Technische Universität Dresden

Skript:

Numerik

Verfasser	Franziska Kühn
Daten	Prof. Dr. Karsten Eppler Wintersemester 2009/10 Grundstudium

Inhaltsverzeichnis

1 Interpolation und Approximation

1.1 Aufgabenstellung & Grundlagen

- Vorgegeben: Meßreihe (x_i, f_i) für $i = 0(1)n$
- Aufgabe: Bestimmen eines $y = y(x)$ derart, dass (1) $y(x_i) = f_i$ für $i = 0(1)n$ (Interpolation) bzw. (2) $y(x_i) \approx f_i$ für $i = 0(1)n$ („Beste“ Approximation)
- Vorgehensweise: Ansatz
$$y(x) = \sum c_j \varphi_j(x)$$
mit φ_j als Ansatzfunktionen. Interpolationsbedingung:
$$\sum c_j \cdot \varphi(x_i) \stackrel{!}{=} f_i \qquad i = 0(1)n$$
- Allgemeinere Formulierung: gegeben $f \in U$, Abbildung $T : Y \to U$, gesucht $y \in Y$ mit
$$T \cdot y = f$$
für Interpolation bzw.
$$\|T \cdot y - f\| \leq \|T \cdot v - y\| \; \forall v \in Y$$
bei Approximation.
- im obigen Beispiel: $U \subset \mathbb{R}^{n+1}$, $Y \subset C(\overline{\Omega})$. $Y = C(\overline{\Omega})$ ist nicht sinnvoll. Einschränkungen erforderlich, z.B. $\dim Y < \infty$
- Beispiel: Polynominterpolation
$$Y = \{y : y(x) = \sum_{j=0}^{n} c_j \cdot x^j\}$$
allgemein:
$$Y = \text{span}\{\varphi_j\}_{j=0}^{n} = \{y(x) = \sum_{j=0}^{n} \varphi_j \cdot c_j, c_j \in \mathbb{R}\}$$
- Definition: Ist T linear und $Y = \text{span}\{\varphi_j\}_{j=0}^{n}$, so liegt eine Aufgabe der linearen Interpolation bzw. linearen Approximation vor.
- Ein Funktionensystem $\{\varphi_j\}$ genügt der Haarschen Bedingung , wenn gilt:
$$\{\varphi \in \text{span}(\varphi_j), \varphi \neq 0, \varphi_j \text{ linear unabhängig}\} \Rightarrow \varphi \text{ besitzt höchstens n Nullstellen}$$
(Haarscher Raum). In diesem Fall Forderung (1) äquivalent zur Lösung eines linearen Gleichungssystems:
$$\begin{aligned} \sum_{j=0}^{n} c_j \cdot \varphi_j(x_i) &= f_i \qquad i = 0(1)n \\ A \cdot y &= f \end{aligned}$$

mit

$$A = \begin{pmatrix} \varphi_0(x_0) & \dots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \dots & \varphi_n(x_n) \end{pmatrix} \quad y = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \quad f = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

Forderung: A regulär für beliebige Wahl von x_i

- Approximation: $f \in C(\overline{\Omega})$, gesucht $y \in \text{span}\{\varphi_j\}$ mit

$$\begin{aligned} \|y - f\|_\infty &\leq \|v - f\| \; \forall v \in Y \\ \|v\|_\infty &= \max_{x \in \Omega} |v(x)| \end{aligned}$$

(Tschebyscheff-Approximation)

1.2 Polynominterpolation

$$\varphi_j(x) = x^j \qquad j = 0(1)n \qquad x \in [a, b]$$

Lemma: Es seien x_i für $i = 0(1)n$ paarweise verschieden und $\varphi_j(x) = x^j$. Dann ist A regulär.

Beweis:

- Sei

$$A = \begin{pmatrix} \varphi_0(x_0) & \dots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \dots & \varphi_n(x_n) \end{pmatrix}$$

für gewisse $a = x_0 < x_1 < \ldots < x_n = b$ singulär. Dann existiert ein $y \neq 0$ mit $A \cdot y = 0$. Also hat

$$\sum_{j=0}^{n} y_j \cdot x^j =: p(x)$$

(n+1) Nullstellen. Fundamentalsatz der Algebra: $p = 0$. Aber auch: $y = 0$. Widerspruch!

Bemerkungen:

- Dies korrespondiert mit der Haarschen Bedingung.
- Bestimmen der Koeffizienten direkt aus dem linearen Gleichungssystem i.A. nicht sinnvoll.

1.2.1 Wahl spezieller Basen

1. Lagrange-Interpolation

$$\begin{aligned} \varphi_j(x_i) &= \delta_{ij} \\ \varphi_k(x) &= \frac{\prod_{k \neq j}(x - x_j)}{\prod_{k \neq j}(x_k - x_j)} \qquad k = 0(1)n \end{aligned}$$

Folgerung: A=E

2. Newton-Interpolation

$$\begin{aligned} \varphi_j(x_i) &= 0 \text{ für } i < j \\ \varphi_j(x) &= \prod_{k<j}(x - x_k) \end{aligned}$$

A ist Dreiecksmatrix

$$p(x) = c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0) \cdot (x - x_1) + \ldots + c_n \cdot \prod_{k=0}^{n-1}(x - x_k)$$

Darstellung (Berechnung) der c_j mittels Steigungen möglich. Dazu:

$$\begin{aligned} f[x_i] &:= f(x_i) = f_i \\ f[x_i, \ldots, x_k] &:= \frac{f[x_{i+1}, \ldots, x_k] - f[x_i, \ldots, x_{k-1}]}{x_k - x_i} \\ c_k &= f[x_0, \ldots, x_k] \qquad k = 0(1)n \end{aligned}$$

Beispiel:

x_k	-1	0	1	2
f_k	-4	-4	0	14

Ergebnis:

$$p(x) = -4 + 0 + 2 \cdot (x+1) \cdot x + x \cdot (x-1) \cdot (x+1)$$

Begründung des Steigungsschemas über Rekursion (Beweis siehe *Numerische Mathematik*, R. Plato, S.8): Sei $p(f|x_i, \ldots, x_{i+k}) \in P_k$ mit

$$p(f|x_i, \ldots, x_{i+k})(x_j) = f_j$$

für $j = i, \ldots, i+k$.

Lemma: Es sei $q \in P_{k+1}$ definiert durch

$$q(x) := \frac{(x_i - x) \cdot p(f|x_{i+1}, \ldots, x_{i+k+1}) - (x_{i+k+1} - x) \cdot p(f|x_i, \ldots, x_{i+k})}{x_i - x_{i+k+1}}$$

Dann gilt:

$$q = p(f|x_i, \ldots, x_{i+k+1})$$

Beweis:

$$\begin{aligned} q(x_j) &= \frac{(x_i - x_j) \cdot f_j - (x_{i+k+1} - x_j) \cdot f_j}{x_i - x_{i+k+1}} = f_j \qquad j = i+1, \ldots, k \\ q(x_i) &= \frac{(x_i - x_i) \cdot p(\ldots) - (x_{i+k+1} - x_i) \cdot f_i}{x_i - x_{i+k+1}} = f_i \end{aligned}$$

analog für x_{i+k+1}

Nutzung der Rekursion führt zu Neville-Schema. Berechnung beim Newton-Schema lässt sich analog zum Horner-Schema realisieren:

$$p(x) = c_0 + (x - x_0) \cdot (c_1 + (x - x_1) \cdot (c_2 \ldots (c_n)))$$

Start mit c_n

1.2.2 Interpolationsfehler

Sei f hinreichend glatt, $x_k \in [a, b]$, $f_i := f(x_i)$ für $i = 0(1)n$.

Satz: Es sei f (n+1)mal differenzierbar. Dann gilt:

$$f(x) = p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \prod_{k=0}^{n} (x - x_k)$$

Beweis:

- Hilfsfunktion

$$g(t) := f(t) - p_n(t) - (f(x) - p_n(x)) \cdot \frac{(t-x_0)\cdot(t-x_1)\ldots(t-x_n)}{(x-x_0)\cdot(x-x_1)\ldots(x-x_n)}$$

für festes $x \neq x_k$ mit $x \in [a,b]$. Offenbar gilt:

$$\begin{aligned} g(x_k) &= 0 \qquad k = 0(1)n \\ g(x) &= 0 \end{aligned}$$

g besitzt also (n+2) Nullstellen. Satz von Rolle: Es existiert $\xi \in (a,b)$ mit $g^{(n+1)}(\xi) = 0$. Folglich gilt:

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - \frac{f(x) - p_n(x)}{(x-x_0)\ldots(x-x_n)} \cdot (n+1)!$$

Bemerkungen:

- Falls $f \in C^\infty$ und alle Ableitungen $f^{(n)}$ von f sind auf $[a,b]$ gleichmäßig beschränkt. Dann

$$\|f - p_n(x)\|_{C[a,b]} \to 0 \qquad (n \to \infty)$$

Dabei ist p_n das Interpolationspolynom vom Grad n mit

$$\begin{aligned} x_i &= a + i \cdot \frac{b-a}{n} \qquad i = 0(1)n \\ f_i &= f(x_i) \end{aligned}$$

- Aber: Beispiel von Runge

$$\begin{aligned} f(x) &= \frac{1}{1+25x^2} \qquad x \in [-1,1]; x_k = -1 + k \cdot \frac{2}{n} \\ p_n(x) &\not\to f \end{aligned}$$

Ausweg: Spline-Interpolation.

- Analyse des Fehlerterms bei der Polynominterpolation

$$q_n(x) = \prod_{j=0}^{n} (x - x_j)$$

Idee: Interpolationsstellen nicht a priori festlegen, sondern $\{x_j\}_{j=0}^n \subset [a,b]$ so wählen, dass

$$\max_{x \in [a,b]} |q_n(x)| \to \min$$

1.2.3 Tschebysheff-Polynome

Referenzintervall $[-1,1]$

$$\begin{aligned} T_n(x) &:= \cos(n \arccos x) \qquad n = 0(1)... \\ T_0(x) &= 1 \\ T_1(x) &= x \end{aligned}$$

- Ist T_n ein Polynom für $n > 1$? Additionstheoreme: ($n \in \mathbb{N}$):

$$\begin{aligned} \cos((n \pm 1)x) &= \cos x \cdot \cos(nx) \mp \sin x \cdot \sin(nx) \\ \cos((n-1)x) &= 2 \cdot \cos x \cdot \cos(nx) - \cos((n+1)x) \end{aligned}$$

Mit Definition der Polynome ergibt sich:

$$\begin{aligned} T_{n+1}(x) &= \cos((n+1)\arccos x) \\ &= 2 \cdot \cos(\arccos x) \cdot \cos(n \cdot \arccos x) - \cos((n-1)\arccos x) \\ &= 2x \cdot T_n(x) - T_{n-1}(x) \end{aligned}$$

Mit T_0, T_1 ist jedes T_n ein Polynom n-ten Grades mit $T_n(x) = 2^{n-1} \cdot x^n + \ldots$ für $n \geq 1$. Aus Definition folgt auch: $|T_n(x)| \leq 1$ für alle $x \in [-1, 1]$.

- Nullstellen von T_n:

$$\begin{aligned} T_n(x_j) = 0 \quad &\Leftrightarrow \quad n \cdot \arccos x_j = \frac{\pi}{2} + j \cdot \pi \qquad j = 0(1)(n-1) \\ &\Leftrightarrow \quad x_j = \cos\left(\frac{\pi + 2j\pi}{2n}\right) \end{aligned}$$

Übertragung auf allgemeines Intervall $[a, b]$ durch Transformation:

$$\begin{aligned} x &= \frac{a+b}{2} + \xi \cdot \frac{b-a}{2} \qquad \xi \in [-1, 1] \\ \xi &= \frac{2x}{b-a} - \frac{a+b}{b-a} \end{aligned}$$

$\hat{T}_n$ ist transformiertes Tschebysheff-Polynom.

$$q(x) = \prod_{j=0}^{n} (x - x_j)$$

mit $x_j = x(\xi_j)$ mit ξ_j als Nullstellen der Tschebysheff-Polynome. $\hat{T}_n$ hat führenden Koeffizien

$$\left(\frac{2}{b-a}\right)^n \cdot 2^{n-1}$$

und

$$|q_n(x)| \leq \left(\frac{b-a}{2}\right)^n$$

Satz Sei $p_n \in \mathcal{P}_n$ ein Polynom

$$p_n(x) = a_n \cdot x^n + \ldots a_0 \qquad a_n \neq 0$$

Dann existiert $\hat{x} \in [-1, 1]$ mit

$$|p_n(\hat{x})| \geq \frac{a_n}{2^{n-1}} = a_n \cdot 2^{1-n}$$

Beweis:

- Sei $p_n \in \mathcal{P}_n$ mit $a_n = 2^{n-1}$. Annahme: $\forall x \in [-1, 1]$ gilt: $|p_n(x)| < 1$. Mit $T_n \in \mathcal{P}_n$ gilt $T_n - p_n \in \mathcal{P}_n$. Wegen $a_n = 2^{n-1}$ ist $T_n - p_n \in \mathcal{P}_l$ mit $l < n$. Betrachte

$$\overline{x_j} = \cos\frac{j\pi}{n} \qquad j = 0(1)n$$

Dann gilt:

$$\begin{aligned} T_n(\overline{x_{2j}}) &= 1 \qquad T_n(\overline{x_{2j+1}}) = -1 \\ p_n(\overline{x_{2j}}) &< 1 \qquad p_n(\overline{x_{2j+1}}) > -1 \end{aligned}$$

Somit:

$$\begin{aligned} (T_n - p_n)(\overline{x_{2j}}) &> 0 \\ (T_n - p_n)(\overline{x_{2j+1}}) &< 0 \end{aligned}$$

Also wechselt $(T_n - p_n)$ zwischen den (n+1) Punkten $\overline{x_j}$ jeweils das Vorzeichen. $(T_n - p_n)$ besitzt n verschiedene Nullstellen, $T_n - p_n \neq 0$. Damit $T_n - p_n \notin P_{n-1}$. Widerspruch!

- Der Fall beliebiger Polynome wird mit Normierung behandelt.

1.2.4 Interpolationsbedingungen höherer Ordnung

(hier: 1. Ordnung)

- Aufgabenstellung:
$$y(x_i) = f_i \qquad y'(x_i) = d_i (=: f_i') \qquad i = 0(1)n$$

- Seien φ_j die Lagrange-Basisfunktionen. Wir führen die Hermite-Polynome wie folgt ein:
$$\begin{aligned} H_j(x) &:= (1 - 2 \cdot (x - x_j) \cdot \varphi_j'(x_j)) \cdot \varphi_j^2(x) \\ \hat{H}_j(x) &:= (x - x_j) \cdot \varphi_j^2(x) \end{aligned}$$

  Polynome vom maximalen Grad (2n+1)

Lemma Es sei $x_0 < \ldots < x_n$ gegeben und H_j, $\hat{H}_j$ die zugehörigen Hermite-Polynome. Dann gilt:
$$\begin{aligned} H_j(x_i) &= \delta ij \\ H_j'(x_i) &= 0 \\ \hat{H}_j(x_i) &= 0 \\ \hat{H}_j'(x_i) &= \delta ij \end{aligned}$$

für $i = 0(1)n$ und $j = 0(1)n$.

Beweis:
$$\begin{aligned} H_j(x_i) &= \begin{cases} \varphi_j^2(x_j) = 1 & i = j \\ \mu_{ij} \cdot \varphi_j^2(x_i) = 0 & i \neq j \end{cases} \\ H_j'(x) &= -2\varphi_j'(x_j) \cdot \varphi_j^2(x) + (1 - 2(x - x_j) \cdot \varphi_j'(x_j)) \cdot 2\varphi_j(x) \cdot \varphi_j'(x) \\ H_j'(x_j) &= -2\varphi_j'(x_j) + 2\varphi_j'(x_j) = 0 \\ H_j'(x_i) &= 0 \end{aligned}$$

Folgerung: Auf der Grundlage des Lemmas gilt:
$$y(x) = \sum_{j=0}^{n} f_j \cdot H_j(x) + \sum_{j=0}^{n} f_j' \cdot \hat{H}_j(x)$$

Satz Die Interpolationsaufgabe mit disjunkten Stützstellen ist eindeutig lösbar mit einem Polynom y mit Maximalgrad (2n+1). Es sei ferner $f_i = f(x_i)$, $f_i' = f'(x_i)$ für $i = 0(1)n$ eine Funktion die (2n+2) mal stetig differenzierbar ist. Dann gibt es ein $\xi = \xi(x) \in \text{conv}\,\{x_0, \ldots, x_n\}$ mit
$$f(x) = y(x) + \frac{f^{(n+2)}(\xi)}{(2n+2)!} \cdot \prod_{i=0}^{n} (x - x_i)^2$$

Beweis:

- Mit der Hilfsfunktion
$$g(t) := f(t) - y(t) - (f(x) - y(x)) \cdot \frac{(t - x_0)^2 \cdots (t - x_n)^2}{(x - x_0)^2 \cdots (x - x_n)^2}$$

  analog zum Interpolationsfehler bei Polynominterpolation.

1.3 Spline-Interpolation

- Nachteil der Polynominterpolation: hohe Oszillation. Alternative: stückweise Polynomansätze

$$s_\triangle = \begin{cases} s_1 & x \in [x_0, x_1] \\ s_2 & x \in [x_1, x_2] \end{cases}$$

$s_\triangle \in C^{k-1,1}$ realisierbar; mit $C^{k,\alpha}$: k mal stetig differenzierbar, k-te Ableitung Hölderstetig mit Konstante α:

$$|y(x) - y(t)| \leq K \cdot |x - t|^\alpha$$

- Ansatz für kubische Splines:

$$\begin{aligned} s_i(x) &= \frac{x - x_{i-1}}{h_i} \cdot f_i + \frac{x_i - x}{h_i} \cdot f_{i-1} + h_i^2 \cdot \left(\sigma_i \left(\frac{(x - x_{i-1})^3}{h_i^3} - \frac{x - x_{i-1}}{h_i}\right) + \right. \\ &\quad \left. + \sigma_{i-1} \cdot \left(\frac{(x_i - x)^3}{h_i^3} - \frac{x_i - x}{h_i}\right)\right) \end{aligned}$$

mit $h_i = x_i - x_{i-1}$, $s_\triangle(x) = s_i(x)$, $x \in \Omega_i = [x_{i-1}, x_i]$ für $i = 1(1)n$.

- Nach Ansatz:

$$s_\triangle(x_i) = f_i \qquad i = 1(1)n$$

- Bestimmung der σ_i aus der Bedingung der Glattheit der Ableitung:

$$\begin{aligned} s_i'(x_i - 0) &\stackrel{!}{=} s_{i+1}'(x_i + 0) \qquad i = 1(1)(n-1) \\ s_i'(x_i - 0) &= \frac{f_i - f_{i-1}}{h_i} + h_i \cdot \sigma_{i-1} + 2h_i \cdot \sigma_i \\ s_{i+1}'(x_i + 0) &= \frac{f_{i+1} - f_i}{h_{i+1}} - h_{i+1} \cdot \sigma_{i+1} - 2h_{i+1} \cdot \sigma_i \end{aligned}$$

Ferner gilt („per Ansatz“):

$$s_i''(x_i - 0) = s_{i+1}''(x_i + 0) = 6\sigma_i$$

Aus den obigen Bedingungen ergibt sich folgendes lineares Gleichungssystem:

$$h_i \cdot \sigma_{i-1} + 2(h_i + h_{i+1}) \cdot \sigma_i + h_{i+1} \cdot \sigma_{i+1} = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}$$

für $i = 1(1)(n-1)$.

- Diskussion der Freiheitsgrade:

$$s_i \in \mathcal{P}_3 \qquad i = 1(1)n$$

also 4n Paramater abzüglich Interpolationsbedingungen (2n Bedingungen) und Glattheitsforderung (2(n-1) Bedingungen), also 2 freie Parameter. Zusatzbedingungen möglich:

1. natürliche Splines:
$$s''(x_0 + 0) = 0 = s''(x_n - 0)$$
2. eingespannter Spline:
$$s'(x_0 + 0) = f_0' \qquad s'(x_n - 0) = f_n'$$
3. „not a knot“-Spline:
$$s_1^{(3)}(x_1 - 0) = s_2^{(3)}(x_1 + 0) \qquad s_{n-1}^{(3)}(x_{n-1} - 0) = s_n^{(3)}(x_{n-1} + 0)$$

Damit eindeutig lösbares lineares Gleichungssystem mit (n+1) Unbekannten σ_i.

- Zur Konvergenz von kubischen Splines (bei Verfeinerung- bzw. Approximationsgüte bei fixierten Diskretisierungsniveau):
  - Sei $f_0 \in C^4[a,b]$ ($\in C^{3,1}[a,b]$). Dann existiert eine Konstante $c > 0$ derart, dass für beliebiges Gitter $\{x_i\}_{i=0}^n$ $a = x_0 < x_1 < \ldots < x_n = b$ gilt:

$$|s(x) - f(x)| \leq c \cdot h^4 \qquad h = \max_i h_i$$

  - Beweisskizze: (Im wesentlichen fur für innere Intervalle, die Randintervalle müssen getrennt untersucht werden.) In $\Omega_i = [x_{i-1}, x_i]$ genügt s_i der Differentialgleichung

$$\begin{aligned} s_i''(x) &= 6\sigma_i \cdot \frac{x - x_{i-1}}{h_i} + 6\sigma_{i-1} \cdot \frac{x_i - x}{h_i} \qquad (x \in \Omega_i) \\ s_i(x_{i-1}) &= f_{i-1} \qquad s_i(x_i) = f_i \end{aligned}$$

  (Randwertaufgabe). Vergleich der rechten Seite mit $f''(x)$. Bezeichnung:

$$\delta_i := 6\sigma_i - f''(x_i)$$

  (Fehler der 2. Ableitung in Stützstellen). Aus der Gleichung für das lineare Gleichungssystem folgt:

$$\begin{aligned} h_i \cdot \delta_{i-1} + 2(h_i + h_{i+1}) \cdot \delta_i + h_{i+1} \cdot \delta_{i+1} &= 6\frac{f_{i+1} - f_i}{h_{i+1}} - 6\frac{f_i - f_{i-1}}{h_i} - h_i \cdot f_{i-1}'' \\ &\quad -2(h_i + h_{i-1}) f_i'' - h_{i+1} \cdot f_{i+1}'' \end{aligned}$$

  Taylorentwicklung (an x_i bzgl. x_{i-1}):

$$\begin{aligned} f_{i-1} &= f_i - f_i' \cdot h_i + \frac{1}{2} h_i^2 f_i'' - \frac{1}{6} f^{(3)}(\xi_i) \cdot h_i^3 \\ 6\frac{f_i - f_{i-1}}{h_i} &= 6f_i' - 3f_i'' \cdot h_i + f^{(3)}\xi_i) \cdot h_i^2 \end{aligned}$$

  mit $x_i \in (x_{i-1}, x_i)$. Analog Taylor $x_i \to x_{i+1}$:

$$6\frac{f_{i+1} f_i}{h_{i+1}} = 6f_i' + 3f_i'' + f^{(3)}(\eta_i) \cdot h_{i+1}^2$$

  mit $\eta_i \in (x_i, x_{i+1})$. Damit folgt: (mit $\lambda_i \in \Omega_i, \varrho_i \in \Omega_{i+1}$)

$$\begin{aligned} h_i \cdot \delta_{i-1} + 2(h_i + h_{i+1}) \cdot \delta_i + h_{i+1} \cdot \delta_{i+1} &= -f^{(3)}(\xi_i) \cdot h_i^2 + f^{(3)}(\eta_i) \cdot h_{i+1}^2 \\ &\quad h_i \cdot \underbrace{(f_i'' - f_{i-1}'')}_{f^{(3)}(\lambda_i) \cdot h_i} + h_{i+1} \cdot \underbrace{(f_i'' - f_{i+1}'')}_{-f^{(3)}(\varrho_i) \cdot h_{i+1}} \\ &= h_i^2 \cdot (f^{(3)}(\lambda_i) - f^{(3)}(\xi_i)) \\ &\quad + h_{i+1}^2 \cdot (f^{(3)}(\eta_i) - f^{(3)}(\varrho_i)) \\ &=: r_i \end{aligned}$$

  Es gilt:

$$|r_i| \leq L \cdot h \cdot (h_i^2 + h_{i+1}^2)$$

  für $f \in C^{3,1}[a,b]$. Andererseits gilt für ein i mit $|\delta_i| = \delta = \max |\delta_i|$:

$$\begin{aligned} |r_i| &= |h_i \cdot \sigma_{i-1} + 2(h_i + h_{i+1}) \cdot \sigma_i + h_{i+1} \cdot \sigma_{i+1}| \\ &\geq 2(h_i + h_{i+1}) \cdot \delta - h_{i+1} \cdot \delta - h_i \cdot \delta \\ &= (h_i + h_{i+1}) \cdot \delta \end{aligned}$$

  Damit $|\delta_i| \leq |r_i| \leq 2Lh^2$. Abschätzung für $x \in \Omega_i$:

$$s'' - f'' = \frac{x - x_{i-1}}{h_i} \cdot (\delta_i + \underbrace{f_i'' - f''(x)}_{f^{(3)}(\alpha_i)(x_i - x)}) + \frac{x_i - x}{h_i} \cdot (\delta_{i+1} + \underbrace{f_{i-1}'' - f''(x)}_{-f^{(3)}(\beta_i)(x - x_i)})$$

Die Hilfsfunktion $w = s - f$ genügt der Randwertaufgabe

$$\begin{aligned} w'' &= p_i(x) \\ w(x_{i-1}) &= w(x_i) = 0 \end{aligned}$$

Dabei gilt:

$$\begin{aligned} |p_i(x)| &\leq |\delta_i| + |\delta_{i+1}| + |f^{(3)}(\alpha_i) - f^{(3)}(\beta_i)| \cdot \frac{(x_i - x) \cdot (x - x_{i-1})}{h_i} \\ &\leq 2\delta + Lh^2 \leq c_1 \cdot h^2 \end{aligned}$$

Also $|w''| \leq c_1 \cdot h^2$.

Lemma Es sei

$$\begin{aligned} -w'' &\leq -v'' \\ w|_\Gamma &\leq v|_\Gamma \qquad \Gamma = \partial(\Omega) \end{aligned}$$

in Ω. Dann gilt für alle $x \in \Omega$:

$$w(x) \leq v(x)$$

(Maximumprinzip). Somit:

$$\begin{aligned} v^{+''} &= -c_1 \cdot h^2 \\ v(x_{i-1}) &= v(x_i) = 0 \end{aligned}$$

mit

$$v^{\pm} = \mp\frac{1}{2} c_1 \cdot h^2 \cdot (x - x_{i-1}) \cdot (x_i - x)$$

Mit dem Lemma gilt:

$$\begin{aligned} -\frac{1}{2} c_1 \cdot h^2 (x_i - x) \cdot (x - x_{i-1}) &\leq w(x) \leq \frac{1}{2} h^2 \cdot c_1 \cdot (x_i - x) \cdot (x - x_{i-1}) \\ |w(x)| &\leq \frac{c_1}{8} \cdot h^4 \end{aligned}$$

für alle $x \in \Omega$.

– Bemerkung: Bei kubischen Splines gilt:

$$\|s^{(k)} - f^{(k)}\| \leq c \cdot h^{4-k} \qquad k = 0, 1, 2$$

mit

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

Satz Der natürliche Interpolationsspline minimiert das Funktional

$$F[y] := \frac{1}{2} \int_a^b [y''(x)]^2 \, dx$$

bzgl.

$$y \in K := \{y \in C^2(\bar{\Omega}) : y|_{\Omega_i} \in C^4(\bar{\Omega}_i), y(x_i) = f_i\}$$

Beweis:

- Sei $v \in K$ beliebig.

$$\begin{aligned} F[v] &= \frac{1}{2} \int_a^b (y^{(2)} + (v^{(2)} - y^{(2)})) \, dx \\ &= \frac{1}{2} \int_a^b (y^{(2)})^2 \, dx + \frac{1}{2} \underbrace{(v^{(2)} - y^{(2)})^2}_{\geq 0} \, dx + \int_a^b y^{(2)} \cdot (v^{(2)} - y^{(2)}) \, dx \\ &\geq F[y] + \underbrace{\int_a^b y^{(2)} \cdot (v^{(2)} - y^{(2)}) \, dx}_{z.Z.: \geq 0} \end{aligned}$$

Da $v, y \in K$ ist

$$w := v - y \in K_0 := \{v \in C^2(\Omega); v|_{\Omega_i} \in C^4(\Omega_i), v(x_i) = 0\}$$

Damit:

$$\begin{aligned}
\int_\Omega y^{(2)} \cdot w^{(2)}\, dx &= \sum_{i=1}^{n} \left(y^{(2)} \cdot w'|_{x_{i-1}}^{x_i} - \int_{\Omega_i} y^{(3)} \cdot w'\, dx \right) \\
&= \underbrace{y^{(2)} \cdot w'|_a^b}_{0} - \sum_{i=1}^{n} \int_{\Omega_i} y^{(3)} \cdot w'\, dx \\
&= \underbrace{-\sum_{i=1}^{n} y^{(3)} \cdot w|_{x_{i-1}}^{x_i}}_{0} + \sum_{i=1}^{n} \int_{\Omega_i} \underbrace{y^{(4)}}_{0, y \in \mathcal{P}_3(\Omega_i)} \cdot w\, dx \\
&= 0
\end{aligned}$$

Der zweite Umformungsschritt folgt, da $y^{(2)}(a) = y^{(2)}(b) = 0$ (natürlicher Spline). Damit: $F[v] \geq F[y]$ für alle $v \in K$.

Bemerkung:

- Abänderung z.B. auf eingespannten Spline möglich:

$$\begin{aligned}
\tilde{K} &= \{v \in K | v'(a) = f'_a, v'(b) = f'_b\} \\
\tilde{K}_0 &= \{v \in K_0 | v'(a) = v'(b) = 0\}
\end{aligned}$$

  Spline realisiert minimale Biegeenergie für $\tilde{K} \subset K$

1.3.1 Bézier-Kurven

- Konstruktion 2-mal stetig differenzierbarer Kurven durch stückweise kubische Polynome, Interpolationspunkte $(z_i) = (x_i, y_i) \in \mathbb{R}^2$ für $i = 0(1)n$. Gesucht:

$$v_i(t) = \begin{pmatrix} x_i(t) \\ y_i(t) \end{pmatrix} \qquad t \in [0, 1]$$

  Ansatz:

$$v_i(t) = z_{i-1} \cdot (1-t)^3 + a_i \cdot 3 \cdot (1-t)^2 \cdot t + b_i \cdot 3 \cdot (1-t) \cdot t^2 + z_i \cdot t^3$$

  unter Verwendung der Bernstein-Polynome

$$P_{k,n}(t) = \binom{n}{k} \cdot t^k \cdot (1-t)^{n-k}$$

  Es gilt:

$$\begin{aligned}
\dot{v}_i(t) &= -3z_{i-1} \cdot (1-t)^2 - 6 \cdot a_i \cdot (1-t) \cdot t + 3a_i \cdot (1-t)^2 + \\
&\quad +6b_i \cdot t - 9b_i \cdot t^2 + 3z_i \cdot t^2 \\
\ddot{v}_i(t) &= 6z_i \cdot (1-t) + 12a_i \cdot t - 6a_i - 6a_i \cdot (1-t) + 6b_i - 18b_i \cdot t + 6z_i \cdot t
\end{aligned}$$

  Damit:

$$\begin{aligned}
v_i(0) &= z_{i-1} \qquad v_i(1) = z_i \\
\dot{v}_i(0+0) &= 3(a_i - z_{i-1}) \qquad \dot{v}_i(1-0) = 3(z_i - b_i) \\
\ddot{v}_i(0+0) &= 6b_i - 12a_i + 6z_{i-1} \\
\ddot{v}_i(1-0) &= 6z_i + 6a_i - 12b_i
\end{aligned}$$

- Forderung:

$$\begin{aligned} \dot{v}_i(1-0) &= \dot{v}_{i+1}(0+0) \\ \ddot{v}_i(1-0) &= \ddot{v}_{i+1}(0+0) \end{aligned}$$

Man erhält:

$$\begin{aligned} z_i - b_i &= a_{i+1} - z_i \\ \Rightarrow z_i &= \frac{b_i + a_{i+1}}{2} \end{aligned}$$

Außerdem:

$$\begin{aligned} -a_i + 2b_i &= 2a_{i+1} - b_{i+1} =: c_i \\ 2a_i - b_i &= c_{i-1} \\ \Rightarrow a_i &= \frac{1}{3}c_i + \frac{2}{3} \cdot c_{i-1} \\ b_i &= \frac{1}{3}c_{i+1} + \frac{2}{3} \cdot c_i \end{aligned}$$

- Bemerkung: Mit Hermitepolynomen

$$\begin{aligned} H_0(t) &= (1+3t) \cdot (1-t)^2 \\ H_1(t) &= (3-2t) \cdot t^2 \\ \hat{H}_0(t) &= t \cdot (1-t)^2 \\ \hat{H}_1(t) &= (t-1) \cdot t^2 \end{aligned}$$

ergibt sich:

$$v_i(t) = z_{i-1} \cdot H_0(t) + 3(a_i - z_{i-1}) \cdot \hat{H}_0(t) + 3(z_i - b_i) \cdot \hat{H}_1(t) + z_i \cdot H_1(t)$$

- numerisch effiziente und stabile Approximation durch Polygonzüge mittels des Casteljeu-Algorithmus

1.4 Approximationsaufgaben

1.4.1 Methode der kleinsten Quadrate

- Ausgangssituation: Meßreihe (t_i, f_i) für $i = 1(1)m$. Ansatz:

$$y = y(t) = \sum_{i=1}^{n} x_i \cdot \varphi_i(t) \qquad (n << m)$$

- Interpolation führt auf lineares Gleichungssystem

$$A \cdot x = b$$

mit $A = (a_{ij})$, $a_{ij} = \varphi_j(t_i)$ und $b = (f_1, \ldots, f_m)$. Im Allgemeinen ist dieses Gleichungssystem nicht lösbar (überbestimmt).

- Idee: Minimierung des „Fehlerquadrates“

$$\begin{aligned} \triangle_i &= \left| f_i - \sum_{j=1}^{n} x_j \cdot \varphi_j(t_i) \right| \\ \sum_{j=1}^{m} \triangle_i^2 &\rightarrow \min \\ \Leftrightarrow \|A \cdot x - b\|_2^2 =: D(x) &\rightarrow \min \qquad (1) \end{aligned}$$

Satz Sei $A \in \mathbb{R}^{m \times n}$ mit Rang A = n ($n < m$), $b \in \mathbb{R}^m$. Dann sind äquivalent:

1. $\hat{x} \in \mathbb{R}^n$ löst (1), d.h.
$$\forall x \in \mathbb{R}^n : \|A\hat{x} - b\|_2^2 \leq \|Ax - b\|_2^2$$

2. $\hat{x} \in \mathbb{R}^n$ löst das lineare Gleichungssystem
$$A^T \cdot A \cdot \hat{x} = A^T \cdot b$$
(Normalengleichungssystem)

Beweis:

1. Sei $A \in \mathbb{R}^{m \times n}$ mit $n < m$, Rang A =n. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{m \times m}$ ($Q^T = Q^{-1}$) mit
$$Q^T \cdot A = \begin{pmatrix} R \\ 0 \end{pmatrix}$$
mit $R \in \mathbb{R}^{n \times n}$ als obere Dreiecksmatrix mit $v_{ii} \neq 0$ für alle i. Dann gilt:
$$\begin{aligned} \|Ax - b\|_2^2 &= \|Q^T \cdot (Ax - b)\|_2^2 \\ &= \|Q^T \cdot Ax - Q^T \cdot b\|_2^2 \\ &= \| \begin{pmatrix} R \\ 0 \end{pmatrix} \cdot x - \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} \|_2^2 \\ &= \|Rx - \tilde{b}_1\|_2^2 + \|\tilde{b}_2\|_2^2 \end{aligned}$$
mit
$$Q^T \cdot b = \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} \qquad \tilde{b}_1 \in \mathbb{R}^n, \tilde{b}_2 \in \mathbb{R}^{m-n}$$
$\hat{x}$ als Lösung von $R \cdot x = \tilde{b}_1$ minimiert (1). „Minimalabstand“ zur Lösbarkeit: $\|\tilde{b}_2\|_2^2$

2. Sei
$$A = Q \cdot \begin{pmatrix} R \\ 0 \end{pmatrix}$$
Dann existiert R^{-1}.
$$\begin{aligned} A^T \cdot A &= \begin{pmatrix} R^T & 0 \end{pmatrix} \cdot Q^T \cdot Q \cdot \begin{pmatrix} R \\ 0 \end{pmatrix} = R^T \cdot R \\ A^T \cdot b &= \begin{pmatrix} R^T & 0 \end{pmatrix} \cdot Q^T \cdot b = R^T \cdot \tilde{b}_1 \end{aligned}$$
Folglich:
$$\begin{aligned} A^T \cdot A \cdot x = A^T \cdot b \quad &\Leftrightarrow \quad R^T \cdot R \cdot x = R^T \cdot \tilde{b}_1 \\ &\Leftrightarrow \quad R \cdot x = \tilde{b}_1 \\ &\Leftrightarrow \quad \text{x löst (1)} \end{aligned}$$

Bemerkung:
$$\begin{aligned} A^T \cdot A &= (a_{ij}^*) \\ a_{ij}^* &= \sum_{l=1}^{m} \varphi_j(t_l) \cdot \varphi_i(t_l) \\ A^T \cdot b &= (a_i^b) \\ a_i^b &= \sum_{l=1}^{m} \varphi_i(t_l) \cdot f_l \end{aligned}$$

Beispiel:

- Gegebene Messwerte:

1	2	3	4
8	11	9	10

Also m=4, n=2.

$$\begin{aligned} g(t) &= x_1 \cdot t + x_2 \\ A^T \cdot A &= \begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix} \\ A^T \cdot b &= \begin{pmatrix} 97 \\ 38 \end{pmatrix} \\ \Rightarrow \hat{x} &= \begin{pmatrix} 0,4 \\ 8,5 \end{pmatrix} \\ \Rightarrow g(t) &= 0,4t + 8,5 \end{aligned}$$

Bemerkung: Notwendige Optimierungsbedingung:

$$\begin{aligned} \|Ax - b\|_2^2 = D(x) &\rightarrow \min \\ \nabla D(x) &= 0 = 2A^T \cdot (Ax - b) \end{aligned}$$

1.4.2 T-Approximation

- Alternative Approximierung:

$$\|Ax - b\|_\infty = \max_{1 \le i \le m} \left| \sum_{j=1}^{n} x_j \cdot \varphi_j(t_i) - f_i \right| \rightarrow \min \qquad (2)$$

zusätzliche Variable $\sigma \in \mathbb{R}$. Damit ist (2) äquivalent zu $\sigma \rightarrow \min$ bei

$$\forall i : \left| \sum_{j=1}^{n} x_j \cdot \varphi_j(t_i) - f_i \right| \le \sigma$$

Dies ist eine lineare Optimierungsaufgabe mit freien Variablen. Entweder Transformation bzw. Einsatz spezieller Techniken.

- Beispiel:

$$\begin{aligned} -\sigma &\le x_1 + x_2 - 8 \le \sigma \\ -\sigma &\le 2x_1 + x_2 - 11 \le \sigma \\ -\sigma &\le 3x_1 + x_2 - 9 \le \sigma \\ -\sigma &\le 4x_1 + x_2 - 10 \le \sigma \end{aligned}$$

Einschub Vektornormen:

- Normen im $\mathbb{R}^n$:

$$\begin{aligned} \|x\|_2^2 &= \sum_{i=1}^{n} x_i^2 \\ \|x\|_1 &= \sum_{i=1}^{n} |x_i| \\ \|x\|_\infty &= \max_i |x_i| \\ \|x\|_p^p &= \sum_{i=1}^{n} |x_i|^p \qquad 1 \le p < \infty \end{aligned}$$

- Normierter Raum: X linearer Raum mit assoziierten Zahlenkörper und Norm $\|.\| : X \to \mathbb{R}_+$ mit

  1. $\|x\| \geq 0$ für $x \neq 0$
  2. $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$
  3. $\|x + z\| \leq \|x + y\| + \|y + z\|$

  zugehörige Metrik:
  $$d(x, y) = \|x - y\|$$

- Hilbertraum: Raum mit Skalarprodukt $< ., . >: X \times X \to \mathbb{R}$ mit

  1. $< x, x >\geq 0, < x, x >= 0 \Leftrightarrow x = 0$
  2. $< x, y >=< y, x >$
  3. $< \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2, y >= \alpha_1 \cdot < x_1, y > +\alpha_2 \cdot < x_2, y >$

- Dann Norm:
  $$\|.\|_H^2 =< x, x >$$

1.4.3 Ein allgemeines Konzept

- $T : Y \to U$, Y,U (lineare) normierte Räume. Zu gegebenem $f \in U$ ist ein $y \in Y$ zu bestimmen mit
  $$\|Ty - f\| \leq \|Tv - f\| \qquad (2)$$
  für alle $v \in Y$. Falls Y,U Hilberträume: Es existiert eine Lösung y (mit $Ty \in U$) von (2) als Minimum von
  $$J(v) := \frac{1}{2}\|Tv - f\|^2 = \frac{1}{2} < Tv - f, Tv - f > \to \min_{v \in Y} \qquad (1)$$

- Es gilt:
  $$\begin{aligned} y \in Y \text{ löst } (1) \quad &\Leftrightarrow \quad < Ty - f, Tv - f >\geq 0 \qquad \forall v \in Y \qquad (???) \\ &\Leftrightarrow \quad T^* \cdot (Ty - f) = 0 \end{aligned}$$
  (Gaußsche Normalengleichung)

- „klassisches“ Approximationsproblem:
  $$\begin{aligned} [Ty] &:= [Ty]_i = (y(t_i)) \qquad i = 1(1)m \\ \left(\sum x_j T\varphi_j - f_i \middle| T\varphi_k\right) &= 0 \qquad k = 1(1)n \\ \Leftrightarrow \sum_{j=1}^{n} x_j \cdot (T\varphi_j | T\varphi_k) &= (f | T\varphi_k) \qquad k = 1(1)n \end{aligned}$$

- Anderes Beispiel:
  $$\begin{aligned} Y &= \mathcal{P}_n[-1, 1] \qquad Ty = y \\ U &= L_2(-1, 1) := \left\{ y \middle| \int_{-1}^{1} y^2 dx < \infty \right\} \end{aligned}$$
  für U:
  $$< u, v >_2 = \int_{-1}^{1} u(x) \cdot v(x)\, dx \qquad (\forall u, v \in C[a, b])$$
  Damit:
  $$\begin{aligned} U &= C[a, b] \cup \{v | \exists \{v_n\} \subset C[a, b] : \|v_n - v\|_2^2 \to 0 (n \to \infty)\} \\ &=: \overline{C[a, b]}^{L_2} \end{aligned}$$

- Falls eine Orthonormalbasis $\{p_i\}_{i=0}^n$ von Y bekannt ist, so lässt sich eine Lösung der Bestapproximation (1) sofort angeben:

$$\hat{y}(f) = \sum_{i=0}^{n} < p_i, f > p_i$$

Anmerkung: $\{p_i\}_{i=0}^n$ ist Orthonormalbasis von Y, falls

$$< p_i, p_j >= \delta_{ij} = \int_{-1}^{1} p_i(x) \cdot p_j(x)\, dx$$

- Ein (geeignetes) Orthogonalsystem in $L_2(-1,1) \cap \mathcal{P}_n[-1,1]$: Legendre-Polynome. Rekursive Definition:

$$P_{j+1}(x) = \frac{2j+1}{j+1} \cdot x \cdot P_j(x) - \frac{j}{j+1} \cdot P_{j-1}(x)$$

mit $P_0(x) = 1$, $P_1(x) = x$. Explizite Definition:

$$P_n(x) = \frac{1}{2^n} \frac{d^n}{dx^n}(x^2-1)^n$$

Es gilt:

$$\begin{aligned} < p_i, p_j > &= 0 \qquad (i \neq j) \\ < p_j, p_j > &= \frac{2}{j+1} \end{aligned}$$

Bestapproximation mit Polynomen vom Höchstgrad n:

$$\hat{y}(f) = \sum_{j=1}^{n} c_j(f) \cdot P_j(.)$$

mit

$$\frac{2}{j+1} c_j = \int_{-1}^{1} f(x) \cdot p_j(x)\, dx \qquad j = 0(1)n$$

- Allgemeiner (Wichtung): $L_2^w(a,b)$ mit $w(x) > 0$ für alle $x \in (a,b)$, $w \in C[a,b]$. Dann

$$< u, v >_w = \int_a^b w(x) \cdot u(x) \cdot v(x)\, dx$$

- Rekursive Erzeugung der Orthogonalsysteme: P_{n+1} aus $x \cdot P_n$, dann Schmidtsches Orthogonalisierungsverfahren. Resultiert immer in 3-Term-Rekursion (Übung)
- Beispiel: $a = -1$, $b = 1$
  1. $w = 1$, Legendre-Polynome
  2. $w = \frac{1}{\sqrt{1-x^2}}$, Tschebysheff-Polynome

T-Approximation (Approximation in der $\|.\|_\infty$-Norm)

- stetige T-Approximation (gleichmäßige Approximation), $f \in C[a,b]$,

$$\max_{x \in [a,b]} \left| \sum_{j=1}^{n} c_j \cdot \varphi_j(x) - f(x) \right| \to \min \qquad (1)$$

- Spezialfall: gleichmäßige Polynominterpolation:

$$\max_{x \in [a,b]} |p_n(x) - f(x)| \to \min \qquad (2)$$

bei $p_n \in \mathcal{P}_n$.

- Notwendige Optimalitätsbedingung für $f \notin \mathcal{P}_n$. Alternatensatz:

Satz Zu gegebenen $f \in C[a,b]$ sei ein $p_n^* \in \mathcal{P}_n$ gegeben als Lösung von (2). Es sei

$$\begin{aligned} \delta_n &:= \|f - p_n^*\|_\infty (> 0) \\ \delta(x) &= f(x) - p_n(x) \in C[a,b] \end{aligned}$$

Es existieren dann (n+2) Punkte $x_1 < \ldots < x_{n+2}$ mit

$$\begin{aligned} |\delta(x_i)| &= \delta_n \\ \delta(x_i) \cdot \delta(x_{i+1}) &= -\delta_n^2 < 0 \qquad i = 1(1)(n+2) \end{aligned}$$

Beweis:

- Annahme: Höchstens (n+1) „Vorzeichenwechsel“. Dann existieren (n+1) Alternatenpunkte $a \leq x_0 < \ldots < x_n \leq b$. Im Intervall (x_{i-1}, x_i) existiert $\tilde{x}_i$ als einfache Nullstelle von δ. Definiere
$$q(x) := \prod_{i=1}^{n} (x - \tilde{x}_i) \ \in \mathcal{P}_n$$

- Zu zeigen: $p_n^* - \alpha \cdot q \in \mathcal{P}_n$ mit $\|f - (p_n^* - \alpha q)\|_\infty < \delta_n$ für alle $\alpha \in (0, \alpha_0]$. Per Konstruktion gilt:
$$\delta(x) \cdot q(x) > 0 \ \forall x \in M := \{x : |\delta(x)| = \delta_n\}$$
Sei
$$M' = \{x \in [a,b] : \delta(x) \cdot q(x) \leq 0\}$$
M' ist kompakt, $M \cap M' = \emptyset$,
$$0 \leq d := \begin{cases} \max\limits_{x \in M'} |\delta(x)| \\ 0 \qquad \text{sonst} \end{cases}$$
Wähle $\alpha_0 := \frac{\delta_n - d}{2\|q\|_\infty} > 0$. Zu zeigen: $\|(p_n^* + \alpha_0 \cdot q) - f\|_\infty < \delta_n$ im Widerspruch zur Annahme, dass p_n^* Bestapproximation in $\mathcal{P}_n$ bzgl. $[a,b]$.

- Es existiert $\xi \in [a,b]$ mit
$$|\delta(\xi) - \alpha_0 \cdot q(\xi)| = \|\delta - \alpha_0 \cdot q\|_\infty$$

  1. Falls $\xi \in M'$, dann
$$\begin{aligned} |\delta(\xi) - \alpha_0 \cdot q(\xi)| &\leq |\delta(\xi)| + \alpha_0 \cdot |q(\xi)| \\ &\leq d + \frac{1}{2}(\delta_n - d) = \frac{1}{2}(d + \delta_n) < \delta_n \end{aligned}$$

  2. Falls $\xi \notin M'$ ist, gilt:
$$\begin{aligned} \delta(\xi) \cdot q(\xi) &> 0 \\ \Rightarrow |\delta(\xi) - \alpha_0 \cdot q(\xi)| &\overset{!}{<} \max\{|\delta(\xi), \alpha_0 \cdot |q(\xi)|\} \\ &\leq \max\{\delta_n, \frac{1}{2}(\delta_n - d)\} \\ &= \delta_n \end{aligned}$$

Satz Nullstellen orthogonaler Polynome: Sei $\{p_n\}_0^\infty$ ein System von Orthopolynomen in $L_2^w(a,b)$. Dann hat p_k genau k einfache Nullstellen in (a,b).

Beweis:

- Seien $(x_i) \in (a,b)$, $i = 1(1)m$, alle Vorzeichenwechsel von p_k in (a,b). Anahme: $m < k$. Dann:
$$q(x) := \prod_{i=1}^{m} (x - x_i) \in \mathcal{P}_{k-1}$$
Dann gilt aber:
$$(q, p_k) = \int_a^b w(x) \cdot q(x) \cdot p_k(x) > 0 (< 0)$$
im Widerspruch zu $p_k \perp \mathcal{P}_{k-1}$.

Bemerkung zur praktischen Realisierbarkeit von (1)

- Bestimme $c^* \in \mathbb{R}^n$, sodass

$$\max_{x\in[a,b]} \left| \sum_{j=1}^{n} c_j^* \cdot \varphi_j(x) - f(x) \right| \to \min \qquad (1)$$

- Rennes-Verfahren: Vorgabe von Punkten $x_i \in [a,b]$ für $i = 1(1)k$ und lösen diskretes T-Approximationsproblem:

$$\begin{array}{rcll} & & \sigma & \to & \min \\ \left| \sum_{j=1}^{n} c_j \cdot \varphi_j(x_i) - f(x_i) \right| & \geq & \sigma & & i = 1(1)k \end{array}$$

  Aussage: $c^{*,k}$, $\sigma^{*,k}$ ist Lösung von (1), falls

$$\max_{x\in[a,b]} \left| \sum_{j=1}^{n} c_j^{*,k} \cdot \varphi_j(x) - f(x) \right| \leq \sigma^{*,k}$$

  Anderenfalls existiert $x_{k+1} \in [a,b]$ mit

$$\left| \sum_{j=1}^{n} c_j^{*,k} \cdot \varphi_j(x_{k+1}) - f(x_{k+1}) \right| \geq \sigma^{*,k}$$

  Dann x_{k+1} als zusätzliche Stützstelle. Dann Lösung von (2) mit (k+1) Stützstellen.

- Hier sinnvoll: Realisierung mittels dualer Simplexmethode (Einfügen einer Nebenbedingung)

2

Numerische Differentiation und Integration

2.1 Numerische Differentiation

- Seien $x_i = x_0 + i \cdot h$ für $i = 0(1)n$ mit $h > 0$ fix ($h << 1$). Nach Taylor: (falls f hinreichend glatt):

$$\begin{aligned} f_0 &= f_1 - f_1' \cdot h + \frac{f_1''}{2} \cdot h^2 - \frac{f^{(3)}}{6} \cdot h^3 + \frac{f^{(4)}}{24} \cdot h^4 + R_4(\xi) \\ f_2 &= f_1 + f_1' \cdot h + \frac{f_1''}{2} \cdot h^2 + \frac{f^{(3)}}{6} \cdot h^3 + \frac{f^{(4)}}{24} \cdot h^4 + R_4(\xi) \end{aligned}$$

- Rückwärtsdifferentiation:

$$f'(x_1) = \frac{f(x_1) - f(x_0)}{h} + \frac{1}{2} f''(\sigma) \cdot h + O(h^2)$$

- Vorwärtsdifferentiation:

$$f'(x_1) = \frac{f(x_2) - f(x_1)}{h} - \frac{1}{2} f''(\eta) \cdot h + O(h^2)$$

- Zentrale Differentiation: ((2)-(1))

$$f'(x_1) = \frac{f(x_2) - f(x_0)}{2h} + \frac{1}{6} \cdot (f^{(3)}(\sigma') + f^{(3)}(\eta')) \cdot h^2 + O(h^4)$$

- Analog lassen sich Approximationen für Ableitungen höherer Ordnung finden.

$$f''(x_1) = \frac{1}{h^2} \cdot (f_0 - 2f_1 + f_2) - \frac{1}{12} \cdot (f^{(4)}(\xi) + f^{(4)}(\eta)) \cdot h^2$$

  Anderer Weg zur Gewinnung von Ableitungsformeln: Berechnung aus Interpolationspolynom von f.

  Fehler: Nach Appromximationssatz für Polynominterpolation

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

  Für Fehler der Ableitung (formal):

$$f'(x) = p_n'(x) + (f^{(n+1)}(\xi(x)))' \cdot \frac{\prod_{i=0}^{n}(x - x_i)}{(n+1)!} + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \left(\prod_{i=0}^{n} (x - x_i) \right)'$$

  Kritisch für Fehlerkontrolle: $f^{(n+1)}(\xi(x))'$, da keine vernünftige Fehlerkontrolle möglich. Für Gitterstellen $x = x_k$ ist dieser Term 0:

$$f'(x_k) = p_n'(x_k) + \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \prod_{i=0, i \neq k}^{n} (x_i - x_k)$$

- Beispiel:

$$\begin{aligned} p_2(x) &= f_0 \cdot \frac{(x-x_1)\cdot(x-x_2)}{(x_0-x_1)\cdot(x_0-x_2)} + f_1 \cdot \frac{(x-x_0)\cdot(x-x_2)}{(x_1-x_0)\cdot(x_1-x_2)} + f_2 \cdot \frac{(x-x_0)\cdot(x-x_1)}{(x_2-x_0)\cdot(x_2-x_1)} \\ p_2'(x) &= \frac{1}{2h^2}\cdot f_o \cdot (-h) - \frac{1}{h^2}\cdot f_1(0) + \frac{1}{2h^2}\cdot f_2 \cdot h \\ &= \frac{f_2 - f_0}{2h} \end{aligned}$$

Maschinenrealisierung: $\delta_i \to \tilde{f}_i$ (Fehler ε: fix).

$$\begin{aligned} f'(x_1) &= \frac{f(x_1)-f(x_0)}{h} + \frac{1}{2}\cdot f''(\sigma')\cdot h \\ &= \frac{\tilde{f}_1 - \tilde{f}_0}{h} + \underbrace{\frac{(f_1-\tilde{f}_1)-(f_0-\tilde{f}_0)}{h}}_{\leq 2\frac{\varepsilon}{h}} + \underbrace{\frac{f''(\sigma')\cdot h}{2}}_{\leq L\cdot h} \end{aligned}$$

für $h \to 0$: wachsender Fehler. Es existiert ein h_{opt}, für den der Fehler minimal ist.

2.2 Numerische Integration

- Integral als lineares Funktional

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

mit $I : C[a,b] \to \mathbb{R}$. Näherungsweise $J[f] \approx J[p_n]$. Genauer:

$$|J[f] - J[p_n]| = \left| \int_a^b (f(x) - p_n(x))\,dx \right| \leq (b-a)\cdot \|f-p\|_\infty$$

(ziemlich grob geschätzt...)

- Für $f \sim p$ Interpolationspolynom mit den Stützstellen x_i gibt es zwei Varianten:
  1. geschlossene Newton-Cotes-Formeln:

$$x_i = a + i\cdot h \qquad i = 0(1)n \qquad h = \frac{b-a}{n}$$

  2. offene Newton-Cotes-Formeln:

$$x_i = a + i\cdot h \qquad i = 1(1)(n-1) \qquad h = \frac{b-a}{n}$$

- Ansatz:

$$p_n(x) = \sum_{j=0}^{n} \varphi_j(x)\cdot f_j \qquad \varphi_j(x) = \frac{\prod_{k\neq j}(x-x_k)}{\prod_{k\neq j}(x_j-x_k)}$$

Dann:

$$I_n(f) := I(p_n) = \sum_{j=0}^{n} f_j \underbrace{\int_a^b \varphi_j(x)\,dx}_{=:\alpha_j}$$

Damit: numerische Integrationsformel:

$$I(f) \approx \sum_{j=0}^{n} \alpha_j \cdot f_j \qquad \alpha_j \in \mathbb{R}$$

Andere Darstellung für α_j mit $x = a + t\cdot h$, $h = \frac{b-a}{n}$:

$$\alpha_j = \int_a^b \frac{\prod(x-x_k)}{\prod(x_j-x_k)}\,dx = \frac{b-a}{n}\int_0^n \frac{\prod_{k\neq j}(t-k)}{\prod_{k\neq j}(j-k)}\,dt$$

- Spezialfälle (geschlossene NC-Formeln):

  1. n=1: Trapezregel
  $$I_1(f) = \frac{h}{2} \cdot (f_0 + f_1) \qquad \frac{h^3}{12} \cdot \|f''\|_\infty$$
  2. n=2: Simpsonregel
  $$I_2(f) = \frac{h}{3} \cdot (f_0 + 4f_1 + f_2) \qquad \frac{h^5}{90} \cdot \|f^{(4)}\|_\infty$$
  3. n=3: 3/8-Regel
  $$I_3(f) = \frac{3h}{8} \cdot (f_0 + 3f_1 + 3f_2 + f_3) \qquad \frac{3}{80} \cdot h^5 \cdot \|f^{(4)}\|_\infty$$
  4. n=4: Milne-Regel:
  $$I_4(x) = \frac{2h}{45} \cdot (7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) \qquad \frac{8}{945} \cdot h^7 \cdot \|f^{(6)}\|_\infty$$

  Spezialfälle (offene NC-Formel):

  1. n=0: MP-Regel:
  $$I_0(f) = \frac{f(a+b)}{2} \cdot (b-a) \qquad \frac{h^3}{24} \cdot \|f''\|_\infty$$
  2. n=1:
  $$I_1(f) = \frac{b-a}{2} \cdot \left(f\left(\frac{3a+b}{4}\right) + f\left(\frac{3b+a}{4}\right)\right) \qquad \frac{h^3}{4} \cdot \|f''\|_\infty$$

- Fehlerabschätzung: lokaler Fehler bei NC-Formeln:
$$\begin{aligned} f(x) &= p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \prod_{k=0}^{n}(x - x_k) \\ I_f - I_n(f) &= \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x)) \cdot \prod_{k=0}^{n}(x - x_k)\, dx \\ &= \frac{h^{n+2}}{(n+1)!} \int_0^n f^{(n+1)}(a + h \cdot \tau) \cdot \prod_{k=0}^{n}(t-k)\, dt \\ \Rightarrow |I(f) - I_n(f)| &\leq \frac{h^{n+2}}{(n+1)!} \|f^{(n+1)}\|_\infty \cdot \underbrace{\left| \int_0^n \prod_{k=0}^{n}(t-k)\, dt \right|}_{=:c_n} \end{aligned}$$

- Bemerkungen:

  1. offene NC-Formeln: ungebräuchlich
  2. geschlossene NC-Formeln mit $n \geq 8$:
     - Es treten negative Gewichte auf.
     - Fehleraufschaukelung, numerische Instabilität
     - Es gilt:
     $$\sum_{j=0}^{n} |\alpha_j| \to \infty (n \to \infty)$$

     Also Polynomgrad ≤ 6 sinnvoll.

3. Von diesen Quadraturformeln werden Polynome bis zum Grad n exakt integriert. Für $n = 2m$: Polynome vom Grad bis (2m+1) werden exakt integriert.

   Sei
   $$U = \int_0^t \prod_{k=0}^{n} (s-k)\,ds$$
   Beobachtung: $u(0) = u(n) = 0$. Damit:
   $$\begin{aligned}\int_0^n f^{(n+1)}(a+h\cdot\tau)\cdot\prod_{k=0}^{n}(t-k)\,dt &= u(t)\cdot f^{(n+1)}(a+h\cdot\tau)|_0^h - \\ &\quad h\cdot\int_0^n f^{(n+2)}(a+h\cdot\tau)\cdot u(t)\,dt \\ \Rightarrow |I(f)-I_n(f)| &\leq \frac{d_n\cdot h^{n+2}}{(n+1)!}\cdot\|f^{(n+2)}\|_\infty\end{aligned}$$

4. Wegen n beschränkt (≤ 6) kann nicht $h \to 0$ erzwungen werden.

2.2.1 Summierte Quadraturformeln

- Anwendung der Quadraturformeln auf Teilintervalle
- n fixiert für gewisse NC-Formel; Unterteilung des Gesamtintervalls in s Teilintervalle
- Damit Anzahl der Stützstellen $N = n \cdot s$ mit
  $$h = \frac{b-a}{N} = \frac{1}{n}\cdot\frac{b-a}{s} \to 0 (s \to \infty)$$

  Es gilt:
  $$\int_a^b f(x)\,dt = \sum_{j=0}^{s-1}\int_{x_{j\cdot n}}^{x_{(j+1)\cdot n}} f(x)\,dx$$

  Beispiele:

  1. Verallgemeinerte Trapez-Formel:
     $$\tilde{T}_N(f) = h\cdot\left(\frac{1}{2}f_0 + \sum_{k=1}^{N-1} f_k + \frac{1}{2}f_N\right)$$
  2. Verallgemeinerte Simpsonregel:
     $$\hat{s}_N(f) = \frac{h}{3}\cdot(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{N-1} + f_N)$$

Satz Für die verallgemeinerte Trapezregel gilt für $f \in C^2$:
$$\left|\int_a^b f(x)\,dx - \tilde{T}_N(f)\right| \leq \frac{b-a}{12}\cdot h^2\cdot\|f''\|_\infty$$

Beweis:

- Es gilt:
  $$\int_{x_{j-1}}^{x_j} f(x)\,dx - \frac{1}{2}(f(x_j-1) + f(x_j)) \leq \frac{h^3}{12}\cdot\|f''\|_\infty$$

  Aufsummmation:
  $$\begin{aligned}\left|\int_a^b f(x)\,dx - \sum_{j=1}^{N}\frac{1}{2}(f(x_{j-1}) + f(x_j))\right| &\leq \frac{h^2}{12}\cdot\frac{b-a}{N}\sum_{j=0}^{N}\|f''\|_\infty \\ &= \frac{h12}{12}\cdot(b-a)\|f''\|_\infty\end{aligned}$$

Satz Für die verallgemeinerte Simpsonregel gilt:

$$\left| \int_a^b f(x)\,dx - \hat{s}_N(f) \right| \leq \frac{b-a}{180} \cdot h^4 \cdot \|f^{(4)}\|_\infty$$

Bemerkungen:

1. Generell: Verlust einer Approximationsordnung wegen Summation über lokalen Fehler
2. Geschlossene Newton-Cotes-Formeln: Mehrfachzählen der Teilränder
3. Nichtäquidistante Gitter in Abhängigkeit von Fehlerschätzungen sinnvoll: „Verfeinerung“ bei stärkerer Oszillation des Integranden in Teilbereichen
4. Asymptotik der summierten Trapezregel: Sei $f \in C^{2r+2}[a,b]$, $r \in \mathbb{N}_0$. Dann gilt die asymptotische Entwicklung für $h \to 0$

   $$\hat{T}_N(f) = T_0 + T_1 \cdot h^2 + \ldots + T_r \cdot h^{2r} + R_{r+1}(h)$$

   mit

   $$\begin{aligned} T_0 &= \int_a^b f(x)\,dx \\ T_1 &= \frac{b-a}{12} \cdot \|f''\|_\infty \\ R_{r+1}(h) &= O(h^{2r+2}) \end{aligned}$$

   Dabei:

   - nur gerade Potenzen
   - ähnelt Taylorentwicklung bei $h = 0$, aber $\hat{T}_N(f)[h]$ existiert nur für $h > 0$
   - andere Quadraturformeln besitzen ähnliche Asymptotiken

2.3 Gauß-Quadratur

- bisher: Wahl der Stützstellen à priori fixiert, Wahl der Gewichte angepasst.
- Jetzt: Wahl der Stützstellen und Gewichte simultan derart, dass Polynome bis zum Grad (2n-1) exakt integriert werden.
- Ansatz:

  $$I_n(f) = \sum_{i=1}^{n} q_i \cdot f(x_i)$$

  mit $q_i \in \mathbb{R}$, $x_i \in [a,b]$ frei wählbar (2n Parameter).

Satz Es sei $I_n : C[a,b] \to \mathbb{R}$ eine lineare Quadraturformel, für die gilt:

$$|I_n(f)| \leq \tilde{c} \cdot \|f\|_\infty \cdot (b-a) \qquad \tilde{c} \in \mathbb{R}_+$$

Ferner sei $I(p) = I_n(p)$ für alle $p \in \mathcal{P}_n$. Dann existiert $\tilde{c} > 0$ mit

$$|I(f) - I_n(f)| \leq \tilde{c} \cdot \|f^{(n+1)}\|_\infty \cdot (b-a)^{n+2}$$

für alle $f \in C^{n+1}[a,b]$.

Beweis:

- Taylor; $f \in C^{n+1}[a,b]$,

$$\begin{aligned} f(x) &= \underbrace{\sum_{j=0}^{n} \frac{f^{(j)}(a)}{j!} \cdot (x-a)^j}_{=p_n} + \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot (x-a)^{n+1} \\ \Rightarrow \|f - p_n\|_\infty &\leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \cdot (b-a)^{n+1} \end{aligned}$$

  Linearität von I, I_n:

$$\begin{aligned} I(f) - I_n(f) &= I(f) - I(p_n) + \underbrace{I(p_n) - I_n(p_n)}_{0} + I_n(p_n) - I_n(f) \\ |I(f) - I_n(f)| &\leq |I(f) - I(p_n)| + |I_n(f - p_n)| - I_n(f) \end{aligned}$$

  da

$$\begin{aligned} |I_n(f-p)| &\leq \tilde{c}_1 \cdot (b-a)^{k+2} \cdot \|f^{(k+1)}\|_\infty \\ \left| \int_a^b |f(x) - p(x)|\, dx \right| &\leq \tilde{c}_1 \cdot (b-a)^{k+2} \cdot \|f^{(k+1)}\|_\infty \end{aligned}$$

Konstruktion der Gauß-Quadraturformel über orthogonale Polynome:

- damit auch Gauß-Quadaturformel für Integrale mit Gewichten möglich
- Sei $\varphi_0, \ldots, \varphi_n$ das System der orthogonalen Polynome bzgl. $L_2(a,b)$. Erinnerung: φ_k besitzt genau k einfache Nullstellen in (a,b) für $k = 1(1)n$.
- Gauß-Quadraturformel der Ordnung n: Wahl der x_i jeweils als Nullstellen von φ_n.

$$x_i : \varphi(x_i) = 0 \qquad i = 1(1)n \qquad (1)$$

  Wahl der q_i:

$$q_i = \int_a^b \frac{\prod_{k \neq i}(x - x_k)}{\prod_{k \neq i}(x_i - x_k)} \qquad (2)$$

Satz Die Näherungsformel

$$I_n[f] = \sum_{i=1}^{n} q_i \cdot f(x_i)$$

gemäß (1),(2) integriert Polynome bis zum Grad (2n-1) exakt.

Beweis:

- Sei p' ein Polynom vom Höchstgrad (2n-1). Dann gibt es Polynome s, t vom Höchstgrad (n-1) mit

$$p(x) = s(x) \cdot \varphi_n(x) + t(x) \qquad x \in [a,b]$$

  Damit s,t darstellbar als

$$\begin{aligned} s(x) &= \sum_{i=0}^{n-1} \cdot \sigma_i \cdot \varphi_i(x) \\ t(x) &= \sum_{i=0}^{n-1} \tau_i \cdot \varphi_i(x) \end{aligned}$$

Dann gilt:

$$\begin{aligned} I[p] &= \int_a^b p(x)\,dx \\ &= \int_a^b \left(\sum_{i=0}^{n-1} \sigma_i \cdot \varphi_i(x) \cdot \varphi_n(x) \right) dx + \int_a^b t(x)\,dx \\ &= \sum_{i=0}^{n-1} \sigma_i \cdot \underbrace{(\varphi_i, \varphi_n)_2}_{=0} + \int_a^b t(x)\,dx \\ &= \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} q_i \cdot p(x_i) \end{aligned}$$

da

$$p(x_i) = s(x_i) \cdot \underbrace{\varphi_n(x_i)}_{0} + t(x_i) = t(x_i)$$

Bemerkung:

- Quadraturformel (ink. Beweis) lässt sich auf $L_2^w(a,b)$ (bzw. I^w) übertragen (mit modifizierten Gewichten).

Beispiel:

1. Es sei $a = -1, b = 1$ und

$$\varphi_0(x) = 1 \qquad \varphi_1(x) = x \qquad \varphi_2(x) = x^2 - \frac{1}{3} (\Rightarrow x_{1/2} = \pm \frac{1}{\sqrt{3}})$$

Gewichte:

$$q_1 = \int_{-1}^{1} \frac{x - \frac{1}{\sqrt{3}}}{-\frac{2}{\sqrt{3}}}\,dx = 1 = q_2 = \int_{-1}^{1} \frac{x + \frac{1}{\sqrt{3}}}{\frac{2}{\sqrt{3}}}\,dx$$

Damit:

$$Q_2[f] = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Auf [a,b]:

$$Q_2[f] = (b-a) \cdot \left(f\left(\frac{a+b}{2} - \frac{1}{2 \cdot \sqrt{3}} \cdot (b-a) \right) + f\left(\frac{a+b}{2} + \frac{1}{2 \cdot \sqrt{3}} \cdot (b-a) \right) \right)$$

2.4 Genauigkeitserhöhung für asymptotische Formeln durch Extrapolation

(Richardson-Extrapolation/Extrapolation nach Schrittweite 0)

- Grundsituation: Ein zu ermittelnder Wert T wird in Abhängigkeit eines Parameters $h > 0$ (Diskretisierungsschrittweite) durch eine Näherungsformel $T_l(h)$ approximiert.

$$T_l(h) = T + \sum_{j=l}^{m} c_{jl} \cdot h^{p_j} + R_{m+1}(h)$$

mit $0 < p_l < \ldots < p_m$. Ferner gelte

$$\lim_{h \to 0} \frac{R_{m+1}(h)}{h^{p_m}} = 0$$

Dann Gewinnung genauerer Formeln wie folgt:

$$\begin{aligned} T_l(h) &= T + \sum_{j=l}^{m} c_{jl} \cdot h^{p_j} + R_{m+1}(h) \\ T_l(sh) &= T + \sum_{j=l}^{m} c_{jl} \cdot (sh)^{p_j} + R_{m+1}(sh) \qquad s > 0, s \neq 1 \\ \Rightarrow \frac{s^{p_l} \cdot T_l(h) - T_l(sh)}{s^{p_l} - 1} &= T + \sum_{j=l\neq 1}^{m} c_{jl} \cdot \frac{s^{p_l} - s^{p_j}}{s^{p_l} - 1} \cdot h^{p_j} + \hat{R}_{m+1}(h) \\ &=: T_{l+1}(h) \end{aligned}$$

mit

$$\hat{R}_{m+1}(h) = \frac{s^{p_l} \cdot (R_{m+1}(h) - R_{m+1}(sh))}{s^{p_l} - 1}$$

Vorgehensweise im Prinzip wiederholbar.

- Restgliedverhalten:

$$\begin{aligned} \lim_{h\to 0+} \frac{\hat{R}_{m+1}(h)}{h^{p_m}} &= \lim_{h\to 0+} \frac{s^{p_l} \cdot R_{m+1}(h) - R_{m+1}(sh)}{s^{p_l} - 1} \\ &= \lim_{h\to 0+} \frac{s^{p_l}}{s^{p_l} - 1} \cdot \underbrace{\frac{R_{m+1}(h)}{h^{p_m}}}_{\to 0} + \lim_{h\to 0} \frac{s^{p_m+p_l}}{s^{p_l} - 1} \cdot \underbrace{\frac{R_{m+1}(sh)}{(sh)^{p_m}}}_{\to 0} = 0 \end{aligned}$$

Damit besitzt auch $T_{l+1}(h)$ eine Darstellung (1), jedoch eine asymptotisch genauere.

- Praktisches: Berechnung für unterschiedliche Werte h_i, dabei sei $h_{i-1} = s \cdot h_i$, $s = 2 \hat{=}$ „Halbierung“. Wir setzen

$$\begin{aligned} T_{0,i} &:= T_0(h_i) \qquad i = 0(1)k \\ T_{l+1}(h) &= \frac{s^{p_l}}{s^{p_l} - 1} \cdot (T_l(h) - T_l(sh)) \\ &= T_l(h) + \frac{T_l(h) - T_l(sh)}{s^{p_l} - 1} \end{aligned}$$

Rekursion:

$$T_{l+1,i} = T_{l,i} + \frac{T_{l,i} - T_{l,i-1}}{s^{p_l} - 1} \qquad l = 0(1)k \; s = l(1)k$$

- Beispiele:

1. Sei

$$T_0(h) = \frac{1}{h} \cdot (f(x+h) - f(x)) \qquad T = f'(x)$$

Nach Satz von Taylor gilt:

$$T_0(h) = T + \sum_{i=0}^{m} \frac{f^{(j+2)}(x)}{(j+2)!} \cdot h^{j+1} + R(h)$$

mit

$$\lim_{h\to 0} \frac{R(h)}{h^{m+1}} = 0 \qquad p_j = j + 1$$

2. Romberg-Verfahren:

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

Approximation mit summierter Trapez-Regel:

$$\begin{aligned} x_h &= a + k \cdot h \qquad k = 0(1)n \qquad h = \frac{b-a}{n} \text{ n gerade} \\ T_0(h) &= \frac{h}{2} \cdot f_0 + \sum_{k=0}^{n-1} f(x_k) + f(x_n) \end{aligned}$$

- **Satz** Es sei $f : [a, b] \to \mathbb{R}$ eine mindestens (2m+k)-mal stetig differenzierbare Funktion. Dann gilt:

$$T_0(h) = \int_a^b f(x)\,dx + \sum_{j=1}^{m} \frac{B_{2j}}{(2j)!} \cdot \left(f^{(2j-1)}(b) - f^{(2j-1)}(a) \right) \cdot h^{2j} + O(h^{2m+2})$$

mit B ... Bernoulli-Zahlen. (Euler-McLaurin-Formel).

Beweis:

– Sei

$$g(x, h) = \int_x^{x+h} f(t)\,dt$$

Dann:

$$f(x) \cdot h = g(x, h) - \frac{h}{2} \cdot \frac{\partial g}{\partial x} + \sum_{k=1}^{2m} \frac{B_k}{k!} \cdot \frac{\partial^k}{\partial x^k} g(x, h) \cdot h^k + O(h^{2m+2})$$

mit

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k}{k!} \cdot x^k$$

Nach Definition von g folgt:

$$\frac{\partial^k}{\partial x^k} g(x, h) = f^{(k-1)}(x + h) - f^{(k-1)}(x)$$

Eingesetzt ergibt:

$$\begin{aligned} f(x) \cdot h &= \int_x^{x+h} f(t)\,dt - \frac{h}{2} \cdot (f(x + h) - f(x)) + \\ &\quad \sum_{k=1}^{m} \frac{B_{2k}}{(2k!)} \cdot (f^{(2k-1)}(x + h) - f^{(2k-1)}(x)) \cdot h^{2k} + O(h^{2m+2}) \end{aligned}$$

Beziehungsweise

$$\frac{h}{2} \cdot (f(x+h) - f(x)) = \int_x^{x+h} f(t)\,dt + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} \cdot (f^{(2k-1)}(x+h) - f^{(2k-1)}(x)) \cdot h^{2k} + O(h^{2m+2})$$

Aufsummation über Teilintervalle (x_k, x_{k+1}). Dann:

$$\begin{aligned} T_n[f] &= T_0(h) \\ &= \int_x^{x+h} f(t)\,dt + \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} \cdot (f^{(2k-1)}(b) - f^{(2k-1)}(a)) \cdot h^{2k} + O(h^{2m+2}) \end{aligned}$$

- Bemerkungen:

– Falls $f \in C^\infty$, dann „Reihendarstellung“

$$T_n[f] = I[f] + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} \cdot (f^{(2k-1)}(b) - f^{(2k-1)}(a)) \cdot h^{2k}$$

mit $|B_{2k}| \approx 2(2k)! \cdot 2\pi^{-2k}$ für $k \to \infty$.

– Trapezregel: Ergebnis für $h = h_k$:

$$h_{k+1} = \frac{1}{2} \cdot h_k \qquad k = 0(1)N$$

$$\begin{aligned} T_{k,j} &= \frac{4^j \cdot T_{k,j-1} - T_{k-1,j-1}}{4^j - 1} \\ &= T_{k,j-1} + \frac{T_{k,j-1} - T_{k-1,j-1}}{4^j - 1} \end{aligned}$$

1xExtrapolation: Summierte Trapezregel ($\longrightarrow$ Simspon-Regel)
k-fache Extrapolation: „neue“ Formeln

– *Siehe auch „Numerische Mathematik“ - H.R. Schwarz, Seite 314*

2.5 Experimental order of convergence (EOC)

- Zur Überprüfung numerischer Näherungsverfahren ($T_k(h) \approx T$) auf theoretische Konvergenzaussagen

$$|T_k(h) - T| \leq c \cdot h^p \qquad p > 0$$

benutzt man in der Regel zunächst Situtationen, in denen man die exakte Lösung kennt ($T \in \mathbb{R}$ bekannt). Um die Konvergenzordnung zu überprüfen (in der Beobachtung $\hat{=}$ numerisches Experiment) Berechnung von $T_k(h_1), T_k(h_2)$; Bildung des Fehlerfunktionals

$$E(h_i) = T_k(h_i) - T$$

- Experimentelle Konvergenzordnung EOC:

$$EOC = \frac{\ln E(h_1) - \ln E(h_2)}{\ln h_1 - \ln h_2} \approx p$$

- Verallgemeinerung auf komplexere Situationen (Lösung einer Differentialgleichung U; Projektion auf Gitter und Vergleich mit numerischer Lösung)

- „experimentelle“ EOC: $h_1 > \ldots >> h_0$ für Überprüfung $T \approx T_k(h_0)$. Dann:

$$\begin{aligned} \hat{E}(h_1) &= T_k(h_1) - T_k(h_0) \\ \hat{EOC} &= \frac{\ln \hat{E}(h_1) - \ln \hat{E}(h_2)}{\ln h_1 - \ln h_2} \end{aligned}$$

- EOC können benutzt werden, um Vermutungen über Approximations- (Konvergenz-)ordnungen zu erhalten.

3

Numerik linearer Gleichungssysteme

3.1 Direkte Verfahren (Eliminationsverfahren)

- Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$: gesucht ist $x \in \mathbb{R}^n$ mit $A \cdot x = b$
- Gleichungssystem lösbar $\Leftrightarrow$ $b \in$ Im (A). Sei das Gleichungssystem lösbar, $\hat{x} \in \mathbb{R}^n$ Lösung, dann gilt: $x \in \mathbb{R}^n$ Lösung $\Leftrightarrow$ $x - \hat{x} \in \ker(A)$.
- Gaußsches Eliminationsverfahren, Beispiele:

$$
\begin{aligned}
A|b = \begin{pmatrix} 1 & 2 & -1 & 6 \\ -1 & 1 & 2 & -1 \\ 2 & 0 & 3 & -1 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & 2 & -1 & 6 \\ 0 & 3 & 1 & 5 \\ 0 & -4 & 5 & -13 \end{pmatrix} \\
\rightarrow \begin{pmatrix} 1 & 2 & -1 & 6 \\ 0 & 0 & \frac{1}{3} & \frac{5}{3} \\ 0 & 0 & \frac{13}{3} & \frac{-19}{3} \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & 2 & -1 & 6 \\ 0 & 1 & \frac{1}{3} & \frac{5}{3} \\ 0 & 0 & 1 & 1 \end{pmatrix} \\
\rightarrow \begin{pmatrix} 1 & 2 & 0 & 5 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{pmatrix}
\end{aligned}
$$

und ...

$$
\begin{aligned}
A|b = \begin{pmatrix} 1 & 2 & 0 & 1 & 01 & \\ 0 & 0 & 1 & 1 & 0 & 1 \\ -2 & -4 & 0 & -2 & 0 & -2 \\ 1 & 2 & 1 & 2 & 1 & 3 \end{pmatrix} &\rightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 2 \end{pmatrix} \\
&\rightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}
\end{aligned}
$$

Basisvariablen: x_1, x_3, x_5, Nicht-Basisvariablen: x_2, x_4. Damit Lösung:

$$
x = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \cdot s + \begin{pmatrix} -1 \\ -1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \cdot t
$$

Für A^T erhält man die Vektoren, die das Bild aufspannen:

$$
\text{Im}\,(A) = \text{lin}\left\{ \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \right\}
$$

- Gauß-Algorithmus-Transformation mittels Eliminationsmatrizen

$$\begin{aligned} A_0 \cdot x &= b_0 \qquad (A_0 := A, b_0 := b) \\ E_1 \cdot A_0 \cdot x &= E_1 \cdot b_0 \\ \Rightarrow A_{i+1} &= E_{i+1} \cdot A_i \qquad b_{i+1} = E_{i+1} \cdot b_0 \end{aligned}$$

  mit geeigneten regulären $(n \times n)$-Matrizen E_i

$$E_1 = E_n + s_1 \circ e_1^T$$

  mit

$$s_i^j = \begin{cases} 0 & i \leq j \\ -\frac{a_{ij}^{j-1}}{a_{jj}^{j-1}} & i > j \end{cases}$$

  (entspricht Gauß-Algorithmus mit Skalierung auf $a_{jj} = 1$). E_i heißen auch Fröbeniusmatrizen. Ferner gilt:

  - $s^{j^T} \cdot e^j = 0$
  - Ist A_{j-1} bereits $(j-1)$ gestaffelt, dann ist $A_j = E_j \cdot A_{j-1}$ j-gestaffelt.

- Bemerkungen:
  - Seien $u, v \in \mathbb{R}^n$ mit $\|u\|, \|v\| > 0$. Dann Rang $(u \cdot v^T) = 1$.

$$u \cdot v^T = \begin{pmatrix} u_1 \cdot v_1 & \dots & u_1 \cdot v_n \\ \vdots & & \vdots \\ u_n \cdot v_1 & \dots & u_n \cdot v_n \end{pmatrix}$$

    Sprechweise: E_1 sind Rang1-Modifikationen.
  - Sei $u^T \cdot v = 0$, dann gilt:

$$\begin{aligned} (E + u \cdot v^T) \cdot (E - u \cdot v^T) &= E \\ \Rightarrow (E + u \cdot v^T)^{-1} &= (E - u \cdot v^T) \end{aligned}$$

    Verallgemeinerung: Lemma von Sherman-Morisson
  - Insgesamt: Das Gauß-Eliminationsverfahren ist darstellbar als

$$\begin{aligned} A_1 &= E_1 \cdot A_0 = E_1 \cdot A \\ A_2 &= E_2 \cdot A_1 = E_2 \cdot E_1 \cdot A \\ &\vdots \\ A_k &= E_k \cdot A_{k-1} = E_k \cdots E_1 \cdot A \end{aligned}$$

    Bisher stets angenommen, dass $a_{kk}^{k-1} \neq 0$ ist.

Lemma Es gilt:

1. $E_k^{-1} = I - s^k \cdot e^{k^T}$
2. $E_1^{-1} \cdot E_2^{-1} \cdots E_k^{-1} = I - \sum_{j=1}^k s^j \cdot e^{j^T}$

Beweis:

1. Siehe oben

2. Beweis per Induktion: $k \to k+1$

$$\begin{aligned} E_1^{-1} \cdots E_k^{-1} \cdot E_{k+1}^{-1} &= \left(I - \sum_{j=1}^{k} s^j \cdot e^{j^T} \right) \cdot E_{k+1}^{-1} \\ &= \left(I - \sum_{j=1}^{k} s^j \cdot e^{j^T} \right) \cdot \left(I - s^{k+1} \cdot e^{k+1^T} \right) \\ &= I - \sum_{j=1}^{k} s^j \cdot e^{j^T} - s^{k+1} \cdot e^{k+1^T} + \\ &\quad \underbrace{\left(\sum_{i=1}^{k} s^j \cdot e^{j^T} \right) \cdot \left(s^{k+1} \cdot e^{k+1^T} \right)}_{0} \end{aligned}$$

letzter Term 0, da $e^{j^T} \cdot s^{k+1} = 0$ für $j = 1, \ldots, k$.

Satz Durch das Gauß-Verfahren (ohne Pivotisierung) sei die erweiterte Koeffizientenmatrix $(A|b)$ der Dimension $(m, n+1)$ nach l Schritten transformiert zu $(\hat{A}, \hat{b})$ der Form

$$(\hat{A}|\hat{b}) = \begin{pmatrix} R & S & p \\ 0 & 0 & q \end{pmatrix}$$

mit $\hat{a}_{ii} \neq 0$ für $i = 1, \ldots, l$, $R \in \mathbb{R}^{l \times l}$, $S \in \mathbb{R}^{l \times (n-l)}$, $p \in \mathbb{R}^l$, $q \in \mathbb{R}^{m-l}$,

$$R = \begin{pmatrix} \hat{a}_{11} & \ldots & \hat{a}_{1l} \\ & \ddots & \\ 0 & & \hat{a}_{ll} \end{pmatrix} \qquad p = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_l \end{pmatrix} \qquad q = \begin{pmatrix} \hat{b}_{l+1} \\ \vdots \\ \hat{b}_m \end{pmatrix}$$

und $S = (\hat{a}_{ij})$ mit $i = 1, \ldots, l$ $j = l+1, \ldots, n$. Dann besitzt A den Rang l und es gilt:

1. $A \cdot x = b$ lösbar $\Leftrightarrow q = 0$
2. Für $x_R \in \mathbb{R}^l, x_s \in \mathbb{R}^{n-l}$ folgt mit $q = 0$:

$$\begin{aligned} A \cdot x = b &\Leftrightarrow R \cdot x_R + S \cdot x_S = p \\ &\Leftrightarrow x_R = R^{-1} \cdot (p - S \cdot x_s) \text{ mit bel. } x_S \in \mathbb{R}^{n-l} \end{aligned}$$

3. Darstellung des Kerns von A durch

$$\begin{pmatrix} -R^{-1} \cdot S \\ I \end{pmatrix} \qquad \dim \ker A = n - l$$

Spezialfall:

- Es sei $m = n$ und A regulär (und keine Pivotisierung erforderlich). Dann

$$U := A_{n-1} = E_{n-1} \cdots E_1 \cdot A \qquad (^*)$$

mit $U = (u_{ij})$ mit $u_{ij} = 0$ für $i > j$ und $u_{ii} \neq 0$ (obere Dreiecksmatrix). Da E^j regulär für $j = 1, \ldots, n-1$ ist $(^*)$ äquivalent zu

$$A = \underbrace{E_1^{-1} \cdots E_{n-1}^{-1}}_{=:L} \cdot U$$

Aus Lemma folgt: L ist reguläre untere Dreiecksmatrix. Damit kann Gauß-Verfahren ohne Pivotisierung als LU-Faktorisierung $A = L \cdot U$ von A interpretiert werden.

Folgerung Lösung von $A \cdot x = b$ in zwei einfachen Schritten (im Spezialfall):

$$A \cdot x = b \Leftrightarrow L \cdot U \cdot x = b \Leftrightarrow \underbrace{L \cdot v = b}_{(1)} \vee \underbrace{U \cdot x = v}_{(2)}$$

wobei

$$(1) \Leftrightarrow v_i = b_i - \sum_{j=1}^{i-1} l_{ij} \cdot v_j \qquad j = 1, \ldots, m$$

(Vorwärtselimination) und

$$(2) \Leftrightarrow x_i = \frac{1}{u_{ii}} \cdot \left(v_i - \sum_{j=i+1}^{n} u_{ij} \cdot x_j \right) \qquad i = n, \ldots, 1$$

(Rückwärtselimination)

Bemerkung:

- Realisierung: Die Eliminationsmatrizen E_j werden nicht gespeichert, sondern nur die Vektoren s^j.
- Pivotisierung: Wichtig für die Durchführbarkeit, aber noch wichtiger für Reduzierung des Fehlers, der durch Rundungen entsteht.

  prinzipielle Forderung bisher $a_{kk}^{k-1} \neq 0$. Jetzt $A_{k-1} \to A_k = S_k \cdot A_{k-1}$.

1. **Totale Pivotisierung** :

   $$\max \left| a_{ij}^{k-1} \right| =: \left| a_{rs}^{k-1} \right| \qquad k \leq i \leq m;\ k \leq j \leq m$$

   (Suche nach möglichen Pivotelementen: Wahl des Maximums). Damit im nächsten Schritt eine k-gestaffelte Matrix entsteht, werden:

   (a) Zeilen getauscht: $k \leftrightarrow r$

   (b) Spalten getauscht: $k \leftrightarrow s$

   Realisierung durch Permutationsmatrizen:

   $$P_{k,r} = \begin{pmatrix} E_{k-1} & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & E_{m-r-1} \end{pmatrix} \text{ mit } Q = \begin{pmatrix} 0 & & & 1 \\ & 1 & & \\ & & \ddots & \\ & & 1 & \\ 1 & & & 0 \end{pmatrix}$$

   Dann:

   $$\tilde{A}_{k-1} = P_{k,r} \cdot A_{k-1} \cdot P_{k,s}$$

   Nach Vertauschen üblicher Eliminationsschritt. Hinweis: Permutationsmatrizen sind orthonormal, d.h. $P_{ij}^2 = I$. Damit:

   $$A \cdot x = b \Leftrightarrow \underbrace{P_l \cdot A \cdot P_r}_{\tilde{A}} \cdot \underbrace{P_r \cdot x}_{\tilde{x}} = \underbrace{P_l \cdot b}_{\tilde{b}}$$

   Insgesamt hoher Aufwand.

2. **Spaltenpivotisierung**:
$$\max_{k\le i\le m} \left|a_{ik}^{k-1}\right| =: \left|a_{rk}^{k-1}\right|$$

Nur Zeilentausch notwendig, damit kann an Stelle von E_k die Matrix $\tilde{E}_k$ verwendet werden

$$\tilde{E}_k = E_k \cdot P_{k,r}$$

Hinweis: Falls gesamte Restspalte verschwindet, dann weitere Spalten einbeziehen in Pivotisierung. Pivotisierung sichert $\|s^j\|_\infty \le 1$.

Satz Das Gauß-Verfahren ist ohne Pivotisierung durchführbar, wenn einer der beiden Fälle vorliegt:

1. $A = A^T$ und A positiv definit
2. A diagonal dominant, d.h.

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \qquad \text{für } i = 1,\dots,n$$

Beispiel:

1. 2-Punkt-Randwertaufgabe

$$y''(x) = f(x) \qquad y(0) = y(1) = 0 \qquad x \in [0,1]$$

Näherungslösung durch Diskretisierung über äquidistante Gitter $\{x_i\}_{i=0}^N$ mit $x_i = i\cdot h$, $h = \frac{1}{N}$. Näherung $y_i \approx y(x_i)$. Diskretisierung von y'' durch Differenzapproximation

$$y'' \approx \frac{1}{h^2}\cdot(y(x-h) - 2y(x) + y(x+h))$$

Approximation der Randwertaufgabe durch

$$\begin{aligned} -y_{i-1} + 2y_i - y_{i-1} &= h^2 \cdot f_i \qquad i = 1,\dots,N-1 \\ y_0 &= y_N = 0 \end{aligned}$$

Lineares Gleichungssystem mit der tridiagonalen Matrix A (wenn y_0, y_N weggelassen). A symmetrisch und positiv definit.

Bemerkung zur LU-Zerlegung: Ist die Ausgangsmatrix A eine Bandmatrix, d.h. $a_{ij} = 0$ für $|i-j| > p$, dann gilt für die LU-Faktorisierung ohne Pivotisierung $l_{ij} = 0$ für $|i-j| > p$ und $U_{kl} = 0$ für $|k-l| > p$, d.h. L, U sind Bandmatrizen der gleichen Bandbreite.

Hier gilt $p = 1$.

$$L = \begin{pmatrix} 1 & & & 0 \\ \alpha_2 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & \alpha_n & 1 \end{pmatrix} \qquad U = \begin{pmatrix} \beta_1 & \gamma_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \gamma_{n-1} \\ 0 & & & \beta_n \end{pmatrix}$$

mit

$$A = \begin{pmatrix} d_1 & c_1 & & 0 \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & a_n & d_n \end{pmatrix}$$

Dann:

$$\beta_j = d_j \;\gamma_j = c_j \;\alpha_j = \frac{a_j}{\beta_{j-1}} \;\beta_j = d_j - \alpha_j \cdot \gamma_{j-1}$$

d.h. $\approx$ 3n Operationen für LU-Zerlegung. Vorwärtsrechnung $L \cdot v = b$:

$$v_1 = b_1 \qquad v_j = b_j - \alpha_j \cdot v_{j-1} \qquad j = 2, \ldots, n$$

d.h. $\approx$ 2n Operationen für Vorwärtsrechnung. Rückwärtsrechnung $U \cdot x = v$:

$$x_n = \beta_n \cdot v_n \qquad x_j = \frac{1}{\beta_j} \cdot (v_j - \gamma_{j+1} \cdot v_{j+1}) \qquad j = n-1, \ldots, 1$$

d.h. $\approx$ 3n Operationen für Rückwärtsrechnung. Also insgesamt 8n Operationen (falls LU-Zerlegung bekannt: 5n Operationen)

Bemerkung:

- Matrixnorm: Sei $A \in \mathbb{K}^{n \times n}$ (auch: Vektor der Dimension n^2). Vektornorm aus $\mathbb{R}^{n^2}$ möglich. Aber: Zur Vektornorm zugehörige Matrixnorm, falls:

$$\|A \cdot x\|_V \leq \|A\|_M \cdot \|x\|_V$$

Also:

$$\|A\|_M := \min\{K \geq 0 : \|A \cdot x\| \leq K \cdot \|x\|\}$$

Folgerungen:

1. $\|E_n\| = 1$
2. Zeilensummennorm:

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$$

Spektralnorm:

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

λ_{max}: maximaler Eigenwert von $A^T \cdot A$

3.1.1 Problem der Fehlerfortpflanzungen durch Störungen

1. Störungen der rechten Seite:

$$A \cdot x = b \qquad A \cdot \hat{x} = \hat{b}$$

wobei $\hat{x} = x + \triangle x$, $\hat{b} = b + \triangle b$. Dann absoluter Fehler:

$$\begin{aligned} A \cdot \triangle x &= \triangle b \\ \triangle x &= A^{-1} \cdot \triangle b \\ \Rightarrow \|\triangle x\| \leq \|A^{-1} \cdot \triangle b\| &\leq \|A^{-1}\| \cdot \|\triangle b\| \end{aligned}$$

Mit

$$\|b\| = \|A \cdot x\| \leq \|A\| \cdot \|x\|$$

folgt:

$$\|x\| \geq \frac{\|b\|}{\|A\|}$$

Damit relativer Fehler:

$$\frac{\|\triangle x\|}{\|x\|} \leq \underbrace{\|A^{-1}\| \cdot \|A\|}_{:=\mathrm{cond}(A)} \cdot \frac{\|\triangle b\|}{\|b\|}$$

cond(A) gibt es nur für reguläre Matrizen und ist von der Norm abhängig. [Kondition]

2. Störungen in der Koeffizientenmatrix: erfaßbar mit Störungslemma.

Lemma Sei $S \in \mathbb{K}^{n\times n}$ mit $\|I - S\| < 1$. Dann ist S regulär und es gilt:

$$\begin{aligned} S^{-1} &= \sum_{k=0}^{\infty} (I-S)^k \\ \|S^{-1}\| &\leq \frac{1}{1-\|I-S\|} \end{aligned}$$

Beweis:

- S regulär $\Leftrightarrow (S \cdot x = 0 \Rightarrow x = 0) \Leftrightarrow \forall x \neq 0 : S \cdot x \neq 0$. Es gilt:

$$\begin{aligned} \|S \cdot x\| &= \|I + (S - I)\| \geq \|x\| - \|(I - S) \cdot x\| \\ &\geq \|x\| - \|I - S\| \cdot \|x\| \\ &= \underbrace{1 - \|I - S\|}_{>0} \cdot \|x\| \end{aligned}$$

- Falls die angegebene Darstellung gilt, dann

$$\begin{aligned} \|S^{-1}\| &= \left\| \sum_{k=0}^{\infty} (I - S)^k \right\| \\ &\leq \sum_{k=0}^{\infty} \|(I - S)^k\| \\ &\leq \sum_{k=0}^{\infty} \|I - S\|^k \leq \frac{1}{1 - \|I - S\|} \end{aligned}$$

  (Letzter Schritt: geometrische Reihe). Also Reihe konvergent gegen Matrix.

$$S \cdot \underbrace{\sum_{k=0}^{n} (I - S)^k}_{\to \sum_{k=0}^{\infty} (I-S)^k} = I - \underbrace{(I - S)^{n+1}}_{\to 0-\text{Matrix}}$$

  Begründung der verwendeten Formel:

$$\begin{aligned} \sum_{k=1}^{n+1} (I - S)^k &= (I - S) \cdot \sum_{k=0}^{n} (I - S)^k \\ \Rightarrow \sum_{k=0}^{n} (I - S)^k - \sum_{k=1}^{n+1} (I - S)^k &\overset{1}{=} \underbrace{(I - S)^0}_{E_n} - (I - S)^{n+1} \\ &\overset{2}{=} \underbrace{I - (I - S)}_{S} \cdot \sum_{k=0}^{n} (I - S)^k \end{aligned}$$

Störungen:

$$A \cdot x = b \qquad \hat{A} \cdot \hat{x} = b$$

mit $\hat{x} = x + \triangle x$, $\hat{A} = A + \triangle A$.

$$(A + \triangle A) \cdot \triangle x = \triangle A \cdot x$$

Falls $(A + \triangle A)^{-1}$ existiert, dann absoluter Fehler

$$\begin{aligned} \triangle x &= (A + \triangle A)^{-1} \cdot \triangle A \cdot x \\ \Rightarrow \|\triangle x\| < &\leq \| \underbrace{(A + \triangle A)^{-1}}_{\hat{A}^{-1}} \| \cdot \|\triangle A\| \cdot \|x\| \end{aligned}$$

Lemma Sei A regulär und $\|A^{-1} \cdot \triangle A\| < 1$. Dann ist $\hat{A} = A + \triangle A$ regulär und es gilt:

$$\|\hat{A}^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1} \cdot \triangle A\|}$$

Beweis:

- Störungslemma: Wähle $S = A^{-1} \cdot \hat{A}$. Dann

$$\|I - S\| = \|I - A^{-1} \cdot \underbrace{\hat{A}}_{A+\triangle A}\| = \|A^{-1} \cdot \triangle A\| < 1$$

Also:

$$\|S^{-1}\| \leq \frac{1}{1 - \|A^{-1} \cdot \triangle A\|}$$

Mit $\hat{A} = A \cdot S$ folgt die Existenz von A^{-1}, $A^{-1} = S^{-1} \cdot A^{-1}$. Dann

$$\|\hat{A}^{-1}\| \leq \|S^{-1}\| \cdot \|A^{-1}\| = \frac{\|A^{-1}\|}{1 - \|A^{-1} \cdot \triangle A\|}$$

Folgerung Gilt $\|\triangle A\| < \frac{1}{\|A^{-1}\|}$, dann

$$\|\hat{A}^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\triangle A\|}$$

3. Allgemeine Störung, falls $\|\triangle A\| < \frac{1}{\|A^{-1}\|}$:

$$\|\triangle x\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\triangle A\|} \cdot (\|\triangle b\| + \|\triangle A\| \cdot \|x\|)$$

3.1.2 Cholesky-Zerlegung

- Voraussetzungen: A symmetrisch und A positiv definit. (Gauß ohne Pivotisierung möglich)

1. Ansatz:

$$A = L \cdot D \cdot L^T$$

mit D Diagonalmatrix, L untere Dreiecksmatrix mit $l_{jj} = 1$. Aus LU-Zerlegung:

$$U = D \cdot L^T \qquad d_i = u_{ii}$$

2. Ansatz:

$$A = \hat{L} \cdot \hat{L}^T$$

mit $\hat{L}$ als untere Dreiecksmatrix.

$$L \cdot D \cdot L^T = L \cdot D^{\frac{1}{2}} \cdot D^{\frac{1}{2}} \cdot L^T = \underbrace{L \cdot D^{\frac{1}{2}}}_{\hat{L}} \cdot (D^{\frac{1}{2}})^T \cdot L^T$$

- Vorteil: Speicherplatz

3.1.3 Orthogonalisierungsverfahren

- Ziel: Zerlegung von A:

$$A = Q \cdot R$$

mit Q orthonormal ($Q^T = Q^{-1}$), R obere Dreiecksmatrix. Damit:

$$A \cdot x = b \Leftrightarrow Q \cdot R \cdot x = b$$

(einfach lösbar). Zusätzliche Eigenschaft:

$$\begin{aligned} \|Q \cdot x\|_2^2 &= (Q \cdot x)^T \cdot (Q \cdot x) = x^T \cdot Q^T \cdot Q \cdot x \\ &= x^T \cdot x = \|x\|_2^2 \end{aligned}$$

d.h. $\|A \cdot x - b\|_2^2 = \|R \cdot x - Q^T \cdot b\|_2^2$. Ausnutzung bei Fehlerquadratmethode.

- QR-Zerlegung ist nicht eindeutig.

1. Gram-Schmidt-Orthogonalisierung: Sei $A = (a^1 \dots a^n)$. Dann

$$p^k = a^k - \sum_{j=1}^{k-1} r_{jk} \cdot q^j$$

mit

$$r_{jk} = q^{j^T} \cdot a^k \qquad (j = 1, \dots, k-1)$$

Dann:

$$q_k = \frac{p^k}{r_{kk}} \qquad r_{kk} = \|p^k\|_2$$

liefert QR-Zerlegung mit

$$Q = \begin{pmatrix} q^1 & \dots & q^n \end{pmatrix} \in \mathbb{R}^{n \times n} \qquad R = \begin{pmatrix} r_{11} & & r_{1n} \\ & \ddots & \\ 0 & & r_{nn} \end{pmatrix}$$

falls $r_{kk} > 0$. Problem: Numerisch instabil.

2. Householder-Faktorisierung: Householder-Matrizen (zu Vektor u):

$$H_u = I - 2 \cdot u \cdot u^T$$

H_u ist symmetrisch.

Lemma Sei $U^T \cdot u = 1$.

(a) $H_u^T \cdot H_u = I$
(b) $H_u \cdot u = -u$, $H_u(\alpha \cdot u) = -\alpha \cdot u)$
(c) $u^T \cdot v = 0 \Rightarrow H_u \dot{v} = v$

d.h. H_u realisiert Spiegelung an der Hyperebene $u^T \cdot v = 0$.

Ziel: $\tilde{a}_1 = a_1$:

$$H_{u_1} = \tilde{a}_1 = \pm\|\tilde{a}_1\| \cdot e^1$$

Wie u_1 wählen? Realisierung: $a_1 = v_1 + w_1$ mit

$$w_1 = \frac{1}{2} \cdot (\tilde{a}_1 \pm \|\tilde{a}_1\|_2 \cdot e^1) \qquad v_1 = \frac{1}{2} \cdot (\tilde{a}_1 \mp \|\tilde{a}_1\|_2 \cdot e^1)$$

Dann $v_1^T \cdot w_1 = 0$. Wähle

$$u_1 = \frac{w_1}{\|w_1\|_2}$$

Damit:

$$\begin{aligned} H_{u_1} \cdot \tilde{a}_1 &= H_{u_1} \cdot v_1 + H_{u_1} \cdot w_1 = v_1 - w_1 \\ &= \frac{1}{2} \cdot (\tilde{a}_1 \mp \|\tilde{a}_1\|_2 \cdot e^1) - \frac{1}{2} \cdot (\tilde{a}_1 \pm \|\tilde{a}_1\|_2 \cdot e^1) \\ &= \mp\|\tilde{a}_1\| \cdot e^1 \end{aligned}$$

Zur Vermeidung von Stellenauslöschungen:

$$\begin{aligned} w_1 &= \frac{1}{2}(\tilde{a}_1 + \|\tilde{a}_1\|_2 \cdot e^1) \qquad \text{für } a_{11} \geq 0 \\ w_1 &= \frac{1}{2} \cdot (\tilde{a}_1 - \|\tilde{a}_1\|_2 \cdot e^1) \qquad \text{für } a_{11} < 0 \end{aligned}$$

Im 2. Schritt:

$$\tilde{a}_2 = \begin{pmatrix} 0 \\ \hat{a}_{2,R} \end{pmatrix}$$

mit $\hat{a}_2$ als 2.Spalte der Matrix $H_{u_1} \cdot A$ und

$$\hat{a}_{2,R} = \{\hat{a}_{2,i}; i \geq 2\}$$

d.h. Anwendung auf Teilblock. Analog in weiteren Schritten.

3. Givens-Rotation:

$$Q = \begin{pmatrix} E_i & & & & & \\ & s & \dots & t & & \\ & & 1 & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & -t & \dots & s & & \\ & & & & E_j & \end{pmatrix}$$

mit $s^2 + t^2 = 1$ um das Element an der Stelle „$-t$“ in der Matrix zu Null zu machen, d.h. für die Spalte sind (n-1) Givens-Rotationen notwendig, danach. 2. Spalte usw.

3.2 Quadratmittelprobleme + Singulärwertzerlegung

- Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ mit (i.a.) $m > n$. Aufgabe:

$$F(x) = \frac{1}{2} \|A \cdot x - b\|_2^2 \to \min_{x \in \mathbb{R}^n} \qquad (LQ)$$

- **Satz** Das Quadratmittelproblem (LQ) ist stets lösbar, dabei gilt: $\hat{x} \in \mathbb{R}^n$ löst (LQ) $\Leftrightarrow$ $A^T \cdot A \cdot x = A^T \cdot b$ (NG). Sind die Spalten von A linear unabhängig, dann ist $\hat{x}$ eindeutig bestimmt.

- Hinweis: (NG) folgt aus notwendiger Bedingung $F'(x) = 0$.

- Beweis:

  1. Die Menge

  $$\mathcal{R}(A) := \{y \in \mathbb{R}^m; y = A \cdot x, x \in \mathbb{R}^n\}$$

  ist abgeschlossen und konvex. Zu konvex:

  $$y_1, y_2 \in \mathcal{R}(A) \Rightarrow \lambda \cdot y_1 + (1 - \lambda) \cdot y_2 \in \mathcal{R}(A) \,\text{für}\, \lambda \in (0, 1)$$

  Zu abgeschlossen: Sei $\{y^k\} \subset \mathcal{R}(A)$. Dann existiert $\{x^k\} \subset \mathbb{R}^n$ mit $y^i = A \cdot x^i$ für $i = 1, \dots$. Seien o.B.d.A. nur die ersten p Spalten von A linear unabhängig. Dann Wahl von x^k so, dass $x_i^k = 0$ für $i = p + 1, \dots, n$ möglich, d.h. $x^k \mapsto x_p^k$.

  $$A = \begin{pmatrix} A_p & A_{n-p} \end{pmatrix} \qquad y^k = A_p \cdot x_p^k$$

  Damit:

  $$\begin{aligned} A_p^T \cdot y^k &= A_p^T \cdot A_p \cdot x_p^k \\ x_p^k &= (A_p^T \cdot A_p)^{-1} \cdot A_p^T \cdot y^k \end{aligned}$$

  Mit $y_k \to \hat{y}$ folgt:

  $$\begin{aligned} x_p^k \to \hat{x}_p &= (A_p^T \cdot A_p)^{-1} \cdot A_p^T \cdot \hat{y} \\ \hat{y} &= A \cdot \hat{x} \end{aligned}$$

2. (LQ) ist äquivalent zu:

$$\frac{1}{2}\|y-b\|_2^2 \to \min_{y\in\mathcal{R}(A)} \qquad (P)$$

(P) besitzt eine eindeutige Lösung $\hat{y} \in \mathcal{R}(A)$. Also besitzt (LQ) min. 1 Lösung.

- Lösung von (NQ) mit QR-Faktorisierung $A = Q \cdot R$ (denn die direkte Lösung von (NG) mit Cholesky zu ungenau)

$$\begin{aligned} F(x) &= \frac{1}{2}\|A\cdot x-b\|_2^2 = \frac{1}{2}\cdot(A\cdot x-b)^T\cdot(A\cdot x-b) \\ &= \frac{1}{2}\cdot(A\cdot x-b)^T\cdot Q\cdot Q^T\cdot(A\cdot x-b) \\ &= \frac{1}{2}\cdot(Q^T\cdot A\cdot x-Q^T\cdot b)^T\cdot(Q^T\cdot A\cdot x-Q^T\cdot b) \\ &= \frac{1}{2}(R\cdot x-d)^T\cdot(R\cdot x-d) \qquad d:=Q^T\cdot b \\ &= \frac{1}{2}\|R_1\cdot x_1\|_2^2+\frac{1}{2}\|d_2\|_2^2 \end{aligned}$$

mit R als obere Dreiecksmatrix. Damit:

$$F(x) \geq \frac{1}{2}\|d_2\|^2$$

Lösung aus $R_1 \cdot \hat{x} = d_1$. Lösung nach Satz aber nicht notwendig eindeutig (eindeutig für $m > n$, Rang $A = n$). Häufig: Minimalnormlösung gesucht (im rangdefiziten Fall), d.h.

$$\begin{aligned} x_{MN} \in LQ(A,b) = \{\hat{x}\in\mathbb{R}^n : \|A\hat{x}-b\| \leq \|A\cdot x-b\|_2 \forall x\in\mathbb{R}^n\} \\ \text{mit } \|x_{MN}\| \leq \|x\|_2 \forall x \in LQ(A,b) \end{aligned}$$

x_{MN} ist eindeutig bestimmt.

- Bezeichnung: $x_{MN} = A^+ \cdot b$ (A^+: Moore-Penrose-Inverse (Pseudoinverse))
- Lösung von (MN) über Regularisierung von (NG) oder Singulärwertzerlegung.

1. Regularisierung

$$f(x,\alpha) = \|b-A\cdot x\|_2^2+\alpha\cdot\|x\|_2^2 \to \min \qquad \alpha > 0$$

Äquivalent zu: (LQ) von $\hat{A}_\alpha \cdot x = \hat{b}$ mit

$$\hat{A}_\alpha = \begin{pmatrix} A \\ \sqrt{\alpha}\cdot I \end{pmatrix} \qquad \hat{b} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

Wegen $\sqrt{\alpha} \cdot I$ hat die erweiterte Matrix $\hat{A}_\alpha$ im Fall $\alpha > 0$ stetig linear unabhängige Spalten, also Vollrang.

$$(LQ) \Leftrightarrow (NG): \qquad \hat{A}_\alpha^T\cdot\hat{A}_\alpha\cdot x = (A^T\cdot A+\alpha\cdot I)\cdot x = A^T\cdot b$$

eindeutige Lösung:

$$x = (A^T\cdot A+\alpha\cdot I)^{-1}\cdot A^T\cdot b$$

und dann α hinreichend klein.

2. Singulärwertzerlegung

Sei $A \in \mathbb{R}^{m\times n}$ mit Rang $A = r \leq \min\{m,n\}$. Zerlegung

$$A = U\cdot S\cdot V^T$$

mit $U \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{n\times n}$ orthonormal, $S \in \mathbb{R}^{m\times n}$ mit $s_{ij} = 0$ außer für $1 \leq i = j \leq r$, $\sigma_j := s_{jj} > 0$ für $j = 1, \ldots, r$ (Singulärwerte von A). Oft fordert man noch $\sigma_j \geq \sigma_i$ für $j < i$.

$$A = U \cdot S \cdot V^T = \sum_{j=1}^{r} \sigma_j \cdot u^j \cdot {v^j}^T$$

Bestimmung der Singulärwertzerlegung: Orthonormalität von U und V, also

$$\begin{aligned} A^T \cdot A &= V \cdot S^T \cdot S \cdot V^T \\ A \cdot A^T &= U \cdot S \cdot S^T \cdot U^T \end{aligned}$$

Bei beliebiger orthonormaler Matrix Q und beliebiger quadratischer Matrix C gilt: Eigenwerte von C und von $Q \cdot C \cdot Q^T$ sind gleich. Damit Eigenwerte von $A^T \cdot A \in \mathbb{R}^{n,n}$ und $A \cdot A^T \in \mathbb{R}^{m\times m}$ sind gleich und genauer σ_j^2. Auch:

$$\begin{aligned} A \cdot A^T \cdot U &= U \cdot S \cdot S^T \\ A^T \cdot A \cdot V &= V \cdot S^T \cdot S \end{aligned}$$

Spalten von U sind also jeweils normierte Eigenvektoren von $A \cdot A^T$ zu σ^2, Spalten von V normierte Eigenvektoren von $A^T \cdot A$.

Problem: Normierte Eigenvektoren sind nicht eindeutig ($\pm$) und außerdem sind U,V nicht unabhängig zu wählen:

$$A \cdot V = U \cdot S \cdot V^T \cdot V = U \cdot S$$

d.h.

$$A \cdot v^j = \sigma_j \cdot u^j$$

Also falls v^j und σ_j bekannt, folgt u^j. Ebenso:

$$\begin{aligned} A^T \cdot U &= V \cdot S^T \cdot U^T \cdot U = V \cdot S^T \\ A^T \cdot u^j &= \sigma_j \cdot v^j \end{aligned}$$

Hinweis: Für die Zerlegung von A werden und U und V benötigt, für die Darstellung von A aber nur die ersten r Spalten.

Bemerkung: Falls A quadratisch und regulär, dann (für Spektralnorm)

$$\text{cond}(A) = \frac{\sigma_1}{\sigma_n}$$

Es gilt:

$$A^+ = V \cdot S^+ \cdot U^T$$

mit $s_{ij}^+ = 0$ außer $\frac{1}{\sigma_i}$, falls $1 \leq i = j \leq r$, d.h.

$$\begin{aligned} A^+ &= \sum_{j=1}^{r} \frac{1}{\sigma_i} \cdot v^i \cdot ({u^i}^T) \\ \Rightarrow x_{MN} &= A^+ \cdot b = \sum_{j=1}^{r} \frac{{u^i}^T \cdot b}{\sigma_i} \cdot v^i \end{aligned}$$

Begründung, dass Minimallösung: Sei

$$x = \sum_{i=1}^{n} \alpha_i \cdot v^i \qquad b = \sum_{j=1}^{m} \beta_j \cdot u^j$$

Dann:

$$\begin{aligned} A \cdot x - b &= \sum_{j=1}^{r} (\sigma_j \cdot \alpha_j \cdot u^j - \beta_j \cdot u^j) - \sum_{j=r+1}^{m} (\beta_j \cdot u^j) \\ \Rightarrow \alpha_j &= \frac{\beta_j}{\sigma_j} \qquad j = 1, \ldots, r \end{aligned}$$

restliche Komponenten bisher nicht bestimmt. Mit

$$\|x\|_2^2 = \sum_{j=1}^{n} \alpha_j^2 = \sum_{j=1}^{r} \alpha_j^2 + \sum_{j=r+1}^{m} \alpha_j^2$$

folgt $\alpha_j = 0$ für $j = r+1, \ldots, n$ für minimale Norm.

3.3 Iterative Verfahren

$$A \cdot x = b \qquad \text{mit A regulär} \qquad (1)$$

- Ersetzen von (1) durch „benachbachbarte“ Aufgabe

$$B \cdot x = (B - A) \cdot x + b$$

mit $B \in \mathbb{R}^{n \times n}$ regulär, $B \cdot x = d$ ist einfach lösbar (B^{-1} einfach berechenbar) und $\|B^{-1} \cdot (B - A)\|$ „klein“. Iterative Behandlung der Form

$$B \cdot x^{k+1} = (B - A) \cdot x^k + b$$

bzw.

$$x^{k+1} = \underbrace{B^{-1} \cdot (B - A)}_{=:T} \cdot x^k + \underbrace{B^{-1} \cdot b}_{=:t}$$

Iterationsverfahren vom Typ:

$$x^{k+1} = \underbrace{T \cdot x^k + t}_{\Phi(x^k)}$$

mit $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ affin linear.

- Bemerkungen:

1. Einsatz direkter Verfahren: $N = c \cdot 10^5$ ($c \leq 10$) für unstrukturierte Matrizen wegen Aufwand ($c \cdot N^3 + O(N^2)$) und Speicheraufwand
2. Iterative Verfahren für strukturierte Matrizen: $N >> 10^6$, z.B. Wetter auf der Erde

$$12km \cdot \pi \cdot (6300)^2 km^2 \approx 36 \cdot 40.000.000 km^3$$

Iterative Verfahren, weil:

- ein (evtl. 2) Matrix-Vektor-Multiplikation ($0(N^2)$) (für schwach besetzte Matrizen evtl. $O(N)$)
- ein (wenige) Skalarprodukte/Summationen (O(N))
- bei k Iterationen ($k << N$) insgesamt für hinreichend gute Approximation

Bei Matrizen mit spezieller Struktur (symmetrisch, positiv definit, dünn besetzt) besonders effizient (evtl. O(N) ... lineare Komplexität)

3.3.1 Gesamtschritt- und Einzelschrittverfahren

(Jacobi-Verfahren und Gauss-Seidel-Verfahren)

$$A = \underbrace{\begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{NN} \end{pmatrix}}_{=:D} + \underbrace{\begin{pmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{N1} & \ldots & a_{N,N-1} & 0 \end{pmatrix}}_{=:L} + \underbrace{\begin{pmatrix} 0 & a_{12} & \ldots & a_{1N} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N} \\ 0 & \ldots & 0 & 0 \end{pmatrix}}_{=:R}$$

1. Jacobi-Verfahren (Gesamtschritt-Verfahren)

$$\begin{aligned}\sum_{k=1}^{N} a_{jk} \cdot x_k &= b_j \qquad j = 1(1)N \\ \Rightarrow a_{jj} x_j^{m+1} &= b_j - \sum_{k=1, k\not=j}^{N} a_{jk} \cdot x_k^m \qquad (3J)\end{aligned}$$

bzw.

$$\begin{aligned}D \cdot x^{m+1} &= b - (L+R) \cdot x^m \qquad j = 1(1)N \\ x^{m+1} &= -D^{-1} \cdot (L+R) \cdot x^m + D^{-1} \cdot b\end{aligned}$$

mit

$$T_J = -D^{-1} \cdot (L+R) = \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1N}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{a_{N-1,N}}{a_{N-1,N-1}} \\ \frac{a_{N1}}{a_{NN}} & \cdots & \frac{a_{N,N-1}}{a_{NN}} & 0 \end{pmatrix}$$

2. Gauss-Seidel-Verfahren (Einzelschritt-Verfahren)

   Bei Betrachtung von (3J): Zur Berechnung von x_j^{m+1} können die bereits berechneten Näherungen $x_i^{m+1}, i < j$ anstelle von x_i^m benutzt werden.

$$x_j^{m+1} = \frac{1}{a_{jj}} \cdot \left(b_j - \sum_{k=1}^{j-1} a_{jk} \cdot x_k^{m+1} - \sum_{k=j+1}^{N} a_{jk} \cdot x^m \right) \qquad j = 1(1)N; m = 0, \ldots$$

   Ohne Mehraufwand! Dies entspricht in Matrixschreibweise

$$\begin{aligned}(D+L) \cdot x^{m+1} &= b - R \cdot x^m \\ x^{m+1} &= \underbrace{-(D+L)^{-1} \cdot R}_{=:T_{GS}} \cdot x^m + \underbrace{(D+L)^{-1} \cdot b}_{=:t}\end{aligned}$$

Satz Das Iterationsverfahren

$$x^{k+1} = T \cdot x^k + t \qquad k = 0, \ldots$$

ist für beliebiges $x^0 \in \mathbb{R}^n$ und beliebiges $t \in \mathbb{R}^n$ genau dann konvergent, wenn $\varrho(T) < 1$ mit Spektralradius

$$\varrho(T) = \max |\lambda_i(T)|$$

mit λ_i als Eigenwerte von T. $\varrho(T) = \|T\|_2$, falls $T = T^T$.

Beweis:

- Mit Übungsaufgabe 34 und Banachschem Fixpunktsatz. Fixpunktgleichung $x = \varphi(x)$ mit $\varphi : \mathbb{R}^n \to \mathbb{R}^n$,

$$x^{k+1} = \varphi(x^k)$$

Hilfssatz Sei φ kontrahierend, d.h. mit einem $\delta \in (0,1)$ gilt:

$$\|\varphi(x) - \varphi(y)\| \leq \delta \cdot \|x - y\|$$

Dann besitzt $x = \varphi(x)$ einen eindeutigen Fixpunkt $x^* \in \mathbb{R}^n$ und für beliebiges $x^0 \in \mathbb{R}^n$ konvergiert die durch $x^{k+1} = \varphi(x^k)$ erzeugte Folge $\{x^k\}$ gegen x^*. Dabei gilt:

$$\begin{aligned}\|x^0 - x^*\| &\leq \frac{1}{1-\delta} \cdot \|x^1 - x^0\| \\ \|x^k - x^*\| &\leq \frac{\delta^k}{1-\delta} \cdot \|x^1 - x^0\| \\ \|x^k - x^*\| &\leq \frac{\delta}{1-\delta} \cdot \|x^k - x^{k-1}\|\end{aligned}$$

Beweis:

- Sei $\{x^k\}$ erzeugt durch $x^{k+1} = \Phi(x^k)$.

$$\|x^n - x^m\| \leq \|x^n - x^{n-1}\| + \|x^{n-1} - x^{n-2}\| + \ldots + \|x^{m+1} - x^m\|$$

mit

$$\begin{aligned} \|x^{k+1} - x^k\| &= \|\Phi(x^k) - \Phi(x^{k-1})\| \\ &\leq \delta \cdot \|x^k - x^{k-1}\| \leq \delta^k \cdot \|x^1 - x^0\| \end{aligned}$$

Daraus folgt:

$$\begin{aligned} \|x^n - x^m\| &\leq \sum_{j=m}^{n-1} \delta^j \cdot \|x^1 - x^0\| \\ &= \|x^1 - x^0\| \cdot \frac{\delta^m - \delta^n}{1 - \delta} \\ &\to 0 (n, m \to \infty) \end{aligned}$$

Also $\{x^k\}$ Cauchy-Folge und mit Vollständigkeit konvergent.

$$x^k \to x^* \in \mathbb{R}^n (k \to \infty)$$

Mit Stetigkeit von Φ folgt:

$$\begin{aligned} \|\Phi(x^*) - x^*\| &= 0 \\ \Phi(x^*) &= x^* \end{aligned}$$

Eindeutigkeit: Seien x^*,x' Fixpunkte, dann:

$$\begin{aligned} \|x' - x^*\| &= \|\Phi(x') - \Phi(x^*)\| \\ &\leq \delta \cdot \|x' - x^*\| \end{aligned}$$

- Abschätzungen: 1,2 folgt aus der Abschätzung zur Cauchy-Folge mit $n \to \infty$ und pasender Wahl von m. Dritte Gleichung gilt wegen

$$\begin{aligned} \|x^k - x^*\| &= \|\Phi(x^{k-1}) - \Phi(x^*)\| \\ &\leq \delta \cdot \|x^{k-1} - x^*\| \\ &\leq \frac{\delta}{1 - \delta} \cdot \|x^{k-1} - x^k\| \end{aligned}$$

Korollar Sei Φ hinreichend glatt (nichtlinear) und $\Phi^{(j)}(x^*) = 0$ für $j = 1, \ldots, p$.

$$\|x^{k+1} - x^k\| \leq C \cdot \|x^k - x^*\|^{p+1}$$

(Nicht anwendbar für (affin) lineare Abbildungen, da $\varphi^{(1)}(x) = T \neq 0$)

3.3.2 Konvergenzaussagen

Satz Für streng diagonalisierbare Matrizen A ist sowohl das Jacobi-Verfahren als auch das Gauß-Verfahren durchführbar und konvergent.

Beweis:

- Falls

$$0 \leq \sum_{i=1, i \neq j}^{N} |a_{ij}| < |a_{jj}| \qquad \forall j = 1, \ldots, N$$

Also $|a_{jj}| > 0$. Also Verfahren durchführbar, weiter $\|T_J\|_\infty < 1$, d.h. $\varrho(T) \leq \|T_J\|_\infty < 1$.

- Behauptung: $\|T_{GS}\|_\infty < 1$. Dazu $v := T_{GS} \cdot w$

$$\Leftrightarrow \sum_{j \le i} a_{ij} \cdot v_j = -\sum_{j>i} a_{ij} \cdot w_j \qquad i = 1, \ldots, N$$

Wir zeigen rekursiv:

$$|v_l| \le \sigma \cdot \|w\|_\infty$$

wobei

$$\sigma := \max_{1 \le i \le n} \frac{1}{|a_{ii}|} \cdot \sum_{j \ne i} |a_{ij}| < 1$$

für $l = 1, \ldots, n$. Offensichtlich:

$$|v_1| \le \sigma \cdot \|w\|_\infty$$

. Es gelte $|v_l| \le \sigma \cdot \|w\|_\infty$ für $l = k$. Dann gilt für $l = k+1$:

$$\begin{aligned} a_{k+1,k+1} &= -\sum_{j>k+1} a_{k+1,j} \cdot w_j - \sum_{j \le k} a_{k+1,j} \cdot v_j \\ |v_{k+1}| &\le \underbrace{\left(\frac{1}{|a_{k+1,k+1}|} \cdot \sum_{j \ne k+1} |a_{k+1,j}| \right)}_{\le \sigma} \cdot \|w\|_\infty \end{aligned}$$

Bemerkung:

1. Eigentlich gezeigt wurde im zweiten Teil:

$$\|T_{GS}\|_\infty \le \|T_J\| < 1$$

schnellere Konvergenz des Gauss-Seidel-Verfahrens ist zu erwarten, vgl. *Plato Theorem 10.24*

2. Modifikation für Jacobi-Verfahren: Jacobi-Verfahren ist konvergent, falls A^T streng diagonaldominant, d.h.

$$\sum_{j \ne k}^{N} |a_{jk}| < |a_{kk}| \qquad k = 1, \ldots, N$$

Dann

$$\|(L+R) \cdot D^{-1}\|_1 < 1$$

$T_J = D^{-1} \cdot (L+R)$ und $(L+R) \cdot D^{-1}$ sind ähnlich ($A \sim B$, falls $\exists S : B = S \cdot A \cdot S^{-1}$). Also:

$$\begin{aligned} \sigma(T_J) &= \sigma((L+R) \cdot D^{-1}) \\ \Rightarrow \varrho(T_J) &= \varrho((L+R) \cdot D^{-1}) \\ &\le \|(L+R) \cdot D^{-1}\|_1 < 1 \end{aligned}$$

3. Es ist möglich, dass $\varrho(T_{GS}) < 1$ (bzw. $\varrho(T_J) < 1$), aber $\|T_{GS}\|_{1,2,\infty} \ge 1$ (bzw. $\|T_J\|_{1,2,\infty} \ge 1$), d.h. „gängige“ Matrixnormen lassen die Konvergenz nicht erkennen. Auch möglich: Gauß-Seidel-Verfahren konvergent und Jacobi-Verfahren nicht (oder umgekehrt).

Abschwächung der (strengen) Voraussetzungen

- Definition: A heißt (schwach) diagonaldominant, falls gilt:

$$\sum_{i \ne j} |a_{ij}| \le |a_{jj}| \qquad i = 1, \ldots, N$$

mit $l \in \{1, \ldots, N\}$:

$$|a_{ll}| > \sum_{i \ne l} |a_{il}|$$

- Definition: Graph von A: Jeder Variablen x_j wird Knoten p_j zugeordnet. Kante (Bogen) $p_j \to p_k$, falls ein i existiert mit $a_{ij} \neq 0$, $a_{ik} \neq 0$. A heißt nicht zerfallend, wenn der Graph von A zusammenhängend ist. (Äquivalente Definition: A ist irreduzibel)
- **Satz** Ist A schwach diagonaldominant und nicht zerfallend, so gilt $\varrho(T_{GS}) < 1$ und $\varrho(T_J) \leq \|T_J\|_\infty < 1$, also beide Verfahren konvergieren.
- Beispiele:
  1. äquidistante Diskretisierung mit finite Differenzen von
  $$-u''(x) = f(x) \qquad u(0) = u(1) = 0$$
  führt auf $A \cdot u = \hat{f}$ mit $u, \hat{f} \in \mathbb{R}^{n-1}$, $u_i \approx u(i \cdot h)$ für $i = 1, \ldots, N-1$ und $h = \frac{1}{N}$.
  $$\begin{aligned} a_{ii} &= 2 \\ a_{i,i-1} &= a_{i-1,i} = -1 \\ a_{ij} &= 0 \text{ für } |i-j| > 1 \end{aligned}$$
  A schwachdominant und nicht zerfallend. Also Gauss-Seidel-Verfahren und Jacobi-Verfahren konvergieren.
  2. Sei
  $$-u''(x) + c(x) \cdot u(x) = f(x) \qquad c(x) > 0$$
  (Reaktions-Diffusions-Gleichung). Randbedingungen $u(0) = u(1) = 0$. Analoge Diskretisierung wie im Beispiel 1 liefert
  $$\tilde{A} \cdot u = \hat{f}$$
  mit $\tilde{a}_{ii} = 2 + c(i \cdot h) \cdot h^2$. $\tilde{A}$ ist streng diagonaldominant.
  3. Gleichung:
  $$-\Delta u = f \qquad u|_\Gamma = 0$$
  für $\Omega = (0,1)^2$, $\Gamma = \partial(\Omega)$ mit $\Delta u = \text{div}(\text{grad(u)})$. Dann $\forall i, j \in \{1, \ldots, N-1\}$:
  $$\begin{aligned} 4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} &= h^2 \cdot f_{ij} \\ u_{0j} = u_{Nj} = u_{i0} = u_{iN} &= 0 \end{aligned}$$
  System von $(N-1)^2$ Gleichungen für $u_{ij} \approx u(x_i, y_j)$ mit $x_i = i \cdot h$, $y_j = j \cdot h$, Struktur von A:
  $$A = \begin{pmatrix} P & -E & 0 & 0 \\ -E & P & -E & 0 \\ 0 & -E & P & -E \\ 0 & 0 & -E & P \end{pmatrix} \qquad P := \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}$$
  A ist schwach diagonaldominant und besitzt Bandstruktur.
  4. zu Beispiel 1: Es gilt $\|T_{GS}\|_\infty \leq 1 - 2^{1-N}$; Gauss-Seidel-Verfahren liefert
  $$x_i^{k+1} = \frac{1}{2}(x_{i-1}^{k+1} + x_{i+1}^k + b_i) \qquad i = 1(1)N$$
  Sei $v = T \cdot u$, dann gilt:
  $$\begin{aligned} v_i &= \frac{1}{2}(v_{i-1} + u_{i+1}) \qquad v_0 = 0 \\ |v_1| &= \frac{1}{2}|u_1| \leq \frac{1}{2}\|u\|_\infty \\ |v_2| &= \frac{1}{2}(v_1 + u_2) \leq \frac{1}{2} \cdot \left(\frac{1}{2} + 1\right) \cdot \|u\|_\infty \\ |v_j| &= \frac{1}{2}(v_{j-1} + u_{j+1}) \leq \frac{1}{2} \cdot \left(\sum_{k=0}^{j-1} \left(\frac{1}{2}\right)^k\right) \cdot \|u\|_\infty \\ &= \left(1 - \left(\frac{1}{2}\right)^{j-1}\right) \cdot \|u\|_\infty \end{aligned}$$

Verbesserung durch Spektralnorm

$$\|T_{GS}\|_2 = 1 - O(N^{-2})$$

(analog für T_J).

$$\begin{aligned} \|T_h\| &\approx 1 - h^2 \\ \frac{\|T_h\|^k}{1 + \|T_h\|} &= h^{-2} \cdot (1 - h^2)^k \overset{!}{<} \varepsilon \\ \Rightarrow k &> \frac{\ln(\varepsilon \cdot h^2)}{\ln(1 - h^2)} \approx h^{-2} \cdot |\ln(\varepsilon \cdot h^2)| \end{aligned}$$

Für $\varepsilon = 10^{-4}$, $h = 10^{-2}$: $k > 10^4 \cdot |\ln(10^{-8})| \approx 1,8 \cdot 10^5$. Also extrem langsame Konvergenz. Entweder Konvergenzverbesserung oder generell andere iterative Verfahren.

3.3.3 Konvergenzbeschleunigung durch Relaxationsverfahren

- Idee: Die mittels Gauß-Seidel-Verfahren berechneten Werte werden als Hilfsgrößen interpretiert und mit den Werten der vorhergehenden Iteration „interpoliert“.
- Iterationsvorschrift:

$$\begin{aligned} \hat{x}^{(l+1)} &= \frac{1}{a_{jj}} \cdot \left(b_j - \sum_{k=1}^{j-1} a_{jk} \cdot x_k^{(l+1)} - \sum_{k=j+1}^{n} a_{jk} \cdot x_k^{(l)} \right) \qquad j = 1, \ldots, N \\ x_j^{(l+1)} &= w \cdot \hat{x}^{(l+1)} + (1 - w) \cdot x_j^{(l)} \qquad w \in \mathbb{R} \end{aligned}$$

  Zusammengefasst:

$$a_{jj} \cdot x_j^{(l+1)} = w \cdot \left(b_j - \sum_{k=1}^{j-1} a_{jk} \cdot x_k^{(l+1)} - \sum_{k=j+1}^{n} a_{jk} \cdot x_k^{(l)} \right) + (1 - w) \cdot x_j^{(l)} \cdot a_{jj}$$

  Matrixschreibweise:

$$(D + w \cdot L) \cdot x^{(l+1)} = w \cdot b + ((1 - w) \cdot D - w \cdot R) \cdot x^{(l)}$$

  bzw.

$$T(w) = (D + w \cdot L)^{-1} \cdot ((1 - w) \cdot D - w \cdot R) = (E + w \cdot D^{-1} \cdot L)^{-1} \cdot ((1 - w) \cdot E - w \cdot D^{-1} \cdot R)$$

- **Satz** Es gilt:

$$\varrho(T(w)) \geq |w - 1|$$

  für alle $w \in \mathbb{R}$.

  Beweis:

$$\begin{aligned} \prod_{i=1}^{n} \lambda_i &= \det(T(w)) \\ &= \underbrace{\det((E + w \cdot D^{-1} \cdot L)^{-1})}_{1} \cdot \underbrace{\det(((1 - w) \cdot E) - w \cdot D^{-1} \cdot R)}_{((1-w)^n} \\ &= (1 - w)^n \end{aligned}$$

- **Folgerung** Das Relaxationsverfahren ist höchstens für $w \in (0, 2)$ konvergent.
- Bemerkungen:

1. Für $w > 1$: Überrelaxation, für $w < 1$: Unterrelaxation
2. Ziel: Wahl eines optimalen w_0 mit
$$\varrho(T(w_0)) = \min_{w \in (0,2)} \varrho(T(w))$$
3. Für Beispiel 1 aus letztem Abschnitt: Die zugehörige Systemmatrix gehört in die Klasse von Matrizen für die eine optimale Wahl w_0 bekannt ist.

3.3.4 CG-Verfahren (conjugate gradients)

- Gleichung:
$$A \cdot x = b \qquad (1)$$
mit A positiv definit und symmetrisch ($A = A^T$ und $(Ax, x) \geq \gamma \cdot \|x\|^2$ für alle $x \in \mathbb{R}^n$ mit $\gamma > 0$). A ist selbstadjungiert bzgl. des euklidischen Skalarprodukts, d.h.
$$\forall x, y \in \mathbb{R}^n : (Ax, y)_2 = (x, Ay)_2$$
Idee: Zuordnung eines Minimierungsproblems
$$f(x) = \frac{1}{2}(Ax, x) - (b, x)$$
mit $\Delta f(x) = Ax - b$. Damit:
$$\Delta f(\bar{x}) = 0 \Leftrightarrow \bar{x} \text{ löst } (1)$$
f ist konvex.

- **Definition** Das System der Vektoren $\{p_i\}_{i=1}^k$ heißt A-adjungiert (A-konjugiert), falls gilt:
$$\begin{aligned} (Ap^j, p^j) &\neq 0 \\ (Ap^i, p^j) &= 0 \qquad i \neq j \text{ A-orthogonal} \end{aligned}$$

- Zwischenbetrachtung: Aufgabe:
$$f(x) \to \min$$
bei $x \in M_k$,
$$M_k := \{x = \bar{x} + \sum_{j=1}^{k} \alpha_j \cdot p^j; \alpha_j \in \mathbb{R}\} = [\bar{x}] \oplus V_k$$
mit
$$V_k := \text{span}\{p_i\}_{i=1}^k$$
Dann:
$$\begin{aligned} f(x) &= \frac{1}{2}\left(A\left(\bar{x} + \sum_{j=1}^{k} \alpha_j \cdot p^j\right), \left(\bar{x} + \sum_{j=1}^{k} \alpha_j \cdot p^j\right)\right) - \left(b, \left(\bar{x} + \sum_{j=1}^{k} \alpha_j \cdot p^j\right)\right) \\ &= \frac{1}{2}(A\bar{x}, \bar{x}) - (b, \bar{x}) + \left(\sum_{j=1}^{k} \alpha_j \cdot (A\bar{x} - b, p^j) + \frac{1}{2}\alpha_j^2 \cdot (Ap^j, p^j)\right) \\ \Rightarrow \alpha_j^* &= -\frac{(A\bar{x} - b, p^j)}{(Ap^j, p^j)} \end{aligned}$$
Definiere
$$\bar{d} = b - A\bar{x} = -\Delta(\bar{x})$$
(Richtung des steilsten Abstieges). Mit $v^k = \sum_{j=1}^{k} \alpha_j^* \cdot p^j$ folgt:
$$\begin{aligned} \frac{1}{2}(A(\bar{x} + b), \bar{x} + v) - (b, \bar{x} + v) &\to \min \qquad v \in V_k \\ \Leftrightarrow (A(x + v^k) - b, v) &= 0 \\ \Leftrightarrow (\bar{d} - Av^k, v) &= 0 \qquad \forall v \in V_k \end{aligned}$$

- Zwischenbetrachtung: Orthogonalisierung von d^{k+1} (bzgl. $V_k = \text{span}\{d_1, \ldots, d_k\}$). Bestimme $q^k \in V_k$ so, dass

$$(A\underbrace{(d^{k+1} + q^k)}_{=:p^{k+1}}, q^j) = 0 \qquad \forall j = 1, \ldots, k$$

Annahme:

$$\begin{aligned}(Ap^{k+1}, p^{k+1}) = 0 \quad &\Leftrightarrow \quad p^{k+1} = 0 \\ &\Leftrightarrow \quad d^{k+1} = -q^k \in V_k = \text{span}\{p^j\}_{j=1}^k\end{aligned}$$

Dann folgt mit

$$\begin{aligned}(d^{k+1}, v) &= (b - Ax^k - Av^k, v) = (d^k - A^k, v) \\ &= 0 \; \forall v \in V_k\end{aligned}$$

dass $d^{k+1} = 0$. Also x^{k+1} Lösung von (1).

- Zur Berechnung von $q^k \in V_k$: Schmidtsches Orthogonalisierungsverfahren:

$$q_k = \sum_{j=1}^{k} \beta_{kj} \cdot p^j$$

Dann:

$$(Ad^{k+1}, p^i) + \sum_{j=1}^{k} (\beta_{kj} \cdot Ap^j, p^i) \stackrel{!}{=} 0 \qquad \forall i = 1, \ldots, k$$

Dann:

$$\beta_{kj} = -\frac{(Ad^{k+1}, p^j)}{(Ap^j, p^j)} = \frac{(d^{k+1}, Ap^j)}{Ap^j, p^j)}$$

mit

$$\begin{aligned}Ap^j &= \frac{1}{\alpha_j^*} \cdot A \cdot (x^{j+1} - x^j) \\ &= \frac{1}{\alpha_j^*} \cdot (d^j - d^{j+1})\end{aligned}$$

Dann:

$$\begin{aligned}\beta_{kj} &= \frac{1}{\alpha_j^*} \cdot \frac{(d^{k+1}, d^{j+1} - d^j)}{(Ap^j, p^j)} \\ \Rightarrow \beta_{kj} &= 0 \qquad \forall j < k\end{aligned}$$

da $(d^{k+1}, p^j) = 0$ für $j = 1, \ldots, k$. Damit:

$$\beta_{kk} = \frac{1}{\alpha_k^*} \cdot \frac{(d^{k+1}, d^{k+1})}{(Ap^k, p^k)} = \frac{(d^{k+1}, d^{k+1})}{(d^k, p^k)}$$

- Zusammenfassung: CG-Verfahren im quadratischen Fall:
Berechne von einem $x^1 \in \mathbb{R}^n$ ausgehend mit

$$p^1 := d^1 := b - A \cdot x^1$$

rekursiv die folgenden Größen für $k = 1, \ldots$ solange $d^k \neq 0$:

$$\begin{aligned}\alpha_k^* &:= \frac{(d^k, d^k)}{(Ap^k, p^k)} \\ x^{k+1} &:= x^k + \alpha_k^* \cdot p^k \\ d^{k+1} &:= b - A \cdot x^{k+1} \\ \beta_k &:= \frac{(d^{k+1}, d^{k+1})}{(d^k, p^k)} \\ p^{k+1} &:= d^{k+1} + \beta_k \cdot p^k\end{aligned}$$

- **Satz** Nach maximal n Schritten liefert das CG-Verfahren bei exakter Rechnung die Lösung x^* von (1) (mit $A \in \mathbb{R}^{n \times n}$).

- Bemerkung:

  1. Falls (numerisch) $d^n \neq 0$: Verfahren fortsetzen. Aber: Das konterkariert (wieder) die Grundidee iterativer Verfahren ($k << N$).

- **Satz** Für die Konvergenzgeschwindigkeit des CG-Verfahrens (gegen Lösung x^* von (1)) gilt die Abschätzung:

$$\begin{aligned}(A(x^{k+1} - x^*), x^{k+1} - x^*) &\leq 2\left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\right)^k \cdot (A(x^1 - x^*), x^1 - x^*) \\ \|x^{k+1} - x^*\|_A^2 &\leq 2\underbrace{\left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\right)^k}_{\gamma^k} \cdot \|x^1 - x^*\|_A^2\end{aligned}$$

  mit

$$(Ax, x) \geq m \cdot \|x\|_2^2 \qquad M \cdot \|x\|_2^2 \geq (Ax, x)$$

  Für $A = A^T$, positiv definit:

$$m := \min\{\lambda_i\} \qquad M = \max\{\lambda_i\}$$

  Problem: z.B. für $m = 1, M = 10^4$:

$$\gamma = \frac{99}{101}$$

  also sehr schlechte Konvergenzgeschwindigkeit.

- Bemerkung:

  1. Für A regulär, $A \neq A^T$ forme um:

$$A^T \cdot A \cdot x = A^T \cdot b$$

  Dann $A^T \cdot A$ positiv definit, symmetrisch, aber mit $\kappa_0 = \text{cond}(A)_2$ gilt:

$$\gamma = \frac{\kappa_0 - 1}{\kappa_0 + 1} \qquad \text{mit } k_0 >> \sqrt{k_0} >> 1$$

  Also schlechte Konvergenzgeschwindigkeit. (Schlechter als oben, dort gilt:

$$\gamma = \frac{\sqrt{\kappa_0} - 1}{\sqrt{\kappa_0} + 1}$$

3.3.5 Konvergenzverbesserung durch Vorkonditionierung (PCG)

- Verfahrensidee: $B = B^T$ positiv definit mit B^{-1} einfach zu berechnen ($B \cdot y = d$ einfach lösbar), $B \approx A$. Dann:

$$\begin{aligned} A \cdot x &= b \\ \Rightarrow \underbrace{B^{-1} \cdot A}_{=:\tilde{A}} \cdot x &= \underbrace{B^{-1} \cdot b}_{\tilde{b}} \end{aligned}$$

  Durch $B^{-1} \cdot A$ sollen Eigenwerte zu 1 „verschoben“ werden: $\lambda_i \approx 1$. Damit dann bessere Konvergenzgeschwindigkeit.

- $\tilde{A}$ ist bzgl. $(.,.)_B$ selbstadjungiert:

$$\begin{aligned} (\tilde{A} \cdot x, y)_B &= (B^{-1} \cdot A \cdot x)^T \cdot B \cdot y \\ &= x^T \cdot A^T \cdot (B^{-1})^T \cdot B \cdot y = x^T \cdot A^T \cdot y \\ &= y^T \cdot A^T \cdot (B^{-1})^T \cdot B \cdot x \\ &= \ldots = (x, \tilde{A} \cdot y)_B \end{aligned}$$

- Transformation auf Originalaufgabe:

$$\begin{aligned} \tilde{\alpha}_k &:= \frac{(B^{-1} \cdot d^k)^T \cdot d^k}{(B^{-1} \cdot A \cdot p^k)^T \cdot p^k} \\ x^{k+1} &:= x^k + \tilde{\alpha}_k \cdot p^k \\ d^{k+1} &:= b - A \cdot x^{k+1} \\ \tilde{\beta}_k &:= \frac{(B^{-1} \cdot d^{k+1})^T \cdot d^{k+1}}{(B^{-1} \cdot d^k)^T \cdot d^k} \\ p^{k+1} &:= B^{-1} \cdot d^{k+1} + \tilde{\beta}_k \cdot p^k \end{aligned}$$

- Bemerkungen:

  1. $A = A^T$ positiv definit, dann $a_{ii} > 0$ für alle i: Diagonalkonditionierer
  2. Für $A \neq A^T$: $A \cdot x = b$ nicht unbedingt symmetrisieren. Iteratives Konzept: x^k bestimmen aus

$$\|A \cdot x^k - b\|_2^2 = \min_{x \in \kappa_k} \|A \cdot x - b\|_2^2$$

  z.B. Krylov-Unterraum-Verfahren:

$$\kappa_k := \text{span}\{b, A \cdot b, \ldots, A^{k-1} \cdot b\}$$

4

Iterationsverfahren zur Lösung nichtlinearer Gleichungssysteme

- Seien $\varphi, F : \mathbb{R}^n \to \mathbb{R}^n$ gegeben. Gesucht ist die Lösung von $F(x) = 0$ (1) bzw. $x = \varphi(x)$ (2).
- Sukzessive Behandlung von (1) unter Verwendung von Vereinfachungen x^k gegeben,

  $$F(x) = F(x^k) + F'(x_k) \cdot (x - x_k)$$

  falls $\|x - x_k\| < \delta$.
- Newton-Verfahren zur Nullstellenberechnung:

  $$F(x^k) + F'(x_k) \cdot (x - x_k) \stackrel{!}{=} 0$$

  x^{k+1} als Nullstelle der linearisierten Funktion:

  $$\begin{aligned} x^{k+1} &= x^k - (F'(x_k))^{-1} \cdot F(x_k) \qquad k \in \mathbb{N}_0 \\ (\hat{=} x^{k+1} &= \varphi(x_k)) \end{aligned}$$

  mit $\varphi(x) = x - (F'(x))^{-1} \cdot F(x)$.
- Eine Folge $\{x^k\}$ heißt Q-linear bzw. Q-überlinear gegen x^* konvergent, falls

  $$\|x^{k+1} - x^*\| \leq \varrho \cdot \|x^k - x^*\| \qquad k \in \mathbb{N}_0$$

  mit $\varrho \in (0, 1)$ bzw.

  $$\|x^{k+1} - x^*\| \leq \varrho_k \cdot \|x^k - x^*\| \qquad k \in \mathbb{N}_0$$

  mit $\varrho_k \to 0$.
- Die Folge besitzt die Konvergenzordnung $p > 1$, falls diese (mindestens) Q-linear gegen x^* konvergiert und zusätzlich

  $$\|x^{k+1} - x^*\| \leq C \cdot \|x^k - x^*\|^p$$

  mit $C > 0$ erfüllt ist.

4.1 Konvergenz von Fixpunktiterationsverfahren

Satz Sei (X,d) ein vollständiger metrischer Raum und $\varphi : X \to X$ kontrahierend, d.h. es gibt $\varrho \in (0, 1)$ mit

$$d(\varphi(x), \varphi(y)) \leq \varrho \cdot d(x, y)$$

für alle $x, y \in X$. Dann konvergiert für beliebiges $x^0 \in X$ die durch $x^{k+1} = \varphi(x^k)$ erzeugte Folge $\{x^k\}$ gegen einen eindeutigen Fixpunkt $x^* \in X$ von φ. Es gelten (sinngemäß) die bereits formulierten Abschätzungen.

Bemerkungen:

1. Jede abgeschlossene Teilmenge eines linearen normierten vollständigen Raumes ist ein vollständiger metrischer Raum mit $d(x,y) := \|x-y\|$. Metrik heißt translationsinvariant, falls

$$\forall x,y,z : d(x+z, y+z) = d(x,y)$$

bzw. positiv homogen, falls

$$d(\alpha \cdot x, \alpha \cdot y) = |\alpha| \cdot d(x,y)$$

Dann $d(x,0) =: \|x\|$.

2. In-sich-Abbildung ist wesentlich:

$$x = \sqrt{x} \qquad x^* = 1$$

Dann für $M_1 := \left[\frac{1}{2}, \infty\right)$ φ kontrahierend, aber für $M_\varepsilon = [1+\varepsilon, \infty)$ ist $\varphi(M_\varepsilon) \not\subseteq M_\varepsilon$.

Beispiele:

1. Für

$$x^3 + 2x - \cos x = 0$$

definiere

$$x = \frac{\cos x}{2 + x^2} =: \varphi(x)$$

2. Für

$$x \cdot e^x = 1$$

verschiedene Iterationen möglich:

$$\begin{aligned} x &= e^{-x} =: \varphi_1(x) \\ x &= x + \alpha \cdot (x \cdot^x -1) =: \varphi_2(x) \\ x &= x + \alpha(x) \cdot (x \cdot e^x - 1) =: \varphi_3(x) \qquad \alpha(x) = -\frac{1}{f'(x)} \end{aligned}$$

wesentliche Unterschiede in der (für fixierte Genauigkeit) erforderlichen Zahl der Iterationen.

(lokale) Modifikationen des Banachschen Fixpunktsatzes

Satz Es sei x^* Fixpunkt von φ und es existiert $r > 0$ mit

$$\forall x, y \in U_r(x^*) : \|\varphi(x) - \varphi(y)\| \leq \varrho \cdot \|x - y\| \qquad (2)$$

mit $\varrho \in (0,1)$. Dann konvergiert die Fixpunktiteration für beliebige $x_0 \in U_r(x^*)$ gegen x^*.

Beweis:

- Es gilt:

$$\|x^{k+1} - x^*\| \leq \varrho \cdot \|x^k - x^*\|$$

Damit: Für $x^k \in U_r(x^*)$ ist $x^{k+1} \in U_r(x^*)$, also

$$\|x^{k+1} - x^*\| \leq \varrho^{k+1} \cdot \|x^0 - x^*\|$$

Beispiel:

1. $\varphi : \mathbb{R} \to \mathbb{R}$ sei stetig differenzierbar mit $|\varphi'(x)| < 1$. Dann gibt es ein $\varrho \in (0,1)$ und $r > 0$ derart, dass (2) gilt.

Beweis:

- Mittelwertsatz:
$$\varphi(x) - \varphi(y) = \varphi'(\xi) \cdot (x - y)$$
Mit Stetigkeit von φ' folgt: Es existiert $r > 0$ mit $|\varphi'(\xi)| \leq \varrho < 1$ für alle $\xi \in U_r(x^*)$. Also
$$\forall x, y \in U_r(x^*) : |\varphi(x) - \varphi(y)| \leq |\varphi'(\xi)| \cdot |x - y|$$

Bemerkung:

- Konvergenz gut, falls $|\varphi'(x^*)|$ klein bzw. falls
$$|x^{k+1} - x^*| \leq |\varphi'(\xi_k)| \cdot |x^k - x^*|$$
mit $|\varphi'(\xi_k)| \to 0$ für $k \to \infty$, dann Q-überlineare Konvergenz.

2. Im eindimensionalen Fall:
$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$
mit $f(x^*) = 0$, $f'(x^*) \neq 0$. Dann $\varphi'(x^*) = 1 - 1 = 0$.

Satz (lokaler Konvergenzsatz) Zu gegebenem $x^0 \in \mathbb{R}^n$ sei $U_r(x^0)$ eine zugehörige Umgebung ($r > 0$). Weiter sei $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ eine Abbildung, die mit monoton wachsenden Funktionen $s, d : \mathbb{R}^+ \to \mathbb{R}^+$ der Abschätzung
$$\|\varphi(x) - \varphi(y)\| \leq (s(\|\varphi(x) - x\|) + d(\|x - y\|)) \cdot \|x - y\|$$
für alle $x, y \in U_r(x^0)$ genügt. Gilt zusätzlich

1. $\delta_0 := s(\|\varphi(x^0) - x^0\|) + d(\|\varphi(x^0) - x^0\|) < 1$
2. $\frac{1}{1-\delta_0} \cdot \|\varphi(x^0) - x^0\| < r$

so konvergiert $\{x^k\}$ gegen einen Fixpunkt x^* von φ.

Beweis:

1. φ ist stetig, denn nach 1) gilt:
$$\|\varphi(x) - \varphi(y)\| \leq (s(\|\varphi(x) - x\|) + d(2r)) \cdot \|y - x\|$$
Dann für $y \to x$ folgt $\varphi(y) \to \varphi(x)$.

2. Wir zeigen nun rekursiv:
$$\begin{aligned} x^k &\in U_r(x^0) \\ \|x^{k+1} - x^k\| &\leq \delta_0 \cdot \|x^k - x^{k-1}\| \qquad k \in \mathbb{N}_0 \end{aligned}$$
Induktionsanfang: k=1. Wegen 2. gilt:
$$\begin{aligned} \|x^1 - x^0\| &= \|\varphi(x^0) - x^0\| \leq (1 - \delta_0) \cdot r < r \\ \Rightarrow x^1 &\in U_r(x^0) \end{aligned}$$
Außerdem:
$$\begin{aligned} \|x^2 - x^1\| &= \|\varphi(x^1) - \varphi(x^0)\| \\ &\leq (s(\|\varphi(x^0) - x^0\|) + d(\|x^1 - x^0\|)) \cdot \|x^1 - x^0\| \\ &= \delta_0 \cdot \|x^1 - x^0\| \end{aligned}$$
Induktionsschritt: Es gelte die Behauptung für $k = 1, \ldots, l$. Dann gilt:
$$\begin{aligned} \|x^{l+1} - x^0\| &\leq \frac{1}{1 - \delta_0} \cdot \|x^1 - x^0\| < r \\ \Rightarrow x^{l+1} &\in U_r(x^0) \end{aligned}$$

Außerdem:

$$\begin{aligned}\|x^{l+2}-x^{l+1}\| &= \|\Phi(x^{l+1})-\Phi(x^l)\| \\ &\leq (s(\|\Phi(x^l)-x^l\|)+d(\|x^{l+1}-x^l\|))\cdot\|x^{l+1}-x^l\| \\ &\leq \delta_0\cdot\|x^{l+1}-x^l\|\end{aligned}$$

mit $\|x^{l+1}-x^l\| \leq \|x^1-x^0\|$ und Monotonie von s,d. Also $\{x^k\}$ Cauchy-Folge, $x_k \to x^*$,

$$\|x^*-x^0\| \leq \frac{1}{1-\delta_0}\cdot\|x^n-x^0\| < r \qquad (x^* \in U_r(x^0))$$

Wegen Stetigkeit von φ in $U_r(x^0)$: $x^* = \varphi(x^*)$.

Bemerkungen:

1. Der Satz gibt mit (überprüfbaren) lokalen Informationen auch eine Existenzaussage.
2. Falls φ auf $U_r(x^0)$ Lipschitz-stetig differenzierbar, dann

$$\begin{aligned}\varphi(x)-\varphi(y) &= \int_0^1 \varphi'(x+t\cdot(x-y))\cdot(x-y)\,dt \\ \Rightarrow \|\varphi(x)-\varphi(y)\| &\leq \left(\|\varphi'(x)\|+\int_0^1 L\cdot t\cdot\|x-y\|\;dt\right)\cdot\|x-y\|\end{aligned}$$

4.2 Newton- und Quasi-Newton-Verfahren

$$F(x^k)+F'(x^k)\cdot(x^{k+1}-x^k)=0 \qquad k\in\mathbb{N}_0$$

Anwenden des lokalen Konvergenzsatzes: $x \to \varphi(x)$:

$$F(x)+F'(x)\cdot(\varphi(x)-x) \stackrel{!}{=} 0$$

und mit $y \to \varphi(y)$:

$$\begin{aligned}F(y)+F'(y)\cdot(\varphi(y)-y) &\stackrel{!}{=} 0 \\ \Rightarrow F(x)-F(y)+F'(y)\cdot(\varphi(x)-\varphi(y)-(x-y))+\ldots & \\ +(F'(x)-F'(y))\cdot(\varphi(x)-x) &= 0 \\ \Leftrightarrow (F'(x)-F'(y))\cdot(\varphi(x)-x)+F(y)-F(x)+F'(y)\cdot(x-y) &= F'(y)\cdot(\varphi(x)-\varphi(y)) \qquad (1)\end{aligned}$$

Satz Es sei F auf $U_r(x^0)$ Lipschitz-stetig differenzierbar und es gelte:

1. $\|F'(z)\cdot w\| \geq m\cdot\|w\|$ für alle $z\in U_r(x^0)$, $w\in\mathbb{R}^n$ mit $m>0$
2. $\delta_0 := \frac{3}{2}\cdot\frac{L}{m}\cdot\|x^1-x^0\| < 1$
3. $\|x^1-x^0\| < (1-\delta_0)\cdot r$

Dann ist das Newton-Verfahren (Start: x^0) vollständig durchführbar und die erzeugte Folge konvergiert R-quadratisch gegen eine Nullstelle x^* von F.

Beweis:

- Aus (1) folgt mit 1.

$$\|\varphi(x)-\varphi(y)\| \leq \frac{L}{m}\cdot\left(\|\varphi(x)-x\|+\frac{1}{2}\cdot\|x-y\|\right)\cdot\|x-y\| \qquad (2)$$

(Vgl. Beweis lokaler Konvergenzsatz: $s=\frac{L}{m}$, $d(t)=\frac{L}{2m}\cdot t$) Also $\{x^k\}$ konvergent: $x^k\to x^*$. Wenn x^* Fixpunkt von φ:

$$\begin{aligned}F(x^*)+F'(x^*)\cdot(\varphi(x^*)-x^*) &= 0 \\ \Rightarrow F(x^*) &= 0\end{aligned}$$

Mit $x = x^*$, $y = x^k$ in (2) folgt:

$$\|x^{k+1} - x^*\| \leq \frac{L}{2m} \cdot \|x^k - x^*\|^2$$

also quadratische Konvergenz.

Beispiel:

1. Sei
$$F(x_1, x_2) = \begin{pmatrix} e^{-x_1^2} - x_2 \\ \sinh(x_1) - (1 + x_2^2) \cdot x_2 \end{pmatrix} \qquad x^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
Dann:
$$\begin{aligned} F'(x_1, x_2) &= \begin{pmatrix} -2x_1 \cdot e^{-x_1^2} & -1 \\ \cosh(x_1) & -(1 + 3x_2^2) \end{pmatrix} \\ \Rightarrow F(x^0) &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad F'(x^0) = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \end{aligned}$$
Dann:
$$\begin{aligned} F(x^0) + F'(x^0) \cdot d^0 &\stackrel{!}{=} 0 \\ d^0 &= (1, 1)^T \\ \Rightarrow x^1 &= (1, 1)^T \end{aligned}$$
Nächste Iteration:
$$x^2 = \begin{pmatrix} 0,620232 \\ 0,647297 \end{pmatrix}$$
Endergebnis:
$$x^* = \begin{pmatrix} 0,7275 \\ 0,5890 \end{pmatrix}$$

Vereinfachung und Verallgemeinerung der Iterationsvorschrift

$$F(x^k) + A_k \cdot (x^{k+1} - x^k) = 0$$

mit geeigneten Matrizen A_k für $k \in \mathbb{N}_0$.

1. Modifiziertes Newton-Verfahren:
$$A_k := F'(x^0) \qquad k = 0(1)l$$
mit $l > 1$ fixiert. Entsprechend:
$$A_{l+1} := F'(x^l) \qquad k = l + 1(1)2l$$
F' approximiert mittels finiter Differenzen ($\rightarrow n^2$ partielle Ableitungen). Anzustreben: konsistente Approximierung
$$\lim_{k \to \infty} \|A_k - F'(x^k)\| = 0$$

2. Quasi-Newton-Verfahren:
Wahl von A_k derart, dass
$$A_k \cdot (x^k - x^{k-1}) = F(x^k) - F(x^{k-1}) \qquad k \in \mathbb{N}_0 \qquad (1)$$
mit Vorgabe von $A_0 (= F'(x^0))$. Beziehung (1) heißt Quasi-Newton-Gleichung. Basis für (1):
$$F'(x^k) \cdot (x^k - x^{k-1}) = F(x^k) - F(x^{k-1}) + o(\|x^k - x^{k-1}\|)$$

Spezialfall n=1: $A_k = (\alpha_k)$

$$\begin{aligned} \alpha_k \cdot (x^k - x^{k-1}) &= f(x^k) - f(x^{k-1}) \\ \alpha_k &= \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} \end{aligned}$$

bzw.

$$x^{k+1} = x^k - \frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})} \cdot f(x^k)$$

(Regula Falsi/Sekantenverfahren). Verallgemeinerung auf $\mathbb{R}^n$:

$$A_{k+1} \cdot d^l = p^l \qquad l = k - n + 1, \ldots, k \qquad (2)$$

falls $\{d^l\}_{l=1}^n$ Basis des $\mathbb{R}^n$ ist A_{k+1} eindeutig bestimmt.

$$d^k = x^k - x^{k-1} \qquad p^k = F(x^k) - F(x^{k-1})$$

Besser: „Rang1-Update“ in (1) wird nur für Schritt k erfüllt. Ansatz:

$$A = C + r \cdot s^T$$

Aus (2) (nur für $l = k$):

$$\begin{aligned} p &\stackrel{!}{=} A \cdot d = C \cdot d + r \cdot s^T \cdot d \\ \Rightarrow r &= \frac{p - C \cdot d}{s^T \cdot d} \qquad (s^T \cdot d) \neq 0, s \in \mathbb{R}^n \end{aligned}$$

Wir wählen speziell $C = A_k$ mit

$$p^k := F(x^k) - F(x^{k-1})$$

und

$$F(x^k) \neq A_k \cdot d^k \stackrel{!}{=} 0$$

Dazu:

$$A_{k+1} := A_k + \frac{F(x^k) \cdot s^{k^T}}{s^{k^T} \cdot d^k}$$

(Broyden-Verfahren) mit geeigneten Vektoren s^k ($k \in \mathbb{N}_0$) und $A_0 = F'(x^0)$ (auch möglich: $A_0 = E$)

Verfahren:

(a) Vorgabe von A_0, x^0,

(b) $F(x^k) + A_k \cdot d^k = 0$ für $k \in \mathbb{N}_0$

(c) $x^{k+1} := x^k + d^k$

(d) $A_{k+1} := A_k + \frac{F(x^{k+1}) \cdot d^{k^T}}{\|d^k\|_2^2}$

Satz Unter den hinreichenden Bedingungen zur Konvergenz des Newton-Verfahrens konvergiert die vom Broyden-Verfahren erzeugte Folge $\{x^k\}$ Q-überlinear gegen x^* (Nullstelle von F).

Bemerkung:

- Ist F' symmetrisch (z.B. Minimierungsprobleme), dann ist die symmetrische Variante des Broyden-Verfahrens zu empfehlen:

$$A_{k+1} := A_k + \frac{F(x^{k+1}) \cdot (F(x^{k+1}))^T}{(F(x^{k+1}))^T \cdot d^k}$$

Zur Globalisierung schneller lokaler Verfahren:

1. Gedämpfte Verfahren („$x^0 \notin U_r(x^*)$“)

$$x^{k+1} = x^k + \lambda_k \cdot d^k$$

(d^k aus $F(x^k) + A_k \cdot d^k = 0$). Klassisch: $\lambda_k = 1$. Dabei Schrittweite $\lambda > 0$ geeignet wählen, z.B.

$$\|F(x^k + \lambda_k \cdot d^k)\| \stackrel{!}{=} \min_{\lambda>0} \|F(x^k + \lambda \cdot d^k)\|$$

oder $\lambda_k \in \{\delta^j, \delta \in (0,1)\}$ derart, dass

$$\|F(x^k + \lambda_k \cdot d^k)\| < \|F(x^k)\|$$

evtl. Zusatzbedingung, z.B. Goldstein Armijo

Bemerkungen:

- Newton-Richtung ist Anstiegsrichtung für

$$\varphi(x) := \|F(x)\|_2^2$$

da

$$\varphi'(x)d = (F'(x)d)^T \cdot F(x) + F(x)^T \cdot F'(x) \cdot d$$

Mit $F'(x_n)d_n = -f(x)$:

$$\varphi'(x)d_n = -2\|F(x)\|_2^2$$

2. Einbettungsprinzip: $F(x) = 0$

Familie von Problemen

$$H : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n : H(x,t) = 0 \qquad t \in [0,1]$$

mit $H(x,1) = F(x)$ und es existiert $\tilde{x} : H(\tilde{x}, 0) = 0$. Berechnung (für alle $t \in [0,1]$)

$$H(x(t), t) = 0 \longrightarrow x^*$$

Vorgabe (Konstruktion) von Parameterfolge $\{t_k\}$, Behandlung von $H(x, t_k) = 0$. „Warmstart“:

$$x^0(t_k) = x(t_{k-1})$$

Beispiele:

(a) $H(x,t) = (1-t) \cdot (x - \tilde{x}) + t \cdot F(x)$

(b) $H(x,t) = F(x) - (1-t) \cdot F(\tilde{x})$

(c) $H(x,t) = (1-t) \cdot G(x) + t \cdot F(x)$ mit $G(\tilde{x}) = 0$ und $G(x) \approx F(x)$

5 Matrixeigenwertproblem

5.1 Charakterisierung von Eigenwerten und Eigenvektoren

- Sei $A \in \mathbb{R}^{n \times n}$ vorgegeben. Gesucht: $\lambda_i \in \mathbb{C}, x^i \in \mathbb{C}^n$,

$$\begin{aligned} A \cdot x^i &= \lambda_i \cdot x^i \\ p(\lambda) &= \det(A - \lambda \cdot E) \qquad p(\lambda_i) = 0 \end{aligned}$$

  Numerisch unbrauchbar wegen Komplexität und numerischer Instabilität.

- Bemerkung: Links-Eigenvektoren

$$(y^i)^T \cdot A = \lambda_i \cdot (y_i)^T$$

  ($\hat{=} y^i$ Eigenvektoren von A^T). Links- (und Rechts)-Eigenvektoren zu unterschiedlichen Eigenwerten sind zueinander orthogonal. (s. Bemerkung)

- Nach Hauptsatz der Algebra existieren (entsprechend der Vielfachheit) n Eigenwerte. Zu mehrfachen Eigenwerten (algebraische Vielfachheit) müssen nicht entsprechend viele linear unabhängige Eigenvektoren (geometrische Vielfachheit) existieren. Beispiel:

$$\begin{aligned} A &= \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} \\ \lambda_{1/2/3} &= 3 \qquad x^1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T \end{aligned}$$

  Somit zwei weitere Hauptvektoren:

$$(A - \lambda \cdot E)^{k+1} \cdot x = 0$$

  oder

$$(A - \lambda \cdot E) \cdot x^2 = x^1$$

  dann

$$(A - \lambda \cdot E) \cdot x^3 = x^2$$

  usw.

Satz Eigenwertschranken von Gershgorin: Sei λ Eigenwert von A und $\bar{x}$ der zugehörige Eigenvektor. Dann gilt:

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{jj}|$$

wobei i der Index ist mit

$$|x_i| = \max_{1 \leq j \leq n} |x_j|$$

Beweis:

- Es gilt: $A \cdot \bar{x} = \lambda \cdot \bar{x}$. Damit:

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} \cdot x_j &= \lambda \cdot x_i \\ \Rightarrow |\lambda - a_{ii}| \cdot |x_i| &\leq \sum_{j \neq i} |a_{ij}| \cdot |x_j| \\ \Rightarrow |\lambda - a_{ii}| &\leq \sum_{j \neq i} a_{ij}| \end{aligned}$$

  wegen $|x_i| \neq 0$ ($\bar{x} \neq 0$).

Extremalcharakterisierung von Eigenwerten im symmetrischen Fall:

- Sei $A = A^T$, $F(x) := x^T \cdot A \cdot x$. Aufgabenstellung:

$$F(x) \to \min_{\|x\|=1} \qquad (1)$$

  Nach Weierstraß existiert Lösung von (1) ($:= x^1$). Nebenbedingung:

$$g(x) = 1 - \|x\|_2^2$$

  Lagrange-Multiplikatoren-Regel: Es existiert $\lambda_1 \in \mathbb{R}$ mit

$$\operatorname{grad} F(x^1) + \lambda_1 \cdot \operatorname{grad} g(x^1) = 0$$

  (Kuhn-Tucker-Bedingung) mit

$$\operatorname{grad} F(x^1) = 2A \cdot x^1$$

  folgt:

$$A \cdot x^1 = \lambda_1 \cdot x^1$$

  λ_1 ist kleinster Eigenwert.

- Seien $\lambda_1, \ldots, \lambda_k$ bestimmt. Aufgabe:

$$F(x) \to \min_{\|x\|=1} \text{ und } \forall j = 1, \ldots, k : x^T \cdot x^j = 0$$

  Karush-Kun-Tucker-Bedingung: Es existieren $\mu_1, \ldots, \mu_k, \lambda_{k+1}$ mit

$$\begin{aligned} A \cdot x^{k+1} - \lambda_{k+1} \cdot x^{k+1} + \sum_{j=1}^{k} \mu_j \cdot x^j &= 0 \\ \Rightarrow x \in \operatorname{span}\{x^1, \ldots, x^k\} : x^T \cdot \left(A \cdot x^{k+1} - \lambda_{k+1} \cdot x^{k+1} + \sum_{j=1}^{k} \mu_j \cdot x^j \right) &= 0 \end{aligned}$$

  Außerdem:

$$\begin{aligned} 0 &= (x^j)^T \cdot \left(A \cdot x^{k+1} - \lambda_{k+1} \cdot x^{k+1} + \sum_{j=1}^{k} \mu_j \cdot x^j \right) \\ &= \underbrace{\lambda_l \cdot (x^l)^T \cdot x^{k+1}}_{0} - \underbrace{\lambda_{k+1} \cdot (x^l)^T \cdot x^{k+1}}_{0} + \mu_l \cdot \|x^l\|^2 \\ \Rightarrow \mu_l &= 0 \end{aligned}$$

  mit $l \in \{1, \ldots, k\}$. Also λ_{k+1} Eigenwert von A.

Lemma Es seien A, B symmetrische Matrizen und $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ und $\mu_1 \leq \ldots \leq \mu_n$ die (geordneten) Eigenwerte der Matrizen A bzw. A+B. Dann gilt:

$$|\lambda_j - \mu_j| \leq \|B\|_2$$

Bemerkung:

1. Im unsymmetrischen Fall ist diese Aussage nicht gültig.
2. Zum Beweis zu Linkseigenvektoren: Sei $A \in \mathbb{C}^{n\times n}$, $A^* = \bar{A}^T$ adjungierte Matrix. Es gilt: λ_i Eigenwert von A (x^i zugehöriger Eigenvektor) $\Leftrightarrow$ $\bar{\lambda}_i$ Eigenwert von A^* (y^i zugehöriger Eigenvektor). Dann gilt:
$$x^i \perp y^i \qquad (i \neq j)$$
Beweis:
   - Es gilt:
$$\begin{aligned} (y^j, A\cdot x^i) &= (A^*\cdot y^j, x^i) \\ \bar{\lambda}_i \cdot (y^j, x^i) &= \bar{\lambda}_j \cdot (y^j, x^i) \\ 0 &= (\bar{\lambda}_j - \bar{\lambda}_i)\cdot (y^j, x^i) \\ \Rightarrow 0 &= (y^j, x^i) \end{aligned}$$
   Falls $A \in \mathbb{R}^{n\times n}$: y^j Linkseigenvektor $\Leftrightarrow$ y^j ist Eigenvektor zu $A^T = A^*$. Damit:
$$y^j \perp x^i \qquad (Re\lambda_j \neq Re\lambda_i)$$
   falls $\lambda_{i_1} = \bar{\lambda}_{i_2}$:
$$y^{i_1} \perp x^{i_1} \qquad y^{i_2} \perp x^{i_2}$$

5.2 von Misesche Vektoriteration (Potenzmethode)

- A habe ein vollständiges Eigenvektor-System (A ist diagonalisierbar durch Ähnlichkeitstransformation) und besitze einen dominanten Eigenwert, d.h. $|\lambda_n| > |\lambda_{n-1}| \geq \ldots \geq |\lambda_1|$. Dann Potenzmethode (Vektoriteration):
  - Wahl eines Startvektors
$$y^0 = \sum_{i=1}^{n} c_i \cdot v^i \qquad (c_n \neq 0)$$
  - Iteration:
$$\begin{aligned} \tilde{y}^{k+1} &= A\cdot y^k \\ y^{k+1} &= \alpha_k \cdot \tilde{y}^{k+1} \qquad (1) \end{aligned}$$
   mit
$$\alpha_k = \frac{1}{\|A\cdot y^k\|} \qquad \text{oder} \qquad \alpha_k = \frac{1}{|[A\cdot y^k]_l|}$$

Satz Für die durch (1) erzeugte Iteration gilt:

$$\lim_{k\to\infty} \frac{[A\cdot y^k]_l}{y_l^k} = \lambda_n$$

für alle l mit $v_l^n \neq 0$. Und:

$$\lim_{k\to\infty} \frac{(y^k)^T \cdot A \cdot y^k}{(y^k)^T \cdot y^k} = \lambda_n$$

Beweis:

- Sei $\{v^j\}_{j=1}^n$ das (orthonormale) Eigenvektorsystem.

$$\begin{aligned}
y^0 &= \sum_{j=1}^{n} c_j \cdot v^j \\
\Rightarrow A \cdot y^0 &= \sum_{j=1}^{n} c_j \cdot \lambda_j \cdot v^j \\
\Rightarrow y^k &= \prod_{l=0}^{k-1} \alpha_l \cdot \left(\sum_{j=1}^{n} c_j \cdot \lambda_j^k \cdot v^j \right) \\
&= \lambda_n^k \cdot \prod_{l=0}^{k-1} \alpha_l \cdot \left(c_n \cdot v^n + \sum_{j=1}^{n-1} \underbrace{\left(\frac{\lambda_j}{\lambda_n} \right)^k}_{|.|<1} \cdot c_j \cdot v^j \right) \\
\Rightarrow A \cdot y^k &= \lambda_n^{k+1} \cdot \prod_{l=0}^{k-1} \alpha_l \cdot \left(c_n \cdot v^n + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k+1} \cdot c_j \cdot v^j \right)
\end{aligned}$$

Damit:

$$\begin{aligned}
\frac{[A \cdot y^k]_l}{y_l^k} &= \lambda_n \cdot \frac{c_n \cdot v_l^n + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k+1} \cdot c_j \cdot v_l^j}{c_n \cdot v_l^n + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k} \cdot c_j \cdot v_l^j} \\
&= \lambda_n \cdot \frac{1 + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k+1} \cdot \frac{c_j \cdot v_l^j}{c_n \cdot v_l^n}}{1 + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k} \cdot \frac{c_j \cdot v_l^j}{c_n \cdot v_l^n}} \\
&\to \lambda_n \ (k \to \infty)
\end{aligned}$$

Weiter gilt:

$$\begin{aligned}
\frac{A \cdot y^k}{\|y_k\|_2} &= \frac{\lambda_n^{k+1} \cdot \left(c_n \cdot v^n + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k+1} \cdot c_j \cdot v^j \right)}{|\lambda_n|^k \cdot \left\| c_n \cdot v^n + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{k} \cdot c_j \cdot v^j \right\|_2} \\
&\to |\lambda_n| \cdot \frac{c_n \cdot v^n}{\|c_n \cdot v^n\|_2} = \lambda_n \cdot v^n
\end{aligned}$$

Für den Rayleigh-Quotienten gilt:

$$\begin{aligned}
(y^k, A \cdot y^k) &= \lambda_n^{2k+1} \cdot \left(\prod_{l=0}^{k-1} \alpha_l \right)^2 \cdot \left(c_n^2 + \sum_{j=1}^{n-1} \cdot \left(\frac{\lambda_j}{\lambda_n} \right)^{2k+1} \cdot c_j^2 \right) \\
(y^k, y^k) &= \lambda_n^{2k} \cdot \left(\prod_{l=0}^{k-1} \alpha_l \right)^2 \cdot \left(c_n^2 + \sum_{j=1}^{n-1} \cdot \left(\frac{\lambda_j}{\lambda_n} \right)^{2k} \cdot c_j^2 \right) \\
\Rightarrow \frac{(y^k, A \cdot y^k)}{(y^k, y^k)} &= \lambda_n \cdot \frac{1 + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{2k+1} \cdot \left(\frac{c_j}{c_n} \right)^2}{1 + \sum_{j=1}^{n-1} \left(\frac{\lambda_j}{\lambda_n} \right)^{2k} \cdot \left(\frac{c_j}{c_n} \right)^2} \\
&\to \lambda_n
\end{aligned}$$

Die Konvergenz beim Rayleigh-Quotienten ist beschleunigt: Mit $q := |\frac{\lambda_j}{\lambda_n}|$ für $c_j \neq 0$

$$\frac{1 + q^{k+1}}{1 + q^k} \to 1 \qquad (k \to \infty)$$

vs.

$$\frac{1+q^{2k+1}}{1+q^{2k}} \to 1 \qquad (k \to \infty)$$

Es gilt:

$$\begin{aligned} 1 - \frac{1+q^{k+1}}{1+q^k} &= q^k \cdot (1-q) \\ 1 - \frac{1+q^{2k+1}}{1+q^{2k}} &= q^{2k} \cdot (1-q) \end{aligned}$$

beide lineare Konvergenz, aber Faktor q^2 statt q.

Bemerkungen:

1. Zunächst mittels Rayleigh-Quotienten λ_n bestimmen und (2) für die Rekonstruktion des zugehörigen Eigenvektors v^n benutzen. Falls kein vollständiges Orthonormalsystem existiert, ist die Geschwindigkeitssteigerung nicht in gleicher Weise zu erwarten.

2. Es existieren Modifikationen für

   (a) mehrfache Eigenwerte λ_n

   (b) für $\lambda_n = -\lambda_i$ $(i \neq n)$

   (c) für $\lambda_n = \bar{\lambda}_i$ $(i \neq n)$

5.2.1 Inverse Vektoriteration

- Sei A regulär ($\lambda_i \neq 0$ für alle i). Anwenden der Vektoriteration auf A^{-1},

$$\begin{aligned} A \cdot x = \lambda \cdot x \quad &\Leftrightarrow \quad x = \lambda \cdot A^{-1} \cdot x \\ &\Leftrightarrow \quad \lambda^{-1} \cdot x = A^{-1} \cdot x \end{aligned}$$

  um betragsmäßig kleinsten Eigenwert von A zu ermitteln.

- Rekursionsvorschrift:

$$\begin{aligned} A \cdot \tilde{y}^{k+1} &= y^k \\ y^{k+1} &= \alpha_k \cdot \tilde{y}^{k+1} \end{aligned}$$

  hierbei Konvergenzbeschleunigung durch Spektralverschiebung möglich. Für $0 < |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_n|$:

$$\beta_k \approx \lambda_1 \Rightarrow (A - \beta_k \cdot I) \cdot \tilde{y}^{k+1} = y^k$$

  Damit $|\frac{\tilde{\lambda}_1}{\tilde{\lambda}_j}| < |\frac{\lambda_1}{\lambda_j}|$, also schnellere Konvergenz. Sei λ_1^k eine Näherung für den Eigenwert. Dann

$$\begin{aligned} \lambda_1^k &\approx \lambda_1 - \beta_k \\ \Rightarrow \beta_{k+1} &:= \beta_k + \lambda_1^k \end{aligned}$$

  für Update von β_k.

Bemerkungen:

1. Bei lokal guter Startnäherung β_k ist (im Prinzip) jeder (einfache) Eigenvektor durch eine inverse Approximation rekonstruierbar (Verfahren von Wielandt), aber aufwendigr als Potenzmethode.

2. Asymptotische Singularität von $(A - \beta_k \cdot I)$ ist numerisch unproblematisch, da $y^k \in (A - \beta_k \cdot I)$ ebenfalls „asymptotisch“.

3. Möglichkeit zur Gewinnung weiterer Eigenwerte (für $A = A^T$) durch Iteration in orthogonalen Unterraum:

$$\tilde{y}^{k+1} = A \cdot y^k$$

mit $(y^0, x^1) = 0$, dann Reorthogonalisierung von $\tilde{y}^{k+1}$ mittels Gram-Schmidt-Verfahren

$$y^{k+1} = \alpha_k \cdot (\tilde{y}^{k+1} + \sigma_k \cdot x^1)$$

dabei Wahl von σ_k so, dass $(y^{k+1}, x^1) = 0$.

$$\begin{aligned} \Leftrightarrow < \tilde{y}^{k+1}, x^1 > + < x^1, x^1 > \cdot \sigma_k &= 0 \\ \sigma_k &= -\frac{< \tilde{y}^{k+1}, x^1 >}{< x^1, x^1 >} \end{aligned}$$

5.3 Symmetrische Eigenwertprobleme

- Sei $A = A^T$. In allen Verfahren: Systematische Anwendung von (sukzessiven) Ähnlichkeitstransformationen mit orthogonalen Matrizen Q_k:

$$A_{k+1} = Q_k^T \cdot A \cdot Q_k$$

mit $A_0 := A$.

- Für $A = A^T$: Alle Eigenwerte reell, A diagonalisierbar.

5.3.1 Jacobi-Verfahren

- Wahl in 1) derart, dass

$$\lim_{k \to \infty} \|A_k - D_k\| = 0$$

mit $D_k = \text{diag}\,(a_{ii}^k)$, also $a_{ij}^k \to 0$ für $i \neq j$.

- Bezeichnung:

$$S(A) := \sum_{i,j=1, i \neq j}^{n} a_{ij}^2 =: \|A - D(A)\|_F^2$$

mit $D(A) = \text{diag}\,(a_{ii})$.

- **Satz** Seien $\lambda_1 \geq \ldots \geq \lambda_n$ die Eigenwerte von A (der Größe nach geordnet) und die Hauptdiagonalelemente $a_{j_1,j_1} \geq \ldots \geq a_{j_n,j_n}$ seien ebenfalls der Größe nach geordnet. Dann gilt:

$$|a_{j_r,j_r} - \lambda_r| \leq \sqrt{S(A)}$$

für alle $r \in \{1, \ldots, n\}$.

Beweis:

– Störlemma für Eigenwerte bei symmetrischen Matrizen

$$|\mu_j - \lambda_j| \leq \|B\|_2$$

mit λ_j Eigenwerte von A, μ_j Eigenwerte von A+B. Anwendung mit $B := D(A) - A$ (damit $A + B = D(A)$):

$$\begin{aligned} |a_{j_r,j_r} - \lambda_r| &\leq \|A - D(A)\|_2 \\ &\leq \|A - D(A)\|_F = \sqrt{S(A)} \end{aligned}$$

für $r \in \{1, \ldots, n\}$.

- Bemerkung: Es gilt auch

$$\sum_{r=1}^{n}(a_{j_r,j_r}-\lambda_r)^2 \le \|A-D(A)\|_F^2$$

(Folgerung aus Satz von Wielandt/Hoffmann)

- Die Realisierung der Q_k im Jacobi-Verfahren erfolgt mit Givens-Rotationsmatrizen:

$$Q_k = \begin{pmatrix} E_{l-1,m-1} & & & & \\ & s_k & \dots & t_k & \\ & & 1 & & \\ & & \ddots & & \\ & & & 1 & \\ & -t_k & \dots & s_k & \\ & & & & E_{l-1,m-1} \end{pmatrix}$$

mit $s_k^2 + t_k^2 = 1$.

$$\begin{aligned} q_{ll}^k &= q_{mm}^k = s_k \\ q_{lm}^k &= -q_{ml}^k = t_k \end{aligned}$$

Es gilt:

$$Q_k^T \cdot Q_k = I$$

- Die Wirkung von Q_k im Algorithmus:

$$A_{k+1} = A_k + \begin{pmatrix} 0 & x & 0 & x & 0 \\ x & x & x & x & x \\ 0 & x & 0 & x & 0 \\ x & x & x & x & x \\ 0 & x & 0 & x & 0 \end{pmatrix} \qquad (0 : \text{Matrix})$$

mit

$$\begin{aligned} a_{ll}^{k+1} &= s_k^2 \cdot a_{ll}^k + 2s_k \cdot t_l \cdot a_{lm}^k + t_k^2 \cdot a_{mm}^k \\ a_{mm}^{k+1} &= t_k^2 \cdot a_{ll}^k + 2s_k \cdot t_k \cdot a_{lm}^k + s_k^2 \cdot a_{mm}^k \\ a_{lm}^{k+1} &= a_{ml}^{k+1} = t_k \cdot s_k \cdot (a_{mm}^k - a_{ll}^k) + (s_k^2 - t_k^2) \cdot a_{lm}^k \\ a_{jl}^{k+1} &= a_{lj}^{k+1} = s_k \cdot a_{jl}^k + t_k \cdot a_{jm}^k \\ a_{jm}^{k+1} &= a_{mj}^{k+1} = -t_k \cdot a_{jl}^k + s_k \cdot a_{jm}^k \end{aligned}$$

für $j \notin \{l, m\}$. Ansonsten $a_{ij}^{k+1} = a_{ij}^k$.

- **Lemma**

1. Es gilt:

$$\|Q^T \cdot A \cdot Q\|_F = \|A\|_F$$

für alle $A \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{n\times n}$ orthogonal.

2. Es gilt:

$$S(A_{k+1}) = S(A_k) - 2\left((a_{lm}^k)^2 - (a_{lm}^{k+1})^2\right)$$

Ohne Beweis, z.B. siehe Plato

- **Folgerung** Wahl von s_k, t_k derart, dass

$$a_{lm}^{k+1} = a_{ml}^{k+1} = 0$$

(Dann maximale Verringerung von S(A).)

- **Satz**

$$s_k = \sqrt{\frac{1+c}{2}} \qquad t_k = \mathrm{sgn}(a_{lm}^k) \cdot \sqrt{\frac{1-c}{2}}$$

mit

$$c = \frac{a_{ll}^k - a_{mm}^k}{(a_{ll}^k - a_{mm}^k)^2 + 4(a_{lm}^k)^2}$$

(A_k symmetrisch, also $a_{ij}^k = a_{ji}^k$)

- **Satz** Beim klassischen Jacobi-Verfahren

$$\begin{aligned} A_0 &:= A \\ A_{k+1} &= Q_k^T \cdot A_k \cdot Q_k \end{aligned}$$

mit Wahl von l,m aus

$$|a_{lm}^k| = \max_{i,j(i\neq j)} |a_{ij}^k|$$

gilt

$$\lim_{k\to\infty} S(A_k) = 0$$

Beweis:

– Wegen $|a_{ij}^k| \leq |a_{lm}^k|$ für alle $i \neq j$ gilt:

$$S(A_k) = \sum_{i\neq j} (a_{ij}^k)^2 \leq N \cdot (N-1) \cdot (a_{lm}^k)^2$$

Damit

$$\begin{aligned} S(A_{k+1}) &= S(A_k) - 2(a_{lm}^k)^2 \\ &\leq \underbrace{\left(1 - \frac{2}{N\cdot(N-1)}\right)}_{=:q<1} S(A_k) \end{aligned}$$

Bemerkungen:

1. Aufwandbetrachtung: Bei Vorgabe einer Genauigkeit $\eta > 0$ sind m Schritte notwendig mit

$$\begin{aligned} m &\geq -\frac{\ln\left(\frac{\sqrt{S(A)}}{\eta}\right)}{\ln\left(-\frac{2}{n\cdot(n-1)}\right)} \\ &\approx N^2 \cdot \ln\left(\frac{\sqrt{S(A)}}{\eta}\right) \end{aligned}$$

Für vorgegebene Genauigkeit $\eta > 0$: $c \cdot n^2$ Schritte notwendig.

2. Aufwand in jedem Schritt: 4n Multiplikationen, 2n Additionen und O(1) Wurzeln und Divisionen. Aufwand also zunächst O(n), jedoch $\frac{n\cdot(n-1)}{2}$ Vergleichsoperationen zur Bestimmung von l,m . Insgesamt:

$$c \cdot 3 \cdot (n^4 + O(n^3))$$

als Gesamtaufwand für das Verfahren.

3. Beim Abbruch nach k_0 Schritten,

$$A_{k_0+1} \approx \mathrm{diag}\,\{\lambda_1, \ldots, \lambda_n\}$$

Das Produkt $Q_1 \cdots Q_k$ bildet eine Näherung für die Eigenvektoren zu $\lambda_1, \ldots, \lambda_n$. Bekannt: Für $A = A^T$ existiert Q orthogonal mit $D = Q^T \cdot A \cdot Q$ mit $Q = (v^1 \ldots v^n)$.

$$A_{k_0+1} = Q_{k_0}^T \cdots Q_1^T \cdot A \cdot \underbrace{Q_1 \cdots Q_{k_0}}_{\approx Q}$$

5.3.2 „Zyklisches“ Jacobi-Verfahren

- Definition: „Multischritt“
$$A_{m+1} = S_m^T \cdot A \cdot S_m$$
mit
$$S_m = \prod_{p=1}^{n-1} \prod_{q=p+1}^{n} Q_{pg}^m$$
Einträge s_{pg}^m, t_{pg}^m werden aus A_m berechnet.
- Aufwand für einen Multischritt: O(n^3). In der Regel nach $m = O(1)$ Schritten $S(A_m)$ „hinreichend“ klein.
- **Satz** Falls alle Eigenwerte von $A = A^T \in \mathbb{R}^{n\times n}$ einfach sind, so gilt
$$S(A_{m+1}) \leq \frac{S(A_m)^2}{\delta}$$
mit
$$\delta = \min_{\lambda,\mu\in\sigma(A)} |\lambda - \mu|$$

5.3.3 QR-Algorithmus

- Bekannt: Anwendung von Householder-Transformationen zur Berechnung der QR-Zerlegung einer Matrix A, $A = Q \cdot R$, mit $Q = H_1 \cdots H_{n-1}$, R obere Dreiecksmatrix.
$$\begin{aligned} H_i &= I - 2u \cdot u^T \\ H_1 \cdot A &= \begin{pmatrix} x & \\ 0 & \\ \vdots & \tilde{A} \\ 0 & \end{pmatrix} \\ A \cdot H_1 &= \begin{pmatrix} x & 0 & \dots & 0 \\ & & \tilde{A} & \end{pmatrix} \end{aligned}$$
Zur Erzeugung einer Tridiagonal-Matrix ist eine Modifikation notwendig, damit die Struktur der Spalten nicht wieder zerstört wird.
$$(I - 2u \cdot u^T) \cdot A \cdot (I - 2u \cdot u^T) = A - 2u \cdot (Au)^T - 2(Au)^T \cdot u + 4u \cdot u^T \cdot (u^T \cdot A \cdot u)$$
Deshalb im Unterschied zur QR-Faktorisierung:
  - Wahl von $u = \frac{w}{\|w\|}$
  - 1. Schritt:
$$w = \begin{pmatrix} 0 \\ \tilde{w} \end{pmatrix} \qquad \text{mit } \tilde{w} = \begin{cases} \frac{1}{2} \cdot (\tilde{a} + \|\tilde{a}\|_2 \cdot e^1) & \tilde{a}_1 > 0 \\ \frac{1}{2} \cdot (\tilde{a} - \|\tilde{a}\|_2 \cdot e^1) & \tilde{a}_1 < 0 \end{cases}$$
dabei
$$a = \begin{pmatrix} a_{11} \\ \tilde{a} \end{pmatrix} \qquad \tilde{a} = \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$$
Dann
$$H_1 \cdot A \cdot H_1 = \begin{pmatrix} x & x & 0 & \dots & 0 \\ x & x & x & \dots & x \\ 0 & x & & & \\ \vdots & \vdots & & \tilde{A} & \\ 0 & x & & & \end{pmatrix}$$

- **Lemma** Eine symmetrische Matrix kann stets durch (n-2) Householder-Transformationen auf Tridiagonalgestalt gebracht werden. Aufwand:

$$\frac{8}{3}n^3 \cdot \left(1 + o\left(\frac{1}{n}\right)\right)$$

- **Lemma** Zu gegebenen Zahlen $a_1, \ldots, a_n$ und $b_2, \ldots, b_n$ gelten für die charakteristischen Polynome $p_k(\mu) := \det(J_k - \mu \cdot I)$ die folgenden Rekursionsformeln für alle $\mu \in \mathbb{R}$:

$$\begin{aligned} p_0(\mu) &:= 1 \\ p_1(\mu) &= a_1 - \mu \\ p_k(\mu) &= (a_k - \mu) \cdot p_{k-1}(\mu) - b_k^2 \cdot p_{k-2}(\mu) \end{aligned}$$

  Daraus ergeben sich die Rekursionsformeln für die Ableitung:

$$\begin{aligned} p_1' &= -1 \\ p_k'(\mu) &= -p_{k-1}(\mu) + (a_k - \mu) \cdot p_{k-1}'(\mu) - b_k^2 \cdot p_{k-2}'(\mu) \end{aligned}$$

  Damit: Newton-Verfahren zur Eigenwertberechnung möglich. (Eigenwerte sind Nullstellen von $p_k(\mu)$)

- Modifikation dieses QR-Algorithmus für allgemeine Matrizen: Falls $A \neq A^T$ erzeugt Algorithmus Hessenberg-Matrix. Für Hessenberg-Matrizen analoge Rekurstionsformeln wie in Lemma. Aufwand:

$$\frac{10}{3}n^3 \cdot \left(1 + o\left(\frac{1}{n}\right)\right)$$

5.3.4 QR-Algorithmus für symmetrische Matrizen

- Sei $A = A^T$ mit Eigenwerten $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$.
- Algorithmus:

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

- Bemerkung: In jedem Schritt vollständige QR-Zerlegung notwendig.
- **Satz** Unter den obigen Voraussetzungen gilt: Der Algorithmus erzeugt A_k, für die gilt:

$$\forall j > i : a_{ij}^{(k)} = o\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right) \to 0\,(k \to \infty)$$

  d.h. speziell

$$\lim_{k\to\infty} Q_k = I \qquad \lim_{k\to\infty} R_k = D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$$

- Bemerkungen:
  1. Für Tridiagonalmatrizen gilt kubische KOnvergenz (QR-Zerlegung effektiv). ⇒ Householdertransformationen und QR-Algorithmus kombinieren.

6

Numerische Verfahren für Anfangswertaufgaben

6.1 Aufgabenstellung

- gegeben: $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$, $x_0 \in \mathbb{R}$, $y^0 \in \mathbb{R}^m$
- gesucht: Lösung von

$$\begin{aligned} y' &= f(x,y) \qquad y(x_0) = y^0 \\ \Leftrightarrow y(x) &= y^0 + \int_{x_0}^{x} f(s, y(s))\, ds \end{aligned}$$

- Existenz und Eundeutigkeit von Lösungen: Satz von Picard-Lindelöf. Es sei f auf $\mathbb{R} \times \mathbb{R}^m$ stetig und genüge bzgl. des zweiten Arguments einer Lipschitz-Bedingung, d.h. es existiert $L > 0$, sodass

$$\|f(x,u) - f(x,v)\| \leq L \cdot \|u - v\|$$

für alle $u, v \in \mathbb{R}^m$. Dann besitzt das Anfangswertproblem eine eindeutige Lösung für alle $x > x_0$ (globale Existenz). Diese hängt stetig von den Anfangswerten ab: Sei $y(x)$ Lösung von $y' = f(x,y)$, $y(x_0) = y^0$, sei $\tilde{y}$ eine Lösung von $\tilde{y} = f(x,\tilde{y})$, $\tilde{y}(x_0) = \tilde{y}^0$. Dann gilt:

$$\|y(x) - \tilde{y}(x)\| \leq e^{2Ll\cdot(x-x_0)} \cdot \|y^0 - \tilde{y}^0\|$$

Beweisskizze:

1. Wegen Integralidentität und Lipschitz-Eigenschaft: Es existiert $\delta = \delta(y^0) > 0$, sodass auf $[x_0, x_0 + \delta]$ die Abbildung

$$\Phi(\varphi) = y^0 + \int_{x_0}^{x} f(s, \varphi(s))\, ds$$

kontraktiv (in $C[x_0, x_0 + \delta]$) ist, also existiert ein Fixpunkt.

2. Das dynamische System besitzt die Flußeigenschaft (bereits bei f stetig). Sei v_1 eine Lösung des Anfangwertproblems,

$$v_1(x) = y^0 + \int_{x_0}^{x} f(s, v_1(s))\, ds \qquad x \in [x_0, x_0 + \delta]$$

v_2 Lösung von

$$v_2(x) = v_1(x_0 + \delta) + \int_{x_0+\delta}^{x} f(s, v_2(s))\, ds \qquad x \in [x_0 + \delta, x_0 + 2\delta + \delta_1]$$

Dann ist

$$v(x) = \begin{cases} v_1(x) & x \in [x_0, x_0 + \delta] \\ v_2(x) & x \in [x_0 + \delta, x_0 + \delta + \delta_1] \end{cases}$$

Lösung auf dem gesamten Intervall.

3. Mit globaler Lipschitz-Eigenschaft gilt: Es existiert $\delta > 0$, sodass $\delta(y^0) \geq \delta_0$ für alle $y^0 \in \mathbb{R}^m$, also globale Existenz.
4. Nachweis von 2. Aussage mit Gronwell-Lemma

- Beispiele:

1. Anfangswertproblem:
$$y' - 2\sqrt{|y|} = 0 \qquad y(0) = 0$$
Mögliche Lösungen:
$$y_1(x) = 0 \qquad y_2(x) = \operatorname{sgn}(x) \cdot x^2$$
also keine eindeutige Lösung.
2. Anfangswertproblem:
$$y' = \alpha \cdot y \qquad y(0) = y^0$$
Lösung:
$$y(x) = y^0 \cdot e^{\alpha x}$$
(globale Existenz und Eindeutigkeit)
3. Anfangswertproblem
$$y' = \pm y^2 \qquad y(0) = y^0 > 0$$
Für + mit Trennung der Variablen:
$$\begin{aligned} \int \frac{dy}{y^2} &= \int dx \\ -\frac{1}{y} &= x + c \\ \Rightarrow y &= -\frac{1}{x+c} \qquad c = -\frac{1}{y^0} \\ \Rightarrow y(x) &= \frac{1}{\frac{1}{y^0} - x} \end{aligned}$$
für alle $x < \frac{1}{y^0}$. Also keine globale Existenz, da nur lokale Lipschitzbedingung. Für - :
$$y(x) = \frac{1}{x + \frac{1}{y^0}}$$
also globale Existenz.

- Bemerkungen:

1. Anfangswertproblem
$$y' = f_1(x) \cdot f_2(y)$$
Trennung der Variablen:
$$\int \frac{dy}{f_2(y)} = \int \frac{dx}{f_1(x)}$$
Im Allgemeinen nicht explizit ausführbar
2. Anfangswertproblem
$$y'' + a \cdot y' + b \cdot y = f(x)$$
Umwandlung in System von Differentialgleichungen 1. Ordnung (siehe Analysis):
$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} y_2 \\ f - a \cdot y_2 - b \cdot y_1 \end{pmatrix}$$
Für Numerik: Nur Differentialgleichungssysteme 1. Ordnung relevant.

- Im weiteren vorausgesetzt: eindeutige Lösung des Anfangwertproblems existiert. (Es gelte die globale Lipschitz-Bedingung für f) Numerische Lösung: Approximation der stetigen Lösung y für $x \in [a, b]$ auf einem Gitter $\Delta = \{a = x_0 < \ldots < x_n = b\}$, $h_i = x_{i+1} - x_i$ (äquidistant gewählt: $h = \frac{b-a}{n}$) gegeben bzw. konstruiert. Dabei ist (im wesentlichen) die Integralgleichung Ausgangspunkt. Für alle $i = 0(1)(n-1)$:

$$y(x) = y(x_i) + \int_{x_i}^{x} f(s, y(s))\, ds$$

  Ersetzen des Integrals durch numerische Approximation. Mittelwertsatz:

$$y(x_{i+1}) = y(x_i) + h \cdot f(\xi_i, y(\xi_i))$$

  mit $\xi_i \in (x_i, x_{i+1})$.

  1. explizites Euler-Verfahren:
$$y_{i+1} = y_i + h \cdot f(x_i, y_i)$$
  2. implizites Euler-Verfahren:
$$y_{i+1} = y_i + h \cdot f(x_{i+1}, y_{i+1})$$
     bzw.
$$y_{i+1} = y_i + \frac{h}{2} \cdot \big(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\big)$$

  Andere Interpretation mittels Differenzenquotienten:

$$y'(x_i) \approx \frac{y(x_{i+1} - y(x_i)}{h}$$

- Konvergenzuntersuchung auf festem Intervall [a,b]. Dazu seien
  - y_i: Wert der Näherungslösung in x_i
  - $y_h := (y_0, \ldots, y_n) \in \mathbb{R}^{m \times (n+1)}$ Vektor der Näherungslösungen
  - $r_h(y) = (y(x_0), \ldots, y(x_n)) \in \mathbb{R}^{m \times (n+1)}$ Einschränkung der (unbekannten) Lösung y auf das Gitter Δ

  Frage: Was bedeutet hier Konvergenz? Zwei mögliche Wege:

  1. $y_h \to \tilde{y}_h(x)$, z.B. durch Interpolation
  2. Geeigneter: diskrete Konvergenz
$$\|y_h - r_h(y)\|_\infty := \max_{0 \le i \le n} \|y_i - y(x_i)\|_{\infty,2,1} \to 0\,(h \to 0)$$
     Konvergenzordnung p:
$$\|y_h - r_h(y)\|_\infty \le c \cdot h^p$$

- Bemerkungen:

  1. Bestimmung der diskreten Lösung erfolgt im Verfahren b) und c) über die (sukzessive) Lösung von n nichtlinearen Gleichungssystemen der Dimmension m, also implizites Verfahren. In a) „triviales“ Gleichungssystem.

Konvergenzuntersuchung des expliziten Euler-Verfahrens:

- Zur Vereinfachung: Es gelte

$$\|f(x, u) - f(z, v)\| \le \delta \cdot (\|u - v\| + |x - z|)$$

- Es gilt:

$$\begin{aligned} y(x_{k+1}) &= y(x_k) + h \cdot f(\xi_k, y(\xi_k)) \qquad \xi_k \in (x_k, x_{k+1}) \\ y_{k+1} &= y_k + h \cdot f(x_k, y_k) \\ \Rightarrow y(x_{k+1}) - y_{k+1} &\stackrel{?}{=} y(x_k) - y_k + h \cdot \big(f(\xi_k, y(\xi_k)) - f(x_k, y_k)\big) \end{aligned}$$

Sei $\varepsilon_k := \|y_k - y(x_k)\|$. Damit:

$$\begin{aligned} \varepsilon_{k+1} &\leq \varepsilon_k + h \cdot \delta \cdot (\|y(\xi_k) - y_k\| + |\xi_k - x_k|) \\ &\leq \varepsilon_k + h \cdot \delta \cdot (|\xi_k - x_k| + \|y(\xi_k) - y(x_k)\| + \|y(x_k) - y_k\|) \\ &\leq \varepsilon_k + h \cdot \delta \cdot (h + \varepsilon_k + \underbrace{y'(\eta_k)}_{\leq \varrho} \cdot \underbrace{(\xi_k - x_k)}_{\leq h}) \\ &\leq \varepsilon_k + h \cdot \delta \cdot (h + \varepsilon_k + h \cdot \varrho) \end{aligned}$$

Also gilt:

$$\varepsilon_{k+1} \leq \alpha \cdot \varepsilon_k + \beta$$

mit $\varepsilon_0 = 0$,

$$\alpha = (1 + \delta \cdot h) \qquad \beta = h^2 \cdot (\delta + \varrho \cdot \delta) =: c \cdot h^2$$

Es folgt:

$$\begin{aligned} \varepsilon_k &\leq \sum_{j=0}^{k-1} \alpha^j \cdot \beta = \frac{\alpha^k - 1}{\alpha - 1} \cdot \beta \\ &\leq \frac{e^{\delta h k} - 1}{\delta \cdot h} \cdot \delta \cdot (\varrho + 1) \cdot h^2 \\ &\leq \left(e^{\delta \cdot (x_k - x_0)} - 1\right) \cdot (\varrho + 1) \cdot h \end{aligned}$$

Folgerung:

$$\max_{0 \leq k \leq n} \|y_h - y(x_k)\| \leq \tilde{c} \cdot h$$

also Konvergenzordnung 1

- Damit konvergiert die mit dem expliziten Verfahren erzeugte Näherung für $h \to 0$ gegen die Lösung $y(x)$ mit der diskreten Konvergenzordnung 1.

6.2 Explizite Einschrittverfahren

- Verallgemeinerung des expliziten Euler-Verfahrens in der Form

$$y_{i+1} = y_i + h \cdot \psi(x_i, y_i, h) \qquad (1)$$

Dabei sei $\{x_i\}_{i=0}^N$ ein vorgegebenes äquidistantes Gitter für das Intervall [a,b],

$$x_i = a + i \cdot h \qquad h = \frac{b - a}{N}$$

- Formale Interpretation als Gleichungssystem: Für $u_h := (u_0, \ldots, u_N) \in \mathbb{R}^{m \times (N+1)}$ und $F_h(u_h)$ definiert durch

$$\begin{aligned} F_{i+1}(u_h) &:= \frac{u_{i+1} - u_i}{h} - \psi(x_i, u_i, h) \\ F_0(u_h) &:= u_0 - y_0 \end{aligned}$$

Damit ist (1) äquivalent zu $F_h(u_h) = 0$. Wir setzen

$$\|F_h(u_h)\| := \max_{0 \leq i \leq N} \|F_i(u_h)\|$$

und betrachten $F_h(r_h(y))$, d.h.

$$\begin{aligned} F_{i+1}(r_h(y)) &= \frac{y(x_i+1)-y(x_i)}{h} - \psi(x_i, y(x_i), h) \\ F_0(r_h(y)) &= y(x_0) - y_0 = 0 \end{aligned}$$

- Definition: Ein explizites Einschrittverfahren heißt konsistent, falls

$$\lim_{h \to 0+} \|F_h(r_h(y))\| = 0$$

bzw. konsistent mit Ordnung p, falls

$$\|F_h(r_h(y))\| \leq c \cdot h^p$$

mit $c > 0, p > 0$.

- Frage: Ist Konsistenz für Konvergenz hinreichend? Im Allgemeinen nicht, es ist zusätzlich Stabilität notwendig.

- Definition: Ein Einzelschrittverfahren heißt stabil, falls ein $\delta > 0$ existiert mit

$$\|u_h - v_h\| \leq \delta \cdot \|F_h(u_h) - F_h(v_h)\|$$

für alle $u_h, v_h \in \mathbb{R}^{m \times (N+1)}$.

- **Satz** Ist (1) konsistent und stabil, dann konvergiert y_h gegen $r_h(y)$ diskret, d.h.

$$\lim_{h \to 0+} \|y_h - r_h(y)\|_\infty = 0$$

Beweis:

  - Es gilt: (1) $\Leftrightarrow F_h(y_h) = 0$. Damit:

$$\begin{aligned} \|y_h - r_h(y)\| &\leq \delta \cdot \|F_h(y_h) - F_h(r_h(y))\| \\ &= \delta \cdot \|F_h(r_h(y))\| \to 0 \, (h \to 0+) \end{aligned}$$

- **Lemma** Die Verfahrensfunktion genüge einer Lipschitz-Bedingung der Form

$$\|\psi(x, u-h) - \psi(x, v, h)\| \leq L \cdot \|u - v\|$$

für alle $x \in [a, b]$, $u, v \in \mathbb{R}^m$. Dann ist das Verfahren (1) stabil.

Beweis analog zur Konvergenzuntersuchung für das Euler-Verfahren

Modifziertes Eulerverfahren

$$y_{i+1} = y_i + h \cdot f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} \cdot f(x_i, y_i)\right)$$

- Untersuchung der Konsistenzordnung mittels Taylorentwicklung:

$$\begin{aligned} F_i(y) &= \frac{y(x_{i+1}) - y(x_i)}{h} - f\left(x_i + \frac{h}{2}, y(x_i) + \frac{h}{2} \cdot f(x_i, y_i)\right) \\ &= y'(x_i) + \frac{1}{2} h \cdot y''(x_i) - \big(f(x_i, y(x_i)) + \frac{h}{2} \cdot \big(f_x(x_i, y(x_i)) + \\ &\quad + f_y(x_i, y(x_i)) \cdot f(x_i, y(x_i))\big)\big) + o(h^2) \end{aligned}$$

y ist Lösung von $y' = f(x, y)$, also $y'' = f_y + f_y \cdot f$. Damit:

$$F_i(y) = 0 + o(h^2)$$

Die für Stabilität notwendige Lipschitz-Bedingung für die Verfahrensfunktion ψ des (expliziten) Einschrittverfahrens kann aus der Lipschitz-Bedingung für die rechte Seite des Anfangwertproblems gewonnen werden, die für die Existenz (und Eindeutigkeit) der Lösung benötigt wurde.

- Also: Das modifizierte Euler-Verfahren konvergiert mit der Ordnung h^2 und ist explizit.

Runge-Kutta-Verfahren

- klassisches Runge-Kutta-Verfahren: Berechne in jedem Schritt:

$$\begin{aligned} k_1 &:= f(x_i, y_i) \\ k_2 &:= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} \cdot k_1\right) \\ k_3 &:= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} k_2\right) \\ k_4 &:= f(x_i + h, y_i + h \cdot k_3) \end{aligned}$$

  und setze

$$y_{i+1} := y_i + \frac{h}{6} \cdot (k_1 + 2k_2 + 2k_3 + k_4)$$

- Aussage: Das klassische Runge-Kutta-Verfahren ist explizit und besitzt die Konvergenzordnung 4 (bei hinreichender Glattheit von f).

- Beispiel:

  1. Anfangswertproblem:

$$y' = y \qquad y(0) = 1$$

     Ein Schritt Runge-Kutta-Verfahren.

$$\begin{aligned} k_1 &= 1 \\ k_2 &= 1 + \frac{h}{2} \\ k_3 &= 1 + \frac{h}{2} + \frac{h^2}{4} \\ k_4 &= 1 + h + \frac{h^2}{2} + \frac{h^3}{4} \\ \Rightarrow y_1 &= 1 + \frac{h}{6} \cdot \left(1 + 2 \cdot \left(1 + \frac{h}{2}\right) + 2 \cdot \left(1 + \frac{h}{2} + \frac{h^2}{4}\right) + 1 + h + \frac{h^2}{2} + \frac{h^3}{4}\right) \\ &= 1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \frac{h^4}{24} \\ &= \sum_{k=0}^{4} \frac{h^k}{k!} = e^h + o(h^5) \end{aligned}$$

- Verallgemeinerung des klassischen Runge-Kutta-Verfahrens: Allgemeines Runge-Kutta-Verfahren der Stufe $s \in \mathbb{N}, s \geq 2$. Butcher-Feld:

$$\begin{pmatrix} c & A \\ & b^t \end{pmatrix}$$

  Iterationsvorschrift:

$$y_{i+1} = y_i + h \cdot \sum_{j=1}^{s} b_j \cdot f(x_i + c_j \cdot h, k_j)$$

  mit den „Zwischenwerten“

$$k_l = y_i + h \cdot \sum_{j=1}^{s} a_{lj} \cdot f(x_i + c_j \cdot h, k_j)$$

  für $l = 1, \ldots, s$.

  1. modifiziertes Euler-Verfahren ($s = 2$)

$$\begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ & 0 & 1 \end{pmatrix}$$

2. Runge-Kutta-Verfahren, klasisch ($s = 4$)

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

- Klassifikation:
  1. explizites Runge-Kutta-Verfahren: $\forall l \geq j : a_{lj} = 0$
  2. semiimplizites Runge-Kutta-Verfahren: $\forall l > j : a_{lj} = 0$
  3. implizites Runge-Kutta-Verfahren: $\exists l_0 > j_0 : a_{l_0,j_0} \neq 0$
- Bemerkung:
  1. Existieren asymptotische Entwicklungen (genauere Charakterisierungen der Konsistenzabschätzung), so kann das Prinzip der Richardson-Extrapolation zur Verbesserung der Näherung eingesetzt werden. Speziell für das klassische Runge-Kutta-Verfahren gilt bei hinreichender Glattheit eine Darstellung der folgenden Art:

     $$y_h(x) = y(x) + (c_0(x) + c_1(x)) \cdot h^4$$

     für $x \in \{x_i\}_{i=0}^{N(h)}$. Hierbei sind c_0, c_1 stetige Funktionen mit

     $$\|c_0\| \leq c \qquad \|c_1\| \leq c_1(h) \to 0 (h \to 0)$$

     Dann folgt aus Richardson-Prinzip:

     $$\tilde{y}_h(x) = \frac{16 y_{\frac{h}{2}} - y_h(x)}{15}$$

     ist bessere Abschätzung.

Konvergenz von expliziten Einschrittverfahren

- Betrachtung auf nichtäquidistanten Gitter $\Delta = \{a = x_0 < \ldots < x_n = b\}$, $h_i = x_{i+1} - x_i = \Delta x_i$
- Vorschrift:

  $$y_{k+1} = y_k + h \cdot \psi(x_k, y_k)$$

  Bei Einsetzen der Lösung y(x) der Anfangswertaufgabe in die Vorschrift gilt mit lokalen Konsistenzfehler $r_h := r_h(h_k)$,

  $$\begin{aligned} y(x_{k+1}) &= y(x_k) + h_k \cdot \psi(x_k, y(x_k)) + r_h(h_k) \\ y(x_{k+1}) - y_{k+1} &= y(x_k) - y_k + h_k \cdot (\psi(x_k, y(x_k)) - \psi(x_k, y_k)) + r_h(h_k) \end{aligned}$$

  Mit Lipschitz-Bedingung für ψ:

  $$\|\psi(u, v) - \psi(u, x)\| \leq L \cdot \|v - x\|$$

  folgt für $\varepsilon_k := \|y(x_k) - y_k\|$:

  $$\varepsilon_{k+1} \leq \varepsilon \cdot (1 + h_k \cdot L) + r_h(h_k)$$

  Sei $h := \max\{h_i\}$ (Durchmesser der Zerlegung). Ferner gelte $\|r_h(h_k)\| \leq \alpha \cdot h_k^{p+1}$ und $\min h_k > \sigma \cdot h$ mit $\sigma > 0$ fixiert (Eigenschaft des Gitters).

  $$\varepsilon_{k+1} \leq \varepsilon_k \cdot (1 + h \cdot L) + \alpha \cdot h^{p+1}$$

Mit $\varepsilon_0 = 0$ erhält man

$$\begin{aligned}\varepsilon_k &\leq \sum_{j=0}^{k-1}(1+h\cdot L)^j \cdot \alpha \cdot h^{p+1} \\ &\leq \frac{(1+h\cdot L)^k - 1}{(1+h\cdot L)-1} \cdot \alpha \cdot h^{p+1} \\ &\leq \frac{e^{hkL}-1}{L} \cdot \alpha \cdot h^p\end{aligned}$$

Mit

$$h \cdot k \leq \frac{1}{\sigma} \cdot (\sigma \cdot h \cdot k) \leq \frac{1}{\sigma} \cdot \sum_{j=1}^{k} h_k \leq \frac{b-a}{\sigma}$$

folgt

$$\varepsilon_k \leq \frac{e^{\frac{b-a}{\sigma}}-1}{L} \cdot \alpha \cdot h^p \to 0\,(h \to 0)$$

- Im Prinzip analog: Analyse impliziter Einschrittverfahren

  $$y_{k+1} = y_k + h_k \cdot \psi(x_k, y_k, x_{k+1}, y_{k+1})$$

  (besonders bzgl. Konsistenz, Stabilität, Konvergenz)
- Im Allgemeinen besitzen implizite Verfahren bessere Stabilitätseigenschaften.
- Zusätzlich erforderlich: Analyse der Lösung der nichtlinearen Gleichungssysteme (in jedem Schritt), z.B. Fixpunktverfahren, besser Newton-Verfahren

Prinzip der linearen Mehrschrittverfahren (LMSV)

- Mittels Informationen an l Stützstellen $(x_{k-l}.y_{k-l}), \ldots, (x_{k-1}, y_{k-1}) \to (x_k, y_k)$ l-Schritteverfahren (Adams-Moulton, Adams-Bashford).
- „Extrapolation“ einer Interpolierenden auf $[x_{k-l}, x_k]$. Für $l = 2$:

$$y_k = y_{k-1} + \int_{x_{k-1}}^{x_k} p_{l-1,k}(x)\, dx$$

für $l = 2$ lineare Interpolation:

$$\begin{aligned} p_{l-1,k}(x) &= \frac{x_{k-1} - x}{h} \cdot f_{k-2} + \frac{x - x_{k-2}}{h} \cdot f_{k-1} \\ y_k &= y_{k-1} + \int_{x_{k-1}}^{x_k} \left(\frac{x_{k-1} - x}{h} f_{k-2} + \frac{x - x_{k-2}}{h} \cdot f_{k-1} \right) dx \\ &= y_{k-1} - \frac{h}{2} \cdot \left(\frac{x_{k-1} - x}{h} \right)^2 \cdot f_{k-2} \Big|_{x_{k-1}}^{x_k} + \frac{h}{2} \cdot \left(\frac{x - x_{k-2}}{h} \right)^2 \cdot f_{k-1} \Big|_{x_{k-1}}^{x_k} \\ &= y_{k-1} - \frac{h}{2} \cdot f_{k-2} + \frac{3h}{2} \cdot f_{k-1} \end{aligned}$$

Allgemeiner Verfahrensklasse der Mehrschrittverfahren

- Ein l-Schrittverfahren zur näherungsweisen Lösung der Anfangswertaufgabe $y' = f(x, y)$, $y(x_0) = y^0$ besitzt auf einem äquidistanten Gitter die Form

$$\sum_{j=0}^{l} \alpha_j \cdot y_{k+j} = h \cdot \varphi(x_k, y_k, \ldots, x_{k+l}, y_{k+l})$$

für $k = 0, \ldots, N - l$ mit den Startwerten $y_0, \ldots, y_{l-1}$

- explizites Mehrschrittverfahren: φ hängt nicht von x_{k+l}, y_{k+l} ab
- lineares explizites Mehrschrittverfahren: Die Verfahrensfunktion φ ist in der Form

$$\varphi(x, u_1, \ldots, u_l) = \sum_{j=1}^{l-1} \beta_j \cdot f(x + h \cdot j, u_j)$$

Für das gegebene Beispiel (l=2):

$$\alpha_0 = 0 \qquad \alpha_1 = -1 \qquad \alpha_2 = 1 \qquad \beta_1 = -\frac{1}{2} \qquad \beta_2 = \frac{3}{2}$$

Anfangswerte: Berechnung mittels Einschrittverfahren (z.B. Runge-Kutta-Verfahren)

- Definition: Ein Einschrittverfahren heißt nullstabil, falls für das erzeugende Polynom (vom Grad l)

$$p(\xi) = \alpha_l \cdot \xi^l + \ldots + \alpha_0 \in \mathcal{P}_l$$

die Dahlquistsche Wurzelbedingung gilt:

$$p(\xi) = 0 \Rightarrow |\xi| \leq 1$$

und aus $|\xi| = 1$ folgt, dass ξ einfache Nullstelle ist. Im obigen Beispiel:

$$\xi^2 - \xi = -\xi \cdot (1 - \xi)$$

- **Satz** Ein lineares Mehrschrittverfahren ist konvergent, falls es konsistent und nullstabil ist (und die Verfahrensfunktion einer Lipschitz-Bedingung genügt).

Prinzip der „kontinuierlichen Approximation“

$$\begin{aligned} y_N(x) &= \sum_{j=1}^{N} c_j \cdot \varphi_j(x) \\ y'_N(x) &= \sum_{j=1}^{N} c_j \cdot \varphi'_j(x) \approx f\left(x, \sum_{j=1}^{N} c_j \cdot \varphi_j(x)\right) \end{aligned}$$

- Bestimmung der Werte c_j durch Kollakation

$$y'_N(x_i) = f\left(x_i, \sum_{j=1}^{N} c_j \cdot \varphi_j(x)\right)$$

  oder least-squares, ...

- Beispiel:

  1. Anfangswertproblem:

$$y' = y \qquad y(0) = 1$$

     Ansatz:

$$y_h = 1 + c \cdot x \qquad \Rightarrow y'_h = c$$

     least-squares:

$$\begin{aligned} \int_0^h (c - (1 + c \cdot x))^2 \, dx &\rightarrow \min \\ &= \int_0^h (c \cdot (1 - x) - 1)^2 \, dx \\ \Rightarrow c &= \frac{3}{2} \cdot \frac{(1-h)^2 - 1}{(1-h)^3 - 1} \\ &= 1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^4}{18} + o(h^5) \\ y_h(x) &= 1 + c \cdot h \\ &= 1 + h + \frac{h^2}{2} + \frac{h^3}{6} + o(h^4) \end{aligned}$$

7 Diskretisierung von Randwertaufgaben

7.1 Randwertaufgaben für gewöhnliche DGL 2. Ordnung

- lineare Randwertaufgabe 2. Ordnung $[a, b] \subseteq \mathbb{R}$,

$$(Lu)(x) := a_2(x) \cdot u'' + a_1(x) \cdot u' + a_0(x) \cdot u \stackrel{!}{=} f(x)$$

  mit $f(x), a_0(x), a_1(x), a_2(x) \in C[a, b], a_2(x) \neq 0$. Randwert-Bedingungen:

$$Ru := C \cdot \begin{pmatrix} u(a) \\ u'(a) \end{pmatrix} + D \cdot \begin{pmatrix} u(b) \\ u'(b) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

  mit $C, D \in \mathbb{R}^{2\times 2}$.

- Separierte Randbedingung:

$$C = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 0 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 0 & 0 \\ \beta_1 & \beta_2 \end{pmatrix}$$

  mit $\alpha_1^2 + \alpha_2^2 > 0$, $\beta_1^2 + \beta_2^2 > 0$. Für $\eta_1 = \eta_2 = 0$: homogene Randbedingungen. Spezialfälle:

  1. $\alpha_1 = \beta_1 = 1$, $\alpha_2 = \beta_2 = 0$: Dircihlet-Bedingung
  2. $\alpha_1 = \beta_1 = 0$, $\alpha_2 = \beta_2 = 1$: Neuman-Bedingung
  3. allgemein: Robin-Randbedingung

- Eine Anfangswertaufgabe besitzt (generisch) eine eindeutige Lösung (Lipschitz-Bedingung), eine Randwertaufgabe nicht.

- Beispiele:

  1. Anfangswertproblem:

$$\begin{aligned} y'' + y &= 0 \\ y(x) &= c_1 \cdot \cos x + c_2 \cdot \sin x \end{aligned}$$

     Drei Fälle:

     (a) $y(0) = y(\frac{\pi}{2}) = 1$:

$$\begin{aligned} \Rightarrow c_1 &= 1 \qquad c_2 = 1 \\ y(x) &= \cos x + \sin x \end{aligned}$$

     (eindeutige Lösung)

     (b) $y(0) = y(\pi) = 1$:

$$\Rightarrow c_1 = 1 \qquad c_1 = -1$$

     unlösbar

(c) $y(0) = y(2\pi) = 1$:

$$c_1 = 1 = c_1$$

c_2 beliebig, unendlich viele Lösungen

- **Satz** (Fredholsche Alternative) Entweder besitzt das Randwertproblem $Lu = f$, $Ru = \eta$ genau eine klasische Lösung auf [a,b] für jedes $f \in C[a,b]$ und jedes $\eta \in \mathbb{R}^2$ oder das homogene Randwertproblem $Lu = 0$, $Ru = 0$ besitzt nichttriviale Lösungen $u \neq 0$ (analog zu Matrizen),

$$\ker(L,R) := \{u; Lu = 0, Ru = 0\}$$

ker(L,R) ist ein linearer Unterraum

- Modellaufgabe

$$-y''(x) = f(y,(x),y'(x)) \qquad y(a) = y(b) = 0$$

(Semilineares Randwertproblem mit homogenen Dirichlet-Randbedingungen) Greensche Funktion:

$$G(x,t) := \frac{1}{b-a} \begin{cases} (b-x)\cdot(t-a) & a \le t \le x \le b \\ (b-t)\cdot(x-a) & a \le x \le t \le b \end{cases}$$

Dann:

$$y(x) = \int_a^b G(x,t)\cdot f(t,y(t),y'(t))\,dt$$

Das ist eine Fredholmsche Integralgleichung, analog für Systeme.

7.2 Differenzenverfahren

$$\begin{aligned} y'(x) &= \frac{1}{2h}\cdot\big(y(x+h) - y(x-h)\big) \\ y''(x) &= \frac{1}{h^2}\cdot\big(y(x+h) - 2y(x) + y(x-h)\big) \end{aligned}$$

- Äquidistantes Gitter: $x_i = a + i\cdot h, h = \frac{b-a}{N}$. Einsetzen liefert:

$$\begin{aligned} -y_{i-1} + 2y_i - y_{i+1} &= h^2\cdot f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \\ y_0 &= y_N = 0 \end{aligned}$$

(„tridiagonales“ (nichtlineares) Gleichungssystem)

- Aussage: Sei f hinreichend glatt und (1) besitze eine (isolierte) Lösung. Dann folgt für hinreichend kleines $h > 0$ die Existenz einer Lösung y_h von dem obigen Gleichungssystem mit

$$\|y_i^h - y(x_i)\| \le c\cdot h^2$$

für $i = 0,\ldots,N$.

7.3 Schießverfahren

- Zuordnung eines Anfangwertproblems zu (1)

$$\begin{aligned} y'(x) &= v(x) \\ v'(x) &= f(x,y(x),v(x)) \\ y(0) &= 0 \qquad v(a) = s \end{aligned}$$

mit s als freien Parameter. Sei $y(x,s), v(x,s)$ Lösung für fixiertes $s \in \mathbb{R}$.

- **Lemma** Ist y(.,s) eine Lösung von (1) $\Leftrightarrow$ y(b,s) = 0
- Numerische Lösung: Diskretisierung mittels Anfangswertaufgaben-Löser (N fixiert, Gitter vorgegeben)

  $$s \to y_N(s)$$

  mit $y_N : \mathbb{R} \to \mathbb{R}$ nichtlineare Abbildung $\Rightarrow$ Newton-Verfahren vorgegeben (evtl. Regula-Falsi) zur Bestimmung von $y_N(s) = 0$, z.B. mit Euler-Verfahren (explizit)

  $$\begin{aligned} y_{i+1}(s) &= y_i(s) + h \cdot v_i(s) \\ v_{i+1}(s) &= v_i(s) + h \cdot f(x, y_i(s), v_i(s)) \\ y_0(s) &= 0 \qquad v_0(s) = s \end{aligned}$$

  Differentiation nach s: $f = (f_x, f_y, f_v)$

  $$\begin{aligned} y'_{i+1}(s) &= y'_i(s) + v'_i(s) \\ v'_{i+1}(s) &= v'_i(s) + h \cdot (f_y(x,y,v) \cdot y'_i + f_v(x,y,v) \cdot v'_i) \\ y'(0) &= 0 \qquad v'_0(s) = 1 \end{aligned}$$

  Newton-Iteration:

  $$s_{k+1} = s_k - (v'_N)^{-1} \cdot v_N(s_k)$$

7.4 Sturm-Liouvillsche Randwert- und Eigenwertprobleme

- Betrachten lineare Randwertaufgabe 2. Ordnung, $Lu = f, Ru = \eta$
- Definition:
  1. Ist Lu in der Form

     $$Lu = -(p(x) \cdot u')' + q(x) \cdot u = -p \cdot u'' - p' \cdot u' + q \cdot u$$

     gegeben mit $p(x) > 0$, $q(x) \geq 0$, so wird $Lu = f$,

     $$\begin{aligned} \alpha_1 \cdot u(a) + \alpha_2 \cdot u'(a) &= \eta_1 \\ \beta_1 \cdot u(b) + \beta_2 \cdot u'(b) &= \eta_2 \end{aligned}$$

     Sturm-Liouvillsche Randwertaufgabe genannt.
  2. Das zugeordnete Eigenwertproblem hat die Form $Lu = \lambda \cdot u$, $Ru = 0$ ($\lambda \in \mathbb{C}$). Falls für ein $\hat{\lambda} \in \mathbb{R}$ die Randwertaufgabe eine nichttriviale Lösung $\bar{u}(x) \neq 0$ besitzt, so wird $\hat{\lambda}$ Eigenwert, $\bar{u}$ die zugehörige Eigenfunktion genannt.
- Bemerkung: Das Randwertproblem

  $$-a_0(x) \cdot u'' + a_1(x) \cdot' u' = (-a_1 \cdot u)' + (a_1 + a'_0) \cdot u'$$

  ist von Liouvillscher Form $\Leftrightarrow a_1 = -a'_0$
- Beispiel:

  $$-u'' = \lambda \cdot u \qquad u(0) = u(\lambda) = 0$$

  1. $\lambda = 0$:

     $$\begin{aligned} u'' &= 0 \\ \Rightarrow u &= c_1 + c_2 \cdot x \end{aligned}$$

     Anfangswertbedingungen:

     $$\begin{aligned} u(0) &= c_1 \stackrel{!}{=} 0 \\ u(\pi) &= \pi \cdot c_2 + c_1 \stackrel{!}{=} 0 \\ \Rightarrow c_1 &= c_2 = 0 \end{aligned}$$

     Also λ kein Eigenwert

2. $\lambda < 0$:
$$u(x) = c_1 \cdot e^{\sqrt{-\lambda}\cdot x} + c_2 \cdot e^{-\sqrt{-\lambda}\cdot x}$$
Anfangswertbedingungen:
$$\begin{aligned} c_1 + c_2 &= 0 \\ c_1 \cdot e^{\sqrt{-\lambda}\cdot x} + c_2 \cdot e^{-\sqrt{-\lambda}\cdot x} &= 0 \\ \Leftrightarrow \begin{pmatrix} 1 & 1 \\ e^{\sqrt{-\lambda}\cdot x} & e^{-\sqrt{-\lambda}\cdot x} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} &= 0 \\ \Leftrightarrow c_1 &= c_2 = 0 \end{aligned}$$
Also λ kein Eigenwert
3. $\lambda > 0$:
$$u(x) = c_1 \cdot \sin\sqrt{\lambda}\dot{x} + c_2 \cdot \cos\sqrt{\lambda} \cdot x$$
Anfangswertbedingungen:
$$\begin{aligned} u(0) &= c_2 \overset{!}{=} 0 \\ u(\pi) &= c_1 \cdot \sin\sqrt{\lambda}\pi \overset{!}{=} 0 \\ \Leftrightarrow \lambda_n &= n^2 \end{aligned}$$
Also $\lambda_n = n^2$ Eigenwerte des Sturm-Liouvilleschen Operators, abzählbar unendlich viele Eigenwerte. Es gilt:
$$(\sin n_1 x, \sin n_2 x)_{(0,\infty)} = 0$$
„Einbettung“ des Eigenwertproblems in L_2: $D(L) \subset L_2$. Definiere
$$C_0^2[a,b] := \{u \in C^2[a,b] : u(a) = u(b) = 0\}$$

- **Satz** Es gilt:
1. Für alle $u_1, u_2 \in C_0^2[a,b]$:
$$(Lu_1, u_2)_{L_2} = (u_1, Lu_2)_{L_2}$$
2. Es existiert $c > 0$, sodass für alle $u \in C_0^2[a,b]$:
$$(Lu, u)_{L_2} \geq c \cdot (\|u\|_{L_2}^2 + \|u'\|_{L_2}^2)$$
3. Der Operator (L/R) besitzt abzählbar unendlich viele Eigenwerte λ_i mit $0 < \lambda_1 \leq \ldots \to \infty$. Die zugehörige Folge $\{u_n\}_{n=1}^{\infty}$ der Eigenfunktionen bildet eine Orthonormalbasis für $L_2[a,b]$, d.h.
$$L_2[a,b] = \overline{\operatorname{span}\{u_n\}_{n=1}^{\infty}}$$
und $(u_i, u_j)_{L_2} = 0$ für $i \neq j$.

Beweis:

1. Es gilt:
$$\begin{aligned} (Lu_1, u_2)_{L_2} &= \int_a^b \left((-p \cdot u_1')' + q \cdot u_1\right) \cdot u_2 \, dx \\ &= \underbrace{(-p \cdot u_1') \cdot u_2|_a^b}_{0} + \int_a^b p \cdot u_1' \cdot u_2' + q \cdot u_1 \cdot u_2 \, dx \\ &= \underbrace{(-p \cdot u_2') \cdot u_1|_a^b}_{0} + \int_a^b p \cdot u_1' \cdot u_2' + q \cdot u_1 \cdot u_2 \, dx \\ &= \int_a^b \left((-p \cdot u_2')' + q \cdot u_2\right) \cdot u_1 \, dx = (u_1, L \cdot u_2)_{L_2} \end{aligned}$$

2. Mit $p(x) \geq p_0 > 0$ folgt:

$$\begin{aligned}\int_a^b ((-p \cdot u') + q \cdot u) \cdot u \, dx &= \int_a^b (p \cdot u'^2 + q \cdot u^2) \, dx \\ &\geq \frac{p_0}{2} \cdot \|u'\|_{L_2}^2 + \frac{p_0}{2} \cdot \int_a^b u'(x)^2 \, dx \qquad (*)\end{aligned}$$

Mit

$$\begin{aligned}\int_a^b u(x)^2 \, dx &\leq (b-a) \cdot \max |u(x)|^2 \\ &=: (b-a)\dot{|}u(x_0)|^2 \\ \& \leq (b-a) \cdot (x_0 - a) \cdot \textstyle\int_0^{x_0} u'(x)^2 \, dx & \\ &\leq (b-a)^2 \cdot \|u'\|_{L_2}^2\end{aligned}$$

folgt:

$$(*) \geq c \cdot \left(\|u\|_{L_2}^2 + \|u'\|_{L_2}^2 \right)$$

Grundidee des Galerkin-Verfahrens

- Gesucht ist die Lösung des (linearen) Randwertproblems

$$Lu = f \qquad u(a) = u(b) = 0 \qquad (1)$$

dabei sei L in Sturm-Liouville-Form gegeben.

- Differenzenverfahren: 2fache Kollokation, diskrete Approximierung der Lösung auf Gitter, diskrete Approximierung von (1) auf Gitter. (Erfordert klassische Lösung)

- zweifache Relaxation beim Galerkin-Verfahren: Vorgabe 2er Basen („Ansatzfunktionen“) des L_2:

$$L_2 = \overline{\operatorname{span}\{\varphi_j\}_{j=1}^{\infty}}^{L_2} = \overline{\operatorname{span}\{\psi_j\}_{j=1}^{\infty}}^{L_2}$$

endlich dimensionale Ansatzräume:

$$\begin{aligned}V_N &:= \operatorname{span}\{\varphi_j\}_{j=1}^N \\ W_N &:= \operatorname{span}\{\psi_i\}_{i=1}^N\end{aligned}$$

Ansatz: $u_N \in V_N$, d.h.

$$u_N = \sum_{i=1}^N c_i \cdot \varphi_i$$

Die Projektion des Defektes $Lu - f$ von (1) in W_N soll verschwinden (Testfunktionen) d.h.

$$\begin{aligned}(Lu_N - f, \psi_j)_{L_2} &= 0 \\ \Leftrightarrow \left(L \left(\sum_{i=1}^N c_i \cdot \varphi_i \right) - f, \psi_j \right) &= 0 \\ \Leftrightarrow \sum_{i=0}^N c_i \cdot (L \cdot \varphi_i, \varphi_j) &= (f, \psi_j)\end{aligned}$$

also lineares Gleichungssystem zur Bestimmung von c_i, $A := ((L\varphi_i, \psi_j))$

- Bemerkungen:

1. Mit Teil 1) im Beweis des Satzes

$$(L\varphi_i, \psi_j) = \int_a^b (p \cdot \varphi_i' \cdot \psi_j' + q \cdot \varphi_i \cdot \psi_j) \, dx$$

Also A symmetrisch (im Prinzip)

2. Ritz-Galerkin-Verfahren: $V_N = W_N$
3. „Ideal“ wäre $V_N = \text{span}\{u_i\}_{i=1}^N$ mit u_i als Eigenfunktionen von $Lu_i = \lambda \cdot u_i$, dann $\hat{f}_i = (f_i, u_i)_{L_2}$ und damit A = diag(λ_i). Dann

$$u_N = \frac{f_i}{\lambda_i} \cdot u_i$$

Die Eigenwerte/-funktionen sind jedoch nur in Spezialfällen bekannt, Methode der finiten Elemente. Ansatzfunktionen:

$$\varphi_j(x) = \begin{cases} \frac{1}{n} \cdot (x - x_j) & x \in [x_{j-1}, x_j] \\ \frac{1}{n} \cdot (x_{j+1} - x) & x \in [x_j, x_{j+1}] \end{cases}$$

Speziell für p=q=1:

$$u_N = \sum_{i=1}^{N-1} c_j \cdot \varphi_j(x)$$