x}) - \Phi(\widetilde{x})| \le q\,|\bar{x} - \widetilde{x}|$$

und daher

$$0 \le (1-q)\,|\bar{x} - \widetilde{x}| \le 0$$

und wegen $q < 1$ folgt mit $\bar{x} = \widetilde{x}$ die Eindeutigkeit des Fixpunktes.
Mit

$$|x_k - \bar{x}| = |\Phi(x_{k-1}) - \Phi(\bar{x})| \le q\,|x_{k-1} - \bar{x}| \le q\,[|x_{k-1} - x_k| + |x_k - \bar{x}|]$$

folgt schließlich

$$(1-q)\,|x_k - \bar{x}| \le q\,|x_k - x_{k-1}|$$

und somit die a posteriori Fehlerabschätzung. ■

Beispiel 4.4 *Für die Iterationsvorschrift des Babylonischen Wurzelziehens,*

$$\Phi(x) \;=\; \frac{1}{2}\Big(x + \frac{a}{x}\Big),$$

gilt

$$\Phi(x) - \Phi(y) \;=\; \frac{1}{2}\left(1 - \frac{a}{xy}\right)(x-y)\,.$$

Daraus folgt eine Lipschitz–Abschätzung mit

$$c_L \;=\; \frac{1}{2}\max_{x,y\in D}\left|1 - \frac{a}{xy}\right|\,.$$

Für eine Kontraktion in D mit $c_L = q < 1$ muß also

$$\left|1 - \frac{a}{xy}\right| \;<\; 2$$

bzw.

$$-2 \;<\; 1 - \frac{a}{xy} \;<\; 2$$

für alle $x, y \in D$ gelten. Die obere Abschätzung ist für Argumente $a > 1$ und $x, y > 0$ stets erfüllt. Die untere Abschätzung ist äquivalent zu

$$\frac{a}{xy} \;\leq\; 3\,.$$

Dies ist zum Beispiel erfüllt für alle $x, y \geq \sqrt{a}$. Damit ist $\Phi(x) \;=\; \frac{1}{2}\Big(x + \frac{a}{x}\Big)$ Kontraktion auf $D = [\sqrt{a}, \infty)$.

Der Banachsche Fixpunktsatz liefert nur eine lineare Konvergenzabschätzung. Beim Babylonischen Wurzelziehen ergab sich jedoch eine quadratische Konvergenzordnung. Diese kann auch aus dem folgenden Satz abgeleitet werden.

Satz 4.2 *Sei Φ eine in D p–mal stetig differenzierbare Funktion und im Fixpunkt $\bar{x} = \Phi(\bar{x})$ gelte*

$$\Phi'(\bar{x}) = \Phi''(\bar{x}) = \ldots = \Phi^{(p-1)}(\bar{x}) = 0, \quad \varphi^{(p)}(\bar{x}) \neq 0.$$

Dann gibt es eine Umgebung von $\bar{x}$, $U_\delta(\bar{x}) \;=\; \{y \in D \,:\, |y - \bar{x}| < \delta\}$, so daß

$$|x_{k+1} - \bar{x}| \;\leq\; \frac{1}{p!}\max_{y\in U_\delta(\bar{x})}|\varphi^{(p)}(y)|\;|x_k - \bar{x}|^p$$

für $x_k \in U_\delta(\bar{x})$ gilt.

Beweis: Die Taylor–Entwicklung von $x_{k+1} = \Phi(x_k)$ um den Fixpunkt $\bar{x}$ ergibt

$$x_{k+1} \;=\; \Phi(x_k) \;=\; \Phi(\bar{x}) + \sum_{n=1}^{p-1}\frac{\Phi^{(n)}(\bar{x})}{n!}(x_k - \bar{x})^n + \frac{\Phi^{(p)}(\xi)}{p!}(x_k - \bar{x})^p$$

mit einer Zwischenwertstelle $\xi \in (x_k, \bar{x})$ für $x_k < \bar{x}$ bzw. $\xi \in (\bar{x}, x_k)$ für $\bar{x} < x_k$. Wegen $\Phi^{(n)}(\bar{x}) = 0$ für $n = 1, \ldots, p-1$ und $\bar{x} = \Phi(\bar{x})$ gilt also

$$x_{k+1} \;=\; \bar{x} + \frac{\Phi^{(p)}(\xi)}{p!}(x_k - \bar{x})^p$$

Aus $\Phi^{(p)}(\bar{x}) \neq 0$ und wegen der Stetigkeit von $\Phi^{(p)}(x)$ existiert eine δ–Umgebung $U_\delta(\bar{x})$ von $\bar{x}$ mit $\Phi^{(p)}(\xi) \neq 0$ für $\xi \in U_\delta(\bar{x})$. Daraus folgt die Behauptung. ∎

Beispiel 4.5 *Für die Rekursionsvorschrift*

$$\Phi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

und den Fixpunkt $\bar{x} = \sqrt{a}$ ist

$$\Phi'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)$$

und somit $\Phi'(\bar{x}) = 0$, aber

$$\Phi''(x) = \frac{1}{4}\frac{a}{x^3}, \quad \Phi''(\bar{x}) = \frac{1}{4\sqrt{a}} \neq 0\,.$$

Mit $p = 2$ ergibt sich also die quadratische Konvergenz des Babylonischen Wurzelziehens.

4.4 Sekantenmethode und Newton–Verfahren

Wesentlich für die bisherigen Betrachtungen war die Herleitung der Fixpunktgleichung $x = \Phi(x)$ bzw. die daraus resultierende Rekursionsvorschrift

$$x_{k+1} = \Phi(x_k)\,.$$

Beim Babylonischen Wurzelziehen konnte diese durch einfache Transformationen aus der zu lösenden Gleichung gewonnen werden. Im folgenden soll ein allgemeiner Ansatz zur Herleitung von $\Phi(x)$ verfolgt werden.
Gesucht ist $\bar{x} \in D = [a, b]$ als Lösung der nichtlinearen Gleichung $f(\bar{x}) = 0$. Ist $f(x)$ stetig differenzierbar, so liefert die Taylor–Entwicklung in $\bar{x}$

$$0 = f(\bar{x}) = f(x) + (\bar{x} - x)\, f'(\xi)$$

mit einer Zwischenwertstelle $\xi \in (x, \bar{x}) \cup (\bar{x}, x)$. Daraus folgt

$$\bar{x} = x - \frac{f(x)}{f'(\xi)},$$

wobei $f'(\xi)$ geeignet zu approximieren bleibt. Die Wahl

$$f'(\xi) \approx \frac{f(b) - f(a)}{b - a}$$

führt dann auf die **Sehnen–Methode**

$$x_{k+1} = x_k - \frac{b - a}{f(b) - f(a)}\, f(x_k) = \Phi(x_k)\,.$$

Für zwei schon berechnete Näherungslösungen x_k und x_{k-1} ergibt die alternative Approximation

$$f'(\xi) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

die **Sekantenmethode**

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) = \Phi(x_k)\,.$$

Wird anstelle der vorherigen Näherungslösung x_{k-1} eine Näherungslösung x_ℓ mit $f(x_k)f(x_\ell) < 0$ gewählt, so ergibt sich die modifizierte Sekantenmethode (**Regula Falsi**)

$$x_{k+1} = x_k - \frac{x_k - x_\ell}{f(x_k) - f(x_\ell)} f(x_k) = \Phi(x_k).$$

Gilt im Fixpunkt $\bar{x} = f(\bar{x})$ für die erste Ableitung $f'(\bar{x}) \neq 0$, so kann in einer Umgebung von $\bar{x}$ die Approximation $f(\xi) \approx f'(x_k)$ gewählt werden. Daraus folgt das **Newton–Verfahren**

$$x_{k+1} \;=\; x_k - \frac{f(x_k)}{f'(x_k)} \;=\; \Phi(x_k).$$

Für die Bildungsvorschrift

$$\Phi(x) \;=\; x - \frac{f(x)}{f'(x)}$$

mit $f'(x) \neq 0$ ist

$$\Phi'(x) \;=\; 1 - \frac{f'(x)}{f'(x)} + \frac{f(x)}{[f'(x)]^2}\, f''(x) \;=\; \frac{f(x)f''(x)}{[f'(x)]^2}$$

und wegen $f(\bar{x}) = 0$ gilt im Fixpunkt $\Phi'(\bar{x}) = 0$. Weiterhin ist

$$\Phi''(x) \;=\; \frac{f''(x)[f'(x)]^2 - 2f(x)[f''(x)]^2 + f(x)f'(x)f^{(3)}(x)}{[f'(x)]^3}$$

und somit

$$\Phi''(\bar{x}) \;=\; \frac{f''(\bar{x})}{f'(\bar{x})}\,.$$

Für $f''(\bar{x}) \neq 0$ folgt somit $\Phi''(\bar{x}) \neq 0$ und mit $p = 2$ ergibt sich ein quadratisches Konvergenzverhalten des Newton–Verfahrens in einer Umgebung des Fixpunktes $\bar{x}$.

Am Beispiel des Sekantenverfahrens soll nun gezeigt werden, daß für die Konvergenzordnung p nicht nur natürliche Zahlen in Frage kommen.
Für die Iterationsvorschrift

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\, f(x_k)$$

ergibt sich für den Fehler

$$\begin{aligned} e_{k+1} \;&=\; x_{k+1} - \bar{x} \;=\; x_k - \bar{x} - \frac{x_k - \bar{x} + \bar{x} - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) \\ &=\; e_k - \frac{e_k - e_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) \;=\; \frac{e_{k-1} f(x_k) - e_k f(x_{k-1})}{f(x_k) - f(x_{k-1})} \end{aligned}$$

und somit

$$\frac{e_{k+1}}{e_k e_{k-1}} \;=\; \frac{1}{f(x_k) - f(x_{k-1})} \left[\frac{f(x_k)}{x_k - \bar{x}} - \frac{f(x_{k-1})}{x_{k-1} - \bar{x}} \right] \;=\; \frac{g(x_k) - g(x_{k-1})}{f(x_k) - f(x_{k-1})}$$

mit der Funktion

$$g(x) = \frac{f(x)}{x - \bar{x}}.$$

Mit dem Mittelwertsatz der Integralrechnung gilt

$$\begin{aligned} g(x_k) - g(x_{k-1}) \;&=\; \int\limits_{x_{k-1}}^{x_k} g'(x)dx \;=\; g'(\eta)(x_k - x_{k-1}), \\ f(x_k) - f(x_{k-1}) \;&=\; \int\limits_{x_{k-1}}^{x_k} f'(x)dx \;=\; f'(\xi)(x_k - x_{k-1}) \end{aligned}$$

mit Zwischenwertstellen $\eta, \xi \in (x_{k-1}, x_k)$ und somit folgt

$$\frac{e_{k+1}}{e_k e_{k-1}} \;=\; \frac{g'(\eta)}{f'(\xi)}\,.$$

Die Ableitung von $g(x)$ ist

$$g'(x) \;=\; \frac{f'(x)}{x-\bar{x}} - \frac{f(x)}{(x-\bar{x})^2} \;=\; \frac{f'(x)(x-\bar{x})-f(x)}{(x-\bar{x})^2}$$

und damit

$$g'(\eta) \;=\; \frac{f'(\eta)(\eta-\bar{x})-f(\eta)}{(\eta-\bar{x})^2}.$$

Aus der Taylor–Entwicklung

$$0 \;=\; f(\bar{x}) = f(\eta) + f'(\eta)(\bar{x}-\eta) + \frac{1}{2}f''(\bar{\xi})(\bar{x}-\eta)^2$$

mit einer Zwischenwertstelle $\bar{\xi}$ folgt dann

$$g'(\eta) \;=\; \frac{1}{2}f''(\bar{\xi})$$

und somit

$$\frac{e_{k+1}}{e_k e_{k-1}} \;\le\; \frac{1}{2}\,\frac{f''(\bar{\xi})}{f'(\xi)} \;\le\; M$$

in einer Umgebung des Fixpunktes $\bar{x}$. Es gilt also

$$e_{k+1} \;\le\; M\, e_k\, e_{k-1}$$

bzw.

$$\bar{e}_{k+1} \;\le\; \bar{e}_k\, \bar{e}_{k-1}$$

mit $\bar{e}_\ell = M\, e_\ell$. Seien Anfangsnäherungen x_0 und x_1 gegeben, so daß

$$K \;=\; \max\{\bar{e}_0, \sqrt[q]{\bar{e}_1}\} \;<\; 1$$

mit $q \in I\!R$ gilt. Dann ist

$$\bar{e}_0 \;\le\; K \;=\; K^{q^0}, \quad \bar{e}_1 \;\le\; K^q \;=\; K^{q^1}$$

und durch vollständige Induktion folgt

$$\bar{e}_{k+1} \;\le\; \bar{e}_k\, \bar{e}_{k-1} \;\le\; K^{q^k}\, K^{q^{k-1}} \;=\; K^{q^{k-1}(q+1)} \;=\; K^{q^{k+1}},$$

falls $q+1 = q^2$ erfüllt ist,

$$q_{1/2} \;=\; \frac{1\pm\sqrt{5}}{2}.$$

Damit gilt

$$|x_k - \bar{x}| \;\le\; \frac{K^{q^k}}{M}, \quad q = \frac{1+\sqrt{5}}{2} \;\approx\; 1.618.$$

4.5 Nichtlineare Gleichungssysteme

Gesucht ist der Vektor $\underline{\bar{x}} \in I\!R^n$ als Lösung von n nichtlinearen Gleichungen $F(\underline{\bar{x}}) = \underline{0}$,

$$\begin{aligned} f_1(\bar{x}_1,\ldots,\bar{x}_n) &= 0, \\ \vdots \qquad & \quad \vdots \\ f_n(\bar{x}_1,\ldots,\bar{x}_n) &= 0. \end{aligned}$$

Die Taylor–Entwicklung der Funktionen $f_i(\underline{\bar{x}})$ um eine gegebene Näherungslösung $\underline{x}^k$ führt bei Vernachlässigung des jeweiligen Restgliedes auf

$$0 = f_i(\underline{\bar{x}}) \approx f_i(\underline{x}^k) + \sum_{j=1}^n (\bar{x}_j - x_j^k)\frac{\partial}{\partial x_j} f_i(\underline{x}^k) \quad \text{für } i = 1,\ldots,n.$$

Damit kann eine neue Näherungslösung $\underline{x}^{k+1}$ aus dem linearen Gleichungssystem

$$\sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i(\underline{x}^k)(x_j^{k+1} - x_j^k) \;=\; -f_i(\underline{x}^k) \quad \text{für } i = 1, \ldots, n$$

bestimmt werden. Dies ergibt das **Newton–Verfahren** im $\mathbb{R}^n$,

$$\underline{x}^{k+1} \;=\; \underline{x}^k - \Big(\frac{\partial F}{\partial \underline{x}}(\underline{x}^k)\Big)^{-1} F(\underline{x}^k)\,.$$

Die Konvergenz folgt analog zum eindimensionalen Fall aus dem Banachschen Fixpunktsatz, d.h. dem Nachweis der Kontraktion in einer geeignet gewählten Umgebung des Lösungsvektors, siehe zum Beispiel [5].

Übungsaufgaben

4.1. Zu bestimmen ist die Nullstelle $\bar{x} \in [1,2]$ der Funktion $f(x) = x^3 - x - 1$. Man bestimme die Anzahl der notwendigen Iterationsschritte des Bisektionsverfahrens, um eine relative Genauigkeit von $\varepsilon = 10^{-2}$ zu erreichen. Man berechne diese Näherungslösung.

4.2. Gegeben sei die Iterationsvorschrift

$$x_{k+1} \;=\; \frac{1}{3}\left[2x_k + \frac{a}{x_k^2}\right] \quad \text{für } k = 0, 1, 2, \ldots, \quad w_0 \;=\; a.$$

a) Für welche $a \in \mathbb{R}$ mit $a \geq 1$ ist die Iterationsvorschrift eine Kontraktion?

b) Wieviele Iterationen sind notwendig, um $\sqrt[3]{a}$ mit einer relativen Genauigkeit von ε zu berechnen?

4.3. Zur Berechnung der Nullstelle $\bar{x} > 0$ von $f(x) = (\sinh x)^5 - 3$ führe man einen Schritt des Newton–Verfahrens mit $x_0 = 3 > \bar{x}$ durch. Weiter berechne man $f''(x)$ und beweise mittels vollständiger Induktion, daß $x_k > \bar{x}$ für alle $k \in \mathbb{N}$ gilt. Man zeige außerdem, daß die Folge $\{x_k\}_{k\in\mathbb{N}}$ streng monoton fallend ist.

4.4. Für die näherungsweise Berechnung der Nullstellen von

$$f(x) \;=\; \frac{2x^2 - 3x - 2}{x - 1}$$

untersuche man die Fixpunktiterationen $x_{k+1} = \Phi_i(x_k)$ mit

$$\Phi_1(x) \;=\; \frac{3x^2 - 4x - 2}{x - 1}, \quad \Phi_2(x) \;=\; x - 2 + \frac{x}{x - 1}$$

auf ihre Konvergenz und bestimme gegebenenfalls die Konvergenzordnung.

4.5. Man verwende ein geeignetes Näherungsverfahren zur Berechnung aller Nullstellen von

$$f(x) \;=\; x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}.$$

4.6. Zur Nullstellenbrechnung einer dreimal stetig differenzierbaren Funktion kombiniere man die Taylor–Entwicklungen erster und zweiter Ordnung zu einem verbesserten Newton–Verfahren. Dabei werden die jeweiligen Restglieder vernachlässigt. Zur Berechnung von $\sqrt{a}$ gebe man eine geeignete Fehlerabschätzung an. Anschließend führe man drei Iterationsschritte für $x_0 = a = 4$ durch.

Kapitel 5

Gewöhnliche Differentialgleichungen

Betrachtet werde das **Anfangswertproblem** (Cauchy–Problem) zur Bestimmung einer im Intervall $I = [a, b]$ stetig differenzierbaren Funktion $y \in C^1(I)$,

$$y'(x) \;=\; f(x, y(x)) \quad \text{für } x \in I, \quad y(x_0) \;=\; y_0 \quad \text{für } x_0 \in I.$$

Hierbei ist $f : I \times (-\infty, +\infty) \to I\!R$ eine gegebene skalare Funktion. Die Integration der Differentialgleichung ergibt unter Berücksichtigung der Anfangsbedingung $y(x_0) = y_0$ die **Integralgleichung**

$$y(x) \;=\; y(x_0) + \int\limits_{x_0}^{x} f(s, y(x))ds \quad \text{für } x \in I.$$

Daraus ableitbar ist ein analytisches Näherungsverfahren, die **Methode der sukzessiven Iteration (Picard–Iteration)**

$$\begin{array}{rcl} y_0(x) & = & y_0, \\ y_{n+1}(x) & = & y_0 + \displaystyle\int\limits_{x_0}^{x} f(s, y_n(s))ds \quad \text{für } x \in I,\; n = 0, 1, 2, \ldots. \end{array}$$

Hinreichend für die Konvergenz und damit für die eindeutige Lösbarkeit der Integralgleichung ist die **Lipschitz–Stetigkeit** von $f(x, y)$ bezüglich dem zweiten Argument y,

$$|f(x, y_1) - f(x, y_2)| \;\leq\; L\,|y_1 - y_2| \quad \text{für alle } (x, y_i) \in I \times (-\infty, +\infty),\; i = 1, 2.$$

In diesem Kapitel sollen numerische Verfahren zur näherungsweisen Bestimmung der Lösung $y(x)$ des Anfangswertproblems im Intervall $I = [a, b]$ hergeleitet und analysiert werden. Mit

$$x_k = a + kh \quad \text{für } k = 0, \ldots, n, \quad h = \frac{b-a}{n}$$

seien $n+1$ **gleichmäßig** verteilte **Stützstellen** mit der **Schrittweite** h gegeben. In den Stützstellen x_k sei $y(x_k)$ der exakte Lösungswert und y_k eine zu bestimmende Näherungslösung. Für $k = 0$ stimmt diese offensichtlich mit der Anfangsbedingung $y(x_0) = y_0$ überein. Zu bestimmen sind also Näherungswerte y_k für $k \geq 1$. Bei **Einschrittverfahren** ergibt sich die neue Näherungslösung y_{k+1} allein aus der Kenntnis des vorherigen Näherungswertes y_k. Im Gegensatz hierzu fließen bei **Mehrschrittverfahren** mehrere bekannte Näherungswerte in die Berechnung der neuen Näherung ein.

5.1 Einschrittverfahren

Für $k = 0, \ldots, N-1$ wird zunächst die Lösung $y(x_k)$ in der Stützstelle x_k als bekannt vorausgesetzt. Die Integration der Differentialgleichung $y'(x) = f(x, y(x))$ über $x \in [x_k, x_{k+1}]$ ergibt dann

$$y(x_{k+1}) \;=\; y(x_k) + \int\limits_{x_k}^{x_{k+1}} f(x, y(x)dx\,.$$

Die Herleitung eines Näherungsverfahrens beruht nun auf einer geeigneten **numerischen Approximation** des Integrals. Mit der (linksseitigen) **Rechteckregel** folgt

$$\widetilde{y}_{k+1} \;=\; y(x_k) + h\, f(x_k, y(x_k))$$

und Ersetzen des unbekannten Lösungswertes $y(x_k)$ durch den bereits berechneten Näherungswert y_k ergibt das vorwärtige **Eulersche Polygonzugverfahren**

$$y_{k+1} \;=\; y_k + h\, f(x_k, y_k) \quad \text{für } k = 0, 1, \ldots, n-1.$$

Da aus der Kenntnis der vorherigen Näherungslösung y_k die neue Näherungslösung y_{k+1} allein durch Auswertung der rechten Seite $y_k + h\, f(x_k, y_k)$ gewonnen werden kann, liegt ein **explizites** Verfahren vor. Im Gegensatz hierzu führt die Verwendung der rechtsseitigen Rechteckregel,

$$\widetilde{y}_{k+1} \;=\; y(x_k) + h\, f(x_{k+1}, y(x_{k+1})),$$

auf das **implizite** Polygonzugverfahren

$$y_{k+1} \;=\; y_k + h\, f(x_{k+1}, y_{k+1}) \quad \text{für } k = 0, 1, \ldots, n-1,$$

welches zur Bestimmung von y_{k+1} im allgemeinen die Lösung einer nichtlinearen Gleichung erfordert.
Wird für die numerische Integration die **Trapezregel** verwendet,

$$\widetilde{y}_{k+1} \;=\; y(x_k) + \frac{1}{2} h \left[f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1})) \right],$$

so ergibt sich mit

$$y_{k+1} \;=\; y_k + \frac{1}{2} h \left[f(x_k, y_k) + f(x_{k+1}, y_{k+1}) \right] \quad \text{für } k = 0, 1, \ldots, n-1$$

ein implizites **Crank–Nicolson–Verfahren**.
Das Einsetzen des vorwärtigen Eulerschen Polygonzugverfahrens in das Crank–Nicolson–Verfahren liefert dann die Rekursionsvorschrift der **Heunschen Methode**,

$$y_{k+1} \;=\; y_k + \frac{1}{2} h \left[f(x_k, y_k) + f(x_{k+1}, y_k + h f(x_k, y_k)) \right] \quad \text{für } k = 0, 1, \ldots, n-1.$$

Allgemein sind explizite Einschrittverfahren von der Form

$$y_{k+1} \;=\; y_k + h\, \Phi(x_k, y_k, h) \quad \text{für } k = 0, 1, \ldots, n-1$$

und somit gilt

$$\frac{y_{k+1} - y_k}{h} \;=\; \Phi(x_k, y_k, h) \quad \text{für } k = 0, 1, \ldots, n-1$$

sowie

$$\widetilde{y}_{k+1} \;=\; y(x_k) + h\, \Phi(x_k, y(x_k), h) \quad \text{für } k = 0, 1, \ldots, n-1.$$

Das numerische Näherungsverfahren heißt **konsistent** von der **Ordnung** p, falls

$$\left| \frac{y(x_{k+1}) - y(x_k)}{h} - \Phi(x_k, y(x_k), h) \right| \;\le\; c_K\, h^p$$

mit einer Konstanten c_K gilt.

Der Fehler der berechneten Näherungslösung y_{k+1} setzt sich zusammen aus dem Fehler der numerischen Integration und dem fehlerbehafteten Übergang von den unbekannten Lösungswerten $y(x_k)$ zu den bereits berechneten Näherungswerten y_k,

$$e_{k+1} \;=\; |y(x_{k+1}) - y_{k+1}| \;\leq\; |y(x_{k+1}) - \widetilde{y}_{k+1}| + |\widetilde{y}_{k+1} - y_{k+1}|\,.$$

Für den **Integrationsfehler** ergibt sich aus der Abschätzung des Konsistenzfehlers

$$\begin{aligned}
|y(x_{k+1}) - \widetilde{y}_{k+1}| \;&=\; \left| \int\limits_{x_k}^{x_{K+1}} f(x,y(x))dx - h\,\Phi(x_k,y(x_k),h) \right| \\
&=\; \left| \int\limits_{x_k}^{x_{k+1}} y'(x)dx - h\,\Phi(x_k,y(x_k),h) \right| \\
&=\; |y(x_{k+1}) - y(x_k) - h\,\Phi(x_k,y(x_k),h)| \\
&=\; h\left| \frac{y(x_{k+1}) - y(x_k)}{h} - \Phi(x_k,y(x_k),h) \right| \;\leq\; c_K\,h^{p+1}\,.
\end{aligned}$$

Die Lipschitz–Stetigkeit von $f(x,y)$ bezüglich dem zweiten Argument y induziere die Lipschitz–Stetigkeit von $\Phi(x,y,h)$ bezüglich y, d.h. es gelte

$$|\Phi(x,y_1,h) - \Phi(x,y_2,h)| \;\leq\; c_L\,|y_1 - y_2| \quad \text{für alle } (x,y_i) \in I \times (-\infty,+\infty),\; i = 1,2.$$

Daraus ergibt sich für den durch die Störung der Eingangsdaten hervorgerufenen Fehler die **Stabilitätsabschätzung**

$$\begin{aligned}
|\widetilde{y}_{k+1} - y_{k+1}| \;&=\; |y(x_k) - y_k + h\,\Phi(x_k,y(x_k),h) - h\,\Phi(x_k,y_k,h)| \\
&\leq\; |y(x_k) - y_k| + h\,|\Phi(x_k,y(x_k),h) - \Phi(x_k,y_k,h)| \\
&\leq\; (1 + c_L\,h)\,|y(x_k) - y_k|\,.
\end{aligned}$$

Insgesamt gilt also

$$e_{k+1} \;\leq\; (1 + c_L\,h)\,e_k + c_K\,h^{p+1} \quad \text{für } k = 0,1,\ldots,n-1$$

mit

$$e_0 \;=\; |y(x_0) - y_0| \;=\; 0\,.$$

Für $k = 0$ gilt also

$$e_1 \;\leq\; c_K\,h^{p+1}\,.$$

Durch vollständige Induktion folgt dann

$$e_{k+1} \;\leq\; \left(\sum_{i=0}^{k} (1 + c_L h)^i \right) c_K\,h^{p+1} \;=\; \frac{1 - (1 + c_L h)^{k+1}}{1 - (1 + c_L h)}\,c_K\,h^{p+1} \;\leq\; \frac{c_K}{c_L}\,(1 + c_L h)^{k+1}\,h^p$$

für $k = 0,\ldots,n-1$. Insbesondere für $k = n-1$ ergibt sich mit $h = (b-a)/n$ die Fehlerabschätzung

$$|y(b) - y_n| \;\leq\; \frac{c_K}{c_L}\,\left(1 + \frac{c_L(b-a)}{n} \right)^n\,h^p \;\leq\; \frac{c_K}{c_L}\,e^{c_L(b-a)}\,h^p\,.$$

Für ein konsistentes Näherungsverfahren ist also die Stabilität hinreichend für die Konvergenz des Verfahrens, wobei die Ordnungen von Konsistenz– und Konvergenzabschätzung übereinstimmen.

Beispiel 5.1 *Für eine zweimal stetig differenzierbare Lösung $y(x)$ liefert die Taylor–Entwicklung*

$$\begin{aligned} y(x_{k+1}) \;=\; y(x_k+h) &= y(x_k) + h\,y'(x_k) + \frac{1}{2}\,h^2\,y''(\xi_k) \\ &= y(x_k) + h\,f(x_k, y(x_k)) + \frac{1}{2}\,h^2\,y''(\xi_k) \end{aligned}$$

mit einer Zwischenwertstelle $\xi_k \in (x_k, x_{k+1})$.
Andererseits ist für das vorwärtige Eulersche Polygonzugverfahren $\Phi(x_k, y_k, h) = f(x_k, y_k)$. Damit folgt für den Konsistenzfehler

$$\left| \frac{y(x_{k+1}) - y(x_k)}{h} - \Phi(x_k, y(x_k), h) \right| = \left| \frac{1}{2}\,h\,y''(\xi_k) \right| \le c_K\,h$$

mit

$$c_K = \frac{1}{2} \max_{\xi \in (a,b)} |y''(\xi)|$$

und somit die Konsistenz– und Konvergenzordnung $p = 1$.

Zur Herleitung von Verfahren höherer Konsistenzordnung p wird für die Lösung $y(x)$ der Differentialgleichung $y'(x) = f(x, y(x))$ eine Taylor–Entwicklung höherer Ordnung betrachtet, zum Beispiel ist

$$y(x_{k+1}) = y(x_k) + h\,y'(x_k) + \frac{1}{2}\,h^2 y''(x_k) + \frac{1}{6}\,h^3\,y'''(\xi_k)$$

mit einer Zwischenwertstelle $\xi_k \in (x_k, x_{k+1})$. Einsetzen der Differentialgleichung ergibt

$$y(x_{k+1}) = y(x_k) + h\,f(x_k, y(x_k)) + \frac{1}{2}\,h^2\,\left.\frac{d}{dx} f(x, y(x))\right|_{x=x_k} + \frac{1}{6}\,h^3\,y'''(\xi_k).$$

Mit

$$\begin{aligned} \frac{d}{dx} f(x, y(x)) &= f_x(x, y(x)) + f_y(x, y(x))\,y'(x) \\ &= f_x(x, y(x)) + f_y(x, y(x)) f(x, y(x)) \end{aligned}$$

und

$$\widetilde{y}_{k+1} = y(x_k) + h\,f(x_k, y(x_k)) + \frac{1}{2}\,h^2\,[f_x(x_k, y(x_k)) + f_y(x_k, y(x_k)) f(x_k, y(x_k))]$$

folgt daraus

$$y(x_{k+1}) = \widetilde{y}_{k+1} + \frac{1}{6}\,h^3\,y'''(\xi_k)\,.$$

Für die Abschätzung des Integrationsfehlers ergibt sich

$$|y(x_{k+1}) - \widetilde{y}_{k+1}| \le \frac{1}{6}\,h^3 \max_{\xi \in (x_k, x_{k+1})} |y'''(\xi)|\,.$$

Die Berechnung von $\widetilde{y}_{k+1}$ verlangt die Auswertung der partiellen Ableitungen $f_x(x_k, y(x_k))$ und $f_y(x_k, y(x_k))$. Durch den Ansatz

$$\widetilde{y}_{k+1} = y(x_k) + h\,\Phi(x_k, y(x_k), h)$$

mit

$$\begin{aligned} \Phi(x_k, y(x_k), h) &= f(x_k, y(x_k)) + \frac{1}{2}\,h\,[f_x(x_k, y(x_k)) + f_y(x_k, y(x_k)) f(x_k, y(x_k))] \\ &= a_1 f(x_k, y(x_k)) + a_2 f(x_k + \alpha, y(x_k) + \beta) + R(h) \end{aligned}$$

und noch zu bestimmenden reellen Parametern a_1, a_2, α, β und einem zusätzlichen Restglied $R(h)$ soll die Auswertung von $\widetilde{y}_{k+1}$ auf eine alleinige Auswertung der gegebenen Funktion $f(x,y)$ zurückgeführt werden. Die Anwendung der **Taylor–Formel** in zwei Veränderlichen,

$$f(x_k+\alpha, y(x_k)+\beta) \;=\; f(x_k,y(x_k)) + \alpha f_x(x_k,y(x_k)) + \beta f_y(x_k,y(x_k)) + \mathcal{O}(\alpha^2+\alpha\beta+\beta^2),$$

ergibt

$$\begin{aligned}
\Phi(x_k,y(x_k),h) &= f(x_k,y(x_k)) + \frac{1}{2}\,h\,[f_x(x_k,y(x_k)) + f_y(x_k,y(x_k))f(x_k,y(x_k))] \\
&= a_1 f(x_k,y(x_k)) + a_2\,[f(x_k,y(x_k)) + \alpha f_x(x_k,y(x_k)) + \beta f_y(x_k,y(x_k))] \\
&\qquad + a_2\mathcal{O}(\alpha^2+\alpha\beta+\beta^2) + R(h) \\
&= (a_1+a_2) f(x_k,y(x_k)) + a_2\alpha f_x(x_k,y(x_k)) + a_2\beta f_y(x_k,y(x_k)) \\
&\qquad + a_2\mathcal{O}(\alpha^2+\alpha\beta+\beta^2) + R(h)
\end{aligned}$$

und durch Koeffizientenvergleich folgt

$$a_1+a_2 \;=\; 1, \quad a_2\alpha \;=\; \frac{1}{2}\,h, \quad a_2\beta \;=\; \frac{1}{2}\,h\,f(x_k,y(x_k))$$

sowie

$$R(h) \;=\; -a_2\mathcal{O}(\alpha^2+\alpha\beta+\beta^2)\,.$$

Eine mögliche Lösung ist

$$a_1 \;=\; a_2 \;=\; \frac{1}{2}, \quad \alpha \;=\; h, \quad \beta \;=\; h\,f(x_k,y(x_k))$$

und somit

$$\Phi(x_k,y(x_k),h) \;=\; \frac{1}{2}\,f(x_k,y(x_k)) + \frac{1}{2}\,f(x_k+h, y(x_k)+hf(x_k,y(x_k))) + R(h).$$

Für

$$\bar{y}_{k+1} \;=\; y(x_k) + \frac{1}{2}\,h\;[f(x_k,y(x_k)) + f(x_k+h, y(x_k)+hf(x_k,y(x_k)))]$$

folgt dann die Fehlerabschätzung

$$|y(x_{k+1}) - \bar{y}_{k+1}| \;\le\; |y(x_{k+1}) - \widetilde{y}_{k+1}| + |\widetilde{y}_{k+1} - \bar{y}_{k+1}| \;\le\; \frac{1}{6}\,h^3 \max_{\xi\in(x_k,x_{k+1})} |y'''(\xi)| \;+\; h|R(h)|\,.$$

Mit

$$|R(h)| \;=\; |a_2\mathcal{O}(\alpha^2+\alpha\beta+\beta^2)| \;\le\; c\,h^2$$

ergibt sich für die Abschätzung des Integrationsfehlers

$$|y(x_{k+1}) - \bar{y}_{k+1}| \;\le\; c\,h^3,$$

falls die Lösung $y(x)$ dreimal stetig differenzierbar ist. Für die **Konsistenzordnung** folgt also $p=2$.

Werden in der Berechnung von $\bar{y}_{k+1}$ die exakten Lösungswerte $y(x_k)$ durch die vorher berechneten Näherungslösungen y_k ersetzt, so ergibt sich die **Heunsche Methode**

$$y_{k+1} \;=\; y_k + \frac{1}{2}\,h\;[f(x_k,y_k) + f(x_k+h, y_k+hf(x_k,y_k))]$$

und aus der Stabilitätsabschätzung folgt die Fehlerabschätzung

$$|y(b) - y_n| \;\le\; c\,h^2\,.$$

Eine zweite Lösung der Gleichungen des Koeffizientenvergleichs

$$a_1 + a_2 = 1, \quad a_2\alpha = \frac{1}{2}\,h, \quad a_2\beta = \frac{1}{2}\,h\,f(x_k, y(x_k))$$

ist gegeben durch

$$a_1 = 0, \quad a_2 = 1, \quad \alpha = \frac{1}{2}\,h, \quad \beta = \frac{1}{2}\,h\,f(x_k, y(x_k)).$$

Daraus ergibt sich das **Mittelpunktverfahren** (modifiziertes Euler–Verfahren)

$$y_{k+1} = y_k + h\,f(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hf(x_k, y_k))$$

wiederum als Verfahren 2. Ordnung.
Allgemein können Einschrittverfahren durch Integration der Differentialgleichung,

$$y(x_{k+1}) = y(x_k) + \int\limits_{x_k}^{x_{k+1}} f(x, y(x))dx,$$

und Anwendung einer numerischen Integrationsformel

$$\widetilde{y}_{k+1} = y(x_k) + h\sum_{i=1}^{m} c_i f(x_k + a_i h, y(x_k + a_i h))$$

und anschließende Ersetzung der Lösungswerte $y(x_k)$ durch die bereits berechneten Näherungswerte y_k hergeleitet werden.
Die Bestimmung der Lösung $y(x)$ in den Integrationspunkten $x_k + a_i h$ erfolgt rekursiv aus

$$\begin{aligned} y(x_k + a_i h) &= y(x_k) + \int\limits_{x_k}^{x_k + a_i h} f(s, y(s))ds \\ &\approx y(x_k) + \sum_{j=1}^{i-1} b_{ij} f(x_k + a_j h, y(x_k + a_j h))\,. \end{aligned}$$

Für die neue Näherungslösung y_{k+1} ergibt sich dann die Rekursionsvorschrift

$$y_{k+1} = y_k + h\sum_{i=1}^{m} c_i k_i$$

mit

$$k_i = f(x_k + a_i h, y_k + h\sum_{j=1}^{i-1} b_{ij} k_j) \quad \text{für } i = 1, \ldots, m.$$

Die Parameter a_i, b_{ij} und c_i ergeben sich wie bei der Herleitung der Heunschen Methode durch eine entsprechende Taylor–Entwicklung der Lösung $y(x)$ und Koeffizientenvergleich. Die resultierenden Verfahren sind die **Runge–Kutta–Verfahren**, welches zum Beispiel für $m = 4$ durch

$$\begin{aligned} k_1 &= f(x_k, y_k) \\ k_2 &= f(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_1) \\ k_3 &= f(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_2) \\ k_4 &= f(x_k + h, y_k + hk_3), \\ y_{k+1} &= y_k + \frac{1}{6}\,h\,[k_1 + 2k_2 + 2k_3 + k_4] \end{aligned}$$

ein Verfahren 4. Ordnung beschreibt.

Das alles sind recht mühselige Rechnungen und die Freude des scharfsinnigen Kopfes an sich selbst ist manchmal die alleinige Ursache dessen, dass man weiterrechnet. (F. Kafka)

5.2 Mehrschrifttverfahren

Die Herleitung von Einschrittverfahren zur näherungsweisen Lösung des Anfangswertproblems

$$y'(x) \;=\; f(x, y(x)) \quad \text{für } x \in I, \quad y(x_0) \;=\; y_0 \quad \text{für } x_0 \in I$$

basiert einerseits auf einer Approximation der Ableitung durch Differenzenquotienten,

$$y'(x_k) \;\approx\; \frac{y(x_{k+1}) - y(x_k)}{h} \;=\; f(x_k, y(x_k)),$$

kann aber auch andererseits durch numerische Integration von

$$y(x_{k+1}) \;=\; y(x_k) + \int\limits_{x_k}^{x_{k+1}} f(x, y(x))dx \;\approx\; y(x_k) + h\, f(x_k, y(x_k))$$

begründet werden. In beiden Fällen ergibt sich das Eulersche Polygonzugverfahren

$$y_{k+1} = y_k + h\, f(x_k, y_k).$$

Bei Einschrittverfahren erfolgt die Berechnung der Näherungslösung y_{k+1} in x_{k+1} also nur aus der Kenntnis der Lösung y_k in x_k. Werden mehrere Paare (x_ℓ, y_ℓ) zur Berechnung von y_{k+1} verwendet, so führt dies auf **Mehrschrittverfahren**. Analog zu Einschrittverfahren können diese sowohl durch eine Approximiation des Differenzenquotienten als auch durch eine numerische Integration hergeleitet werden. Hier sollen vor allem **Zweischrittverfahren** behandelt werden, auf eine Diskussion allgemeiner Mehrschrittverfahren soll hier verzichtet werden.

Die Verwendung des zentralen Differenzenquotienten

$$y'(x_k) \;\approx\; \frac{y(x_{k+1}) - y(x_{k-1})}{2h} \;=\; f(x_k, y(x_k))$$

bzw. die numerische Integration von

$$y(x_{k+1}) = y(x_{k-1}) + \int\limits_{x_{k-1}}^{x_{k+1}} f(x, y(x))dx$$

durch die Mittelpunktregel ergibt mit

$$y_{k+1} \;=\; y_{k-1} + 2\, h\, f(x_k, y_k)$$

ein erstes Zweischrittverfahren. Neben der durch die Anfangsbedingung $y(x_0) = y_0$ gegebenen Startnäherung muß die Näherung y_1 durch ein geeignetes Einschrittverfahren bestimmt werden.

Die Taylorentwicklung von $g(x) = f(x, y(x))$ liefert

$$f(x, y(x)) \;=\; g(x) \;=\; g(x_k) + (x - x_k)g'(x_k) + \frac{1}{2}\, g''(\xi)\, (x - x_k)^2$$

für $x \in (x_{k-1}, x_{k+1})$ mit einer Zwischenwertstelle $\xi \in (x_{k-1}, x_{k+1})$ und somit

$$\begin{aligned}
\int\limits_{x_{k-1}}^{x_{k+1}} f(x, y(x))dx \;&=\; \int\limits_{x_{k-1}}^{x_{k+1}} [g(x_k) + (x - x_k)g'(x_k) + \frac{1}{2} g''(\xi)(x - x_k)^2]dx \\
&=\; 2h\, g(x_k) + g'(x_k) \int\limits_{x_{k-1}}^{x_{k+1}} (x - x_k)dx + \frac{1}{2} \int\limits_{x_{k-1}}^{x_{k+1}} g''(x)(x - x_k)^2 dx \\
&=\; 2h\, f(x_k, y(x_k)) + \frac{1}{2} \int\limits_{x_{k-1}}^{x_{k+1}} g''(x)(x - x_k)^2 dx.
\end{aligned}$$

Daraus ergibt sich die Fehlerabschätzung

$$\left| \int_{x_{k-1}}^{x_{k+1}} f(x,y(x))dx - 2h\, f(x_k, y(x_k)) \right| = \left| \frac{1}{2} \int_{x_{k-1}}^{x_{k+1}} g''(x)(x-x_k)^2 dx \right| \le \frac{1}{3} h^3 \max_{\xi \in (x_{k-1}, x_{k+1})} |g''(\xi)|$$

und somit die Konsistenzordnung $p = 2$.

Die Approximation der Ableitung $y'(x_k)$ durch **Konvexkombinationen** des vorwärtigen und rückwärtigen Differenzenquotienten liefert

$$y'(x_k) \approx \alpha \frac{y(x_{k+1}) - y(x_k)}{h} + (1-\alpha) \frac{y(x_k) - y(x_{k-1})}{h}$$

für beliebige Parameter $\alpha \in I\!R$. Dieses Vorgehen führt auf die **BDF–Verfahren** (Backward Difference Formula). Für $\alpha = 3/2$ ergibt sich zum Beispiel

$$3y_{k+1} = 4y_k - y_{k-1} + 2\,h\,f(x_k, y_k),$$

während für $\alpha = -1/2$

$$y_{k+1} = 4y_k - 3y_{k-1} - 2\,h\,f(x_k, y_k)$$

folgt.

Beispiel 5.2 *Für die näherungsweise Lösung des Anfangswertproblems $y'(x) = 0, y(0) = 0$ wird das konsistente Zweischrittverfahren*

$$y_{k+1} = 4y_k - 3y_{k-1}$$

mit den Startwerten $y_0 = 0$ und einem gestörten Näherungswert $y_1 = \varepsilon \neq 0$ betrachtet. Dann folgt

$$y_2 = 4\varepsilon = \frac{1}{2}(3^2 - 1)\,\varepsilon$$

und allgemein gilt

$$y_k = \frac{1}{2}(3^k - 1)\,\varepsilon.$$

Dies folgt durch vollständige Induktion,

$$\begin{aligned} y_{k+1} &= 4y_k - 3y_{k-1} = 4\,\frac{1}{2}(3^k - 1)\,\varepsilon - 3\,\frac{1}{2}(3^{k-1} - 1)\,\varepsilon \\ &= \frac{1}{2}[4 \cdot 3^k - 4 - 3 \cdot 3^{k-1} + 3]\,\varepsilon = \frac{1}{2}[4 \cdot 3^k - 3^k - 1]\,\varepsilon = \frac{1}{2}(3^{k+1} - 1)\,\varepsilon. \end{aligned}$$

Damit folgt $y_k \to \infty$ für $k \to \infty$ und somit ergibt sich keine Konvergenz des Näherungsverfahrens. Ursache hierfür ist die fehlende Stabilität des Verfahrens.

Zur Bestimmung der Lösung $y_k = y(x_k)$ kann auch die Differenzengleichung

$$y_{k+1} - 4y_k + 3y_{k-1} = 0$$

betrachtet werden. Der Ansatz $y_k = \lambda^k$ ergibt

$$0 = \lambda^{k+1} - 4\lambda^k + 3\lambda^{k-1} = \lambda^{k-1}[\lambda^2 - 4\lambda + 3]$$

und somit die charakteristische Gleichung

$$\lambda^2 - 2\lambda + 3 = 0$$

mit den Nullstellen $\lambda_1 = 1$ und $\lambda_2 = 3$. Die allgemeine Lösung der Differenzengleichung ist dann gegeben durch

$$y_k = c_1 + c_2 3^k \quad \text{für } k = 0, 1, 2, \ldots.$$

Aus den Anfangsbedingungen $y_0 = 0$ und $y_1 = \varepsilon$ folgt $c_1 + c_2 = 0$ und $c_1 + 3c_2 = \varepsilon$ und somit $c_1 = -\varepsilon/2$ bzw. $c_2 = \varepsilon/2$. Damit ist wie oben

$$y_k = \frac{1}{2}(3^k - 1)$$

und die Divergenz des Näherungsverfahrens folgt wegen $\lambda_2 = 3 > 1$.

Wie das vorherige Beispiel zeigt, ist für die Stabilität eines Mehrschrittverfahrens die sogenannte **Wurzelbedingung** hinreichend und notwendig, d.h. alle im allgemeinen komplexen Nullstellen der charakteristischen Gleichung des Näherungsverfahrens sind betragsmäßig nicht größer als Eins, mehrfache Wurzeln sind echt kleiner Eins.

Beispiel 5.3 *Für die Rekursionsvorschrift*

$$3y_{k+1} = 4y_k - y_{k-1} + 2hf(x_k, y_k)$$

lautet das charakteristische Polynom

$$\sigma(\lambda) = 3\lambda^2 - 4\lambda + 1 = (3\lambda - 1)(\lambda - 1)$$

mit den Nullstellen $\lambda_1 = 1/3$ und $\lambda_2 = 1$. Damit ist die Wurzelbedingung erfüllt, woraus die Stabilität des Verfahrens folgt.

Als alternativer Zugang zu Mehrschrittverfahren werden im folgenden numerische Integrationsformeln zur Auswertung von

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))dx$$

betrachtet. Die Interpolation von $g(x) = f(x, y(x))$ in den Knoten x_{k-1} und x_k kann mit Lagrange–Polynomen durch

$$g_2(x) = g(x_{k-1})L_{k-1}(x) + g(x_k)L_k(x) = g(x_{k-1})\frac{x - x_k}{x_{k-1} - x_k} + g(x_k)\frac{x - x_{k-1}}{x_k - x_{k-1}}$$

beschrieben werden. Einsetzen des Interpolationspolynoms ergibt

$$\begin{aligned}
\widetilde{y}_{k+1} &= y(x_k) + \int_{x_k}^{x_{k+1}} g_2(x)dx \\
&= y(x_k) + g(x_{k-1})\int_{x_k}^{x_{k+1}} \frac{x - x_k}{x_{k-1} - x_k}\,dx + g(x_k)\int_{x_k}^{x_{k+1}} \frac{x - x_{k-1}}{x_k - x_{k-1}}\,dx \\
&= y(x_k) - \frac{1}{2}\,h\,g(x_{k-1}) + \frac{3}{2}\,h\,g(x_k) \\
&= y(x_k) + \frac{3}{2}\,h\,f(x_k, y(x_k)) - \frac{1}{2}\,h\,f(x_{k-1}, y(x_{k-1})).
\end{aligned}$$

Ersetzen der unbekannten Lösungswerte $y(x_\ell)$ durch die bereits berechneten Näherungslösungen y_ℓ ergibt das explizite **Adams–Bashfort**–Verfahren

$$y_{k+1} = y_k + \frac{3}{2}\,h\,f(x_k, y_k) - \frac{1}{2}\,h\,f(x_{k-1}, y_{k-1}).$$

Im Gegensatz hierzu liefert die Interpolation in den Stützstellen x_k und x_{k+1} mit

$$g_2(x) = g(x_k)L_k(x) + g(x_{k+1})L_{k+1}(x) = g(x_k)\frac{x - x_{k+1}}{x_k - x_{k+1}} + g(x_{k+1})\frac{x - x_k}{x_{k+1} - x_k}$$

das implizite **Adams–Moulton**–Verfahren

$$y_{k+1} = y_k + \frac{1}{2}\,h\,[f(x_k, y_k) + f(x_{k+1}, y_{k+1})].$$

5.3 Stabilität

Die Lösung des Anfangswertproblems

$$y'(x) \;=\; f(x,y(x)) \;=\; \lambda y(x) \quad \text{für } x \in (0,a), \quad y(0) = y_0$$

mit einem beliebigen Parameter $\lambda \in I\!R$ ist gegeben durch

$$y(x) \;=\; y_0 e^{\lambda x}.$$

Die Bestimmung einer Näherungslösung mit dem Eulersches Polygonzugverfahren führt auf die Rekursionsvorschrift

$$y_{k+1} \;=\; y_k + h\, f(x, y_k) \;=\; y_k + h\,\lambda\, y_k \;=\; (1 + h\,\lambda) y_k \;=\; (1 + h\,\lambda)^{k+1} y_0$$

für $k = 0, \ldots, n-1$ mit der Schrittweite $h = a/n$. Für den Fehler der durch das Eulersche Polygonzugverfahren gilt die Abschätzung

$$|y(a) - y_n| \;\le\; \frac{1}{2}\,\frac{C}{L}\,(e^{La} - 1)\,h$$

mit den durch die Lipschitz–Bedingung

$$|f(x,y_1) - f(x,y_2)| \;\le\; L\,|y_1 - y_2|$$

und die Regularität der Lösung

$$C = \max_{x \in [0,a]} |y''(x)|$$

gegebenen Konstanten. Für $f(x,y) = \lambda\, y$ ist

$$|f(x,y_1) - f(x,y_2)| \;=\; |\lambda|\,|y_1 - y_2|$$

und somit $L = |\lambda|$. Für $y(x) = y_0 e^{\lambda x}$ ist $y''(x) = y_0 \lambda^2 e^{\lambda x}$ und insbesondere für $\lambda < 0$ ist

$$C \;=\; |y_0|\,\lambda^2.$$

Damit lautet die Fehlerabschätzung

$$|y(a) - y_n| \;\le\; \frac{1}{2}\,\frac{|y_0|\,\lambda^2}{|\lambda|}\,(e^{|\lambda| a} - 1)\,h \;=\; \frac{1}{2}\,|y_0|\,(e^{|\lambda| a} - 1)\,|\lambda|\,h$$

und somit folgt Konvergenz für $h \to 0$ bei festem λ. Andererseits zeigt die obige Fehlerabschätzung die praktische Forderung

$$|\lambda|\,h \;<\; 1,$$

d.h. mit $h < |\lambda|^{-1}$ muß die Schrittweite h in Abhängigkeit von $|\lambda|$ gewählt werden. Dies gilt entsprechend für komplexe Parameter λ mit $\operatorname{Re}\lambda < 0$. Gesucht sind deshalb Verfahren, die die Konvergenz unabhängig von λ gewährleisten.
Ein numerisches Verfahren heißt ***A*-stabil**, wenn für $\operatorname{Re}\lambda < 0$ die durch das Verfahren erzeugten Näherungslösungen des Anfangswertproblems

$$y'(x) \;=\; \lambda y(x) \quad \text{für } x \in I\!R, \quad y(x_0) \;=\; y_0$$

beschränkt sind.
Betrachtet werden zunächst Einschrittverfahren der allgemeinen Form

$$y_{k+1} \;=\; y_k + h\,\Phi(x_k, y_k, h)\,.$$

Speziell für $f(x,y) = \lambda y$ ergibt sich daraus die Rekursionsvorschrift

$$y_{k+1} \;=\; R(\lambda h) y_k$$

mit der **Stabilitätsfunktion** $R(z)$ und $z = \lambda h$. Dann ist die A–Stabilität äquivalent zu der Forderung

$$|R(z)| \;\le\; 1 \quad \text{für alle } z \text{ mit } \operatorname{Re} z < 0\,.$$

Beispiel 5.4 *Für das Eulersche Polygonzugverfahren*

$$y_{k+1} = y_k + h\,f(x_k, y_k) = y_k + h\,\lambda\,y_k = (1 + h\lambda)y_k = R(h\lambda)y_k$$

mit der Stabilitätsfunktion

$$R(z) = 1 + z\,.$$

Damit ergibt sich die A–Stabilität des Eulerschen Polygonzugverfahrens für

$$|1 + z| \le 1\,.$$

Beispiel 5.5 *Für das Runge–Kutta-Verfahren 4. Ordnung ergibt sich mit*

$$\begin{aligned}
k_1 &= f(x_k, y_k) = \lambda\,y_k \\
k_2 &= f(x_k + \frac{h}{2}, y_k + \frac{h}{2}k_1) = \lambda(y_k + \frac{h}{2}k_1) = (\lambda + \frac{1}{2}\,h\,\lambda^2)y_k \\
k_3 &= f(x_k + \frac{h}{2}, y_k + \frac{h}{2}k_2) = \lambda(y_k + \frac{h}{2}k_2) = (\lambda + \frac{1}{2}\,h\,\lambda^2 + \frac{1}{4}\,h^2\,\lambda^3)y_k \\
k_4 &= f(x_k + h, y_k + hk_3) = \lambda(y_k + hk_3) = (\lambda + h\,\lambda^2 + \frac{1}{2}\,h^2\,\lambda^3 + \frac{1}{4}\,h^3\,\lambda^4)y_k
\end{aligned}$$

für die Berechnung der neuen Näherungslösung

$$\begin{aligned}
y_{k+1} &= y_k + \frac{h}{6}\,[k_1 + 2k_2 + 2k_3 + k_4] \\
&= [1 + (h\lambda) + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4]y_k = R(h\lambda)y_k
\end{aligned}$$

mit der Stabilitätsfunktion

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4.$$

Für die Bestimmung des Stabilitätsgebietes $\{z\,:\,|R(z)| \le 1\}$ *führt der Ansatz*

$$z = -1 + r(\varphi)e^{i\varphi}$$

auf eine nichtlineare Gleichung in r*,*

$$\begin{aligned}
& r^8 + 12r^6\cos 2\varphi + 16r^5\cos 3\varphi + 18r^4(\cos 4\varphi + 2) + 96r^3\cos\varphi \\
& \qquad + 64r^2(1 + 2\cos 2\varphi) + 144r\cos\varphi - 495 = 0\,.
\end{aligned}$$

Durch Anwendung Newton–Verfahrens kann daraus das in Abbildung 5.1 angegebene Stabilitätsgebiet des Runge–Kutta–Verfahrens 4. Ordnung bestimmt werden.

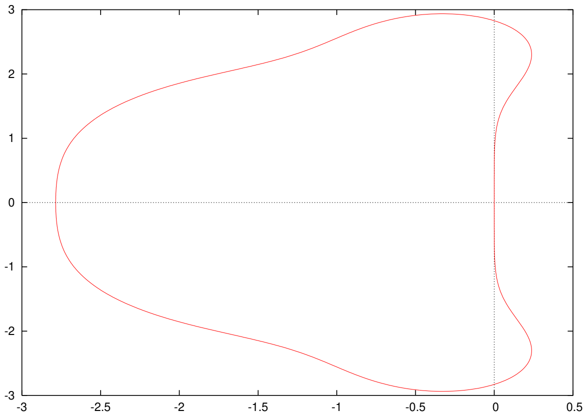

Abbildung 5.1: Stabilitätsgebiet des Runge–Kutta–Verfahrens 4. Ordnung.

Beispiel 5.6 *Betrachtet wird das aus der Trapezregel resultierende implizite Näherungsverfahren*

$$y_{k+1} = y_k + \frac{1}{2} h \left[f(x_k, y_k) + f(x_{k+1}, y_{k+1}) \right].$$

Für $f(x,y) = \lambda y$ ergibt sich

$$y_{k+1} = y_k + \frac{1}{2} \lambda h \left[y_k + y_{k+1} \right]$$

bzw.

$$y_{k+1} = \frac{2 + \lambda h}{2 - \lambda h} y_k = R(\lambda h) y_k$$

mit der Stabilitätsfunktion

$$R(z) = \frac{2+z}{2-z}.$$

Dann gilt $|R(z)| \leq 1$ für $Re\, z \leq 0$ und somit ist dieses Näherungsverfahren A–stabil.

Beispiel 5.7 *Als Beispiel für ein Mehrschrittverfahren wird die Mittelpunktregel*

$$y_{k+1} = y_{k-1} + 2hf(x_k, y_k)$$

betrachtet. Für $f(x,y) = \lambda y$ ergibt dies

$$y_{k+1} = y_{k-1} + 2h\lambda y_k$$

bzw. mit $z = \lambda h$

$$y_{k+1} - 2z y_k - y_{k-1} = 0.$$

Bei Mehrschrittverfahren folgt die Stabilität aus der Wurzelbedingung: Die Nullstellen der charakteristischen Gleichung

$$\lambda^2 - 2z\lambda - 1 = 0$$

sind

$$\lambda_{1/2} = z \pm \sqrt{z^2 + 1}.$$

Für $Re\, z < 0$ folgt $|\lambda_2| > 1$, d.h. die Mittelpunktregel ist **nicht** *A–stabil.*

Es zeigt sich sogar, daß kein explizites Mehrschrittverfahren A–stabil ist. Andererseits besitzen A–stabile implizite Mehrschrittverfahren die maximale Konsistenzordnung $p = 2$.

Abschließend werde ein lineares System $\underline{u}'(x) = A\underline{u}(x)$ gewöhnlicher Differentialgleichungen betrachtet,

$$\begin{aligned} u_1'(x) &= a_{11} u_1(x) + a_{12} u_2(x), \quad u_1(x_0) = u_1^0, \\ u_2'(x) &= a_{21} u_1(x) + a_{22} u_2(x), \quad u_2(x_0) = u_2^0. \end{aligned}$$

Die Eigenwerte und zugehörigen Eigenvektoren der Koeffizienmatrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

seien durch $(\lambda_1, \alpha_1, \beta_1)$ und $(\lambda_2, \alpha_2, \beta_2)$ gegeben. Für verschiedene Eigenwerte $\lambda_1 \neq \lambda_2$ lautet die allgemeine Lösung

$$\underline{u}(x) = C_1 \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} e^{\lambda_1 x} + C_2 \begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} e^{\lambda_2 x},$$

wobei die Konstanten C_1 und C_2 durch die Anfangsbedingung eindeutig bestimmt sind.

Analog zum Eulerschen Polygonzugverfahren führt die Approximation der Ableitungen zu einem numerischen Näherungsverfahren

$$\begin{aligned}\frac{u_1^{k+1}-u_1^k}{h} &= a_{11}u_1^k + a_{12}u_2^k,\\ \frac{u_2^{k+1}-u_2^k}{h} &= a_{21}u_1^k + a_{22}u_2^k,\end{aligned}$$

bzw.

$$\underline{u}^{k+1} = (I+hA)\underline{u}^k = (I+hA)^{k+1}\underline{u}^0 .$$

Aus der Entwicklung des Startvektors $\underline{u}$ in die Eigenvektoren von A,

$$\underline{u}^0 = \gamma_1 \begin{pmatrix}\alpha_1\\ \beta_1\end{pmatrix} + \gamma_2 \begin{pmatrix}\alpha_2\\ \beta_2\end{pmatrix} = \gamma_1\underline{v}^1 + \gamma_2\underline{v}^2,$$

folgt

$$\underline{u}^1 = (I+hA)\underline{u}^0 = (I+hA)(\gamma_1\underline{v}^1+\gamma_2\underline{v}^2) = \gamma_1(1+h\lambda_1)\underline{v}^1 + \gamma_2(1+h\lambda_2)\underline{v}^2$$

und durch rekursive Anwendung ergibt sich

$$\underline{u}^{k+1} = \gamma_1(1+h\lambda_1)^{k+1}\underline{v}^1 + \gamma_2(1+h\lambda_2)^{k+1}\underline{v}^2 .$$

Notwendig für Beschränktheit der Lösung ist also die Forderung

$$|1+h\lambda_i| \le 1 \quad \text{für } i=1,2.$$

Für unterschiedliche Eigenwerte $\lambda_1 \neq \lambda_2$ mit $\mathrm{Re}\,\lambda_i \le 0$ wird die Schrittweite h durch die Bedingung

$$|1+h\lambda_{\max}| \le 1$$

bestimmt. Die zugehörige Lösungskomponente

$$\begin{pmatrix}\alpha_{\max}\\ \beta_{\max}\end{pmatrix} e^{\lambda\max}$$

klingt für $x \to \infty$ schnell ab und leistet zur Darstellung der Lösung einen vernachlässigbaren Beitrag. Im Widerspruch dazu bestimmt $\lambda_{\max}$ die Schrittweite. Differentialgleichungssysteme heißen **steif**, wenn die Lösung abklingende Komponenten mit stark unterschiedlichem Abklingverhalten enthält. Anwendungen dafür finden sich zum Beispiel in der Reaktionskinetik oder bei der Modellierung elektrischer Schaltkreise. Die numerische Lösung steifer Systeme erfordert die Verwendung A–stabiler Verfahren bzw. von Verfahren mit möglichst großem Stabilitätsgebiet.

5.4 Zweipunkt–Randwertprobleme

Zu bestimmen ist die Lösung einer gewöhnlichen Differentialgleichung 2. Ordnung im Intervall (a,b),

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad \text{für } x \in (a,b),$$

wobei in den Randpunkten a und b geeignete Randbedingungen an die Funktion u bzw. an ihre Ableitung u' zu formulieren sind. Bei **Dirichlet**–Randbedingungen erfolgt eine Vorgabe der Funktion in den Randpunkten,

$$u(a) = u_a, \quad u(b) = u_b,$$

während bei einer **Neumann**–Randbedingung die Ableitung vorgeben wird,

$$u'(a) = u_a, \quad u'(b) = u_b .$$

Bei einer **Robin**–Randbedingung ist die Ableitung proportional zur Differenz der gesuchten Funktion zu einem vorgegebenen Vergleichswert,

$$u'(a) \;=\; \alpha[u(a) - u_a], \quad u'(b) \;=\; \beta[u(b) - u_b]\,.$$

Werden in den beiden Randpunkten Randbedingungen unterschiedlichen Typs vorgeschrieben, so spricht man von **gemischten** Randbedingungen.
Der Einfachheit halber wird im folgenden eine Differentialgleichung zweiter Ordnung mit homogenen Dirichlet–Randbedingungen betrachtet,

$$-u''(x) + b(x)u'(x) + c(x)u(x) \;=\; f(x) \quad \text{für } x \in (a,b), \quad u(a) \;=\; u(b) \;=\; 0\,.$$

Bei einer Approximation der Differentialgleichung mit **finiten Differenzen** werden in den **Stützstellen**

$$x_k \;=\; a + k\,h \quad \text{für } k = 0,\ldots,n \quad \text{mit } h \;=\; \frac{b-a}{n}$$

die auftretenden Ableitungen durch vorwärtige, rückwärtige oder zentrale Differenzenquotienten

$$u'(x_k) \;\approx\; \frac{u_{k+1}-u_k}{h}, \quad u'(x_k) \;\approx\; \frac{u_k - u_{k+1}}{h}, \quad u'(x_k) \;\approx\; \frac{u_{k+1}-u_{k-1}}{2h}$$

approximiert. Für die Approximation der zweiten Ableitung ergibt sich durch die Kombination von vorwärtigem und rückwärtigem Differenzenquotienten

$$u''(x_k) \;=\; \frac{u'(x_{k+1}) - u'(x_k)}{h} \;=\; \frac{\dfrac{u(x_{k+1}) - u(x_k)}{h} - \dfrac{u_k(x_k) - u_{k-1}}{h}}{h} \;=\; \frac{u_{k+1} - 2u_k + u_{k-1}}{h^2}\,.$$

Damit folgt für die Approximation der Differentialgleichung in den **inneren** Punkten x_k

$$-\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + b(x_k)\frac{u_{k+1}-u_{k-1}}{2h} + c(x_k)u_k \;=\; f(x_k) \;=\; f_k \quad \text{für } k = 1,\ldots,n-1\,.$$

Aus den Randbedingungen folgt weiterhin

$$u_0 \;=\; 0, \quad u_n \;=\; 0\,.$$

Damit ergeben sich insgesamt $n+1$ Gleichungen für $n+1$ Unbekannte $u_0,\ldots,u_n$. Zu untersuchen bleibt die **eindeutige Lösbarkeit** des zugeordneten linearen Gleichungssystems $L_h\underline{u} = \underline{f}$ sowie die **Stabilität** und die **Konsistenz** des numerischen Näherungsverfahrens. Auf diese Betrachtungen soll an dieser Stelle jedoch verzichtet werden, siehe zum Beispiel [5, 15].
Insbesondere für das Randwertproblem

$$-u''(x) \;=\; f(x) \quad \text{für } x \in (a,b), \quad u(a) \;=\; u(b) \;=\; 0$$

ergibt sich zur Bestimmung der Näherungslösung u_k das lineare Gleichungssystem

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \end{pmatrix}.$$

Am Beispiel dieses linearen Gleichungssystems sollen später effiziente Lösungsverfahren diskutiert werden sollen. Obwohl die Diskretisierung des eindimensionalen Modellproblems auf eine **tridiagonale** Systemmatrix führt, deren Invertierung direkt mit optimalen Aufwand realisiert werden kann, lassen sich die bei der Untersuchung von Iterationsverfahren gewonnenen Aussagen auf den mehrdimensionalen Fall übertragen.

Betrachtet wird zunächst ein anderer Zugang zur Lösung des Randwertproblems

$$-u''(x) + b(x)u'(x) + c(x)u(x) \;=\; f(x) \quad \text{für } x \in (a,b), \quad u(a) \;=\; u(b) \;=\; 0.$$

Die Multiplikation mit einer hinreichend glatten Testfunktion $v(x)$ mit $v(a) = v(b) = 0$ liefert nach der Integration über (a,b) die Geichheit

$$\int\limits_a^b [-u''(x) + b(x)u'(x) + c(x)u(x)]v(x)dx \;=\; \int\limits_a^b f(x)v(x)dx\,.$$

Durch partielle Integration von

$$\int\limits_a^b [-u''(x)]v(x)dx \;=\; \int\limits_a^b u'(x)v'(x)dx - u'(x)v(x)|_a^b \;=\; \int\limits_a^b u'(x)v'(x)dx$$

ergibt sich die gesuchte Funktion u als Lösung des **Variationsproblems**

$$a(u,v) = F(v)$$

für alle Testfunktionen v mit $v(a) = v(b) = 0$ mit der **Bilinearform**

$$a(u,v) \;=\; \int\limits_a^b u'(x)v'(x)dx + \int\limits_a^b b(x)u'(x)v(x)dx + \int\limits_a^b c(x)u(x)v(x)dx$$

und mit der **Linearform**

$$F(v) \;=\; \int\limits_a^b f(x)v(x)dx\,.$$

Dabei ist die Dirichlet–Randbedingung $u(a) = u(b) = 0$ explizit zu fordern.
Eine Approximation des Variationsproblems ergibt sich durch eine **Approximation** des Funktionenraumes zur Bestimmung der Lösung u. Der Ansatz

$$u_h(x) \;=\; \sum_{k=0}^{n} u_k\varphi_k(x) \;=\; \sum_{k=1}^{n-1} u_k\varphi_k(x)$$

mit den in Abschnitt 1.1.5 definierten stückweise linearen Basisfunktionen $\varphi_k(x)$ bezüglich einer Diskretisierung des Intervalles (a,b) führt unter Verwendung der Testfunktionen $v(x) = \varphi_\ell(x)$ für $\ell = 1,\ldots,n-1$ auf das **Galerkin–Variationsproblem** zur Bestimmung von u_h als Lösung von

$$a(u_h,\varphi_\ell) \;=\; F(\varphi_\ell) \quad \text{für alle } \ell = 1,\ldots,n-1.$$

Aus der Linearität der Bilinearform $a(\cdot,\cdot)$ ergibt sich

$$\sum_{k=1}^{n-1} u_k a(\varphi_k,\varphi_\ell) \;=\; F(\varphi_\ell) \quad \text{für alle } \ell = 1,\ldots,n-1$$

und somit das lineare Gleichungssystem $A_h\underline{u} = \underline{f}$ mit der durch die Einträge

$$A_h[\ell,k] = a(\varphi_k,\varphi_\ell)$$

für $k,\ell = 1,\ldots,n-1$ erkärten **Steifigkeitsmatrix** A_h und dem durch

$$f_\ell \;=\; F(\varphi_\ell)$$

für $\ell = 1,\ldots,n-1$ gebildeten **Lastvektor** $\underline{f}$.

Für die stückweise linearen Basisfunktionen $\varphi_k(x)$ ergibt sich

$$\varphi_k'(x) \;=\; \begin{cases} \dfrac{1}{h} & \text{für } x \in (x_{k-1}, x_k), \\[2ex] -\dfrac{1}{h} & \text{für } x \in (x_k, x_{k+1}), \\[2ex] 0 & \text{sonst.} \end{cases}$$

Für das Modellproblem mit $b(x) = c(x) = 0$ für $x \in (a,b)$ folgt dann

$$A_h[\ell,k] \;=\; \int\limits_a^b \varphi_k'(x)\varphi_\ell'(x)dx \;=\; \int\limits_{(x_{k-1},x_{k+1})\cap(x_{\ell-1},x_{\ell+1})} \varphi_k'(x)\varphi_\ell'(x)dx \;=\; 0$$

für $\ell \neq k, k\pm 1$. Für die Diagonaleinträge $A_h[k,k]$ ergibt sich

$$A_h[k,k] \;=\; \int\limits_a^b [\varphi_k'(x)]^2 dx \;=\; \int\limits_{x_{k-1}}^{x_k} \frac{1}{h^2}\,dx + \int\limits_{x_k}^{x_{k+1}} \frac{(-1)^2}{h^2}\,dx \;=\; \frac{2}{h},$$

während für $\ell = k \pm 1$

$$\begin{aligned} A_h[k+1,k] \;&=\; \int\limits_a^b \varphi_k'(x)\varphi_{k+1}'(x)dx \;=\; \int\limits_{x_k}^{x_{k+1}} \frac{1}{h}(-\frac{1}{h})\,dx \;=\; -\frac{1}{h}, \\ A_h[k-1,k] \;&=\; -\frac{1}{h} \end{aligned}$$

folgt. Bei einer gleichmäßigen Unterteilung des Intervalles (a,b) ergibt sich also

$$A_h \;=\; \frac{1}{h}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Lemma 5.1 *Für die Steifigkeitsmatrix A_h sind die Eigenvektoren $\underline{v}^k$ für $k = 1,\ldots,n-1$ gegeben durch die Einträge*

$$v_\ell^k \;=\; \sin\frac{k\ell\pi}{n} \quad \text{für } \ell = 1,\ldots,n-1$$

und die zugehörigen Eigenwerte sind

$$\lambda_k(A_h) \;=\; \frac{4}{h}\sin^2\frac{k\pi}{2n}\,.$$

Beweis: Ausgangspunkt ist das kontinuierliche Eigenwertproblem

$$-u''(x) \;=\; \lambda\, u(x) \quad \text{für } x \in (a,b), \quad u(a) \;=\; u(b) \;=\; 0$$

mit den Eigenlösungen

$$v_k(x) \;=\; \sin\frac{k\pi(x-a)}{b-a}$$

und den zugehörigen Eigenwerten

$$\lambda_k \;=\; \left(\frac{k\pi}{b-a}\right)^2 \qquad \text{für } k = 1,2,3,\ldots.$$

Dies motiviert die Definition der Vektoren $\underline{v}^k$ mit

$$v_\ell^k \;=\; v_k(x_\ell) \;=\; \sin\frac{k\pi(x_\ell - a)}{b-a} \;=\; \sin\frac{k\ell\pi}{n} \quad \text{für } \ell = 1,\ldots,n-1.$$

Zu zeigen bleibt, daß die Vektoren $\underline{v}^k$ die Eigenvektoren von A_h sind: Zu berechnen ist also

$$A_h\underline{v}^k = \frac{1}{h}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}\begin{pmatrix} v_1^k \\ v_2^k \\ \vdots \\ v_{n-2}^k \\ v_{n-1}^k \end{pmatrix} = \frac{1}{h}\begin{pmatrix} 2v_1^k - v_2^k \\ -v_1^k + 2v_2^k - v_3^k \\ \vdots \\ -v_{n-3}^k + 2v_{n-2}^k - v_{n-1}^k \\ -v_{n-2}^k + 2v_{n-1}^k \end{pmatrix}$$

und wegen $v_0^k = v_n^k = 0$ genügt für $\ell = 1,\ldots,n-1$ die Auswertung von

$$-v_{\ell-1}^k + 2v_\ell^k - v_{\ell+1}^k \;=\; -\sin\frac{k(\ell-1)\pi}{n} + 2\sin\frac{k\ell\pi}{n} - \sin\frac{k(\ell+1)\pi}{n}.$$

Mit

$$\sin\alpha - \sin\beta \;=\; 2\cos\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$

und

$$\cos\alpha - \cos\beta \;=\; -2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$

folgt daraus

$$\begin{aligned} -v_{\ell-1}^k + 2v_\ell^k - v_{\ell+1}^k \;&=\; 2\cos\frac{k(2\ell-1)\pi}{2n}\sin\frac{k\pi}{2n} - 2\cos\frac{k(2\ell+1)\pi}{2n}\sin\frac{k\pi}{2n} \\ &=\; 2\sin\frac{k\pi}{2n}\left[\cos\frac{k(2\ell-1)\pi}{2n} - \cos\frac{k(2\ell+1)\pi}{2n}\right] \\ &=\; 4\sin^2\frac{k\pi}{2n}\sin\frac{k\ell\pi}{n} \;=\; 4\sin^2\frac{k\pi}{2n}\,v_\ell^k \end{aligned}$$

und somit die Behauptung. ■

Folgerung 5.1 *Die extremalen Eigenwerte der Steifigkeitsmatrix A_h sind*

$$\begin{aligned} \lambda_{\min}(A_h) \;=\; \lambda_1(A_h) \;&=\; \frac{4}{h}\sin^2\frac{\pi}{2n} \;=\; \frac{4}{h}\sin^2\frac{\pi h}{2(b-a)} \\ &=\; \frac{4}{h}\left[\left(\frac{\pi h}{2(b-a)}\right)^2 + \mathcal{O}(h^4)\right] \;=\; \frac{\pi^2}{(b-a)^2}\,h + \mathcal{O}(h^3) \end{aligned}$$

sowie

$$\lambda_{\max}(A_h) \;=\; \lambda_n(A_h) \;=\; \frac{4}{h}\sin^2\frac{n\pi}{2n} \;=\; \frac{4}{h}.$$

Damit ergibt sich für die spektrale Konditionszahl von A_h

$$\kappa_2(A_h) \;=\; \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} \;=\; \frac{4(b-a)^2}{\pi^2 + \mathcal{O}(h^2)}\,\frac{1}{h^2} \;=\; \mathcal{O}(n^2)\,.$$

Eine Verdoppelung der Freiheitsgrade n zieht also eine Vervierfachung der spektralen Konditionszahl $\kappa_2(A_h)$ nach sich.

Für die iterative Lösung des aus der Approximation resultierenden linearen Gleichungssystem $A_h\underline{u} = \underline{f}$ wird das ω–Jacobi–Verfahren

$$\underline{u}^{k+1} \;=\; \underline{u}^k - \omega D_h^{-1}(A_h\underline{u}^k - \underline{f}).$$

mit einer beliebigen Startnäherung $\underline{u}^0$ betrachtet. Dabei ist $D_h = \operatorname{diag} A_h \;=\; 2h^{-1}\, I_{n-1}$. Für den zugehörigen Fehler ergibt sich

$$\underline{u}^{k+1} - \underline{u} \;=\; (I - \omega\frac{h}{2}A_h)(\underline{u}^k - \underline{u}) \;=\; (I - \omega\frac{h}{2}A_h)^{k+1}(\underline{u}^0 - \underline{u}).$$

Wird der Anfangsfehler $\underline{e}^0 \;=\; \underline{u}^0 - \underline{u}$ in die in Lemma 5.1 angegebenen Eigenvektoren $\underline{v}^k$ der Steifigkeitsmatrix A_h entwickelt,

$$\underline{e}^0 \;=\; \sum_{\ell=1}^{n-1} \alpha_\ell \underline{v}^\ell,$$

dann folgt

$$\begin{aligned}
\underline{e}^{k+1} \;&=\; \sum_{\ell=1}^{n-1} \alpha_\ell (I - \omega\frac{h}{2}A_h)^{k+1}\underline{v}^\ell \;=\; \sum_{\ell=1}^{n-1} \alpha_\ell [1 - \omega\frac{h}{2}\lambda_\ell(A_h)]^{k+1}\underline{v}^\ell \\
&=\; \sum_{\ell=1}^{n-1} \alpha_\ell \left[1 - 2\omega \sin^2\frac{\ell\pi}{2n}\right]^{k+1} \underline{v}^\ell .
\end{aligned}$$

In diesem Fall konvergiert das ω–Jacobi–Verfahren für alle $\omega \in (0,1]$, wobei die Konvergenzrate durch den zu $\ell = 1$ gehörenden Eigenwert

$$1 - 2\omega \sin^2\frac{\pi}{2n} \;=\; 1 - 2\omega\sin^2\frac{\pi h}{2(b-a)} \;=\; 1 - \mathcal{O}(h^2)$$

dominiert wird. Andererseits ist für $\ell = n$

$$1 - 2\omega\sin^2\frac{\ell\pi}{2n} \;=\; 1 - 2\omega\sin^2\frac{\pi}{2} \;=\; 1 - 2\omega$$

bzw. für $\ell = n/2$

$$1 - 2\omega\sin^2\frac{\ell\pi}{2n} \;=\; 1 - 2\omega\sin^2\frac{\pi}{4} \;=\; 1 - 2\omega\left(\frac{\sqrt{2}}{2}\right)^2 \;=\; 1 - \omega\,.$$

Für

$$\omega \;=\; \frac{2}{3}$$

folgt dann

$$\left|1 - 2\omega\sin^2\frac{\ell\pi}{2n}\right| \;\leq\; \frac{1}{3} \quad \text{für } \ell = \frac{n}{2}, \ldots, n-1.$$

In diesem Fall werden bei der Anwendung des ω–Jacobi–Verfahrens mit $\omega = 2/3$ die Koeffizienten α_ℓ des Anfangsfehlers $\underline{e}^0$ mit $\ell \;=\; n/2, \ldots, n-1$ besonders stark gedämpft. Die verbleibenden Koeffizienten a_ℓ für $\ell = 1, \ldots, n/2 - 1$ entsprechen aber gerade wieder dem Ausgangsproblem für eine Diskretisierung mit doppelter Schrittweite. Mit anderen Worten, der verbleibende Fehler nach einigen Schritten des ω–Jacobi–Verfahrens kann auf einem gröberen Gitter dargestellt werden. Die rekursive Anwendung führt dann auf **Mehrgitterverfahren** [4, 7].

Übungsaufgaben

5.1. Gegeben sei das Anfangswertproblem

$$y'(x) = y(x) - x^2 + 2x \quad \text{für } x > 0, \quad y(0) = 0$$

mit der exakten Lösung $y(x) = x^2$. Man gebe die Rekursionsvorschriften des expliziten und des impliziten Euler–Verfahrens an und verwende diese zur Berechnung der Näherungslösung in $x = 0.2$ mit den Schrittweiten $h_1 = 0.1$ und $h_2 = 0.05$.

5.2. Betrachtet werde das Anfangswertproblem

$$u''(x) = f(x, u, u') \quad \text{für } x > 0, \quad u(x_0) = u_0, \quad u'(x_0) = u_1.$$

Man überführe die Differentialgleichung in ein System gewöhnlicher Differentialgleichungen 1. Ordnung und leite für dieses ein explizites Eulersches Polygonzugverfahren her.
Mit einer Schrittweite von $h = 0.1$ berechne man damit die Näherungslösung in $x = 0.2$ des Anfangswertproblems

$$u''(x) + u(x) = 0, \quad u(0) = 0, \quad u'(0) = 1.$$

5.3. Ausgehend von

$$y(x_{k+1}) = y(x_k) + \int\limits_{x_k}^{x_{k+1}} f(x, y(x))dx$$

konstruiere man ein Mehrschrittverfahren durch quadratische Interpolation des Integranden in den Stützstellen x_{k-1}, x_k und x_{k+1}.

5.4. Man untersuche die Konsistenzordnung und die Stabilität des Mehrschrittverfahrens von Milne–Simpson,

$$y_{k+1} = y_{k-1} + \frac{1}{3}h\left[f(x_{k-1}, y_{k-1}) + 4f(x_k, y_k) + f(x_{k+1}, y_{k+1})\right].$$

5.5. Für die Heunsche Methode

$$y_{k+1} = y_k + \frac{1}{2}h\left[f(x_k, y_k) + f(x_k + h, y_k + hf(x_k, y_k))\right]$$

bestimme man die Stabilitätsfunktion $R(z)$ und das zugehörige Stabilitätsgebiet.

Literaturverzeichnis

[1] J. Douglas Faires, R. L. Burden: Numerische Methoden. Spektrum, Heidelberg, 1994.

[2] K. Eriksson, D. Estep, P. Hansbo, C. Johnson: Computational Differential Equations. Cambridge University Press, 1996.

[3] G. H. Golub, C. F. van Loan: Matrix Computations. The John Hopkins University Press, Baltimore, 1989.

[4] W. Hackbusch: Iterative Lösung grosser schwachbesetzter Gleichungssysteme. B. G. Teubner, Stuttgart, 1993.

[5] M. Hanke–Bourgeois: Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens. B. G. Teubner, Stuttgart, 2002.

[6] M. Hestenes, E. Stiefel: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Stand. 49 (1952) 409–436.

[7] M. Jung, U. Langer: Methode der Finiten Elemente für Ingenieure. B. G. Teubner, Stuttgart, Leipzig, Wiesbaden, 2001.

[8] G. Maess: Vorlesungen über Numerische Mathematik I/II. Birkhäuser, Basel, 1985, 1988.

[9] A. Meister: Numerik linearer Gleichungssysteme. Eine Einführung in moderne Verfahren. Vieweg, Braunschweig, 1999.

[10] A. Quarteroni, R. Sacco, F. Saleri: Numerical Mathematics. Springer, New York, 2000.

[11] H.–G. Roos, H. Schwetlick: Numerische Mathematik. B. G. Teubner, Stuttgart, Leipzig, 1999.

[12] Y. Saad, M. H. Schultz: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. 7 (1986) 856–869.

[13] R. Schaback, H. Werner: Numerische Mathematik. Springer, Berlin, 1992.

[14] O. Steinbach: Lösungsverfahren für Lineare Gleichungssysteme. Algorithmen und Anwendungen. B. G. Teubner, Stuttgart, Leipzig, Wiesbaden, in Vorbereitung.

[15] J. Stoer: Numerische Mathematik 1. Springer, Berlin, 1972.

[16] J. Stoer, R. Bulirsch: Numerische Mathematik 2. Springer, Berlin, 1973.

[17] W. L. Wendland, O. Steinbach: Analysis. B. G. Teubner, Stuttgart, Leipzig, Wiesbaden, in Vorbereitung.

Erschienene Preprints ab Nummer 2005/1

2005/1 *O. Steinbach*: Numerische Mathematik 1. Vorlesungsskript.