Vorlesungsmitschrift
Höhere Numerische Mathematik

von Michael Schaefer

10. Juli 2007

Zusammenfassung

Bei dem vorliegenden Dokument handelt es sich um die von mir mit LaTeXgesetzte Version meiner Vorlesungsmitschriften aus der Vorlesung **Höhere Numerische Mathematik** im Sommersemester 2007, gelesen von **Prof. Dr. Helmut Maurer** an der **Westfälischen Wilhelms-Universität Münster**. Ich tue dies, damit ich beim Wiederholen des Stoffs, aber insbesondere auch zur Vorbereitung der Klausur, nicht auf eine lose Blattsammlung von handgeschriebenen und mehr oder weniger lesbaren Zetteln angewiesen bin. Wer Fehler in diesem Dokument findet, den bitte ich, mir diese per E-Mail mitzuteilen:

michael.schaefer@uni-muenster.de

Wichtig: Da dieses Dokument ständigen Korrekturen unterliegt, bietet sich ein Ausdruck erst am Ende des Semesters an. **Den Stand des Dokumentes entnehme man bitte dem weiter oben angeführten Datum.**

Inhaltsverzeichnis

V Eigenwertprobleme

§ 16 Theoretische Grundlagen, Potenzmethode

Sei $A = (a_{i,k})$ eine $n \times n$-Matrix mit $a_{ik} \in \mathbb{C}$. Eine Zahl $\lambda \in \mathbb{C}$ heißt Eigenwert (EW) von A, falls $x \in \mathbb{C}^n, x \neq 0$ existiert mit $Ax = \lambda x$, d.h. $(A - \lambda E)x = 0$. Ein solcher Vektor $x \in \mathbb{C}^n$ heißt (Rechts-)Eigenvektor (EV) zum EW λ. Der Eigenraum

$$L(\lambda) := \{x \in \mathbb{C}^n \mid (A - \lambda E)x = 0\}$$

ist ein linearer Teilraum des $\mathbb{C}^n$ der Dimension

$$\rho(\lambda) = n - \operatorname{rang}(A - \lambda E).$$

ρ nennt man auch Vielfachheit des EW λ. Jeder EW λ ist Nullstelle des charakteristischen Polynoms

$$\begin{aligned} \varphi(\lambda) &= \det(A - \lambda E) = (-1)^n(\lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_1\lambda + \alpha_0) \\ &= (-1)^n(\lambda - \lambda_1)^{\sigma_1} \cdot \ldots \cdot (\lambda - \lambda_k)^{\sigma_k} \end{aligned}$$

Für $\sigma(\lambda_i) = \sigma_i$, $i = 1, \ldots, k$, gilt $1 \leq \rho(\lambda_i) \leq \sigma(\lambda_i)$ für $i = 1, \ldots, k$.

Beispiel. Für

$$J = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}$$

eine $n \times n$-Matrix mit $\lambda \in \mathbb{C}$ gilt $\varphi(\mu) = \det(J - \mu E) = (\lambda - \mu)^n$. Also ist λ der einzige EW von J mit $\sigma(\lambda) = n, \quad \rho(\lambda) = 1$, da $\operatorname{rang}(J - \lambda E) = n - 1$.

Sei $A^H := \overline{A}^*$. Wegen

$$\det(A - \lambda E) = \det(A^* - \lambda E) \text{ und } \overline{\det(A - \lambda E)} = \det(A^H - \overline{\lambda} E)$$

folgt: Ist λ ein EW von A, so ist λ auch ein EW von A^* und $\overline{\lambda}$ ist EW von A^H.
Zwischen den zugehörigen Eigenvektoren x, y, z mit

$$Ax = \lambda x, \quad A * y = \lambda y, \quad A^H z = \overline{\lambda} z$$

gelten die Beziehungen

$$\overline{y} = z, \quad y^* = z^H, \quad z^H A = \lambda z^H.$$

Sei nun T eine reguläre $n \times n$-Matrix. Die Transformation

$$B := T^{-1}AT$$

heißt Ähnlichkeitstransformation. Aus $Ax = \lambda x$ folgt

$$By = \lambda y, \qquad y := T^{-1}x.$$

Es gilt

$$\det(B - \lambda E) = \det(T^{-1}(A - \lambda E)T) = \det(T^{-1}) \det(A - \lambda E) \det(T) = \det(A - \lambda E).$$

Also haben A und B dieselben Eigenwerte λ. Die Vielfachheiten $\rho(\lambda)$ und $\sigma(\lambda)$ sind invariant.
Bei den wichtigsten Verfahren zur EW-Berechnung werden eine Reihe von Ähnlichkeitstranformationen durchgeführt:

$$A_1 := A, \qquad A_{k+1} := T_k^{-1} A_k T_k, \quad k = 1, 2, \ldots \tag{16.1}$$

Damit wird A auf eine "einfachere" Gestalt gebracht, deren EW und EV leichter zu berechnen sind, z.B. nimmt man $T_k = L_k$ (Elementarmatrizen) oder $T_k = Q_k$ (HOUSEHOLDER-Matrizen).
Eine wichtige Klasse von Matrizen sind die normalen Matrizen A, für die gilt

$$A^H A = A A^H.$$

(16.2) Satz. Eine $n \times n$-Matrix A ist genau dann normal, wenn es eine unittäre Matrix U gibt mit

$$U^{-1}AU = U^H AU = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

sowie $U^H U = E$.
Normale Matrizen sind diagonalisierbar und besitzen n linear unabhängige, zueinander orthogonale Eigenvektoren $x_1, \ldots, x_n$ mit $Ax_i = \lambda_i x_i$ für $i = 1, \ldots, n$. Dies sind die Spalten von U: $U = (x_1, \ldots, x_n)$.

Beweis. vgl. Lineare Algebra.

Es ist festzuhalten, dass insbesondere alle hermitschen und symmetrischen Matrizen normal sind.

Die Potenzmethode

Voraussetzung: A habe n linear unabhängige EV $x_1, \ldots, x_n$. Für die EWe gelte $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. Als Startwert der Iteration wähle man ein $x^{(0)} \in \mathbb{R}^n$. Die Iteration lautet dann

$$x^{(k+1)} = Ax^{(k)} = A^{k+1}x^{(0)}, \quad k = 0, 1, \ldots$$

Aufgrund der Darstellung $x^{(0)} = \sum_{i=1}^n c_i x_i$ erhält man

$$x^{(k)} = A^k x^{(0)} = \sum_{i=1}^n c_i A^k x_i = \sum_{i=1}^n c_i \lambda_i^k x_i = \lambda_1^k (c_1 x_1 + r_k)$$

mit dem Rest

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

Berechnung von λ_1 und x_1. Als Voraussetzung muss gelten $c_1 \neq 0$ und $x_{1,j} \neq 0$ für ein $j \in \{1, \ldots, n\}$. Es folgt dann

$$\frac{x_j^{()k+1)}}{x_j^{(k)}} = \lambda_1 \frac{(c_1 x_1 + r_{k+1})_j}{(c_1 x_1 + r_k)_j} = \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right),$$

$$\frac{x^{(k)}}{x_j^{(k)}} = \frac{x_1}{x_{1,j}} + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right).$$

Beispiel. Sei

$$A = \begin{pmatrix} 90 & 231 & 70 \\ 110 & 336 & 110 \\ 70 & 231 & 90 \end{pmatrix}.$$

Es folgt dann

$$x^{(0)}, x^{(1)}, x^{(2)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 391 \\ 556 \\ 391 \end{pmatrix}, \quad \begin{pmatrix} 190756 \\ 272836 \\ 190756 \end{pmatrix}.$$

Damit berechnen wir zunächst

$$\frac{x_2^{(1)}}{x_2^{(0)}} = 556, \quad \frac{x_2^{(2)}}{x_2^{(1)}} = \frac{272836}{556} \approx 490.71.$$

Weiterhin bestimmen wir die normierten Vektoren zu

$$\frac{x^{(k)}}{x_2^{(k)}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0.703237 \\ 1 \\ 0.703237 \end{pmatrix}, \quad \begin{pmatrix} 0.699160 \\ 1 \\ 0.699160 \end{pmatrix}.$$

Die exakten Werte lauten $\lambda_1 = 490$, $\lambda_2 = 20$ und $x_1 = \begin{pmatrix} 0.7 \\ 1 \\ 0.7 \end{pmatrix}$. Die Begründung für die schnelle Konvergenz lautet, dass der Quotient $\frac{\lambda_2}{\lambda_2} \approx 0.04$ sehr klein ist.

Die inverse Potenzmethode

Als Voraussetzung gelte $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$. Ziel ist die Berechnung von λ_2. Wir wählen nun ein $\mu \in \mathbb{R}$, sodass gilt $|\lambda_2 - \mu| < |\lambda_i - \mu|$ für $i = 1, 3, \ldots, n$. Dann hat die Matrix $T := (A - \mu E)^{-1}$ die Eigenwerte $(\lambda_i - \mu)^{-1}$.
Also hat T den größten Eigenwert $\frac{1}{\lambda_2 - \mu}$. Wir wenden nun die Potenzmethode

$$x^{(k+1)} = Tx^{(k)}, \quad k = 0, 1, \ldots, n$$

auf die Matrix T an. Nun löse man das LGS $(A - \mu E)x^{(k+1)} = x^{(k)}$, indem man die LR-Zerlegung von $A - \mu E$ bestimmt (diese hängt nicht von k ab) und damit das LGS löst.

§ 17 Transformationsmethoden

Wir betrachten Sequenzen von Ähnlichkeitsabbildungen:

$$\begin{aligned} A &= A_1 \to A_2 \to \ldots \to A_{m+1} \\ A_{k+1} &= T_k^{-1} A_k T_k, \quad k = 1, 2, \ldots \\ B &= A_{m+1} = T^{-1} A T, \quad T := T_1 T_2 \cdots T_m \end{aligned} \tag{17.1}$$

Ziele:

(i) B hat möglichst einfache Form.

(ii) das EW-Problem für B sei nicht schlechter konditioniert als dasjenige für A.

zu (i). B ist entweder eine HESSENBERG-Matrix

$$B = \begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix} = (b_{ij}), \quad b_{ij} = 0 \quad \text{für } i > j+1$$

oder eine Tridiagonalmatrix (symmetrisch, falls A symmetrisch ist)

$$B = \begin{pmatrix} \delta_1 & \gamma_2 & & \\ \gamma_2 & \delta_2 & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \gamma_n & \delta_n \end{pmatrix}.$$

zu (ii). Zunächst gilt

$$\begin{aligned} B &= T^{-1} A T \\ B + \Delta B &= T^{-1}(A + \Delta A) T \\ \Delta A &= T \Delta B T^{-1} \end{aligned}$$

Wir erhalten die Abschätzungen

$$\begin{aligned} ||B|| &\le \left|\left|T^{-1}\right|\right| \cdot ||A|| \cdot ||T|| = cond(T)\, ||A|| \\ ||\Delta A|| &\le cond(T)\, ||\Delta B|| \end{aligned}$$

und damit

$$\frac{||\Delta A||}{||A||} \le cond(T)^2 \frac{||\Delta B||}{||B||}.$$

Desweiteren gilt

$$cond(T) = cond(T_1 T_2 \cdots T_m) \le cond(T_1) \cdots cond(T_m).$$

Folgerung: Man wähle T_k, sodass die Kondition von T_k nicht zu groß wird.

Beispiele. Elementarmatrizen

$$T_k = L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,k} & & 1 \end{pmatrix} \tag{17.2a}$$

mit $|l_{ik}| \le 1$. Dann gilt $cond_\infty(T_k) \le 4$.

HOUSEHOLDER-Matrix

$$(17.2b) \quad T_k = E - 2w_k w_k^*$$

mit $w_k^* w_k = 1$. Dann gilt $cond_2(T_k) = 1$.

GIVENS-Rotationen.

§ 17.1 Transformation auf HESSENBERG-Form

Wir betrachten eine Sequenz $A = A_1 \to A_2 \to \ldots \to A_{n-1} = B$ mit $A_{k+1} = T_k^{-1} A_k T_k$. Wir benutzen die elementare Permutation

$$P_{rs} = \begin{pmatrix} 1 & & & & & & & & & 0 \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & \ldots & \ldots & \ldots & 1 & & \\ & & & \vdots & 1 & & & \vdots & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & \vdots & & & 1 & \vdots & & \\ & & & 1 & \ldots & \ldots & \ldots & 0 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ 0 & & & & & & & & & 1 \end{pmatrix}$$

Es gilt $P_{rs}^{-1} = P_{rs}$, $P_{rs}A$ vertauscht die Zeilen r und s von A und AP_{rs} vertauscht die Spalten r und s von A.

Man wählt nun die Elementarmatrix L_k in (17.2a) und erhält im Übergang

$$A_k = \begin{pmatrix} * & & & & * & \ldots & * \\ * & \ddots & & & \vdots & & \vdots \\ & \ddots & \ddots & & \vdots & & \vdots \\ & & * & * & \vdots & & \vdots \\ 0 & \ldots & 0 & * & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & * & * & \ldots & * \end{pmatrix} = \left(\begin{array}{c|c|c} B_k & * \\ \hline 0 \;|\; a & \tilde{A}_k \end{array}\right)$$

mit B_k HESSENBERG-Matrix. Ziel ist eine Transformation

$$a = \begin{pmatrix} \alpha_{k+1,k} \\ \vdots \\ \alpha_{nk} \end{pmatrix} \to \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Bestimme dazu zunächst ein $r \geq k+1$ mit $|\alpha_{r,k}| = \max_{k+1 \leq j \leq n} |\alpha_{jk}|$ und unterscheide zwei Fälle:

(i) $\alpha_{rk} = 0$: Setze $T_k = E_n$.

(ii) $a_{rk} \neq 0$: Vertausche Zeile und Spalte r und $k+1$ in A_k, d.h. setze $A_k' := P_{r,k+1}^{-1} A_k P_{r,k+1} = (\alpha_{ij}')$. Definiere die Elementarmatrix L_{k+1} (Index $k+1$ wegen HESSENBERG-Form) mit

$$l_{j,k+1} = \frac{\alpha_{jk}'}{\alpha_{k+1,k}'}, \quad j = k+1, \ldots, n.$$

Beachte: $|l_{j,k+1}| \leq 1$. Setze nun $A_{k+1} := L_{k+1}^{-1} A_k' L_{k+1} = T_k^{-1} A_k T_k$ mit $T_k := P_{r,k+1} L_{k+1}$.

Ergebnis: $B = A_{n-1}$ hat gewünschte HESSENBERG-Form. Die Anzahl der benötigten Operationen ergibt sich zu $\frac{5}{6}n^3 + O(n^2)$.

Weiterhin benutzt man die Methode von HYMAN, welche eine "einfache" Berechnung des charakteristischen Polynoms

$$p(\lambda) = \det(B - \lambda E)$$

sowie des Quotienten $\frac{p(\lambda)}{p'(\lambda)}$ liefert. Anschließend wendet man das NEWTON-Verfahren an (vgl. STOER: Numerische Mathematik 1).

§ 17.2 Transformation einer symmetrischen Matrix auf Tridiagonalform

Sei A symmetrisch, d.h. $A = A^*$. Wähle HOUSEHOLDER-Matrizen

$$T_k = E - 2ww^* = E - \beta u_k u_k^*$$

mit $||w_k||_2 = 1$. Dann ist $A_{k+1} = T_k^{-1} A_k T_k$ symmetrisch. Für den Übergang $A_k \to A_{k+1}$ gilt

$$A_k = \left(\begin{array}{c|c} J_k & \begin{matrix}0\\ a_k^*\end{matrix} \\ \hline \begin{matrix}0 & a_k\end{matrix} & \tilde{A}_k \end{array}\right), \quad J_k = \begin{pmatrix} \delta_1 & \gamma_2 & & \\ \gamma_2 & \delta_2 & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \gamma_n & \delta_n \end{pmatrix}$$

Nach §6 gibt es eine (n,n)-HOUSEHOLDER-Matrix $\tilde{T}_k = E_{n-k} - \beta_k \tilde{u}_k \tilde{u}_k^*$ mit $\tilde{u}_k \in \mathbb{R}^{n-k}$ für die gilt

$$\tilde{T}_k a_k = c e_1, \quad e_1 \in \mathbb{R}^{n-k}, \quad |c| = ||a_k||_2 .$$

Wir ergänzen $\tilde{T}_k$ zu einer (n,n)-Matrix

$$T_k = \left(\begin{array}{c|c} E_k & 0 \\ \hline 0 & \tilde{T}_k \end{array}\right) = E_n - \beta_k u_k u_k^*, \qquad u_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{u}_k \end{pmatrix} \in \mathbb{R}^n.$$

Dann gilt mit $\gamma_{k+1} := c$

$$A_{k+1} = T_k^{-1} A_k T_k = T_k A_k T_k = \left(\begin{array}{c|c} J_k & \begin{matrix}0\\ \gamma_{k+1}\,0\cdots 0\end{matrix} \\ \hline \begin{matrix}0\end{matrix} & \begin{matrix}\gamma_{k+1}\\ 0\\ \vdots\\ 0\end{matrix} \quad \tilde{T}_k \tilde{A}_k \tilde{T}_k \end{array}\right)$$

Dabei ist

$$\tilde{T}_k \tilde{A}_k \tilde{T}_k = \tilde{A}_k - \tilde{U}_k q_k^* - q_k \tilde{u}_k^*$$

mit $q_k := p_k - \frac{1}{2}(p_k^* \tilde{u}_k)\,\tilde{u}_k$ und $p_k := \beta_k \tilde{A}_k \tilde{u}_k$. Damit ist nun $B = J_n = A_{n-1}$ tridiagonal.

Anwendung des NEWTON-Verfahrens zur Berechnung der Eigenwerte der Tridiagonalmatrix

$$J = J_n = \begin{pmatrix} \delta_1 & \gamma_2 & & \\ \gamma_2 & \delta_2 & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \gamma_n & \delta_n \end{pmatrix}.$$

Das charakteristische Polynom lautet

$$p_n(\lambda) = \det(J_n - \lambda E_n).$$

Für $p_k(\lambda) = \det(J_k - \lambda E_k)$ gilt die folgende Drei-Term-Rekursion:

$$p_0(\lambda) = 1, \quad p_1(\lambda) = \delta_1 - \lambda, \quad p_k(\lambda) = (\delta_k - \lambda)p_{k-1}(\lambda) - \gamma_k^2 p_{k-2}(\lambda) \text{ für } k = 2, \ldots, n \tag{17.3}$$

oBdA sei $\gamma_i \neq 0$ für $i = 2, \ldots, n$, d.h. J_n sei unzerlegbar. Setze für $\lambda \in \mathbb{R}$:

$$q(\lambda) = \begin{pmatrix} q_0(\lambda) \\ \vdots \\ q_{n-1}(\lambda) \end{pmatrix} \in \mathbb{R}^n$$

mit $q_0(\lambda) = 1$ und $q_k(\lambda) = \frac{(-1)^k p_k(\lambda)}{\gamma_2 \cdot \ldots \cdot \gamma_{k+1}}$ für $k = 1, \ldots, n$ sowie $\gamma_{n+1} = 1$. Dann ist (17.3) äquivalent zu

$$(J_n - \lambda E_n)q(\lambda) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q_n(\lambda) \end{pmatrix} \tag{17.4}$$

Für die Eigenwerte λ_k von J_n gilt $p_n(\lambda_k) = 0$, also $q_n(\lambda_k) = 0$ nach Definition. Damit gilt

$$(J_n - \lambda_k E_n)q(\lambda_k) = 0, \quad q(\lambda_k) \neq 0 \in \mathbb{R}^n$$

wegen $q_0(\lambda) \equiv 1$. Daher ist $q(\lambda_k) \in \mathbb{R}^n$ ein Eigenvektor zum Eigenwert λ_k. Differenziert man (17.4) bezüglich λ, so erhält man

$$-q(\lambda) + (J_n - \lambda E_n)q'(\lambda) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q_n'(\lambda) \end{pmatrix}.$$

Multiplikation mit $q^*(\lambda_k)$ für $\lambda = \lambda_k$ liefert wegen $q(\lambda_k)^*(J_n - \lambda_k E_n) = 0$ die Abschätzung

$$0 < q(\lambda_k)^* q(\lambda_k) = q_{n-1} q_n'(\lambda_k) = q_{n-1}(\lambda_k) \frac{(-1)^k p_n'(\lambda_k)}{\gamma_2 \cdot \ldots \cdot \gamma_{n+1}}.$$

Insbesondere folgt $p_n'(\lambda_k) \neq 0$. Also ist jeder Eigenwert λ_k eine einfache NUllstelle von $p_n(\lambda)$.

(17.5) Satz. Es gelte $\gamma_i \neq 0$ für $i = 2, \ldots, n$. Dann besitzt J_n genau n verschiedene Eigenwerte $\lambda_n < \ldots < \lambda_1$. □

Die Berechnung von $p_n'(\lambda)$ erfolgt durch Differentiation von (17.3):

$$\left.\begin{aligned} p_0'(\lambda) = 0, \quad p_1'(\lambda) = -1 \\ p_k'(\lambda) = -p_{k-1}(\lambda) + (\delta_k - \lambda)p_{k-1}'(\lambda) - \gamma_k^2 p_{k-2}'(\lambda) \end{aligned}\right\} \tag{17.6}$$

für $k = 2, \ldots, n$.

Newton-Verfahren zur Berechnung von λ_1: Es gilt die Iteration

$$\lambda^{(k+1)} = \lambda^{(k)} - \frac{p_n(\lambda^{(k)})}{pn'(\lambda^{(k)})}$$

Für den Startwert $\lambda^{(0)}$ macht man folgende Schätzung:

$$|\lambda_j| \leq \rho(J_n) \neq ||J_n||_\infty = \max_k \{|\gamma_k| + |\delta_k| + |\delta_{k+1}\} =: \lambda^{(0)}$$

Methode von MAEHLY zur Berechnung der Eigenwerte $\lambda_2, \lambda_3, \ldots$: Man beachte zunächst, dass das Polynom

$$p_{n,j}(\lambda) := \frac{p_n(\lambda)}{(\lambda - \lambda_1) \cdot \ldots \cdot (\lambda - \lambda_j)}$$

die Nullstellen $\lambda_{j+1}, \ldots, \lambda_n$ besitzt. Nun wendet man die NEWTON-Iteration zur Berechnung der größten Nullstelle λ_{j+1} von $p_{n,j}(\lambda)$ an:

$$\lambda^{(k+1)} = \lambda^{(k)} - \frac{p_{n,j}(\lambda^{(k)})}{p'_{n,j}(\lambda^{(k)})}.$$

Hierbei gilt nun

$$p'_{n,j}(\lambda) = \frac{p'_n(\lambda)}{(\lambda - \lambda_1) \cdot \ldots \cdot (\lambda - \lambda_j)} - \frac{p_n(\lambda)}{(\lambda - \lambda_1) \cdot \ldots \cdot (\lambda - \lambda_j)} \sum_{i=1}^{j} \frac{1}{\lambda - \lambda_i}$$

und damit

$$\frac{p_{n,j}(\lambda)}{p'_{n,j}(\lambda)} = \frac{p_n(\lambda)}{p'_n(\lambda) - \sum_{i=1}^{j} \frac{p_n(\lambda)}{\lambda - \lambda_i}}$$

Beispiel. Es sei

$$J := \begin{pmatrix} 12 & 1 & 0 & 0 & 0 \\ 1 & 9 & 1 & 0 & 0 \\ 0 & 1 & 6 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Die Eigenwerte von J_5 sind symmetrisch zu $\lambda_3 = 6$, d.h. es gilt $\lambda_i + \lambda_{6-i} = 12$ für $i = 1, \ldots, 5$. Es sind daher nur λ_1 und λ_2 zu berechnen. Startwert für die Berechnung von λ_1 ist $x^{(0)} = ||J_5||_2 = 13$. Es ergibt sich

$$\lambda_1 = 12.316876, \quad \lambda_2 = 9.0161363, \quad \lambda_3 = 6, \quad \lambda_4 = 12 - \lambda_2, \quad \lambda_5 = 12 - \lambda_1.$$

§ 18 Das QR-Verfahren

Historisches: Das QR-Verfahren (FRANCIS, 1961) ist eine Weiterentwicklung des LR-Algorithmus von RUTISHAUSER.

Das QR-Verfahren wird hauptsächlich auf HESSENBERG- bzw. auf Tridiagonalmatrizen angewandt. Man bildet eine Sequenz von Matrizen

$$\left.\begin{array}{l} A_1 := A, \quad A_k = Q_k R_k, \quad Q_k \text{ orthogonal, } Q_k Q_k = E_n, \quad R_k \text{ r. o. Dreiecksmatrix} \\ A_{k+1} = R_k Q_k, \quad k = 1, 2, \ldots \end{array}\right\} \tag{18.1}$$

Die QR-Zerlegung $A_k = Q_k R_k$ existiert nach Satz (6.4). Beachte: Diese ist nicht eindeutig bestimmt!

(18.2) Lemma. Sei A regulär und $A := QR$ mit einer orthogonalen Matrix Q und einer rechten oberen Dreiecksmatrix R. Dann sind Q und R eindeutig bestimmt bis auf Multiplikation mit einer orthogonalen Diagonalmatrix D, d.h. falls

$$A = Q_1 R_1 = Q_2 R_2,$$

so gibt es $D = \operatorname{diag}(\pm 1)$ mit $Q_1 = Q_2 D$ und $R_1 = D R_2$.

Beweis. Als Übungsaufgabe.

Das QR-Verfahren ist eine Sequenz von Ähnlichkeitstransformationen.

(18.3) Lemma. Die Matrizen Q_k und R_k seien gemäß (18.1) definiert. Mit $P_k := Q_1 \cdot \ldots \cdot Q_k$ und $U_k := R_k \cdot \ldots \cdot R_1$ gelten

(i) A_{k+1} ist ähnlich zu A_k, denn $A_{k+1} = Q_k^{-1} A_k Q_k$,

(ii) $A_{k+1} = P_k^{-1} A_1 P_k$,

(iii) $A^k = P_k U_k$.

Beweis.

zu (i) Aus $A_k = Q_k R_k$ und $A_{k+1} = R_k Q_k$ folgt

$$Q_k^{-1} A_k Q_k = \underbrace{Q_k^{-1} Q_k}_{=E_n} R_k Q_k = R_k Q_k = A_{k+1}.$$

zu (ii) Rekursiv erhält man

$$A_{k+1} = Q_k^{-1} A_k Q_k = \ldots = (Q_1 \cdot \ldots \cdot Q_k)^{-1} A_1 (Q_1 \cdot \ldots \cdot Q_k) = P_k^{-1} A_1 P_k.$$

zu (iii) Nach (b) gilt $P_{k-1} A_k = A P_{k-1}$. Wir erhalten darauf eine Zerlegung

$$\begin{aligned} P_k U_k &= Q_1 \cdot \ldots \cdot Q_{k-1} \underbrace{Q_k R_k}_{=A_k} R_{k-1} \cdot \ldots \cdot R_1 \\ &= P_{k-1} A_k U_{k-1} = A P_{k-1} U_{k-1} \\ &= \ldots = A^{k-1} P_1 U_1 = A^{k-1} \underbrace{Q_1 R_1}_{=A_1 = A} = A^k \end{aligned}$$

Damit sind alle Behauptungen gezeigt. □

(18.4) Satz. (Konvergenz des QR-Verfahrens) Die reelle Matrix A habe betragsmäßig verschiedene Eigenwerte λ_i mit $0 < |\lambda_n| < \ldots < |\lambda_1|$. Der Faktor R_k in der Zerlegung $A_k = Q_k R_k$ habe positive Diagonalelemente. Dann gibt es eine Permutation P der Eigenwerte λ_i,

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

mit

$$\lim_{k\to\infty} A_k = \begin{pmatrix} \mu_1 & & * \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix}.$$

Beweis. Die Eigenvektoren von A seien $x_1, \ldots, x_n$. Mit $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n\times n}$ gilt

$$A = XDY, \quad Y = X^{-1}, \quad D = \operatorname{diag}(\lambda_i)$$

(vgl. Diagonalisierung). Betrachte die Dreieckszerlegung von Y, d.h. $PY = LU$ mit P Permutationsmatrix, L linke untere Dreiecksmatrix mit $L_{ii} = 1$ für $i = 1, \ldots, n$ sowie U rechte obere Dreiecksmatrix. Es gelte oBdA $P = E_n$. Die QR-Zerlegung von X laute $X = QR$. Dann folgt

$$A^k = XD^kY = QRD^kLU = QR(D^kLD^{-k})D^kU.$$

Mit der Voraussetzung $0 < |\lambda_n| < \ldots < |\lambda_1|$ folgt

$$D^kLD^{k-1} = E_n + I_k, \quad (I_k)_{i,j} = \mathcal{O}\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right) \text{ für } i > j \tag{18.5}$$

Achtung: I_k hängt von k ab und ist **nicht** die Einheitsmatrix. Es gilt $\lim_{k\to\infty} I_k = 0$. Man erhält eine Zerlegung

$$A_k = QR(E_n + I_k)D^kU = Q(E_n + RI_kR^{-1})RD^kU = Q(E_n + F_k)RD^kU,$$

wobei $F_k := RI_kR^{-1} \to 0$ für $k \to \infty$. Weiterhin erhalten wir eine QR-Zerlegung $E_n + F_k = \tilde{Q}_k\tilde{R}_k$ mit $\tilde{Q}_k \to E_n$ und $\tilde{R}_k \to E_n$ für $k \to \infty$. Insgesamt ergibt sich

$$A^k = (Q\tilde{Q}_k)(\tilde{R}_kRD^kU).$$

Sei nun weiter $D = |D|D_1$ mit $D_1^2 = E_n$, $|D| = \operatorname{diag}(|\lambda_i|)$ und $U = D_2(D_2^{-1}U)$ mit $D_2^2 = E_n$, $D_2 = \operatorname{diag}(\pm 1)$. Die Matrizen $|D|$ und $D_2^{-1}U$ haben positive Diagonalelemente. Es gilt

$$A^k = Q\tilde{Q}_kD_2D_1^k\left((D_2D_1^k)^{-1}\tilde{R}_kR(D_2D_1^k)|D|D_2^{-1}U\right).$$

Aus den Lemmata (18.2) und (18.3) folgt mit der Voraussetzung, dass P_k positive Diagonalelemente hat:

$$P_k := Q\tilde{Q}_kD_2D_1^k, \quad U_k := (D_2D_1^k)^{-1}\tilde{R}_kR(D_2D_1^k)|D|D_2^{-1}U.$$

Dies gilt wegen der Eindeutigkeit der Zerlegung $A^k = P_kU_k$. Nach Definition von P_k gilt

$$\begin{aligned} Q_k &= P_{k-1}^{-1}P_k = D_1^{-(k-1)}D_2^{-1}\tilde{Q}_{k-1}^{-1}\underbrace{Q^{-1}Q}_{=E_n}\tilde{Q}_kD_2D_1^k \\ &= D_1 + D_1^{-k+1}D_2^{-1}(\tilde{Q}_{k-1}\tilde{Q}_k - E_n)D_2D_1^k. \end{aligned}$$

Dies konvergiert für $k \to \infty$ gegen D_1, da dann $\tilde{Q}_k \to E_n$ gilt. Ebenso folgt

$$\operatorname{diag}(R_k) = \operatorname{diag}(U_kU_{k-1}^{-1}) = \operatorname{diag}(\tilde{R}_k)\operatorname{diag}(\tilde{R}_{k-1}^{-1}) \cdot |D|.$$

Dies konvergiert für $k \to \infty$ gegen $|D|$, da dann $\tilde{R}_k \to E_n$ gilt. Insgesamt erhalten wir

$$A_k = Q_kR_k \overset{k\to\infty}{\longrightarrow} D_1 \cdot |D| \cdot \begin{pmatrix} 1 & & * \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

für $k \to \infty$. Es ergibt sich lineare Konvergenz: $\mathcal{O}\left(\left|\frac{\lambda_i}{\lambda_j}\right|\right)$ für $i > j$. □

Anwendung von Shift-Techniken

Für ein $s_k \in \mathbb{R}$ betrachtet man eine Zerlegung

$$A_k - s_k E_n = Q_k R_k$$

und definiert damit eine neue Iteration durch

$$A_{k+1} := R_k Q_k + s_k E_n.$$

Übliche Shift-Techniken sind

(i) $s_k = a_{nn}^{(k)}$, falls $A_k = (a_{ij}^{(k)})$.

(ii) s_k sei derjenige Eigenwert λ von $\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix}$, für den $|\lambda - a_{nn}^{(k)}|$ am kleinsten ist.

Man kann zeigen: Für unzerlegbare symmetrische Tridiagonalmatrizen konvergiert das QR-Verfahren mit Shift-Technik im Gegensatz zum normalen QR-Verfahren sogar quadratisch.

Ergänzung: Sind mehrere Eigenwerte betragsgleich, d.h. gilt

$$|\lambda_1| > \ldots > |\lambda_r| = \ldots = |\lambda_s| > \ldots > |\lambda_n|,$$

so gilt eine modifizierte Konvergenzaussage: (mit $P = E_n$)

$$A \overset{k}{\to} \begin{pmatrix} \lambda_1 & & & & & & * \\ & \ddots & & & & & \\ & & \lambda_{r-1} & & & & \\ & & & A_k^{r,s} & & & \\ & & & & \lambda_{s+1} & & \\ & & & & & \ddots & \\ 0 & & & & & & \lambda_n \end{pmatrix}.$$

Die Elemente von $A_k^{r,s}$ konvergieren im Allgemeinen **nicht**. Jedoch konvergieren die Eigenwerte von $A_k^{r,s}$ gegen $\lambda_r, \ldots, \lambda_s$. Dieser Fall tritt bei reellen, symmetrischen Matrizen auf.

§ 19 Eigenwert-Abschätzungen

Sei $||\cdot||$ eine Norm im $\mathbb{R}^n$ oder $\mathbb{C}^n$ und sei $||A|| = \max\{||Ax|| \mid ||x|| = 1\}$ die zugeordnete Matrix-Norm. Eine einfache Eigenwert-Abschätzung ist

$$\rho(A) = \max\{|\lambda| \mid \lambda \text{ ist Eigenwert von } A\} \leq ||A||\,.$$

Eine bessere Abschätzung liefert das folgende

(19.1) Lemma. Sei B eine (n,n)-Matrix. Dann gilt für alle Eigenwerte λ von A, die **nicht** Eigenwerte von B sind, die Abschätzung

$$1 \leq \big|\big|(\lambda E_n - B)^{-1}(A-B)\big|\big| \leq \big|\big|(\lambda E_n - B)^{-1}\big|\big| \cdot ||A-B||$$

Beweis. Sei $Ax = \lambda x$ mit $x \neq 0 \in \mathbb{R}^n$. Durch Subtraktion von Bx auf beiden Seiten erhalten wir

$$(A-B)x = (\lambda E_n - B)x.$$

Da λ kein Eigenwert von B ist, existiert auf der rechten Seite die inverse Matrix und es gilt

$$(\lambda E_n - B)^{-1}(A-B)x = x.$$

Es sei nun oBdA $||x|| = 1$. Dann folgt insgesamt

$$1 \leq \big|\big|(\lambda E_n - B)^{-1}(A-B)\big|\big| \leq \big|\big|(\lambda E_n - B)^{-1}\big|\big| \cdot ||A-B||$$

Das war zu zeigen. □

Folgerung: Man wähle

$$B = A_D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix} = \operatorname{diag}(a_{ii}).$$

Mit $\lambda \neq a_{ii}$ für $i = 1, \ldots, n$ folgt

$$1 \leq \big|\big|(\lambda E_n - A_D)^{-1}(A - A_D)\big|\big|_\infty = \max_{1 \leq i \leq n} \frac{1}{|\lambda - a_{ii}|} \sum_{k \neq i} |a_{ik}|.$$

Darauf folgt dann insbesondere

$$|\lambda - a_{ii}| \leq \sum_{k \neq i} |a_{ik}|$$

für ein i.

(19.2) Satz (Abschätzung von Gerschgorin). Es gelten die folgenden Aussagen:

(i) Die Vereinigung aller Kreisscheiben $K_i := \left\{\lambda \mid |\lambda - a_{ii}| \leq \sum_{k \neq i} |a_{ik}|\right\}$ für $i = 1, \ldots, n$ enthält alle Eigenwerte von A.

(ii) Ist die Vereinigung M_1 von k Kreisen K_i disjunkt von der Vereinigung M_2 der übrigen Kreise K_j, so enthält M_1 genau k und M_2 genau $n-k$ Eigenwerte von A.

Beweis.

(i) vgl. Folgerung.

(ii) **Idee:** Sei $A = A_D + A_R$ mit $A_D = \operatorname{diag}(a_{ii})$. Setze nun $A_t := A_D + t \cdot A_R, \quad 0 \leq t \leq 1$. Es folgt dann $A_0 = A_D$ und $A_1 = A$.

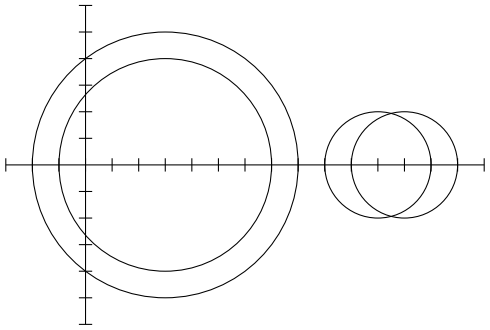
Für $t = 0$ ist die Behauptung richtig, wie man leicht sieht. Da die Eigenwerte von A_t stetig bezüglich t sind, folgt mit $t = 1$ die Behauptung aus Stetigkeitsgründen.

Damit sind alle Behauptungen gezeigt. □

Beispiel. Es sei

$$A = \begin{pmatrix} 3 & 2 & 1 & -2 \\ 1 & 11 & 0 & 1 \\ -1 & 0 & 12 & -1 \\ -3 & 1 & 0 & 3 \end{pmatrix}, \quad r_i := \sum_{k \neq i} |a_{ik}|.$$

Dann gilt $r_1 = 5, r_2 = 2, r_3 = 2, r_4 = 4$.



Verbesserung durch Skalierung: Man betrachtet statt A nun eine Matrix A' mit

$$A' = D^{-1}AD, \quad D = \operatorname{diag}(d_i), \quad d_i > 0.$$

Es folgt dann

$$K_i' = \left\{ \lambda \mid |\lambda - a_{ii}| \leq \sum_{k \neq i} |a_{ik}| \cdot \frac{d_k}{d_i} \right\}.$$

Durch geeignete Wahl von d_i, d_k kann $r_i' := \sum_{k \neq i} |a_{ik}| \frac{d_k}{d_i}$ verkleinert werden.

Kondition des Eigenwertproblems

Wir betrachten zunächst ein Beispiel für schlechte Konditionierung. Es sei

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

Für $\varepsilon \neq 0$ ist diese Matrix nicht symmetrisch. Betrachte die folgenden zwei Fälle:

$\varepsilon = 0$**.** A_0 hat einen Eigenwert $\lambda = 0$ und einen Eigenvektor $x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

$\varepsilon > 0$**.** A_ε hat zwei Eigenwerte $\lambda_1 = \sqrt{\varepsilon}$, $\lambda_2 = -\sqrt{\varepsilon}$ und reelle Eigenvektoren $x_1(\varepsilon) = \begin{pmatrix} \sqrt{\varepsilon} \\ 1 \end{pmatrix}$, $x_2(\varepsilon) = \begin{pmatrix} -\sqrt{\varepsilon} \\ 1 \end{pmatrix}$.

Wir stellen fest, dass gilt

$$\Delta A = A_\varepsilon - A_0 = \mathcal{O}(\varepsilon),$$

aber

$$\Delta\lambda_i = \lambda_i(\varepsilon) - \lambda_i(0) = \pm\sqrt{\varepsilon} = \mathcal{O}(\sqrt{\varepsilon}), \quad \Delta x_i = x_i(\varepsilon) - x_i(0) = \mathcal{O}(\sqrt{\varepsilon})$$

jeweils für $i = 1, 2$. Es folgt

$$\frac{|\Delta\lambda|}{||\Delta A||} = \mathcal{O}(\varepsilon^{-\frac{1}{2}}),$$

was für kleine ε sehr groß wird. Daher ist dieses spezielle Eigenwertproblem schlecht konditioniert.

Definition. Für $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ sei

$$|x| := \begin{pmatrix} |x_1| \\ \vdots \\ |x_n| \end{pmatrix}.$$

Eine Norm $||\cdot||$ im $\mathbb{R}^n$ heißt absolute Norm, wenn gilt

$$||| x ||| = ||x|| \; \forall x \in \mathbb{R}^n.$$

Beispielsweise sind $||\cdot||_\infty$ und $||\cdot||_2$ absolute Normen. Man überlege sich, dass für absolute Normen

$$||\mathrm{diag}(d_1, \ldots, d_n)|| = \max_{1 \le i \le n} |d_i|$$

gilt.

(19.3) Satz (Störungssatz). A und B seien (n,n)-Matrizen. Dabei sei B diagonalisierbar mit $B = PDP^{-1}$ und $D = \mathrm{diag}(\lambda_1(B), \ldots, \lambda_n(B))$. Dann gibt es zu jedem Eigenwert $\lambda(A)$ von A einen Eigenwert $\lambda(B)$ von B mit

$$| \lambda(A) - \lambda(B) | \le \mathrm{cond}(P) \cdot ||A - B||$$

für absolute Normen $||\cdot||$ im $\mathbb{R}^n$.

Beweis. Für $\lambda(A) \neq \lambda_i(B)$, $i = 1, \ldots, n$, gilt

$$\begin{aligned} \left|\left|(\lambda(A)E_n - B)^{-1}\right|\right| &= \left|\left|P(\lambda(A)E_n - D)^{-1}P^{-1}\right|\right| \\ &\le ||P|| \cdot \left|\left|P^{-1}\right|\right| \cdot \left|\left|(\lambda(A)E_n - D)^{-1}\right|\right| \\ &= \mathrm{cond}(P) \cdot \max_{1 \le i \le n} \frac{1}{| \lambda(A) - \lambda_i(B) |} \\ &= \mathrm{cond}(P) \cdot \frac{1}{\min_{1 \le i \le n} | \lambda(A) - \lambda_i(B) |} \end{aligned}$$

Wegen $1 \le \left|\left|(\lambda(A)E_n - B)^{-1}\right|\right| \cdot ||A - B||$ nach (19.1) folgt, dass es ein $i \in \{1, \ldots, n\}$ gibt mit

$$| \lambda(A) - \lambda_i(B) | \le \mathrm{cond}(P) \cdot ||A - B||$$

□

Folgerung. Ist B eine normale Matrix, dann kann P mit $\mathrm{cond}_2(P) = 1$ gewählt werden. Es gilt dann die Abschätzung

$$| \lambda(A) - \lambda(B) | \le ||A - B||_2 \tag{19.4}$$

für ein $\lambda(B)$. Für symmetrische Matrizen ist das Eigenwertproblem gut konditioniert, d.h. für symmetrische Matrizen A und ΔA gilt

$$| \lambda(A + \Delta A) - \lambda(A) | \le ||\Delta A||_2 .$$

Sei A eine nicht-symmetrische Matrix. Wir betrachten eine Störung $A \to A + \varepsilon C$ mit einer (n,n)-Matrix C. Man kann zeigen, dass eine Abschätzung

$$| \lambda(A + \varepsilon C) - \lambda(A) | \le K \cdot |\varepsilon|^{\frac{1}{\nu}} \tag{19.5}$$

gilt, wobei ν die maximale Dimension eines zu $\lambda(A)$ gehörenden JORDAN-Blockes ist. Als eine weitere Folgerung erhalten wir, dass für Diagonalmatrizen gilt

$$\min_{||x||_2 = 1} ||Dx||_2 = \min_{1 \le i \le n} |d_i|.$$

Es sei nun A eine normale Matrix. Dann gibt es eine orthogonale Matrix U mit

$$A = U^* D U, \quad D = \mathrm{diag}(\lambda_i(A)).$$

Sei $f(\lambda)$ ein Polynom, d.h. $f(\lambda) = \alpha_0 + \alpha_1\lambda + \ldots + \alpha_k\lambda_k$. Das entsprechende Matrix-Polynom lautet dann $f(A) = \alpha_0 + \alpha_1 A + \ldots + \alpha_k A^k$. Man überlegt sich leicht, das gilt $f(A) = f(U^*DU) = U^* f(D) U$. Wegen $||Ux||_2 = ||x||_2$ folgt für $||x||_2 = 1$:

$$||f(A)x||_2 = ||U^* f(D) U x||_2 = ||f(D)Ux||_2 \ge \min_{||y||_2 = 1} ||f(D)y||_2 = \min_{1 \le i \le n} | f(\lambda_i(A)) | .$$

(19.6) Satz. Sei A normal und $x \in \mathbb{R}^n$ mit $||x||_2 = 1$. Sei $f(\lambda)$ ein beliebiges Polynom. Dann gibt es einen Eigenwert $\lambda(A)$ von A mit

$$| f(\lambda(A)) | \leq ||f(A)x||_2 .$$

Folgerung. Sei $x \in \mathbb{R}^n$ mit $||x||_2 = 1$ und sei $f(\lambda)$ das lineare Polynom $f(\lambda) = \lambda - x^*Ax$. Dann gilt

$$\begin{aligned} ||f(A)x||_2^2 &= ||(A - x^*AxE_n)x||_2^2 = x^*(A^* - x^*AxE_n)(A - x^*AxE_n)x \\ &= x^*A^*Ax - (x^*A^*x)(x^*Ax) \end{aligned}$$

Es folgt: Es gibt $\lambda(A)$ von A mit

$$| \lambda(A) - x^*Ax |^2 \leq x^*A^*Ax - (x^*A^*x)(x^*Ax).$$

Insbesondere erhält man für symmetrische Matrizen den folgenden

(19.7) Satz (Bogolyubov, Krylov, Weinstein). Ist A symmetrisch und $x \in \mathbb{R}^n$ mit $||x||_2 = 1$, so gilt

$$\min_{1 \leq i \leq n} | \lambda_i(A) - x^*Ax | \leq \sqrt{(x^*A^2x) - (x^*Ax)^2}.$$

Statistische (Quantenmechanische) Deutung

x seien Zustände, ohne Einschränkung $||x||_2 = 1$. Weiterhin seien $\lambda_i(A)$ Messwerte und x^*Ax sei der Erwartungswert des "Operators" A. Dieser hat den Wert λ, falls x ein Eigenvektor von A zum Eigenwert λ ist. Dann ist

$$\sqrt{(x^*A^2x) - (x^*Ax)^2} = ||A - (x^*Ax)x||_2$$

die Unschärfe von A bezüglich x. Diese hat den Wert 0 für jeden Eigenvektor x. In diesem Fall besagt die Abschätzung aus (19.7) grob gesprochen

$$| \text{Messwert - Erwartungswert} | \leq \text{ Unschärfe.}$$

VI Lineare Ausgleichsprobleme

§ 20 Approximation in normierten Räumen

§ 20.1 Funktionalanalytische Grundlagen

Definition. Sei V ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$. Eine Norm $||\cdot||$ für V ist eine Abbildung $||\cdot|| : V \to \mathbb{R}$ oder $\mathbb{C}$ mit den Eigenschaften

(i) $||f|| > 0$ für alle $f \in V$ mit $f \neq 0$.

(ii) $||\alpha f|| = |\alpha| \cdot ||f||$ für alle $f \in V$ und $\alpha \in \mathbb{R}$ oder $\mathbb{C}$.

(iii) $||f + g|| \leq ||f|| + ||g||$ für alle $f, g \in V$.

Ein Paar $(V, ||\cdot||)$ heißt normierter Vektorraum. Eine Norm heißt streng oder strikt, wenn gilt

$$||f + g|| = ||f|| + ||g|| \Rightarrow f, g \text{ sind linear abhängig}$$

Beispiel. Sei $V = C[a, b] = \{f : [a, b] \to \mathbb{R} \text{ oder } \mathbb{C} \mid f \text{ stetig } \}$. Für alle Zahlen $1 \leq p < \infty$ ist dann

$$||f||_p := \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}$$

eine Norm für V. Man kann zeigen, dass gilt

$$\lim_{p \to \infty} ||f||_p = ||f||_\infty := \max_{a \leq x \leq b} |f(x)|.$$

Die Norm $||f||_2$ ist streng aufgrund der Cauchy-Schwarz-Ungleichung. Die Norm $||f||_\infty$ ist nicht streng. Beispielsweise gilt für $f(x) \equiv 1$ und $g(x) = x$, jeweils definiert auf dem Intervall $[0, 1]$,

$$||f + g||_\infty = \max_{0 \leq x \leq 1} (1 + x) = 2 = ||f||_\infty + ||g||_\infty ,$$

aber f und g sind offensichtlich linear unabhängig.

Eine Folge $\{f_n\}_{n \in \mathbb{N}} \subset V$ heißt Cauchy-Folge, wenn es zu jedem $\varepsilon > 0$ ein $n(\varepsilon) \in \mathbb{N}$ gibt mit

$$||f_k - f_l|| < \varepsilon \text{ für alle } k, l \geq n(\varepsilon).$$

Konvergiert jede Cauchy-Folge eines normierten Raumes $(V, ||\cdot||)$ gegen ein Element von V, so heißt V vollständig. Ein solcher vollständig normierter Vektorraum heißt Banach-Raum. Man kann zeigen: Jeder endlich-dimensionale normierte Raum ist ein Banach-Raum.

Beispiele. Der Raum $(C[a, b], ||\cdot||_\infty)$ ist vollständig, also ein Banach-Raum, während der Raum $(C[a, b], ||\cdot||_2)$ nicht vollständig ist. Dessen "Vervollständigung" ist der Raum $(L^2[a, b], ||\cdot||_2)$.

Definition. Eine bilineare Abbildung $(\cdot, \cdot) : V \times V \to \mathbb{R}$ oder $\mathbb{C}$ heißt Inneres Produkt, wenn für alle $f, g, h \in V$ und $\alpha \in \mathbb{R}$ oder $\mathbb{C}$ folgende Eigenschaften gelten:

(i) $(f + g, h) = (f, h) + (g, h)$ (Linearität)

(ii) $(\alpha f, g) = \alpha (f, g)$ (Homogenität)

(iii) $(f, g) = \overline{(g, f)}$ (Symmetrie)

(iv) $(f, f) > 0$ für alle $f \neq 0$.

Durch $||f|| := \sqrt{(f, f)}$ wird eine Norm für V definiert. Es diese gilt die Cauchy-Schwarz-Ungleichung

$$| (f, g) | \leq ||f|| \cdot ||g||$$

für alle $f, g \in V$. Gleichheit gilt genau dann, wenn f und g linear abhängig sind. *Beweisidee: Betrachte* $(\alpha f + g, \alpha f + g) \geq 0$ *für* $\alpha = -\frac{(g,f)}{(f,f)}$.

Bemerkung. Die Dreiecksungleichung für $||f|| = \sqrt{(f,f)}$ folgt mit CAUCHY-SCHWARZ.

Der normierte Vektorraum $(V, ||\cdot||)$ mit $||f|| = \sqrt{(f,f)}$ heißt Prä-HILBERT-Raum. Ist dieser Raum vollständig, so heißt V HILBERT-Raum. Prä-HILBERT-Räume sind stets streng normierte Räume. Begründung:

$$\begin{aligned} ||f+g||^2 &= (f+g, f+g) \\ &= (f,f) + (g,g) + (f,g) + (g,f) \\ &\leq ||f||^2 + ||g||^2 + 2|(f,g)| \\ &\overset{C.S.}{\leq} ||f||^2 + ||g||^2 + 2\,||f|| \cdot ||g|| \\ &= (||f|| + ||g||)^2 \end{aligned}$$

Gleichheit kann nur dann gelten, wenn schon $|(f,g)| = ||f|| \cdot ||g||$ gilt. Dies ist aber nur dann der Fall, wenn f und g linear abhängig sind.

Beispiele.

(i) Der Raum $(\mathbb{C}^n, ||\cdot||_2)$ ist ein HILBERT-Raum mit

$$(x,y) := \sum_{i=1}^{n} x_i \overline{y_i}, \quad x, y \in \mathbb{C}^n, \quad ||x||_2 := \sqrt{(x,x)}.$$

(ii) Sei $w : [a,b] \to \mathbb{R}$ eine stetige Gewichtsfunktion mit $w(x) > 0$ für $a < x < b$. Ein inneres Produkt auf $C[a,b]$ erhält man durch

$$(f,g) := \int_a^b f(x) g(x) w(x) dx, \quad f, g \in C[a,b].$$

Es folgt dann

$$||f||_2 = \left(\int_a^b f(x)^2 w(x) dx \right)^{\frac{1}{2}}.$$

Der Raum $(C[a,b], ||\cdot||_2)$ ist ein Prä-HILBERT-Raum, aber nicht vollständig!

§ 20.2 Das allgemeine Approximationsproblem

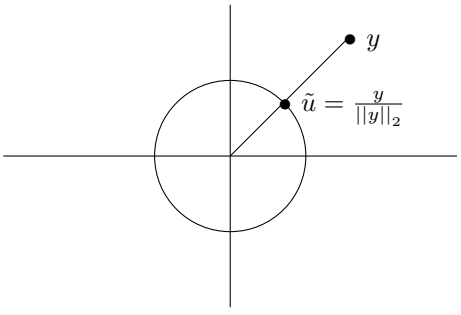
Sei $(V, ||\cdot||)$ ein normierter Raum und sei $T \subset V$ eine Teilmenge. Zu einem Element $v \in V$ suchen wir die beste Näherung oder Approximation (Proximum) $\tilde{u} \in T$ von v bezüglich T. Mit

$$||v - \tilde{u}|| \leq ||v - u|| \text{ für alle } u \in T \tag{20.1}$$

Bemerkung. Existenz und Eindeutigkeit von $\tilde{u} \in T$ hängen wesentlich von der Norm $||\cdot||$ und der Menge T ab.

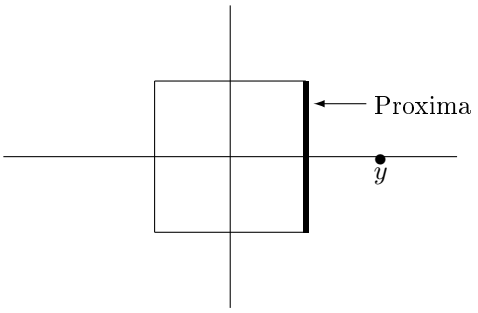
Beispiele.

(i) Sei $V = \mathbb{R}^2$ mit $||\cdot||_2$ und sei $T = \{x \in \mathbb{R}^2 \mid ||x||_2 \leq 1\}$. T ist kompakt und für $y \in \mathbb{R}^2$ mit $||y||_2 > 1$ gibt es ein eindeutiges Proximum $\tilde{u} \in T$, nämlich $\tilde{u} = \frac{y}{||y||_2}$.



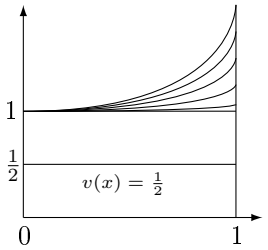
Die Existenz des Proximum folgt aus der Kompaktheit, die Eindeutigkeit aus der strengen Konvexität von T.

(ii) Sei $(V, ||\cdot||_\infty)$, $V = \mathbb{R}^2$, $||x||_\infty := \max\{|x_1|, |x_2|\}$ gegeben und sei $T := \{x \in \mathbb{R}^2 \mid ||x||_\infty \leq 1\}$. Dann ist T kompakt. Für $y = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \in \mathbb{R}^2$ ist das Proximum $\tilde{u} \in T$ nicht eindeutig bestimmt, denn der Abstand $||x - y||_\infty = \max\{|x_1 - 2|, |x_2|\}$ wird minimal für alle Punkte der Kante $\{(1, x_2) \mid |x_2| \leq 1\}$.



Die nicht-Eindeutigkeit resultiert daraus, dass $||\cdot||_\infty$ keine strenge Norm ist.

(iii) Sei $V = C[0,1]$, $||\cdot|| = ||\cdot||_\infty$, $T = \{u \in V \mid u(x) = e^{\beta x},\ \beta > 0\}$. Gesucht ist ein Proximum $\tilde{u} \in T$ an die Funktion $v(x) \equiv \frac{1}{2}$, $x \in [0,1]$.



Für $u(x) = e^{\beta x}$ gilt

$$||u - v||_\infty = \max_{0 \leq x \leq 1} \left| e^{\beta x} - \frac{1}{2} \right| = e^\beta - \frac{1}{2}$$

Wegen $\beta > 0$ wird ein Minimum **nicht** angenommen. Begründung: T ist nicht kompakt.

Definition. Zu $v \in V$ heißt

$$e_T(v) = \inf_{u \in T} ||v - u|| \tag{20.2}$$

Minimalabstand von v zu T. Ein Proximum $\tilde{u} \in T$ genügt

$$e_T(v) = ||v - \tilde{u}|| \,.$$

Nach Definition von $e_T(v)$ gibt es eine Minimalfolge $\{u_n\} \subset T$ mit $e_T(v) = \lim_{n \to \infty} ||v - u_n||$.

(20.3) Lemma. Es gilt:

(i) Jede Minimalfolge ist beschränkt.

(ii) Jeder in T liegende Häufungspunkt einer Minimalfolge ist ein Proximum.

Beweis. **zu (i)**: Es gibt wegen $e_T(v) = \lim_{n \to \infty} ||v - u_n||$ ein $n_0 \in \mathbb{N}$ mit $e_T(v) \leq ||v - u_n|| \leq e_T(v) + 1$ für alle $n \geq n_0$. Es folgt

$$||u_n|| \leq ||u_n - v|| + ||v|| \leq e_T(v) + 1 + ||v|| =: K_2$$

für alle $n \geq n_0$. Setze

$$K_1 := \max_{n < n_0} \{||u_n||\}.$$

Dann gilt $||u_n|| \leq K$ mit $K = \max\{K_1, K_2\}$ für alle $n \in \mathbb{N}$.

zu (ii): Es gelte $\tilde{u} = \lim_{n\to\infty} u_{k(n)} \in T$. Wir erhalten die Abschätzung

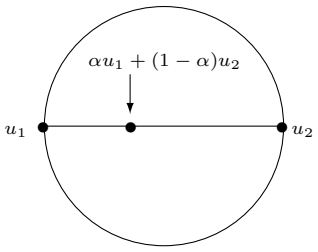
$$||v - \tilde{u}|| \leq \underbrace{\left|\left|v - u_{k(n)}\right|\right|}_{\overset{n\to\infty}{\longrightarrow} e_T(v)} + \underbrace{\left|\left|u_{k(n)} - \tilde{u}\right|\right|}_{\overset{n\to\infty}{\longrightarrow} 0}.$$

Da der Ausdruck links des Gleichheitszeichens von n unabhängig ist, gilt bereits $||v - \tilde{u}|| = e_T(v)$, also ist $\tilde{u} \in T$ ein Proximum von v bezüglich T. □

Definition. Die Menge T heißt streng konvex, wenn für alle $u_1, u_2 \in T$ mit $u_1 \neq u_2$ gilt

$$\alpha u_1 + (1-\alpha)u_2 \in \mathring{T}$$

für alle $\alpha \in]0, 1[$.



(20.4) Satz (1. Existenz- und Eindeutigkeitssatz). Sei $T \subset V$ eine kompakte Teilmenge. Dann gibt es zu jedem $v \in V$ ein Proximum $\tilde{u} \in T$. Ist außerdem T streng konvex, so ist $\tilde{u} \in T$ eindeutig bestimmt.

Beweis. In zwei Schritten:

Existenz: Eine Minimalfolge $\{u_n\} \subset T$ enthält wegen der Kompaktheit von T einen Häufungspunkt $\tilde{u} \in T$. Nach dem Lemma (20.3) (ii) ist $\tilde{u} \in T$ ein Proximum von v bezüglich T.

Eindeutigkeit: Seien $u_1, u_2 \in T$ Proxima von v bezüglich T mit $u_1 \neq u_2$. Es folgt

$$\left|\left|\frac{1}{2}(u_1 + u_2) - v\right|\right| \leq \frac{1}{2}\underbrace{||u_1 - v||}_{=e_T(v)} + \frac{1}{2}\underbrace{||u_2 - v||}_{=e_T(v)} = e_T(v).$$

Da T als streng konvex vorausgesetzt wurde, gilt $\frac{1}{2}(u_1 + u_2) \in \mathring{T}$. Betrachte nun $\tilde{u} := \frac{1}{2}(u_1 + u_2) + \alpha(v - \frac{1}{2}(u_1 - u_2)) \in T$ für $\alpha > 0$ genügend klein. Damit erhalten wir

$$||v - \tilde{u}|| = \left|\left|\frac{1}{2}(1-\alpha)(u_1 + u_2) - (1-\alpha)v\right|\right| = (1-\alpha)\underbrace{\left|\left|\frac{1}{2}(u_1 + u_2) - v\right|\right|}_{=e_T(v)} < e_T(v)$$

für $\alpha > 0$ genügend klein. Widerspruch zur Optimalität von u_1 bzw. u_2! Folglich muss schon $u_1 = u_2$ gelten. □

(20.5) Satz (2. Existenz- und Eindeutigkeitssatz). Sei $U \subset V$ ein endlich-dimensionaler Unterraum von V. Dann gibt es zu jedem $v \in V$ ein Proximum $\tilde{u} \in U$. Ist außerdem $(V, ||\cdot||)$ streng normiert, dann ist $\tilde{u} \in U$ eindeutig bestimmt.

Beweis. In zwei Schritten:

Existenz: Nach Lemma (20.3) (i) ist jede Minimalfolge beschränkt und besitzt daher einen Häufungspunkt $\tilde{u}$. Nun ist $U \subset V$ endlich-dimensional und somit abgeschlossen, also gilt schon $\tilde{u} \in U$. Somit ist $\tilde{u}$ nach (20.3) (ii) ein Proximum.

Eindeutigkeit: Sei $v \in V$ und seinen $u_1, u_2 \in U$ Proxima von v bezüglich U. Wie im Beweis von (20.4) folgt

$$\left|\left|\frac{1}{2}(u_1 + u_2) - v\right|\right| = e_U(v)$$

und damit

$$||(v - u_1) + (v - u_2)|| = 2e_U(v) = ||v - u_1|| + ||v - u_2||\,.$$

Da $||\cdot||$ als streng vorausgesetzt ist, folgt

$$v - u_1 = \alpha(v - u_2)$$

für ein $\alpha \in \mathbb{R}$, d.h. $v - u_1$ und $v - u_2$ sind linear abhängig und wir erhalten

$$(1 - \alpha)v = u_1 - \alpha u_2 \in U.$$

Für den nicht-trivialen Fall $v \notin U$ kann dies aber nur für $\alpha = 1$ erfüllt sein, woraus direkt $u_1 = u_2$ folgt. □

Basisdarstellung

Es gilt $U = \operatorname{span}(u_1, \ldots, u_n) = \left\{\sum_{k=1}^n \alpha_k u_k \middle| \alpha_k \in \mathbb{R} \text{ oder } \mathbb{C}\right\}$. Betrachte die Funktion

$$\text{(20.6a)} \quad F(\alpha) := \left|\left|v - \sum_{k=1}^n \alpha_k u_k\right|\right|, \quad \alpha = \begin{pmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{pmatrix} \in \mathbb{R}^n \text{ oder } \mathbb{C}^n.$$

Die Approximationsaufgabe (20.1) ist für $T = U$ äquivalent zu einer Optimierungsaufgabe

$$\text{(20.6b)} \quad \text{Bestimme } \tilde{\alpha} \in \mathbb{R}^n \text{ oder } \mathbb{C}^n \text{ mit } F(\tilde{\alpha}) = \min_{\alpha \in \mathbb{R}^n/\mathbb{C}^n} F(\alpha).$$

Im Fall $V = C[a, b]$ mit $[a, b] \subset \mathbb{R}$ heißt (20.6b)

- Chebyshev-Problem für $||\cdot|| = ||\cdot||_\infty$, vgl. Remez-Algorithmus.
- Gauss-Approximation für $||\cdot|| = ||\cdot||_2$, Approximation im quadratischen Mittel.

§ 21 Approximation in Prä-HILBERT-Räumen

Sei V ein Prä-HILBERT-Raum mit dem inneren Produkt $(\cdot,\cdot)$ und der durch $||f|| := \sqrt{(f,f)}$, $f \in V$, induzierten Norm. Sei $U = \text{span}(u_1, \ldots, u_n) \subset V$ ein endlich-dimensionaler Unterraum. Zu $f \in V$ gibt es nach Satz (20.5) genau ein $\tilde{u} \in U$ mit

$$||f - \tilde{u}|| = \min_{u \in U} ||f - u|| .$$

Dies ist äquivalent zum Optimierungsproblem

$$||f - \tilde{u}||^2 = \min_{u \in U} ||f - u||^2 = \min_{u \in U} (f - u, f - u) .$$

Mit einer Basisdarstellung $u = \sum_{k=1}^n \alpha_k u_k$ lautet das Optimierungsproblem

$$\min_{\alpha \in \mathbb{R}^n} \left|\left| f - \sum_{k=1}^n \alpha_k u_k \right|\right|^2 , \quad \alpha = (\alpha_1, \ldots, \alpha_n)^t \in \mathbb{R}^n.$$

(21.1) Satz (Orthogonalität und Normalgleichungen). $\tilde{u} \in U$ ist genau dann ein Proximum an $f \in V$, wenn

$$(f - \tilde{u}, u) = 0$$

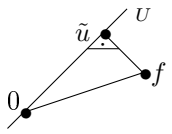
für alle $u \in U$ gilt. In der Basisdarstellung $\tilde{u} = \sum_{k=1}^n \tilde{\alpha_k} u_k$ sind die Koeffizienten $\tilde{\alpha_k}$ die eindeutig bestimmte Lösung der Normalgleichungen

$$\sum_{i=1}^n \tilde{\alpha_i} (u_i, u_k) = (f, u_k) \; \forall k = 1, \ldots, n.$$

Die Abweichung (Residuum) erfüllt

$$||f - \tilde{u}||^2 = ||f||^2 - ||\tilde{u}||^2 = ||f||^2 - \sum_{k=1}^n \tilde{\alpha_k} (f, u_k) .$$

Beweis. Nach Pythagoras erhält man folgendes Schema:



(V)

Es gelte $(f - \tilde{u}, u) = 0$ für alle $u \in U$. Wegen $\tilde{u} \in U$ ist auch $u - \tilde{u} \in U$. Daher gilt $(f - \tilde{u}, u - \tilde{u}) = 0$ für alle $u \in U$ und wir folgern

$$\begin{aligned} ||f - u||^2 &= (f - u, f - u) = (f - \tilde{u} + \tilde{u} - u, f - \tilde{u} + \tilde{u} - u) \\ &= (f - \tilde{u}, f - \tilde{u}) + (\tilde{u} - u, \tilde{u} - u) \\ &= ||f - \tilde{u}||^2 + ||u - \tilde{u}||^2 \geq ||f - \tilde{u}||^2 \end{aligned}$$

Die Eindeutigkeit dieses Residuums folgt direkt, da Gleichheit bereits $\tilde{u} = u$ impliziert. Also ist $\tilde{u}$ das eindeutig bestimmte Proximum von f in U.

Noch zu zeigen ist, dass die Orthogonalitätsrelation für ein $\tilde{u} = \sum_{k=1}^n \tilde{\alpha_k} u_k$ lösbar ist. Die Relation $(f - \tilde{u}, u) = 0 \, \forall u \in U$ besagt

$$\left(f - \sum_{i=1}^n \tilde{\alpha_i} u_i, u_k \right) = 0$$

für alle $k = 1, \ldots, n$. Dies ergibt die Normalgleichungen

$$\sum_{i=1}^n \tilde{\alpha_i} (u_i, u_k) = (f, u_k) \quad \text{für alle } k = 1, \ldots, n.$$

Da $(u_1, \ldots, u_n)$ eine Basis von U ist, folgt, dass die GRAMsche Matrix $G = ((u_i, u_k))_{1 \le i,k \le n}$ regulär, symmetrisch und positiv definit ist. Somit sind die Normalgleichungen eindeutig lösbar. Für die Abweichung erhält man mit $(f - \tilde{u}, \tilde{u}) = 0$ die Aussage

$$||f||^2 = ||f - \tilde{u} + \tilde{u}||^2 = ||f - \tilde{u}||^2 + ||\tilde{u}||^2 \quad \Rightarrow \quad ||f - \tilde{u}||^2 = ||f||^2 - ||\tilde{u}||^2 .$$

Wegen

$$||\tilde{u}||^2 = (\tilde{u} - f + f, \tilde{u}) = \underbrace{(\tilde{u} - f, \tilde{u})}_{=0} + (f, \tilde{u}) = \sum_{k=1}^{n} \tilde{\alpha_k} (f, u_k)$$

folgt auch die letzte Behauptung. □

Beispiel. Es sei $V = C[-1, 1]$, $(f, g) := \int_{-1}^{1} f(x)g(x)dx$ sowie $f(x) = e^x$ und $U = \text{span}(1, x, x^2) = \text{span}(u_0, u_1, u_2)$. Es gilt dann

$$(u_i, u_k) = \left(x^i, x^k\right) = \int_{-1}^{1} x^{i+k} dx = \frac{1 + (-1)^{i+k}}{i + k + 1}$$

und wir erhalten

$$G = ((u_i, u_k))_{0 \le i,k \le 2} = \begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} .$$

Weiterhin ist

$$\begin{aligned} (f, u_0) &= \int_{-1}^{1} e^x dx = e - \frac{1}{e}, \\ (f, u_1) &= \int_{-1}^{1} x e^x dx = \frac{2}{e}, \\ (f, u_2) &= \int_{-1}^{1} x^2 e^x dx = e - \frac{5}{e}. \end{aligned}$$

Wir erhalten die Normalgleichungen

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \cdot \begin{pmatrix} \tilde{\alpha_0} \\ \tilde{\alpha_1} \\ \tilde{\alpha_2} \end{pmatrix} = \begin{pmatrix} e - 1/e \\ 2/e \\ e - 5/e \end{pmatrix} .$$

Als Lösung ergibt sich $\tilde{\alpha_0} = 0.99629$, $\tilde{\alpha_1} = 1.10364$ und $\tilde{\alpha_2} = 0.53672$. Man berechnet

$$\max_{-1 \le x \le 1} \left| e^x - (\tilde{\alpha_0} + \tilde{\alpha_1} x + \tilde{\alpha_2} x^2 \right| \approx 0.082.$$

Im Fall eines Orthonormalsystems (ONS) $u_1, \ldots, u_n$, d.h. wenn gilt

$$(u_i, u_k) = \delta_{i,k} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} ,$$

vereinfachen sich die Normalgleichungen zu

$$\tilde{u} = \sum_{k=1}^{n} (f, u_k) u_k, \; \tilde{\alpha_k} = (f, u_k) . \tag{21.2}$$

Die Koeffizienten $\tilde{\alpha_k} = (f, u_k)$ heißen verallgemeinerte FOURIER-Koeffizienten. Aus der Abweichung

$$||f - \tilde{u}||^2 = ||f||^2 - \sum_{k=1}^{n} \tilde{\alpha_k}^2 \ge 0 \tag{21.3}$$

folgt die sogenannte BESSELsche Ungleichung

$$\sum_{k=1}^{n} \tilde{\alpha_k}^2 \le ||f||^2 . \tag{21.4}$$

(21.5) Definition. Das ONS $\{u_n\}_{n\in\mathbb{N}}$ heißt vollständig, wenn es zu jedem $f \in V$ eine Folge $\{f_n\}_{n\in\mathbb{N}}$ gibt mit

$$f_n \in \operatorname{span}(u_1, \ldots, u_n) \text{ und } \lim_{n\to\infty} ||f - f_n|| = 0.$$

(21.6) Satz (Vollständigkeitsrelation). Notwendig und hinreichend für die Vollständigkeit des ONS $\{u_n\}_{n\in\mathbb{N}}$ ist die Vollständigkeitsrelation (PARSEVAL-Gleichung)

$$\sum_{k=1}^{\infty} \tilde{\alpha_k}^2 = ||f||^2 .$$

Beweis. Mit (21.4) und der Abweichung in (21.1) trivial. □

Trigonometrische Approximation

Gegeben sei eine stückweise stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ mit der Periode 2π, d.h. mit $f(x + 2\pi) = f(x)$ für alle $x \in \mathbb{R}$.

Es sei nun $V := C^{-1}[-\pi, \pi] = \{f : [-\pi, \pi] \to \mathbb{R} \text{ stückweise stetig }\}$ und eine Norm für V gegeben durch $||f|| := \left(\int_{-\pi}^{\pi} f(x)^2 dx\right)^{\frac{1}{2}}$.

Durch Nachrechnen sieht man, dass die Funktionen $u_0, \ldots, u_{2m}$, definiert durch

$$u_0(x) = \frac{1}{\sqrt{2\pi}}, \quad u_{2k-1}(x) = \frac{1}{\sqrt{\pi}} \sin(kx), \quad u_{2k} = \frac{1}{\sqrt{\pi}} \cos(kx)$$

für $k = 1, \ldots, m$, ein ONS in $V = C^{-1}[-\pi, \pi]$ bilden. Das Proximum von $f \in V$ bezüglich $U = \operatorname{span}(u_0, \ldots, u_{2m})$ ist nach (21.2) gegeben durch

$$\left.\begin{aligned} \tilde{u}(x) &= \tfrac{a_0}{2} + \textstyle\sum_{k=1}^{m}(a_k \cos(kx) + b_k \sin(kx)) \\ a_k &= \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} f(x)\cos(kx)\, \forall k = 0, \ldots, m \\ b_k &= \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} f(x)\sin(kx),\ \forall k = 1, \ldots, m \end{aligned}\right\} \tag{21.7}$$

Die BESSELsche Ungleichung (21.4) liefert

$$\frac{a_0^2}{2} + \sum_{k=1}^{m}(a_k^2 + b_k^2) \le \frac{1}{\pi} ||f||_2^2 .$$

Eine Aussage über die Konvergenz der Reihe (21.7) liefert (ohne Beweis) der folgende Satz:

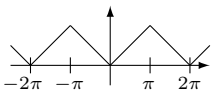
(21.8) Satz. Sei $f \in V = C^{-1}[-\pi, \pi]$ periodisch mit Periode 2π.

(i) Die Folge (21.7) der Proxima von f bezüglich $U = \operatorname{span}(u_0, \ldots, u_{2m})$ konvergiert für $m \to \infty$ im quadratischen Mittel gegen f, d.h. bezüglich $||\cdot||_2$.

(ii) Existiert zusätzlich die Ableitung $f' \in C^{-1}[-\pi, \pi]$, dann konvergiert die Reihe (21.7) punktweise gegen $\lim_{h\to 0+} \frac{1}{2} f(x + h) + f(x - h))$ für alle $x \in [-\pi, \pi]$.

Beispiele.

(i) Für die stetige, gerade Funktion $f(x) = |x|$, $f(x + 2\pi) = f(x)$



berechnet man

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| dx = \frac{2}{\pi}\int_{0}^{\pi} x dx = \pi$$

sowie

$$a_k = \frac{2}{\pi}\int_{0}^{\pi} x\cos(kx) dx = \frac{2}{\pi k^2}\left((-1)^k - 1\right),\ k = 1, 2, \ldots$$

und

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\sin(kx)dx = 0,\ \forall k.$$

Damit erhält man die FOURIER-Reihe

$$|x| \approx \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{\cos((2k+1)x)}{(2k+1)^2}.$$

(ii) Die Funktion

$$f(x) = \begin{cases} -1, & -\pi \le x < 0 \\ 0, & x = 0 \\ 1, & 0 < x \le \pi \end{cases}$$

ist unstetig in $x = 0$ und ungerade. Es folgt $a_k = 0$ für alle k und

$$b_k = \frac{2}{\pi}\int_0^{\pi}\sin(kx)dx = \begin{cases} \frac{4}{\pi k}, & k \text{ gerade} \\ 0, & \text{sonst} \end{cases}.$$

Orthogonalpolynome

Sei $V = C[a,b]$ mit Innerem Produkt $(f,g) := \int_a^b f(x)g(x)\omega(x)dx$ für $f, g \in V$ und einer Gewichtsfunktion $\omega : [a,b] \to \mathbb{R}$ mit $\omega(x) > 0$ für $a < x < b$. Dann ist $V = C[a,b]$ ein Prä-HILBERT-Raum. Wendet man nun das GRAM-SCHMIDTsche Orthogonalisierungsverfahren auf die Monome $u_n(x) = x^n$ an, so erhält man Orthogonalpolynome mit Höchstkoeffizient 1, d.h.

$$\tilde{f}_n \in \tilde{\Pi}_n = \left\{x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \mid a_0, \ldots, a_{n-1} \in \mathbb{R}\right\}, \quad n = 0, 1, \ldots.$$

Als Ergebnis erhalten wir, dass die Polynome $\tilde{p}_0, \ldots, \tilde{p}_n \in \tilde{\Pi}_n$ eine Basis von $\Pi_n = \left\{\sum_{k=0}^n a_k x^k \mid a_k \in \mathbb{R}\right\}$ bilden. Weiterhin genügen die Polynome der Drei-Term-Rekursion

$$\left.\begin{aligned} \tilde{p}_0(x) = 1, \quad \tilde{p}_1(x) = x - \delta_1 \\ \tilde{p}_{n+1}(x) = (x - \delta_{n+1})\tilde{p}_n(x) - \gamma_{n+1}^2\tilde{p}_{n-1}(x) \text{ für } n = 1, 2, \ldots \\ \delta_{n+1} = \frac{(x\tilde{p}_n, \tilde{p}_n)}{(\tilde{p}_n, \tilde{p}_n)}, \quad \gamma_{n+1}^2 = \frac{(\tilde{p}_n, \tilde{p}_n)}{(\tilde{p}_{n-1}, \tilde{p}_{n-1})} \end{aligned}\right\} \tag{21.9}$$

Beweis. Durch Nachrechnen, dass $(\tilde{p}_n, \tilde{p}_k) = 0$ für alle $k < n$ gilt. □

(21.10) Nullstellensatz. Das Orthogonalpolynom $\tilde{p}_n \in \tilde{\Pi}_n$ hat in (a,b) genau n einfache Nullstellen.

Beweis. Seien $x_1, \ldots, x_m$ die Punkte, in denen das Vorzeichen von $\tilde{p}_n$ auf (a,b) wechselt. Zu zeigen ist dann $m = n$. Mit dem Polynom

$$q(x) = \prod_{i=1}^{m}(x - x_i)$$

hat das Polynom $q(x)\tilde{p}_n(x)$ in (a,b) konstantes Vorzeichen. Würde man nun $m < n$ annehmen, so folgt

$$0 = (q, \tilde{p}_n) = \int_a^b q(x)\tilde{p}_n(x)\omega(x)dx$$

wegen $(\tilde{p}_k, \tilde{p}_k) = 0\,\forall k < n$. Daher müsste dann aber aufgrund des konstanten Vorzeichens schon $q(x)\tilde{p}_n(x) = 0$ in (a,b) gelten. Dies kann aber nicht sein, da wegen $m < n$ das Polynom $\tilde{p}_n$ mindestens Grad 1 hat, und auch q nach Definition nicht das Nullpolynom sein kann. Daher muss schon $m = n$ gelten. □

Das Proximum von f aus $C[a,b]$ in $U = \operatorname{span}(\tilde{p}_0, \ldots, \tilde{p}_n)$ ist nach (21.1) die FOURIER-Reihe

$$\tilde{u} = \sum_{k=0}^{n}\tilde{\alpha}_k\tilde{p}_k, \quad \tilde{\alpha}_k = \frac{(f, \tilde{p}_k)}{(\tilde{p}_k, \tilde{p}_k)} \tag{21.11}$$

Es folgt der

(21.12) Konvergenzsatz. Für jede Gewichtsfunktion $\omega \in C[a,b]$ mit $\omega(x) > 0$ für $a < x < b$ sind die Funktionen $\{p_0, \ldots, p_n\}$, $p_n := \tilde{p}_n \cdots \frac{1}{\sqrt{(\tilde{p}_n, \tilde{p}_n)}}$ ein vollständiges Orthonormalsystem in $V = C[a,b]$. Die Folge der Proxima konvergiert für $n \to \infty$ im Mittel gegen f.

Beweis. in HÄMMERLIN, HOFFMANN: Numerische Mathematik, Kapitel 4 §5.6. □

Beispiele.

(i) LEGENDRE-Polynome. Es sei $[a,b] = [-1,1]$ und $\omega(x) \equiv 1$ für $x \in [-1,1]$. Man prüft nach, dass die LEGENDRE-Polynome

$$\tilde{L}_n(x) := \frac{n!}{(2n)!} \frac{d^n(x^2-1)^n}{dx^n} = x^n + a_{n-1}x^{n-1} + \ldots$$

für $n = 0, 1, \ldots$ orthogonal sind, d.h., dass $\left(\tilde{L}_k, \tilde{L}_n\right) = 0$ für alle $k < n$ gilt. Hierzu benutzt man partielle Integration.

Weiterhin gilt $\left(\tilde{L}_n, \tilde{L}_n\right) = \frac{2n+1}{2} \cdot \frac{1}{(2^n \cdot n!)^2}$ und man hat

$$\tilde{L}_1(x) = x, \quad \tilde{L}_2(x) = x^2 - \frac{1}{3}, \quad \tilde{L}_3(x) = x^3 - \frac{3}{5}x.$$

Beachte den Nullstellensatz (21.10)!

(ii) CHEBYCHEV-Polynome: Es sei $[a,b] = [-1,1]$ und $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ für $-1 < x < 1$. Man definiert rekursiv die CHEBYCHEV-Polynome $T_n \in \Pi_n$:

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

$$T_0(x) = 1,\ T_1(x) = x$$

Man substituiert $x = \cos(\Theta), \Theta = \arccos(x)$ und erhält die Darstellung

$$T_n(x) = \cos(n\Theta) \quad \Theta = \arccos(x).$$

Dies schließt man aus dem Additionstheorem

$$\cos((n+1)\Theta) = 2\cos(\Theta)\cos(n\Theta) - \cos((n-1)\Theta).$$

Die Orthogonalität von $T_n(x)$ bezüglich $\omega(x)$ ergibt sich wegen

$$\int_{-1}^{1} T_i(x)T_k(x)\frac{dx}{\sqrt{1-x^2}} \overset{x=\cos(\Theta)}{=} \int_0^\pi \cos(i\Theta)\cos(k\Theta)\frac{\sin(\Theta)}{\sin(\Theta)}d\Theta = \begin{cases} 0, & i \neq k \\ \pi, & i = k = 0 \\ \frac{\pi}{2}, & i = k \neq 0 \end{cases}.$$

Die Nullstellen der $T_n(x)$ nennt man CHEBYCHEV-Abszissen. Wegen $T_n(x) = \cos(n\Theta)$ gilt für diese

$$x_k = \cos\left(\frac{2k-1}{n}\frac{\pi}{2}\right) \in (-1,1), \quad k = 1, \ldots, n.$$

Die Extremalstellen berechnen sich zu

$$x_k^{(ex)} = \cos\left(k\frac{\pi}{n}\right), \quad k = 0, \ldots, n,\ n \geq 1.$$

Für die Funktionswerte gilt $T_n(e_k^{(ex)}) = (-1)^k$.

Beispielsweise erhält man $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$ und allgemein $T_n(x) = 2^{n-1}x^n + \ldots$. Normiert man zum Höchstkoeffizienten 1, so erhält man

$$\tilde{T}_n(x) = \frac{1}{2^{n-1}}T_n(x) \in \tilde{\Pi}_n.$$

Die FOURIER-Entwicklung (21.11) einer Funktion $f \in C[a,b]$ lautet

$$\tilde{u} = \frac{c_0}{2} + \sum_{k=1}^{n} c_k T_k(x), \quad c_k = \frac{2}{\pi}\int_{-1}^{1} f(x)T_k(x)\frac{dx}{\sqrt{1-x^2}} \overset{x=\cos(\Theta)}{=} \frac{2}{\pi}\int_0^\pi f(\cos(\Theta))\cos(k\Theta)d\Theta.$$

Also sind c_k die FOURIER-Koeffizienten (21.7) der 2π-periodischen Funktion $F(\Theta) := f(\cos(\Theta))$.

Weitere Orthogonalpolynome finden sich beispielsweise in M. Abramovitz, F. Stegun: Handbook of Mathematical Functions.

VII Numerische Integration

§ 22 Die Integrationsformel von Newton-Cotes

Gegeben sei ein $f \in C[a,b]$ mit $[a,b] \subset \mathbb{R}$. Ziel ist die Berechnung von $I(f) = \int_a^b f(x)dx$. Man kann äquidistante Stützstellen $x_i = a + ih$ mit $h = \frac{b-a}{n}$ für $i = 0, \ldots, n$ wählen, es ist dann $x_0 = a$ und $x_n = b$. Es sei $p_n(x) \in \Pi_n$ das f interpolierende Polynom mit

(i) grad $(p_n) \leq n$,

(ii) $p_n(x_i) = f_i := f(x_i)$ für $i = 0, \ldots, n$.

Ein Näherungswert für $I(f)$ ist dann

$$I_n(f) := \int_a^b p_n(x)dx.$$

Der Fehler lautet

$$R_n(f) = I(f) - I_n(f) = \int_a^b f(x) - p_n(x)dx.$$

Mit der Formel von Lagrange gilt

$$p_n(x) = \sum_{i=0}^{n} L_i(x) f_i \quad L_i(x) = \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k}.$$

Wir erhalten

$$I_n(f) = \int_a^b p_n(x)dx = \sum_{i0}^{n} f_i \int_a^b L_i(x)dx$$

und definieren

$$A_i := \int_a^b L_i(x)dx = \int_a^b \prod_{k=0, k\neq i}^{n} \frac{x - x_k}{x_i - x_k} dx.$$

Die Formeln von Newton-Cotes ergeben sich daraus durch die Substitution $x = a + sh$, also $dx = hds$ für $0 \leq s \leq n$:

$$\left.\begin{array}{l} I_n(f) = \sum_{i=0}^{n} A_i f_i, \quad f_i = f(a+ih) \\ A_i = h \int_0^n \prod_{k=0, k\neq i}^{n} \frac{s-k}{i-k} ds =: h \cdot a_i, \quad a_i \in \mathbb{Q} \end{array}\right\} \tag{22.1}$$

Beispiele.

(i) $n = 1$. Trapezregel. Es ist $h = b - a$ und $a_0 = a_1 = \frac{1}{2}$. Wir erhalten $I_1 = \frac{1}{2}(f(a) + f(b))$.

(ii) $n = 2$. Simpsonsche Regel. Es ist $h = \frac{b-a}{2}$ und $a_0 = a_2 = \frac{1}{3}$, $a_1 = \frac{4}{3}$. Wir erhalten $I_2(f) = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$.

Allgemein gilt $a_i \in \mathbb{Q}$ und $\sum_{i=0}^{n} a_i = n$. Dazu betrachte man, dass insbesondere $p_n(x) \equiv 1$ für $f(x) \equiv 1$ gilt. Man erhält:

n	a_0	a_1	a_2	a_3	a_4	Skalierung	Bezeichnung
1	1	1				$\frac{1}{2}$	Trapezregel
2	1	4	1			$\frac{1}{3}$	Simpsonsche Regel
3	1	3	3	1		$\frac{3}{8}$	Newtonsche 3/8-Regel
4	7	32	12	32	7	$\frac{2}{45}$	Milne-Regel

Beachte: Für $n \geq 8$ können negative Gewichte a_i auftreten. Es kann dann für $I_n(f) = \sum_{i=0}^{n} ha_i f_i$ zu Auslöschungen kommen.

Beispiel. Sei $f(x) = e^x$. Dann ist $I(f) = \int_0^1 e^x dx = e - 1 \approx 1.7183$. Approximativ erhalten wir

$$\begin{aligned} I_1 &= \frac{1}{2}(1+e) \approx 1.8591 \\ I_2 &= \frac{1}{6}(1 + 4e^{\frac{1}{2}} + e) \approx 1.7189 \\ I_3 &= \frac{1}{8}(1 + 3e^{\frac{1}{3}} + 3e^{\frac{1}{3}} + e) \approx 1.7185 \end{aligned}$$

Eine Abschätzung für den Fehler $R_n(f) = I(f) - I_n(f)$ liefert der folgende

(22.2) Satz.

(i) Für $f \in C^{n+1}[a,b]$ gilt

$$|R_n(f)| \leq h^{n+2} c_n \left|\left| f^{(n+1)} \right|\right|_\infty$$

mit $c_n = \frac{1}{(n+1)!} \int_0^n \prod_{i=0}^n |s-i| ds$,

(ii) Für n gerade und $f \in C^{n+2}[a,b]$ gilt

$$|R_n(f)| \leq h^{n+3} c_n^* \left|\left| f^{(n+2)} \right|\right|_\infty$$

mit $c_n^* = \frac{n}{2} c_n$, c_n wir in (i).

Beweis.

(i) Es ist $R_n(f) = \int_a^b (f(x) - p_n(x)) dx$. Die Restgliedformel der Interpolation (vgl. (13.8)) besagt, dass es zu $x \in [a,b]$ ein $\xi_x \in [a,b]$ gibt mit $f(x) - p_n(x) = \frac{L(x)}{(n+1)!} f^{(n+1)}(\xi_x)$, wobei $L(x) = \prod_{i=0}^n (x - x_i)$. Es folgt

$$|R_n(f)| \leq \frac{1}{(n+1)!} \left(\int_a^b |L(x)| dx \right) \left|\left| f^{(n+1)} \right|\right|_\infty .$$

Durch Substitution mit $x = a + sh, 0 \leq s \leq n, dx = h ds$ erhalten wir

$$\int_a^b |L(x)| dx = \int_a^b \prod_{i=0}^n |x - x_i| dx = h^{n+2} \int_0^n \prod_{i=0}^n |s-i| ds.$$

Daraus folgt die Behauptung.

(ii) Für n gerade ist $L(x) = \prod_{i=0}^n (x - x_i)$ schiefsymmetrisch bezüglich der Intervallmitte $c = \frac{a+b}{2}$. Es gilt daher $\int_a^b L(x) dx = 0$. Die TAYLOR-Entwicklung von $f^{(n+1)}$ in $x = c = \frac{a+b}{2}$ liefert

$$f^{(n+1)}(\xi_x) = f^{(n+1)}(c) + (\xi_x - c) f^{(n+2)}(\eta_x)$$

für ein geeignetes $\eta_x \in [a,b]$. Nun gilt für den Fehler

$$\begin{aligned} \int_a^b (f(x) - p_n(x)) dx &= \frac{1}{(n+1)!} \int_a^b L(x) (f^{(n+1)}(c) + (\xi_x - c) f^{(n+2)}(\eta_x)) dx \\ &= \frac{1}{(n+1)!} \int_a^b L(x) (\xi_x - c) f^{(n+2)}(\eta_x) dx. \end{aligned}$$

Wegen $|\xi_x - c| \leq \frac{b-a}{2} = \frac{n}{2} h$ folgt

$$|R_n(f) \leq h^{n+2} c_n \frac{n}{2} h \left|\left| f^{(n+2)} \right|\right|_\infty = h^{n+3} c_n^* \left|\left| f^{(n+2)} \right|\right|_\infty$$

mit $c_n^* = \frac{n}{2} c_n$. □

Beispiele.

(i) $n = 1$ (Trapezregel). Es ist

$$I_1(f) = \frac{1}{2}(f(a) + f(b)).$$

Als Abschätzung gilt

$$|R_n(f)| \leq \frac{h^3}{12} \max_{a \leq x \leq b} |f''(x)|,\ h = b - a.$$

Dies gilt wegen

$$c_1 = \frac{1}{2}\int_0^1 |s| \cdot |s-1| ds = \frac{1}{2}\int_0^1 s(1-s)dx = \frac{1}{2}\left[\frac{1}{2}s^2 - \frac{1}{3}s^3\right]_0^1 = \frac{1}{2}\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{1}{12}.$$

(ii) $n = 2$ (SIMPSONsche Regel). Nach (22.2) gilt

$$|R_n(f)| \leq \frac{h^5}{90} \max_{a \leq x \leq b} |f^{(4)}|,\ h = \frac{b-a}{2}.$$

(iii) $n = 3$ (NEWTONsche $\frac{3}{8}$-Regel). Es gilt $I = \int_0^1 \frac{\sin(x)}{x} dx = 0.94608307$ sowie

$$\begin{aligned} I_3 &= \frac{1}{8}\left(f(0) + 3f\left(\frac{1}{3}\right) + 3f\left(\frac{2}{3}\right) + f(1)\right) = 0.94611 \\ |R_3(f)| &\leq \frac{3}{80}\left(\frac{1}{3}\right)^5 \left\|f^{(4)}\right\|_\infty \leq 3.1 \cdot 10^{-5} \\ |I - I_3| &= 2.8 \cdot 10^{-5} \end{aligned}$$

§ 23 Zusammengesetzte Trapezregel und Extrapolationsverfahren

Gegeben seien äquidistante Stützstellen $x_i = a + hi$, $h = \frac{b-a}{n}$, $i = 0, \ldots, n$. Man wendet nun die Trapezregel auf die Teilintervalle $[x_i, x_{i+1}]$ an:

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{h}{2}\left(f(x_i) + f(x_{i+1})\right).$$

Dies ergibt die zusammengesetzte Trapezregel

$$\int_a^b f(x)dx \approx T(h) = \sum_{i=0}^{n-1} \frac{h}{2}(f(x_i) + f(x_{i+1})) = h\left(\frac{f(a)}{2} + \sum_{i=1}^{n-1} f(x_i) + \frac{f(b)}{2}\right). \tag{23.1}$$

Für den Gesamtfehler gilt nach (22.2) (i) mit $n = 1$

$$\left|T(h) - \int_a^b f(x)dx\right| \leq \frac{1}{2}\sum_{i=0}^{n-1} h^3 \max_{x_i \leq x \leq x_{i+1}} |f''(x)| \leq \frac{n}{12}h^3 \max_{a \leq x \leq b} |f''(x)| = \frac{b-a}{12}h^2 \, ||f''||_\infty \tag{23.2}$$

wegen $h = \frac{b-a}{n}$. Diese Fehlerabschätzung ist Motivation für die asymptotische Entwicklung von $T(h)$:

(23.3) Satz (Euler-Maccaurinsche Summenformel). Für $f \in C^{2m+2}[a, b]$ gilt die Entwicklung

$$T(h) = \tau_0 + \tau_1 h^2 + \tau_2 h^4 + \ldots + t_m h^{2m} + \alpha_{m+1}(h)h^{2m+2}$$

mit

(i) $\tau_0 = \int_a^b f(x)dx$,

(ii) $\tau_k = \frac{(-1)^{k+1}}{(2k)!} B_k \left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right)$. Hierbei sind die B_k die sogenannten Bernoulli-Zahlen, $B_1 = \frac{1}{6}, B_2 = \frac{1}{30}, B_3 = \frac{1}{42}, \ldots$.

(iii) $\alpha_{m+1}(h) = \frac{1}{(2m+2)!}\int_a^b f^{(2m+2)}(x) K_{2m+2}\left(\frac{x-a}{h}\right) dx$, wobei $K_{2m+2} \in C[0, n]$ ist mit

$$\int_a^b K_{2m+2}\left(\frac{x-a}{2}\right) dx = (-1)^m B_{m+1}(b-a).$$

Beweis. siehe Stoer, "Numerische Mathematik 1, §3.2 □

Extrapolation

Man wählt zunächst eine Schrittweitenfolge h_j mit $h_0 = b - a$, $h_0 > h_1 > \ldots$.

Beispiele.

(i) Romberg-Folge: $h_j = \frac{h_0}{2^j} = \frac{b-a}{2^j}$ für $j = 0, 1, \ldots$.

(ii) Bulirsch-Folge: $h_0 = b - a$, $h_1 = \frac{h_0}{2}, h_2 = \frac{h_0}{3}, h_3 = \frac{h_0}{4}, h_4 = \frac{h_0}{6}, h_5 = \frac{h_0}{8}, \ldots$

Berechne die Trapezsummen

$$T_{j,0} := T(h_j), \quad j = 0, 1, \ldots, m.$$

Sei $\tilde{T}_{m,m}(h)$ das interpolierende Polynom in $x = h^2$ mit

$$\left.\begin{array}{rcl} \tilde{T}_{m,m}(h) & = & a_0 + a_1 h^2 + \ldots + a_m h^{2m} \\ \tilde{T}_{m,m}(h_j) & = & T(h_j),\ j = 0, 1, \ldots, m \end{array}\right\} \tag{23.4}$$

Idee: Der WErt für $h = 0$,

$$\tilde{T}_{m,m}(0) = a_0 \approx \tau_0 = \int_a^b f(x)dx$$

ist eine gute Näherung für τ_0.
Es gilt

$$\tilde{T}_{m,m}(0) = \sum_{j=0}^{m} L_j(0) T(h_j)$$

Beispiel. $h_0 = b - a$, $h_1 = \frac{b-a}{2}$. Dann ist

$$T_{1,1} := \tilde{T}_{1,1}(0) = L_0(0) \cdot T(h_0) + L_1(0) \cdot T(h_1)$$

mit

$$L_i(x) = \prod_{k=0, k\neq i}^{1} \frac{x - x_k}{x_i - x_k}, \quad x_i = h_i^2, \quad i = 0, 1$$

und es ergibt sich für $x = 0$

$$L_0(0) = \frac{-h_1^2}{h_0^2 - h_1^2} = -\frac{1}{3}, \qquad L_1(0) = \frac{-h_0^2}{h_1^2 - h_0^2} = \frac{4}{3}.$$

Es folgt also

$$\begin{aligned} T_{11} &= \tilde{T}_{11}(0) = -\frac{1}{3} \underbrace{\frac{b-a}{2}\left(f(a) + f(b)\right)}_{T(h_0)} + \frac{4}{3} \underbrace{\frac{b-a}{2}\left(\frac{f(a)}{2} + f\left(\frac{a+b}{2}\right) + \frac{f(b)}{2}\right)}_{T(h_1)} \\ &= \frac{h_1}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \end{aligned}$$

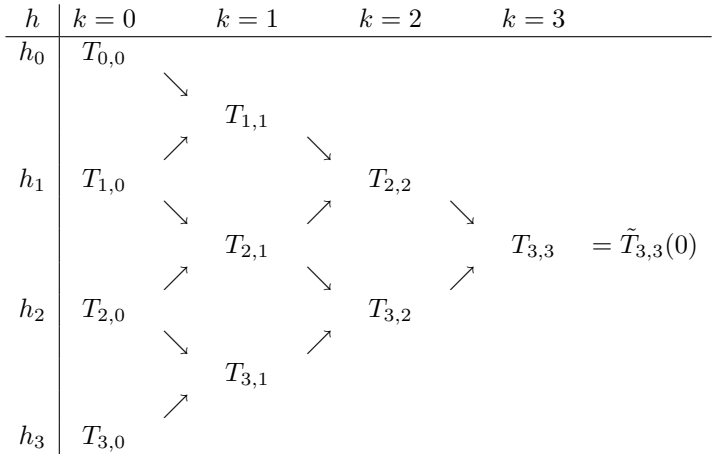
Die Berechnung des Wertes $\tilde{T}_{m,m}(0) = a_0$ erfolgt mit dem Algorithmus von NEVILLE: Sei $\tilde{T}_{ik}(h)$ dasjenige Polynom in $x = h^2$ mit

$$\tilde{T}_{i,k}(h_j) = T_{j,0} = T(h_j) \quad \text{für } j = i-k, \ldots, i,$$

d.h. man betrachte den Abschnitt $h_{i-k}, \ldots, h_i$. Dann ergibt sich folgende Rekursion für $T_{i,k} = \tilde{T}_{i,k}(0)$:

$$T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1} \quad \text{für } 1 \leq k \leq i \leq m.$$

Die Berechnung erfolgt spaltenweise in einem Tableau



Es ergibt sich

$$\left(\frac{h_{i-k}}{h_i}\right)^2 - 1 = 2^{2k} - 1$$

und damit

$k = 1$ $T_{i,1} = T_{i,0} - \frac{T_{i,0} - T_{i-1,0}}{3}$,

$k = 2$ $T_{i,2} = T_{i,1} - \frac{T_{i,1} - T_{i-1,1}}{15}$.

Beispiel. $I = I(f) = \int_0^1 e^x dx = 1.718281828$ mit $m = 3$.

h_i	$T_{i,0}$	$T_{i,1}$	$T_{i,2}$	$T_{i,3}$
1	1.859140914			
		1.718861151		
$\frac{1}{2}$	1.753931092		1.718282687	
		1.718318841		1.71828182
$\frac{1}{4}$	1.727221904		1.718281842	
		1.718284155		
$\frac{1}{8}$	1.720518592			

Nächstes Ziel ist ein Ausdruck für den Fehler $\tilde{T}_{m,m} - \int_a^b f(x)dx$.

(23.5) Hilfssatz. Seien x_i, $i = 0, \ldots, m$ paarweise verschiedene Zahlen (Knoten) und sei

$$L_i = \prod_{k=0, k\neq i}^{m} \frac{x - x_k}{x_i - x_k}$$

für $i = 0, \ldots, m$. Dann gilt

$$\sum_{i=0}^{m} x_i^k L_i(0) = \begin{cases} 1, & k = 0 \\ 0, & k = 1, \ldots, m \\ (-1)^m x_0 \cdot \ldots \cdot x_m, & k = m+1 \end{cases}.$$

Beweis. Setze $x = 0$ in die folgenden Gleichungen ein:

$$x^k \equiv \sum_{i=0}^{m} x_i^k L_i(x) \, \forall x \in \mathbb{R}, \, k = 0, \ldots, m$$

sowie

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

Die zweite Gleichung folgt mit $f(x) = x^{m+1}$, also $f^{(m+1)}(\xi)\frac{1}{(m+1)!} = 1$ aus der Restgliedformel der Interpolation. □

Durch Substitution mit $x = h^2$, $x_i = h_i^2$ in (23.5) erhält man

$$\sum_{i=0}^{m} h_i^{2k} L_i(0) = \begin{cases} 1, & k = 0 \\ 0, & k = 1, \ldots, m \\ (-1)^m h_0^2 \cdot \ldots \cdot h_m^2, & k = m+1 \end{cases}. \tag{23.6}$$

Das Polynom $\tilde{T}_{m,m}(h)$ in (23.4) interpoliert die Werte $T(h_i)$, $i = 0, \ldots, m$:

$$T_{m,m} = \tilde{T}_{m,m}(0) = \sum_{i=0}^{m} L_i(0) T(h_i). \tag{23.7}$$

Mit $T(h) = \tau_0 + \tau_1 h^2 + \ldots + \tau_m h^{2m} + \alpha_{m+1}(h) h^{2m+2}$ folgt

$$T_{m,m} = \sum_{i=0}^{m} L_i(0) T(h_i) = \tau_0 + \frac{1}{(2m+2)!} \int_a^b f^{(2m+2)} K(x) dx$$

mit

$$K(x) = \sum_{i=0}^{m} L_i(0) h_i^{2m+2} K_{2m+2}\left(\frac{x-a}{h_i}\right)$$

Man kann zeigen, dass die Funktion $K(x)$ auf $[a, b]$ für die ROMBERG- und BULIRSCH-Folge h_i stets das gleiche Vorzeichen hat. Nach dem Mittelwertsatz der Integralrechnung gilt dann

$$\int_a^b f^{(2m+2)}(x) K(x) dx = f^{(2m+2)}(\xi) \int_a^b K(x) dx$$

an einer geeigneten Stelle $\xi \in [a,b]$. Weiter ist

$$\begin{aligned}\int_a^b K(x)dx &= \sum_{i=0}^{m} L_i(0) h_i^{2m+2} \underbrace{\int_a^b K_{2m+2}\left(\frac{x-a}{2}\right) dx}_{(-1)^m (b-a) B_{m+1}} \\ &= (-1)^m h_0^2 \cdots h_m^2 (-1)^m (b-a) B_{m+1} = h_0^2 \cdots h_m^2 (b-a) B_{m+1}.\end{aligned}$$

Insgesamt gilt dann

$$T_{mm} - \int_a^b f(x)dx = (b-a) h_0^2 \cdots h_m^2 B_{m+1} \frac{f^{(2m+2)}(\xi)}{(2m+2)!} \tag{23.8}$$

Für $m = 0$ erhält man wegen $B_1 = \frac{1}{6}$ die Abschätzung (23.2). In der Praxis genügt es oft, $m \leq 6$ zu betrachten.

§ 24 Allgemeines über Extrapolationsverfahren

Zur näherungsweise Berechnung der Lösung eines Problems wendet man Diskretisierungsverfahren an:

- wähle Schrittweite $h > 0$.
- Reusltat der Rechnung sei $T(h)$.
- Asymptotische Entwicklung

$$T(h) = \tau_0 + \tau_1 h^\gamma + \tau_2 h^{2\gamma} + \ldots + \tau_m h^{m\gamma} + \alpha_{m+1}(h) h^{(m+1)\gamma} \tag{24.1}$$

Hierbei ist τ_i unabhängig von h und $\alpha_{m+1}(h)$ in h beschränkt, d.h. $\alpha_{m+1}(h) = \tau_{m+1} + \mathcal{O}(h)$. Weiterhin ist $\tau_0 = \lim_{h \searrow 0} T(h)$.

Beispiel. Numerische Differentiation. Sei $f \in C^{m+2}[x-a, x+a]$.

(i) Für $h > 0$ ist $T(h) = \frac{f(x+h)-f(x)}{h}$ eine Approximation der Ableitung $f'(x)$. Durch TAYLOR-Entwicklung erhält man

$$T(h) = \tau_0 + \tau_1 h + \ldots + \tau_m h^m + h^{m+1}(\tau_{m+1} + \mathcal{O}(h)) \text{ mit } \tau_k = \frac{f^{(k+1)}(x)}{(k+1)!},\ k = 0, \ldots, m+1$$

Es gilt $\gamma = 1$.

(ii) Eine bessere Approximation erhält man durch $T(h) = \frac{f(x+h)-f(x-h)}{2h}$ für $h \neq 0$. Für $f \in C^{2m+3}[x-a, x+a]$ mit $a > 0$ erhält man die asymptotische Entwicklung

$$T(h) = \tau_0 + \tau_1 h^2 + \ldots + \tau_m h^{2m} + h^{2m+2}(\tau_{m+1} + \mathcal{O}(h)) \text{ mit } \tau_k = \frac{f^{(2k+1)}(x)}{(2k+1)!},\ k = 0, \ldots, m+1.$$

Hier gilt $\gamma = 2$.

Extrapolationsverfahren

Man wähle eine Schrittweitenfolge $h_0 > \ldots > h_m > \ldots > 0$ und berechne $T(h_i)$ für $i = 0, \ldots, m$. Sei $\tilde{T}_{i,k}(h)$ für $k \leq i$ dasjenige Polynom in $x = h^\gamma$, für das

$$\tilde{T}_{i,k}(h) = b_0 + b_1 h^\gamma + \ldots + b_m h^{m\gamma},\ m = i - k,$$

mit $\tilde{T}_{i,k}(h_j) = T(h_j)$ für $j = i-k, \ldots, i$ gilt. Die extrapolierten Werte

$$T_{i,k} := \tilde{T}_{i,k}(0) = b_0$$

sind Näherungswerte für

$$\tau_0 = \lim_{h \searrow 0} T(h).$$

Für die Berechnung nach dem Algorithmus von NEVILLE setzt man darin nun $x_i = h_i^\gamma$ und erhält

$$T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^\gamma - 1} \tag{24.2}$$

für $1 \leq k \leq i \leq m$. Für die ROMBERG-Folge gilt $h_i = \frac{h_0}{2^i}$ und $\left(\frac{h_{i-k}}{h_i}\right)^\gamma - 1 = 2^{k\gamma} - 1$.

Beispiel. Numerische Differentiation. Berechne $f'(0)$ für $f(x) = \tan\left(\frac{\pi}{2}x\right)$. Es gilt $f'(0) = \frac{\pi}{2} \approx 1.5707963$. Wir betrachten die Extrapolation dieses Wertes mit

(i) $T(h) = \frac{f(x+h)-f(x)}{h} = \frac{f(h)-f(0)}{h}$,

(ii) $T(h) = \frac{f(x+h)-f(x-h)}{2h}$.

Wegen $x = 0$ und $f(0) = 0$ erhält man - da f eine ungerade Funktion ist - dieselben Startwerte $T(h_i)$. Methode (i):

h_i	$T_{i,0}$	$T_{i,1}$	$T_{i,2}$	$T_{i,3}$
0.5	2.00000			
		1.31370		
0.25	1.65685		1.59643	
		1.52575		1.56382
0.125	1.59130		1.57197	
		1.56042		
0.0625	1.57586			

Methode (ii):

h_i	$T_{i,0}$	$T_{i,1}$	$T_{i,2}$	$T_{i,3}$
0.5	2.00000			
		1.54247		
0.25	1.65685		1.57125	
		1.56945		1.57079
0.125	1.59130		1.57080	
		1.57072		
0.0625	1.57586			

Für eine Abschätzung des Fehlers $T_{i,k} - \tau_0$ benutzen wir die Methode aus §23. Es gilt

$$T_{i,k} = \tilde{T}_{i,k}(0) = \sum_{j=i-k}^{i} L_j(0) T(h_j),\ L_j(x) = \prod_{l=i-k, l\neq j}^{i} \frac{x - x_l}{x_j - x_l}.$$

Setzt man dort $x = h^\gamma$, so kann man zeigen

$$\sum_{j=i-k}^{i} h_j^{r\gamma} L_j(0) = \begin{cases} 1, & r = 0 \\ 0, & r = 1, \dots, k \\ (-1)^k h_{i-k}^\gamma \cdots h_i^\gamma, & r = k+1 \end{cases}.$$

Die Fehlerabschätzung lautet dann

$$T_{i,k} - \tau_0 = (-1)^k h_{i-k}^\gamma \cdots h_i^\gamma (\tau_{k+1} + \mathcal{O}(h_{i-k}) \tag{24.3}$$

§ 25 Die Gauss'sche Integrationsmethode

Sei $f \in C[a,b]$ und sei $w(x)$ eine stetige Gewichtsfunktion mit $w(x) > 0$ für $a < x < b$. Gesucht ist eine Intgrationsformel für

$$I(f) = \int_a^b f(x)w(x)dx. \tag{25.1}$$

Dazu betrachtet man die nicht äquidistanten Stützstellen $x_i \in [a,b]$, $i = 1,\dots,n$, $x_1 < x_2 < \dots < x_n$. Mit Π_j bezeichnen wir die Polynome vom Grad $\leq j$, $\tilde{\Pi}_j = \{p \in \Pi_j \mid p(x) = x^j + a_{j-1}x^{j-1} + \dots + a_0\}$.

Idee: Eine Integrationsformel der Form

$$G_n(f) = \sum_{i=1}^n A_i f(x_i) \tag{25.2}$$

hat $2n$ freie Parameter A_i, x_i $(i = 1,\dots,n)$.

Forderung: $G_n(f) = I(f)$ für alle $f \in \Pi_{2n-1}$, d.h. $G_n(f)$ ist exakt in Π_{2n-1}. Dies ergibt $2n$ Bedingungen für $2n$ Parameter.

(25.3) Satz. Es gibt keine Formel des Typs (25.2), die exakt in Π_{2n} ist.

Beweis. Annahme: $G_n(f) = \int_a^b f(x)w(x)dx$, $\forall f \in \Pi_{2n}$.

Mit $f(x) := \prod_{i=1}^n (x - x_i)^2$, $f \in \Pi_{2n}$, erhält man einen Widerspruch:

$$G_n(f) = \sum_{i=1}^n A_i \underbrace{f(x_i)}_{=0} = 0,$$

$$I(f) = \int_a^b \prod_{i=1}^n (x - x_i)^2 w(x)dx > 0$$

□

Wir entwickeln nun die Konstruktion einer Π_{2n-1} exakten Formel

$$G_n(f) = \sum_{i=1}^n A_i f(x_i).$$

Seien $\tilde{p}_n \in \tilde{\Pi}_n$ die zu $w(x)$ gehörenden Orthogonalpolynome (vgl. §16) bzgl. des Inneren Produktes $\int_a^b f(x)g(x)w(x)dx$, $f,g \in C[a,b]$. Nach Satz (16.10) hat dann $\tilde{p}_n \in \tilde{\Pi}_n$ genau n Nullstellen $x_1,\dots,x_n \in]a,b[$.

(25.4) Satz. Seien $x_1,\dots,x_n \in]a,b[$ die Nullstellen des Orthogonalpolynoms $\tilde{p}_n \in \tilde{\Pi}_n$ und sei

$$L_i(x) = \prod_{\substack{k=1 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k}, \qquad i = 1,\dots,n.$$

Dann ist die Integrationsformel

$$G_n(f) = \sum_{i=1}^n A_i f(x_i)$$

exakt in Π_{2n-1}. Es gilt

$$A_i = \int_a^b L_i(x)^2 w(x)dx > 0.$$

Beweis. Interpolationsformel von LAGRANGE.

$$f(x) = \sum_{i=1}^{n} L_i(x) f(x_i) \qquad \forall f \in \Pi_{n-1}.$$

$$\Rightarrow G_n(f) = I(f) = \sum_{i=1}^{n} \left(\int_a^b L_i(x) w(x) dx \right) f(x_i)$$

Faktorisierung eines Polynoms $f \in \Pi_{2n-1}$ durch $f = q \cdot \tilde{p}_n + r$, $q, r \in \Pi_{n-1}$

$$\begin{aligned} \Rightarrow I(f) &= \int_a^b f(x) w(x) dx \\ &= \underbrace{\int_a^b q(x) \tilde{p}_n(x) w(x) dx}_{=(q,\tilde{p}_n)=0,\ \text{da } q \in \Pi_{n-1} \text{ und } \tilde{p}_n orthogonal} + \int_a^b r(x) w(x) dx \\ &= \int_a^b r(x) w(x) dx = G_n(r), \qquad \text{da } r \in \Pi_{n-1}. \end{aligned}$$

Es gilt

$$G_n(r) = G_n(r) + \underbrace{G_n(q \cdot \tilde{p}_n)}_{=0,\ \text{da } \tilde{p}_n(x_i) = 0} = G_n(r + q \cdot \tilde{p}_n) = G_n(f)$$

$$\Rightarrow I(f) = G_n(f) \qquad \forall f \in \Pi_{2n-1}.$$

Es ist $L_i^2 \in \Pi_{2n-2}$

$$\Rightarrow \int_a^b L_i(x)^2 w(x) dx = G_n(L_i^2) = \sum_{k=1}^{n} A_k L_i(x_k)^2 = \sum_{k=1}^{n} A_k \delta_{i,k}^2 = A_i$$

Insbesondere $A_i > 0$. □

Beispiel. $[a, b] = [-1, 1]$, $w(x) \equiv 1$, $\tilde{p}_n(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n}(x^2 - 1)^n$
LEGENDRE-Polynome $\tilde{p}_1(x) = x$, $\tilde{p}_2(x) = x^2 - \frac{1}{3}$, $\tilde{p}_3(x) = x^3 - \frac{3}{5}x$

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

$f(x) = e^x$, $I(f) = \int_{-1}^{1} e^x dx = 2.350402$.
SIMPSON-Regel $I_2(f) = 2.362054$.
GAUSS-Integration mit gleich vielen Funktionsauswertungen: $G_3(f) = 2.350337$.

Fazit. Die GAUSS-Formeln liefern im Vergleich zu den NEWTON-COTES-Formeln die genaueren Ergebnisse. Nachteil: Beim Übergang $n \to n+1$ können die berechneten Funktionswerte $f(x_i)$ nicht benutzt werden, da die Nullstellen x_i verschieden sind für n und $n+1$.

Fehlerabschätzung.

(25.5) Satz. Sei $f \in C^{2n}[a, b]$ und sei $G_n(f)$ die in Satz (25.4) definierte Integrationsformel. Dann gilt

$$I(f) - G_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!} (\tilde{p}_n, \tilde{p}_n)$$

mit einem $\xi \in [a, b]$. *Beweis.* vgl.STOER, Numerische Mathematik 1. □

VIII Gewöhnliche Differentialgleichungen

§ 26 Theoretische Grundlagen gewöhnlicher Differentialgleichungen

§ 26.1 Typen von Differentialgleichungen (DGL)

(1) Explizite DGL n-ter Ordnung
Sei $D \subset \mathbb{R}^{n+1}$ $(n \geq 1)$ uns sei $f : D \to \mathbb{R}$. Sei $I \subset \mathbb{R}$ ein Intervall.
Eine n-mal stetig diff'bare Funktion $y : I \to \mathbb{R}$, d.h. $y \in C^n(I, \mathbb{R})$, heißt Lösung der expliziten DGL n-ter Ordnung

$$y^{(n)} = f(x, y, y', \ldots, y^{(k)}, \ldots, y^{(n-1)}) \tag{26.1}$$

im Intervall I, wenn gilt:

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right),$$

$$\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right) \in D \qquad \forall x \in I.$$

Die DGL (26.1) heißt linear, falls

$$y^{(n)} = a_0(x)y + a_1(x)y' + \ldots + a_{n-1}(x)y^{(n-1)} + b(x), \tag{26.2}$$

mit skalaren Funktionen $a_i(\cdot) \in C(I, \mathbb{R})$, $i = 0, 1, \ldots, n-1$, $b(\cdot) \in C(I, \mathbb{R})$
$n = 1$: skalare DGL 1.Ordnung

$$y' = f(x, y) \tag{26.3}$$

Beispiele. $n = 1$:

(1)

$$y' = \lambda y, \ \lambda \in \mathbb{R},$$

allgemeine Lösung

$$y(x) = ce^{\lambda x}, \ x \in \mathbb{R}, \ c \in \mathbb{R}.$$

(2) RICCATI-DGL

$$y' = 1 + y^2$$

allgemeine Lösung

$$y(x) = \tan(x - c), \ c \in \mathbb{R}$$

$$y(x) \to \pm\infty \text{ für } x \to c \pm \frac{\pi}{2}$$

(3)

$$y' = \frac{1}{y}, \ D = \mathbb{R}x]0, \infty[,$$

allgemeine Lösung

$$y(x) = \sqrt{2x + c}, \ c \in \mathbb{R},$$

nur definiert für $2x + c \geq 0$

$n = 2$:

(1)

$$y''(x) = ay'(x)^2 + by(x) + c, \ a, b, c \in \mathbb{R},$$

allgemeine Lösung: ?

(2)

$$y'' + y = 0 \text{ (lineare DGL)},$$

allgemeine Lösung

$$y(x) = c_1 \sin(x) + c_2 \cos(x),\ c_1, c_2 \in \mathbb{R}.$$

(3)

$$y'' + \sin(y) = 0$$

allgemeine Lösung: ?

(2) Systeme von DGL
Sei $D \subset \mathbb{R}^{n+1}$ und sei $f : D \to \mathbb{R}^n$. $f = (f_1, f_2, \ldots, f_n)^*$. Sei $I \subset \mathbb{R}$ ein Intervall. Eine stetig diff'bare Funktion $y : I \to \mathbb{R}^n$, d.h. $y \in C^1(I, \mathbb{R}^n)$, heißt Lösung des Systems von DGL 1.Ordnung,

$$y' = f(x, y) \tag{26.4}$$

im Intervall I, wenn gilt

$$y'(x) = f(x, y(x)),\ (x, y(x)) \in D \text{ für alle } x \in I.$$

Komponentenweise mit $y(x) = (y_1(x), y_2(x), \ldots, y_n(x))^*$

$$\begin{aligned} y_1'(x) &= f_1(x, y_1(x), \ldots, y_n(x)) \\ &\vdots \\ y_k'(x) &= f_k(x, y_1(x), \ldots, y_n(x)) \\ &\vdots \\ y_n'(x) &= f_n(x, y_1(x), \ldots, y_n(x)) \end{aligned}$$

System linearer DGL, falls f_i affin-linear in $y_1, \ldots, y_n$ ist, d.h.

$$f_i(x, y_1, \ldots, y_n) = a_{i1}(x)y_1 + a_{i2}(x)y_2 + \ldots + a_{in}(x)y_n + b_i(x), \qquad \forall i = 1, \ldots, n.$$

Setze

$$A(x) = \begin{pmatrix} a_{11}(x) & \ldots & a_{1n}(x) \\ \vdots & & \vdots \\ a_{n1}(x) & \ldots & a_{nn}(x) \end{pmatrix} n \times n - Matrix,$$

$$b(x) = \begin{pmatrix} b_1(x) \\ \vdots \\ b_n(x) \end{pmatrix}$$

System linearer DGL in vektorieller Form

$$y' = A(x)y + b(x) \tag{26.5}$$

Die explizite skalare DGL (26.1) n-ter Ordnung ist äquivalent zu einem System 1.Ordnung: Setze

$$y_1 = y,\ y_2 = y',\ y_3 = y'',\ \ldots,\ y_k = y^{(k-1)},\ \ldots,\ y_n = y^{(n-1)},$$

System

$$\begin{cases} y_1' &= y_2, \\ y_2' &= y_3, \\ &\vdots \\ y_k' &= y_{k+1}, \\ &\vdots \\ y_{n-1}' &= y_n, \\ y_n' &= f(x, y_1, y_2, \ldots, y_n). \end{cases}$$

(3) Anfangswertaufgaben (AWA) für Systeme von DGL
Gegeben sei das System von DGL 1.Ordnung

$$y' = f(x, y).$$

Sei $I \subset \mathbb{R}$ ein Intervall und seien $x_0 \in I$ und $y_0 \in \mathbb{R}^n$ mit $(x_0, y_0) \in D$.
Eine Funktion $y \in C^1(I, \mathbb{R}^n)$ heißt Lösung der AWA

$$y' = f(x, y),\ y(x_0) = y_0, \tag{26.6}$$

wenn $y(\cdot)$ eine Lösung der DGL $y' = f(x, y)$ in I ist und y die Anfangsbedingung $y(x_0) = y_0$ erfüllt.
Für eine explizite DGL n-ter Ordnung lauten die Anfangsbedingungen

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}),$$
$$y^{(i)}(x_0) = y_{i,0} \text{ für } i = 0, 1, \ldots, n-1.$$

Andere Bezeichnungen für dynamische Systeme: Ersetze

$$x \to t \text{ (Zeit)} \qquad y(x) \to x(t) \text{ (Zustand eines Systems zur Zeit } t)$$

$$\dot{x} = \frac{dx}{dt} = f(t, x),\ x(t_0) = x_0 \in \mathbb{R}^n \tag{26.6a}$$

Beispiele.

(1) Exponentielles Wachstum bzw. Zerfall. Es bezeichne $y(t)$ die Menge einer radioaktiven Substanz oder die Größe einer Population zur Zeit t. Gegeben ist dann eine DGL $\dot{y}(t) = \frac{dy}{dt} = ay(t)$ mit Anfangswert $y(t_0) = y_0 > 0$, $a \in \mathbb{R}$. Als Lösung erhält man $y(t) = y_0 e^{a(t-t_0)}$.

(2) Logistisches Wachstum. Sei $y(t)$ die Größe einer Population zur Zeit t. Man hat eine DGL $\dot{y}(t) = a_y - b_y^2 = a_y\left(1 - \frac{b}{a}y\right)$ mit Anfangswert $y(0) = y_0 > 0$. Diese löst man mit der Methode der Trennung der Variablen:

$$y(t) = k\frac{y_0}{y_0 + (k - y_0)e^{-at}},\ k = \frac{b}{a}$$
$$\lim_{t\to\infty} y(t) = k \text{ für alle } y_0 > 0$$

Diese DGL treten z.B. beim Schätzung der Parameter a und b aus Bevölkerungsdaten auf.

(3) AWA. Gegeben ist eine DGL $y' = y^2$ mit Anfangswert $y(0) = c > 0$. Dann ist $y(x) = \frac{c}{1-cx}$ die eindeutige Lösung im maximalen Existenzintervall $I^* =]-\infty, \frac{1}{c}[$ (offenes Intervall).

(4) NEWTONsche Bewegungsgleichungen. Es bezeichne $x(t)$ die Position eines Wagens zur Zeit t und $v(t)$ die Geschwindigkeit zur Zeit t. Man erhält ein System

$$\dot{x} = v, \qquad \dot{v} = f(t, y, v)$$

a) freier Fall. Es ist $f(t, y, \dot{y}) = mg$ mit $g = 9.81\frac{m}{s^2}$. Weiterhin ist $\dot{y} = v$ und $m\dot{v} = mg$, also $\dot{v} = g$. Mit $y(0) = y_0$ und $v(0) = v_0$ erhalten wir

$$y(t) = \frac{1}{2}gt^2 + v_0 t + y_0,$$
$$v(t) = gt + v_0$$

als Lösungen der AWA.

b) gedämpfte Schwingung. Es wirkt eine Kraft $f(t, y, \dot{y}) = mg - ky$ mit $k > 0$. Daraus ergibt sich $m\dot{v} = f(t, y, \dot{y}) = mg - ky$, also erhalten wir $\dot{v} = g - \frac{k}{m}y$ und $\dot{y} = v$.

(5) Räuber-Beute-Modell. Wir bezeichnen mit $x(t)$ die Beute- und mit $y(t)$ die Räuberpopulation, jeweils zum Zeitpunkt t. Man erhält das dynamische Modell von VOLTERRA, LOTKA

$$\dot{x}(t) = \frac{dx}{dt} = ax(t) - b \cdot x(t)y(t),$$
$$\dot{y}(t) = -cy(t) + d \cdot x(t)y(t)$$

jeweils mit Anfangswerten $x(0) = x_0 > 0$ und $y(0) = y_0 > 0$.

§ 26.2 Existenz und Eindeutigkeit der Lösung von Anfangswertaufgaben

Sei $D \subset \mathbb{R}^{n+1}$ und sei $f : D \to \mathbb{R}^n$ stetig. Weiterhin sei $||\cdot||$ eine Norm im $\mathbb{R}^n$, z.B. $||\cdot|| = ||\cdot||_\infty$.

(26.7) Definition.

(i) f genügt auf D eine LIPSCHITZ-Bedingung bezüglich $y \in \mathbb{R}^n$, wenn ein $L > 0$ existiert mit $||f(x, y_1) - f(x, y_2)|| \leq L\,||y_1 - y_2||$ für alle $(x, y_1), (x, y_2) \in D$.

(ii) f heißt LIPSCHITZ-stetig bezüglich y auf D, wenn es zu jedem $(x, y) \in D$ eine Umgebung U gibt, so dass die Einschränkung $f\big|_{U \cap D}$ einer LIPSCHITZ-Bedingung bezüglich y auf $D \cap U$ genügt.

Beispiele.

(i) Die Abbildung $x \mapsto A(x) \in \mathbb{R}^{n \times n}$ sei stetig. Betrachtet man ein System linearer DGL

$$y'(x) = A(x)y(x) + b(x) = f(x, y(x)),$$

$$f(x, y) = A(x)y + b(x),$$

so folgt

$$||f(x, y_1) - f(x, y_2)|| = ||A(x) \cdot (y_1 - y_2)|| \leq ||A(x)||\,||y_1 - y_2||\,.$$

Sei $D = I \times \mathbb{R}^n$ für $I \subset \mathbb{R}$ kompakt. Setze $L := \max_{x \in I} ||A(x)|| < \infty$, so folgt

$$||f(x, y_1) - f(x, y_2)|| \leq L\,||y_1 - y_2||$$

für alle $(x, y_1), (x, y_2) \in D = I \times \mathbb{R}^n$.

(ii) Man betrachte die DGL $y' = 1 + y^2 = f(x, y)$. Dann genügt $f(x, y)$ keiner LIPSCHITZ-Bedingung auf $D = I \times \mathbb{R}$. Aber $f(x, y) = 1 + y^2$ genügt einer LIPSCHITZ-Bedingung auf $D = I \times J$ mit $J \subset \mathbb{R}$ kompakt. Außerdem ist $1 + y^2$ LIPSCHITZ-stetig auf $D = I \times \mathbb{R}$.

(iii) Sei $f : \mathbb{R}^2 \to \mathbb{R}$ mit $f(x, y) = |y|^{\frac{1}{2}}$. f ist LIPSCHITZ-stetig auf $\mathbb{R} \times (0, \infty)$, genügt aber keiner LIPSCHITZ-Bedingung.

Wie in der Vorlesung "Einführung in die Numerische Mathematik" zeigt man das folgende Kriterium:

(26.8) Satz (Kriterium für LIPSCHITZ-Bedingung).

(i) Ist $D \subset \mathbb{R}^{n+1}$ konvex und sind die partiellen Ableitungen $\frac{\partial f_i}{\partial x_j}(x, y)$, $1 \leq i, j \leq n$ stetig und beschränkt in D, so genügt f einer LIPSCHITZ-Bedingung bezüglich y in D.

(ii) Ist $D \subset \mathbb{R}^{n+1}$ ein Gebiet und ist f in D stetig differenzierbar, so ist f LIPSCHITZ-stetig bezüglich y in D.

Beweis. vgl. "Einführung in die Numerische Mathematik". □

(26.9) Existenz- und Eindeutigkeitssatz von PICARD-LINDELÖF. Die Funktion $f : D \to \mathbb{R}^n$ sei auf dem Streifen $D = I \times \mathbb{R}^n$, $I \subset \mathbb{R}$ Intervall, stetig und genüge einer LIPSCHITZ-Bedingung bezüglich y auf D. Dann hat die AWA

$$y' = f(x, y),\ y(x_0) = y_0$$

für alle $(x_0, y_0) \in D$ eine eindeutig bestimmte Lösung $y : I \to \mathbb{R}^n$.

Beweisidee. Setze $V = C(I, \mathbb{R}^n)$. Betrachte dann den Operator

$$y \in C(I, \mathbb{R}^n) \to Ty \in C(I, \mathbb{R}^n),\ (Ty)(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt.$$

Für einen Fixpunkt $y \in C(I, \mathbb{R}^n)$ mit $y = Ty$ gilt dann

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt\,\forall x \in I$$

sowie

$$y(x_0) = y_0, \quad y'(x) = f(x, y(x)),$$

y ist also eine Lösung der AWA. Dazu zeigt man, dass der Operator auf $C(I, \mathbb{R}^n)$ kontrahierend ist. Über eine gewichtete L_∞-Norm auf $C(I, \mathbb{R}^n)$ folgt dann die Behauptung. □

Anwendung auf lineare DGL $y' = A(x)y + b(x)$ liefert die Existenz der Lösung für alle $I \subset \mathbb{R}$. **Achtung:** $y' = 1 + y^2$, also $f(x, y) = 1 + y^2$ erfüllt keine LIPSCHITZ-Bedingung in einem Streifen.

Mit $a > 0$ und $b > 0$ betrachtet man einen Quader

$$Q = \{(x, y) \in \mathbb{R} \times \mathbb{R}^n \mid |x - x_0| \leq a, \, ||y - y_0|| \leq b\}.$$

(26.9a) Lokaler Existenz- und Eindeutigkeitssatz. Die Funktion f sei auf dem "Quader" $Q := \{(x, y) \in \mathbb{R} \times \mathbb{R}^n \mid |x - x_0| \leq a, \, ||y - y_0|| \leq b\}$ mit $a > 0$, $b > 0$ geeignet, stetig und genüge auf Q einer LIPSCHITZ-Bedingung bezüglich y. Sei $M := \max_{(x,y)\in Q} ||f(x, y)||$ und sei $\alpha := \min\{a, \frac{b}{M}\}$. Dann existiert genau eine Lösung der AWA $y' = f(x, y)$, $y(x_0) = y_0$ im Intervall $[x_0 - \alpha, x_0 + \alpha]$.

Anwendungen von (26.9a).

(i) $y' = e^{-x^2} + y^3$, $y(0) = 1$. Wähle $a = b = 1$ und $Q = [0, 1] \times [0, 2]$. Es ist

$$M = \max_{(x,y)\in Q} \left(e^{-x^2} + y^3\right) = 1 + 8 = 9, \quad \alpha = \min\left\{a, \frac{b}{M}\right\} = \min\left\{1, \frac{1}{9}\right\} = \frac{1}{9}.$$

Die Lösung der AWA existiert daher mindestens in $[0, \frac{1}{9}]$.

(ii) $y' = 1 + y^2$, $y(0) = 0$. Es ist $y(x) = \tan(x)$ für $x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$ die Lösung der DGL. Wählt man $a > 0$ und $b > 0$, so gilt

$$\alpha = \min\left\{a, \frac{b}{1 + b^2}\right\}.$$

Man maximiert nun $\frac{b}{1+b^2}$ nach b und erhält für $b = 1$ ein Maximum. Dann ist $\alpha = \frac{1}{2}$ für $b = 1$ und $a \geq \frac{1}{2}$ beliebig.

(26.10) Satz (Maximales Existenzintervall). Sei $D \subset \mathbb{R}^{n+1}$ offen und sei $f \in C(D, \mathbb{R}^n)$ LIPSCHITZ-stetig bezüglich y auf D. Zu $(x_0, y_0) \in D$ gibt es ein eindeutig bestimmtes maximales Existenzintervall $I_{\max} = (x_-, x_+)$ mit $-\infty \leq x_- < x_0 < x_+ \leq +\infty$, so dass die AWA $y' = f(x, y)$, $y(x_0) = y_0$ genau eine Lösung im offenen Intervall $I_{\max}$ besitzt. Die Lösung $y : I_{\max} \to \mathbb{R}^n$ kommt nach links und rechts "dem Rand von D beliebig nahe", d.h. es gilt eine der folgenden drei Bedingungen:

(i) $x_+ = +\infty$ oder $x_- = -\infty$;

(ii) $x_+ < +\infty$ und $\lim_{x\uparrow x_+} ||y(x)|| = \infty$;

(iii) $x_+ < +\infty$ und $\liminf_{x\uparrow x_+} d((x, y(x)), \delta D) = 0$.

Bei (ii) und (iii) gelten alternativ analoge Aussagen für x_-.

Erläuterungen.

(i) Sei $y' = A(x)y + b(x)$, so existiert die Lösung in $I_{\max} = \mathbb{R}$.

(ii) Sei $y' = 1 + y^2$, so existiert die Lösung $y(x) = \tan(x)$ in $I_{\max} = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

(iii) Sei $y' = \frac{1}{y}$, $y(0) = 1$, so existiert die Lösung $y(x) = \sqrt{2x + 1}$ in $I_{\max} = \left(-\frac{1}{2}, +\infty\right)$.

§ 27 Einschrittverfahren, Grundbegriffe

Sei $D \subset \mathbb{R}^{n+1}$ und sei $f : D \to \mathbb{R}^n$ eine C^p-Funktion mit $p \in \mathbb{N}$. Für $(x_0, y_0) \in D$ sei $y(x)$ die Lösung der AWA

$$y' = f(x, y),\ y(x_0) = y_0 \tag{27.1}$$

in einem geeigneten Intervall $I \subset \mathbb{R}$.

Diskretisierung von AWA

Man wählt eine Schrittweite $h \neq 0$ und ein Gitter $I_h := \{x_i = x_0 + ih \in I,\ i = 0, 1, \ldots\}$. Die exakten Werte lauten dann $y(x_i)$ für $x_i \in I_h$. Die approximierten Werte bezeichnet man mit $y_i \approx y(x_i)$.

Motivation: EULERsches Polygonzugverfahren

Für $x \in I$ gilt näherungsweise

$$\frac{y(x+h) - y(x)}{h} \approx y'(x) = f(x, y(x)).$$

Für die Näherungswerte y_i von $y(x_i)$ in $x_i = x_0 + ih$ erhält man die Rekursion

$$(27.2)\ \begin{cases} y_0 &= y(x_0), \\ y_{i+1} &= y_i + hf(x_i, y_i),\ i = 0, 1, \ldots \end{cases}$$

Begründung: $y(x_i + h) - y(x_i) \approx hf(x_i, y(x_i))$.

Beispiel. Wir betrachten $y' = 1 + y^2$, $y(0) = 0$ mit Schrittweite $h = 0.1$.

i	x_i	$y(x_i)$	y_i (EULER)	y_i (EULER verbessert)
0	0	0	0	0
1	0.1	0.1003	0.1000	0.1005
2	0.2	0.2027	0.2010	0.2030
3	0.3	0.3093	0.3050	0.3098
4	0.4	0.4228	0.4143	0.4234
5	0.5	0.5463	0.5315	0.5470

Allgemeine Einschrittverfahren.

Beispiele für Einschrittverfahren sind

- Verfahren von HEUN. $y_{i+1} = y_i + h \cdot \frac{1}{2} \cdot (f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i)))$.
- Modifiziertes EULER-Verfahren. $y_{i+1} = y_i + hf\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i)\right)$.
- RUNGE-KUTTA-Verfahren. Zunächst definiert man Funktionen

$$\begin{aligned} f_1 &= f(x_i, y_i), \\ f_2 &= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f_1\right), \\ f_3 &= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f_2\right), \\ f_4 &= f(x_i + h, y_i + hf_3). \end{aligned}$$

  Damit erhält man dann die Rekursion

$$y_{i+1} = y_i + \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4).$$

Man wähle eine Schrittweitenfunktion $f_h(x, y)$ (später mehr dazu) und erhält die Rekursion

$$(27.3)\ \begin{cases} y_0 &= y(x_0) \quad \text{wie in (27.1)} \\ y_{i+1} &= y_i + h \cdot f_h(x_i, y_i) \end{cases}$$

Es ergibt sich dann eine Näherungslösung $y_h(x)$ durch $y_h(x_i) = y_i$. $y_h(x)$ ist dabei für $x \in I_h$ definiert bzw. für festes $x \in I$ für

$$h \in H_x := \left\{ h = \frac{x - x_0}{m},\ m = 1, 2, \ldots \right\}.$$

Daraus ergibt sich die Rekursion

$$\text{(27.3a)} \left\{ \begin{array}{rcl} y_h(x_0) & = & y_0, \\ y_h(x+h) & = & y_h(x) + h f_h(x, y_h(x)) \end{array} \right.$$

Für den lokalen Diskretisierungsfehler des Verfahres (27.3) gilt mit $x \in I$

$$\tau_h(x) := \frac{y(x+h) - y(x)}{h} - f_h(x, y(x_i)), \tag{27.4}$$

d.h. für $x = x_i$ erhalten wir

$$\tau_h(x_i) = \frac{y(x_i + h - y(x_i)}{h} - f_h(x_i, y(x_i)).$$

Interpretation: $\tau_h(x_i)$ gibt an, wie gut die exakte Lösung $y(x)$ das Einschrittverfahren erfüllt.

(27.5) Definition.

(i) Das Einschrittverfahren (27.3) heißt konsistent, wenn gilt

$$\lim_{h \to 0} f_h(x, y) = f(x, y)\ \forall x \in I, y \in \mathbb{R}^n \text{ in einer geeigneten Umgebung von } y(x).$$

(ii) Das Einschrittverfahren (27.3) hat die Ordnung $p \in \mathbb{N}$, falls $||\tau_h(x)|| = \mathcal{O}(|h|^p)$, d.h. falls es ein $N \in \mathbb{R}_+$ gibt mit $||\tau_h(x)|| \leq N \cdot |h|^p$ für alle $x \in I$.

Bestimmung der Ordnung $p \in \mathbb{N}$.

Sei $y' = f(x, y(x))$ gegeben. Weiteres Differenzieren liefert

$$y'' = \frac{d}{dt} f(t, y(t))\big|_{t=x} = f_x(x, y(x)) + f_y(x, y(x)) \cdot \underbrace{y'(x)}_{= f(x, y(x))} = (f_x + f_y f)(x, y(x))$$

usw. zur Berechnung von $y^{(k)}(x)$ für $k = 3, 4, \ldots$.

Beispiel. Sei $y' = xy^2$. Wir erhalten dann

$$y'' = y^2 + x 2 y y' = y^2 + 2xyxy^2 = y^2 + 2x^2 y^3$$

und

$$y''' = 2yy' + 4xy^3 + 6x^2 y^2 y' = 2xy^3 + 4xy^3 + 6x^3 y^4 = 6xy^3(1 + x^2 y).$$

(Hinweis: Verwende die Produktregel $(y^2)' = (yy)' = yy' + y'y = 2yy'$, da y eine Funktion ist!)

Die TAYLOR-Entwicklung von $y(x+h)$ liefert

$$\frac{y(x+h) - y(x)}{h} = y'(x) + \frac{h}{2!} y''(x) + \ldots + \frac{h^{p-1}}{p!} y^{(p)}(x + \vartheta h) \tag{27.6}$$

mit $0 < \vartheta < 1$ geeignet. Für das EULER-Verfahren gilt $f_h(x, y) = f(x, y)$; daher gilt für den lokalen Diskretisierungsfehler

$$\tau_h(x) = \frac{y(x+h) - y(x)}{h} - \underbrace{f_h(x, y(x))}_{= y'} \overset{(26.7)}{=} \frac{h}{2} (f_x + f_y f)(x, y(x)) + \mathcal{O}(h^2) = \mathcal{O}(h).$$

Generelle Idee zur Gewinnung von Einschrittverfahren höherer Ordnung.

Man nehme für $f_h(x,y)$ einen Abschnitt der TAYLOR-Entwicklung in (27.6). Sei zum Beispiel

$$f_h(x,y) = f(x,y) + \frac{h}{2}\left(f_x(x,y) + f_y(x,y)f(x,y)\right),$$

so erhält man $p = 2$ als Ordnung des lokalen Diskretisierungsfehlers.

Ansatz für ein Einschrittverfahren 2. Ordnung.

Man setze

$$f_h(x,y) = c_1 f(x,y) + c_2 f(x + \alpha_2 h, y + h\beta_2 f(x,y))$$

mit Konstanten $c_1, c_2, \alpha_2, \beta_2 \geq 0$. TAYLOR-Entwicklung bezüglich h in $h = 0$ liefert

$$\begin{aligned} f_h(x,y) &= c_1 f(x,y) + c_2 f(x,y) + c_2 h\left(\alpha_2 f_x(x,y) + \beta_2 f_y(x,y) f(x,y)\right) + \mathcal{O}(h^2) \\ &= (c_1 + c_2) f(x,y) + h(c_2\alpha_2 f_x + c_2\beta_2 f_y f)(x,y) + \mathcal{O}(h^2). \end{aligned}$$

Ein Vergleich mit (27.6) ergibt folgende Bedingungen für ein Verfahren 2. Ordnung:

$$c_1 + c_2 = 1, \quad c_2\alpha_2 = \frac{1}{2}, \quad c_2\beta_2 = \frac{1}{2}.$$

Für das Verfahren von HEUN setzt man $c_1 = c_2 = \frac{1}{2}$ und $\alpha_2 = \beta_2 = 1$. Dann gilt

$$f_h(x,y) = \frac{1}{2}(f(x,y) + f(x+h, y + hf(x,y)). \tag{27.7}$$

Für das modifizierte (verbesserte) EULER-Verfahren setzt man $c_1 = 0$, $c_2 = 1$ und $\alpha_2 = \beta_2 = \frac{1}{2}$. Dann gilt

$$f_h(x,y) = f\left(x + \frac{h}{2}, y + \frac{h}{2} f(x,y)\right). \tag{27.8}$$

Das Verfahren von RUNGE-KUTTA.

Dieses Verfahren wird definiert über

$$(27.9)\quad \begin{cases} f_h(x,y) &= \frac{1}{6}\left(f_1(x,y) + 2f_2(x,y) + 2f_3(x,y) + f_4(x,y)\right) \\ f_1(x,y) &= f(x,y) \\ f_2(x,y) &= f\left(x + \frac{h}{2}, y + \frac{h}{2} f_1(x,y)\right) \\ f_3(x,y) &= f\left(x + \frac{h}{2}, y + \frac{h}{2} f_2(x,y)\right) \\ f_4(x,y) &= f(x + h, y + h f_3(x,y)) \end{cases}$$

Man kann zeigen, dass dieses Verfahren die Ordnung $p = 4$ besitzt. Das RUNGE-KUTTA-Verfahren lässt sich auch allgemeiner definieren:

$$\begin{aligned} f_h(x,y) &= \left(c_1 f_1 + c_2 f_2 + \ldots + c_m f_m\right)(x,y) \\ f_1(x,y) &= f(x,y) \\ f_2(x,y) &= f\left(x + \alpha_2 h, y + \beta_{2,1} f_1(x,y)\right) \\ &\vdots \\ f_m(x,y) &= f(x + \alpha_m h, y + h(\beta_{m,1} f_1(x,y) + \ldots + \beta_{m,m-1} f_{m-1}(x,y)) \end{aligned}$$

Koeffizientenschema (BUTCHER-Schema)

$$\begin{array}{c|ccccc} 0 & & & & & \\ \alpha_2 & \beta_{2,1} & & & & \\ \vdots & & & & & \\ \alpha_m & \beta_{m,1} & \beta_{m,2} & \cdots & \beta_{m,m-1} & \\ \hline & c_1 & c_2 & \cdots & c_{m-1} & c_m \end{array}$$

Es gilt immer $\sum_{i=1}^m c_i = 1$ und $a_k = \sum_{j=1} k - 1\beta_{k,j}$ für $k = 2, \ldots, m$.

Beispiele.

$\underline{m=1}$

0	
	1

EULER-Schema $y_{i+1} = y_i + hf(x_i, y_i)$ mit Ordnung $p = 1$.

$\underline{m=2}$

0		
1	1	
	1/2	1/2

HEUN-Verfahren $y_{i+1} = y_i + \frac{h}{2}\left(f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i))\right)$ mit Ordnung $p = 2$.

$\underline{m=3}$

0			
1/2	1/2		
1	−1	2	
	1/6	4/6	1/6

$y_{i+1} = y_i + \frac{h}{6}\left(f_1 + 4f_2 + f_3\right)(x_i, y_i)$ mit

$$\begin{aligned} f_1(x,y) &= f(x,y), \\ f_2(x,y) &= f\left(x + \frac{h}{2}, y + \frac{h}{2} f(x,y)\right), \\ f_3(x,y) &= f\left(x + h, y - hf(x,y) + 2hf\left(x + \frac{h}{2}, y + \frac{h}{2} f(x,y)\right)\right) \end{aligned}$$

und Ordnung $p = 3$.

$\underline{m=4}$

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

RUNGE-KUTTA-Verfahren mit Ordnung $p = 4$.

§ 28 Konvergenz von Einschrittverfahren

Sei $f : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ eine C^p-Funktion, $p \in \mathbb{N}$.Sei $y(x)$ eine Lösung der AWA $y' = f(x,y)$, $y(x_0) = y_0$ mit $x_0 \in [a,b]$ und $y_0 \in \mathbb{R}^n$. Wählt man ein $x \in I := [a,b]$ fest, so erhält man den sogenannten globalen Diskretisierungsfehler

$$e(x,h) = y_h(x) - y(x). \tag{28.1}$$

Dabei ist $y_h(x)$ die Näherungslösung definiert für

$$h' \in H_x := \left\{ \frac{x - x_0}{m} = h,\ m = 1,2,\ldots \right\}.$$

Erinnerung: $y_h(\tilde{x} + h) = y_h(\tilde{x}) + h f_h(\tilde{x}, y_h(\tilde{x}))$ für $\tilde{x} \in [a,b]$.

Das Einschrittverfahren (27.3) heißt konvergent, wenn für alle $x \in I = [a,b]$ gilt

$$\lim_{m\to\infty} (e(x,h_m)) = y_{h_m}(x) - y(x) = 0,$$

wobei $h_m = \frac{x-x_0}{m}$. Erinnerung: Für den lokalen Diskretisierungsfehler gilt

$$\tau_h(x) = \frac{1}{h}(y(x+h) - y(x)) - f_h(x, y(x)).$$

Wir wollen zeigen: Falls $||\tau_h(x)|| = \mathcal{O}(|h|^p)$, so gilt auch

$$||e(x,h_m)|| = \mathcal{O}(|h_m|^p).$$

(28.2) Hilfssatz. Für die Zahlen $d_i \geq 0$, $\delta > 0$ und $c \geq 0$ gelte die Abschätzungh

$$d_{i+1} \leq (1+\delta)d_i + c$$

für $i = 0, 1, \ldots$. Dann folgt die Abschätzung

$$d_m \leq e^{m\delta} d_0 + \frac{e^{m\delta - 1}}{\delta} c.$$

Beweis. Rekursiv erhält man zunächst

$$\begin{aligned} d_1 &\leq (1+\delta)d_0 + c \\ d_2 &\leq (1+\delta)d_1 + c \leq (1+\delta)^2 d_0 + c(1+\delta) + c \end{aligned}$$

womit induktiv folgt

$$\begin{aligned} d_m &\leq (1+\delta)^m d_0 + c(1 + (1+\delta) + \ldots + (1+\delta)^{m-1}) = (1+\delta)^m d_0 + c\frac{(1+\delta)^m - 1}{\delta} \\ &\leq e^{m\delta} d_0 + c\frac{e^{m\delta} - 1}{\delta} \end{aligned}$$

□

(28.3) Konvergenzsatz. Die Funktion $f_h(x,y)$ sei stetig auf

$$G := \{(x,y,h) \mid x \in I = [a,b], ||y - y(x)|| \leq \alpha, |h| \leq h_0\}$$

mit $\alpha > 0$, $h_0 > 0$ geeignet. Es existieren Konstanten $N > 0$ und $M > 0$, so dass die Abschätzungen

(i) $||f_h(x,\tilde{y} - f_h(x,y)|| \leq M \cdot ||\tilde{y} - y||$ für alle $(x,\tilde{y},h)$,$(x,y,h) \in G$,

(ii) $||\tau_h(x)|| \leq N \cdot |h|^p$ für ale $x \in I = [a,b]$ und $|h| \leq h_0$,

gelten. Dann gibt es eine Schrittweite $\bar{h} > 0$, so dass für den globalen Diskretisierungsfehler die Abschätzung

$$||e(x,h_m)|| \leq \frac{e^{M\cdot|x-x_0|-1}}{M} \cdot \max_i ||\tau_{h_m})x_i)|| \leq \frac{e^{M\cdot|x-x_0|-1}}{M} \cdot N \cdot |h_m|^p$$

für alle $x \in I$ und $h_m = \frac{x-x_0}{m}$ mit $|h_m| \leq \bar{h} \leq h_0$ gilt.

Beweis. Sei $x \in I$ fest, $x \neq x_0$, $h = h_m = \frac{x-x_0}{m}$ für $m = 1, 2, \ldots$. Für den lokalen Diskretisierungsfehler gilt

$$\tau_h(x_i) = \frac{1}{h}\left(y(x_i + h) - y(x_i)\right) - f_h(x_i, y(x_i)).$$

Dies bedeutet

$$y(x_{i+1}) = y(x_i) + h f_h(x_i, y(x_i)) + h\tau_h(x_i).$$

Aus dem Einschrittverfahren ergibt sich

$$y_{i+1} = y_i + h f_h(x_i, y_i).$$

Definiert man nun $d_i := ||y_i - y(x_i)||$, er ergibt sich durch Subtraktion obiger Gleichungen die Abschätzung

$$d_{i+1} \leq d_i + |h| \cdot ||f_h(x_i, y_i) - f_h(x_i, y(x_i))|| + |h| \cdot ||\tau_h(x_i)||\,.$$

Falls $||y_i - y(x_i)|| \leq \alpha$ gilt, so folgt mit Voraussetzung (a)

$$d_{i+1} \leq d_i + |h| \cdot M \cdot d_i + |h| \cdot \max_i ||\tau_h(x_i)||\,.$$

Wir setzen dann $\delta := 1 + |h| \cdot M$ und $c = |h| \cdot \max_i ||\tau_h(x_i)||$. Weiterhin ergibt sich $d_0 = ||y_0 - y(x_0)|| = 0$ und $m\delta = m \cdot |h_m| \cot M = |x - x_0| \cdot M$. Daraus schließen wir mit Hilfssatz (28.2) auf die gesuchte Abschätzung:

$$\begin{aligned} d_m &= ||e(x, h_m)|| \leq \frac{e^{m\delta} - 1}{\delta} c = \frac{e^{M|x-x_0|} - 1}{|h|M} |h| \cdot \max_i ||\tau_h(x_i)|| \\ &= \frac{e^{M|x-x_0|} - 1}{M} \max_i ||\tau_h(x_i)|| \overset{(b)}{\leq} \frac{e^{M|x-x_0}}{M} \cdot N \cdot |h|^p \end{aligned}$$

Nachtrag: Es gilt $\frac{e^{M|x-x_0|}-1}{M} \cdot N \cdot |h|^p \leq \alpha$ für alle $|h| \leq \bar{h}$, $\bar{h} > 0$ geeignet, da bis auf h alle Werte konstant sind. □

Es drängt sich die Vermutung auf, dass für den Fehler eine asymptotische Entwicklung

$$y_h(x) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \ldots \tag{28.4}$$

existiert.

(28.5) Satz (von Gragg). Sei f eine C^{N+2}-Funktion und sei $y_h(x)$ die von einem Einschrittverfahren der Ordnung p gelieferte Näherungslösung für die Lösung $y(x)$ der AWA $y' = f(x, y)$ mit $y(x_0) = y_0$. Dann gibt es Funktionen $e_k(x)$, $k = p, p+1, \ldots, N$ und $E_{N+1}(x, h)$ mit

$$y_h(x) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \ldots + h^N e_N(x) + h^{N+1} E_{N+1}(x, h)$$

für alle $x \in I = [a, b]$ und $h = h_m = \frac{x-x_0}{m}$, $m = 1, 2, \ldots$. Das Restglied $E_{N+1}(x, h)$ ist bei festem $x \in I$ beschränkt für alle $h = h_m$.

Bemerkung. Wendet man diesen Satz auf Extrapolationsaufgaben an, so erhält man die Entwicklung (24.1) mit $\gamma = 1$.

§ 29 Spezielle lineare Mehrschrittverfahren

Beispiel. $y_{j+2}-y_j = 2hf(x_{j+1}, y_{j+1})$. Allgemein hat ein lineares Mehrschrittverfahren (MSV) die Form

$$\sum_{k=0}^{m} a_k y_{j+k} = h \cdot \sum_{k=0}^{m} b_k f(x_{j+k}, y_{j+k}) \tag{29.1}$$

für $j = 0, 1, \ldots$. Dies ist ein MSV der Stufe m. Als Startwerte wählt man y_0 (Anfangswert der AWA) und $y_i = \bar{y}_i$ für $i = 1, \ldots, m-1$.

Beispiel. Mit $m = 1$, $a_1 = 1$, $a_0 = -1$, $b_1 = 0$ und $b_0 = 1$ erhält man das EULER-Verfahren

$$y_{j+1} - y_j = hf(x_j, y_j).$$

Es sei in (29.1) ohne Einschränkung $a_m = 1$. Wir erhalten die folgenden Fälle:

- $b_m = 0$. Dies nennt man ein explizites MSV, da man y_{j+m} direkt aus den $y_j, \ldots, y_{j+m-1}$ berechnen kann.
- $b_m \neq 0$. Dies nennt man ein implizites MSV, da y_{j+m} auf beiden Seiten der Gleichung (29.1) auftritt. Als Lösungsidee erhält man, dass y_{j+m} die Lösung einer nichtlinearen Fixpunktgleichung

  $$y_{j+m} = -a_{m-1}y_{j+m-1} - \ldots - a_0 y_j + h(b_m f(x_{j+m}, y_{j+m}) + \ldots + b_0 f(x_j, y_j)) =: g(y_{j+m}) \tag{29.2}$$

  ist. Wie in der Numerik I kann man eine Fixpunktiteration betrachten:

  $$y_{j+m}^{(k+1)} = g\left(y_{j+m}^{(k)}\right),\ k = 0, 1, \ldots$$

  Wegen

  $$\frac{\partial g}{\partial y}\left(y_{j+m}^{(k)}\right) = hb_m \underbrace{\frac{\partial f}{\partial g}\left(x_{j+m}, y_{j+m}^{(k)}\right)}_{\text{beschränkt}}$$

  folgt

  $$\left|\left|\frac{\partial g}{\partial y}\left(y_{j+m}^{(0)}\right|\right|\right) \leq q < 1$$

  für $|h|$ genügend klein. Den Startwert $y_{j+m}^{(0)}$ kann man durch ein explizites MSV gewinnen.

Das explizite Verfahren heißt auch Prädikator-Verfahren, das implizite nennt man Korrektor-Verfahren.

Konstruktion von speziellen linearen MSV

Es seien jeweils geeignete Indizes q, p und k gewählt.

Integration.

$$y(x_{j+k}) - y(x_{j-p}) = \int_{x_{j-p}}^{x_{j+k}} f(x, y(x))dx$$

Interpolation. Für den Index q sei $P_q(x)$ das interpolierende Polynom mit

(i) grad $(P_q) \leq q$,

(ii) $P_q(x_i) = f(x_i, y(x_i))$, $i = j, j-1, \ldots, j-q$.

Nach der LAGRANGE-Interpolation ergibt sich dann

$$P_q(x) = \sum_{i=0}^{q} f(x_{j-i}, y(x_{j+i}))L_i(x),$$

$$L_i(x) = \prod_{l=0, l\neq i}^{q} \frac{x - x_{j-l}}{x_{j-i} - x_{j-l}}$$

Als Näherung erhält man

$$\begin{aligned} y(x_{j+k}) - y(x_{j-p}) &\approx \sum_{i=0}^{q} f(x_{j-i}, y(x_{j-i})) \cdot \int_{x_{j-p}}^{x_{j+k}} L_i(x)dx \\ &= h \cdot \sum_{i=0}^{q} \beta_{q,i} f(x_{j-i}, y(x_{j-i})) \end{aligned}$$

wobei $\beta_{q,i} = \frac{1}{h} \int_{x_{j-p}}^{x_{j+k}} L_i(x)dx$ ist. Als Ansatz für die Näherungswerte y_i von $y(x_i)$ wählt man

$$y_{j+k} - y_j = h \sum_{i=0}^{q} \beta_{q,i} f(x_{j-i}, y_{j-i}). \tag{29.3}$$

Im Folgenden sei $f_k := f(x_k, y_k)$. Wir erhalten nun weitere Verfahren.

Die Verfahren von Adams-Bashforth. Setze $k = 1$, $p = 0$, $q = 0, 1, 2, \ldots$.

$$y_{j+1} - y_j = h(\beta_{q,0} f_j + \beta_{q,1} f_{j-1} + \ldots + \beta_{q,q} f_{j-q}). \tag{29.4}$$

Dies ist ein explizites MSV der Stufe $m = q + 1$. Für $q = 0$ gewinnt man das Euler-Verfahren zurück. Für die β erhält man die Zahlenwerte

$\beta_{q,i}/i$	0	1	2	3
$\beta_{0,i}$	1			
$2 \cdot \beta_{1,i}$	3	-1		
$12 \cdot \beta_{2,i}$	23	-16	5	
$24 \cdot \beta_{3,i}$	55	-59	37	-9

Die Verfahren von Adams-Houlton. Setze $k = 0$, $p = 1$, $q = 0, 1, 2, \ldots$ und ersetze dann $j \to j+1$:

$$y_{j+1} - y_j = h(\beta_{q,0} f_{j+1} + \beta_{q,1} f_j + \ldots + \beta_{q,q} f_{j+1-q}). \tag{29.5}$$

Dies ist ein implizites MSV der Stufe $m = q$. Für die β erhält man die Zahlenwerte

$\beta_{q,i}/i$	0	1	2	3
$\beta_{0,i}$	1			
$2 \cdot \beta_{1,i}$	1	1		
$12 \cdot \beta_{2,i}$	5	8	-1	
$24 \cdot \beta_{3,i}$	9	19	-5	1

Erinnerung: y_{j+1} kann durch Fixpunktiteration bestimmt werden.

Die Verfahren von Nyström. Setze $k = 1$, $p = 1$, $q = 0, 1, 2, \ldots$.

$$y_{j+1} - y_{j-1} = h(\beta_{q,0} f_j + \beta_{q,1} f_{j-1} + \ldots + \beta_{q,q} f_{j-q}). \tag{29.6}$$

Dies ist ein explizites MSV der Stufe $m = q + 1$. Im Spezialfall $q = 0$ und $\beta_{0,0} = 2$ erhält man mit

$$y_{j+1} - y_{j-1} = 2hf(x_j, y_j) \tag{29.7}$$

die bekannte Mittelpunktsregel zurück.

§ 30 Allgemeine lineare Mehrschrittverfahren

Ein lineares Mehrschrittverfahren hat die Form

$$\sum_{k=0}^{m} a_k y_{j+k} = h \cdot \sum_{k=0}^{m} b_k f(x_{j+k}, y_{j+k}) \tag{30.1}$$

mit oBdA $a_m = 1$. Der lokale Diskretisierungsfehler ergibt sich als Erweiterung von (27.4) zu

$$\tau_h(x) = \frac{1}{h} \sum_{k=0}^{m} a_k y(x+kh) - \sum_{k=0}^{m} b_k f(x+kh, y(x+kh)), \tag{30.2}$$

wobei $y(x)$ die exakte Lösung meint. Wir führen eine abkürzende Schreibweise

$$L_h(y(x)) = \sum_{k=0}^{m} a_k y(x-kh) - h \sum_{k=0}^{m} b_k f(x+kh, y(x+kh))$$

ein. Damit gilt dann

$$\tau_h(x) = \frac{1}{h} L_h(y(x)).$$

Eine TAYLOR-Entwicklung von $L_h(y(x))$ und $\tau_h(x)$ bezüglich h in $h = 0$ führt zu

$$L_h(y(x)) = c_0 y(x) + c_1 h y'(x) + \ldots + c_p h^p y^{(p)}(x) + c_{p+1} h^{p+1} y^{(p+1)}(x)\,(1 + \mathcal{O}(h))$$

mit

$$\left.\begin{array}{rcl} c_0 & = & a_0 + \ldots + a_m \\ c_1 & = & a_1 + 2a_2 + \ldots + m a_m - (b_0 + \ldots + b_m) \\ c_p & = & \frac{1}{p!}\left(a_1 + 2^p a_2 + \ldots + m^p a_m\right) - \frac{1}{(p-1)!}\left(b_1 + 2^{p-1} b_2 + \ldots + m^{p-1} b_m\right) \end{array}\right\} \tag{30.3}$$

für $p \geq 1$.

Beispiel. Setze $m = 1$, $a_1 = 1$, $a_0 = 0$, $b_1 = 0$ und $b_0 = 1$, so erhält man mit

$$c_0 = a_0 + a_1 = 0, \quad c_1 = a_1 - (b_0 + b_1) = 0$$

das bekannte EULER-Verfahren zurück.

Insgesamt erhält man

$$\tau_h(x) = \frac{1}{h} L_h(y(x)) = \frac{1}{h} c_0 y(x) + c_1 y'(x) + c_2 h y''(x) + \ldots + c_p h^{p-1} y^{(p)}(x) + c_{p+1} h^p y^{(p+1)}\,(1 + \mathcal{O}(h))\,.$$

(30.4) Definition. Das lineare MSV (30.1) hat die Ordnung p (ist konsistent von der Ordnung p), wenn gilt

$$c_0 = \ldots = c_p = 0, \quad c_{p+1} \neq 0,$$

d.h. wenn gilt

$$||\tau_h(x)|| \in \mathcal{O}(|h|^p)$$

für alle $x \in I = [a, b]$.

Eine zentrale Rolle spielen die Polynome

$$\rho(\lambda) := \sum_{k=0}^{m} a_k \lambda^k, \quad \sigma(\lambda) := \sum_{k=0}^{m} b_k \lambda^k, \quad \lambda \in \mathbb{C} \tag{30.5}$$

Für die Ordnung $p = 1$ gilt $c_0 = \rho(1) = 0$ und $c_1 = \rho'(1) - \sigma(1) = 0$. Als zentrale Idee betrachtet man die $2m + 1$ freien Parameter $a_0, \ldots, a_{m-1}, b_0, \ldots, b_m$ ($a_m = 1$!) und bestimmt diese so, dass die Ordnung p maximal wird.

Beispiel. Gesucht ist ein explizites lineares MSV mit $m = 3$ und maximaler Ordnung mit $a_0 = a_2 = 0$. Die Relationen (30.3) ergeben daher ein Lineares Gleichungssystem für a_1, b_0, b_1 und b_2, wobei $a_3 = 1$ und $b_3 = 0$ gilt:

$$\begin{aligned} c_0 &= a_1 + a_3 = 0 \\ c_1 &= a_1 + 3a_3 - (b_0 + b_1 + b_2) = 0 \\ 2c_2 &= a_1 + 9a_3 - 2(b_1 + 2b_2) = 0 \\ 6c_3 &= a_1 + 27a_3 - 3(b_1 + 4b_2) = 0 \end{aligned}$$

Es ergibt sich daraus $a_1 = -1$, $b_0 = \frac{1}{3}$, $b_1 = -\frac{2}{3}$ und $b_2 = \frac{7}{3}$. Die Ordnung des Verfahrens ist 4, da man $c_4 \neq 0$ nachrechnet. Ersetzt man weiter $j + 3 \to j + 1$, so erhält man mit

$$y_{j+1} - y_{j-1} = \frac{h}{3}(7f_j - 2f_{j-1} + f_{j-2})$$

ein NYSTRÖM-Verfahren aus (29.2) mit $q = 2$.

Konvergenz und Stabilität von linearen MSV

Die Fehler in den Startdaten $y_0, \ldots, y_{m-1}$ seien ε_i, $i = 0, \ldots, m-1$ mit $\varepsilon_i := y_i - y(x_i)$. Weiter sei $\varepsilon := (e_0, \ldots, e_{m-1})^t \in \mathbb{R}^m$. Die Näherungslösung $y_{k,\varepsilon}(x)$ ist definiert für alle $x \in I_h = \{x_o + ih,\, i = 0, 1, \ldots\}$ bzw. für festes $x \in [a, b]$ und $h = \frac{x - x_0}{k}$ mit $k = 1, 2, \ldots$.

(30.6) Definition. Das lineare MSV (30.1) heißt konvergent von der Ordnung $p \in \mathbb{N}_+$, wenn

$$||y_{k,\varepsilon}(x) - y(x)|| \in \mathcal{O}(|h|^p)$$

für $x \in [a, b]$, $h = \frac{x - x_0}{k}$ mit $k = 1, 2, \ldots$ und $\varepsilon = \varepsilon(h) \in \mathcal{O}(|h|^p)$ für die Fehler in den Eingabedaten gilt. **Achtung:** Von der Konsistenz zur Ordnung p kann man im Allgemeinen nicht auf Konvergenz zur Ordnung p schließen!

Gegenbeispiel. Das $(m = 2)$-Schrittverfahren

$$y_{i+2} + 4y_{i+1} - 5y_i = h(4f_{j+1} + 2f_j)$$

hat die Ordnung $p = 3$, was man mit (30.3) zeigen kann. Man wende dieses Verfahren auf die AWA $y' = -y$ mit $y(0) = 1$ an. Die exakte Lösung lautet $y(x) = e^{-x}$. Man erhält damit Startwerte $y_0 = 1$ und $y_1 = e^{-h}$. Betrachten wir $h = \frac{1}{100}$, so ergibt sich

j	$y_j - y(x_j)$
2	$-0.164 \cdot 10^{-8}$
4	$-0.300 \cdot 10^{-7}$
$\vdots$	$\vdots$
97	$0.512 \cdot 10^{58}$
98	$-0.257 \cdot 10^{59}$
99	$0.129 \cdot 10^{60}$
100	$-0.652 \cdot 10^{60}$

Als Begründung für dieses Verhalten stellen wir fest, dass $\lambda = -5$ eine Nullstelle von $\rho(\lambda)$ ist. Als Fazit ziehen wir, dass die Nullstellen des Polynoms $\rho(\lambda)$ in der Stabilitätsbetrachtung eine zentrale Rolle spielen.

(30.7) Definition. Das lineare MSV (30.1) heißt stabil, wenn für die Nullstellen $\lambda \in \mathbb{C}$ des Polynoms $\rho(\lambda) = \sum_{k=0}^{m} a_k \lambda^k$ gilt:

(i) $|\lambda| \leq 1$,

(ii) $|\lambda| = 1 \Rightarrow \lambda$ ist eine einfache Nullstelle von $\rho(\lambda)$.

Beispiele.

1. Adams-Bashforth (29.4)

   $$y_{j+m} - y_{j+m-1} = h(...).$$

   Es ist

   $$\rho(\lambda) = \lambda^m - \lambda^{m-1} = \lambda^{m-1}(\lambda - 1)$$

   und damit sind $\lambda_1 = 0$, $\lambda_2 = 1$ die Nullstellen. Da λ_2 eine einfache Nullstelle ist, ist das Verfahren stabil.

2. Nyström (29.7)

   $$y_{j+m} - y_{j+m-2} = h(...).$$

   Es ist

   $$\rho(\lambda) = \lambda^m - \lambda^{m-2} = \lambda^{m-2}(\lambda^2 - 1)$$

   und damit sind $\lambda_1 = 0$, $\lambda_2 = 1$ und $\lambda_3 = -1$ die Nullstellen. Da λ_2 und λ_3 jeweils einfache Nullstellen sind, ist das Verfahren stabil.

Man beachte, dass die Stabilitätsbedingung die Ordnung p einschränkt:

(30.8) Satz (von Dahlquist). Für die Ordnung p eines stabilen MSV gilt

$$p \leq \begin{cases} m+1, & \text{falls } m \text{ ungerade,} \\ m+2, & \text{falls } m \text{ gerade} \end{cases}.$$

Beweis. s. R.D. Grigorieff: Numerik gewöhnlicher DGL Bd. 2. □

Beispiele.

1. Nyström mit $q = 0$:

   $$y_{j+1} - y_{j-1} = 2hf_j \quad \text{Mittelpunktsregel}$$

   Dann ist $m = 2, p = 3$, es handelt sich also nicht um ein maximales Verfahren.

2. Adams-Moulton mit $m = q$ ungerade ist von der Ordnung $p = q + 1$ und somit maximal.

Im Folgenden untersuchen wir die homogene Differenzengleichung

$$\sum_{k=0}^{m} a_k z_{j+k} = 0 \tag{30.9}$$

Dabei sei ab sofort $n = 1$, also $z_{j+k} \in \mathbb{C}$ (n = Dimension des Systems). Als Lösungsansatz wählen wir $z_j = \lambda^j$ mit $\lambda \in \mathbb{C}$. Dies ist eine Lösung von (30.9), falls

$$0 = \sum_{k=0}^{m} a_k \lambda^{j+k} = \lambda^j \sum_{k=0}^{m} a_k \lambda^k = \lambda^j \rho(\lambda)$$

gilt. Es folgt dann $\rho(\lambda) = 0$, da $\lambda = 0$ uninteressant ist.

Im Spezialfall, dass λ eine zweifache Nullstelle von $\rho(\lambda)$ ist, gilt

$$\rho'(\lambda) = \sum_{k=1}^{m} a_k k \lambda^{k-1} = 0.$$

Damit berechnet man

$$\begin{aligned} \sum_{k=0}^{m} a_k (j+k) \lambda^{j+k} &= \lambda^j \left(j \sum_{k=0}^{m} a_k \lambda^k + \lambda \sum_{k=1}^{m} a_k k \lambda^{k-1} \right) \\ &= \lambda^j \left(j \underbrace{\rho(\lambda)}_{=0} + \lambda \underbrace{\rho'(\lambda)}_{=0} \right) = 0 \end{aligned}$$

Folgerung: $z_j = j\lambda^j$ ist eine Lösung von (30.9), falls λ eine zweifache Nullstelle von $\rho(\lambda)$ ist.

Im allgemeinen Fall einer r-fachen Nullstelle λ von $\rho(\lambda)$ sind $z_j = \lambda^j, \ldots, z_j = j\lambda^j, \ldots, z_j = j^{r-1}\lambda^j$ Lösungen von (30.9). Damit sind bereits alle Lösungen von (30.9) gefunden.

(30.10) Satz. Seien $\lambda_1, \ldots, \lambda_n$ die Nullstellen von $\rho(\lambda)$ mit Vielfachheiten $\nu_1, \ldots, \nu_l$ und sei $m = \nu_1 + \ldots + \nu_l$. Dann sind

$$\lambda_k^j, j\lambda_k^j, \ldots, j^{\nu_k - 1}\lambda_k^j$$

für $k = 1, \ldots, l$ Lösungen der Differenzengleichung (30.9). Jede weitere Lösung z_j von (30.9) ist eine Linearkombination dieser Lösungen.

(30.12) Korollar. Das lineare MSV (30.1) ist genau dann stabil, wenn alle Lösungen der Differenzengleichung $\sum_{k=0}^m a_k z_{j+k}$ für $j \to \infty$ beschränkt bleiben.

(30.12) Satz. Ist das MSV (30.1) konvergent für die AWA $y' = 0$, $y(0) = 0$, so ist es stabil.

Beweis. Sei λ eine Nullstelle von $\rho(\lambda)$ mit Vielfachheit ν. Vorgegeben seien Anfangswerte $y_j = \varepsilon_j(h) := j^{\nu-1}\lambda^j h$ für $j = 0, \ldots, m-1$. Das MSV für $y' = 0$, $y(0) = 0$ lautet

$$\sum_{k=0}^m a_k y_{j+k} = 0.$$

Nach Satz (30.10) muss gelten

$$y_j = j^{\nu-1}\lambda^j h$$

für alle $j = 0, 1, \ldots$. Sei $x \in [a, b]$ fest gewählt und sei $h = h_n = \frac{x}{n}$ für $n = 1, 2, \ldots$. Da das MSV konvergent ist, folgt

$$\lim_{n\to\infty} \left(y_n = n^{\nu-1}\lambda^n \frac{x}{n} \right) = 0.$$

Dies ist nur möglich, wenn $|\lambda| \leq 1$ und $\nu = 1$ für $|\lambda| = 1$ gilt, d.h. das MSV ist stabil. □

(30.13) Satz. Zu Vereinfachung sei $n = 1$. Die Funktion f genüge auf

$$G := \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, |y - y(x)| \leq \alpha\}, \quad \alpha > 0$$

einer Lipschitz-Bedingung bezüglich y. Das MSV sei stabil und für den Fehler $\varepsilon := (\varepsilon_0, \ldots, \varepsilon_{m-1})^t \in \mathbb{R}^m$ mit $\varepsilon_i := y_i - y(x_i)$ und den lokalen Diskretisierungsfehler $\tau_h(x)$ gelten die Abschätzungen

$$\begin{aligned} |\varepsilon_i| \leq c(h),\ i = 0, \ldots, m-1 \quad &\text{und} \quad \lim_{h\to 0} c(h) = 0, \\ |\tau_h(x)| \leq d(h),\ x \in [a, b] \quad &\text{und} \quad \lim_{h\to 0} d(h) = 0. \end{aligned}$$

Dann gibt es Konstanten c_1, c_2 und eine Schrittweite $h_0 > 0$ mit $|y_{k,\varepsilon}(x) - y(x)| \leq c_1 \cdot e^{c_2|x-x_0|} \cdot (c(h) + d(h))$ für alle $x \in [a, b]$ und $h = \frac{x-x_0}{n}$, $n = 1, 2, \ldots$, mit $|h_n| \leq h_0$.

Beweis. Es gelten die Relationen

$$\begin{aligned} \sum_{k=0}^m a_k y_{j+k} &= h \sum_{k=0}^m b_k f(x_{j+k}, y_{j+k}), \\ \sum_{k=0}^m a_k y(x_{j+k}) &= h \sum_{k=0}^m b_k f(x_{j+k}, y(x_{j+k})) + h\tau_h(x_j) \end{aligned}$$

nach Definition des lokalen Diskretisierungsfehlers $\tau_h(x_j)$. Für den Fehler $e_j := y_j - y(x_j)$, $i = 0, 1, \ldots$ gilt $e_j = \varepsilon_j$ für $j = 0, \ldots, m-1$. Es folgt durch Differenzenbildung

$$\sum_{k=0}^m a_k e_{j+k} = h \sum_{k=0}^m b_k \left(f(x_{j+k}, y_{j+k}) - f(x_{j+k}, y(x_{j+k})) \right) - h\tau_h(x_j) =: h \cdot g_j \tag{30.14}$$

Sei $L > 0$ eine Lipschitz-Konstante von f bezüglich y. **Voraussetzung:** Es gelte $|e_k| = |y_k - y(x_k)| \leq \alpha$ für $k = j, \ldots, j+m$. Diese Behauptung werden wir später im Beweis verifizieren. Unter dieser Voraussetzung ergibt sich aus (30.14) die Abschätzung

$$|g_j| \leq M \cdot \sum_{k=0}^m |e_{j+k}| + d(h) \tag{30.15}$$

mit $M := L \cdot \max_{0\le k\le m} |b_k|$ und $\tau_h(x)| \le d(h)$. Man setzt nun

$$E_j := \begin{pmatrix} e_j \\ \vdots \\ e_{j+m-1} \end{pmatrix} \in \mathbb{R}^m, \quad B := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^m, \quad A := \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & \cdots & -a_{m-1} \end{pmatrix} \in \mathbb{R}^{m\times m}.$$

Die Rekursion (30.14) lautet dann in vektorieller Schreibweise

$$E_{j+1} = AE_j + hg_jB \tag{30.16}$$

und es gilt

$$\det(\lambda I_n - A) = \sum_{k=0}^{m} a_k\lambda^k = \rho(\lambda),$$

da wir bei linearen MSV von $a_m = 1$ ausgegangen sind. Da das MSV zusätzlich nach Voraussetzung stabil ist, gilt

(i) $|\lambda| \le 1$ für alle Nullstellen,

(ii) $|\lambda| = 1 \Rightarrow \lambda$ ist eine einfache Nullstelle.

Nach §12 gibt es eine Vektornorm $||\cdot||$ im $\mathbb{R}^n$, so dass für die zugeordnete Matrixnorm

$$||A|| = \rho(A) = \max\{|\lambda| \mid \lambda \text{ ist Eigenwert von } A\} \le 1$$

gilt. Aus (30.16) folgt dann

$$||E_{j+1}|| \le ||A|| \cdot ||E_j|| + |h| \cdot |g_j| \cdot ||B|| \le ||E_j|| + |h| \cdot |g_j| \cdot ||B||\,. \tag{30.17}$$

Wegen der Äquivalenz aller Normen im $\mathbb{R}^n$ gibt es weiterhin eine Konstante $\tilde{c} > 0$ mit

$$\frac{1}{\tilde{c}}\,||y|| \le ||y||_1 = \sum_{k=0}^{m} |y_k| \le \tilde{c}\,||y||$$

für alle $y \in \mathbb{R}^n$. Es folgt

$$\frac{1}{\tilde{c}}\,||E_j|| \le ||E_j||_1 = \sum_{k=0}^{m-1} |e_{j+k}| \le \tilde{c}\,||E_j||\,.$$

Mit

$$\sum_{k=0}^{m} |e_{j+k}| \le ||E_j||_1 + ||E_{j+1}||_1 \le \tilde{c}\,(||E_j|| + ||E_{j+1}||)$$

wird aus (30.15) die Abschätzung

$$|g_j| \le M \sum_{k=0}^{m} |e_{j+k}| + d(h) \le M\tilde{c}\,(||E_j|| + ||E_{j+1}||) + d(h).$$

Da wir $||B||_1 = 1$ und damit $||B|| \le \tilde{c}\,||B||_1 = \tilde{c}$ haben, ergibt sich aus (30.17)

$$||E_{j+1}|| \le ||E_j|| + |h|M\tilde{c}^2\,(||E_j|| + ||E_{j+1}||) + \tilde{c}d(h)|h|,$$

woraus wir

$$(1 - |h|M\tilde{c}^2) \cdot ||E_{j+1}|| \le (1 + |h|M\tilde{c}^2)\,||E_j|| + \tilde{c}d(h)|h|$$

folgern. Man überlege sich nun zunächst, dass für alle $|t| \le \frac{1}{2}$ die Abschätzung $\frac{1+t}{1-t} \le 1 + 4t$ gilt. Setzt man nun $t := |h|M\tilde{c}^2$, was für genügend kleines h kleiner als $\frac{1}{2}$ wird, so ergibt die letzte Ungleichung

$$||E_{j+1}|| \le (1 + 4|h|M\tilde{c}^2)\,||E_j|| + 2|h|\tilde{c}d(h).$$

Wendet man nun den Hilfssatz (28.2) an, so gilt $||E_0|| = ||\varepsilon|| \leq \tilde{c}c(h)$ und

$$||E_n|| \leq \tilde{c}e^{4|x-x_0|M\tilde{c}^2}c(h) + \frac{e^{4|x-x_0|M\tilde{c}^2} - 1}{4|h|M\tilde{c}^2} \cdot 2|h|\tilde{c}d(h),$$

wobei $h = h_n = \frac{x-x_0}{n}$. Mit geeigneten Konstanten $c_1, c_2 > 0$ gilt daher

$$|e_n| = |y_n - y(x)| \leq c_1 e^{c_2|x-x_0|} \cdot (c(h) + d(h)) .$$

Die rechte Seite ist kleiner als ein gegebenes α für alle $|h| \leq h_0$ mit entsprechend geeignetem $h_0 > 0$. Damit ist auch die gemachte Voraussetzung nachträglich verifiziert und der Beweis abgeschlossen. □