Höhere Numerische Mathematik

Vorlesung vom SS 2008

Mario Ohlberger

Institut für Numerische und Angewandte Mathematik
Fachbereich Mathematik und Informatik
Westfälische Wilhelms-Universität Münster

Dieses Skript beruht auf meiner Vorlesung *Höhere Numerische Mathematik* vom Sommersemester 2008 an der Westfälische Wilhelms-Universität Münster.
Es ist eine üeberarbeitete und ergänzte Version meines Skripts zur Vorlesung "Numerik II" gelesen an der Albert-Ludwigs-Universität Freiburg im Sommersemester 2005. Das Skript wurde in seiner ersten Fassung von Stefan Armbruster getippt und kann in dieser Version von seiner Webpage heruntergeladen werden: `http://www.informatik.uni-freiburg.de/~armbruss/skriptum/`.
Ich danke Herrn Daniel Lengeler, der das Skript basierend auf meinen handschrifftlichen Notizen überarbeitet hat.
Es besteht keine Garantie auf Richtigkeit und/oder Vollständigkeit des Manuskripts.

Mario Ohlberger

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 0

Einleitung

Viele Fragestellungen aus den Naturwissenschaften, der Ökonomie und Medizin führen auf mathematische Probleme, die numerisch gelöst werden müssen. In dieser Vorlesung wird die Theorie und Praxis grundlegender numerischer Algorithmen zur Behandlung gewöhnlicher Differentialgleichungen behandelt. Dazu gehören Einschritt- und Mehrschrittverfahren zur Approximation von Anfangswertproblemen, sowie Finite Differenzen und Finite Elemente Verfahren zur Diskretisierung von Randwertproblemen. Weitere Themen sind Gradientenverfahren, Eigenwertprobleme und Grundzüge der Approximation.

Beginnen wir mit ein paar Grundbegriffen und Beispielen zu gewöhnlichen Differentialgleichungen:

Definition 0.1 (Gewöhnliche Differentialgleichung)
Seien $n \in \mathbb{N}$ und $I \subset \mathbb{R}$ ein Intervall und $F : I \times \mathbb{R}^{n+1} \to \mathbb{R}$ gegeben.
Unter einer skalaren gewöhnlichen Differentialgleichung (DGL) n-ter Ordnung für eine Funktion $y \in C^n(I)$ versteht man eine Gleichung der Form

$$F(x, y(x), y'(x), y''(x), \ldots, y^{(n)}(x)) = 0, \quad (x \in I). \tag{1}$$

Falls eine Funktion $f : I \times \mathbb{R}^n \to \mathbb{R}$ existiert, so dass (1) in der Form

$$y^{(n)}(x) = f(x, y(x), y'(x), y''(x), \ldots, y^{(n-1)}(x)), \quad (x \in I) \tag{2}$$

geschrieben werden kann, so heißt (1) explizit (sonst implizit). Ein Funktion $y \in C^n(I)$, die (1) erfüllt, heißt Lösung der gewöhnlichen Differentialgleichung.

Beispiel 0.2 (Freier Fall/Gravitation)
Sei $t \in I := [0, \infty]$ die Zeit, $x(t) \in \mathbb{R}$ der Ort eines Massepunktes zur Zeit t, $v(t) = x'(t)$ die Geschwindigkeit des Massepunktes und $a(t) = v'(t) = x''(t)$ die Beschleunigung des Massepunktes, sowie $K = K(t, x(t), v(t))$ eine äußere Kraft, die auf den Massepunkt wirkt.

Newtonsches Kraftgesetz:

$$\begin{aligned} & mx''(t) = K(t, x(t), x'(t)), \\ \Longrightarrow \quad & x''(t) = \tfrac{1}{m} K(t, x(t), x'(t)). \end{aligned} \tag{3}$$

Für gegebene Kraft K ist (3) eine explizite DGL zweiter Ordnung.

A) Freier Fall ohne Luftwiderstand:

Erbeschleunigung: $g = 9.81\frac{m}{s^2} \Longrightarrow K(t, x(t), x'(t)) = -mg$.
Einsetzen in (3) ergibt:

$$\begin{aligned} x''(t) = &\quad -\frac{1}{m}mg = -g, \\ \Longrightarrow \quad x'(t) = &\quad -gt + c_1, \\ \Longrightarrow \quad x(t) = &\quad -\frac{1}{2}gt^2 + c_1 t + c_2, \quad \text{mit } c_1, c_2 \in \mathbb{R}. \end{aligned}$$

Die Lösung dieser DGL ist also nicht eindeutig bestimmt.

B) Freier Fall mit Luftwiderstand:
Hier ist die Kraft gegeben durch:

$$K(t, x(t), x'(t)) = -\sigma x'(t) - mg, \quad \text{mit } \sigma > 0.$$

Wir erhalten die DGL:

$$x''(t) = -\frac{\sigma}{m}x'(t) - g.$$

Mit $x'(t) = v(t)$ ergibt sich eine DGL erster Ordnung für v:

$$v'(t) = -\frac{\sigma}{m}v(t) - g. \tag{4}$$

Die Funktionen $v(t) = c_1 \exp(-\frac{\sigma}{m}t) - \frac{mg}{\sigma}$ sind für alle $c_1 \in R$ Lösungen von (4), denn

$$\begin{aligned} v'(t) &= -c_1\tfrac{\sigma}{m}\exp(-\tfrac{\sigma}{m}t), \\ \text{und} \quad -\tfrac{\sigma}{m}v(t) - g &= -c_1\tfrac{\sigma}{m}\exp(-\tfrac{\sigma}{m}t) + g - g. \end{aligned}$$

Bemerkung: Man kann zeigen, dass alle Lösungen von (4) von dieser Form sein müssen.

Für die Lösung von (3) gilt nach Integration:

$$x(t) = -c_1\frac{m}{\sigma}\exp(-\frac{\sigma}{m}t) - \frac{mg}{\sigma}t + c_2.$$

Auch hier ist die Lösung nicht eindeutig.

C) Freier Fall aus großer Höhe:
Da sich die Gravitation mit der Höhe ändert, müssen wir in diesem Fall das Gravitationsgesetz ansetzen als:

$$K(t, x(t), x'(t)) = -mg\frac{R^2}{x^2(t)}$$

Dabei bezeichnet R den Erddurchmesser. Wir erhalten die explizite DGL zweiter Ordnung:

$$x''(t) = -g\frac{R^2}{x^2(t)}. \tag{5}$$

Ansat zur Lösung: $x(t) = at^b$. Damit erhlaten wir: $x'(t) = abt^{b-1}$ und $x''(t) = ab(b-1)t^{b-2}$. Einsetzen in (5) ergibt:

$$ab(b-1)t^{b-2} = -gR^2\frac{1}{a^2t^{2b}} = -\frac{gR^2}{a^2}t^{-2b}.$$

Also muß gelten:

1.) $b - 2 = -2b \implies b = \frac{2}{3}$,

2.) $ab(b-1) = -\frac{gR^2}{a^2} \implies a = \left(\frac{9gR^2}{2}\right)^{1/3}$.

Also ist

$$x(t) = \left(\frac{9gR^2}{2}\right)^{1/3} t^{2/3}$$

eine Lösung von (5).

Beispiel 0.3
Für die explizite DGL erster Ordnung

$$y'(x) = x^2 + (y(x))^2$$

kann keine lösung in geschlossener Form angegeben werden. Es existieren jedoch Lösungen!

Bemerkung: Wir halten fest:

1. Die Bestimmung einer Lösung (in Formelgestalt) ist sehr oft nicht möglich.
2. Falls eine lösung existiert, ist sie i.A. nicht eindeutig.
3. Die *freien Konstanten* $(_1, c_2, \ldots)$ können durch zusätzliche Bedingungen festgelegt werden: Anfangswerte oder Randwerte.

In Kapitel 1 werden wir uns mit numerischen Verfahren zur Lösung von gewöhnlichen Differentialgleichungen beschäftigen. Dabei werden wir uns zunächst mit Anfangswertproblemen auseinandersetzen.

Kapitel 1

Numerik Gewöhnlicher Differentialgleichungen

1.1 Exkurs zur Theorie gewöhnlicher Differentialgleichungen

Definition 1.1 (Anfangswertproblem (AWP))
Sei folgende Voraussetzung erfüllt:

(V) *Seien* $I := [a, b],\ b < a\ , G \subset \mathbb{R}^n$ *zusammenhängende und offene Teilmenge (Gebiet),* $S := I \times G$ *und* $f : S \longrightarrow \mathbb{R}^n$ *stetig und* $y_0 \in G$.

Dann heißt $y : I \longrightarrow \mathbb{R}^n$ *Lösung des* **AWP**, *g.d.w.*

(i) $y \in C^1(I, \mathbb{R}^n)$ *und* $y(I) \subseteq G$

(ii) $y'(x) = f(y, y(x))\ \ \forall x \in I$

(iii) $y(a) = y_0$

Satz 1.2

Unter der Voraussetzung (V) sind folgende Aussagen äquivalent:

a) $y : I \longrightarrow \mathbb{R}^n$ löst (AWP).

b) $y \in C^0(I, G)$ und $y(x) = y_0 + \int_a^x f(s, y(s))ds\ \ \forall x \in I$.

Beweis: siehe Ü.A.

Definition 1.3 (Picard-Lindelöf Iteration)
Definiere Operator $T : C^0(I, \mathbb{R}^n) \longrightarrow C^0(I, \mathbb{R}^n)$ *durch*

$$(Ty)(x) := y_0 + \int_a^x f(s, y(s))ds.$$

Dann ist 1.2 b) äquivalent zu $Ty = y$.

Fixpunktiteration: $y^{(0)} \in C^0(I, \mathbb{R}^n)$ *gegeben,* $y^{(n+1)} = Ty^{(n)}$.

Frage: Wann konvergiert diese Iteration?

Satz 1.4 (Picard-Lindelöf, lokale Version)
Es gelten die Voraussetzungen (V). Erfüllt f auf S die Lipschitzbedingung

$$(L) \qquad \|f(x,y) - f(x,z)\|_\infty \le L \|y - z\|_\infty \quad \forall (x,y),(x,z) \in S,$$

so hat das AWP lokal eine eindeutige Lösung $\tilde{y}$, d.h. $\exists \varepsilon > 0,\ \exists \tilde{y} \in C^1(I_\varepsilon, \mathbb{R}^n)$ mit $\tilde{y}$ löst AWP auf $I_\varepsilon := [a, a + \varepsilon]$.

Beweis: Es ist

$$\begin{aligned} \|(Ty)(x) - (Tz)(x)\|_\infty &\overset{(L)}{\le} L \int_a^x \|y(s) - z(s)\|_\infty \, \mathrm{d}s \\ &\le L \|y - z\|_\infty (x - a) \\ &\le L\varepsilon \|y - z\|_\infty, \text{ für } x \in I_\varepsilon. \end{aligned}$$

Wähle $\varepsilon > 0$, so dass $L\varepsilon < 1$. Dann folgt, dass T Kontraktion ist und die Aussage folgt mit dem Banachschen Fixpunktsatz. □

Satz 1.5 (Picard-Lindelöf, globale Version)
Gelte (V) und (L). Erfüllt f auf S die Bedingung

$$(M) \qquad \|f(x,y)\|_\infty \le M \qquad \forall (x,y) \in S$$

und G erfülle zusätzlich $G \supseteq \overline{B_\sigma(y_0)} := \{x \in \mathbb{R}^n \mid \|x - y_0\|_\infty \le \sigma\}$, wobei $\sigma \ge (b-a)M$ sei.
Dann gilt:

a) Das AWP hat auf I genau eine Lösung $\tilde{y}$.

b) $\forall x \in I$ gilt $(x, \tilde{y}(x)) \in K_M = K_M(a, y_0) := \{(x,y) \in \mathbb{R} \times \mathbb{R}^n | \, \|y - y_0\|_\infty \le M|x - a|\} \cap S$ (siehe Abb. 1.1)).

c) Die Fixpunktiteration von Picard-Lindelöf konvergiert gleichmäßig auf I gegen $\tilde{y}$. Für den Fehler gilt:

$$\left\|\tilde{y}(x) - y^{(k)}(x)\right\|_\infty \le \frac{L^k (x-a)^k}{k!} e^{L(x-a)} \qquad \forall x \in I.$$

Beweis: siehe Ü.A.

Satz 1.6 (Satz von Peano)

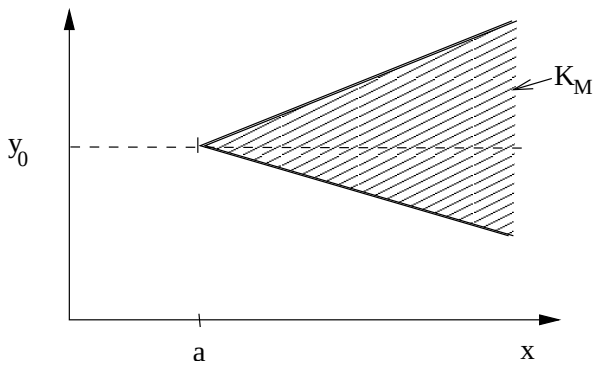

Abbildung 1.1: Picard-Lindelöf: Grafische Darstellung von K_M.

Sei S ein Rechteckgebiet, das $K_M(a, y_0)$ enthält und es gelten (V) und (M). Dann ex. eine Lösung $\tilde{y}$ des AWP auf I.

Beweis: (siehe z.B. Walter [9])

Lemma 1.7 (Lemma von Gronwall)
Sei $p, q \in C^0([a, b])$ mit $p, q \geq 0$. Erfüllt die Funktion $e : [a, b] \longrightarrow \mathbb{R}$ die Integralbedingung

$$0 \leq e(x) \leq p(x) + \int_a^x q(s)e(s)ds \qquad \forall x \in [a, b],$$

so gilt:

$$0 \leq e(x) \leq p(x) + \int_a^x q(s)p(s) \exp\left(\int_s^x q(t)dt\right) ds.$$

Beweis: siehe Ü.A.

Satz 1.8 (Stetigkeitssatz für AWP)
Es gelten die Vor. (v) und (L). Sei $\tilde{y}$ Lösung des AWP und $\tilde{z}$ sei Lösung des gestörten AWP:

$$z'(x) = f(x, z(x)) + \varepsilon(x) \ , \quad z(a) = z_0 \in G.$$

Gelte $z_0 - y_0 \leq \tilde{\varepsilon}$ und sei $\varepsilon(x) \in C^0(I, \mathbb{R}^n)$ mit $\|\varepsilon(x)\|_\infty \leq \varepsilon \qquad \forall x \in I.$
Dann gilt:

$$\|\tilde{z}(x) - \tilde{y}(x)\|_\infty \leq (\tilde{\varepsilon} + \varepsilon(x - a))e^{L(x-a)}.$$

Beweis: Folgt direkt aus 1.7 mit $e(x) := \|\tilde{z}(x) - \tilde{y}(x)\|_\infty$.

$$\begin{aligned} \Longrightarrow \quad e(x) &= \left\| z_0 + \int_a^x f(\tau, \tilde{z}(\tau)) - \varepsilon(\tau)\mathrm{d}\tau - y_0 - \int_a^x f(\tau, \tilde{y}(\tau))\mathrm{d}\tau \right\| \\ &\leq \|z_0 - y_0\| + \varepsilon(x - a) + \int_a^x L \ e(\tau)\mathrm{d}\tau \\ &= p(t) + \int_a^t q(\tau)e(\tau)\mathrm{d}\tau \text{ mit} \\ &\quad p(\tau) = \tilde{\varepsilon} + \varepsilon(x - a), \ q(\tau) = L \end{aligned}$$

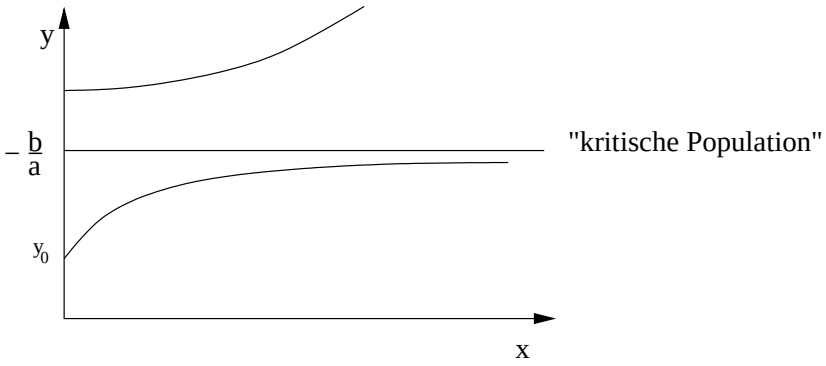

Abbildung 1.2: Lineare (AWP):Anwendung, Bakterienwachstum.

Aus dem Lemma von Gronwall folgt:

$$e(x) \leq \tilde{\varepsilon} + \varepsilon(x-a) + L\int_a^x (\tilde{\varepsilon} + \varepsilon(\tau - a))e^{L(\tau-a)}\mathrm{d}\tau.$$

Durch Berechnung des Integrals folgt die Behauptung. □

Satz 1.9 Methode der Trennung der Variablen
Läßt sich die Differentialgleichnung $y' = f(x,y)$ als $y' = \frac{p(x)}{q(y)}$ schreiben und ist $\frac{dq}{dx} = \frac{dp}{dy}$ erfüllt, so gilt: $y(x)$ ist gegeben durch

$$F(x, y(x)) = konst$$

mit

$$F(x,y) = \int_a^x p(t)dt + \int_{y_0}^y q(s)ds.$$

Beweis: Nachrechnen.

Definition 1.10 (Lineare AWP)
Sei $I = [a,b]$. Gesucht $y \in C^1(I,\mathbb{R})$:

(L-AWP) $\left|\begin{array}{ll} y'(x) = \alpha y(x) + \beta, & \alpha, \beta \in \mathbb{R} \\ y(a) = y_0 \end{array}\right.$

Die Lösung des homogenen AWPs $z' = \alpha z$, $z(a) = z_0$ ist dann gegeben durch $z(x) = z_0 e^{\alpha(x-a)}$.
Durch "Variation der Konstanten" folgt:

$$\underline{\textit{Ansatz:}} \quad y(x) = z(x)v(x) \quad \Longrightarrow \quad y(x) = (\tfrac{\beta}{\alpha} + y_0)e^{\alpha(x-a)} - \tfrac{\beta}{\alpha}.$$

Anwendung: Bakterienwachstum
$y_0 \hat{=}$ # Bakterien am Anfang, $-\beta > 0$ Sterberate, $\alpha > 0$ Wachstumsfaktor. Das qualitative Verhalten der Lösung ist in Abb. 1.2 dargestellt.

Definition 1.11 (Systeme linearer AWP)
Sei $I = [a, b]$ und $A \in \mathbb{R}^{n\times n}$ diagonalisierbar, $y_0, B \in \mathbb{R}^n$. Gesucht $y \in C^1(I, \mathbb{R}^n)$:

$$\textit{(SL-AWP)} \qquad \left| \begin{array}{l} y'(x) = Ay(x) + B \\ y(a) = y_0 \end{array} \right.$$

Wir definieren die "Matrixexponentielle" $exp(A)$ durch

$$exp(A) := \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

Dann löst auch hier $z(x) = z_0 exp(A(x-a))$ das homogene System $z' = Az$, $z(a) = z_0$.
Eine Lösung von SL-AWP erhält man durch Variation der Konstanten mit dem Ansatz:

$$y(x) = exp(A(x-a))v(x).$$

Definition 1.12 (Differentialgleichungen höherer Ordnung)
Geg: $I = [a, b]$; $G \subset \mathbb{R}^{m-1}$ *Gebiet;* $S = I \times G$, $f : S \longrightarrow \mathbb{R}$.
Ges: $y \in C^m(I, \mathbb{R}^n)$; $y^{(k)}(I) \subseteq P_k(G)$; P_k *Projektion*

$$y^{(m)}(x) = f(x, y'(x), \ldots, y^{(m-1)}(x)) \ \forall x \in I,$$
$$y^{(k)}(a) = y_0^{(k)} \subset P_k(G) \ , k = 0, \ldots, m-1.$$

Reduktion auf System 1. Ordnung:
Setze $y_k := y^{(k)}$; $k = 0, \ldots, m-1$
und $F(x, y_0, \ldots, y_{m-1}) = (y_0, \ldots, y_{m-1}, f(x, y_0, \ldots, y_{m-1}))^T$.
Dann folgt mit $y = (y_0, \ldots, y_{m-1})^T$:

$$y' = F(x, y),$$
$$y(a) = (y_0^{(0)}, \ldots, y_0^{(m-1)})^T.$$

Generalvereinbarung:

1) Alle Sätze werden im Folgenden für skalare AWP 1. Ordnung formuliert.

2) Sollte ein Satz für Systeme nicht gelten, so wird dies explizit angegeben.

1.2 Einschrittverfahren

Wir betrachten das skalare AWP:

$$\text{(AWP)} \qquad \left| \begin{array}{l} y'(x) = f(x,y), \\ y(a) = y_0. \end{array} \right.$$

Es seien stets die Bedingungen (V), (M), (L) erfüllt und es sei stets $S \subseteq K_M(a,y_0)$ ein Rechteckgebiet, wobei

$$K_M = K_M(a,y_0) := \{(x,y) \in I \times \mathbb{R} |\ |y-y_0| \leq M(x-a) + \varepsilon_0\}, \ \ \varepsilon_0 > 0.$$

Wir bezeichnen mit $\tilde{y}$ die eindeutig bestimmte Lösung von (AWP).

Definition 1.13 (Diskretes Verfahren zur Lösung von (AWP))
Definiere Gitter $I_h = \{x_0, \ldots, x_n\} \subseteq I, \ \ x_{j+1} = x_j + h_j, \ \ h_j > 0, \ j = 0, \ldots, n-1$. *Setze* $x_0 = a, x_n = b$.
Schrittweitenvektor: $\overline{h} = (h_0, \ldots, h_{n-1})^T$.
I_h *heisst zulässiges Gitter auf* I. *Definiere* $h := max_{j=0,\ldots,n-1} h_j$ *als Feinheit von* I_h.
Ein numerisches Verfahren Φ

- *findet ein Gitter* I_h
- *ordnet* I_h *eine Gitterfunktion* $u_h : I_h \longrightarrow \mathbb{R}$ *zu.*

u_h *ist auf* I_h *diskrete Näherungslösung von* $\tilde{y}$.

Definition 1.14 (Globaler Fehler von Verfahren Φ)
Für $x \in I_h$ *setze* $e_h(x) := \tilde{y}(x) - u_h(x)$ *("Fehlerfunktion"). Definiere*

1) $e_i := e_h(x_i) \ , \ \ x_i \in I_h$,

2) $\|e_h\|_h := max_{x \in I_h} |e_h(x)|$.

$\|e_h\|_h$ *heist Diskretisierungsfehler von* Φ *auf* I_h.

Definition 1.15 (Konvergenz/Konvergenzordnung)

1) Φ *heisst konvergent auf* I, *falls* $\|e_h\|_h \longrightarrow 0$ *für* $h \longrightarrow 0$.

2) Φ *hat auf* I *die Konvergenzordnung* $p > 0$, *falls für genügend glattes* f *gilt:*

$$\|e_h\|_h = O(h^p) \quad für\ h \longrightarrow 0.$$

Definition 1.16 (Explizite Einschrittverfahren (ESV))
Ein Verfahren Φ heisst explizites Einschrittverfahren, falls es eine Verfahrensfunktion $\varphi = \varphi(x,u,h)$, $(x,u) \in K_M(a,y_o)$, $h \in (0, b-a]$, $\varphi(x,u,h) \in [-M,M]$ gibt, so dass u_h auf zulässigem Gitter I_h definiert ist durch:

$$\left|\begin{array}{ll} u_0 & := y_0 + \varepsilon_0, \\ u_{j+1} & := u_j + h_j \varphi(x_j, u_j, h_j), \quad j = 0, \ldots n-1, \\ u_h(x_j) & := u_j, \qquad \forall x_j \in I_h. \end{array}\right.$$

Beispiel 1.17
1) Eulerverfahren (explizit): $\varphi(x,u,h) = f(x,u)$.

2) Verbessertes Eulerverfahren: $\varphi(x,u,h) = f(x + \frac{h}{2}, u + \frac{h}{2} f(x,u))$.

Definition 1.18 (Abschneidefehler/Konsistenz)

1) Für $h \in (0, b-a]$, $x \in [a, b-h]$, $y : [x, x+h] \longrightarrow \mathbb{R}$ mit $y'(\xi) = f(\xi, y(\xi))$ für $\xi \in [x, x+h]$ sowie $(x, y(x)) \in K_M$, heißt

$$\tau_h(x,y) := \frac{y(x+h) - y(x)}{h} - \varphi(x, y(x), h)$$

lokaler Abschneidefehler oder Diskretisierungsfehler von φ.

2) φ heißt konsistent mit (AWP), falls gilt

$$|\tau_h(x, \tilde{y})| \overset{h \to 0}{\longrightarrow} 0 \quad \forall x \in I \setminus \{b\}.$$

φ hat Konsistenzordnung $p \in \mathbb{N}$, falls für genügend glattes f gilt: $|\tau_h(x,\tilde{y})| = O(h^p)$, $h \longrightarrow 0$ $\forall x \in I \setminus \{0\}$

3) Das ESV Φ heißt konsistent mit (AWP), falls $e_0 \overset{h \to 0}{\longrightarrow} 0$ und φ konsistent mit (AWP) ist.
Φ hat Konsistenzordnung $p \in \mathbb{N}$, falls

$$|e_0| = O(h^p) \ , \quad h \longrightarrow 0$$

und φ die Konsistenzordnung p hat.

Lemma 1.19 (Diskretes Lemma von Gronwall)
Seien $(p_n)_{n\in\mathbb{N}}, (q_n)_{n\in\mathbb{N}}, (e_n)_{n\in\mathbb{N}}$ positive Folgen mit $e_{n+1} \leq (1+q_n)e_n + p_n$ für $n < N$. Dann gilt:

$$e_n \leq \Big(e_0 + \sum_{j=1}^{n-1} p_j\Big) \exp\Big(\sum_{j=0}^{n-1} q_j\Big) \qquad \text{für } n < N.$$

Beweis: Durch vollständige Induktion.

Satz 1.20 (Konvergenzsatz für ESV)
Φ sei ESV mit Verfahrensfunktion $\varphi = \varphi(x,u,h)$. φ sein bzgl. u auf K_M global Lipschitz stetig. Dann gilt:
Ist Φ mit (AWP) konsistent (mit Ordnung p), so ist Φ auf I konvergent (mit Ordnung p).

Beweis: Sei $a \le x < x+h \le b$. Es ist

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

Für $\tilde{y}_j = \tilde{y}(x_j)$; $x_j \in I_h$; $\tau_j := \tau_h(x_j,\tilde{y})$ gilt also

$$\tilde{y}_{j+1} - \tilde{y}_j = h_j\varphi(x_j,\tilde{y}_j,h_j) + h_j\tau_j \;\; ; \;\; j = 0,\dots,n-1$$

Ist $\tilde{L}$ die Lipschitzkonstante von φ bzgl. u, so gilt:

$$|e_{j+1}| \le |e_j|(1+h_j\tilde{L}) + |\tau_j|h_j$$

$$\overset{1.19}{\Longrightarrow} \quad \|e_h\|_h \le (|e_0| + (b-a)\max_{j=0,\dots,n-1}|\tau_j|)e^{(b-a)\tilde{L}}$$

Nach Voraussetzung ist Φ konsistent (Ordnung p), also folgt

1) $\max_{j=0,\dots,n-1}|\tau_j| \longrightarrow 0 \quad (= O(h^p))$

2) $|e_0| \longrightarrow 0 \quad (= O(h^p))$

$\Longrightarrow$ Behauptung. □

Beispiel 1.21

1) Das Eulerverfahren ist konvergent mit der Ordnung 1. ($\varphi(x,u,h) = f(x,u)$)

2) Das verbesserte Eulerverfahren ist konvergent mit Ordnung 2.

Definition 1.22 (Implizites ESV)
Ein Verfahren Φ zur Lösung von (AWP) heiß implizites ESV, falls es eine Verfahrensfunktion $\tilde{\varphi} = \tilde{\varphi}(x,u,v,h)$, (x,u), $(x+h,v) \in K_M$, $h \in (0, b-a]$, $\tilde{\varphi}(x,u,v,h) \in [-M,M]$ gibt, so dass für zulässige Gitter I_h mit genügend kleiner Feinheit h u_h definiert ist durch:

$$\left|\begin{array}{ll} u_0 & = y_0 + \varepsilon_0, \\ u_{j+1} & = u_j + h_j\tilde{\varphi}(x_j,u_j,u_{j+1},h_j), \quad j = 0,\dots,n-1, \\ u_h(x_j) & := u_j, \qquad \forall x_j \in I_h. \end{array}\right.$$

Beispiel 1.23

1) Implizites Eulerverfahren:

$$\tilde{\varphi}(x,u,v,h) = f(x+h,v) \quad \Longrightarrow \quad u_{j+1} = u_j + h_j f(x_{j+1},u_{j+1}).$$

Dies entspricht der Rechteckregel:

$$\frac{1}{h}\int_x^{x+h} f(t,y(t))dt \approx f(x+h,y(x+h)).$$

2) Crank-Nicholson-Verfahren oder Trapezverfahren:

$$\tilde{\varphi}(x,u,v,h) = \frac{1}{2}[f(x,u) + f(x+h,v)].$$

Dies entsricht einer Integration mit der Trapezregel.

Satz 1.24
Sei Φ ein implizites ESV mit $\tilde{\varphi} = \tilde{\varphi}(x,u,v,h)$ Lipschitz-stetig bzgl. u und v auf dem Definitionsbereich von $\tilde{\varphi}$. Dann existiert für genügend kleines h eine bzgl. u global Lipschitz-stetige Funktion $v(x,u,h)$, so das gilt:

$$u_{j+1} = u_j + h_j\tilde{\varphi}(x_j,u_j,v(x_j,u_j,h_j),h_j) \;\; j = 0,\ldots,n-1.$$

D.h. Φ ist mit $\varphi(x,u,h) := \tilde{\varphi}(x,u,v(x,u,h),h)$ lokal ein explizites ESV im Sinne von 1.16.

Beweis:

1) Für feste x,u,h sei T_u definiert als
$T_u : \mathbb{R} \longrightarrow \mathbb{R}\,, \;\; v \longmapsto u + h\tilde{\varphi}(x,u,v,h)$
Sei L_1 die Lipschitzkonstante von $\tilde{\varphi}$ bzgl v.

$$|T_u v_1 - T_u v_2| \leq \underbrace{h \cdot L_1}_{<1,\ falls\ h\ klein\ genug} |v_1 - v_2|$$

$\Longrightarrow$ T_u ist kontrahierende Selbstabb. auf $\mathbb{R}$.
Nach Banachschen Fixpunktsatz existiert genau ein $\tilde{v}$, so dass $T_u\tilde{v} = \tilde{v}$
$\Longrightarrow$ $\tilde{v} = \tilde{v}(x,u,h)$ ist wohldefiniert und $(x+h,\tilde{v}) \in K_M$.

2) Seien v_1, v_2 die eindeutigen Fixpunkte von T_{u_1}, T_{u_2}. L_2 sei die Lipschitzkonstante von $\tilde{\varphi}$ bzgl. u. Dann gilt:

$$\begin{aligned} |v_1 - v_2| &= |T_{u_1}v_1 - T_{u_2}v_2| \leq |T_{u_1}v_1 - T_{u_1}v_2| + |T_{u_1}v_2 - T_{u_2}v_2| \\ &\leq hL_1|v_1 - v_2| + |u_1 - u_2| + h \cdot L_2|u_1 - u_2| \\ \Longrightarrow\ |v_1 - v_2| &= \underbrace{\frac{1+hL_2}{1-hL_1}}_{:=\tilde{L}_H} |u_1 - u_2|. \end{aligned}$$

Also folgt für $v_1 = v(x,u_1,h), v_2 = v(x,u_2,h)$:

$$|v(x,u_1,h) - v(x,u_2,h)| \leq \tilde{L}_H|u_1 - u_2|, \quad h \leq H < \frac{1}{L_1}. \qquad \square$$

Definition 1.25 Taylorverfahren
Das Taylorverfahren der Ordnung p ist gegeben durch:

$$\varphi_p(x,u(x),h) := f(x,u(x)) + \frac{h}{2}\frac{d}{dx}f(x,u(x)) + \ldots + \frac{h^{p-1}}{p!}\frac{d^{p-1}}{dx^{p-1}}f(x,u(x)).$$

Satz 1.26

Sei $f \in C^p(S)$ mit $p \in \mathbb{N}$. Dann ist das Taylorverfahren Φ_p mit Verfahrensfunktion φ_p konvergent mit Ordnung p, falls $|e_0| = O(h^p)$, $h \to 0$.

Beweis:

1) Für x, y, h folgt mit der Taylorentwicklung mit Integralrestglied:

$$\begin{aligned}\tau_h(x,y) &= \frac{y(x+h)-y(x)}{h} - \varphi_p(x, y(x), h) = \frac{1}{h}\frac{1}{p!}\int_x^{x+h}(x-\xi)^p y^{(p+1)}(\xi)d\xi \\ &= \frac{h^p}{p!}\int_0^1 (1-t)^p y^{(p+1)}(x+th)dt \\ &= O(h^p)\end{aligned}$$

2) $\frac{d^k}{dx^k}f(x,u,h)$ habe Lipschitz-Konstante L_k , $k = 0, \ldots, p-1$.
Dann gilt:

$$|\varphi_p(x, u_1(x), h) - \varphi_p(x, u_2(x)h)| \leq \underbrace{\sum_{j=0}^{p-1} \frac{(b-a)^j}{j!} L_j}_{=:L} |u_1(x) - u_2(x)|.$$

Aus 1) und 2) folgt die Behauptung mit Satz 1.20. □

Bemerkung:

1) Satz 1.26 zeigt, dass es ESV mit beliebig hoher Konvergenzordnung gibt.

2) Taylorverfahren sind in der Praxis unbedeutend, da

- φ_p nicht unabhängig vom Richtungsfeld f ist,

- höhere Ableitungen von f "teuer" auszurechnen sind.

Definition 1.27 (Explizite Runge-Kutta Verfahren)
Das ESV Φ mit Verfahrensfunktion φ heißt explizites Runge-Kutta Verfahren, falls gilt:

$$\left|\begin{array}{ll} u_0 & := y_0 + \varepsilon_0, \\ u_{j+1} & := u_j + h_j \varphi(x_j, u_j, h_j), \quad j = 0, \dots, n-1, \\ & \\ \varphi(x,u,h) & := \sum_{i=1}^m \gamma_i k_i(x,u,h) \text{ mit } \gamma_i \in \mathbb{R}\,, \ \sum_{i=1}^m \gamma_i = 1, \\ k_1(x,u,h) & := f(x,u), \\ k_2(x,u,h) & := f(x, \alpha_2 h, u + \beta_{2,1} k_1(x,u,h)), \\ \vdots & \\ k_m(x,u,h) & := f\left(x + \alpha_m h, u + h \sum_{l=1}^{m-1} \beta_{m,l} k_l(x,u,h)\right), \\ \text{mit } \alpha_i \in [0,1]. & \end{array}\right.$$

Die Koeffizienten $\alpha_i, \gamma_i, \beta_{m,l}$ sind von x, u, h unabhängig und so gewählt, dass die Konsistenzordnung von φ maximal wird. m heißt Stufenzahl des Runge-Kutta-Verfahrens und k_i heißt i-te Stufenfunktion des Verfahrens Φ.

Notation: *(Butcher-Tableau)*

$$\begin{array}{c|cccc} \alpha_1 & 0 & & & 0 \\ \alpha_2 & \beta_{2,1} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ \alpha_m & \beta_{m,1} & \dots & \beta_{m,m-1} & 0 \\ \hline & \gamma_1 & & \gamma_{m-1} & \gamma_m \end{array} \qquad \textit{bzw.} \qquad \begin{array}{c|c} \alpha & \beta \\ \hline & \gamma^t \end{array}$$

Bemerkung 1.28

1) $\sum_{i=1}^m \gamma_i = 1 \iff$ Konsistenz des entsprechenden R.-K. Verfahrens.

2) Oft wird zusätzlich gefordert, dass $\sum_{i=0}^{l-1} \beta_{l,i} = \alpha_l \ \ \forall l = 2, \dots, m$ gilt, damit $k_l(x_j, u_j, h_j)$ bei genügend glattem f die Ableitung $\tilde{y}'(x_j + \alpha_l h)$ bis auf Größen der Ordnung $O(h^2)$ annähert.

Beispiel 1.29

1) Eulerverfahren: $m = p = 1$ $\quad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

$\implies k_1(x,u,h) = f(x,u)\,; \ \ \varphi(x,u,h) = k_1(x,u,h) = f(x,u)$

2) Verbessertes Eulerverfahren: $m = p = 2$ $\quad \begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$

$k_1(x,u,h) = f(x,u)$
$k_2(x,u,h) = f(x + \frac{1}{2}h, u + \frac{1}{2}hk_1(x,u,h))$
$\varphi(x,u,h) = 1 \cdot k_2(x,u,h) = f(x + \frac{h}{2}, u + \frac{h}{2} f(x,u))$

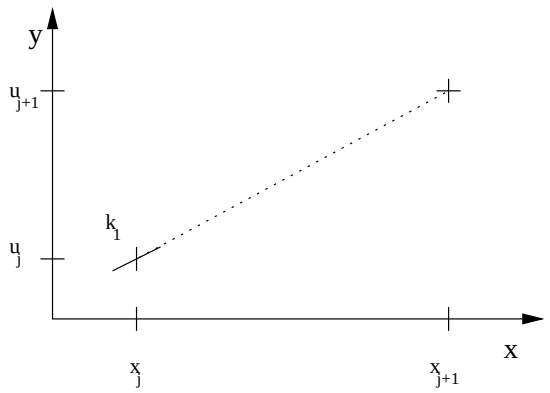

Abbildung 1.3: Runge-Kutta: Eulerverfahren

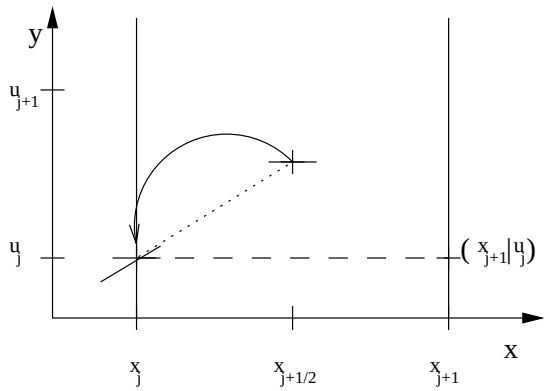

Abbildung 1.4: Runge-Kutta: verbessertes Eulerverfahren

3) Verfahren von Heun: $m = p = 3$

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

$k_1(x,u,h) = f(x,u)$
$k_2(x,u,h) = f(x+\frac{1}{3}h, u+\frac{1}{3}hk_1)$
$k_3(x,u,h) = f(x+\frac{2}{3}h, u+\frac{2}{3}hk_2)$
$\varphi(x,u,h) = \frac{1}{4}k_1 + \frac{3}{4}k_3$

4) Klassisches R.-K. Verfahren: $m = p = 4$

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Bemerkung 1.30

1) Im Spezialfall $f(x,y) = f(x)$ werden expl. R.-K. Verfahren zu zusammengesetzten Quadraturen auf I: z.B.

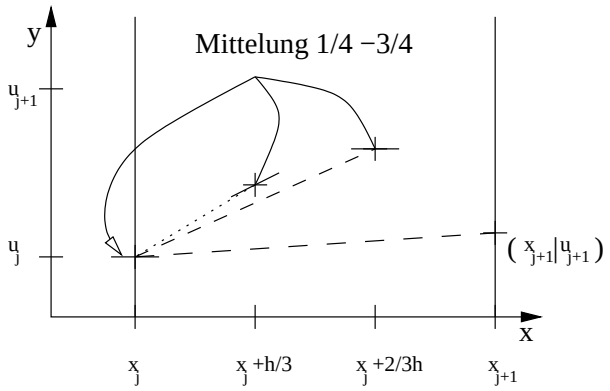

Abbildung 1.5: Runge-Kutta: Verfahren von Heun

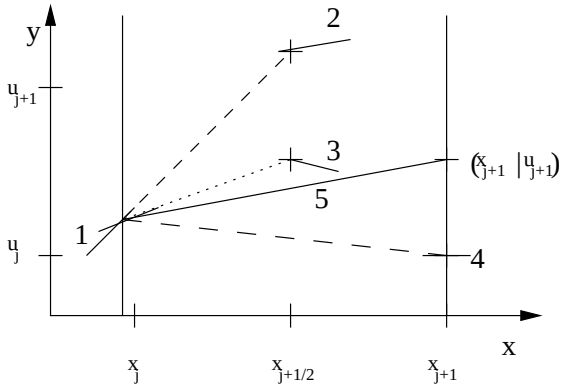

Abbildung 1.6: Runge-Kutta: klassisch

Eulerverfahren $\longleftrightarrow$ Rechteckregel
verb. Eulerverfahren $\longleftrightarrow$ Tangententrapezregel
klass. R-K. Verfahren $\longleftrightarrow$ Simpsonregel

Idee: $\frac{1}{h_j}\int_{x_j}^{x_{j+1}} f(t)\,dt$ wird ersetzt durch $\frac{1}{6}f(x_j) + \frac{2}{3}f(x_j + \frac{h_j}{2}) + \frac{1}{6}f(x_{j+1})$.

2) Zur Konstruktion von R.-K. Verfahren genügt es Spezialfälle $f(x,y) = f(y)$ zu betrachten (autonome Differentialgleichungen). Denn

(i) Koeff. von R.-K. Verfahren bleiben für Systeme von Differentialgleichungen 1. Ordnung unverändert.

(ii) Jede skalare Differentialgleichung 1. Ordnung lässt sich als autonomes System schreiben:

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

$$\Longleftrightarrow \quad \left| \begin{array}{ll} y_1'(x) = f(y_2, y_1) & , y_1(a) = y_0 \\ y_2'(x) = 1 & , y_2(a) = a \end{array} \right.$$

Hinweis: i) gilt nur wenn $\sum_{i=0}^{l-1} \beta_{e,i} = \alpha_l$.

3) Stufenzahl m und Konsistenzordnung $p = p(m)$ sind im Allgemeinen nicht korreliert:

m	1	2	3	4	5	6	7	8	9	10	11
$p(m)$	1	2	3	4	4	5	6	6	7	7	8

Für $m \geq 10$ gilt immer $p(m) \leq m-3$. z.B. $m = 17, p = 10$.

1.31 Konstruktion von R.-K. Verfahren mit zwei Stufen
Es gilt für ESV: $\tau_h(x,y) := \frac{y(x+h)-y(x)}{h} - \varphi(x, y(x), h)$.

Ansatz für $m = 2$ und $f(x,y) = f(y)$:

$$\begin{aligned} \varphi(x, y(x), h) &= \gamma_1 k_1(x, y(x), h) + \gamma_2 k_2(x, y(x), h) \\ &= \gamma_1 f(y(x)) + \gamma_2 f\left(y(x) + h\beta_{2,1} f(y(x))\right) \end{aligned}$$

$$\overset{Taylor}{\Longrightarrow} \quad \varphi(x,y,h) = (\gamma_1+\gamma_2) f(y) + h\beta_{2,1}\gamma_2 (ff')(y) + \tfrac{h^2}{2}\gamma_2\beta_{2,1}^2 (f^2 f'')(y) + O(h^3)$$

wobei $y = y(x)$, $(f \cdot g)(y) = f(y)g(y)$; $f^{(k)} = \frac{d^k}{dy^k} f(y)$.

Andererseits folgt auch mit Taylorentwicklung

$$\frac{y(x+h)-y(x)}{h} = f(y) + \frac{h}{2}(ff')(y) + \frac{h^2}{6}(f''f^2 + f'^2 f)(y) + O(h^3)$$

Also folgt:

$$\begin{aligned} \tau_h(x,y) &= \Big(1 - (\gamma_1+\gamma_2)\Big) f(y) + h\Big(\tfrac{1}{2} - \beta_{2,1}\gamma_2\Big)(f'f)(y) \\ &\quad + \tfrac{h^2}{2}\Big(\tfrac{1}{3}(f''f^2 + f'^2 f) - \gamma_2\beta_{2,1}(f^2 f'')\Big)(y) + O(h^3) \end{aligned}$$

Bedingungen:
Für $p = 1$: $\gamma_1 + \gamma_2 = 1$
Für $p = 2$: $\frac{1}{2} - \beta_{2,1}\gamma_2 = 0$
Für $p = 3$: nicht möglich.

Wähle $\alpha_2 = \beta_{2,1} \in [0,1]$ frei!
Spezialfälle:

1) $\gamma_1 = 0, \gamma_2 = 1 \implies \beta_{2,1} = \frac{1}{2}$,

2) $\gamma_1 = \gamma_2 = \frac{1}{2} \implies \beta_{2,1} = 1$.

Bemerkung 1.32
Die Vorgehensweise für $m > 2$ ist analog, wobei für den Aufwand gilt:

Ordnung p	1	2	3	4	5	6	7	8	9	10
# Bedingungen	1	2	4	8	16	37	85	200	486	1205

Diese Bedingungen sind nichtlineare Gleichungen!

$\rightsquigarrow$ Systematisch Behandlung von Butcher 1964: "grafische Methode".

Satz 1.33

1) Hat ein Runge-Kutta Verfahren die Konsistenzordnung p, so gilt für hinreichend glattes f:

$$\tau_h(x,y) = h^p \overline{\tau}(x,y) + O(h^{p+1}) \ , \ h \longrightarrow 0$$

wobei $\overline{\tau}(x,y) = \sum_k \varepsilon_k D_k f(x,y)$ mit $D_k f \mathrel{\hat{=}}$ Produkt partieller Ableitungen, $\varepsilon_k \mathrel{\hat{=}}$ Fehlerkoeffizient.

2) Explizite R.-K. Verfahren mit Konistenzordnung p sind konvergent mit Ordnung p.

Beweis:

1) Klar nach Konstruktion in 1.31.

2) Es genügt zu zeigen, dass die k_l, $l = 1, \ldots, n$ Lipschitz-stetig sind (dann folgt die Beh. mit Satz 1.20).
Beweis durch Induktion über l:
I.A. $l = 1:\ \ k_1(x,u,h) = f(x,y) \implies L_1 = L$
I.S. $l \to l+1$: Seien $k_1, \ldots, k_l$ Lipschitz-stetig mit Konstanten $L_1, \ldots, L_l$. Dann folgt:

$$\begin{aligned}
&|k_{l+1}(x,u_1,h) - k_{l+1}(x,u_2,h)| \\
&\quad = \quad |f(x+\alpha_{l+1}h, u_1 + h\textstyle\sum_{j=1}^{l} k_j(x,u_1,h)\beta_{l+1,j}) - \\
&\qquad\quad f(x+\alpha_{l+1}h, u_2 + h\textstyle\sum_{j=1}^{l} k_j(x,u_2,h)\beta_{l+1,j})| \\
&\quad \overset{\text{I.V.}}{\leq} \quad L|u_1-u_2| + Lh\textstyle\sum_{j=1}^{l} |\beta_{l+1,j}|L_j|u_1-u_2| \\
&\quad \leq \quad \underbrace{L(1 + (b-a)\sum_{j=1}^{l} |\beta_{l+1,j}|L_j)}_{=:L_{l+1}} |u_1-u_2|.
\end{aligned}$$

Damit folgt die Behauptung. □

Definition 1.34 (Implizite Runge-Kutta Verfahren)
Φ *sei ESV mit Verfahrensfunktion* φ *gegeben durch*

$$\varphi(x,u,h) := \sum_{i=1}^{m} \gamma_i k_i(x,u,h) \ \textit{mit} \ \sum_{i=1}^{m} \gamma_i = 1.$$

Φ *heißt implizites R.-K. Verfahren mit* m *Stufen, falls*

$$k_j(x,u,h) = f\left(x + \alpha_j h, u + h\sum_{l=1}^{m} \beta_{j,l} k_l(x,u,h)\right) \ , \quad j = 1, \ldots, m$$

wobei $\gamma_j, \alpha_j, \beta_{j,l}$ *so gewählt sind, dass* φ *maximale Konsistenzordnung hat.*

Satz 1.35

a) Das implizite R.-K. Verfahren mit m Stufen ist wohldefiniert, falls für die Schrittweite h gilt:

$$hL \max_{k=1,\dots,m} \sum_{j=1}^{m} |\beta_{k,j}| < 1,$$

wobei L die Lipschitz Konstante von f bzgl. y ist.

b) Sei $f(x,y) = f(x)$. Dann stimmt das implizite R.-K. Verfahren mit m Stufen überein mit zusammengesetzten Gauß'schen Quadraturformeln auf I mit jeweils m Knoten in den Teilintervallen $[x_j, x_{j+1}]$, $j = 0, \dots, n-1$.

c) m-stufige implizite R.-K. Verfahren haben Konsistenz- und Konvergenzordnung $2m$, $m \geq 1$.

Beweis:

a) folgt aus der Kontraktionseigenschaft im Banach'schem Fixpunktsatz.

b) & c) siehe Deuflhard/Bornemann §6.2 und §6.3.

Bemerkung 1.36 (Vorteil impliziter R.-K. Verfahren)
Qualitative Eigenschaften der Lösung von (AWP), wie etwa das Monotonieverhalten, werden bei gröberen Gittern bereits wiedergegeben (Nachteil: Es sind nicht-lineare Gleichungssysteme zu lösen).

Beispiel 1.37

1) $m = 1$, $p = 2$:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

2) $m = 2$, $p = 4$:

$$\begin{array}{c|cc} \frac{1}{2} - \frac{1}{\sqrt{12}} & \frac{1}{4} & \frac{1}{4} - \frac{1}{\sqrt{12}} \\ \frac{1}{2} + \frac{1}{\sqrt{12}} & \frac{1}{4} + \frac{1}{\sqrt{12}} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Nächstes Ziel: Schrittweitensteuerung bei ESV.
Dazu ist es notwendig eine genauere Abschätzung des globalen Fehlers zu erhalten, als diejenige, die durch das Lemma von Gronwall entsteht (vgl. Beweis von Satz 1.20).
Wir benötigen eine asymptotische Entwicklung von e_h, um

1.) den tatsächlichen Fehler zu schätzen,

2.) Extrapolation anwenden zu können,

3.) Schrittweiten steuern zu können.

Satz 1.38 (Hauptterm der asymptotischen Fehlerentwicklung bei ESV)
Φ Sei ein ESV auf äquidistantem Gitter I_h. Φ habe Konsistenzordnung p mit

$$e_0 = \rho_0 h^p + \mathcal{O}(h^{p+1}) \quad (h \longrightarrow 0),$$

$$\tau_h(x,y) = \overline{\tau}(x)h^p + \mathcal{O}(h^{p+1}).$$

φ und f seien auf ihrem Definitionsbereich in C^2, $\overline{y}$ sei die eindeutig bestimmte Lösung der linearen Störgleichung

$$(\text{SG}) \left| \begin{array}{ll} \overline{y}'(x) & = \rho(x)\overline{y}(x) - \overline{\tau}(x), \\ \overline{y}(a) & = \rho_0, \end{array} \right|$$

mit $\rho(x) := f_y(x, \tilde{y}(x)) \ \ \forall x \in I$

Dann gilt für $x_j \in I_h$:

$$u_j = \tilde{y}(x_j) + \overline{y}(x_j)h^p + \mathcal{O}(h^{p+1}) \quad (h \longrightarrow 0).$$

Beweis: Für $e_j := u_j - \tilde{y}(x_j)$, $\overline{\tau}_j = \overline{\tau}(x_j)$ gilt (vgl 1.20)

$$e_{j+1} = e_j + h(\varphi(x_j, u_j, h) - \varphi(x_j, \tilde{y}_j, h)) - h^{p+1}\overline{\tau}_j + O(h^{p+2}).$$

Mit Taylorentwicklung folgt:

$$\varphi(x_j, u_j, h) = \varphi(x_j, \tilde{y}_j + e_j, h) = \varphi(x_j, \tilde{y}_j, h) + e_j\varphi_y(x_j, \tilde{y}_j, h) + \frac{1}{2}e_j^2\varphi_{yy}(x_j, \eta_j, h)$$

für ein $\eta_j \in I(\tilde{y}_j, \tilde{y}_j + e_j)$.
Wegen der Konsistenz gilt weiter: $\lim\limits_{h \longrightarrow 0} \varphi(x,y,h) = f(x,y)$

$$\begin{array}{lrl} \Longrightarrow & \varphi_y(x,y,0) & = \quad f_y(x,y) \\ \Longrightarrow & \varphi_y(x_j,\tilde{y}_j,h) & \overset{Taylor}{=} \varphi_y(x_j,\tilde{y}_j,0) + O(h) \\ & & = \quad f_y(x_j,\tilde{y}_j) + O(h) \\ & & \overset{Vor.}{=} \rho(x_j) + O(h). \end{array}$$

Einsetzen in die Gleichung für e_{j+1} ergibt:

$$e_{j+1} = e_j(1 + h\rho(x_j)) - h^{p+1}\overline{\tau}_j + O(h^2e_j) + O(he_j^2) + O(h^{p+2}.)$$

Setze $\overline{e}_j = h^{-p}e_j$, so folgt

$$\underbrace{\bar{e}_{j+1} = \bar{e}_j(1 + h\rho(x_j)) - h\bar{\tau}_j}_{\triangleq \, Euler-Verfahren \, angewendet \, auf \, (SG)} \quad + O(h^2\bar{e}_j) + O(h^{p+1}\bar{e}_j^2) + O(h^2)$$

Konsistenz des Euler-Verfahrens

$$\Longrightarrow \quad \bar{e}_j = \bar{y}(x_j) + O(h).$$

Also folgt

$$e_j = u_j - \tilde{y}(x_j) = h^p\bar{e}_j = h^p\bar{y}(x_j) + O(h^{p+1}). \qquad \square$$

Beispiel 1.39
Sei Φ das Eulerverfahren und $f(x,y) = \lambda y$, $\lambda \in \mathbb{R}$ y_0 gegeben. Sei $I = [0,b]$, $b > 0$ und $u_0 = y_0$.

$$\Longrightarrow \quad \tilde{y}(x) = y_0e^{\lambda x}. \quad (*)$$

Für den Abschneidefehler gilt:

$$\begin{array}{rcl} \tau_h(x,\tilde{y}) & = & \frac{h}{2}\tilde{y}''(x) + O(h^2) \\ & (*) & \\ & = & h\underbrace{\frac{1}{2}\lambda^2 y_0 e^{\lambda x}}_{=:\bar{\tau}(x)} + O(h^2) \end{array}$$

sowie $\rho(x) = f_y(x,\tilde{y}) = \lambda$ und $\rho_0 = 0$.
Also lautet (SG):

$$\left| \begin{array}{l} \bar{y}'(x) = \lambda\bar{y}(x) - y_0\frac{\lambda^2}{2}e^{\lambda x} \\ y(0) = 0 \end{array} \right.$$

Lösung von (SG) ist $\bar{y}(x) = -y_0\frac{\lambda^2}{2}xe^{\lambda x}$

Mit Satz 1.38 folgt:

$$\begin{array}{rcl} u_j & = & \tilde{y}_j + \bar{y}(x_j)h + O(h^2) \\ & = & \tilde{y}_j - y_0\frac{\lambda^2}{2}x_j e^{\lambda x_j}h + O(h^2) \end{array}$$

Direkte Verifikation: Nach dem Euler-Verfahren ist

$$u_{j+1} = (1+\lambda h)u_j.$$

Induktion $\implies$ $u_{j+1} = (1+\lambda h)^j y_0$ für $j = 1,\ldots,n-1$.
Weiter ist

$$\begin{array}{rcl} (1+h\lambda) & \overset{\text{Taylor für } e^{\lambda h}}{=} & e^{\lambda h} - \frac{h^2}{2}\lambda^2 + O(h^3) \\ \implies (1+h\lambda)^j & \overset{\text{Binomi}}{=} & e^{\lambda x_j} - x_j\frac{\lambda^2}{2}e^{\lambda x_j}h + O(h^2) \\ \implies \underbrace{(1+h\lambda)^j y_0}_{=u_j} & = & \underbrace{e^{\lambda x_j}y_0}_{=\tilde{y}(x_j)}\underbrace{-y_0 x_j\frac{\lambda^2}{2}e^{\lambda x_j}}_{=\bar{y}(x_j)}\cdot h + O(h^2). \end{array}$$

Folgerung 1.40
Gilt spezieller als in Satz 1.38

$$e_0 = \rho_0 h^p + \rho_1 h^{p+1} + \ldots + \rho_k h^{p+k} + O(h^{p+k+1})$$

und

$$\tau_n(x,y) = \bar{\tau}_o(x)h^p + \bar{\tau}_1(x)h^{p+1} + \ldots + \bar{\tau}_k h^{p+k} + O(h^{p+k+1}),$$

so gibt es Funktionen $\bar{y}_0,\ldots,\bar{y}_k$, so dass

$$u_j = \tilde{y}(x_j) + \bar{y}_0(x_j)h^p + \ldots + \bar{y}_k(x_j)h^{p+k} + O(h^{p+k+1}) \quad .$$

Die Funktionen $\bar{y}_0,\ldots,\bar{y}_k$ genügen den Störgleichungen

$$(SG_i)\left| \begin{array}{ll} \bar{y}_i(x) = \rho(x)\bar{y}_i(x) - \bar{\tau}_i(x), & \forall i = 0,\ldots,k. \\ \bar{y}_i(a) = \rho_i, & \end{array} \right.$$

Beweis: Die Behauptung folgt induktiv unter Verwendung von 1.38.

1.41 Extrapolation
Seien $I_h, I_{h'}$ zwei äquidistante Gitter auf I mit $h' = qh$, $0 < q < 1$. Sei weiter $x \in I_h \cap I_{h'}$ und es gelten die Voraussetzungen von Satz 1.38.
Seien $u_h, u_{h'}$ die Nährungslösungen des gleichen Verfahrens Φ auf I_h bzw. $I_{h'}$. Dann gelten:

1) $u_h(x) = \tilde{y}(x) + \bar{y}(x)h^p + O(h^{p+1})$,

2) $u_{h'}(x) = \tilde{y}(x) + \bar{y}(x)q^p h^p + O(h^{p+1})$.

Setzt man nun $u_h^{(1)} := \alpha u_h(x) + \beta u_{h'}(x)$ mit $\alpha = -\frac{q^p}{1-q^p}$ und $\beta = 1 - \alpha$, so folgt

$$u_h^{(1)}(x) = \tilde{y}(x) + O(h^{p+1}).$$

(Dies ist die Idee der Richardson-Extrapolation!)

1.42 Schrittweitensteuerung beim ESV
Φ sei ESV mit Verfahrensfunktion φ, I_h Gitter auf I und $u_h : I_h \longrightarrow \mathbb{R}$ Näherungslösung.
Nach Beweis von Satz 1.20 gilt:

$$\|e_h\|_h \leq \frac{1}{L}\tau e^{L(b-a)} + |e_0|e^{L(b-a)},$$

wobei τ obere Schranke für den Abschneidefehler ist.
Sei $|e_0| \leq \varepsilon$. Ist dann $\tau \leq \frac{1}{2}\frac{L\tilde{\varepsilon}}{e^{L(b-a)}} =: \eta$ und $e_0 \leq \frac{1}{2}\frac{\tilde{\varepsilon}}{e^{L(b-a)}}$, so gilt:

$$\|e_h\|_h \leq \tilde{\varepsilon}.$$

Idee: Wähle in jedem Schritt die Gitterweite $h_j > 0$ so, dass h_j maximal ist und $\tau \leq \eta$ erfüllt ist!

Darstellung des Verfahrensfehlers:
$\hat{y}$ genüge dem lokalen AWP

$$\left| \begin{array}{l} y' = f(x,y), \\ y(x^*) = z^*, \end{array} \right.$$

wobei f die Voraussetzungen aus Satz 1.38 erfülle.
Dann gilt für $x \in \{x^*, x^* + h\}$

$$u_h = \hat{y}(x) + \bar{y}_0(x)h^p + \bar{y}_1(x)h^{p+1} + O(h^{p+2}).$$

Es folgt:

$$\begin{array}{rcl} \tau_h(x^*, y) & = & -\bar{y}_0(x^* + h)h^{p-1} - \bar{y}_1(x^* + h)h^p + O(h^{p+1}) \\ & \overset{\text{Taylor}}{=} & -h^p\ \bar{y}_0'(x^*) + O(h^{p+1}). \end{array}$$

Ansatz: Berechne h näherungsweise durch

$$h^p\ |\bar{y}_o'(x^*)| = \eta \qquad (A)$$

1.Variante: Schätzung von h mittels Extrapolation.

1.) Bestimme Näherungslösung v_1 in $x^* + \tilde{h}$ ausgehend von (x^*, z^*) bei Schrittweite $\tilde{h}$.

2.) Bestimme Näherungslösung v_2 in $x^* + \tilde{h}$ ausgehend von (x^*, z^*) bei Schrittweite $\frac{\tilde{h}}{2}$.

Mit 1.38 folgt

(1) $\hat{y}(x^* + \tilde{h}) - v_1 = -\bar{y}_o'(x^*)\tilde{h}^{p+1} + O(\tilde{h}^{p+2})$,

(2) $\hat{y}(x^* + \tilde{h}) - v_2 = -\bar{y}_o'(x^*)\left(\frac{\tilde{h}}{2}\right)^{p+1} + O(\tilde{h}^{p+2})$.

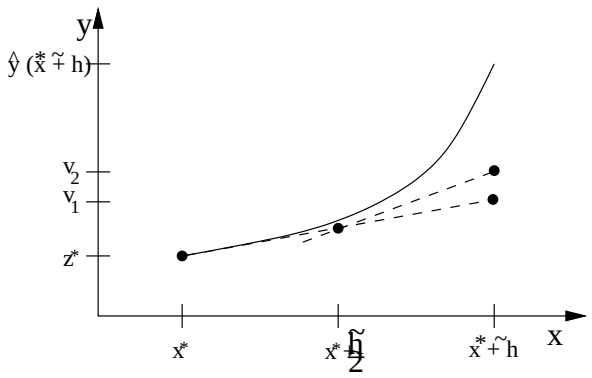

Abbildung 1.7: Schätzung mittels Extrapolation

$$\overset{(2)-(1)}{\Longrightarrow} \quad \tilde{h}^{p+1}\bar{y}_0'(x^*) = \tfrac{v_1-v_2}{1-2^{-(p+1)}} + O(\tilde{h}^{p+2}),$$
$$\Longrightarrow \quad \bar{y}_0'(x^*) = \tilde{h}^{-(p+1)}\tfrac{v_1-v_2}{1-2^{-(p+1)}} + O(\tilde{h}).$$

Einsetzen in (A) ergibt für h die Schätzung

$$h = \tilde{h}\sqrt[p]{\frac{2^{p+1}-1}{2^{p+1}}\frac{\eta\tilde{h}}{|v_1-v_2|}}.$$

Empfehlung:

Ist $h \in \left[\frac{\tilde{h}}{2}, 2\tilde{h}\right]$, so arbeite weiter mit $h_j := h$,

andernfalls führe eine neue Schätzung mit $\tilde{h} := h$ durch.

Grund: Die Schrittweite soll nicht zu stark schwanken.

Sinnvoll: Abbruchbedingung: Wenn $h < 10^{-d}$, $d > 0$, dann Abbruch.

2. Variante: Schätzung von h mit Verfahren höherer Konsistenzordnung.

Seien $\Phi, \tilde{\Phi}$ Verfahren mit Konsistenzordnung p, q, $q > p$ und seien für $\Phi, \tilde{\Phi}$ die Voraussetzungen von Satz 1.38 erfüllt.

1.) Berechne v_1 in $x^* + \tilde{h}$ ausgehend von (x^*, z^*) mit Φ und Schrittweite $\tilde{h}$.
2.) Berechne v_2 in $x^* + \tilde{h}$ ausgehend von (x^*, z^*) mit $\tilde{\Phi}$ und Schrittweite $\tilde{h}$.

Mit Satz 1.38 folgt:

$$(1) \quad \hat{y}(x^* + \tilde{h}) - v_1 = -\bar{y}_0'(x^*)\tilde{h}^{p+1} + O(h^{p+2}),$$
$$(2) \quad \hat{y}(x^* + \tilde{h}) - v_2 = -\bar{y}_0'(x^*)\tilde{h}^{q+1} + O(h^{q+2}).$$

$$\overset{(2)-(1)}{\Longrightarrow} \quad \bar{y}_0'(x^*) = \tilde{h}^{-(p+1)}\tfrac{v_1-v_2}{1-\tilde{h}^{q-p}} + O(\tilde{h}.$$

Einsetzen in (A) ergibt:

$$h = \tilde{h}\sqrt[p]{\frac{\tilde{h}\eta(1-\tilde{h}^{q-p})}{|v_1-v_2|}}.$$

Aufwandsvergleich:

Variante 1: 100% Mehraufwand (Zwischenpunkt)

Variante 2: Bei verschiedenen Verfahren i.A. Summe aus Aufwand von Φ und $\tilde{\Phi}$ (für die Praxis besser!).

Realisierung von Variante 2 mit eingebetteten R.-K. Verfahren

Idee: m-Stufen

$$\begin{array}{c|cccc} 0 & & & & \\ \alpha_2 & \beta_{2,1} & & & \\ \vdots & \vdots & \ddots & & \\ \alpha_m & \beta_{m,1} & \cdots & \beta_{m,m-1} & \\ \hline & \gamma_1 & \cdots & \gamma_{m-1} & \gamma_m \\ & \tilde{\gamma}_1 & \cdots & \tilde{\gamma}_{m-1} & \tilde{\gamma}_m \end{array}$$

wobei Φ definiert durch $\gamma_i, i = 1, \ldots, m$ die optimale Konsistenzordnung
und $\tilde{\Phi}$ definiert durch $\tilde{\gamma}_i, i = 1, \ldots, m$ eine um 1 kleinere Konsistenzordnung liefert.

Vorteil: Die Stufenfunktionen k_i müssen für Φ und $\tilde{\Phi}$ nur einmal berechnet werden.

Beispiel: Das Verfahren von Dormand-Prince ($m = 7$)

$$\begin{array}{c|ccccccc} 0 & & & & & & & \\ \frac{1}{5} & \frac{1}{5} & & & & & & \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5130}{18656} & & \\ 1 & \frac{35}{348} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\ \hline \gamma & \frac{35}{348} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \tilde{\gamma} & \frac{5179}{57900} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \end{array}$$

$\Longrightarrow \quad p(\gamma) = 5 \ , \ \ p(\tilde{\gamma}) = 4.$

1.3 Mehrschrittverfahren

Um Mehrschrittverfahren zu untersuchen benötigen wir einige Sätze aus der Theorie linearer Differenzengleichungen, die wir im folgenden Abschnitt formulieren.

1.3.1 Theorie der linearen Differenzengleichungen

Definition 1.43 (Differenzengleichung)
Sei $U := \{u : \mathbb{N}_0 \longrightarrow \mathbb{C}\}$ *die Menge der komplexen Zahlenfolgen mit* $u_i = u(i)$. *Seien* $k \in \mathbb{N}_0$, $f \in U$, $\alpha_i \in \mathbb{C}$, $i = 0, \ldots, k-1$ *gegeben. Dann heißt:*

$$(DG_k) \qquad u_{n+k} + \alpha_{k-1} u_{n+k-1} + \ldots + \alpha_0 u_n = f_n \;\; ; \;\; n \in \mathbb{N}_0$$

Differenzengleichung der Ordnung k.
(DG_k) *heißt homogen, falls* $f = 0$, *sonst inhomogen.*

Lemma 1.44
Sei eine homogene Differenzengleichung mit Ordnung k gegeben.

1) Seien $u_0, \ldots, u_{k-1} \in \mathbb{C}$ gegebene Anfangswerte einer Folge $u \in U$. Dann gilt: Es gibt genau eine Folge $u \in U$ mit diesen Anfangsdaten, welche der Differenzengleichung genügt.

2) Die Menge aller Lösungen der Differenzengleichungen bildet einen k-dimensionalen Unterraum von U. Basislösungen sind definiert durch:

$$\left|\begin{array}{l} u^{(j)} = (u_n^{(j)})_{n \in \mathbb{N}_0}, \\ u_n^{(j)} = \delta_{nj} \quad 0 \le n, j \le k-1, \\ u^{(j)} \text{ erfüllt die Differenzengleichung für } 0 \le j \le k-1, \end{array}\right.$$

Beweis: 1) Durch vollständige Induktion, 2) nachrechnen.

Definition und Satz 1.45 (Charakterisisches Polynom)
Für (DG_k) homogen heißt

$$\rho(t) = t^k + \alpha_{k-1} t^{k-1} + \ldots + \alpha_0$$

charakterisisches Polynom. Seien $\lambda_1, \ldots, \lambda_m$ Nullstellen des Poloynoms ρ mit Vielfachheit $m_1, \ldots, m_m$ mit $\sum_{i=1}^m m_i = k$. Dann sind äquivalent:

1) u ist eine Lösung von (DG_k),

2) $u = (u_n)_{n \in \mathbb{N}_0}$, $u_n = \sum_{i=1}^m p_i(n) \lambda_i^n$, $n \in \mathbb{N}_0$, wobei $p_i \in \mathbb{P}_{m_i - 1}$ sind für $i = 1, \ldots, m$.

Beweis:

1) $\Rightarrow$ 2) $u_0, \dots, u_{k-1}$ seien Anfangswerte einer gegebenen Lösung von (DG_k). Setze

$$\text{A:}= \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & 0 & 1 \\ -\alpha_0 & \dots & \dots & -\alpha_{k-1} \end{pmatrix} \in \mathbb{C}^{k\times k} \text{ "Begleitmatrix von } \rho\text{"}$$

und $\vec{u}_n := \begin{pmatrix} u_n \\ \vdots \\ u_{n+k-1} \end{pmatrix} \in \mathbb{C}^k$ "Folgenabschnitt n".

Dann gilt:

$$\vec{u}_n = A\vec{u}_{n-1} \overset{\text{Induktion}}{\Longrightarrow} \vec{u}_n = A^n\vec{u}_0.$$

Nach linearer Algebra ex. eine invertierbare Matrix S, so dass $A = SJS^{-1}$, wobei J die Jordansche Normalform von A ist.

Also $J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_n \end{pmatrix}$ mit $J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{m_i\times m_i}$.

Dann folgt induktiv: $A^n = SJ^nS^{-1}$, wobei

$$J_i^n = \begin{pmatrix} \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \binom{n}{2}\lambda_i^{n-2} & \dots & \binom{n}{m_i-1}\lambda_i^{n-m_i+1} \\ 0 & \ddots & \binom{n}{1}\lambda_i^{n-1} & \dots & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \lambda_i^n \end{pmatrix}$$

mit $\binom{n}{j} = 0$ für $j > n$.
$\binom{n}{j}$ ist für festes j und $n \in \mathbb{N}_0$ ein Polynom vom Grad j in n.
$\Longrightarrow$ J_i^n enthält Koeffizienten nur vom Typ $p_i(n)\lambda_i^n$ mit Grad von $p_i \le m_i - 1$.
Also hat u_n die angegebene Form.

2) $\Rightarrow$ 1) Ü.A.

Beispiel 1.46

$u_{n+3} - 4u_{n+2} + 5u_{n+1} - 2u_n = 0$ mit Startwerten $u_0 = 1,\ u_1 = 2,\ u_2 = -1$.
Das charakteristische Polynom ist:

$$\begin{aligned} \rho(t) &= t^3 - 4t^2 + 5t - 2 \\ &= (t-1)^2(t-2), \end{aligned}$$

$\Longrightarrow$ $\lambda_1 = 1$ mit $m_1 = 2$, $\lambda_2 = 2$ mit $m_2 = 1$.

Mit Satz 1.45 folgt:

$$u_n = p_1(n)1^n + p_2(n)2^n.$$

Mit $p_1(n) = \alpha + \beta n$ und $p_2(n) = \gamma$ folgen die Bedingungen:

$$\left.\begin{array}{rl} 1 & = u_0 = \alpha + \gamma \\ 2 & = u_1 = \alpha + \beta + 2\gamma \\ -1 & = u_2 = \alpha + 2\beta + 4\gamma \end{array}\right\} \implies \begin{array}{l} \alpha = 5 \\ \beta = 5 \\ \gamma = -4 \end{array}$$

Also ist die eindeutige Lösung der Differenzengleichunggegeben durch:

$$u_n = 5(1+n) - 2^{n+2}.$$

Bemerkung 1.47
Für die Matrix A gilt stets $Rg(A - \lambda I) \geq k - 1$ für $\lambda \in \mathbb{C}$.
Ist λ Eigenwert von A, so gilt $\dim \mathrm{Eig}(\lambda) = \dim \ker(A - \lambda I) \leq 1$.
$\implies$ Für jeden Eigenwert λ gibt es höchstens einen Jordanblock.

Satz 1.48 (Wurzelbedingung von Dahlquist)
Sei (DG_k) eine homogene, lineare Differenzengleichung mit charakteristischem Polynom ρ. Dann sind äquivalent:

1) Jede Lösung von (DG_k) ist beschränkt.

2) Die Nullstellen λ_i von ρ erfüllen

   (i) $|\lambda_i| \leq 1$,
   (ii) $|\lambda_i| = 1 \implies \lambda_i$ ist einfache Nullstelle.

Beweis:

2) $\Rightarrow$ 1) Die Behauptung folgt direkt aus Satz 1.45, da

$$\lim_{n\to\infty} |u_n| = \lim_{n\to\infty} \sum_{i=1}^{m} |p_i(n)||\lambda_i|^n \leq C.$$

1) $\Rightarrow$ 2) Sei $u_n = \lambda_i^n$. Dann folgt aus $\lim_{n\to\infty} |u_n| \leq C$ auch $\lim_{n\to\infty} |\lambda_i|^n \leq C$ und somit $|\lambda_i| \leq 1$.
Sei weiter λ_i NST mit Vielfachheit $m_i \geq 2$. Dann ist $u_n = n\lambda_i^n$ eine Lösung und mit Satz 1.45 folgt:
$\lim_{n\to\infty} |u_n| \leq C \implies \lim_{n\to\infty} n|\lambda_i^n| \leq C \implies |\lambda_i| < 1.$ □

Definition 1.49 (Verschiebeoperatoren)
$E : U \longrightarrow U\ ,\ u \longmapsto Eu$ *mit* $(Eu)_n = u_{n+1}$

$$\implies (u_0, u_1, u_2, \ldots) \longmapsto (u_1, u_2, u_3, \ldots)$$

$E^{-1} : U \longrightarrow U\ ,\ u \longmapsto E^{-1}u$ *mit* $(E^{-1}u)_n = \begin{cases} u_{n-1}\ , & n \geq 1 \\ 0\ , & n = 0 \end{cases}$

$$\implies (u_0, u_1, u_2, \ldots) \longmapsto (0, u_0, u_1, u_2, \ldots)$$

Bemerkung 1.50

1) $EE^{-1}u = u$, aber $E^{-1}Eu = u - u_0 e^{(0)}$ mit $e_n^{(j)} := \delta_{jn}$ für $j, n \in \mathbb{N}_0$.

2) Es ist $E^j := \underbrace{E \cdot \ldots \cdot E}_{j-mal}$
$\implies \rho(E)u = f$ "Differenzengleichung" (DG_k).

Lemma 1.51
Sei (DG_k) gegeben. Definiere $v^{(j)} := E^{-j-1}u^{(k-1)}$ mit $u^{(k-1)}$ Basislösung von $\rho(E)u = 0$. Dann gilt:

$$\rho(E)v^{(j)} = e^{(j)}, \quad \text{für } j \in \mathbb{N}_0.$$

Beweis: Ü.A.

Satz 1.52
Sei $\rho(E)u = f$ inhomogene Differenzengleichung der Ordnung k. Dann gilt:

(i) $\bar{u} := \sum_{n=0}^{\infty} f_n v^{(n)}$ ist eine Lösung.

(ii) Alle Lösungen haben die Gestalt:
$u = \tilde{u} + \bar{u}$ mit $\tilde{u}$ ist Lösung von $\rho(E)u = 0$.
($\implies$ Lösungsmenge ist k-dimensionale Untermannigfaltigkeit von U)

(iii) Für gegebenes AWP $\rho(E)u = f$ mit $u_i = \beta_i$ für $i = 0, \ldots, k-1$ gilt:

$$u = \tilde{u} + \bar{u},$$

wobei $\tilde{u}$ Lösung von $\rho(E)u = 0$ ist mit $\tilde{u}_i = \beta_i$, $i = 0, \ldots, k-1$.

Beweis: Nachrechnen.

Bemerkung 1.53
Es gilt $v_n^{(j)} = 0$ für $n \leq k + j - 1$, also insbesondere $\bar{u}_n := \sum_{j=0}^{n-k} f_j u_{n-j-1}^{(k-1)}$.

Beispiel 1.54

1) Sei $\rho(t) = t^2 - 1$; $f = (f_n)_{n\in\mathbb{N}}$ mit $f_n := n$
$\implies$ NST von $\rho(t)$: $\lambda_{1/2} = \pm 1$

a) Homogene Lösung: $\tilde{u}_n = \alpha 1^n + \beta(-1)^n$ (nach Satz 1.45).

b) Spezielle Lösung des inhomogenen Problems:

1. Ansatz: $\bar{u}_n = An + B$

$$\implies \begin{array}{rcl} n = f_n & = & (\rho(E)\bar{u})_n = \bar{u}_{n+2} - \bar{u}_n \\ & = & A(n+2) + B - An - B = 2A \end{array}$$

Ansatz schlägt fehl, da A unabhängig von n gewählt werden muss!

2. Ansatz: $\bar{u}_n = An^2 + Bn$
$\Longrightarrow \quad n = f_n = \bar{u}_{n+2} - \bar{u}_n = A \cdot (n+2)^2 + B(n+2) - An^2 - Bn$
Sortiere Terme mit Faktor $n^2, n, 1$. Dann folgt durch Koeffizientenvergleich:

$$\left.\begin{array}{ll} \text{Terme mit } n^2: & \diagup \\ \text{Terme mit } n: & 4A + B - B = 1 \\ \text{Terme mit } 1: & 4A + 2B = 0 \end{array}\right\} \Longrightarrow \begin{array}{l} A = \frac{1}{4} \\ B = -\frac{1}{2} \end{array}$$

Eine spezielle Lösung ist also:

$$\bar{u}_n = \frac{1}{4}n^2 - \frac{1}{2}n$$

c) Wir erhalten eine allgemeinere Lösung: $u = \tilde{u} + \bar{u}$

$$u_n = \frac{1}{4}n^2 - \frac{1}{2}n + \alpha + \beta(-1)^n$$

2) Sei $\rho(t) = t^2 - 1$ und $f_n = n + 2^n$

a) Homogene Lösung wie in 1).

b) Inhomogene Lösung:
Ansatz: $\bar{u}_n = An^2 + Bn + C2^n$
Koeffizientenvergleich: $c = \frac{1}{3}\ ,\ A = \frac{1}{4}\ ,\ B = -\frac{1}{2}$

c) Allgemeine Lösung:

$$u_n = \frac{1}{4}n^2 - \frac{1}{2}n + \frac{1}{3}2^n + \alpha + \beta(-1)^n$$

Lösung von Differenzengleichungen im Hauptfall:

Seien $f_n := \sum_{j=1}^{i} q_j(n)\mu_j^n$ mit $\mu_j \in \mathbb{C}\, q_j \in \mathbb{P}$ mit $\text{Grad}(q_j) =: g_j$. μ_j sei m_j-fache NST von ρ mit $0 \le m_j \le k$ ($m_j = 0 \implies \mu_j$ keine NST).

Ansatz: $\bar{u}_n = \sum_{j=1}^{i} n^{m_j}\tilde{q}_j(n)\mu_j^n$ mit $\tilde{q}_j$ allgemeines Polynom vom Grad g_j.

In Beispiel 1): $i = 1\ ;\ \mu_1 = 1\ ;\ q_1(n) = n.$

In Beispiel 2): $i = 2\ ;\ \mu_1 = 1\ ;\ \mu_2 = 2\ ;\ q_1(n) = n\ ;\ q_2(n) = 1.$

1.3.2 Lineare k-Schrittverfahren

Es gelten die Voraussetzungen aus Abschnitt 1.2 für (AWP) und zusätzlich sei I_h äquidistant, d.h. $h_j = h \ \ \forall j$.

Definition 1.55 (k-Schrittverfahren)

a) Ein Verfahren Φ zur Lösung des (AWP) auf I heißt k-Schrittverfahren mit $k \in \mathbb{N}$, falls es eine Verfahrensfunktion $\varphi = \varphi(x, v_0, v_1, \ldots, v_k, h)$ mit $(x + hi, v_i) \in K_M \ \forall i = 0, \ldots, k$, $h \in (0, \frac{b-a}{k}]$, $\varphi(x, v_0, v_1, \ldots, v_k, h) \in [-M, M]$ gibt, so dass für (äquidistante) zulässige Gitter I_h mit genügend kleiner Feinheit die Näherungslösung $u_h(x_j) =: u_j$ durch die Differenzengleichung k-ter Ordnung

$$\sum_{i=0}^{k} a_i u_{j+i} = h\varphi(x_j, u_j, u_{j+1}, \ldots, u_{j+k}, h)$$

wohldefiniert ist für geeignete Startwerte, $u_0, \ldots, u_{k-1}$.

b) Eine k-Schrittverfahren heißt linear, falls φ die Form

$$\varphi(x, u_0, u_1, \ldots, u_k, h) = \sum_{i=0}^{k} b_i f(\underbrace{x + ih}_{=x_i}, u_i)$$

hat.

Definition und Bemerkung 1.56

1) Ist $\rho(t) = \sum_{i=0}^{k} a_i t^i$ das charakteristische Polynom von Φ, so schreiben wir

$$\Phi = (\rho, \varphi).$$

Ist weiter $\sigma(t) := \sum_{i=0}^{k} b_i t^i$ das so genannte 2. charakteristische Polynom von Φ, so schreiben wir

$$\Phi = (\rho, \sigma).$$

2) Implizitheit ist grundsätzlich zulässig.

3) Für $k = 1$ erhalten wir dei Definition von Einschrittverfahren.

Definition 1.57 (Abschndeidefehler und Konsistenz)

1) Für $h \in (0, \frac{b-a}{k}]$, $x \in [a, b-kh]$ *und* $y \in C^1(I) : y'\xi) = f(\xi, y(\xi))$ *auf* $[x, x+kh]$ *mit* $(x+hi, y(x+hi)) \in K_M, j = 0, \ldots, k$ *heißt*

$$\tau_h(x,y) := \frac{1}{h}\sum_{i=0}^{k} a_i y(x+hi) - \varphi(x, y(x), y(x+h), \ldots, y(x+kh), h)$$

der lokale Abschneidefehler.

b) Ist $\tau_h(x, \tilde{y}) = o(1)$ *für* $h \longrightarrow 0$*, so heißt* Φ *konsistent mit (AWP), falls die Anfangswerte* $u_0, \ldots, u_{k-1}$ *konsistent sind, d.h.* $u_i \overset{h\to 0}{\longrightarrow} y_0$ *für* $i = 0, \ldots, k-1$.
Ist $\tau_h(x, \tilde{y}) = O(h^p)$, $p \in \mathbb{N}$ *für* $h \longrightarrow 0$*, so hat* Φ *die Konsistenzordnung* p*, falls die Anfangswerte die Konsistenzordnung* p *haben, d.h.* $|u_i - y_0| = O(h^p)$ *für* $i = 0, \ldots, k-1$.

1.58 Mehrschrittverfahren resultierend aus Quadraturen

Ansatz: Interpolationsquadratur des Richtungsfeldes.
Aus $y' = f(x,y)$ folgt für $x_j, x_{j+k} \in [a,b]$

$$y(x_{j+k}) = y(x_{j+q}) + \int_{x_{j+q}}^{x_{j+k}} f(t, y(t))dt$$

für $0 \le q \le k-1$.

Idee: Approximation mit geeigneter Quadraturformel.
Approximiere $\int_{x_{j+q}}^{x_{j+k}} f(t,y(t))dt \approx h\sum_{i=0}^{s} b_i \underbrace{f(x_{j+i}, u_{j+i})}_{=:f_{j+i}}$ mit $0 \le s \le k$.

Erhalte MSV:

$$u_{j+k} = u_{j+q} + h\sum_{i=0}^{s} b_i f_{j+i}.$$

Ist $s < k$, oder $b_k = 0 \implies$ Verfahren ist explizit.
Ist $s = k$ und $b_k \neq 0 \implies$ Verfahren ist implizit.

Ansatz zu Bestimmung der b_i**:** Ersetze $f(t, y(t)$ durch Polynom p_j vom Grad s, $0 \le s \le k$, so dass p_j die Daten (x_{j+i}, f_{j+i}) $i = 0, \ldots, s$ interpoliert.
Berechne die Koeffizienten b_i durch Lagrange-Ansatz:
$p_j(t) = \sum_{i=0}^{s} L_{ij}(t) f_{j+i}$ mit $L_{ij}(t) = \prod_{l=0, l\neq i}^{s} \frac{t-x_{j+l}}{x_{j+i}-x_{j+l}} \quad \implies L_{ij}(x_{j+l}) = \delta_{il}$

Dann ist:

$$b_i = \frac{1}{h}\int_{x_{j+q}}^{x_{j+k}} L_{ij}(t)dt.$$

Bemerke: Auf äquidistanten Gittern hängen die b_i nicht von h und j ab.
Im Sinne von 1.55 und 1.56 folgt:

$$u_{j+k} - u_{j+q} = h\varphi(x_j, u_j, \ldots, u_{j+k}, h)$$

mit $\varphi(x, v_0, v_i, \ldots, v_k, h) = \sum_{i=0}^{s} b_i f(x + ih, v_i)$.

$\Longrightarrow$ $\rho(t) = t^k - t^q$ 1. charakteristisches Polynom,

$\Longrightarrow$ $\sigma(t) = \sum_{i=0}^{s} b_i t^i$ 2. charakteristisches Polynom,

$\Longrightarrow$ $\Phi = (\rho, \sigma)$.

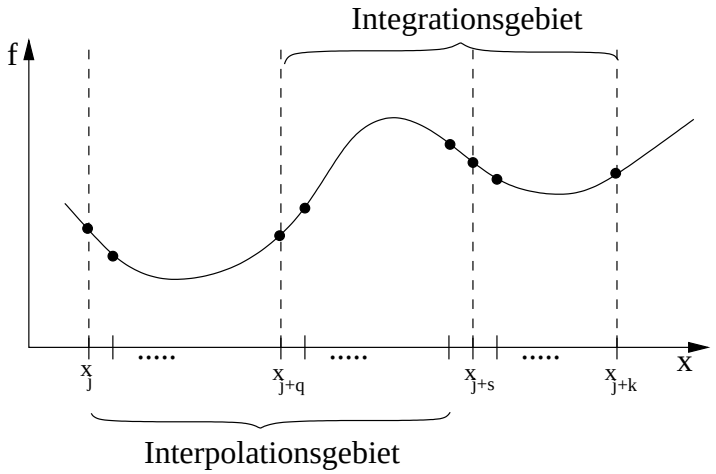

Abbildung 1.8: spezielle lineare Mehrschrittverfahren

Satz 1.59
Seien $f \in C^{s+1}(S)$ und $\Phi = (\rho, \sigma)$ ein k-Schrittverfahren der Klasse 1.58. Dann hat Φ die Konsistenzordnung $s + 1$, falls die Anfangswerte diese Konsistenzordnung haben.
(Für k gerade, $q = 0$ und $s = k$ gilt sogar "s+2"!)

Beweis: Setze $t_i = x + ih$, $i = 0, \ldots, s$. Sei $p(t_i) = y'(t_i) = f(t_i, y(t_i))$

$$\Longrightarrow \quad p(t) - f(t, y(t)) \overset{\text{Numerik I}}{=} \frac{1}{(s+1)!} \prod_{i=0}^{s} (t - t_i)\ y^{(s+2)}(\xi_t) \qquad (*)$$

für ein $\xi_t \in (t_0, t_s)$.

$$\begin{aligned}
\Longrightarrow \quad \tau_h(x, \tilde{y}(x)) &= \tfrac{1}{h}\left[\tilde{y}(x + kh) - \tilde{y}(x + qh)\right] - \tfrac{1}{h} \sum_{i=0}^{s} b_i f(x + ih, \tilde{y}(x + ih)) \\
&= \tfrac{1}{h} \int_{x+qh}^{x+kh} \tilde{y}'(t)dt - \tfrac{1}{h} \int_{x+qh}^{x+kh} p(t)dt \\
&= \tfrac{1}{h} \int_{x+qh}^{x+kh} \Big(f(t, \tilde{y}(t)) - p(t)\Big)dt \\
&\overset{(*)}{=} \tfrac{1}{h} \tfrac{1}{(s+1)!} \int_{x+qh}^{x+kh} \prod_{i=0}^{s} (t - t_i)\ \tilde{y}^{(s+2)}(\xi_t)dt, \\
\Longrightarrow \quad |\tau_h(x, \tilde{y}(x))| &\leq \tfrac{1}{h} \tfrac{1}{(s+1)!} \left\|\tilde{y}^{(s+2)}\right\|_{\infty, I} (sh)^{s+1}(k - q)h \\
&\leq \underbrace{C(\tilde{y}, s)}_{konstant} h^{s+1}. \qquad \square
\end{aligned}$$

1.60 Berechnung der Koeffizienten b_i
Es ist $b_{s-r} = (-1)^r \sum_{i=r}^{s} \binom{i}{r} b_i^*$, $0 \le r \le s$ mit $b_i^* = \int_{q-k}^{k-s} \binom{t+i-1}{i} dt$.
Mit $a := q-k$, $b := k-s$ entstehen die $b_i^* = b_i^*(a,b)$ als Taylorkoeffizienten der Funktion $g(z,a,b)$ bei der Entwicklung um $z=0$, wobei

$$g(z,a,b) := \frac{(1-z)^{-a}}{\ln(1-z)} - \frac{(1-z)^{-b}}{\ln(1-z)}$$

$\Longrightarrow$ $g(z,a,b) = \sum_{i=0}^{\infty} b_i^*(a,b) z^i$ mit $a < b \le 1$, $|z| \le R < 1$. g heißt erzeugende Funktion für b_i^*.

Definition 1.61

1) $q = k-1$ *"Adams-Verfahren"*

2) $q = k-1$, $s = k-1$ *"Adams-Bashforth-Verfahren", explizit*

3) $q = k-1$, $s = k$ *"Adams-Moulton-Verfahren", implizit*

4) $q = k-2$, $s = k-1$ *"Nyström-Verfahren", explizit*

5) $q = k-2$, $s = k$ *"Milne-Simpson-Verfahren", implizit*

1.62 MSV resultierend aus Differentiationsformeln (BDF-Verfahren)

Ansatz: Interpoliere die Daten $(x_j, u_j), \dots, (x_{j+k}, u_{j+k})$ mit $p(x)$ und verwende $p'(x_{j+q})$ als Approximation für $y'(x_{j+q}) = f(x_{j+q}, y(x_{j+q}))$.

$q = k \implies$ implizites Verfahren
$0 \le q \le k-1 \implies$ explizites Verfahren

Lagrangeansatz für $p(x)$:

$$h \cdot p'(x_{j+q}) = \sum_{i=0}^{k} \underbrace{L'_{ij}(x_{j+q})h}_{=:a_i} \cdot u_{j+i} = h \underbrace{f(x_{j+q}, u_{j+q})}_{=:\varphi},$$

wobei a_i unabhängig von j und von h ist.

$\Longrightarrow$ $\rho(t) = \sum_{i=0}^{k} a_i t^i$, $\sigma(t) = t^q$, $\Longrightarrow$ $\Phi = (\rho, \sigma)$ "BDF-k Verfahren".

Satz 1.63
Sei $f \in C^k(S)$. Dann gilt für das BDF-k Verfahren aus 1.62:
Die Konsistenzordnung ist $p = k$ bei hinreichend konsistenten Startwerten.

Beweis: Ü.A.

Satz 1.64 (Charakterisierung der Konsistenz von MSV)
Sei $\Phi(\rho, \varphi)$ k-Schrittverfahren mit $k \ge 1$. Seien $u_0, \dots u_{k-1}$ Startwerte. Dann sind äquivalent:

1) Φ ist konsistent mit (AWP).

2) a) $u_i \longrightarrow y_0$ für $h \longrightarrow 0$; $i = 0, \ldots, k-1$,

b) $\rho(1)\tilde{y}(x) = 0$,

c) $\varphi(x, \tilde{y}(x), \ldots, \tilde{y}(x+kh), h) - \rho'(1) f(x, \tilde{y}(x)) \longrightarrow 0$ für $h \longrightarrow 0$.

Beweis:

"1) $\Longrightarrow$ 2)"

a) folgt direkt aus der Definition von Konsistenz.

b) & c):

$$\begin{aligned} \sum_{i=0}^{k} a_i \tilde{y}(x+ih) &= \sum_{i=0}^{k} a_i \left(\tilde{y}(x) + ih\tilde{y}'(x) + o(h)\right) \\ &= \rho(1)\tilde{y}(x) + \rho'(1)h\tilde{y}'(x) + o(h) \end{aligned}$$

Es folgt:

$$\tau_h(x, \tilde{y}(x)) = \frac{1}{h}\left[\rho(1)\tilde{y}(x) + \rho'(1)h\tilde{y}(x)\right] - \varphi(x, \tilde{y}(x), \ldots, \tilde{y}(x+kh), h) + o(1) \quad (*)$$

$$\begin{aligned} \left|\tfrac{\rho(1)\tilde{y}(x)}{h}\right| &\leq |\tau_h(x, \tilde{y}(x))| + |\rho'(1) f(x, \tilde{y}(x)) - \varphi(\ldots)| + o(1) \\ &\leq C \text{ unabhängig von h} \end{aligned}$$

Also folgt $\rho(1)\tilde{y}(x) = 0$.
Einsetzen in (*) ergibt dann:
$\rho'(1) f(x, \tilde{y}(x)) - \varphi(x, \tilde{y}(x), \ldots, \tilde{y}(x+kh), h) = o(1)$ $(h \to 0)$. $\Longrightarrow$ 2c).

"2) $\Longrightarrow$ 1)" Folgt direkt aus (*). □

Folgerung 1.65
Ist $\Phi = (\rho, \sigma)$ ein lineares k-Schrittverfahren, so gilt:
Φ ist konsistent mit (AWP), genau dann wenn 1.64 2a) und $\rho(1) = 0$, $\rho'(1) = \sigma(1)$ erfüllt ist.

Beweis: Zu zeigen: $\rho'(1) = \sigma(1) \iff$ 1.64 2c) gilt (2a) und 2b) gelten sofort).

Da $\varphi(x, \tilde{y}(x), \ldots, \tilde{y}(x+kh), h) = \sum_{i=0}^{k} b_i f(x+ih, \tilde{y}(x+ih)) \longrightarrow \underbrace{\sum_{i=0}^{k} b_i}_{=\sigma(1)} f(x, \tilde{y}(x))$

für $h \longrightarrow 0$, da $f, \tilde{y}$ stetig.
$\Longrightarrow$ 2c) ist erfüllt $\iff$ $\rho'(1) = \sigma(1)$. □

Bemerkung 1.66
Konsistenz und Lipschitzstetigkeit von φ bzgl. $v_0, \ldots, v_k$ ist *nur* für $k = 1$ hinreichend für die Konvergenz! (vgl. Übungsaufgabe).

Definition 1.67 (Asymptotische Stabilität)
Sei $\Phi = (\rho, \varphi)$ k-Schrittverfahren mit $k \geq 1$. Sei $C_h := \{u : I_h \longrightarrow \mathbb{R}\}$ der Funktionenraum der Gitterfunktionen. Sei $F_h : C_h \longrightarrow C_h$ für $v \in C_h$ definiert durch

$$(F_h(v))_i := v_i - u_i \ , \quad i = 0 \ldots k-1 \ , \quad u_i \text{ Startwerte,}$$
$$(F_h(v))_{j+k} := \frac{1}{h} \sum_{i=0}^{k} a_i v_{i+j} - \varphi(x_j, v_j, \ldots, v_{j+k}, h) \text{ für } j = 0, \ldots, n-k.$$

F_h heißt Defektfunktion.

Das Verfahren Φ heißt asymptotisch stabil, genau dann wenn
$\exists K, H > 0$: $\forall h \in (0, H)$ und $\forall v, w \in C_h$:

$$\|v - w\|_h \leq K \, \|F_h(v) - F_h(w)\|_h \, .$$

Bemerkung 1.68

1) Durch $F_h(u) = 0$ wird die Verfahrenslösung des k-Schrittverfahrens Φ für (AWP) implizit beschrieben, falls $u_0, \ldots, u_{k-1}$ Startwerte sind.

2) Ist $F_h(v_h) = \varepsilon_h$ mit ε_h Störtertm (etwa durch Rundungsfehler), so folgt aus der Stabilität:

$$\|u_h - v_h\|_h \leq K \| \underbrace{F(u_h)}_{=0} - \underbrace{F(v_h)}_{=\varepsilon_h} \|_h = K \, \|\varepsilon_h\|_h$$

3) $F_h(\tilde{y}_{j+k}) = \tau_h(x_{j+k}, \tilde{y})$

Satz 1.69 (Charakterisierung stabiler k-Schrittverfahren)
Sei $\Phi = (\rho, \varphi)$ k-Schrittverfahren mit $k \geq 1$. Gelte:

(i) $f = 0 \implies \varphi = 0$ (bei linearen Verfahren stets erfüllt),

(ii) φ sei Lipschitzstetig bzgl. $v_o, \ldots, v_k$, d.h. es ex. ein $\tilde{L} > 0$ mit

$$|\varphi(x, v_o, \ldots, v_k, h) - \varphi(x, w_o, \ldots, w_k, h)| \leq \tilde{L} \max_{0 \leq i \leq k} |v_i - w_i|$$

(bei linearen Verfahren erfüllt, falls L-Bedingung gilt.)

Dann sind äquivalent:

a) Φ ist asymptotisch stabil.

b) ρ erfüllt die Wurzelbedingung von Dahlquist (vgl. Satz 1.48).

Beweis: Gelte stets (i), (ii).

a) $\Longrightarrow$ b) Sei Φ asymptotisch stabil. Dann folgt für $f = 0$ $\varphi = 0$ nach (i) und es ist

$$(F_h(v))_i = v_i - u_i\,, \quad i = 0, \dots, k-1,$$
$$(F_h(v))_{j+k} = \tfrac{1}{h} \sum_{i=0}^{k} a_i v_{j+i}\,, \quad j = 0, \dots, n-k.$$

Nach a) $\exists K, H > 0$:

$$\|u - v\|_h \le K \|F_h(u) - F_h(v)\|_h = K \|F_h(u - v)\|_h$$

$\forall h \in (0, H)$ und $\forall u, v \in C_h$.
Also gilt für alle $w \in C_h$: $\|w\|_h \le K \|F_h(w)\|_h$.
Für die Lösung $\tilde{u}$ der Gleichung $F_h(\tilde{u})_{j+k} = 0$, $j = 0, \dots, n-k$ (homogene Gleichung) folgt:

$$\|\tilde{u}\|_h \le K \max_{0 \le i \le k-1} |\tilde{u}_i|\,.$$

$\Longrightarrow$ Für beliebige Startwerte $\tilde{u}_0, \dots, \tilde{u}_{k-1}$ ist die Lösung der homogenen k-stufigen Differenzengleichung $\sum_{i=0}^{k} a_i v_{j+i} = 0$ und $v_i = \tilde{u}_i$ für $i = 0, \dots, k-1$ beschränkt.
Also folgt mit Satz 1.48, dass ρ die Wurzelbedingung von Dahlquist erfüllt.

b) $\Longrightarrow$ a) Seien $v, w \in C_h$, dann gilt:

$$(F_h(v) - F_h(w))_i = v_i - w_i\,, \quad i = 0, \dots, k-1$$
$$(F_h(v) - F_h(w))_{j+k} = \tfrac{1}{h} \textstyle\sum_{i=0}^{k} a_i (v_{j+i} - w_{j+i})$$
$$-\varphi(x_j, v_j, \dots, v_{j+k}, h) + \varphi(x_j, w_j, \dots, w_{j+k}, h) \text{ für } j = 0, \dots, n-k.$$

Setze $\delta_\mu := v_\mu - w_\mu\,, \quad \mu = 0, \dots, n$ und $d_\mu := (F_h(v) - F_h(w))_\mu$, $\mu = 0, \dots, n$ sowie

$$(*) \quad \eta_\mu := d_{\mu+k} + \varphi(x_\mu, v_\mu, \dots v_{\mu+k}, h) - \varphi(x_\mu, w_\mu, \dots w_{\mu+k}, h)$$

Dann gilt $\sum_{i=0}^{k} a_i \delta_{i+j} = h\eta_j$ für $j = 0, \dots, n-k$.
Sei δ_μ vorgegeben für $\mu = 0, \dots, k-1$. Mit Satz 1.52 folgt

$$(**) \quad \delta_{j+k} = \sum_{i=0}^{k-1} \delta_i u_{j+k}^{(i)} + \sum_{\mu=0}^{j} h\eta_\mu u_{j+k-\mu-1}^{(k-1)},$$

wobei $u^{(\nu)}$ für $\nu = 0, \dots, k-1$ Basislösung der homogenen Differenzengleichung ist, d.h. $\sum_{i=0}^{k} a_i u_{j+i}^{(\nu)} = 0$ und $u_n^{(\nu)} = \delta_{n,\mu}$ für $0 \le n, \mu \le k-1$.

Nach Voraussetzung a) und Satz 1.48 folgt $\left|u_n^{(\nu)}\right| \le M, \quad n \in \mathbb{N}_0, \quad 0 \le \nu \le k-1$. Mit (*) gilt

$$|\eta_\mu| \overset{(ii)}{\le} |d_{\mu+k}| + \tilde{L} \underbrace{\max_{0 \le i \le k} |\delta_{\mu+i}|}_{=:\varepsilon_\mu}\,.$$

Also folgt aus (**) und $\delta_i = d_i$ für $i = 0, \dots, k-1$:

$$|\delta_{j+k}| \le Mdk + Mh \sum_{\mu=0}^{j} (d + \tilde{L}\varepsilon_\mu) = Md(k + (j+1)h) + Mh\tilde{L} \sum_{\mu=0}^{j} \varepsilon_\mu,$$

mit $d := \|F_h(v) - F_h(w)\|_h = \max_{0 \le \mu \le n} |d_\mu|$.
Beachte: Die rechte Seite ist monoton wachsend in j. Also folgt

$$\varepsilon_j \le Md(k + (j+1)h) + hM\tilde{L} \sum_{\mu=0}^{j} \varepsilon_\mu,$$

$$\Longrightarrow \quad \varepsilon_j(1 - hM\tilde{L}) \leq Md(k + (j+1)h) + hM\tilde{L}\sum_{\mu=0}^{j-1}\varepsilon_\mu.$$

Für $h \leq H := \frac{1}{2M\tilde{L}}$ ist $hM\tilde{L} \leq \frac{1}{2}$ und wir erhalten

$$\varepsilon_j \leq \underbrace{2M(k + (b-a))}_{=:c}\, d + \underbrace{2M\tilde{L}}_{=:D}\sum_{\mu=0}^{j-1} h\varepsilon_\mu.$$

Mit diskretem Lemma von Gronwall folgt:

$$\varepsilon_j \leq ce^{D(b-a)}d, \quad \text{für } j = 0, \ldots, n-k.$$

Mit $K := ce^{D(b-a)}$ folgt hieraus schließlich

$$\|v - w\|_h \leq K \|F_h(v) - F_h(w)\|_h . \qquad \square$$

Folgerung 1.70

(i) Für L-stetige φ sind ESV stabil.

(ii) $\Phi(\rho, \varphi)$ und φ L-stetig, dann gilt:
Φ stabil $\iff$ $\Phi_0 = (\rho, 0)$ stabil.

Satz 1.71 (Konvergenzsatz von Dahlquist)
Sei $\Phi = (\rho, \varphi)$ k-Schrittverfahren mit $k \geq 1$. Sei Φ konsistent mit (AWP) von der Ordnung $p \geq 1$ und seien die Voraussetzungen aus Satz 1.69 erfüllt. Dann sind äquivalent:

(i) Φ ist konvergent mit Ordnung p.

(ii) Φ ist asymptotisch stabil.

Beweis:

(i) $\Longrightarrow$ (ii) Indirekter Beweis. Annahme: Φ ist nicht stabil.
Dann folgt mit den Sätzen 1.48 und 1.69, dass ρ die Wurzelbedingung nicht erfüllt. Folglich hat $\rho(E)u = 0$ eine unbeschränkte Lösung.
Wähle $f = 0, \varphi = 0$ und konsistente Startwerte, so dass u_n divergiert für $h \longrightarrow 0$.Dies ist ein Widerspruch zu (i).

(ii) $\Longrightarrow$ (i) Sei $\tilde{y}_h := \tilde{y}|_{I_h}$ Dann gilt:
$(F_h(\tilde{y}_h))_i = \tilde{y}(x_i) - u_i\ , \quad 0 \leq i \leq k-1,$
$(F_h(\tilde{y}_h))_{j+k} = \tau_h(x_j, \tilde{y})\ , \quad j = 0, \ldots, n-k.$
Also folgt:

$$F_h(\tilde{y}_h) - F_h(u_h) = F_h(\tilde{y}_h) - 0$$

und somit

$$\|\tilde{y} - u_h\|_h \overset{\Phi\text{stabil}}{\leq} K \|F_h(\tilde{y}_h) - F_h(u_h)\|_h = K \|F_h(\tilde{y}_h)\|_h = (O(h^p)),$$

da Φ konsistent von der Ordnung p ist. $\square$

Bemerke: Aus dem Beweis des Konvergenzsatzes von Dahlquist folgt für ein asymptotisch stabiles k-Schrittverfahren $\Phi = (\rho, \varphi)$ die Fehlerabschätzung

$$\|\tilde{y} - u_h\|_h \leq K \max\Big(\max_{i=0}^{k-1} |\tilde{y}_i - u_i|, \max_{i=0}^{n-k} \tau_h(x_i, \tilde{y}) \Big).$$

Folgerung 1.72

1) Für MSV der Klasse 1.58 gilt:
Ist $0 \leq q < k$ und $k - 1 \leq s \leq k$, so ist das Verfahren konvergent mit Ordnung $s + 1$.

2) Für BDF-Verfahren (1.62) gilt:
Die Verfahren sind konvergent, falls

a) $q = k \leq 6$ (implizit),

b) $q = k - 1 \leq 1$ (explizit).

Beweis:

1) Nach Satz 1.59 sind die MSV mit Ordnung $s + 1$ konsistent für $f \in C^{s+1}(S)$ und es ist $\rho(t) = t^k - t^q = t^q(t^{k-q} - 1)$.
$\Longrightarrow$ NST von ρ sind: $\lambda_1 = 0$ (q-fache NST) und $\lambda_{j+2} = e^{i\frac{2\pi}{k-q}j}$ für $j = 0, \ldots, k - q - 1$ (einfache NSTen).
$\Longrightarrow$ ρ erfüllt die Wuzelbedingung.

$\overset{1.68}{\Longrightarrow}$ Verfahren ist asymptotisch stabil.

$\overset{1.70}{\Longrightarrow}$ Konvergenz mit Ordnung $s + 1$.

2) Ü.A.

Beispiel 1.73 (Milne-Simpson für $k = 2$)
Verfahren:

$$u_{j+2} - u_j = \frac{h}{3}[f(x_{j+2}, u_{j+2}) + 4f(x_{j+1}, u_{j+1}) + f(x_j, u_j)]$$

Betrachten wir dieses Verfahren für $y' = \lambda y, \lambda < 0, y(0) = 1$:
$\Longrightarrow$ $\tilde{y}(x) = e^{\lambda x}$.
Für $t := \lambda h$ folgt für das Verfahren:

$$u_{j+2}(1 - \frac{t}{3}) - \frac{4}{3}tu_{j+1} - (1 + \frac{t}{3})u_j = 0,$$

$\Longrightarrow$ $\rho_t(\mu) = \mu^2(1 - \frac{t}{3}) - \mu\frac{4}{3}t - (1 + \frac{t}{3})$.

Nullstellen:

$$\mu_1 = 1 + t + O(t^2) = e^t(1 + O(t^2)),$$
$$\mu_2 = -1 + \tfrac{t}{3} + O(t^2) = -e^{-\frac{t}{3}}(1 + O(t^2))$$

Mit den Startwerten $u_0 = 1, u_1 = 1 + t$ folgt:

$$\begin{aligned} u_j &= [1 + \frac{\lambda h}{2} + O(\lambda^2 h^2)] \underbrace{e^{\lambda x_j}}_{guter\ Term\ (konvergent)} (1 + O(\lambda h)) \\ &\quad + [-\frac{\lambda h}{2} + O(\lambda^2 h^2)] \underbrace{e^{-\lambda x_j}(-1)^j}_{parzitärer\ Term\ (divergent)} (1 + O(\lambda h)). \end{aligned}$$

Der parazitäre Term wird für $\lambda < 0$ exponentiell groß und dominiert das Verhalten.

Satz 1.74 (Konsistenzordnungskriterien für lineare MSV)
Sei $\Phi = (\rho, \sigma)$ lineares MSV mit $k \geq 1$, d.h.

$$\sum_{i=0}^{k} a_i u_{j+i} = h \sum_{i=0}^{k} b_i f_{j+i} \quad \text{mit} \quad f_{j+i} := f(x_{j+1}, u_{j+i}).$$

Für $g \in C^1(\tilde{I}), \tilde{I} = [0, b-a]$ definiere Operator

$$K(g, h) := \frac{1}{h} \sum_{i=0}^{k} (a_i g(ih) - h b_i g'(ih)).$$

Sei $f \in C^p(S)$. Dann sind äquivalent:

(i) $\tau_h(x, \tilde{y}) = O(h^p)$,

(ii) $K(t^s, 1) = 0$ für $s = 0, \ldots, p$,

(iii) $\lambda = 0$ ist $(p+1)$-fache NST von $\rho(e^\lambda) - \lambda\sigma(e^\lambda)$,

(iv) $\lambda = 1$ ist p-fache NST von $\frac{\rho(\lambda)}{\ln(\lambda)} - \sigma(\lambda)$

(v) Die Konsistenzordnung für $y' = y, y(0) = 1$ ist p.

(vi) Die Konsistenzordnung für $y' = sx^{s-1}, y(0) = 1$ ist p für alle $s = 0, \ldots, p$.

Beweis: Wir zeigen die Äquivalenz von (i), (ii) und (iii). Für die Äquivalenz zu (iv), (v) und (vi) siehe Ü.A.

(i) $\iff$ (ii): Ist $f \in C^p(S)$, so folgt $\tilde{y} \in C^{p+1}(I)$
Taylorentwicklung von $\tilde{y}$ und $\tilde{y}'$ liefert:

(*)
$$\begin{aligned} &\tilde{y}(x + ih) = \textstyle\sum_{s=0}^{p} \frac{\tilde{y}^{(s)}(x)}{s!} i^s h^s + O(h^{p+1}), \\ &\tilde{y}'(x + ih) = \textstyle\sum_{s=1}^{p} \frac{\tilde{y}^{(s)}(x)}{s!} \cdot s i^{s-1} h^{s-1} + O(h^p). \end{aligned}$$

Mit $x_i = x + ih$ folgt für den Abschneidefehler:

$$\begin{aligned} h\tau_h(x,\tilde{y}) &= \sum_{i=0}^{k} a_i\tilde{y}(x_i) - hb_i\tilde{y}'(x_i) \\ &\overset{(*)}{=} \rho(1)\tilde{y}(x) + \sum_{s=1}^{p}\left(\tfrac{y^{(s)}(x)}{s!}h^s(\sum_{i=0}^{k}(a_i i^s - b_i s i^{s-1}))\right) + O(h^{p+1}) \\ &\overset{\text{Def. } K(g,h)}{=} \sum_{s=0}^{p} h^s\tilde{y}^{(s)}(x)\tfrac{1}{s!}K(t^s,1) + O(h^{p+1}). \end{aligned}$$

Also folgt die Behauptung.

(ii) $\Longleftrightarrow$ (iii): Nach Teil 1 gilt:

$$h\tau_h(x,e^x) = e^x\sum_{s=0}^{p}\frac{h^s}{s!}K(t^s,1) + O(h^{p+1}).$$

Andererseits folgt aus der Definition:

$$\begin{aligned} h\tau_h(x,e^x) = hK(e^{x+t},h) &= e^x\sum_{i=0}^{k}(a_i e^{ih} - hb_i e^{ih}) \\ &= e^x(\rho(e^h) - h\sigma(e^h)). \end{aligned}$$

Also folgt

$$\underbrace{\rho(e^h) - h\sigma(e^h)}_{=:I(h)} = \sum_{s=0}^{p}\frac{h^s}{s!}K(t^s,1) + O(h^{p+1}).$$

Entwicklung von I in $h = 0$ ergibt die Behauptung. □

Folgerung 1.75

1) Ist $g(t) = y(x+t)$, so folgt $K(g,h) = \tau_h(x,y)$.

2) Ist $f \in C^q(S)$ und $q \geq p$, p Konsistenzordnung, so gilt:

$$\tau_h(x,\tilde{y}) = \sum_{s=p+1}^{q} h^{s-1}\frac{\tilde{y}^{(s)}(x)}{s!}K(t^s,1) + O(h^q).$$

$K(t^{p+1},1)$ heißt "Hauptfehlerkoeffizient".

Bemerkung 1.76

1) Für *spezielle Klassen* von linearen MSV kann analog zu Satz 1.38 eine asymptotische Entwicklung des globalen Fehlers angegeben werden.

2) Insbesondere gilt bei MSV: Die Existenz einer asymptotischen Entwicklung hängt von der Wahl der Startwerte ab.
Siehe hierzu auch Stoer/Bulirsch: Numerische Mathematik II [8].

1.3.3 Das Extrapolationsverfahren von Gragg

Definition 1.77 (Mittelpunktverfahren)
Sei $u_h : I_h \longrightarrow \mathbb{R}$ *definiert durch* $u_h(x_i) = u_i\ ,\ \ i = 0, \ldots, n$ *mit*

$$\left|\begin{array}{l} u_0 = y_0, \\ u_1 = u_0 + hf(x_0, u_0), \\ u_{i+1} - u_{i-1} = 2hf(x_i, u_i),\ i = 1, \ldots, n-1. \end{array}\right.$$

Satz 1.78 (Satz von Gragg)

Für genügend glattes f gilt für den globalen Fehler des Mittelpunktverfahrens:

$$u_j = \tilde{y}(x_j) + \sum_{i=1}^{N} h^{2i}(v_{2i}(x_j) + \underbrace{(-1)^j w_{2i}(x_j)}_{\text{"oszillierender Term"}}) + \mathcal{O}(h^{2N+2})$$

mit $n \in \mathbb{N}$ und v_{2i}, w_{2i} unabhängig von h.

(ohne Beweis)

Problem: Der oszillierende Term führt zu numerischer Auslöschung.

Definition 1.79 (Graggsche Funktion)
Setze

$$s_j := S(x_j, h) := \frac{1}{2}[u_j + u_{j-1} + hf(x_j, u_j)],$$

wobei u_j *Lösung des Mittelpunktverfahrens ist.*
Dann folgt aus Satz 1.78:

$$s_j = \tilde{y}(x_j) + h^2[v_{21}(x_j) + \frac{1}{4}\tilde{y}''(x_j)] + \mathcal{O}(h^4)$$

D.h. die Graggsche Funktion "glättet" den oszillierenden Term.

Algorithmus 1.80 (Algorithmus von Gragg)

Seien $f(x,y), x_0, y_0, H > 0$ und $n \in \mathbb{N}$ gegeben.
Setze $\bar{x} = x_0 + H, h := \frac{H}{n}, x_i := x_0 + ih$, also $x_n = \bar{x}$.
Gesucht ist eine Näherung für $\tilde{y}(\bar{x})$.
Definiere dazu:

$$\begin{aligned} u_0 &= y_0, \\ u_1 &= u_0 + hf(x_0, u_0), \\ u_{i+1} &= u_{i-1} + 2hf(x_i, u_i). \end{aligned}$$

Dann ist

$$S(\bar{x}, h) = \frac{1}{2}[u_n - u_{n-1} + hf(x_n, u_n)]$$

eine gute Näherung von $\tilde{y}(\bar{x})$.

Extrapolationsschema:

Sei $(n_i)_{i\in\mathbb{N}}$ eine Folge mit $0 < n_0 < n_1 < n_2 < \ldots$, z.B. Rombergfolge $n_i := 2^i$.
Setze $h_i = \frac{H}{n_i}$ für $i \in \mathbb{N}_0$

Berechne: $S(\bar{x}, h_0), S(\bar{x}, h_1), S(\bar{x}, h_2), \ldots$

Mit dem Neville-Aitken Schema berechnet man dann

$$\begin{array}{lllll} S(\bar{x}, h_0) = & T_{00} & \searrow & & \\ S(\bar{x}, h_1) = & T_{10} & \to & T_{11} & \\ \vdots & \vdots & & & \ddots \\ S(\bar{x}, h_i) = & T_{i0} & \to & T_{1i} \quad \ldots & T_{ii} \end{array}$$

$$T_{ij} = \frac{-h_{i-j}T_{i,j-1} + h_i T_{i-1,j-1}}{h_i - h_{i-j}}$$

T_{ii} ist Extrapolation in $h = 0 \implies T_{ii} \approx \tilde{y}(\bar{x})$.

1.3.4 Prädiktor-Korrektor-Verfahren

Definition 1.81
Sei $\Phi^* = (\rho^*, \sigma^*)$ *ein explizites lineares k-Schrittverfahren*

$$\sum_{i=0}^{k} a_i^* u_{j+i} = h \sum_{i=0}^{k-1} b_i^* f_{j+i}$$

mit $a_k^* = 1$. Φ^* *heißt Prädiktorformel.*

Sei $\Phi = (\rho, \sigma)$ *implizites lineares k-Schrittverfahren*

$$\sum_{i=0}^{k} a_i u_{j+i} = h \sum_{i=0}^{k} b_i f_{j+i}$$

mit $a_k = 1, b_k \neq 0$. Φ *heißt Korrektorformel*

Algorithmus:

(P) $u_{j+k}^{(0)} + \sum_{i=0}^{k-1} a_i^* u_{j+i} = h \sum_{i=0}^{k-1} b_i^* f_{j+i}$
Für $\mu = 1, \ldots, l$:
  (E) $f_{j+k}^{(\mu-1)} := f(x_{j+k}, u_{j+k}^{(\mu-1)})$
  (C) $u_{j+k}^{(\mu)} + \sum_{i=0}^{k-1} a_i u_{j+i} = h b_k f_{j+k}^{(\mu-1)} + h \sum_{i=0}^{k-1} b_i f_{j+i}$
(E) Setze $u_{j+k} := u_{j+k}^{(l)}$; $f_{j+k} := f(x_{j+k}, u_{j+k}^{(l)})$

Das Verfahren heißt $P(EC)^l E$*-Verfahren.* $(P) \hat{=}$ *explizite Näherung,* $(EC)^l \hat{=}$ *Fixpunktiteration für das implizite Verfahren.*

Bemerkung 1.82
1) Die $P(EC)^l E$-Verfahren sind explizite, i. A. nichtlineare Mehrschrittverfahren.

2) Alternativ zum letzten E-Schritt kann man setzen:

$$u_{j+k} := u_{j+k}^{(l)}, \quad f_{j+k} := f(x_{j+k}, u_{j+k}^{(l-1)}) = f_{j+k}^{(l-1)}.$$

Dieses Verfahren heißt $P(EC)^l$-Verfahren und ist für die Praxis wichtiger.

Satz 1.83
Sei Φ^* explizites k-Schrittverafahren mit Konsistenzordnung $p^* \geq 1$ und Φ sei implizites k-Schrittverfahren mit Konsistenzordnung $p \geq 1$. Dann gilt für das zugehörige $P(EC)^l E$-Verfahren:

1) $P(EC)^l E$ ist konsistent mit Ordnung $\bar{p} := \min\{p^* + l, p\}$

2) Erfüllt ρ die Wurzelbedingung, so ist das $P(EC)^l E$- Verfahren konvergent mit Ordnung $\bar{p}$, falls die Startwerte konvergent mit Ordnung $\bar{p}$ sind.

3) Ist $p < p^* + l$, so hat das $P(EC)^l E$-Verfahren den Hauptfehlerkoeffizienten des Korrektorverfahrens.

(ohne Beweis)

Bemerkung 1.84

1) Eine analoge Aussage zu Satz 1.83 gilt auch für die $P(EC)^l$-Verfahren.

2) In der Praxis wählt man häufig Adams-Bashforth (P) und Adams-Moulton (C) der gleichen Konsistenzordnung und führt nur $l = 1$ Iterationen durch. Man profitiert von dem deutlich kleineren Hauptfehlerkoeffinzienten.

3) Kennt man die Hauptfehlerkoeffizienten zu (P) und (C), so kann man anlog zum Vorgehen bei ESV eine Schrittweitenkontrolle für die $P(EC)^l E$ erhalten.

1.4 Steife Differentialgleichungen und Stabilitätsbegriffe

Beispiel 1.85
Betrachte AWP

$$\left| \begin{array}{l} y'(x) = q(y - g(x)) + g'(x) \\ y(a = 0) = y_0 \; ; \;\; g_0 := g(a) \end{array} \right.$$

Sei $q < 0$, $|q| \gg 0$, $|g'(x)| \ll 1$.
Dann ist die exakte Lösung gegeben durch:

$$\tilde{y}(x) = g(x) + e^{qx}(y_0 - g_0).$$

Für $\tilde{y}_j := \tilde{y}(x_j)$; $g_j := g(x_j)$ folgt:

$$\tilde{y}_{j+1} - g_{j+1} = \underbrace{e^{qh}}_{<1}(\tilde{y}_j - g_j).$$

$\Longrightarrow$ Für die exakte Lösung wird der Abstand von $\bar{y}$ und g monoton kleiner.

Ziel: Das numerische Verfahren soll dieses Verhalten reproduzieren.

a) Euler explizit: Startwert $u_0 = y_0$ und $h > 0$:

$$\Longrightarrow \quad u_{j+1} - g_{j+1} = (1 + hq)(u_j - g_j) + O(h^2)$$

Bei der Näherung erhält man nur dann eine Dämpfung, wenn $|1 + hq| < 1$ $\Longrightarrow$ Bedingung an h.

b) Euler implizit: Startwert $u_0 = y_0$ und $h > 0$:

$$\Longrightarrow \quad u_{j+1} - g_{j+1} = \underbrace{\frac{1}{1 - hq}}_{<1,\ \forall h>0}(u_j - g_j) + O(h^2)$$

$\Longrightarrow$ Immer Dämpfung, unabhängig von h!

Allgemeine Situation 1.86
Seien u und v Lösungen des Systems

$$y' = f(x, y) \;\; \text{im} \;\; \mathbb{R}^n$$

mit $u(a) = u_0$, $v(a) = v_0$. Dann ist:

$$u'(x) - v'(x) = J(x)(u(x) - v(x))$$

mit

$$J(x) = \int_0^1 f_y(x, tu(x) + (1 - t)v(x))dt.$$

[Für $n = 1$, z.B. $f(x, y) = qy$, $v(x) = g(x)$ (siehe Beispiel 1.85)]

Definition 1.87 (Steife Differentialgleichung)
Das AWP $y' = f(x, y)$ heißt steife Differentialgleichung *rechts von a, falls gilt:*

1) λ EW von $J(x)$ $\implies$ Re $\lambda < 0$.

2) Für $n \geq 2$ gilt:
$\exists$ EW λ_1 mit $\lambda_1 < 0$ und $|\lambda_1| \gg 0$ und $\exists$ EW λ_2 mit $|\lambda_2| \ll |\lambda_1|$.

Beispiel 1.88

y_i $i = 1, 2, 3$ seien Konzentrationen von Substanzen zur Zeit t, k_i seien Reaktionsraten:

$$\begin{aligned} y_1' &= -k_1 y_1 + k_2 y_2 y_3, \\ y_2' &= k_1 y_1 - k_2 y_2 y_3 - k_3 y_2^2, \\ y_3' &= k_3 y_2^2. \end{aligned}$$

Seien $y_1(0) = 1$, $y_2(0) = y_3(0) = 0$.

Dann folgt für die EW von $J(x)$

x	λ_1	λ_2	λ_3
0	0	0	$-0,04$
10^{-2}	0	$-0,36$	-2180
100	0	$-0,0048$	-4240
∞	0	0	-10^4

$\implies$ "sehr steifes System"

Definition 1.89 ((Absolute Stabilität))
Sei $y' = qy$ Testgleichung. *Sei* $\Phi = (\rho, \sigma)$ *lineares k-Schrittverfahren, d.h.*

$$(*) \quad \sum_{i=0}^{k} (a_i - hqb_i) u_{j+i} = 0 \; ; \quad j = k, \ldots, n-k$$

und $u_0, \ldots, u_{k-1}$ *gegeben.*

Definiere das Stabilitätspolynom

$$\rho_t(\lambda) = \rho(\lambda) - t \cdot \sigma(\lambda)$$

D.h. $\rho_t(\lambda)$ *ist charakteristisches Polynom von* $(*)$ *mit* $t = hq$.

Φ *heißt* absolut stabil *für* $t \in \mathbb{C}$ *falls gilt:*

$$t \in D_s := \{t \in \mathbb{C} \mid \rho_t(\lambda) = 0 \; für \; ein \; \lambda \in \mathbb{C} \implies |\lambda| < 1\}.$$

D_s *heißt Stabilitätsgebiet von* Φ.
Für ESV $\Phi = \varphi$ *der Form* $u_{j+1} = g(t)u_j$ *ist das Stabilitätspolynom gegeben durch*

$$\rho_t(\lambda) = \lambda - g(t).$$

Beispiel 1.90 (Stabilitätsgebiete für explizite Runge-Kutta-Verfahren)
Siehe Folie aus Deuflhard, Bornemann [4, Seite 230].

Beispiel 1.91
Das Verfahren von Milne-Simpson ist asymptotisch stabil, aber nicht absolut stabil. Dies wurde in Beispiel 1.73 gezeigt.

Weitere Forderung an ” gute” numerische Verfahren:

A) Ein fallender Exponentialterm der Lösung von $y' = qy$ soll stets (für alle h) fallend genähert werden

Dies führt auf den Begriff der A-Satabilität

Definition 1.92 ((A-Stabilität))
Φ *heißt A-stabil, gdw.* $D_s \supset H_- = \{t \in \mathbb{C} \mid \; Re \; t < 0\}$.

Abschwächungen dazu sind:
$A(\alpha)$*-Stabilität:* $\iff \; D_s \supset \{t \in \mathbb{C} \mid \; |arg(-t)| < a\}$
$A(0)$*-Stabilität:* $\iff \; D_s \supset \mathbb{R}^-$

Satz 1.93
Sei $\Phi = (\rho, \sigma)$ lineares MSV mit $k \geq 2$ und Φ sei A-stabil. Dann gilt

1) Φ ist implizit,

2) Φ hat Konvergenzordnung 2.

Das Trapezverfahren hat unter allen A-stabilen MSV die kleinste Fehlerschranke.

(ohne Beweis)

Satz 1.94

1) Die BDF-k Verfahren sind A-stabil für $1 \leq k = s \leq 2$.

2) Implizite Runge-Kutta Verfahren sind A-stabil.

(ohne Beweis)

1.5 Numerische Lösung von Randwertproblemen

Beispiel 1.95 (Wärmeleitender Stab)

Gegeben: Stab $\hat{=}\ [a,b]$
Wärmequellendichte $f(x),\ x \in [a,b]$
Randwerte: $y(a) = y_l, y(b) = y_r$
Wärmleitfähigkeitskoefffizient $k(x),\ x \in [a,b]$

Gesucht: Temperatur $y(x)$ mit

$$(*) \quad \left| \begin{array}{l} -(k(x)y'(x))' = f(x) \ \ \forall x \in (a,b), \\ y \in C^2((a,b)) \cap C^0([a,b]), \\ y(a) = u_l,\ y(b) = u_r. \end{array} \right.$$

Formulierung als System 1. Ordnung: Setze $y_1 = y(x),\ y_2(x) = y'(x)$

$$(*) \implies \left| \begin{array}{l} y_1'(x) = y_2(x), \\ y_2'(x) = -\frac{k'(x)}{k(x)} y_2(x) - \frac{f(x)}{k(x)} \end{array} \right.$$

mit $y_1(a) = u_l,\ y_1(b) = u_r$.

Definition 1.96 (Lineare Randwertprobleme)
Seien $B_a, B_b \in \mathbb{R}^{n\times n},\ A \in C^0(I, \mathbb{R}^{n\times n}),\ I := [a,b],\ f \in C^0(I,\mathbb{R}^n),\ g \in \mathbb{R}^n$.
Dann heißt $y : I \longrightarrow \mathbb{R}^n$ *Lösung des* linearen Randwertproblems (RWP), *falls gilt:*

1) $y \in C^1(I, \mathbb{R}^n)$,

2) $y'(x) = Ay + f$ *in* I,

3) $B_a y(a) + B_b y(b) = g$.

Beispiel 1.97
Die Differentialgleichung

$$y''(x) + y(x) = 0,\ \ x \in [0,\pi] \iff \left| \begin{array}{l} y_1'(x) - y_2(x) = 0 \\ y_2'(x) + y_1(x) = 0 \end{array} \right.$$

hat die allgemeine Lösung $y(x) = c_1 \sin(x) + c_2 \cos(x)$.
Für verschiedene Randbedingungen ergibt sich unterschiedliches Vehalten:

1) $y(0) = y(\pi);\ y'(0) = y'(\pi) \implies$ ergibt die eindeutige Lösung $y \equiv 0$.

2) $y(0) = y(\pi) = 0 \implies$ ergibt unedlich viele Lösungen $y(x) = c_1 \sin(x)$.

3) $y(0) = 0,\ y(\pi) = 1 \implies$ ergibt keine Lösung.

Ziel: Bedingung für eindeutige Lösbarkeit!

Definition 1.98 (Fundamentalsystem)
Für

$$(*) \qquad y' = Ay + f \quad in\ I$$

definiere das Fundametalsystem $\{y_1, \ldots, y_n\}$, $y_i : I \longrightarrow \mathbb{R}^n$, *durch*

1) $y_i \in C^1(I, \mathbb{R}^n)$,

2) $y_i' = Ay_i$ *in* I,

3) $y_i(a) = e_i$, *mit* e_i *i-ter Einheitsvektor im* $\mathbb{R}^n$.

Die Matrix

$$Y(x) := \begin{pmatrix} y_{11}(x) & \ldots & y_{n1}(x) \\ \vdots & \ddots & \vdots \\ y_{1n}(x) & \ldots & y_{nn}(x) \end{pmatrix}$$

heißt Fundamentalmatrix.
Ist $y_0 \in C^1(I, \mathbb{R}^n)$, $y_0(a) = 0$ *eine spezielle Lösung von* $(*)$, *so läßt sich jede Lösung von* $(*)$ *darstellen als*

$$y(x) = y_0(x) + Y(x) \cdot s$$

mit einem Vektors $\in \mathbb{R}^n$.

Satz 1.99 (Existenz und Eindeutigkeit linearer RWPe)
Die folgenden Aussagen sind äquivalent:

1) (RWP) besitzt eine eindeutige Lösung.

2) Das zugehörige homogene RWP mit $f = 0$, $g = 0$ besitzt nur die triviale Lösung.

3) Die Matrix $B_a + B_b Y(b) \in \mathbb{R}^{n \times n}$ ist regulär.

Beweis: Seien $y_0(x)$ und $Y(x)$ wie in Definition 1.98 definiert.
Dann gilt $y_0(a) = 0$ und $Y(a) = I$.
Also ist $y(x) = y_0(x) + Y(x)s$ genau dann Lösung von (RWP), wenn gilt

$$[B_a + B_b Y(b)]s = g - B_b y_0(b)$$

Also ist $s \in \mathbb{R}^n$ genau dann eindeutig bestimmt, wenn 2) oder 3) gelten. □

1.5.1 Sturm-Liouville Probleme

Definition 1.100 (Sturm-Liouville Probleme)
Seien $p \in C^1(I),\ q, f \in C^0(i),\ I = [a,b]$ *gegeben mit*

$$q(x) \geq 0\ \forall x \in I,\ p(x) \geq p_0 > 0\ \forall x \in I$$

Dann heißt $y \in C^2((a,b)) \cap C^0([a,b]) \cap C^1((a,b])$ *Lösung des* Sturm-Liouville Problems (SLP) *mit homogenen Dirichlet und Neumann Randwerten, falls gilt:*

$$(SLP) \left| \begin{array}{l} -(p(x)y'(x))' + q(x)y(x) = f(x)\ \ \forall x \in I \\ y(a) = 0;\ y'(b) = 0 \end{array} \right.$$

Bemerkung: Allgemein Randbedingungen sind: $\left| \begin{array}{l} \alpha_1 y'(a) + \alpha_0 y(a) = g_a \\ \beta_1 y'(b) + \beta_0 y(b) = g_b \end{array} \right.$

Definition 1.101 (Variationsproblem)
Für $v \in X := \{u \in C^1([a,b]) \mid u(a) = 0\}$ *definiere das sogenannte* Energiefunktional $I : X \to \mathbb{R}$ *durch*

$$I(v) = \frac{1}{2}\int_a^b p(x)(v'(x))^2 dx + \frac{1}{2}\int_a^b q(x)(v(x))^2 dx - \int_a^b f(x)v(x)dx.$$

Ziel: Finde ein $y \in X$ mit $I(y) = \inf_{v \in X} I(v)$.

Lemma 1.102 (Eulergleichung und natürliche Randbedingung)
Ist $y \in X$, so dass

$$I(y) \leq I(v)\ \ \forall v \in X,$$

so gilt $\forall \varphi \in X$:

$$(*)\quad \int_a^b py'\varphi' + qy\varphi = \int_a^b f\varphi$$

$(*)$ heißt schwache Formulierung der Eulergleichung.

Ist zusätzlich $y \in C^2((a,b))$, so folgt weiter:

$$(**) \left| \begin{array}{l} -(py')' + qy = f \quad \text{in } I, \\ y(a) = 0;\ y'(b) = 0. \end{array} \right.$$

Beweis:
Teil 1: Sei $y \in X$ mit $I(y) \leq I(v)\ \ \forall v \in X$.
$\implies I(y) \leq I(y + \varepsilon\varphi)\ \ \forall \varepsilon \in \mathbb{R}\ ;\ \ \forall \varphi \in X.$
$\implies G(0) \leq G(\varepsilon)\ \ \forall \varepsilon \in \mathbb{R}$, wobei $G(\varepsilon) := I(y + \varepsilon\varphi)$.
Da G differezierbar in ε is, folgt $G'(0) = 0$.
Es ist

$$\begin{aligned} G(\varepsilon) &= \tfrac{1}{2}\textstyle\int_a^b p(y' + \varepsilon\varphi')^2 + q(y + \varepsilon\varphi)^2 - 2f(y + \varepsilon\varphi) \\ &= I(y) + \varepsilon\textstyle\int_a^b py'\varphi' + qy\varphi - f\varphi + \varepsilon^2(I(\varphi) + \int_a^b f\varphi) \end{aligned}$$

$$\implies G'(0) = \textstyle\int_a^b py'\varphi' + qy\varphi - f\varphi \overset{!}{=} 0, \quad \forall \varphi \in X.$$

Teil 2: Ist $y \in C^2((a,b))$, so folgt mit partieller Integration

$$\int_a^b py'\varphi' = -\int_a^b (py')'\varphi + \Big[py'\varphi\Big]_a^b.$$

Einsetzen in $(*)$ ergibt

$$(\dagger) \quad -\int_a^b (py')'\varphi + \int_a^b qy\varphi - f\varphi + [py'\varphi]_a^b = 0.$$

Also folgt für alle $\varphi \in C_0^1((a,b))$:

$$\int_a^b [-(py')' + qy - f]\varphi = 0.$$

$$\Longrightarrow \quad -(py')' + qy = f.$$

Durch Einsetzen in $(\dagger)$ erhält man für alle $\varphi \in X$:

$$0 = [py'\varphi]_a^b = p(b)y'(b)\varphi(b) - p(a)y'(a)\underbrace{\varphi(a)}_{=0} = p(b)y'(b)\varphi(b)$$

Wähle $\varphi(b) \neq 0 \implies$, so folgt $y'(b) = 0$, da $p(b) > 0$. $y'(b) = 0$ heißt natürliche Randbedingung. □

Nächstes Ziel: Existenz einer schwachen Lösung, d.h.

$$\exists y \in X \text{ mit } I(y) = \inf_{v \in X} I(v).$$

Dazu benötigen wir folgenden Hilfssatz.

Satz 1.103 (Poincaré Ungleichung)
Für alle $v \in X$ gilt mit einer Konstanten $c_p \leq \frac{1}{2}(b-a)^2$:

$$\int_a^b (v(x))^2 dx \leq c_p \int_a^b (v'(x))^2 dx$$

Beweis: Es ist $|v(x)| = |v(x) - \underbrace{v(a)}_{=0}| \leq \int_a^x |v'(s)|\, ds$.

$\Longrightarrow \quad (v(x))^2 \leq \left(\int_a^x |v'(s)|\, ds\right)^2 \leq (x-a)\int_a^x |v'(s)|^2\, ds.$

$\Longrightarrow \quad \int_a^b (v(x))^2 dx \leq \int_a^b (x-a)\int_a^b |v'(s)|^2\, ds = \frac{1}{2}(b-a)^2 \int_a^b |v'(s)|^2\, ds.$ □

Lemma und Definition 1.104
Definiere auf X die Norm

$$\|v\|_X := \left(\int_0^1 (v(x))^2 + (v'(x))^2 dx\right)^{\frac{1}{2}}.$$

Sei $(v_n)_{n\in\mathbb{N}}$, $v_n \in X$ eine Minimalfolge, d.h.

$$\lim_{n \longrightarrow \infty} I(v_n) = \inf_{v \in X} I(v).$$

Dann ist $(v_n)_{n\in\mathbb{N}}$ eine Cauchyfolge in X, d.h.

$$\|v_n - v_m\|_X \longrightarrow 0 \quad (n, m \longrightarrow \infty).$$

Beweis: Setze $d := \inf_{v\in X} I(v)$.

1. Schritt: Zeige $d > -\infty$: Es ist

$$\begin{array}{rcl} I(v) & \overset{\text{Vor. in (SLP)}}{\geq} & \frac{p_0}{2}\int_a^b (v'(x))^2 dx - \left(\int_a^b (f(x))^2\right)^{\frac{1}{2}} \left(\int_a^b (v(x))^2\right)^{\frac{1}{2}} \\ & \geq & \frac{p_0}{2}\int_a^b (v'(x))^2 dx - \underbrace{\sqrt{c_p}\left(\int_a^b (f(x))^2\right)^{\frac{1}{2}}}_{=:a} \underbrace{\left(\int_a^b (v'(x))^2\right)^{\frac{1}{2}}}_{=:b}. \end{array}$$

Mit der Ungleichung $|ab| \leq \frac{\delta}{2}a^2 + \frac{1}{2\delta}b^2 \ \ \forall a, b \in \mathbb{R}$ folgt

$$I(v) \geq \left(\frac{p_0}{2} - \frac{\delta}{2}\right)\int_a^b (v'(x))^2 dx - \frac{\sqrt{c_p}}{2\delta}\|f\|^2_{L^2(a,b)}.$$

Für $\delta = p_0$ folgt: $I(v) \geq -\frac{\sqrt{c_p}}{2p_0}\|f\|^2_{L^2(a,b)}$. Also folgt $d > -\infty$, da $f \in C^0([a,b])$.

2. Schritt: Sei $(v_n)_{n\in\mathbb{N}}$ Folge in X mit $\lim_{n\to\infty} I(v_n) = d > -\infty$.

Zeige: (v_n) ist Cauchyfolge in X.

Notation: Definiere die Bilinearform $B : X \times X \longrightarrow \mathbb{R}$ durch

$$B(v, w) := \int_a^b pv'w' + \int_a^b qvw.$$

Dann gilt $B(v, v) = \int_a^b p\,(v')^2 + \int_a^b qv^2 \geq p_0\int_a^b (v'(x))^2 dx$.

Aufgrund der Poincaré Ungleichung gilt weiter:

$$\begin{array}{rcl} B(v, v) & \geq & p_0\left(\int_a^b (v'(x))^2 dx + c_p\int_a^b (v'(x))^2 dx\right)\frac{1}{1+c_p} \\ & \geq & \frac{p_0}{1+c_p}\|v\|^2_X. \end{array}$$

Damit folgt:

$$\begin{array}{rcll} \frac{p_0}{1+c_p}\|v_n - v_m\|_X & \leq & B(v_n - v_m, v_n - v_m) & \\ & \overset{\text{Parallelogrammidentität}}{=} & 2B(v_n, v_n) + 2B(v_m, v_m) - 4B(\frac{v_n+v_m}{2}, \frac{v_n+v_m}{2}) & \\ & \overset{I(v)=\frac{1}{2}B(v,v)-\int_a^b fv}{=} & 4[I(v_n) + I(v_m) - 2I(\frac{v_n+v_m}{2})] & \\ & \leq & 4[I(v_n) + I(v_m) - 2d] & \\ & \longrightarrow & 4[d + d - 2d] = 0 \quad \text{für } n, m \to \infty. & \end{array}$$

Also ist $(v_n)_{n\in\mathbb{N}}$ Cauchyfolge. □

Problem: Der Raum $(X, \|\cdot\|_X)$ ist nicht vollständig!

Lösung: Vervollständige X bzgl. der Norm $\|\cdot\|_X$ und erhalte vollständigen Raum $\bar{X}^{\|\cdot\|_X}$.

Beispiel aus der Analysis III: Ist $Y = C^0([a,b])$, so ist $\bar{Y} = L^2((a,b))$ die Vervollständigung von Y bzgl. $\|v\|_Y = \left(\int_a^b v^2\right)^{\frac{1}{2}}$.

Um $\bar{X}$ zu charakterisieren, benötigen wir den Begriff der schwachen Ableitung.

Definition 1.105 (Schwache Ableitung)
$v \in L^1(a,b)$ besitzt eine schwache Ableitung k-ter Ordnung $D^k v \in L^1(a,b)$, falls für alle $\varphi \in C_0^\infty(a,b)$ gilt:

$$\int_a^b v(x)\varphi^{(k)}(x)dx = (-1)^k \int_a^b (D^k v)(x)\varphi(x)dx.$$

Beispiel 1.106
Sei $v(x) = |x|$ mit $x \in [-1,1]$.
Dann gilt: $D^1(v)(x) = \operatorname{sign}(x)$, denn für $\varphi \in C_0^\infty(-1,1)$ gilt:

$$\begin{aligned}\int_{-1}^1 v(x)\varphi'(x)dx &= \int_{-1}^0(-x)\varphi'(x)dx + \int_0^1 x\varphi(x)dx \\ &= [(-x)\varphi(x)]_{-1}^0 + \int_{-1}^0 \varphi(x)dx + [x\varphi(x)]_0^1 - \int_0^1 \varphi(x)dx \\ &= -\int_{-1}^0(-1)\varphi(x)dx - \int_0^1 \varphi(x)dx \\ &= -\int_{-1}^1 \operatorname{sign}(x)\varphi(x)dx.\end{aligned}$$

$\rightsquigarrow$ Allgemein gilt in einer Raumdimension: Stückweise differenzierbare Funktionen, die global stetig sind, sind schwach differenzierbar. Die schwache Ableitung ist durch die stückweise definierte Ableitung gegeben.

Satz 1.107 (Eindimensionale Sobolevräume)

1) Der Sobolevraum

$$\begin{aligned}H^m(a,b) := \{&v \in L^2(a,b) \mid v \text{ hat schwache Ableitungen} \\ &D^k v \in L^2(a,b)\ \forall k = 0,\dots,m\}\end{aligned}$$

ist mit dem Skalarprodukt

$$(u,v)_{H^m(a,b)} := \sum_{k=0}^m \int_a^b D^k u D^k v$$

ein Hilbertraum. Durch $\|u\|_{H^m(a,b)} := \sqrt{(u,u)_{H^m(a,b)}}$ erhalten wir eine Norm auf $H^m(a,b)$.

2) $u \in H^m(a,b)$ ist fast überall gleich einer Funktion $\tilde{u} \in C^{m-1}([a,b])$ und es ist

$$\|\tilde{u}\|_{C^{m-1}([a,b])} \le c\,\|u\|_{H^m(a,b)}.$$

3) Zu $u \in H^m(a,b)$ gibt es eine Folge $(u_j)_{j\in\mathbb{N}}$, $u_j \in C^m([a,b])$, so dass

$$\|u - u_j\|_{H^m(a,b)} \longrightarrow 0 \qquad (j \to \infty).$$

Ist $\tilde{u}(a) = 0$, so kann man u_j so wählen, dass $u_j(a) = 0$ ist.

Beweis: z.B [Alt. Lineare Funktionalanalysis, Springer, 1992].

Folgerung 1.108
Die Vervollständigung von X bzgl. $\|\cdot\|_X$ ist gegeben durch

$$\bar{X} = \{v \in H^1(a,b) \mid v(a) = 0\}.$$

Satz 1.109 (Existenz und Eindeutigkeit einer schwachen Lösung von (SLP))
Seien die Voraussetzungen aus Definition 1.100 erfüllt. Dann gibt es genau ein $y \in \bar{X}$, so dass $\forall \varphi \in \bar{X}$ gilt:

$$\int_a^b py'\varphi' + \int_a^b qy\varphi = \int_a^b f\varphi.$$

Beweis:

Existenz: Sei $(v_n)_{n\in\mathbb{N}}$ Minimalfolge in $\bar{X}$ (anstelle von X).
Analog zu Lemma 1.104 folgern wir, dass $(v_n)_{n\in\mathbb{N}}$ Cauchyfolge in $\bar{X}$ ist.
Da $\bar{X}$ vollständig ist (Satz 1.107) $\exists y \in \bar{X}$ mit $\|v_n - y\|_{\bar{X}} \to 0 \ \ (n \to \infty)$.

Zeige: $I(y) = d \ (= \inf_{v\in\bar{X}} I(v))$.

Es ist

$$\begin{aligned}
&|I(\ y\ ) - I(v_n)| = \\
&\quad = \left|\tfrac{1}{2}\int_a^b p((y')^2 - (v_n')^2) + q(y^2 - v_n^2) - \int_a^b f(y - v_n)\right| \\
&\quad \le \tfrac{1}{2}\|p\|_\infty \int_a^b |y'^2 - v_n'^2| + \tfrac{1}{2}\|q\|_\infty \int_a^b |y^2 - v_n^2| + \int_a^b |f|\,|y - v_n| \\
&\quad \le \tfrac{1}{2}\|p\|_\infty \int_a^b |y' - v_n'|\,(|y'| + |v_n'|) + \tfrac{1}{2}\|q\|_\infty \int_a^b |y - v_n|\,(|y| + |v_n|) \\
&\qquad + \int_a^b |f|\,|y - v_n| \\
&\overset{\text{C.S.}}{\le} \underbrace{\left[\left(\frac{1}{2}\|p\|_\infty + \frac{1}{2}\|q\|_\infty\right)(\|y\|_{\bar{X}} + \|v_n\|_{\bar{X}}) + \|f\|_{L^2(a,b)}\right]}_{\le C < \infty} \underbrace{\|y - v_n\|_{\bar{X}}}_{\to 0 \ (n\to\infty)}
\end{aligned}$$

Also folgt $I(y) = \lim_{n\to\infty} I(v_n) = d$ und damit die Existenz der Lösung.

Eindeutigkeit: Seien $y_1,\ y_2$ Lösungen, so folgt für $v = y_1 - y_2$:

$$\int_a^b pv'\varphi' + \int_a^b qv\varphi \ \ \forall \varphi \in \bar{X}.$$

Wähle $\varphi = v$:

$$\overset{1.106}{\Longrightarrow} \quad 0 = B(v,v) \ge \frac{p_0}{1 + c_p}\|v\|_X \quad \Longrightarrow \quad \|v\|_X = 0$$

$$\Longrightarrow \quad v = 0 \quad \Longrightarrow \quad y_1 = y_2 \qquad \square$$

Satz 1.110 (A priori Abschätzung)
Für jede schwache Lösung $y \in \bar{X}$ von (SLP) gilt:

1) $\|y'\|_{L^2(a,b)} \leq C_1 \|f\|_{L^2(a,b)}$.

2) Ist $y \in H^2(a,b)$, so gilt
$\|y''\|_{L^2(a,b)} \leq C_2 \|f\|_{L^2(a,b)}$.

Dabei sind $C_1 = \frac{\sqrt{c_p}}{p_0}$, $C_2 = \frac{\sqrt{3}}{p_0}\left(\frac{\sqrt{c_p}}{p_0}\|p'\|_\infty + \frac{c_p}{p_0}\|q\|_\infty + 1\right)$
Beweis: Ü.A.

1.5.2 Das Ritz-Galerkin Vefahren

Definition 1.111 (Ritz-Galerkin Verfahren für (SLP))
Sei $X := \{v \in H^1(a,b) \mid v(a) = 0\}$ und $I : X \longrightarrow \mathbb{R}$ das Energiefunktional aus Definition 1.101. Sei $X_h \in X$ ein endlichdimensionaler Teilraum. Dann heißt $u_h \in X_h$ Ritz-Galerkin Approximation *der schwachen Lösung $y \in X$ von (SLP), g.d.w*

$$I(u_h) = \inf_{v_h \in X_h} I(v_h).$$

Idee: Minimierung von I auf endlichen Teilräumen.

Folgerung 1.112 (Schwache diskrete Differentialgleichung)

1) Da $X_h \in X$ ist, folgt $I(u) \leq I(u_h)$.

2) Ist $u_h \in X_h$ Ritz-Galerkin Approximation von (SLP), so gilt:

$$(D - SLP) \qquad B(u_h, \varphi_h) = \int_a^b f\varphi_h, \qquad \forall \varphi_h \in X_h$$

mit $B(u,v) := \int_a^b pu'v' + \int_a^b quv$.

Beweis: Analog zum kontinuierlichen Fall in Lemma 1.102.

Bemerkung 1.113 (Formulierung von (D-SLP) als lineares Gleichungssystem)
Seien $N := \dim(X_h)$ und $\{\varphi_1, \ldots, \varphi_N\}$ eine Basis von X_h. Dann läßt sich u_h darstellen als

$$u_h(x) = \sum_{j=1}^{N} u_j \varphi_j(x)$$

mit Koeffizienten $u_j \in \mathbb{R},\ j = 1, \ldots, N$.
Damit ist (D-SLP) äquivalent zu

$$\sum_{j=1}^{N} B(\varphi_j, \varphi_k) u_j = \int_a^b f\varphi_k \quad \forall k = 1, \ldots, N.$$

Mit den Definitionen $u := (u_1, \ldots, u_N);\ S_{kj} := B(\varphi_j, \varphi_k),\ k, j = 1, \ldots, N;\ b_k := \int_a^b f\varphi_k;\ b := (b_1, \ldots, b_N)$ ist somit (D-SLP) äquivalent zu dem linearen Gleichungssystem

$$\mathbf{Su} = \mathbf{b}$$

Die Matrix S wird "Steifigkeitsmatrix" genannt.

Satz 1.114 (Abstrakte Fehlerabschätzung)
Seien X ein normierter Raum, $X_h \subset X$ ein Teilraum. Weiter sei $f \in X' = L(X; \mathbb{R})$ ein lineares Funktional auf X.
Gilt dann

$$u \in X : B(u, \varphi) = f(\varphi), \qquad \forall \varphi \in X,$$

$$u_h \in X_h : B(u_h, \varphi_h) = f(\varphi_h) \qquad \forall \varphi_h \in X_h$$

mit einer Bilinearform $B : X \times X \longrightarrow \mathbb{R}$ die koerziv ist, d.h.

$$\exists C_0 > 0, \text{ so dass } \forall \varphi \in X : B(\varphi, \varphi) \geq C_0 \|\varphi\|_X^2$$

und stetig, d.h.

$$\exists C_1 \geq 0, \text{ so dass } \forall \varphi, \psi \in X : |B(\varphi, \psi)| \leq C_1 \|\varphi\|_X \cdot \|\psi\|_X,$$

so gilt die Fehlerabschätzung

$$\|u - u_h\|_X \leq \frac{C_1}{C_0} \inf_{v_h \in X_h} \|u - v_h\|_X$$

und es ist

$$B(u - u_h, \varphi_h) = 0 \qquad \forall \varphi_h \in X_h. \quad \text{"Galerkin-Orthogonalität"}$$

Beweis: Wegen $X_h \subset X$, folgt $\forall \varphi_h \in X_h$
$B(u, \varphi_h) = f(\varphi_h)$ und $B(u_h, \varphi_h) = f(\varphi_h)$
$\Longrightarrow\ B(u - u_h, \varphi_k) = 0\ \forall \varphi_h \in X_h$ (Galerkin-Orthogonalität $\surd$).

Mit Koerzivität und Stetigkeit von B folgt nun:

$$\begin{aligned}
C_0 \|u - u_h\|_X^2 &\leq B(u - u_h, u - u_h) \\
&= B(u - u_h, u) - \underbrace{B(u - u_h, u_h)}_{=0} \\
&\overset{v_h \in X_h}{=} B(u - u_h, u) - \underbrace{B(u - u_h, v_h)}_{=0} \\
&= B(u - u_h, u - v_h) \\
&\overset{\text{stetig}}{\leq} C_1 \|u - u_h\|_X \|u - v_h\|_X
\end{aligned}$$

Also folgt: $\|u - u_h\|_X \leq \frac{C_1}{C_0} \|u - v_h\|_X \quad \forall v_h \in X_h$. □

Bemerkung 1.115
Für (SLP) ist

$$B(u,v) = \int_a^b pu'v' + \int_a^b quv$$

und

$$f(v) := \int_a^b fv.$$

Ist $X := \{v \in H^1(a,b) \,|\, v(a) = 0\}$, so ist $f \in X'$, da $\forall v \in X$ gilt

$$|f(v)| \leq \|f\|_{L^2(a,b)} \|v\|_{L^2(a,b)} \leq \|f\|_{L^2(a,b)} \|v\|_X .$$

Beispiel 1.116 (Wahl von X_h)

1) Polynomapproximation: $X_h = \mathbb{P}_k$ ($\hat{=}$ Polynome von Grad $\leq k$)
   $\Longrightarrow \quad N = \dim(X_h) = k + 1.$

2) Eigenräume: $X_h := \text{span}\{\varphi_1, \ldots, \varphi_N\}$, wobei $\varphi_1, \ldots, \varphi_N$ die ersten N normierten Eigenfunktionen des Operators

$$Lu := -(pu')' + qu \,, \quad u(a) = 0, \; u'(b) = 0$$

   sind, d.h.

$$L\varphi_j = \lambda_j \varphi_j \,; \quad 0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N.$$

3) Stückweise Polynome: $X_h = \{\varphi_h \in C^0([a,b]) \,|\, \varphi_h(a) = 0, \varphi_h \restriction_{[x_j, x_{j+1}]} \in \mathbb{P}_k\}$ wobei $a = x_0 < x_1 < \ldots < x_n = b$ eine Zerlegung von $[a,b]$ ist.
   $\Longrightarrow$ Finite Elemente!

1.5.3 Finite Elemente Verfahren

Definition 1.117 (Finite Elemente Verfahren)
1) Sei $I := [a,b]$ *und* $I_h := \{x_0, \dots, x_n\} \subset I$ *ein Gitter mit* $x_0 = a$, $x_n = b$, $x_{j+1} = x_j + h_j$, $h_j > 0$. *Definiere* $I_j := (x_j, x_{j+1})$ *und* $h := \max_{j=0,\dots,n-1} h_j$. *Für festes* $k \in \mathbb{N}$ *wählen wir*

$$X_h := \{\varphi_h \in C^0([a,b]) \,|\, \varphi_h(a) = 0, \varphi_h \restriction_{I_j} \in \mathbb{P}_k \; \forall j = 0, \dots, n-1\}.$$

Ein $u_h \in X$ *heißt Finite Elemente Approximation von (SLP), falls für alle* $\varphi_h \in X_h$ *gilt:*

$$B(u_h, \varphi_h) = f(\varphi_h) \quad \forall \varphi_h \in X_h.$$

Dabei sind B, f *wie in Bemerkung 1.115 definiert.*

2) (Finite Elemente Basis für $k = 1$*)*
Sei $k = 1$. *Für* $j = 1, \dots, n$ *sei* $\varphi_j \in X_h$ *definiert durch*

$$\varphi_j(x_i) = \delta_{ij} \quad \forall x_i \in I_h.$$

D.h. für $j = 1, \dots, n$ *ist*

$$\varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_{j-1}} & x \in I_{j-1} \\ \frac{x_{j+1} - x}{h_j} & x \in I_j \\ 0 & sonst \end{cases}.$$

Es ist $supp(\varphi_j) = I_j \cup I_{j-1}$ *und somit folgt für die Steifigkeitsmatrix* $S_{kj} := B(\varphi_j, \varphi_k) = 0$, *falls* $|j - k| \geq 2$.

3) (Fehlerabschätzung für $k = 1$*)*
Sei $k = 1$. *Definiere Interpolierende* $\tilde{u}_h \in X_h$ *zu* $y \in X$ *durch*

$$\tilde{u}_h(x_j) := y(x_j) \quad \forall j = 0, \dots, n \implies \tilde{u}_h(x) = \sum_{j=1}^{n} y(x_j)\varphi_j(x).$$

Dann gilt die Fehlerabschätzung (nach Satz 1.114)

$$\|y - u_h\|_X \leq \frac{C_1}{C_0} \inf_{v_h \in X_h} \|y - v_h\|_X \leq \frac{C_1}{C_0} \|y - \tilde{u}_h\|_X .$$

D.h. der Fehler der Finite Elemente Approximation ist durch den Interpolationsfehler abgeschätzt.

Satz 1.118 (Interpolationsfehler auf dem Einheitselement)
Für $v \in H^1(0,1)$ sei $\tilde{v}_h \in \mathbb{P}_1$ die lineare Interpolierende zu v, d.h. $\tilde{v}_h \in \mathbb{P}_1$, $\tilde{v}_h(0) = v(0)$, $\tilde{v}_h(1) = v(1)$. Dann gilt:

1) $\|v - \tilde{v}_h\|_{L^2(0,1)} \leq \sqrt{c_p} \|v'\|_{L^2(0,1)}$,

2) $\|\tilde{v}_h'\|_{L^2(0,1)} \leq \|v'\|_{L^2(0,1)}$.

Ist zusätzlich $v \in H^2(0,1)$, so ist

3) $\|v - \tilde{v}_h\|_{L^2(0,1)} \leq c_p \|v''\|_{L^2(0,1)}$,

4) $\|(v' - \tilde{v}_h')\|_{L^2(0,1)} \leq \sqrt{c_p} \|v''\|_{L^2(0,1)}$.

Beweis: Analog zu Hilfssatz 1.103 (Poincaré Ungleichung) zeigt man , dass für alle $w \in H^1(0,1)$ mit $w(0) = w(1) = 0$ gilt

$$(*) \quad \|w\|^2_{L^2(0,1)} \leq c_p \left\|w'\right\|^2_{L^2(0,1)},$$

wobei $c_p \leq \frac{1}{2}$ ist. Setze $w = v - \tilde{v}_h$. Dann folgt aus $(*)$ unter Verwendung von

$$(**) \quad \tilde{v}_h'(x) = v(1) - v(0) = \int_0^1 v'(s)ds :$$

$$\begin{array}{rcl} \|v - \tilde{v}_h\|^2_{L^2(0,1)} & \leq & c_p \|v - \tilde{v}_h)'\|^2_{L^2(0,1)} = c_p \int_0^1 (v'(x) - \tilde{v}_h'(x))^2 dx \\ & = & c_p \int_0^1 (v'(x))^2 - 2\tilde{v}_h'(x)v'(x) + (\tilde{v}_h'(x))^2 dx \\ & \overset{(**)}{=} & c_p \int_0^1 (v'(x))^2 dx - 2(v(1) - v(0))^2 + (v(1) - v(0))^2 \\ & = & c_p \int_0^1 (v'(x))^2 dx - (v(1) - v(0))^2 \\ & \leq & c_p \int_0^1 (v'(x))^2 dx \end{array}$$

Also haben wir 1) gezeigt.
2) folgt direkt aus $(**)$, da

$$\left\|\tilde{v}_h'\right\|_{L^2(0,1)} = |v(1) - v(0)| \leq \left\|v'\right\|_{L^2(0,1)}$$

3) folgt aus 1) und $(*)$, da $(*)$ auch für $w \in H^1(0,1)$ mit $\int_0^1 w(x)dx = 0$ gilt.

4) folgt analog zur Herleitung von Gleichung 2), wenn man beachtet, dass $\tilde{v}_h'' = 0$ gilt. □

Folgerung 1.119 (Interpolationsabschätzung)
Sei $y \in X$ und $\tilde{u}_h \in X_h$ die Interpolierende von y für $k = 1$ aus Definition 1.117, 3). Dann gilt, falls $y \in H^2(a,b)$:

1) $\|y - \tilde{u}_h\|_{L^2(a,b)} \leq c_p h^2 \|y''\|_{L^2(a,b)}$,

2) $\|(y - \tilde{u}_h)'\|_{L^2(a,b)} \leq \sqrt{c_p} h \|y''\|_{L^2(a,b)}$.

Beweis: (Folgt aus 1.118 mit Skalierungsargument)

zu 1): Es ist $\|y - \tilde{u}_h\|^2_{L^2(a,b)} = \sum_{j=0}^{n-1} \|y - \tilde{u}_h\|^2_{L^2(I_j)}$.

Durch die Transformation $x = F(\bar{x}) = h_j \bar{x} + x_j$ und $\bar{y}(\bar{x}) = y(F(\bar{x}))$ folgt

$$I_j = F((0,1))$$

und

$$\begin{array}{rcl} \tilde{u}_h|_{I_j} & = & y(x_j) + \frac{x - x_j}{h_j}(y(x_{j+1}) - y(x_j)) \\ & = & \bar{y}(0) + \bar{x}(\bar{y}(1) - \bar{y}(0)) =: \tilde{\bar{y}}_h(\bar{x}) \end{array}$$

Also gilt wegen $F'(\bar{x}) = h_j$ und $\bar{y}''(\bar{x}) = y''(F(\bar{x})) \cdot (F'(\bar{x}))^2$

$$\begin{aligned} \|y - \tilde{u}_h\|^2_{L^2(I_j)} &= h_j \, \|\bar{y} - \tilde{\bar{y}}_h\|^2_{L^2(0,1)} \\ &\overset{1.118\ 3)}{\leq} c_p^2 h_j \, \|\bar{y}''\|^2_{L^2(0,1)} \\ &\leq c_p^2 h_j^4 \, \|y''\|^2_{L^2(I_j)} \, . \\ \Longrightarrow \quad \|y - \tilde{u}_h\|^2_{L^2(a,b)} &\leq c_p^2 \sum_{j=0}^{n-1} h_j^4 \, \|\bar{y}''\|^2_{L^2(I_j)} \\ &\leq c_p^2 h^4 \, \|y''\|^2_{L^2(a,b)} \, . \end{aligned}$$

Also ist 1) gezeigt.

zu 2): (Analoge Vorgehensweise!)

$$\|(y - \tilde{u}_h)'\|^2_{L^2(a,b)} = \sum_{j=0}^{n-1} \|(y - \tilde{u}_h)'\|^2_{L^2(I_j)} \, .$$

$$\begin{aligned} \|(y - \tilde{u}_h)'\|^2_{L^2(I_j)} &= \tfrac{1}{h_j} \, \|(\bar{y} - \tilde{u}_h)\|^2_{L^2(0,1)} \\ &\overset{1.118\ 4)}{\leq} c_p \tfrac{1}{h_j} \, \|\bar{y}''\|^2_{L^2(0,1)} \\ &= c_p h_j^2 \, \|y''\|^2_{L^2(I_j)} \, . \\ \Longrightarrow \quad \|(y - \tilde{u}_h)'\|^2_{L^2(a,b)} &\leq c_p \sum_{j=0}^{n-1} h_j^2 \, \|y''\|^2_{L^2(I_j)} \\ &\leq c_p h^2 \, \|y''\|^2_{L^2(a,b)} \, . \end{aligned}$$ □

Satz 1.120 (Fehlerabschätzung für lineare Finite Elemente)
Sei $y \in X$ die schwache Lösung von (SLP) und es gelten die Voraussetzungen aus Definition 1.100. Sei zusätzlich $f \in L^2(a,b)$, $p, p', q \in L^\infty(a,b)$ und es gelte $y \in H^2(a,b)$.
Dann existiert eine Konstante $c > 0$, so dass für die stückweise lineare Finite Elemente Approximation $u_h \in X_h$ (mit $k = 1$) gilt:

$$\|y - u_h\|_X \leq c \, h \, \|f\|_{L^2(a,b)}$$

Beweis: Nach 1.117, 3) gilt $\|y - u_h\|_X \leq \frac{c_1}{c_0} \|y - \tilde{u}_h\|_X$, wobei $\tilde{u}_h$ die Interpolierende von y in X_h ist und

$$c_1 := \max\{\|p\|_\infty, \|q\|_\infty\}, \; c_0 := \frac{p_0}{1 + c_p}.$$

Dann folgt weiter mit Interpolationsfehlerabschätzung 1.119:

$$\begin{aligned} \|y - u_h\|_X &\leq \tfrac{c_1}{c_0} \left(\|y - \tilde{u}_h\|^2_{L^2(a,b)} + \|(y - \tilde{u}_h)'\|^2_{L^2(a,b)} \right)^{\frac{1}{2}} \\ &\overset{1.119}{\leq} \tfrac{c_1}{c_0} (c_p^2 h^4 + c_p h^2)^{1/2} \, \|y''\|_{L^2(a,b)} \\ &\overset{1.110}{\leq} \underbrace{\frac{c_1 c_2}{c_0} \sqrt{c_p} (c_p h^2 + 1)^{1/2}}_{\leq C \text{ (z.B. für } h \leq h_{\max})} h \, \|f\|_{L^2(a,b)} \, . \end{aligned}$$ □

Bemerkung 1.121

1) Man kann zeigen, dass mit den Voraussetzungen $p \in C^1(a,b),\ q, f \in C^0(a,b);\ p, p', q \in L^\infty(a,b)$ und $f \in L^2(a,b)$ folgt, dass $y \in H^2(a,b)$ ist.

2) Für $k > 1$ glit die Fehlerabschätzung:

$$\|y - u_h\|_X \leq c \cdot h^k \left\| y^{(k+1)} \right\|_{L^2(a,b)}$$

falls p, q, f und y regulär genug sind.

3) Weiteres zu Finiten Elementen findet man z.B. in den Büchern von Ciarlet [3] oder Braess [2].

Bemerkung 1.122 (Weitere Diskretisierungsverfahren für (SLP))

1) Finite Differenzenverfahren:
Idee: Ersetze Ableitungen durch Differenzenquotienten, z.B.

$$y' \approx \frac{y_j - y_{j-1}}{h},\ \ y'' \approx \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2}.$$

2) Schießverfahren:
Berechne mit ESV oder MSV Approximationen der Anfangswertprobleme für $y_0,\ Y$ aus Definition 1.98 und erhalte $y_{0,h},\ Y_h$.
Wie im kontinuierlichen Fall ist dann

$$y_h := y_{0,h} + Y_h s_h$$

wobei s_h Lösung des Gleichnungssystems

$$[B_a + B_b Y_h(b)] s_h = g - B_b y_{0,h}(b)$$

ist (vgl. Satz 1.99 mit Beweis).

Kapitel 2

Gradientenverfahren

Generalvereinbarung 2.1
In diesem Kapitel gelte stets

1) $A \in \mathbb{R}^{m\times m}$ mit $A = A^T$,

2) A sei positiv definit und

3) $b \in \mathbb{R}^m$.

Ziel: Löse das lineare Gleichungssystem $Ax = b$.

Dazu wollen wir das Problem zunächst in ein Minimierungsproblem überführen.

Definition 2.2 (Minimierungsaufgabe)
Sei $F(x) := \langle A^{-1}(Ax - b), Ax - b\rangle\,,\ x \in \mathbb{R}^m$.
x heißt Lösung der Minimierungsaufgabe (M), g.d.w.

$$(M) \qquad F(x) = \min_{y\in\mathbb{R}^m} F(y).$$

Lemma 2.3
A, b seineen mit den Eigenschaften der Generalvereinbarung 2.1 gegeben.
Dann sind äquivalent:

1) $x \in \mathbb{R}^m$ löst $Ax = b$,

2) $x \in \mathbb{R}^m$ löst (M).

Beweis: A positiv definit $\implies$ A^{-1} positiv definit. Also gilt $F(y) \geq 0\ \ \forall y \in \mathbb{R}^m$.

1) $\Rightarrow$ 2): Gilt $Ax = b$, so folgt $F(x) = \left\langle A^{-1}0, 0\right\rangle = 0$.
Folglich löst x die Minimierungsaufgabe (M).

2) $\Rightarrow$ 1): x löst (M) $\implies$ $\nabla F(x) = 0$ $\iff$ $2(Ax - b) = 0$ $\iff$ $Ax - b = 0$ $\implies$ 1) $\square$

Idee der Gradientenverfahren 2.4
Aus Lemma 2.3 folgt: $Ax = b \iff F(x) = 0.$

Idee: Sei $(z_n)_{n\in\mathbb{N}}$ eine Folge im $\mathbb{R}^m$ definiert durch

$$z_{n+1} := z_n + \alpha_n t_n \quad n = 1, 2, \ldots$$

mit Koeffizienten $\alpha_n \in \mathbb{R}$ und Richtungsvektoren $t_n \in \mathbb{R}^m \setminus \{0\}$.
Wähle α_n, t_n so, dass $F(z_n) \longrightarrow 0 \ (n \to \infty)$.

Ansatz für die Wahl von α_n:

$$\alpha_n := \beta_n \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle}$$

mit $\beta_n \in \mathbb{R}$ und $r_n := b - Az_n$ der "Residuenvektor".

Dann gilt:

$$(*) \left| \begin{array}{rcl} F(z_n) - F(z_{n+1}) &=& \left\langle A^{-1}r_n, r_n\right\rangle - \left\langle A^{-1}(r_n - A\alpha_n t_n), r_n - A\alpha_n t_n\right\rangle \\ &=& 2\alpha_n \langle t_n, r_n\rangle - \alpha_n^2 \langle At_n, t_n\rangle \\ &=& 2\beta_n \frac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle} - \beta_n^2 \frac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle} \\ &=& (2 - \beta_n)\beta_n \underbrace{\frac{\langle t_n, r_n\rangle^2}{\langle At_n, t_n\rangle}}_{\geq 0} \overset{0\leq\beta_n\leq 2}{\geq} 0. \end{array} \right.$$

$\Longrightarrow$ Für $0 \leq \beta_n \leq 2$ ist $F(z_n) \leq F(z_{n+1})$.
$\Longrightarrow$ $(F(z_n))_{n\in\mathbb{N}}$ ist monoton fallend.

Also folgt: $\lim\limits_{n\ to\infty} F(z_n) = \lambda \geq 0$ existiert.
$\Longrightarrow$ $(F(z_n) - F(z_{n-1}))_{n\in\mathbb{N}}$ ist Nullfolge.

Man nennt β_n den Relaxationsparameter der Gradientenverfahren.
Die Gleichung (*) zeigt:
Für $\beta_n = 1$ wird $F(z_n) - F(z_{n+1})$ maximal und F nimmt auf der Geraden $z_n + \alpha_n t_n$ ein Minimum an.

Definition 2.5 (Allgemeines Gradientenverfahren)
Seien $(\beta_n)_{n\in\mathbb{N}}$ *und* $(t_n)_{n\in\mathbb{N}}$ *gegeben mit* $\beta_n \in [0,2]$ *und* $t_n \in \mathbb{R}^m$. *Dann heißt die Folge* $(z_n)_{n\in\mathbb{N}}$, $z_n \in \mathbb{R}^m$ *Lösung des Gradientenverfahren mit Startwert* $z_1 \in \mathbb{R}^m$ *wenn gilt:*

$$r_1 = b - Az_1$$

und für $n = 1, 2, \ldots$ *gilt*

$$\alpha_n = \beta_n \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle},$$
$$z_{n+1} = z_n + \alpha_n t_n,$$
$$r_{n+1} = b - Az_{n+1} = r_n - \alpha_n A t_n.$$

Definition 2.6 (Konvergenz)
Ein Gradientenverfahren heißt konvergent, falls gilt

$$r_n \longrightarrow 0 \quad (n \to \infty)$$

Dies ist äquivalent zu

$$\begin{aligned} & z_n \longrightarrow A^{-1}b \quad (n \to \infty) \\ \iff \quad & z_n \longrightarrow x \quad\;\; (n \to \infty) \text{ mit } Ax = b. \end{aligned}$$

2.1 Eigentliches Gradientenverfahren

Definition 2.7 (Eigentliches Gradientenverfahren)
Das Gradientenverfahren 2.5 mit $t_n = r_n$ *und* $\beta_n = 1$ *heißt* eigentliches Gradientenverfahren. *Die Richtungsvektoren werden in Richtung des Gradienten von* F *gewählt:*

$$r_n = b - Az_n = -\frac{1}{2}\nabla F(z_n).$$

Satz 2.8 (Konvergenz)
Das eigentliche Gradientenverfahren 2.7 ist konvergent.

Beweis: Es gilt $\forall x \in \mathbb{R}^m : \langle Ax, x\rangle \overset{\text{Schwarzsche Ungleichung}}{\leq} \left(\sum_{i,j=1}^{m} a_{ij}^2\right)^{\frac{1}{2}} \langle x, x\rangle .$

Setze $k := \left(\sum_{i,j=1}^{m} a_{ij}^2\right)^{\frac{1}{2}}$. Dann gilt $\forall x \in \mathbb{R}^m \setminus \{0\}$:

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} \leq k.$$

Mit $(*)$ aus 2.4 folgt:

$$F(z_n) - F(z_{n+1}) = \frac{\langle r_n, r_n\rangle^2}{\langle Ar_n, r_n\rangle} \geq \frac{1}{k}\langle r_n, r_n\rangle .$$

Aus 2.4 wissen wir, dass $F(z_n) - F(z_{n+1}) \to 0$ für $(n \to \infty)$. Also folgt für $k \neq 0$: $r_n \to 0 \ \ (n \to \infty)$.

□

Bemerkung 2.9
Je zwei aufeinanderfolgende Residuenvektoren des eigentlichen Gradientenverfahrens stehen senkrecht aufeinander:

$$\begin{aligned} \langle r_n, r_{n+1} \rangle &= \langle r_n, r_n - \alpha_n A t_n \rangle \\ &= \left\langle r_n, r_n - \tfrac{\langle r_n, r_n \rangle}{\langle Ar_n, r_n \rangle} A r_n \right\rangle \\ &= \langle r_n, r_n \rangle - \langle r_n, r_n \rangle = 0. \end{aligned}$$

Definition 2.10 (Gradientenverfahren bezüglich der kanonischen ON-Basis des $\mathbb{R}^m$)
Wähle $\beta_n = 1$ und $t_{i+jm} = e_i$ für $i = 1, \ldots, m$; $j = 0, 1, 2, \ldots$, wobei $e_i \in \mathbb{R}^m$ der i-te Einheitsvektor ist. Dann folgt mit $n = i + jm$:

$$\begin{aligned} \alpha_n = \tfrac{\langle e_i, r_n \rangle}{\langle Ae_i, e_i \rangle} &= \tfrac{1}{a_{ii}} \langle e_i, b - Az_n \rangle \\ &= \tfrac{1}{a_{ii}} (b_i - \sum_{l=1}^{m} a_{il} z_{n,l}) \end{aligned}$$

$$\Longrightarrow \quad z_{n+1} = z_n + \frac{1}{a_{ii}} (b_i - \sum_{l=1}^{m} a_{il} z_{n,l}) e_i.$$

D.h. z_{n+1} und z_n unterscheiden sich nur in der i-ten Komponente.

Das ist das Einschrittverfahren (siehe Numerik I).

Satz 2.11 (Konvergenz)
Das Gradientenverfahren 2.10 ist konvergent.

Beweis: Setze $\gamma := \max\limits_{i=1,\ldots,m} a_{ii}$. Dann ist mit $(*)$ aus 2.4

$$F(z_n) - F(z_{n+1}) \overset{n=i+jm}{=} \frac{\langle e_i, r_n \rangle^2}{\langle Ae_i, e_i \rangle} \geq \frac{1}{\gamma} (r_{n,i})^2 \geq 0.$$

Da nach 2.4 $F(z_n) - F(z_{n+1}) \to 0$, folgt $r_{n,i} \longrightarrow 0 \ \forall i = 1, \ldots, m$ und somit $r_n \to 0 \ (n \to \infty)$. □

2.2 *Conjugate Direction* Verfahren (CD)

Definition 2.12 (A-orthogonal)
Ein System von k Vektoren $q_1, \ldots, q_k \in \mathbb{R}^m$ $(k \leq m)$ *heißt* A-orthogonal *oder* A-konjugiert, *wenn gilt:*

$$\langle Aq_i, q_j \rangle = 0 \;\; i \neq j; \;\; i, j = 1, \ldots, k,$$

$$\langle Aq_i, q_i \rangle \neq 0 \;\; \forall i = 1, \ldots k.$$

Für $k = m$ *bilden A-orthogonale Systeme eine Basis des* $\mathbb{R}^m$. *Wir setzen* $\langle x, y \rangle_A = \langle Ax, y \rangle$, $\|x\|_A = \sqrt{\langle x, x \rangle_A} = \sqrt{\langle Ax, x \rangle}$.

Definition 2.13 (cd-Verfahren) "conjugate direction"
Sei $\{q_i\}_{i=1}^m$ *A-orthogonal. Wähle im allgemeinen Gradientenverfahren 2.5:*

$$t_n = q_n \text{ und } \beta_n = 1 \text{ für } n = 1, \ldots, m.$$

Satz 2.14 (Konvergenz der cd-Methode)
Das cd-Verfahren ist für beliebige Startvektoren $z_1 \in \mathbb{R}^m$ ein endliches Verfahren. Es gilt:

$$r_{m+1} = 0$$

und somit

$$z_{m+1} = A^{-1}b = x.$$

Beweis: Laut Gradientenverfahren folgt induktiv für $1 \leq n \leq m+1$:

$$r_n = r_{n-1} - \alpha_{n-1}At_{n-1} = r_{n-2} - \alpha_{n-2}At_{n-2} - \alpha_{n-1}At_{n-1} = \ldots = r_{n-l} - \sum_{i=n-l}^{n-1} \alpha_i At_i.$$

Für $l = n - j$, $1 \leq j \leq n$ folgt

$$r_n = r_j - \sum_{i=j}^{n-1} \alpha_i Aq_i.$$

$$\Longrightarrow \quad \langle q_j, r_n \rangle \overset{\text{A-orthogonal}}{=} \langle q_j, r_j \rangle - \alpha_j \langle Aq_j, q_j \rangle$$
$$\overset{\text{Def. von } a_j}{=} \langle q_j, r_j \rangle - \langle q_j, r_j \rangle = 0.$$

$\Longrightarrow$ $\langle q_j, r_{m+1} \rangle = 0 \;\; \forall j = 1, \ldots, m.$
$\Longrightarrow$ $r_{m+1} = 0$, da $\{q_i\}_{i=1}^m$ Basis des $\mathbb{R}^m$. $\square$

2.3 *Conjugate Gradient* Verfahren (CG)

Definition 2.15 (cg-Verfahren) "conjugate gradient"
Wähle $\beta_n = 1 \ \forall n = 1, \ldots, m+1, \ t_1 = r_1$ *und*

$$t_n = r_n + \gamma_{n-1} t_{n-1}$$

mit $\gamma_{n-1} := -\frac{\langle Ar_n, t_{n-1}\rangle}{\langle At_{n-1}, t_{n-1}\rangle}$ *für* $n = 2, 3, \ldots$. *Idee: Das A-orthogonale System wird mit Hilfe des Schmitd'schen Orthogonalisierungsverfahrens bzgl.* $\langle \cdot, \cdot \rangle_A$ *schrittweise aufgebaut.*

Satz 2.16 (Konvergenz des cg-Verfahrens)
Sei $z_1 \in \mathbb{R}^m$ und $t_1 = r_1 = b - Az_1$. Dann lautet das cg-Verfahrenfür $n = 1, \ldots, l$ mit $l \leq m$:

$$\left|\begin{array}{l} a_n = \frac{\langle t_n, r_n\rangle}{\langle At_n, t_n\rangle} \\ z_{n+1} = z_n + \alpha_n t_n \\ r_{n+1} = b - Az_{n+1} \\ \gamma_n = -\frac{\langle Ar_{n+1}, t_n\rangle}{\langle At_n, t_n\rangle} \\ t_{n+1} = r_{n+1} + \gamma_n t_n \end{array}\right.$$

Es erfüllt die Gleichungen
a) $\langle At_i, t_j\rangle = 0 \ \ 1 \leq j \leq i-1$

b) $\langle r_i, r_j\rangle = 0 \ \ 1 \leq j \leq i-1$

c) $\langle t_i, r_j\rangle = \langle r_j, r_j\rangle \ \ \ 1 \leq j \leq i$

und $l \leq m$ ist so gewählt, dass gilt $r_{l+1} = 0$.

Beweis: Folgt aus linearer Algebra und Satz 2.14.

Bemerkung 2.17
Für Gleichungssysteme resultierend aus Diskretisierungsverfahren, wie z.B. der Finit Elemente Methode, ist m so groß, dass man in der Praxis weniger als m Schritte iterieren wird. Dass dies Sinn mach zeigt die folgende Fehlerabschätzung.

Satz 2.18
Sei $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ die Kondition von $A \in \mathbb{R}^{n\times n}$. Dann gilt für das cg-Verfahren für $Ax = b$:

$$\|z_n - x\|_A \leq 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^n \|z_1 - x\|_A$$

wobei $\|y\|_A := \sqrt{\langle Ay, y\rangle}$ definiert ist.
(ohne Beweis)

Bemerkung 2.19

1) Vorkonditionierung: Anstelle von $Ax = b$ löst man $C^{-1}A(C^{-1})^T y = C^{-1}b$, wobei $y = C^T x$ ist und C auch folgende Anforderungen erfüllt

   1) $C \in \mathbb{R}^{n\times n}$ regulär.
   2) Gleichungsysteme mit C sollen einfach zu lösen sein.
   3) $\kappa(C^{-1}A(C^{-1})^T$ sollt möglich nahe an 1 liegen.

2) Ist A nicht symmetrisch und positiv definit, so kann man z.B. $A^T Ax = A^T b$ anstelle von $Ax = b$ lösen, denn $A^T A$ ist positiv definit und symmetrisch.
   Dieser Ansatz führt auf das bicg-Verfahren.

3) Für eine gegbene Toleranz TOL, kann

$$\langle r_n, r_n \rangle \leq TOL$$

   als Abbruchbedingung verwendet werden.

Kapitel 3

Eigenwertprobleme

Definition 3.1 (Eigenwertproblem)
Sei $A \in \mathbb{K}^{n\times n}$ mit $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Dann heißt $\lambda \in \mathbb{K}$ Eigenwert von A, wenn ein Eigenvektor $x \in \mathbb{K}^n$ mit $x \neq 0$ existiert mit der Eigenschaft $Ax = \lambda x$.

Vollständiges Eigenwertproblem: Finde alle EW (und EV) von A

Partielles Eigenwertproblem: Finde einzelne EW (und EV) von A (z.B. den größten und kleinsten EW)

3.1 Grundbegriffe der linearen Algebra und theoretische Grundlagen

Notation und Grundlagen 3.2

1) Für $x, y \in \mathbb{K}^n$ bezeichne $\langle x, y\rangle$ und $\|x\|$ das euklidische Skalarprodukt und die euklidische Norm.

2) $\|A\|$ bezeichnet die Spektralnorm von A. Bemerke: $\rho(A) \leq \|A\|_*$ für alle Normen $\|\cdot\|_*$.

3) Für $\mathbb{K} = \mathbb{C}$ ist $A^H := \bar{A}^T$. A heißt hermitesch, falls $A^H = A$.

4) Die EWe von A sind die Nullstellen des charakteristischen Polynoms

$$\rho_A(\lambda) := \det(A - \lambda I).$$

5) Ist ein EW λ bekannt, so findet man alle EVen zu λ durch Lösung des singulären homogenen Gleichungssystems

$$(A - \lambda I)x = 0$$

Umgekehrt bestimmt ein EV $x \neq 0$ den zugehörigen EW λ durch den Rayleigh Quotienten

$$R(x) = \frac{\langle Ax, x\rangle}{\|x\|^2},$$

d.h. $\lambda = R(x)$.

6) $\sigma(A) := \{\lambda \,|\, \lambda \text{ EW von } A\}$ heißt Spektrum von A.

7) Das charakteristische Polynom besitzt die Darstellung

$$\rho_A(x) = \prod_{i=1}^{s} (x - \lambda_i)^{\sigma_i},$$

wobei λ_i die paarweise verschiedenen Eigenwerte von A sind. Es gilt $\sum_{i=1}^{s} \sigma_i = n$ und σ_i heißt algebraische Vielfachheit von λ_i.
Die Eigenvektoren zu λ_i (Vereinigt mit dem Nullvektor) bilden den sogenannten Eigenraum $E_i := \text{Kern}(A - \lambda_i I)$. Ist $\rho_i := \dim(E_i)$, so heißt ρ_i geometrische Vielfachheit von λ_i.

Ähnlichkeitstransformationen 3.3
Ist $T \in \mathbb{K}^{n\times n}$ regulär, so heißte $B := T^{-1}AT$ Ähnlichkeitstransformation und B ähnlich zu A.

1) Ähnliche Matrizen besitzen dieselben EWe λ, denn mit $y := T^{-1}x$ folgt:

$$T^{-1}ATy = T^{-1}Ax = \lambda T^{-1}x = \lambda y.$$

D.h. λ ist EW zu B mit EV $y = T^{-1}x$.

2) Ähnliche Matrizen besitzen dasselbe char. Polynom, denn

$$\begin{aligned} \det(T^{-1}AT - \lambda I) &= \det(T^{-1}(A - \lambda I)T) \\ &= \det(T^{-1})\det(A - \lambda I)\det(T) \\ &= \det(A - \lambda I). \end{aligned}$$

Idee einer numerischen Methode 3.4
Wende eine Folge von Ähnlichkeitstransformationenen an, um A in einfachen Gestalt zu transformieren, d.h.

$$A^{(0)} := A,$$
$$A^{(i)} := T_i^{-1}A^{(i-1)}T_i, \quad i = 1, 2, 3, \ldots$$

und geeignete Matrizen T_i.

Satz 3.5 (Gerschgorinscher Kreissatz)
Sei $A \in \mathbb{C}^{n\times n}$. Für $i = 1, \ldots, n$ definiere die sogenannten Gerschgorin Kreise

$$G_i := \{z \in \mathbb{C} \mid |z - a_{ii}| \le r_i\}, \quad r_i := \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Dann gilt:

(i) Ist λ EW von A, so ist $\lambda \in \bigcup_{i=1}^{n} G_i$, d.h. $\sigma(A) \subset \bigcup_{i=1}^{n} G_i$.

(ii) Hat die Vereinigung $\hat{G}$ von m Kreisen G_i einen leeren Schnitt mit den restlichen $n-m$ Kreisen, so enthält $\hat{G}$ genau m EWe von A (gezählt mit ihren algebraischen Vielfachheiten).

Beweis:

zu (i): Sei λ EW von A mit EV $x \neq 0$.
Aus $Ax = \lambda x$ folgt:

$$(\lambda - a_{ii})x_i = \sum_{j=1, j\neq i}^{n} a_{ij}x_j \quad \forall i = 1, \ldots, n.$$

Für $i \in \{1, \ldots, n\}$mit $|x_i| = \max_{i=1,\ldots,n} |x_j|$ folgt:

$$|\lambda - a_{ii}| = \left| \sum_{j=1, j\neq i}^{n} \frac{a_{ij}x_j}{x_i} \right| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

$\Longrightarrow \quad \lambda \in G_i \subset \bigcup_{j=1}^{n} G_j.$

zu (ii): Wir setzen $D = (a_{ii}\delta_{ij})_{i,j=1,\ldots,m}$ und betrachten

$$B(t) := D + t(A - D), \quad 0 \leq t \leq 1$$

mit Gerschgorin Kreisen

$$G_i(t) := \left\{ z \in \mathbb{C} \,|\, |z - a_{ii}| \leq t \cdot \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\} \quad i = 1, \ldots, n.$$

Es ist $B(0) = D$ und $B(1) = A$. Die EW von $B(t)$ sind die Nullstellen von $\rho_{B(t)}$ und hängen daher stetig von t ab. Wende (i) auf $B(t)$ an und lasse t von 0 nach 1 laufen.
Dabei wird der Radius der Kreise bei festem Mittelpunkt immer größer.
Die Anzahl der EWe in einem Kreis $G_i(t)$ kann sich erst dann ändern, wenn dieser einen anderen Kreis trifft. Daraus folgt die Beh. □

Folgerung 3.6
Da A und A^H dieseleben Eigenwerte besitzen, gilt der Satz 3.5 auch mit

$$G_i' := \{z \in \mathbb{C} \,|\, |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ji}|$$

$$\Longrightarrow \quad \sigma(A) \subset \left(\bigcup_{i=1}^{n} G_i \right) \cap \left(\bigcup_{i=1}^{n} G_i' \right).$$

Satz 3.7
Sei $A \in \mathbb{C}^{n\times n}$ und $D = (a_{ii}\delta_{ij})_{i,j}$ die Diagonalmatrix von A. Dann gilt für die EWe von A

$$1 \leq \left\| (\lambda I - D)^{-1}(A - D) \right\|,$$

falls keines der Diagonalelemente a_{ii} ein EW von A ist.

Beweis: Sei $Ax = \lambda x, \; x \neq 0 \quad \Longrightarrow \quad \lambda x - Dx = (A - D)x$

$\Longrightarrow \quad x = (\lambda I - D)^{-1}(A - D)x.$

Also ist 1 ein EW von $(\lambda I - D)^{-1}(A - D)$, d.h.

$$1 \leq \left\|(\lambda I - D)^{-1}(A - D)\right\|. \qquad \square$$

Folgerung 3.8
Da die Spektralnorm kleiner als z.B. die Zeilensummennorm ist, folgt:

$$1 \leq \max_{i=1,\ldots,n} \sum_{k \neq j} \left| \frac{a_{jk}}{\lambda - a_{ii}} \right|$$

$\Longrightarrow \quad \exists j$ mit $|\lambda - a_{jj}| = \sum_{k \neq j} |a_{jk}| \quad \Longrightarrow \quad \lambda \in G_j \subset \bigcup_{i=1}^{n} G_i$

Satz 3.9 (Kondition des Eigenwertproblems)
Sei $B \in \mathbb{C}^{n \times N}$ diagonalisierbar, d.h. es ex eine reguläre Matrix $P \in \mathbb{C}^{n \times n}$ mit

$$P^{-1}BP = D = diag(\lambda_1(B), \ldots, \lambda_n(B)).$$

Dann gibt es zu jedem EW $\lambda_j(B)$ einen EW $\lambda_j(A)$ von $A \in \mathbb{C}^{n \times n}$ mit

$$|\lambda_j(A) - \lambda_j(B)| \leq \kappa(P) \left\|A - B\right\|.$$

Dabei ist $\kappa(P) = \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}$ die Kondition von P.

Beweis:

$$\begin{aligned} \left\|(\lambda I - B)^{-1}\right\| &= \left\|P(\lambda I - D)^{-1}P^{-1}\right\| \leq \left\|(\lambda I - D)^{-1}\right\| \kappa(P) \\ &\leq \max_j \frac{1}{|\lambda - \lambda_j(B)|} \kappa(P) \\ &= \frac{1}{\min_j |\lambda - \lambda_j(B)|} \kappa(P). \end{aligned}$$

Analog zu Satz 3.7 folgt:

$$1 \leq \left\|(\lambda I - B)^{-1}\right\| \left\|A - B\right\|.$$

Also folgt:

$$1 \leq \frac{1}{\min_j |\lambda - \lambda_j(B)|} \kappa(P) \left\|A - B\right\|$$

$\Longrightarrow \quad \min_j |\lambda - \lambda_j(B)| \leq \kappa(P) \left\|A - B\right\|$
$\Longrightarrow \quad \exists \lambda = \lambda_j(A)$ mit $|\lambda_j(A) - \lambda_j(B)| \leq \kappa(P) \left\|A - B\right\|$ $\square$

Beispiel 3.10 (Anwendung der Gerschgorin-Kreise)

$$A = \begin{pmatrix} 0,9 & 0 & \\ 0 & 0,4 & 0 \\ 0 & 0 & 0,2 \end{pmatrix} + 10^{-5} \begin{pmatrix} 0,1 & 0,4 & -0,2 \\ -0,1 & 0,5 & 0,1 \\ 0,2 & 0,1 & 0,3 \end{pmatrix}$$

Nach Satz 3.9 erwartet man, dass die EW einer Matrix A mit kleinen Außendiagonalelementen

ungefähr mit dem Diagonalelementen übereinstimmen.
Die 3 Gerschgorinkreise sind disjunkt, daher besitzt A die EW $\lambda_1, \lambda_2, \lambda_3$ mit

$$\begin{vmatrix} \lambda_1 - (0,9 + 0,1 \cdot 10^{-5}) \\ \lambda_2 - (0,4 + 0,5 \cdot 10^{-5}) \\ \lambda_3 - (0,2 + 0,3 \cdot 10^{-5}) \end{vmatrix} \leq \begin{matrix} 0,6 \cdot 10^{-5}, \\ 0,2 \cdot 10^{-5}, \\ 0,3 \cdot 10^{-5}. \end{matrix}$$

Diese Abschätzungen können noch wesentlich verbessert werden.
Sei $P := diag(10^5, 1, 1)$. Dann ist

$$P^{-1}AP = \begin{pmatrix} 0,9 & 0 & 0 \\ 0 & 0,4 & 0 \\ 0 & 0 & 0,2 \end{pmatrix} + \begin{pmatrix} 0,1 \cdot 10^{-5} & 0,1 \cdot 10^{-10} & -0,2 \cdot 10^{-10} \\ -0,1 & 0,5 \cdot 10^{-5} & 0,1 \cdot 10^{-5} \\ 0,2 & 0,1 \cdot 10^{-5} & 0,3 \cdot 10^{-5} \end{pmatrix}.$$

Der erste Gerschgorin Kreis ist noch disjunkt zu den beiden anderen, die nicht mehr disjunkt sind.
Also Folgt für λ_1:

$$\left|\lambda_1 - (0,9 + 0,1 \cdot 10^{-5})\right| \leq 0,6 \cdot 10^{-10}.$$

Entsprechend kann die Abschätzung für λ_2, λ_3 verbessert werden (siehe Abbildung 3.1).

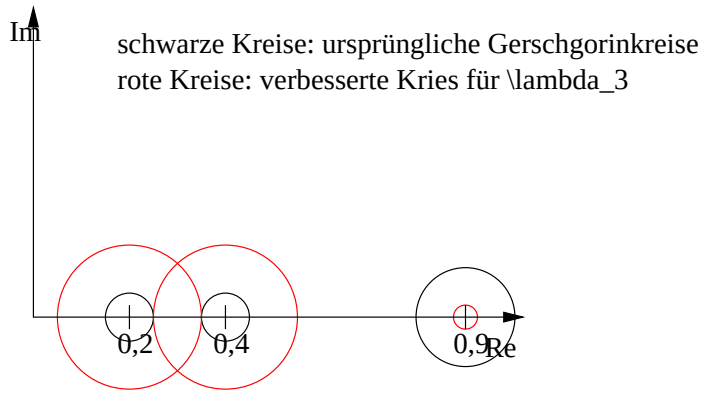

Abbildung 3.1: Gerschgorinkreise: Beispiel, Radien nicht maßstabsgetreu!!

Satz 3.11 (Bauer-Fike)
Sein $A \in \mathbb{C}^{n\times n}$ digonalisierbar, d.h. es existiert eine reguläre Matrix $P \in \mathbb{C}^{n\times n}$ mit $P^{-1}AP = D := diag(\lambda_1, \ldots, \lambda_n)$. Ferner sei $A + E \in \mathbb{C}^{n\times n}$ eine Störung von A und λ EW von $A + E$. Dann gilt:

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \left\|P^{-1}EP\right\| \leq \kappa(P) \left\|E\right\|.$$

Beweis: Ist λ auch EW von A, so ist die Beh. richtig.

Sei $\lambda \neq \lambda_j$ für $j = 1, \ldots, n$. Sei x EV zu λ, d.h. $(A + E)x = \lambda x$ mit $x \neq 0$.
$\Longrightarrow$ $Ex = \lambda x - Ax = (\lambda I - A)x = (\lambda I - PDP^{-1})x = P(\lambda I - D)P^{-1}x$.
$\Longrightarrow$ $P^{-1}x = (\lambda I - D)^{-1}P^{-1}Ex = (\lambda I - D)^{-1}P^{-1}EP(P^{-1}x)$.
$\Longrightarrow$ $\left\|P^{-1}x\right\| \leq \frac{1}{\min_{j=1,\ldots,n} |\lambda - \lambda_j|} \left\|P^{-1}EP\right\| \left\|P^{-1}x\right\|$.
$\Longrightarrow$ $\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \left\|P^{-1}EP\right\| \leq \kappa(P) \left\|E\right\|$. $\square$

Bemerkung 3.12
Nach dem Satz von Bauer-Fike existiert also ein EW λ_i von A mit

$$|\lambda - \lambda_i| \leq \kappa(P)\,\|E\|\,.$$

Die Kondition von P bestimmt also die Störanfälligkeit der EW von A.

Satz 3.13 (Satz von Schur)
Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ existiert eine unitäre Matrix $U \in \mathbb{C}^{n\times n}$ (d.h. $U^H U = I$) mit

$$U^H A U = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Beweis: (Induktion über n)
Ind. Anf: $n = 1$ $\surd$
Ind. Vor: Für $A \in \mathbb{C}^{(n-1)\times(n-1)}$ gilt die Behauptung.
Ind. Beh: Sei $A \in \mathbb{C}^{n\times n}$. Zeige: $\exists U$ unitär mit $U^H A U = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$.
Sei $\lambda \in \mathbb{C}$ EW von A und $z \in \mathbb{C}^n \setminus \{0\}$ ein EV mit $\|z\| = 1$.
Wir können z zu einer Orthonormalbasis auf $\mathbb{C}^n$ ergänzen, d.h. wir können eine unitäre Matrix $\hat{V}$ finden, so dass

$$V = (z, \hat{V}) \text{ mit } V^H V = I.$$

Sei $B = V^H A V$. Wegen $VBe_1 = AVe_1 = Az = \lambda z = \lambda V e_1$
$\Longrightarrow$ $Be_1 = \lambda e_1$, d.h. die erste Spalte von B ist ein λ-faches des ersten Einheitsvektors.
Also $B = V^H A V = \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & C \end{array}\right)$.
Nach Ind. Vor. ex. eine unitäre Matrix W mit $W^H C W = T$ und T hat obere Dreiecksgestalt.
$\Longrightarrow$ $A = V \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & W^H C W \end{array}\right) V^H = \underbrace{V \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W \end{array}\right)}_{=:U} \left(\begin{array}{c|c} \lambda & b \\ \hline 0 & T \end{array}\right) \underbrace{\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & W^H \end{array}\right) V^H}_{:=U^H}$.
Daraus folgt die Behauptung. □

Folgerung 3.14 (Schur für hermitesche Matrix)
Sei $A \in \mathbb{C}^{n\times n}$ hermitesch. Dann ex. eine unitäre Matrix $U = (u_1, \ldots, u_n)$ mit

$$U^H A U = diag(\lambda_1, \ldots, \lambda_n).$$

$\lambda_1, \ldots, \lambda_n$ sind EW von A, die reell sind und u_i die EVen zu λ_i.
Insbesondere ist A Matrix von n linear unabhängigen zu einander orthogonalen EVen.

Beweis: A hermitesch $\Longrightarrow$ $A^H = A$ und somit

$$(U^H A U)^H = U^H A (U^H)^H = U^H A U.$$

$\Longrightarrow$ $U^H AU$ ist selbst wieder hermitesch.
$\Longrightarrow$ Beh. mit 3.13. □

Folgerung 3.15 (Bauer-Fike für hermitesche Matrizen)
Ist $A \in \mathbb{C}^{n\times n}$ hermitesch und $A+E$ eine Störung von A und sind $\lambda_1, \ldots, \lambda_n$ EW von A, so gilt für einen EW von $A+E$:

$$\min_{j=1,\ldots,n} |\lambda - \lambda_j| \leq \|E\| .$$

Beweis: Nach 3.14 existiert eine unitäre Matrix P mit $P^{-1}AP = diag(\lambda_1, \ldots, \lambda_n)$.
Da P unitär ist, folgt $\kappa(P) = 1$ und somit folgt die Beh. aus dem Satz von Bauer Fike, 3.11. □

3.2 Variationsprinzip für Eigenwerte hermitescher Matrizen

Satz 3.16 (Rayleighsches Maximumsprinzip)
Sei $A \in \mathbb{K}^{n\times n}$ hermitesch. Die EW von A seien $\lambda_1 \geq \ldots \geq \lambda_n$. Sei $U = (u_1, \ldots, u_n)$ unitäre Matrix mit $U^H AU = diag(\lambda_1, \ldots, \lambda_n) =: \Lambda$.
Für $j = 1, \ldots, n$ definiere den $(n+1-j)$-dimensionalen Teilraum

$$M_j := \{x \in \mathbb{K}^n \mid \langle u_i, x\rangle = 0 \, \forall i = 1, \ldots, j-1\} = (\text{span}(u_1, \ldots, u_n))^{\perp}.$$

Dann gilt:

$$\lambda_j = \max_{x \in M_j \setminus \{0\}} R(x)$$

mit dem Rayleigh-Quotienten $R(x) := \frac{\langle Ax, x\rangle}{\langle x, x\rangle}$.

Beweis: Sei $j \in \{1, \ldots, n\}$ und $x \in M_j \setminus \{0\}$ beliebig.
setze $y := U^H x$. Dann folgt:

$$y_i = \langle u_i, x\rangle = 0, \quad i = 1, \ldots, j-1.$$

Wegen $\lambda_i \leq \lambda_j$ für $i = j, \ldots, n$ gilt:

$$R(x) = \frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\langle U\Lambda U^H x, x\rangle}{\langle x, x\rangle} = \frac{\langle \Lambda U^H x, U^H x\rangle}{\langle x, x\rangle} = \frac{\langle \Lambda y, y\rangle}{\langle y, y\rangle} = \frac{\sum_{i=1}^n \lambda_i |y_i|^2}{\sum_{i=1}^n |y_i|^2} \leq \lambda_j$$

und somit

$$\sup_{x \in M_j \setminus \{0\}} R(x) \leq \lambda_j.$$

Andererseits ist $u_j \in M_j \setminus \{0\}$ und $R(u_j) = \lambda_j$.
$\Longrightarrow$ $\max_{x \in M_j \setminus \{0\}} R(x) = \lambda_j$. $\square$

Bemerkung 3.16a
Definiert man $f : \mathbb{R} \longrightarrow \mathbb{R}$ bei festem $x \in \mathbb{K}^n \setminus \{0\}$ durch

$$f(\lambda) = \frac{1}{2} \|Ax - \lambda x\|^2 = \frac{1}{2}\lambda^2 \|x\|^2 - \lambda \langle Ax, x\rangle + \frac{1}{2} \|Ax\|^2$$

so nimmt f in $\lambda = R(x)$ sein Minimum an.
Ist daher x nähreungsweise ein EV von A, so ist $R(x)$ eine gute Näherung des zugehörigen Eigenwertes.

Satz 3.17 (Courantsches Minimum-Maximum Prinzip)
Sei $A \in \mathbb{K}^{n\times n}$ hermitesch mit EWen $\lambda_1 \geq \ldots \geq \lambda_n$. Für $j = 1, \ldots, n$ definiere

$$\mathcal{N}_j := \{N_j \subset \mathbb{K}^n \mid N_j \text{ ist linearer Teilraum der Dimension } n+1-j\}$$

Dann gilt:

$$\lambda_j = \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R(x), \quad j = 1, \ldots, n.$$

Beweis: Sei $U = (u_1, \ldots, u_n)$ die unitäre Matrix von EVen zu den EWen $\lambda_1, \ldots, \lambda_n$ von A. Sei

$j \in \{1, \ldots, n\}$. Definiere $L_j = \text{span}(u_1, \ldots, u_j)$ und wähle $N_j \in \mathcal{N}_j$ beliebig.
Wegen

$$\begin{aligned} \dim(L_j \cap N_j) &= \dim(L_j) + \dim(N_j) - \dim(L_j \cup N_j) \\ &= n + 1 - \underbrace{\dim(L_j \cup N_j)}_{\leq n} \geq 1 \end{aligned}$$

existiert ein $x \in L_j \cap N_j$ mit $x \neq 0$.
Da $x \in L_j$, folgt: $x = \sum_{i=1}^{j} \alpha_i u_i$.
$\Longrightarrow \quad R(x) = \frac{\sum_{i=1}^{j} \lambda_i |\alpha_i|^2}{\sum_{i=1}^{j} |\alpha_i|^2} \geq \lambda_j$, da $\lambda_i \geq \lambda_j$ für $i = 1, \ldots, j$

$$\Longrightarrow \quad \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R(x) \geq \lambda_j$$

Wählt man andererseits $N_j = M_j$, so gilt nach Satz 3.16

$$\max_{x \in M_j \setminus \{0\}} R(x) = \lambda_j \qquad \square$$

Folgerung 3.18
Seien $A, B \in \mathbb{K}^{n \times n}$ hermitesch und gelte $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$ sowie $\lambda_1(B) \geq \ldots \geq \lambda_n(B)$.
Dann gilt die Abschätzung:

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|, \quad j = 1, \ldots, n.$$

Beweis: A, B hermitesch $\Longrightarrow$ $E = A - B$ hermitesch. Sei $x \in \mathbb{K}^n \setminus \{0\}$.
Dann gilt:

$$R_E(x) = \frac{\langle (A - B)x, x \rangle}{\langle x, x \rangle} \leq \|A - B\|.$$

Somit folgt

$$(*) \qquad R_A(x) \leq R_B(x) + \|A - B\|.$$

Sei $j \in \{1, \ldots, n\}$ und $N_j \in \mathcal{N}_j$ (vgl. 3.17), so folgt aus $(*)$:

$$\min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R_A(x) \leq \min_{N_j \in \mathcal{N}_j} \max_{x \in N_j \setminus \{0\}} R_B(x) + \|A - B\|.$$

Mit dem Courantschen Min-Max Prinzip folgt $\lambda_j(A) \leq \lambda_j(B) + \|A - B\|$.
Vertauscht man die Rollen von A und B, so folgt auch

$$\lambda_j(B) \leq \lambda_j(A) + \|A - B\|$$

$$\Longrightarrow \quad |\lambda_j(B) - \lambda_j(A)| \leq \|A - B\|. \qquad \square$$

3.3 Transformation auf Hessenberg-Form

Definition 3.19 (Hessenberg-Matrizen)
Eine Matrix $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times n}$ *heißt eine (obere)* Hessenberg-Matrix, *wenn* $a_{ij} = 0$ *für* $1 \le j \le i-2$, *d.h.*

$$A = \begin{pmatrix} * & \dots & \dots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}$$

Eine Hessenberg-Matrix heißt unreduziert, falls sämtliche Subdiagonalelemente $a_{i+1,i}$, $i = 1,\dots,n-1$ *ungleich 0 sind.*
Eine symmetrische Hessenberg-Matrix ist somit eine symmetrische Tridiagonalmatrix.

Erinnerung an die Numerik I 3.20 (Householder-Matrix)
Ist $x \in \mathbb{R}^n\backslash\{0\}$ und definiert man $u := x + \text{sign}(x_1)\,\|x\|\,e_1 \in \mathbb{R}^n$, so ist durch $P = I - \frac{2uu^T}{\langle u,u\rangle} = I - \beta uu^T$ mit $\beta := \frac{2}{\langle u,u\rangle} = \frac{1}{\|x\|(\|x\|+|x_1|)}$ eine Householder Matrix definiert mit $Px = -sgn(x_1)\,\|x\|\,e_1$.

Satz 3.21 (Householder-Transformationen auf Hessenberg-Form)
Zu einer Matrix $A \in \mathbb{R}^{n\times n}$ existieren $n-2$ Hoseholder-Matrizen $P_1,\dots,P_{n-2}$, so dass

$$P_{n-2}\dots P_1 A P_1 \dots P_{n-2}$$

eine zu A orthogonal-ähnliche Hessenberg-Matrix ist.

Beweis: (Induktion über $k = 1,\dots,n-2$)

Angenommen es seien schon $k-1$ Householder-Matrizen $P_1,\dots,P_{k-1}$ bestimmt, so dass

$$P_{k-1}\dots P_1 A P_1 \dots P_{k-1} = \left(\begin{array}{c|c|c} H_k & B_k \\ \hline 0 \;|\; a_k & C_k \end{array}\right) \begin{matrix}\}k \\ \}n-k\end{matrix},$$

wobei $H \in \mathbb{R}^{k\times k}$ eine Hessenberg-Matrix, $B_k \in \mathbb{R}^{k\times k}, C \in \mathbb{R}^{(n-k)\times(n-k)}$ und $a_k \in \mathbb{R}^{n-k}$ ist.

Für P_k machen wir folgenden Ansatz:

$$P_k = \text{diag}(I_k, \tilde{P}_k) = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right)$$

mit der Identität $I_k \in \mathbb{R}^{k\times k}$ und einer $(n-k)\times(n-k)$ Householder-Matrix $\tilde{P}_k$. Dann folgt:

$$P_k P_{k-1}\dots P_1 A P_1 \dots P_{k-1} P_k = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right)\left(\begin{array}{c|c|c} H_k & B_k \\ \hline 0 \;|\; a_k & C_k \end{array}\right)\left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right)$$

$$= \left(\begin{array}{c|c} H_k & B_k\tilde{P}_k \\ \hline 0 \;|\; \tilde{P}_k a_k & \tilde{P}_k C_k \tilde{P}_k \end{array}\right) = \left(\begin{array}{c|c|c} H_{k+1} & B_{k+1} \\ \hline 0 \;|\; a_{k+1} & C_{k+1} \end{array}\right) \begin{matrix}\}k \\ \}n-k\end{matrix}$$

mit einer Hessenberg-Matrix $H_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$, wenn die Householder-Matrix $\tilde{P}_k$ so gewählt wird, dass $\tilde{P}_k a_k$ ein Vielfaches des ersten Einheitsvektors in $\mathbb{R}^{n-k}$ ist (siehe 3.20). □

Algorithmus 3.22 (Householder-Transformation auf Hessenberg-Form)
Input: $A \in \mathbb{R}^{n\times n}$
Für $k = 1, \ldots, n-2$

Falls $a_k = (a_{k+1,k}, \ldots, a_{n,k})^T \neq 0$

dann Berechne Householder Matrix $\tilde{P}_k$ durch
$u_k := (a_{k+1,k} + \mathtt{sign}(a_{k+1,k}) \|a_k\|, a_{k+2,k}, \ldots, a_{n,k})^T$
$\beta_k := \|a_k\|^{-1} (\|a_k\| + |a_{k+1,k}|)^{-1}$
$\tilde{P}_k := I_{n-k} - \beta_k u_k (u_k)^T$ und berechne $A := P_k A P_k$

sonst setze $P_k := I$

Output: Die Ausgangsmatrix A wird in $n-2$ Schritten mit orthogonal-ähnlichen Transformationen in eine Hessenberg-Matrix $P^T AP, P = P_1 \cdot \ldots \cdot P_{n-2}$ überführt.

Bemerkung:
Da orthogonale Ähnlichkeitstransformationen die Symmetrie erhalten, wird eine symmetrische Matrix A durch $n-2$ Schritte auf eine Tridiagonalmatrix transformiert.

3.4 Eigenwertbestimmung für Hessenberg-Matrizen

Grundlegende Idee 3.23
Bestimme das charakteristische Polynom $\rho_A(\lambda)$ einer Hessenberg-Matrix A, sowie die Ableitung $\rho'_A(\lambda)$, so dass man die Eigenwerte durch Nullstellensuchen, etwa mit dem Newtonverfahren, berechnen kann.

Satz 3.24 (Berechnung von $\rho_A(\lambda), \rho'_A(\lambda)$ für Tridiagonalmatrizen)
Sei $T \in \mathbb{R}^{n\times n}$ eine symmetrische Tridiagonalmatrix, d.h.

$$T = \begin{pmatrix} b_1 & c_1 & & & 0 \\ c_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ 0 & & & c_{n-1} & b_n \end{pmatrix}$$

und $\rho_A(\lambda) = det(T - \lambda I)$.
Für $k = n, \ldots, 0$ seien $f_k(\lambda), g_k(\lambda)$ definiert durch:

$$\begin{aligned} f_n(\lambda) &= 1, & g_n(\lambda) &= 0, \\ f_{n-1}(\lambda) &= b_1 - \lambda, & g_{n-1}(\lambda) &= -1, \\ f_{n-i-1}(\lambda) &= (b_{i+1} - \lambda) f_{n-i}(\lambda) - |c_i|^2 f_{n-i+1}(\lambda), \\ g_{n-i-1}(\lambda) &= -f_{n-i}(\lambda) + (b_{i+1} - \lambda) g_{n-i}(\lambda) - |c_i|^2 g_{n-i+1}(\lambda). \end{aligned}$$

Dann gilt:

$$\begin{aligned} f_0(\lambda) &= \rho_A(\lambda), \\ g_0(\lambda) &= \rho'_A(\lambda). \end{aligned}$$

Beweis: Wir zeigen durch vollst. Induktion über i, dass $f_{n-i}(\lambda)$ die Determinate des i-ten Hauptminors von $T - \lambda I$ ist:

Für $i = 0, 1$ ist dei Aussage richtig.
Sei die Aussage richtig für alle j mit $1 \le j \le i$, $2 \le i$.
Dann folgt mit Determinanten-Entwicklungssatz

$$\det \left(\begin{array}{ccccc|c|c} b_1 - \lambda & c_1 & & & & & 0 \\ c_1 & \ddots & \ddots & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & c_{i-1} & & \\ \hline & & & c_{i-1} & b_i - \lambda & c_i \\ \hline 0 & & & & c_i & b_{i+1} - \lambda \end{array} \right) =$$

$$\begin{aligned} &= (b_{i+1} - \lambda) f_{n-i}(\lambda) - c_i c_i f_{n-(i-1)}(\lambda) \\ &= f_{n-i-1}(\lambda). \end{aligned}$$

Da $f'_{n-i}(\lambda) = g_{n-i}(\lambda)$ für $i = 0, \ldots, n$, folgt die Behauptung. □

Bemerkung 3.25
Das Verfahren lässt sich einfach erweitern auf allgemeine Tridiagonalmatrizen $T = \text{tridiag}(a, b, c)$, falls im Algorithmus $|c_i|^2$ durch $a_i c_i$ ersetzt wird.

Beispiel 3.26
Sei T gegeben durch

$$T = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 4 \\ 0 & 4 & 5 \end{pmatrix}.$$

Dann ist
$f_3(\lambda) = 1,\ f_2(\lambda) = 1 - \lambda,\ f_1(\lambda) = \lambda^2 - 4\lambda - 1,\ f_0(\lambda) = -\lambda^3 + 9\lambda^2 - 3\lambda - 21.$
Aus $f_0(\lambda) = 0$ ergeben sich die Eigenwerte.
Weiter ist $g_3(\lambda) = 0,\ g_2(\lambda) = -1,\ g_1(\lambda) = 2\lambda - 4,\ g_0(\lambda) = -3\lambda^2 + 18\lambda - 3.$

Bemerkung 3.27
Gilt $c_k \neq 0$ für $1 \leq k \leq n-1$, so haben die Polynome f_{n-i} i reelle einfache Nullstellen $\lambda_j^{(i)}, j = 1, \ldots, i$ und die Nullstellen von f_{n-i} trennen die Nullstellen von f_{n-i-1}. Daher bilden die Polynome $(f_n, \ldots, f_0)$ eine Sturmsche Kette.
Dadurch gilt für ein beliebiges Intervall $[a, b]$ mit $f_0(a) f_0(b) \neq 0$:

Ist n_a die Anzahl von Vorzeichenwechsel der Sequenz

$$(f_n(a), \ldots, f_0(a)$$

und n_b die Anzahl von Vorzeichenwechsel der Sequenz

$$(f_n(b), \ldots, f_0(b),$$

so besitzt f_0 auf $[a, b]$ genau $n_a - n_b$ Nullstellen. Mit Intervallhalbierung kann man dann alle Nullstellen in $[a, b]$ finden.

Motivaton 3.28 (Bestimmung von $\rho_A(\lambda)$ für unreduzierte Hessenberg-Matrizen)
Sei $H = (h_{ij})_{i,j} \in \mathbb{K}^{n \times n}$ unreduzierte Hessenberg-Matrix, d.h. $h_{i+1,i} \neq 0\ \forall i = 1, \ldots, n-1$.

Vorschlag von Hyman (1957)
Betrachte das lineare Gleichnugssystem

$$(H - \lambda I)x = -c\ e_1, c \in \mathbb{K}.$$

Setzte man $x_n = 1$, so kann man durch Rückwärtseinsetzen nacheinander $x_{n-1}, x_{n-2}, \ldots, x_1$ berechnen und schließlich c bestimmen.

Andererseits kann x_n mit der Cramerschen Regel berechnet werden durch:

$$1 = x_n = \frac{(-1)^n c h_{2,1} h_{3,2} \ldots h_{n,n-1}}{\det(H - \lambda I)}.$$

Also folgt

$$\rho_H(\lambda) = \det(H - \lambda I) = (-1)^n c h_{2,1} h_{3,2} \ldots h_{n,n-1}.$$

Mit diesem Vorgehen erhalten wir folgenden Algorithmus.

Satz 3.29 (Verfahren nach Hyman)
Sei $H \in \mathbb{K}^{n\times n}$ unreduzierte Hessenberg-Matrix. Für H mit charakteristischem Polynom

$$\rho_H(\lambda) = (-1)^n h_{2,1}h_{3,2}\dots h_{n,n-1}\varphi(\lambda)$$

liefert der Algorithmus

$$\begin{aligned}
&\underline{\text{Input:}}\ \lambda \in \mathbb{C}\\
&h_{1,0} = 1, x_n = 1, y_n = 0\\
&x_{n-i} = \tfrac{1}{h_{n-i+1,n-i}}\Big(\lambda x_{n-i+1} - \sum_{j=n-i+1}^{n} h_{n-i+1,j}x_j\Big)\\
&y_{n-i} = \tfrac{1}{h_{n-i+1,n-i}}\Big(x_{n-i+1} + \lambda y_{n-i+1} - \sum_{j=n-i+1}^{n} h_{n-1+i,j}y_j\Big)
\end{aligned}$$

für $1 \le i \le n$ das Ergebnis

$$\begin{aligned}
x_0 &= \varphi(\lambda)\ \ (= c),\\
y_0 &= \varphi'(\lambda).
\end{aligned}$$

Ist λ ein Eigenwert von H, so ist $x = (x_1, \dots, x_n)^T$ ein zugehöriger Eigenvektor.

Beweis: Folgt aus 3.28 und $y_{n-i} = x'_{n-i}$. □

3.5 Vektoriteration für partielle Eigenwertprobleme

Definition 3.30 (Vektoriteration nach von Mises)
Sei $A \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix mit dominaten Eigenwert λ_1, d.h.

$$(D-EW) \qquad |\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Dann erhält man eine Folge von Approximationen $\lambda_k, k = 1, 2, \ldots$ von λ_1 durch folgenden Algorithmus:

Input: $z^0 \in \mathbb{C}^n$ mit $\|z^0\| = 1$ und $l \in \{1, \ldots, n\}$

Für $k = 1, 2, \ldots$ berechne

$$\begin{aligned} \tilde{z}^k &= Az^{k-1} \\ z^k &= \frac{1}{\|\tilde{z}^k\|}\tilde{z}^k \\ \lambda^k &= \frac{(Az^k)_l}{z^k_l} \end{aligned}$$

Satz 3.31 (Konvergenz der Vektoriteration nach von Mises)
Sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar mit einer Basis $\{u_1, \ldots, u_n\}$ von normierten Eigenvektoren zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$. Sei λ_1 ein dominanter Eigenwert von A, d.h. es gelte $(D-EW)$. Der Startwert $z^0 \in \mathbb{C}$ habe eine nichttriviale Komponente in Richtung u_1, d.h.

$$z^0 = \sum_{i=1}^{n} \alpha_i u_i \quad \text{mit } \alpha_1 \neq 0.$$

Dann gilt für z^k, λ^k aus der Vektoriteration nach von Mises (Def. 3.30), falls $u_{1,l} \neq 0$ ist:

1) $\|z^k - \sigma_k u_1\| = \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}|^k\Big) \quad (k \to \infty)$ mit $\sigma_k := \frac{\lambda_1^k \alpha_1}{|\lambda_1^k \alpha_1|}$, also $|\sigma_k| = 1$.

2) $\lambda^k - \lambda_1 = \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}|^k\Big) \quad (k \to \infty)$.

Beweis: Für die Iterierten z^k zeigt man durch vollständige Induktion, dass

$$z^k = \frac{A^k z^0}{\|A^k z^0\|}.$$

Mit $z^0 = \sum_{i=1}^{n} \alpha_i u_i$ folgt weiter

$$A^k z^0 = \sum_{i=1}^{n} \alpha_i \lambda_i^k u_i = \lambda_1^k \alpha_1 \Big(u_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_i\Big).$$

Wegen $(D-EW)$ folgt dann

$$A^k z^0 = \lambda_1^k \alpha_1 \Big(u_1 + \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}|^k\Big) \quad (k \to \infty).$$

und hieraus

$$z^k = \frac{\lambda_1^k \alpha_1 \Big(u_1 + \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}\Big|^k\Big)}{|\lambda_1^k \alpha_1| \Big\|u_1 + \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}\Big|^k\Big)\Big\|} = \sigma_k u_1 + \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}\Big|^k\Big).$$

Also ist 1) gezeigt. Weiter gilt

$$\begin{aligned}
\lambda^k &= \frac{(Az^k)_l}{z_l^k} \\
&= \frac{(A^{k+1}z^0)_l}{\|A^k z^0\|} \frac{\|A^k z^0\|}{(A^k z^0)_l} \\
&= \frac{\lambda_1^{k+1}\Big(\alpha_1 u_{1,l} + \sum_{i=2}^n \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^{k+1} u_{i,l}\Big)}{\lambda_1^k \Big(\alpha_1 u_{1,l} + \sum_{i=2}^n \alpha_i \Big(\frac{\lambda_i}{\lambda_1}\Big)^k u_{i,l}\Big)} \\
&\overset{u_{1,l}\neq 0}{=} \lambda_1 + \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}\Big|^k\Big) \quad (k \to \infty).
\end{aligned}$$

□

Bemerkung 3.32

1) Die Konvergenz der Vektoriteration nach von Mises ist also umso besser, je weiter der dominante Eigenwert λ_1 von den anderen Eigenwerten entfernt ist.

2) Varainte für $A \in \mathbb{C}^{n\times n}$ hermitesch:
Ist A hermitesch, so erhält man eine bessere Näherung von λ_1, wenn man zur Berechnung von λ^k den Rayleigh-Quotienten verwendet, d.h.

$$\lambda^k := R_A(z^k).$$

In diesem Fall gilt:

$$\lambda^k - \lambda_1 = \mathcal{O}\Big(|\frac{\lambda_2}{\lambda_1}\Big|^{2k}\Big) \quad (k \to \infty).$$

Definition 3.33 (Inverse Iteration nach Wielandt)
Sei $A \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix mit

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

so erhält man eine Näherung des Kehrwertes $\frac{1}{\lambda_n}$ des kleinsten Eigenwertes von A, indem man in der Vektoriteration nach von Mises A durch A^{-1} ersetzt.

Bemerkung 3.34 (Inverse Iteration mit Diagonal-Shift)
Ist $A \in \mathbb{C}^{n\times n}$ diagonalisierbar und $\mu \neq \lambda_i$ für alle $i = 1, \ldots, n$ eine gute Näherung eines Eigenwertes λ_j mit

$$|\lambda_j - \mu| \ll |\lambda_i - \mu| \qquad \forall i \neq j,$$

so kann die Näherung μ durch inverse Iteration für die Matrix $B := A - \mu I$ verbessert werden (Diagonal-Shift).

3.6 Das QR-Verfahren

Ziel: Verwende Ähnlichkeitstransformationen, um A sukzessive auf obere Dreiecksform zu transformieren:

$$\begin{aligned} A^0 &:= A, \\ A^i &:= T_i^{-1} A^{i-1} T_i, \quad i = 1, 2, ldots \end{aligned}$$

Beim QR-Verfahren wird T_i unitär gewählt!

Definition 3.35 (QR-Verfahren)
Sei $A \in \mathbb{C}^{n\times n}$. Dann ist das QR-Verfahren definiert durch:

$$\begin{aligned} A^0 &:= A, \\ A^i &:= Q^i R^i, && (QR\text{-Zerlegung von } A^i) \\ A^{i+1} &:= R^i Q^i. \end{aligned}$$

Hierbei ist Q^i unitär und R^i obere Dreiecksmatrix (siehe Numerik I, Kapitel 2.2). Wegen

$$A^{i+1} = R^i Q^i = (Q^i)^H A^i Q^i$$

sind alle Iterierten A^i ähnlich zu A.

Satz 3.36 (Konvergenz des QR-Verfahrens)
Die Eigenwerte von $A \in \mathbb{C}^{n\times n}$ seien betragsmäßig getrennt, d.h.

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

Dann gilt für die Diagonalelemente a_{jj}^i der Matrizen A^i des QR-Verfahrens 3.35:

$$\{\lim_{i\to\infty} a_{jj}^i |\ j = 1, \ldots, n\} = \{\lambda_1, \ldots, \lambda_n\}.$$

Weiter gilt

$$\lim_{i\to\infty} a_{jk}^i = 0 \text{ für } j > k.$$

Beweis: Siehe z.B. Stoer, Bulirsch [8].

Bemerkung 3.37
Da die Berechnung der QR-Zerlegung für allgemeine Matrizen A sehr aufwendig ist, bringt man in den Anwendungen A zunächst durch Housholder-Transformation auf Hessenberg-Form. Die Berechnung der QR-Zerlegung einer Hessenberg-Matrix kann dann mit Hilfe der sogenannten Givens-Rotation durchgeführt werden.

Algorithmus 3.38 (QR-Verfahren für Hessenberg-Matrizen)
Sei $A \in \mathbb{R}^{n\times n}$ eine Hessenberg-Matrix. Der folgende Algorithmus bestimmt in Schritt 1 (implizit) die QR-Zerlegung $A = QR$ und überschreibt in Schritt 2 A mit $A = RQ$.

1. Für $k = 1, \ldots, n-1$:

   Bestimme $c_k = \cos\Phi_k$ und $s_k = \sin\Phi_k$ mit

$$\begin{pmatrix} c_k & -s_k \\ s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,k} \\ a_{k+1,k} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

   Für $j = k, \ldots, n$ setze

$$\begin{pmatrix} a_{k,j} \\ a_{k+1,j} \end{pmatrix} := \begin{pmatrix} c_k & -s_k \\ s_k & c_k \end{pmatrix} \begin{pmatrix} a_{k,j} \\ a_{k+1,j} \end{pmatrix}.$$

2. Für $k = 1, \ldots, n-1$:
   Für $j = k, \ldots, n$ setze

$$(a_{j,k}, a_{j,k+1}) = (a_{j,k}, a_{j,k+1}) \begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix}.$$

Bemerkung 3.39 (LR-Verfahren)
Ersetzt man in Definition 3.35 die QR-Zerlegung durch die LR-Zerlegung, so erhält man das LR-Verfahren. Unter geeigneten Vorausetzungen gilt

$$\lim_{i\to\infty} A^i = \lim_{i\to\infty} R^i = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

und

$$\lim_{i\to\infty} L^i = I.$$

Nachteil:

1) Eventuell ist Pivotisierung notwendig und gilt nur $P^iA^i = L^iR^i$ mit einer Permutationsmatrix P^i, so ist die Konvergenz nicht gesichert.

2) L^i ist nicht unitär und das Verfahren konvergiert schlechter als das QR-Verfahren.

Voteil:

Für Hessenberg-Matrizen $A \in \mathbb{R}^{n\times n}$ ist die Berechnung der LR-Zerlegung mit dem Gauß-Algorithmus etwa doppelt so schnell wie die Berechnung der QR-Zerlegung.

Kapitel 4

Approximation

4.1 Allgemeine Approximation in normierten Räumen

Definition 4.1 (Beste Approximation/Proximum)
Sei $(X, \|\cdot\|)$ ein normierter Raum und $T \subset X$ eine beliebige Teilmenge. Zu einem $v \in X$ definiere

$$J_v(u) := \|v - u\| .$$

Dann heißt ein $u \in T$ beste Approximation *oder* Proximum *von v in T, g.d.w.*

$$J_v(u) = \inf_{w \in T} J_v(w).$$

Die Zahl $E_v(T) := \inf_{w \in T} J_v(w)$ heißt Minimalabstand von $v \in X$ zur Teilmenge T.

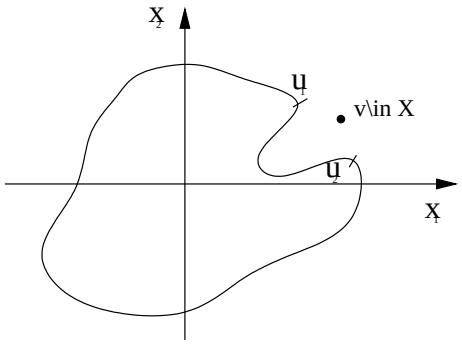

Abbildung 4.1: Proximum: Beispiel

Beispiel 4.2

1) Sei $X = \mathbb{R}^2$ und $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm. Sei $T := \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$. Dann existiert zu jedem $v \in X$ eine beste Approximation $u \in T$:
$\Longrightarrow$ Hier existiert zu jedem $v \in X$ genau ein Proximum.

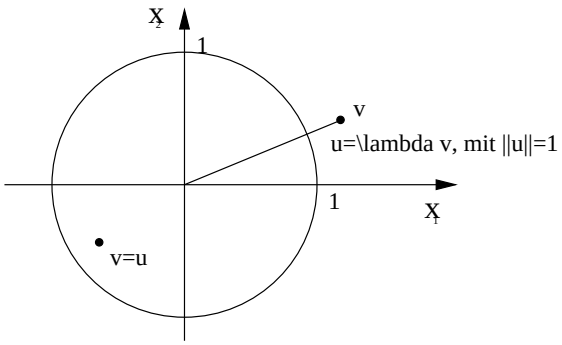

Abbildung 4.2: Proximum: Beispiel 1)

2) Sei $X = \mathbb{R}^2, \|\cdot\| = \|\cdot\|_2$ und $T := \{x \in \mathbb{R}^2 \mid \|x\| < 1\}$.
Ist $v \in X \setminus T$, so existiert keine beste Approximation $u \in T$ von V, denn $E_v(T) = \|v\| - 1$ und für alle $w \in T$ gilt $\|w - v\| > \|v\| - 1$
$\Longrightarrow$ Hier existiert für $x \in X \setminus T$ kein Proximum.

3) Sei $X = \mathbb{R}^2, \|\cdot\| = \|\cdot\|_\infty$, d.h. $\|x\|_\infty = \max\{|x_1|, |x_2|\}$ und $T := \{(x_1, 0) \mid x_1 \in \mathbb{R}\}$. Sei $v = (0, 1) \in X$. Dann gilt $u \in [-1, 1] \times \{0\} \implies u$ ist Proximum, da $J_v(u) = 1 = E_v(T)$
$\Longrightarrow$ Hier existieren unendlich viele beste Approximationen.

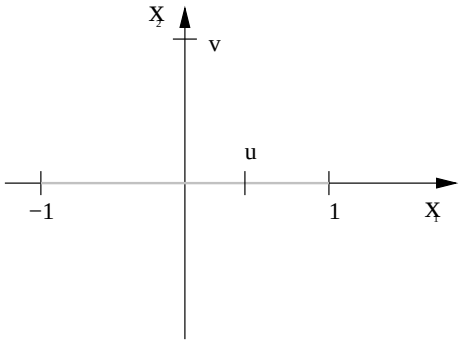

Abbildung 4.3: Proximum: Beispiel 3)

Satz 4.3 (Existenz eines Proximums)
Sei $T \subset X$ eine kompakte Teilmenge. Dann existiert zu jedem $v \in X$ ein Proximum $u \in T$.

Beweis: Sei $(u_n)_{n\in\mathbb{N}}$ eine Minimalfolge in T für $v \in X$, d.h. $\lim_{n \longrightarrow \infty} J_v(u_n) = E_v(T)$. Da T kompakt ist, enthält $(u_n)_{n\in\mathbb{N}}$ eine Teilfolge, die in T konvergiert, d.h. $\exists u \in T$ mit

$$\lim_{j \longrightarrow \infty} u_{n_j} = u.$$

Zu zeigen: u ist Proximum, d.h. $J_v(u) = E_v(T)$.

Es gilt:

$$\begin{array}{rccc} \|v-u\| \leq & \|v-u_{n_j}\| + & \|u_{n_j}-u\| & \\ & \downarrow (j\to\infty) & \downarrow (j\to\infty) & \\ & E_v(T) & 0 & \end{array}$$

$\Longrightarrow \quad \|v-u\| \leq E_v(T)$.

Da $E_v(T) = \inf_{w\in T} J_v(w) = \inf_{w\in T} \|v-w\|$, folgt:

$$J_v(u) = \|v-u\| = E_v(T). \qquad \square$$

Definition 4.4 (Konvexe Teilmengen)
$T \subset X$ heißt konvex, g.d.w.

$$K_{u_1,u_2} := \{\lambda u_1 + (1-\lambda)u_2 \,|\, \lambda \in (0,1)\} \subset T$$

für alle $u_1, u_2 \in T$.

$T \subset X$ heißt streng konvex, g.d.w.

$$K_{u_1,u_2} \subset \overset{\circ}{T}\ \forall u_1, u_2 \in T$$

mit $u_1 \neq u_2$ (vgl. Abbildung 4.4).

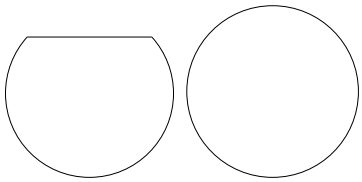

Abbildung 4.4: Konvexe und streng konvexe Mengen.

Satz 4.5 (Eindeutigkeit von Proxima)
Sei $T \subset X$ kompakt und streng konvexe Teilmenge des normierten Raumes X. Dann gibt es zu jedem $v \in X$ genau ein Proximum $u \in T$.

Beweis: Existenz ist klar nach Satz 4.3. Eindeutigkeit: Seien u_1, u_2 mit $u_1 \neq u_2$ Proxima von $v \in X$ in T. Dann gilt:

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

T konvex $\implies \left\|\frac{1}{2}(u_1 + u_2) - v\right\| = E_v(T)$.

Da T streng konvex ist, folgt $\frac{1}{2}(u_1 + u_2) \in \mathring{T}$. Also existiert ein $\tilde{\lambda} \in (0, 1)$, so dass

$$\tilde{u} = \frac{1}{2}(u_1 + u_2) + \tilde{\lambda}(v - \frac{1}{2}(u_1 + u_2)) \in T$$

ist. Dann gilt:

$$\begin{aligned} \|\tilde{u} - v\| &= \left\|\tfrac{1}{2}(1 - \tilde{\lambda})(u_1 + u_2) - (1 - \tilde{\lambda})v\right\| \\ &= (1 - \tilde{\lambda}) \left\|\tfrac{1}{2}(u_1 + u_2) - v\right\| \\ &= (1 - \tilde{\lambda}) E_v(T) \\ &< E_v(T). \end{aligned}$$

Dies ist ein Widerspruch zur Definition von $E_v(T) \implies u_1 = u_2$ □

Für Anwendungen ist vor allem der Fall wichtig, dass T ein endlich dimensionaler Teilraum von X ist.

Satz 4.6 (Fundamentalsatz der Approximationstheorie in normierten Räumen)
Sei $T \subset X$ ein endlichdimensionaler linearer Teilraum des normierten Vektorraums X. Dann existiert zu jedem $v \in X$ ein Proximum $u \in T$.

Beweis: Sei $(u_n)_{n\in\mathbb{N}}$ Minimalfolge für $v \in X$.
Wir zeigen zunächst: $(u_n)_{n\in\mathbb{N}}$ ist beschränkt.

Es gilt:

$$E_v(T) \leq \|v - u_n\| \leq E_v(T) + 1 \ \ \forall n \geq N.$$

Also ist $\|u_n\| \leq \|v - u_n\| + \|v\| \leq E_v(T) + 1 + \|v\| =: K_1 \ \ \forall n \geq N$.

Sei nun $K_2 \geq \|u_n\|$ für $n < N$ und setzte $K = \max\{K_1, K_2\}$, so folgt $\|u_n\| \leq K \ \forall n \in \mathbb{N}$.

Da T endlich dimensionaler linearer Teilraum ist, existiert also eine Teilfolge $(u_{n_j})_{j\in\mathbb{N}}$, die gegen ein $u \in T$ konvergiert.
Analog zum Beweis von Satz 4.3 zeigt man, dass u ein Proximum ist. □

Definition 4.7 (Streng normierter Raum)
Sei $(X, \|\cdot\|)$ ein normierter Raum. $\|\cdot\|$ heißt strenge Norm *und X* streng normiert*, g.d.w.*

$$(\|f + g\| = \|f\| + \|g\| \ \textit{für } f, g \in X \textit{ mit } f, g \neq 0) \implies (\exists \lambda \in \mathbb{C} \textit{ mit } g = \lambda f)\,.$$

Satz 4.8 (Eindeutigkeit in streng normierten Räumen)

Ist $(X, \|\cdot\|)$ ein streng normierter Raum und $T \subset X$ ein endlichdimensionaler linearer Teilraum, so existiert zu jedem $v \in X$ genau ein Proximum $u \in T$.

Beweis: Ist $v \in T$, so ist $u = v$ das eindeutige Proximum.

Sei also $v \in X \setminus T$. Sind u_1, u_2 verschiedene Proxima zu v, so gilt wie in Beweis von Satz 4.5:

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

also $\|(v - u_1) + (v - u_2)\| = \|v - u_1\| + \|v - u_2\|$

Da $\|\cdot\|$ strenge Norm ist, $\exists \lambda \in \mathbb{C}$ mit $v - u_1 = \lambda(v - u_2)$
$\Longrightarrow \quad (1 - \lambda)v = u_1 - \lambda u_2$.
Da $v \notin T$, folgt $\lambda = 1$ und somit $0 = u_1 - u_2$. Dies ist ein Widerspruch zur Annahme. □

4.2 Der Satz von Weierstraß: Approximation durch Polynome

Motivation 4.9
Bei der Approximation durch Polynome wurde in der Numerik I und in der Analysis immer von der Regularität der Funktion f gebrauch gemacht. So z.B. bei der Approximation von f durch die Taylorreihe, oder bei der Approximation durch Polynominterpolation.
In diesem Kapitel stellen wir uns die Frage, ob eine beliebig gute Polynomapproximation auch für Funktiontn $f \in C([a,b])$ möglich ist.

Satz 4.10 (Approximationssatz von Weierstraß)
Sei $X = C([a,b])$ und $\|\cdot\| = \|\cdot\|_\infty$ für $|a|, |b| < \infty$.
Dann gibt es zu jedem $f \in X$ und $\varepsilon > 0$ ein $n \in \mathbb{N}$ und ein $p \in \mathbb{P}_n$, so dass $\|f - p\|_\infty < \varepsilon$ ist

Beweis: (Konstruktiv!)
Ohne Einschränkung sei $[a,b] = [0,1]$ (sonst Transformation).
Wir zeigen, dass die Folge der Bernstein-Polynome

$$(B_n f)(x) := \sum_{i=0}^{n} f(\frac{i}{n}) \binom{n}{i} x^i (1-x)^{n-i}$$

für $n \longrightarrow \infty$ auf $[0,1]$ gleichmäßig gegen f konvergieren.

Definiere $q_{ni} := \binom{n}{i} x^i (1-x)^{n-i}$, dann ist

$$1 = (x + (1-x))^n = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{i=0}^{n} q_{ni}.$$

Also folgt:

$$f(x) - (B_n f)(x) = \sum_{i=0}^{n} (f(x) - f(\frac{i}{n})) q_{ni}(x)$$

$$\implies |f(x) - (B_n f)(x)| \leq \sum_{i=0}^{n} \left| f(x) - f(\frac{i}{n}) \right| q_{ni}(x) \quad \forall x \in [0,1].$$

Da f gleichmäßig stetig ist, existiert zu jedem $\varepsilon > 0$ ein δ, so dass $\forall x, i$

$$\left| x - \frac{i}{n} \right| < \delta \implies \left| f(x) - f(\frac{i}{n}) \right| < \frac{\varepsilon}{2}.$$

Sei $x \in [0,1]$. Setze

$$N_< := \{i \in \{0, \ldots, n\} \mid \left| x - \frac{i}{n} \right| < \delta\},$$

$$N_\geq <:= \{i \in \{0, \ldots, n\} \mid \left| x - \frac{i}{n} \right| \geq \delta\}.$$

Wir erhalten:

$$\sum_{i \in N_<} \left| f(x) - f(\frac{i}{n}) \right| q_{ni}(x) \leq \frac{\varepsilon}{2} \sum_{i \in N_<} q_{ni}(x) \leq \frac{\varepsilon}{2}.$$

und

$$\begin{aligned}\sum_{i\in N_{\geq}} \left|f(x)-f(\frac{i}{n})\right| q_{ni}(x) &\leq \sum_{i\in N_{\geq}} \left|f(x)-f(\frac{i}{n})\right| q_{ni}(x)\frac{(x-\frac{i}{n})^2}{\delta^2}\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\sum_{i\in N_{\geq}} q_{ni}(x)(x-\frac{i}{n})^2\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\sum_{i=0}^{n} q_{ni}(x)(x^2-2x\frac{i}{n}+(\frac{i}{n})^2).\end{aligned}$$

Es ist

1) $\sum\limits_{i=0}^{n} q_{ni}(x)x^2 = x^2$,

2) $\sum\limits_{i=0}^{n} q_{ni}(x)2x\frac{i}{n} = 2x\cdot x\underbrace{\sum_{i=1}^{n}\binom{n-1}{i-1}x^{i-1}(1-x)^{(n-1)-(i-1)}}_{=1} = 2x^2$,

3)

$$\begin{aligned}\sum_{i=0}^{n} q_{ni}(x)(\frac{i}{n})^2 &= \frac{x}{n}\sum_{i=1}^{n}(i-1)\binom{n-1}{i-1}x^{i-1}(1-x)^{(n-1)-(i-1)}+\frac{x}{n}\\ &= \frac{x^2}{n}(n-1)\underbrace{\sum_{i=2}^{n}\binom{n-2}{i-2}x^{i-2}(1-x)^{(n-2)-(i-2)}}_{=1}+\frac{x}{n}\\ &= x^2(1-\frac{1}{n})+\frac{x}{n} = x^2+\frac{x}{n}(1-x).\end{aligned}$$

Mit 1), 2) und 3) folgt:

$$\begin{aligned}\sum_{i\in N_{\geq}} \left|f(x)-f(\frac{i}{n})\right| q_{ni}(x) &\leq \frac{2\,\|f\|_\infty}{\delta^2}(\underbrace{x^2-2x+x^2}_{=0}+\underbrace{\frac{x}{n}(1-x)}_{\leq\frac{1}{4n}}\\ &\leq \frac{2\,\|f\|_\infty}{\delta^2}\frac{1}{4n}\\ &< \frac{\varepsilon}{2} \text{ für } n>\frac{M}{\delta^2\varepsilon}.\end{aligned}$$

Insgesamt folgt also

$$|f(x)-(B_nf)(x)| \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon \;\; \forall x\in[0,1]. \qquad \square$$

Bemerkung 4.11
Satz 4.10 zeigt, dass sich $f\in C([a,b])$ in folgender Weise entwickeln lässt:

$$f(x) = (B_1f)(x)+[(B_2f)(x)-(B_1f)(x)]+\ldots+[(B_nf)(x)-(B_{n-1}f)(x)]+\ldots$$

Die Reihe konvergiert gleichmäßig, lässt sich aber im allgemeinen nicht zu einer Potenzreihe umordnen!

Satz 4.12 (Fehlerabschätzung für die Approximation mit Bernsteinpolynomen)
Seien die Voraussetzungen von Satz 4.10 erfüllt und sei

$$w_f(\delta) := \sup_{|x'-x''|\leq\delta, x', x''\in[a,b]} \left|f(x') - f(x'')\right|$$

das Stetigkeitsmodul von f bezgl. δ. Dann gilt die Abschätung:

$$|f(x) - (B_n f)(x)| \leq \frac{5}{4} w_f(\frac{1}{\sqrt{n}}).$$

Beweis: Sei $\lambda = \lambda(x', x'', \delta) := \left\lceil \frac{|x'-x''|}{\delta} \right\rceil$ das größte Ganze.
Mit der Definition von $wf(\delta)$ folgt:

$$\delta_1 \leq \delta_2 \implies w_f(\delta_1) \leq w_f(\delta_2).$$

Damit folgt:

$$\left|f(x') - f(x'')\right| \leq w_f(\left|x' - x''\right|) \leq w_f((\lambda+1)\delta).$$

Aus $w_f(\mu\delta) \leq \mu\ w_f(\delta)$ für $\mu \in \mathbb{N}$ folgt:

$$\left|f(x') - f(x'')\right| \leq (\lambda+1) w_f(\delta).$$

Setze $N_{\geq} := \{i \in \{0,\ldots,n\} \,|\, \lambda(x, \frac{i}{n}, \delta) \geq 1\}$ und $N_<$ entsprechend. Jetzt gehen wir analog zum Beweis von Satz 4.10 vor:

$$\begin{aligned} |f(x) - (B_n f)(x)| &\leq \sum_{i=0}^{n} \left|f(x) - f(\frac{i}{n})\right| q_{ni} \\ &\leq w_f(\delta) \sum_{i=0}^{n} (1 + \lambda(x, \frac{i}{n}, \delta)) q_{ni}(x). \end{aligned}$$

Da $\lambda(x, \frac{i}{n}, \delta) = 0$ für alle $i \in N_<$ gilt, folgt weiter:

$$\begin{aligned} |f(x) - (B_n f)(x)| &\leq w_f(\delta) \left(1 + \sum_{i\in N_{\geq}} \lambda(x, \frac{i}{n}, \delta) q_{ni}(x)\right) \\ &\leq w_f(\delta) \left(1 + \frac{1}{\delta} \sum_{i\in N_{\geq}} \left|x - \frac{i}{n}\right| q_{ni}(x)\right) \\ &\overset{\frac{|x-\frac{i}{n}|}{\delta} \geq 1 \ \forall i \in N_{\geq}}{\leq} w_f(\delta) \left(1 + \frac{1}{\delta^2} \sum_{i\in N_{\geq}} (x - \frac{i}{n})^2 q_{ni}(x)\right) \\ &\overset{\text{Analog zum Beweis von 4.10}}{\leq} w_f(\delta) \left(1 + \frac{1}{4n\delta^2}\right). \end{aligned}$$

Wählen wir $\delta = \frac{1}{\sqrt{n}}$, so folgt die Abschätzung des Satzes. □

Bemerkung 4.13

1) Abhängig vom Stetigkeitsmodul kann die Schranke in Satz 4.12 beliebig langsam konvergieren. Bei höheren Anforderungen an die Stetigkeit von f kann andererseits eine schnellere Konvergenz erwartet werden.

2) In der Praxis ist die Approximation durch Bernstein-Polynome nicht von Bedeutung. Im nächsten Abschnitt werden wir wirkungsvollere Verfahren kennen lernen!

4.3 Gleichmäßige Approximation / Tschebyschev Approximation

Motivation 4.14
In §4.2 haben wir gesehen, dass sich jede stetige Funktion $f \in C^0([a,b])$ durch Polynome $p_n \in \mathbb{P}_n$ approximieren lassen. Die Frage, welches Polynom $p \in \mathbb{P}_n$ das Proximum zu f in $T = \mathbb{P}_n$ ist, wurde jedoch nicht beantwortet. Ist $\|\cdot\| = \|\cdot\|_\infty$, so beantworten wir diese Frage in diesem Abschnitt. $\|\cdot\|_\infty$ wird auch Tschebyschev-Norm genannt.

Definition 4.15 (best approximatin polynomial (BAP))
Sei $X = C^0(I)$, $I \subset \mathbb{R}$ beschränktes Intervall ausgestattet mit der ∞-Norm $\|\cdot\| = \|\cdot\|_\infty$.

Dann heißt $p_n \in \mathbb{P}_n$ best approximating polynomial (BAP) vom Grad $\leq n$ von $f \in C^0(I)$, g.d.w. p_n Proximum von f in $\mathbb{P}_n$ ist.

Bemerkung 4.16
1) Der Fundamentalsatz der Approximationstheorie 4.6 liefert die Existenz eines (BAP), da $\mathbb{P}_n$ ein endlichdimensionaler linearer Teilraum von $C^0(I)$ ist.

2) Eindeutigkeit des (BAP) können wir zunächst nicht erwarten, da $\|\cdot\|_\infty$ keine strenge Norm ist.

Satz 4.17 (Charakterisierung von BAP)
Sei $f \in C^0(I), I = [a,b]$ beschränkt und $p \in \mathbb{P}_n$. Es gebe $n+2$ Punkte $a \leq x_0 < x_1 < \ldots < x_n < x_{n+1} \leq b$, so dass

1) $|f(x_i) - p(x_i)| = J_f(p) = \|f - p\|_\infty$ für $i = 0, \ldots, n+1$,

2) $f(x_{i+1}) - p(x_{i+1}) = -(f(x_i) - p(x_i))$ für $i = 0, \ldots, n$.

Dann ist p BAP vom Grad $\leq n$ an f.

Beweis: Sei $p^* \in \mathbb{P}_n$ und $M := \{x \in I \mid |f(x) - p^*(x)| = J_f(p^*)\}$.
Ist p^* kein BAP, so existiert ein $\bar{p} \in \mathbb{P}_n, \bar{p} \neq 0$, so dass $p^* + \bar{p}$ BAP ist (nach Satz 4.6). Dann gilt:

$$|(f(x) - p^*(x)) - \bar{p}(x)| = |f(x) - (p^*(x) + \bar{p}(x))| < |f(x) - p^*(x)| \quad \forall x \in M$$

$$\Longrightarrow \quad (f(x) - p^*(x))\bar{p}(x) > 0 \ \ \forall x \in M \qquad (*)$$

Sei nun $p \in \mathbb{P}_n$ und es gelten die Voraussetzungen 1) und 2).
Dann kann es kein $\bar{p} \neq 0, \bar{p} \in \mathbb{P}_n$ geben, so dass $(*)$ erfüllt ist. Denn dazu müsste $\bar{p}$ in $[a,b]$ mindestens $(n+1)$-mal das Vorzeichen wechseln (wegen 2)), also mindestens $n+1$ Nullstellen besitzen. Nach dem Fundamentalsatz der Algebra ist das nicht möglich.
Da es zu p kein $\bar{p}$ gibt, so dass $(*)$ gilt, muß p bereits BAP sein. □

Bemerkung 4.18
Im Beweis von Satz 4.17 haben wir lediglich verwendet, dass Polynome vom Grad $\leq n$ höchstens n Nullstellen haben. Da diese Eigenschaft auch in anderen Funktionsklassen erfüllt wird (Bsp: $1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots$), machen wir folgende Definition.

Definition 4.19 (Haarscher Raum, Tschebyschev Systeme)
Besitzen $n+1$ linear unabhängige Elemente $g_0, \ldots, g_n \in C^0([a,b])$ die Eigenschaft, dass jedes Element $g \in U = span(g_0, \ldots, g_n), g \neq 0$ in $[a,b]$ höchsten n verschiedene Nullstellen besitzt, so heißt U Haarscher Raum. *Eine Basis $\{g_o, \ldots, g_n\}$ eines Haarschen Raumes nennt man* Tschebyschev System.

Definition 4.20 (Alternante)
Eine Menge von Punkten $a \leq x_0 < x_1 < \ldots < x_n <_{n+1} \leq b$ heißt Altenante *der Länge $n+2$ für $f \in C^0([a,b])$ und $p \in U$, U Harrscher Raum, g.d.w. $\exists\ \sigma \in \{-1,1\}$ mit*

$$sign(f(x_i) - p(x_i)) = \sigma(-1)^i \ \ \forall i = 0, \ldots, n+1.$$

Satz 4.21 (Alternantensatz von Tschebyschev)
Sei $U \subset C^0([a,b])$ ein Haarscher Raum der Dimsneion $n+1$ und sei $f \in C^0([a,b])$. Dann ist $p \in U$ genau dann Proximum von f bezgl. $\|\cdot\|_\infty$, wenn eine Alternante $\{x_0, \ldots, x_{n+1}\}$ der Länge $n+2$ existiert, so dass

$$|f(x_i) - p(x_i)| = J_f(p) \ \ \forall i = 0, \ldots, n+1.$$

Beweis: Die hinreichende Aussage folgt aus dem Charakterisierungssatz 4.17 und der Bemerkung 4.18.

Beweis der notwendigen Aussage:

Zu zeigen: $p \in U$ Proximum $\implies$ $\exists$ Alternante der Länge $n+2$ mit $|f(x_i) - p(x_i)| = J_f(p) \ \ \forall i = 0, \ldots, n+1$.

Beweis: Sei $p \in U$ Proximum und wir nehmen an, daß es kein Alternante der Länge $n+2$ gibt, so dass $|f(x_i) - p(x_i)| = J_f(p) \forall i = 0, \ldots, n+1$ gilt.
Dann exisitieren also höchstens $n+1$ Werte $x_i \in I$ mit $|f(x_i) - p(x_i)| = J_f(p)$ und $\text{sign}(f(x_i) - p(x_i) = \sigma(-1)^i, i = 0, \ldots, k,\ k \leq n$.

Sei $M := \{x_0, \ldots, x_k\}$ die Menge dieser Extremalpunkte.
Da U Haarscher Raum der Dimension $n+1$ ist, existiert ein $p^* \in U$ mit den einfachen Nullstellen $\xi_0, \ldots, k, k \leq n$, wobei $x_i < \xi_i < x_{i+1}$ ist und

$$\text{sign}(p^*(x_i)) = \text{sign}(f(x_i) - p(x_i))$$

für $i = 0, \ldots, k$ gilt.

O.E. können wir annehmen, dass $|p^*(x)| \leq 1$ ist für alle $x \in I$ (sonst verwende $\bar{p}^* := \frac{1}{\|p^*\|_\infty} p^*$).
Also gilt für p, f, p^*:

$$(f(x) - p(x))p^*(x) > 0 \ \ \forall x \in M.$$

Wir setzen: $M' := \{x \in I : (f(x) - p(x))p^*(x) \leq 0\}$.
Dann gilt:

1) M' ist abgeschlossen,

2) $M \cap M' = \emptyset$,

3) $d := \max_{x \in M'} |f(x) - p(x)| < \max_{x \in M} |f(x) - p(x)|$ (bzw. $d := 0$ falls $M' = \emptyset$).

Definiere: $\theta = \frac{1}{2}(\max_{x \in I} |f(x) - p(x)| - d) > 0$ (nach 3)) und sei $q := p + \theta p^*$.

Sei $\xi \in I$ ein Wert, für den gilt

$$|f(\xi) - q(\xi)| = \max_{x \in I} |f(x) - q(x)|$$

1. Fall: $\xi \in M'$

$$\begin{aligned} \max_{x \in I} |f(x) - q(x)| &= \max_{x \in I} |(f(x) - p(x)) - \theta p^*(x)| \\ &= |(f(\xi) - p(\xi)) - \theta p^*(\xi)| \\ &\leq \underbrace{|f(\xi) - p(\xi)|}_{\leq d,\ da\ \xi \in M'} + \theta \underbrace{|p^*(\xi)|}_{\leq 1} \\ &\leq d + \theta \leq \tfrac{1}{2}(\max_{x \in I} |f(x) - p(x)| + d) \\ &\overset{(3)}{<} \max_{x \in I} |f(x) - p(x)| . \end{aligned}$$

Dies ist ein Widerspruch dazu, dass p BAP ist.

2. Fall $\xi \notin M'$:

Dann gilt wegen $\operatorname{sign}(f(x) - p(x)) = \operatorname{sign}(p^*(x)$ für $x \in M$:

$$\begin{aligned} |(f(\xi) - p(\xi)) - \underbrace{\theta}_{>0} p^*(\xi)| &< \max\{|f(\xi) - p(\xi)|, \theta |p^*(\xi)|\} \\ &\leq \max_{x \in I} |f(x) - p(x)| , \end{aligned}$$

da $\theta \leq \max_{x \in I} |f(x) - p(x)|$. Dies ist ein Widerspruch dazu, dass p BAP ist. □

Satz 4.22 (Eindeutigkeit)
Sei $U := \operatorname{span}(g_o, \ldots, g_n)$ Haarscher Unterraum von $C^0([a,b])$. Dann ist das Proximum $p \in U$ an ein $f \in C^0([a,b])$ eindeutig bestimmt.

Beweis: Seien p_1 und p_2 Proxima aus U an $f \in C^0([a,b])$.
Dann ist auch $\frac{1}{2}(p_1 + p_2)$ Proximum (Satz 4.5) und nach 4.21 exisitert eine Alternante der Länge $n+2$, so dass

$$f(x_i) - \frac{1}{2}(p_1(x_i) + p_2(x_i)) = \sigma(-1)^i E_f(U).$$

Also ist $\frac{1}{2}(f(x_i) - p_1(x_i)) + \frac{1}{2}(f(x_i) - p_2(x_i)) = \sigma(-1)^i E_f(U)$.
Da p_1, p_2 Proxima sind, gilt $|f(x_i) - p_k(x_i)| \leq E_f(U)$ $k = 1,2$ und $i = 0, \ldots, n+1$. Daraus folgt

$$f(x_i) - p_1(x_i) = f(x_i) - p_2(x_i) \ \ \forall i = 0, \ldots, n+1$$

$$\Longrightarrow \ \ p_1(x_i) = p_2(x_i) \ \ \forall i = 0, \ldots, n+1.$$

Da U Haarscher Unterraum ist, folgt hieraus $p_1 - p_2 \equiv 0$, da $p_1 - p_2$ mindestens $n+2$ Nullstellen hat. □

Satz 4.23 (Abschätzung für den Minimalabstand)
Sei $U := span(g_0, \ldots, g_n)$ ein Haarscher Unterraum von $C^0([a,b])$. Für $f \in C^0([a,b])$ und ein $p \in U$ sei $x_0, \ldots, x_{n+1}$ eine Alternante, dann gilt:

$$\delta \leq E_f(U) \leq \Delta$$

mit

$$\delta := \min_{i=0,\ldots,n+1} |f(x_i) - p(x_i)|,$$
$$\Delta := \max_{i=0,\ldots,n+1} |f(x_i) - p(x_i)|.$$

Beweis: Die Abschätzung nach oben ist klar, da $E_f(U)$ Minimalabstand ist.

Nehmen wir an, es gelte $E_f(U) < \delta$. Dann gilt für das eindeutig bestimmte Proximum $p^* \in U$ an f:

$$J_f(p^*) = E_f(U),$$

also

$$(*) \qquad \max_{i=0,\ldots,n+1} |f(x_i) - p^*(x_i)| \leq J_f(p^*) < \delta = \min_{i=0,\ldots,n+1} |f(x_i) - p(x_i)|.$$

Somit folgt mit

$$p^* - p = (f - p) - (f - p^*),$$

dass gilt

$$\operatorname{sign}(p^*(x_i) - p(x_i)) = \operatorname{sign}(f(x_i) - p(x_i)) = \sigma(-1)^i,$$

für $i = 0, \ldots, n+1$.
Also hat $p^* - p \in U$ in jedem der $n+1$ Teilintervalle (x_i, x_{i+1}), $i = 0, \ldots, n$, mindestens eine Nullstelle. Da U Haarscher Unterraum ist, muss also $p = p^*$ sein.
Dies ist Widerspruch zu $(*)$. □

Der Charakterisierungssatz 4.17 bildet die Grundlage für ein Verfahren zur Approximation des BAP zum einem $f \in C^0([a,b])$. Das Verfahren ist allgemein auch für Haarsche Unterräume über $\mathbb{R}$ durchführbar, wird hier jedoch nur für die Approximation durch Polynome formuliert.

(Austauschalgorithmus von Remez) 4.24
Sei $f \in C^0([a,b])$. Dann erzeugt der folgende Algorithmus von Remez eine Folge von Polynomen $p^{(k)} \in \mathbb{P}_n$ mit

$$\lim_{k \longrightarrow \infty} \left\| p^{(k)} - p_n \right\|_\infty = 0,$$

wobei p_n das eindeutig bestimmte BAP von f in $\mathbb{P}_n$ ist.
Als Abbruchktiterium kann die Bedingung

$$\Delta - \delta \leq \varepsilon$$

gewählt werden, wobei δ, Δ bezüglich $p^{(k)}$ wie in 4.23 definiert seien.

Remez-Algorithmus:

Input: Startzerlegung $I^{(0)}$ von $[a,b]$, d.h.

$$I^{(0)} := \{x_0, \ldots, x_{n+1}\}$$

mit $a \leq x_0 < x_1 < \ldots < x_{n+1} \leq b$.

Iteration für $k = 0, 1, 2, \ldots$

1. Schritt: Zu $I^{(k)}$ bestimme Polynom $p^{(k)} \in \mathbb{P}_n$, so dass $I^{(k)}$ Alternante für $f - p^{(k)}$ ist und für alle $x_i \in I^{(k)}$ gilt:

$$\left| f(x_i) - p^{(k)}(x_i) \right| = \left| \xi^{(k)} \right|$$

mit einer Konstante $\xi^{(k)} \in \mathbb{R}$.
Setze man $p^{(k)}(x) := \sum_{j=0}^{n} \alpha_j^{(k)} x^j$, so führen diese Forderungen auf das Gleichungssystem:

$$(-1)^i \xi^{(k)} + \sum_{j=0}^{n} \alpha_j^{(k)} x_i^j = f(x_i)\,, \quad i = 0, \ldots, n+1$$

für die Unbekannten $\xi^{(k)}, \alpha_0^{(k)}, \ldots, \alpha_n^{(k)}$.

2. Schritt: Setze $r^{(k)}(x) := f(x) - p^{(k)}(x)$ und bestimme $y \in [a, b]$ mit der Eigenschaft:

$$r^{(k)}(y) = \max_{x \in [a,b]} r^{(k)}(x).$$

Ersetze einen Punkt $x_i \in I^{(k)}$ durch y und erhalte $I^{(k+1)}$ gemäß folgender Austauschvorschrift:

1. Fall: $x_j < y < x_{j+1}$ für eine $j \in \{0, \ldots, n\}$.
   Falls $\operatorname{sign}(r^{(k)}(x_j)) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_j durch y, sonst erzetze x_{j+1} durch y.

2. Fall: $y < x_0$.
   Falls $\operatorname{sign}(r^{(k)}(x)) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_0 durch y, sonst ersetze x_{n+1} durch y.

3. Fall: $y > x_{n+1}$.
   Falls $\operatorname{sign}(r^{(k)}(x_{n+1})) = \operatorname{sign}(r^{(k)}(y))$, ersetze x_{n+1} durch y, sonst ersetze x_0 durch y.

Bemerkung: Gilt $y = x_i$ für ein $x_i \in I^{(k)}$, so ist $p^{(k)}$ bereits BAP. (Charakterisierungssatz 4.17)

Beweis: Skizze: Man zeigt:

1) Das Gleichungssystem in Schritt 1 ist stets eindeutig lösbar.

2) Es gilt $\xi^{(k+1)} > \xi^{(k)} \;\; \forall k = 0, 1, 2, \ldots$.

3) $\lim_{m \longrightarrow \infty} I^{(k_m)} = \{\bar{x}_0, \ldots, \bar{x}_{n+1}\}$ für eine Teilfolge.

4) $\lim_{m \longrightarrow \infty} p^{(k_m)} = p$ für diese Teilfolge.

5) p ist BAP an f.

Beispiel 4.25

$f(x) = x^2, x \in [0,1]$, Gesucht ist das BAP in $\mathbb{P}_1$.
Als Startzerlegung wählen wir $I^{(0)} = \{x_0, x_1, x_2\} = \{0, \frac{1}{3}, 1\}$.

k=0:

1. Schritt: Bestimme $p^{(0)}, \xi^{(0)}$ durch lösen von

$$\begin{array}{llll} \xi^{(0)} & +\alpha_0^{(0)} & & = 0 \\ -\xi^{(0)} & +\alpha_0^{(0)} & +\frac{1}{3}\alpha_1^{(0)} & = \frac{1}{9} \\ \xi^{(0)} & +\alpha_0^{(0)} & +\alpha_1^{(0)} & = 1 \end{array} .$$

Man erhält $\alpha_0^{(0)} = -\frac{1}{9}, \alpha_1^{(0)} = 1$ und $\xi^{(0)} = \frac{1}{9}$, d.h. $p^{(0)} = -\frac{1}{9} + x$.

2. Schritt: $\left\|f - p^{(0)}\right\|_\infty = \max\limits_{x\in[0,1]} \left|x^2 - x + \frac{1}{9}\right| = \frac{5}{36} > \frac{1}{9}$.
Das Maximum wird für $y = \frac{1}{2}$ angenommen. Nach dem Austauschkriterium ist $x_1 = \frac{1}{3}$ gegen $y = \frac{1}{2}$ auszutauschen. Wir erhalten $\implies$ $I^{(1)} = \{0, \frac{1}{2}, 1\}$.

k=1:

1. Schritt: Bestimme $p^{(1)}, \xi^{(1)}$ aus

$$\begin{array}{llll} \xi^{(1)} & +\alpha_0^{(1)} & & = 0 \\ -\xi^{(1)} & +\alpha_0^{(1)} & +\frac{1}{2}\alpha_1^{(1)} & = \frac{1}{4} \\ \xi^{(1)} & +\alpha_0^{(1)} & +\alpha_1^{(1)} & = 1 \end{array} .$$

Man erhält $\xi^{(1)} = \frac{1}{8}$ und $p^{(1)}(x) = -\frac{1}{8} + x$.

2. Schritt: Es ist $\left\|f - p^{(1)}\right\|_\infty = \max\limits_{x\in[0,1]} \left|x^2 - x + \frac{1}{8}\right| = \frac{1}{8}$.
Dieser Wert wird für $x_0 = 0, x_1 = \frac{1}{2}, x_2 = 1$ angenommen. Also ist $p^{(1)}$ bereits BAP an f in $\mathbb{P}_1$ (Satz 4.17). Das Verfahren bricht ab.

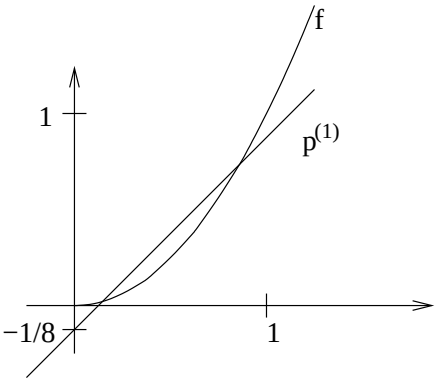

Abbildung 4.5: Beispiel

Motiviert durch das Beispiel wollen wir im folgenden der Frage nachgehen, wie für $f(x) = x^n$ in $[-1,1]$ das BAP in $\mathbb{P}_{n-1}$ aussieht. Dies führt uns zu den Tschebyschev Polynomen 1. Art.

Definition 4.26 (Tschebyschev Polynome 1. Art)
Durch die Rekursion

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

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

werden die sogenannten Tschebyschev Polynome 1. Art *definiert.*

Satz 4.27
Die Tschebyschev-Polynome 1. Art haben die Darstellung $T_n(x) = \cos(n \arccos(x))$, $x \in [-1, 1], n \in \mathbb{N}$. Es gelten die Eigenschaften

1) $|T_n(x)| < 1 \ \ \forall x \in [-1, 1]$,

2) $T_n(x)$ hat in $[-1, 1]$ die Extrempunkte

$$x_i^{(n)} = \cos(\frac{i\pi}{n}), T_n(x_i^{(n)}) = (-1)^i \ \ i = 0, \ldots, n,$$

3) T_n hat n einfache Nullstellen in $[-1, 1]$

$$\tilde{x}_i^{(n)} = \cos(\frac{2i-1}{2n}\pi) \ \ i = 1, \ldots, n,$$

4) Zwischen je zwei Nullstellen von T_{n+1} liegt eine Nullstelle von T_n.

Beweis: Siehe Satz 4.8 aus der Vorlesung *Numerik I.*

Satz 4.28
Sei $f(x) = x^n \in C^0([-1, 1])$. Dann ist

$$p_{n-1} := x^n - \frac{1}{2^{n-1}} T_n(x)$$

BAP zu f in $\mathbb{P}_{n-1}$.

Beweis: 1.) Zeige $p_{n-1} \in \mathbb{P}_{n-1}$.
Unter Verwendung der Rekursionsformel zeigt man induktiv, dass $T_n(x)$ die Form

$$T_n(x) = 2^{n-1}x^n + \sum_{i=1}^{n-1} \alpha_i x^i$$

hat. Dann folgt:

$$p_{n-1} = x^n - \frac{1}{2^{n-1}} T_n(x) = -\frac{1}{2^{n-1}} \sum_{i=1}^{n-1} \alpha_i x^i \implies p_{n-1} \in \mathbb{P}_{n-1}.$$

2.) Zeige p_{n-1} ist BAP zu f.
Es gilt $x^n - p_{n-1}(x) = \frac{1}{2^{n-1}} T_n(x)$. Nach Satz 4.27 (2) ist $\{x_i^{(n)} := cos(\frac{i\pi}{n}) \,|\, i = 0, \ldots, n\}$ eine Alternante der Länge $(n-1)+2$ für $p_{n-1}(x)$ und es gilt nach 4.27 1), 2)

$$\left| \frac{1}{2^{n-1}} T_n(x_n^{(i)}) \right| = \frac{1}{2^{n-1}} = \max_{x \in [-1,1]} \left| \frac{1}{2^{n-1}} T_n(x) \right|.$$

Nach dem Charakterisierungssatz 4.17 muß p_{n-1} BAP zu f sein. □

Satz 4.29
Die Tschebyschev-Polynome bilden bezüglich der Gewichtsfunktion $w(x) := \frac{1}{\sqrt{1-x^2}}$ ein Orthogonalsystem. Es gilt:

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

(ohne Beweis)

Definition 4.30 (Tschebyschev-Entwicklung)
Sei f stetig in $[-1,1]$. Dann heißen

$$a_k(f) := \frac{2}{\pi}\int_{-1}^{1} f(x)T_k(x)\frac{1}{\sqrt{1-x^2}}dx,\ k = 0,1,2,\ldots$$

Tschebyschev-Koeffizienten *von f und die formal gebildete Reihe*

$$S_f^n(x) = \frac{a_0(f)}{2} + \sum_{k=1}^{n} a_k(f)\cdot T_k(x)$$

heißt Tschebyschev-Entwicklung *von f.*

Entwicklungssatz 4.31
Sei $f \in C^2([-1,1])$. Dann konvergiert die Tschebyscheventwicklung für $x \in [-1,1]$ gleichmäßig gegen f, d.h.

$$S_f^\infty := \lim_{n \longrightarrow \infty} S_f^n = f.$$

Für die Tschebyschev-Koeffinzienten gilt die Abschätzung:

$$|a_k(f)| \leq \frac{c}{k^2},\ \ k = 1,2,3,\ldots.$$

Beweis: Mit $\varphi(\theta) := f(\cos\theta)$ erhält man durch Substitution $x = \cos\theta$ und zweimalige partielle Integration:

$$\begin{aligned} a_k(f) &= -\frac{2}{\pi k}\int_0^\pi \frac{d\varphi}{d\theta}\sin(k\theta)\,d\theta \\ &= \frac{2}{\pi k^2}\left[\frac{d\varphi}{d\theta}\cos(k\theta)\right]_0^\pi - \frac{2}{\pi k^2}\int_0^\pi \frac{d^2\varphi}{d\theta}\cos(k\theta)\,d\theta \end{aligned}$$

Folglich existiert ein $c > 0$ mit $|a_k(f)| \leq \frac{c}{k^2}$.

Aufgrund dieser Abschätzung existiert ein $g \in C^0([-1,1])$, so dass gilt

$$\lim_{n \longrightarrow \infty} \left\|S_f^n - g\right\|_\infty = 0.$$

Da die Tschebyschev-Polynome bezgl. $\frac{1}{\sqrt{1-x^2}}$ ein Orthogonalsystem bilden, gilt außerdem

$$\lim_{n \longrightarrow \infty} \left\| S_f^n - f \right\|_2 = 0.$$

Da $\left\| S_f^n - g \right\|_2 \leq \left\| S_f^n - g \right\|_\infty \left(\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} \right)^{\frac{1}{2}}$ ist, folgt

$$\begin{aligned} \|f - g\|_2 &\leq \left\| f - S_f^n \right\|_2 + \left\| S_f^n - g \right\|_2 \\ &\leq \underbrace{\left\| f - S_f^n \right\|_2}_{\longrightarrow 0 \ (n \to \infty)} + \underbrace{\left\| S_f^n - g \right\|_\infty}_{\longrightarrow 0 \ (n \to \infty)} \cdot \underbrace{\left(\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} \right)^{\frac{1}{2}}}_{=c} \\ &\leq 0. \end{aligned}$$

$\Longrightarrow \quad f = g.$ □

Folgerung 4.32
Die Koeffizientenabschätzung in Satz 4.31 zeigt, dass für $f \in C^2([-1,1])$ eine gute Näherung des Proximums $p_n \in \mathbb{P}_n$ durch S_f^n gegeben ist.
Diese Möglichkeit der Approximation bietet sich dann an, wenn die Koeffizienten $a_k(f)$ einfach zu berechnen sind.

Beispiel 4.33
Sei $f(x) = \sqrt{1-x^2}$, $I := [-1,1]$.
Hier erhält man nach Substitution mit t = arccos(x):

$$\begin{aligned} a_k(f) &= \frac{2}{\pi} \int_0^\pi \cos(kt) \sin(t)\, dt \\ &= \begin{cases} \frac{4}{\pi} \frac{1}{1-k^2} & für\ k = 2\kappa \\ 0 & für\ k = 2\kappa + 1 \end{cases}, \kappa \in \mathbb{N}. \end{aligned}$$

Also folgt $S_f^0(x) = \frac{2}{\pi}$; $S_f^2(x) = \frac{2}{3\pi}(5 - 4x^2)$; $S_f^4(x) = \frac{2}{15\pi}(23 - 4x^2 - 16x^4)$; ...

4.4 Approximation im Prä-Hilbertraum

Motivation 4.34
In diesem Abschnitt wollen wir die Approximation in Prä-Hilberträumen untersuchen, d.h. X sei normierter Raum und $\|x\| := \|x\|_2 := \langle x, x\rangle^{\frac{1}{2}}\ \forall x \in X$ mit einem Skalarprodukt $\langle\cdot,\cdot\rangle$ auf X.
Ist $T \subset X$ endlichdimensionaler linearer Teilraum, so gilt da $\|\cdot\|$ eine strenge Norm ist (Ü.A.), dass zu jedem $f \in X$ genau ein Proximum $p \in T$ existiert (Satz 4.8).

Satz 4.35 (Charakterisierung des Proximums in Prä-Hilberträumen)
Sei $(X, \|\cdot\|)$ ein Prä-Hilbertraum und $T \subset X$ einn endlichdimensionaler linearer Teilraum. Dann ist $p \in T$ genau dann Proximum an $f \in X$, wenn für alle $g \in T$ gilt

$$\langle f - p, g\rangle = 0.$$

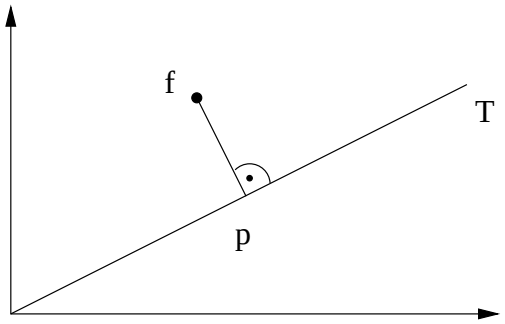

Abbildung 4.6: Proximum im Prä-Hilbertraum

Beweis:

"$\Leftarrow$" Gelte $\langle f - p, g\rangle = 0\ \forall g \in T$.
Definiere $\tilde{g} := g - p \in T$. Dann folgt:

$$\begin{aligned}\|f - g\|^2 &= \|f - p + \underbrace{p - g}_{=-\tilde{g}}\|^2 = \|(f-p) - \tilde{g}\|^2 \\ &= \langle (f-p) - \tilde{g}, (f-p) - \tilde{g}\rangle = \|f-p\|^2 + \|\tilde{g}\|^2 - 2\underbrace{\langle f - p, \tilde{g}\rangle}_{=0},\end{aligned}$$

$\Longrightarrow\ \|f - p\|^2 \leq \|f - g\|^2$. Also ist p Proximum.

"$\Rightarrow$" Sei $p \in T$ Proximum:
Wir nehemen an, es gäbe ein $g \in T$ mit $\langle f - p, g\rangle = c \neq 0$.
Setze $h := p + c\frac{g}{\|g\|^2} \in T$. Dann gilt

$$\|f - h\|^2 = \left\|f - p - c\frac{g}{\|g\|^2}\right\|^2 = \|f-p\|^2 - 2\frac{c}{\|g\|^2}\underbrace{\langle g, f - p\rangle}_{=c} + |c|^2\frac{\|g\|^2}{\|g\|^4}$$

$$\implies \|f-h\|^2 = \|f-p\|^2 \underbrace{- \frac{|c|^2}{\|g\|^2}}_{<0}$$

$$\implies \|f-h\| < \|f-p\| .$$

Dies ist ein Widerspruch dazu, dass p Proximum ist. □

Satz 4.36 (Berechnung des Proximums / Normalengleichungen)
Sei $\{p_0, \dots, p_n\}$ eine Basis von T. Sei $f \in X$.
Dann ist das Proximum $p(x) = \sum_{i=0}^{n} \alpha_i p_i(x)$ definiert durch:
$\alpha := (\alpha_0, \dots, \alpha_n)$ ist Lösung der Normalengleichungen:

$$\sum_{i=0}^{n} \alpha_i \langle p_i, p_j \rangle = \langle f, p_j \rangle \quad 0 \le j \le n$$

$$\iff S\alpha = b$$

mit $S_{ij} = \langle p_i, p_j \rangle$, $b_j = \langle f, p_j \rangle$.

Beweis: (Folgt direkt aus dem Charakterisierungssatz 4.35.)

Bemerkung 4.37 (Orthonormalsysteme)
Ist die Basis $\{p_0, \dots, p_n\}$ orthonormiert, so gilt mit $\alpha_i := \langle f, p_i \rangle$ für $0 \le i \le n$:

1) $p(x) = \sum_{i=0}^{n} \alpha_i p_i(x)$ ist Proximum an f in $\operatorname{span}(p_0, \dots, p_n)$.

2) $\|f-p\|^2 = \|f\|^2 - \langle f, p \rangle = \|f\|^2 - \sum_{i=0}^{n} \alpha_i \langle f, p_i \rangle = \|f\|^2 - \sum_{i=0}^{n} \alpha_i^2$.

Satz 4.38 (Besselsche Ungleichung)
Sei $\{p_i | i \in \mathbb{N}\}$ eine Orthonormalsystem unedlicher Dimension von X und für $f \in X$ sei $\alpha_i := \langle f, p_i \rangle$.
Dann gilt die Besselsche Ungleichung

$$\sum_{i=0}^{\infty} \alpha_i^2 \le \|f\|^2 .$$

Beweis: Aus Bemerkung 4.37 2) folgt für endliche Orthonormalsysteme

$$0 \le \|f\|^2 - \sum_{i=0}^{n} \alpha_i^2 .$$

Diese Aussage bleibt auch für unendliche ONSe richtig. □

Definition 4.39 (Vollständiges ONS)
Sei X ein Prä-Hilbertraum. Ein Orthonormalsystem (ONS) $\{p_i | i \in \mathbb{N}\}$ heißt vollständig, wenn es zu jedem $f \in X$ eine Folge $(f_k)_{k\in\mathbb{N}}$ gibt mit $f_k \in span(p_0, \ldots, p_k)$, so dass gilt

$$\lim_{k\to\infty} \|f - f_n\| = 0.$$

Satz 4.40 (Vollständigkeitsrelation)
Sei X ein Prä-Hilbertraum, $\{p_i | i \in \mathbb{N}\}$ ein ONS und für $f \in X$ sei $\alpha_i := \langle f, p_i\rangle$. Dann sind folgende Aussagen äquivalent:

(i) $\{p_i | i \in \mathbb{N}\}$ ist vollständiges ONS.

(ii) Es gilt $\sum_{i=0}^{n} \alpha_i^2 = \|f\|^2$.

Die Vollständigkeitsrelation 2) wird auch Parsevalsche Gleichung genannt.

Beweis: Sei $\{p_i | i \in \mathbb{N}\}$ ein vollständiges ONS. Wir betrachten die Folge $(f_k)_{k\in\mathbb{N}}$, $f_k \in \text{span}(p_0, \ldots, p_k)$, für die $\lim_{k\to\infty} \|f - f_n\| = 0$ gilt. Weiter sei $(\bar{p}_k)_{k\in\mathbb{N}}$ die Folge von Proxima in $\text{span}(p_0, \ldots, p_k)$ an f. Dann gilt

$$\|f - \bar{p}_k\| \leq \left\|f - \bar{f}_k\right\|$$

für alle $k \in \mathbb{N}$ und nach Bemerkung 4.37 2) $\|f - \bar{p}_k\|^2 = \|f\|^2 - \sum_{i=0}^{k} \alpha_i^2$. Wegen $\lim_{k\to\infty} \|f - f_n\| = 0$ folgt also

$$\lim_{k\to\infty} (\|f\|^2 - \sum_{i=0}^{k} \alpha_i^2) = 0$$

und somit $\lim_{k\to\infty} \sum_{i=0}^{k} \alpha_i^2 = \|f\|^2$.

Ist andereseits

$$\lim_{k\to\infty} (\|f\|^2 - \sum_{i=0}^{k} \alpha_i^2) = 0,$$

so folgt $\lim_{k\to\infty} \|f - \bar{p}_k\| = 0$ und damit die Vollständigkeit von $\{p_i | i \in \mathbb{N}\}$.

Satz 4.41 (Legendre Polynome)

1) Die Legendre Polynome

$$p_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx}(x^2 - 1)^n \quad n = 0, 1, 2, \ldots$$

bilden für das Intervall $[-1, 1]$ ein Orthogonalsystem.
Es gilt für $m, n \in \mathbb{N}$:

$$\int_{-1}^{1} p_n(x) \cdot p_m(x) dx = \begin{cases} 0 & m \neq n \\ \frac{2}{2n+1} & m = n \end{cases}.$$

Außerdem ist $p_n(1) = 1$ und $p_n(-1) = (-1)^n$.

2) Die Legendre Polynome entstehen durch Orthogonalisierung der Monome $1, x, x^2, x^3, x^4, \ldots$.

3) Ist $f(x) = x^n$, $x \in [-1, 1]$, so ist $p(x) := x^n - p_n(t)$ das Proximum in $\mathbb{P}_{n-1}$ an f bezgl. $\|\cdot\|_2$.

(ohne Beweis)

Satz 4.42 (Vollständigkeit der Legendre Polynome)
Die normierten Legendre Polynome

$$p_n(x) := \frac{1}{2^n n!}\sqrt{\frac{2n+1}{2}}\frac{d^n}{dx}(x^2-1)^n \quad n = 0, 1, 2, \ldots$$

bilden ein vollständiges ONS in $(C([-1, 1], \|\cdot\|_2)$.

(ohne Beweis)

Index

Literaturverzeichnis

[1] M. Bollhöfer und V. Mehrmann.*Numerische Mathematik*, Vieweg Verlag, Wiesbaden 2004.

[2] P. Braess. *Finite Elemente*, Springer, Berlin 1992.

[3] P.G. Ciarlet.*The finite element methods for elliptic problems*, North-Holland, Amsterdam 1987.

[4] P. Deuflhard und F. Bornemann. *Numerische Mathematik II*, 3. Auflage, Walter de Gruyter, Berlin 2002.

[5] R. D. Grigorieff. *Numerik gewöhnlicher Differentialgleichungen*, 2 Auflage, Teubner, Stuttgart 1977.

[6] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme. Leitfäden der Angewandten Mathematik und Mechanik, 69*, Teubner Studienbücher Mathematik, Teubner, Stuttgart 1991.

[7] G. Hämmerlin und K.-H. Hoffmann.*Numerische Mathematik*, Springer, Berlin 1989.

[8] J. Stoer und R. Bulirsch. *Einführung in die Theorie der Numerischen Mathematik I & II*, Heidelberger Taschenbücher, Springer, Berlin 2000.

[9] W. Walter. *Gewöhnliche Differentialgleichungen*, 5. Auflage, Springer, Berlin, 1993.