Humboldt-Universität zu Berlin
Institut für Mathematik
Sommersemester 2010

Numerik

Prof. Carstensen und Prof. Schröder

Bodo Graumann

19. Mai 2014

Diese Dokument wurde auf http://bodograumann.de veröffentlicht. Es steht unter der Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) Lizenz.

Der Code wurde mit gvim sowie vim-latex erstellt und mit xelatex kompiliert – all das auf Gentoo Linux. Meinen Dank an die Freie Software Community und die TEX-Kollegen auf TEX.SX für ihre Hinweise und Unterstützung.
Bitte schreibt mir eure Kommentare und Verbesserungsvorschläge zu diesem Dokument! Ihr könnt mir entweder direkt mailen oder das Kontaktformular auf meiner Internetseite benutzen.

Inhaltsverzeichnis

1 Nichtlineare Gleichungen

Motivation Es gibt keine allgemeine Lösungsformel für Polynome mit Grad größer gleich 5.

Motivationsbeispiel für ein Iteratives Verfahren Das Verfahren nach Heron von Alexandria für die Berechnung $\sqrt{a}$:

$$x_0 = 1.4142136, \quad x_{i+1} = \frac{1}{2}\left(x_i + \frac{a}{x_i}\right)$$

Die Wurzel ensteht dann als Fixpunkt der Iterationsfunktion, da $x = \frac{1}{2}\left(x + \frac{a}{x}\right) \Leftrightarrow x^2 = a$ gilt.

Im Folgenden gehen wir davon aus, dass $f \in C^1(D;\mathbb{R}^n)$ mit $D \subset \mathbb{R}^n$ offen und nicht leer. Dann suchen wir in der Regel die Nullstellen von f: $\{\, x \in D \mid f(x) = 0 \,\}$.

1.1 Newton-Verfahren

Für $x_j \in D$ sei $g_j(x)$ das Taylor-Polynom erster Ordnung von f um x_j, das heißt

$$g_j(x) = f(x_j) + f'(x_j)(x - x_j)$$

Sofern $f'(x_j)$ regulär ist, hat $g_j(x)$ genau eine Nullstelle, die wir mit x_{j+1} bezeichnen:

$$x_{j+1} := x_j - f'(x_j)^{-1} f(x_j)$$

1 **Algorithmus: Newton-Verfahren**
Zum Startwert x_0 in D berechne man

$$x_{j+1} := x_j - f'(x_j)^{-1} f(x_j), \quad j \in \mathbb{N}$$

solange $f'(x_j)$ existiert und regulär ist.

2 **Definition: „Durchführbarkeit“**
Das Newton-Verfahren heißt durchführbar mit Startwert x_0 falls es nicht abbricht.

3 **Definition: „Grenzwerte“**

$$(x_j) \to x^* \qquad :\Leftrightarrow \qquad \lim_{j\to\infty} |x^* - x_j| = 0$$

$$(x_j) \to x^* \text{ } q\text{-linear} \qquad :\Leftrightarrow \qquad q := \limsup_{\substack{j\to\infty \\ x^* \neq x_j}} \frac{|x^* - x_{j+1}|}{|x^* - x_j|} < 1$$

$$(x_j) \to x^* \text{ superlinear} \qquad :\Leftrightarrow \qquad q = 0$$

$$(x_j) \to x^* \text{ } q\text{-qradratisch} \qquad :\Leftrightarrow \qquad q_2 := \limsup_{\substack{j\to\infty \\ x^* \neq x_j}} \frac{|x^* - x_{j+1}|}{|x^* - x_j|^2} < \infty$$

4 **Satz: lokale Konvergenz des Newton-Verfahrens**

Sei $x^* \in D$ eine Nullstelle von $f \in C^1(D, \mathbb{R}^n)$ und $f'(x^*)$ regulär. Dann gilt

1. Durchführbarkeit: $\exists \rho > 0 \forall x_0 \in B_\rho(x^*) \subset D$Newton-Verfahren ist durchführbar $\wedge (x_j) \to x^*$
2. Superlineare Konvergenz: $(x_j) \to x^*$ superlinear
3. Quadratische Konvergenz: Wenn f' Lipschitz-stetig ist, so konvergiert $(x_j) \to x^*$ q-quadratisch.

5 **Lemma: Banachsches Störungslemma**

Gegeben sei eine induzierte Matrixnorm $\|\cdot\|$ und zwei Matrizen $A, B \in \mathbb{R}^{n\times m}$ sowie $\alpha > 0$ mit $\|A - B\| \leq \alpha$. Ist A regulär und $\|A^{-1}\| < \frac{1}{\alpha}$, dann ist B regulär mit

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \alpha\|A^{-1}\|}$$

Dies bedeutet insbesondere, dass $(GL(n \times m), \|\cdot\|)$ offen ist.

Beweis (5)

$$B = A - AA^{-1}(A - B) = A(1 - \underbrace{A^{-1}(A - B)}_{=:C})$$

Mit $z \in \mathbb{R}^n \setminus \{0\}$ und $(1 - C)z = 0$ folgt:

$$Cz = z \Rightarrow \frac{|Cz|}{|z|} = 1 \leq \|C\| = \|A^{-1}(A - B)\| \leq \|A^{-1}\|\|A - B\| < 1 \lightning$$

Somit ist $1 - C$ regulär und damit auch B. Mit Dreiecksungleichung erhalten wir:

$$(1 - \|C\|)\|(1 - C)^{-1}\| \leq \|(1 - C)^{-1}\| - \|C(1 - C)^{-1}\| \leq \|(1 - C)(1 - C)^{-1}\| = 1$$

$$\Rightarrow \|B^{-1}\| = \|(1 - C)^{-1}A^{-1}\| \leq \|A^{-1}\|\|(1 - C)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|C\|} \leq \frac{\|A^{-1}\|}{1 - \alpha\|A^{-1}\|}$$

6 **Satz: Integralmittelwertsatz**

Betrachten wir eine Strecke $\overline{ab}$, die vollständig in D enthalten ist. Dann ist

$$f(b) - f(a) = \left(\int_0^1 f'(sb + (1 - s)a)\mathrm{d}s\right)(b - a)$$

(Die Integration erfolgt komponentenweise.)

Beweis (6) Für $j \in \{1, 2, \cdots, n\}$ definieren wir $g_j \in C^1(D, \mathbb{R}), g_j(s) := f_j(sb + (1 - s)a)$. Nach dem Hauptsatz der Differential- und Integralrechnung gilt dann:

$$f_j(b) - f_j(a) = g_j(1) - g_j(0) = \int_0^1 \frac{\partial g_j}{\partial s}(s)\mathrm{d}s = \left(\int_0^1 f_j'(sb + (1 - s)a)\mathrm{d}s\right)(b - a)$$

Beweis (Satz 4 (lokale Konvergenz des Newton-Verfahrens)) f' ist stetig in x^*, somit $\exists \rho > 0 \forall x \in B_\rho(x^*) \subset D \| f'$
Mit $\|f'(x^*)^{-1}\| = \frac{1}{6\varepsilon} < \frac{1}{\varepsilon}$ folgt aus dem Störungslemma, dass $f'(x)$ regulär ist und

$$\forall x \in B_\rho(x^*) \|f'(x)^{-1}\| \leq \frac{\|f'(x^*)^{-1}\|}{1 - \varepsilon \|f'(x^*)^{-1}\|} = \frac{1}{5\varepsilon}$$

Für $x_j \in B_\rho(x^*)$ folgt nach Einsetzen mit dem Integralmittelwertsatz

$$\begin{aligned}
x^* - x_{j+1} &= x^* - x_j + f'(x_j)^{-1}(f(x_j) - f(x^*)) \\
&= f'(x_j)^{-1}(f'(x_j)(x^* - x_j) - \int_0^1 \left(f'(x_j) - f'(sx_j + (1-s)x^*)\right) \mathrm{d}s \, (x^* - x_j)) \\
&= f'(x_j)^{-1} \int_0^1 f'(sx_j + (1-s)x^*) \mathrm{d}s \, (x^* - x_j) \\
\|x^* - x_{j+1}\| &\leq \underbrace{\|f'(x_j)^{-1}\|}_{\leq \frac{1}{5\varepsilon}} \underbrace{\| \int_0^1 \left(f'(x_j) - f'(sx_j + (1-s)x^*)\right) \mathrm{d}s \, \|}_{\substack{\leq \int_0^1 \| \left(f'(x_j) - f'(sx_j + (1-s)x^*)\right) \| \mathrm{d}s \\ \leq \|f'(x^*) - f'(x_j)\| + \max_{0 \leq s \leq 1} \|f'(x^*) - f'(sx_j + (1-s)x^*)\| \\ \leq \varepsilon + \varepsilon}} \|x^* - x_j\| \\
&\leq \frac{2}{5}\|x^* - x_j\| \leq \frac{2}{5}\rho < \rho
\end{aligned}$$

Folglich ist $x_{j+1} \in B_\rho(x^*) \subset D$. Per vollständiger Induktion über j folgt 1 sowie die $q = \frac{2}{5}$-lineare Konvergenz $(x_j) \to x^*$.

Nun definieren wir

$$q_j := \frac{2}{5\varepsilon} \max_{\|x - x^*\| \leq \rho} \|f'(x^*) - f'(x)\|$$

Wegen $\lim_{j\to\infty} \|x^* - x_j\| = 0$ und $f \in C^1$ gilt $\lim_{j\to\infty} q_j = 0$ und dann folgt analog zu obiger Abschätzung

$$\|x^* - x_{j+1}\| \leq \frac{1}{5\varepsilon}\|x^* - x_j\| \cdot 2\frac{q_j}{\frac{2}{5\varepsilon}} \Rightarrow 2$$

Wenn f' in $B_\rho(x^*)$ Lipschitz-stetig ist, dann existiert ein L mit $0 < L < \infty$, sodass

$$\|f'(x_j) - f'(sx_j + (1-s)x^*)\| \leq L\|x_j - sx_j + (1-s)x^*\| = \underbrace{(1-s)}_{\leq 1}\|x^* - x_j\|$$

Damit folgt

$$\|x^* - x_{j+1}\| \leq \underbrace{\frac{6}{5}\|f'(x^*)^{-1}\|L}_{=:C}\|x^* - x_j\|^2$$

also q-quadratische Konvergenz.

7 **Bemerkung:**

1. Die Menge der $x_0 \in D$ die gegen x^* im Newton-Verfahren konvergieren, heißt *Einzugsbereich* der Nullstelle x^* von f.
2. Im allgemeinen sind nur lokale Konvergenzsätze möglich. Für den Spezialfall der konvexen Funktionen kann man globale Konvergenz für eine gedämpfte Version des Newton-Verfahrens beweisen.
3. Die Regularität von $f'(x^*)$ ist im allgemeinen nicht notwendig. Sei beispielsweise $n = 1$, $f(x) = (x-1)^m$, $m > 1$. Dann erhalten wir
$$x_{j+1} = x_j - \frac{(x_j-1)^m}{m(x_j-1)^{m-1}} = x_j - \frac{1}{m}(x_j - 1) = (1 - \frac{1}{m})x_j + \frac{1}{m}$$
$$1 - x_{j+1} = (1 - \frac{1}{m})(1 - x_j)$$
Dies ist also q-lineare, nur nicht quadratische Konvergenz.
4. Abbruch-Kriterium:
   a) $j \geq 100$
   b) $\|f(x_{j+1})\| > \|f(x_j)\|$
   c) $\|f'(x_j)^{-1}\|\|f(x_j)| \ll 1$

1.2 Weierstraß-Verfahren zur simultanen Berechnung aller einfachen Polynomnullstellen

Es sei $f(z) := (z - z_1)\cdots(z - z_n)$ für $z \in \mathbb{C}$ mit den einfachen Nullstellen $z_1, \cdots, z_n \in \mathbb{C}$. Wir wählen einen Startvektor $(x_1, \cdots, x_n) \in D := \left\{ (\zeta_1, \cdots, \zeta_n) \in \mathbb{C}^n \;\middle|\; i \neq j \Rightarrow \zeta_i \neq \zeta_j \right\}$. Dann definieren wir die Abbildung

$$(x_1, \cdots, x_n) \mapsto (x_j - W_j)_{j=1,\cdots,n} \in \mathbb{C}^n \quad \text{mit} \quad W_j = \frac{f(x_j)}{\prod_{\substack{k=1 \\ k\neq j}}^{n} (x_j - x_k)}$$

Aufgabe Man zeige lokale quadratische Konvergenz für dieses Verfahren, indem man zeigt, dass es ein Newton-Verfahren ist.

1.3 Sekantenverfahren

1.3.1 Einschrittverfahren

Wir definieren die Sekante als

$$F(x_j, h)_{k,l} := \begin{cases} f'(x_j)_{k,l} & \text{falls} \quad h_{lk} = 0 \\ \frac{f_k(x_j + h_{k,l}e_l) - f_k(x_j)}{h_{k,l}} & \text{sonst} \end{cases}$$

Wobei $h \in \mathbb{R}^{n\times n}$ mit $x_j + h_{k,l}e_l \in D$ gewählt werden muss. Nach dem Störungslemma ist $F(x_j, h)$ regulär für $\|h\| \ll 1$. Diese Matrix $F(x_j, h)$ setzen wir dann im Newton-Verfahren ein.

1.3.2 Reguli Falsi (2-Schrittverfahren) für $n = 1$

Ausgegangen wird von 2 Startwerten. Für $j \in \mathbb{N}$ und x_j, x_{j+1} bestimmt man x_{j+2} als Nullstelle von

$$g_j(x) := f(x_j) + \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j}(x - x_j)$$

$$\text{d.h.} \quad x_{j+2} = x_j - \frac{x_{j+1} - x_j}{f(x_{j+1}) - f(x_j)} f(x_j) = \frac{x_j f(x_{j+1}) - x_{j+1} f(x_j)}{f(x_{j+1}) - f(x_j)}$$

sofern $f(x_{j+1}) \neq f(x_j)$.

8 **Satz: lokale Konvergenz des Einschrittsekantenverfahrens**

Sei $x^* \in D \subset \mathbb{R}^n$ eine reguläre Nullstelle von $f \in C^1(D, \mathbb{R}^n)$ (d.h. $f(x^*) = 0$ und $f'(x^*)$ regulär). Dann ist das Sekantenverfahren durchführbar, $j \in \mathbb{N}$:

$$\exists \rho, \gamma > 0 \forall x_0 \in \bar{B}_\rho(x^*) \subset D \forall h^{(j)} \in [\gamma, \gamma]^{n \times n} x_{j+1} = x_j - F(x_j, h^{(j)})^{-1} f(x_j)$$

und konvergiert gegen x^*. Ist zusätzlich $\lim_{j \to \infty} \|h^{(j)}\| = 0$, dann ist (x_j) superlinear konvergent.

Beweis (8) (Skizze)

$$x^* - x_{j+1} = x - x_j - F(x_j, h^{(j)})^{-1}(f(x^*) - f(x_j))$$

$$= \underbrace{F(x_j, h^{(j)})^{-1}}_{\text{beschränkt}} \left(\int_0^1 \underbrace{(F(x_j, h^{(j)}) - f'(sx^* + (1-s)x_j))}_{\substack{\text{beschränkt für } h^{(j)} \\ \text{und klein für } \|h^{(j)}\| \ll 1}} \mathrm{d}s \right) (x - x_j)$$

Weiter ähnlich zum Beweis von Satz 4 (lokale Konvergenz des Newton-Verfahrens).

Bemerkung

1. Vorteile des Sekantenverfahrens gegenüber Newton-Verfahren:
   - Funktionswertauswertungen, keine Ableitungen
   - Mehrschrittverfahren benutzen bekannte Funktionswerte
   - $n = 1, \mathbb{R}$, Inklusionen vermöge $f(x_j) < 0 < f(x_{j+1})$
2. Nachteile des Sekantenverfahrens gegenüber dem Newton-Verfahren
   - im Allgemeinen *keine* quadratische Konvergenz
   - numerische Berechnung der Differenzenquotienten numerisch instabil

9 **Satz: lokaler Konvergenzsatz der Regula Falsi**

Sei x^* eine einfache Nullstelle von $f \in C^1(D, \mathbb{C})$ mit einer offenen Menge $D \subset \mathbb{C}$. (Analog für $\mathbb{R}$) Dann gilt

$$\exists \rho > 0 \forall x_0, x_1 \in \bar{B}_\rho(x^*), x_0 \neq x_1 \text{ist die Regula Falsi durchführbar}$$

und die Folge (x_j) ist superlinear konvergent gegen x^*. Sofern f' Lipschitz-stetig ist, so konvergiert (x_j) R-Ordnung $\frac{1+\sqrt{5}}{2}$ in dem Sinne, dass die Folge $(|x^* - x_j|)_{j\in\mathbb{N}}$ eine obere Schranke mit der q-Konvergenzordnung $\frac{1+\sqrt{5}}{2}$ besitzt:

$$\exists C, 0 < \delta < 1 |x^* - x_j| \leq C\delta^{\left(\left(\frac{1+\sqrt{5}}{2}\right)^j\right)}$$

Beweis (9) (Skizze) $\delta_j := |x^* - x_j|$, bisherige Fehleranalysis

$$\delta_{j+2} \leq C\delta_{j+1}(\delta_j + \delta_{j+1})$$

Induktionsbeweis: Wenn $\delta_j < \delta_{j-1}$, dann $\delta_{j+1} < 2C\delta_j\delta_{j-1}$. Ansatz $\delta_j =: \delta^{a_i}$. $\delta^{a_{j+1}} < 2W\delta^{a_j + a_{j+1}}$. $a_{j+1} \geq a_j + a_{j-1} \Rightarrow$ Wachstum der Fibonacci-Zahlen.

Bemerkung

- Regula Falsi ist gegebenenfalls dem Newton-Verfahren vorzuziehen für $n = 1$.
- Mehrschrittverfahren: zu $x_j, \cdots, x_{j-m} \in D$ bestimmt man $A_j \in \mathbb{R}^{n\times n}$ und $g_j(x) = f(x_j) + A_j(x - x_j)$ mit $g_j(x_k) = f(x_k)$. Das führt auf *nm* Gleichungen

  $$f(x_{j-k}) - f(x_j) = A_j(x_{j-k} - x_j)$$

  mit n^2 Unbekannten in A_j und gegebenenfalls zu Stabilitätsproblemen mit fast singulären Matrizen.
- Dämpfung: Für alle $j \in \mathbb{N}_0$ wählt man $0 < \alpha_j \leq 1$:

  $$x_{j+1} = x_j - \alpha_j F(x_j, h^{(j)})^{-1} f(x_j)$$

  z.B. mit $|f(x_{j+1})| < |f(x_j)|$
- Modifizierte Newton-Verfahren: Eine in der Regel nicht streng monoton steigende Folge $(k_j)_{j\in\mathbb{N}}$ mit $0 \leq k_j \leq j$ definiert

  $$x_{j+1} = x_j - F(x_{k_j}, h^{(k_j)})^{-1} f(x_j)$$

  Dann braucht man nicht so viele Auswertungen der Ableitung.

1.4 Quasi-Newton-Verfahren (QNV)

Im Iterationsschritt $j \in \mathbb{N}$ eines QNV definieren wir $g_j(x) = f(x_j) + A_j(x - x_j)$ mit regulären $A_j \in \mathbb{R}^{n\times n}$ zur Berechnung von $x_{j+1} = x_j - A_j^{-1} f(x_j)$. Im Newton-Verfahren ist $A_j = f'(x_j)$.

Quasi-Newton-Bedingung

$$\forall j \in \mathbb{N} g_j(x_{j-1}) = f(x_{j-1}) \tag{1-1}$$

1.4.1 Broyden Update

Die Idee ist, A_j durch eine Rang-1-Modifikation von A_{j-1} zu berechnen mittels

$$A_j(x_j - x_{j-1}) = f(x_j) - f(x_{j-1})$$

was sofort aus der Quasi-Newton-Bedingung 1-1 erwächst. Außerdem fordern wir mit $z_j := x_j - x_{j-1}$

$$\begin{aligned}&\forall z \in \mathbb{R}^n, z \perp z_j (A_j - A_{j-1})z = 0\\ &\Rightarrow A_j = A_{j-1} - u_j z_j^T\end{aligned}$$

für ein $u_j \in \mathbb{R}^n$.

Rang-1-Aufdatierungsformel nach Broyden

$$A_j = A_{j-1} - \frac{(A_j - A_{j-1})z_j z_j^T}{|z_j|^2} = A_{j-1} - \frac{(f(x_j) - f(x_{j-1}) - A_{j-1}z_j)z_j^T}{|z_j|^2}$$

Sherman-Morrison-Formel Sei I die $n \times n$-Einheitsmatrix und $a, b \in \mathbb{R}^n$. Dann ist wegen Determinantenentwicklung nach der letzten Spalte und elementaren Zeilenumformungen

$$|I - ab^T| = \begin{vmatrix} 1 - a_1b_1 & a_1b_2 & \cdots & a_1b_n & 0\\ a_2b_1 & 1 - a_2b_2 & \cdots & a_2b_n & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ b_1 & b_2 & \cdots & b_n & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & \cdots & 0 & a_1\\ 0 & 1 & \cdots & 0 & a_2\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ b_1 & b_2 & \cdots & b_n & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 0 & \cdots & 0 & a_1\\ 0 & 1 & \cdots & 0 & a_2\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & 0 & 1 - a_1b_1 - \cdots - a_nb_n \end{vmatrix} = 1 - a \cdot b$$

$$\Rightarrow (I - ab^T)^{-1} = I_n + \frac{ab^T}{1 - a \cdot b}$$

denn es gilt

$$(I - ab^T)(I + \frac{ab^T}{1 - a \cdot b}) = I - ab^T \left(\underbrace{1 - \frac{1}{1 - a \cdot b} + \frac{a \cdot b}{1 - a \cdot b}}_{=0} \right)$$

10 **Satz: lokale Konvergenz des Broyden-Verfahrens**

Sei x^* eine Nullstelle von $f \in C^1(D, \mathbb{R}^n)$ in $D \subset \mathbb{R}^n$ mit $f'(x^*)$ regulär. Dann $\exists \rho, \delta > 0 \forall x_0 \in \bar{B}_\rho(x^*) \subset D \forall A_0 \in B_\|$ ist das Broyden-Verfahren für J Schritte durchführbar und entweder $J = \infty$ und die berechnete Folge $(x_j) \to x^*$ superlinear oder $J < \infty$ und $x_J = x^*$.

Beweis (10) (siehe Sch79, S. 142 ff.)

Bemerkungen

1. Modifikation mit $A_j = F(x_j, h^{(j)})$ falls $|A_j^{-1} f(x_j)|$ zu klein, weiter mit Broyden-Update
2. Alternativ sind modifizierte Rang-1 bzw. Rang-2 Modifikationen möglich, die gegebenenfalls strukturerhaltend / symmetrieerhaltend sind.

1.5 Homotopie-Methoden

Seien $f, g \in C^1(D, \mathbb{R}^n)$ mit $x_*, x_0 \in D$, sodass $f(x^*) = 0$, $f'(x^*)$ regulär und $g(x_0) = 0$. In der Anwendung ist nun x_0 bekannt und x^* unbekannt. Wir definieren nun

$$H := \begin{cases} D \times [0,1] \to \mathbb{R}^n \\ \quad (x,t) \mapsto tf(x) + (1-t)g(x) \end{cases}$$

Wir setzen jetzt voraus, dass

$$\forall 0 \le t \le 1 \exists x_0(t) \in DH(x_0(t), t) = 0$$

und $x_0(t)$ sei stetig in t. Für eine Partition $0 = t_0 < t_1 < \cdots < t_N = 1$ für $N \in \mathbb{N}$ (welche adaptiv gewählt werden kann) löse man

$$H(x, t_j) = 0$$

für die $j = 1, \cdots, N$ mit k_j Iterationen und Startwert $x_{j-1,k_{j-1}}$ bzw. x_0 für $j = 0$ mit einem beliebigen lokal konvergenten Iterationsverfahren.

11 **Algorithmus: Eingebettete Newton-Verfahren**

Es sei $H\colon D \times [0,1] \to \mathbb{R}^n$ stetig mit $x_0 \in D$ und $H(x_0, 0) = 0$.

$$\forall j = 1, \cdots, N x_{j,0} := \begin{cases} x_0 & \text{falls } j = 1 \\ x_{j-1,k_{j-1}} & \text{sonst} \end{cases}$$

$$\forall k = 0, \cdots, k_j - 1 x_{j,k+1} = x_{j,k} - \left(\frac{\partial H}{\partial x}(x_{j,k}, t_j)\right)^{-1} H(x_{j,k}, t_j)$$

Damit erhalten wir $(x_{j,k_j})_{j=0,\cdots,N}$ als Approximation von $(x_0(t_j))_{j=0,\cdots,N}$.

12 **Satz: Konvergenzsatz**

Zu $H \in C(D \times [0,1], \mathbb{R}^n)$ mit $H_x := \frac{\partial H}{\partial x} \in C(D \times [0,1], \mathbb{R}^{n \times n})$ existiere $x_0(t) \in C([0,1], D)$ sodass $\forall 0 \le t \le 1 H(x_0(t), t) = 0$ und $H_x(x_0(t), t)$ regulär und seien $k = k_1 = k_2 = \cdots = k_N \in \mathbb{N}, 0 = t_0 < t_1 < \cdots < t_N = 1 = t_N$ hinreichend fein und $|x_0(0) - x_0|$ hinreichend klein. Dann ist das eingebettete Newton-Verfahren durchführbar und $x_{N,k}$ liegt im Einzugsbereich der Nullstelle $x^* = x_0(1)$.

Beweis (12) (Skizze) *Behauptung 1*:

$$\exists \rho > 0 \exists 0 < q < 1 \forall 0 \le t \le 1 \forall x_{t,0} \in B_\rho(x_0(t)) |x_0(t) - x_{t,k}| \le q^k |x_0(t) - x_{t,0}|$$

wobei $x_{t,k}$ die Näherung nach k Schritten des Newton-Verfahrens zur Lösung von $H(\xi, t) = 0$ mit dem Startwert $x_{t,0}$ ist.

Beweisidee $\left\{ (x_0(t), t) \;\middle|\; 0 \le t \le 1 \right\} \subset\subset D \times [0,1]$. H, H_x, H_x^{-1} und dort gleichmäßig stetig. Das Störungslemma und die Argumente im Beweis des lokalen Konvergenzsatzes sind daher gleichmäßig in t und liefern die universelle (von t nicht abhängigen) Konstanten q, ρ der Behauptung.

Behauptung 2:

$$\max \left\{ |x_0(t_j) - x_0(t_{j-1})| \;\middle|\; j = 1, 2, \cdots, N \right\} < \delta := (1 - q^k)\rho$$

für hinreichend feine Zerlegungen, da x_0 gleichmäßig stetig ist.

Behauptung 3:

$$\forall j = 0, \cdots, N - 1 x_{j+1} := x_{t_{j+1},0} := x_{t_j,k} \in \bar{B}_\rho(x(t_{j+1}))$$

Beweis: Vollständige Induktion über j: Wenn $x_j \in \bar{B}_\rho(x_0(t_j))$, so

$$|x_{j+1} - x(t_{j+1})| \le \underbrace{|x_0(t_j) - x_{t_j,k}|}_{\le q^k |x_0(t_j) - x_j| \le \rho q^k} + \underbrace{|x_0(t_{j+1}) - x_0(t_j)|}_{< \delta = (1-q^k)\rho}$$

Bemerkung

1. In der Praxis wählen wir kN möglichst klein, z.B. $k = 1$.
2. Adaptive Zeitschrittwahl ($k_j \approx 3 \Rightarrow$ Zeitschrittkorrektur)
3. Homotopie versagt bei Verzweigungspunkte, d.h. wenn H_x fast singulär ist.

1.6 Ausgleichsprobleme

Beispiel Sei $f(x) = Ax - b$ mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ mit vollem $Rang(A) = n \leq m$. Dann suchen wir die Minimalstelle von $\frac{1}{2}|f(x)|^2$ mit der Lösung $x^* = A^\dagger b$ und der Moore-Penrose-Inversen $A^\dagger = (\underbrace{A^T A}_{\text{SPD}})^{-1} A^T$.

Hier: nichtlineare Ausgleichsprobleme mit $f \in C^1(D, \mathbb{R}^m)$ für eine offene Menge $D \subset \mathbb{R}^n$ und der Minmalstelle x^* von $\varphi(x) = \frac{1}{2}|f(x)|^2$ in D mit der euklidischen Norm $|\cdot|$, das heißt $\varphi(x) = \frac{1}{2} f(x) \cdot f(x) = \frac{1}{2} f(x)^T f(x)$.

Für eine Näherung $x_j \in D$ sei

$$g_j(x) = f(x_j) + f'(x_j)(x - x_j) =: b - Ax$$

mit geeigneter Matrix $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Die Lösung dieses lokalen Minimierungsproblems für $\frac{1}{2}|g_j(x)|^2$ ist dann $x_j - f'(x_j)^\dagger f(x_j)$.

13 **Lemma: Abstiegsrichtung**

Sofern $f'(x_j)$ vollen Spaltenrang besitzt und $f(x_j) \neq 0$, ist $z_j := -f'(x_j)^\dagger f(x_j)$ eine Abstiegsrichtung ist, in dem Sinne, dass

$$\exists \varepsilon > 0 \forall \alpha < \varepsilon \varphi(x_j + \alpha z_j) < \varphi(x_j)$$

Beweis (13)

$$\varphi(x_j + \alpha z_j) = \varphi(x_j) + \alpha \varphi'(x_j) \cdot z_j + O(\alpha^2)$$

$$\varphi'(x_j) = \frac{1}{2}\left(f'(x_j)^T f(x_j) + (f(x_j)^T f'(x_j))^T\right) = f'(x_j)^T f(x_j)$$

$$\Rightarrow f(x_j)^T f(x_j) = f(x_j)^T \underbrace{f'(x_j) f'(x_j)^\dagger}_{= f'(x_j)\left(f'(x_j)^T f'(x_j)\right)^{-1} f'(x_j)^T} f(x_j) = -\varphi'(x_j)^T z_j$$

$$\Rightarrow \varphi'(x_j) \cdot z_j < 0$$

Bezeichnung Die Minimierung von $\varphi(x_j + \alpha z_j)$ für $\alpha \in \mathbb{R}$ mit $x_j + \alpha z_j \in D$ heißt *line search*.

14 **Algorithmus: Schrittweitenwahl nach Armijo**

Eingabe: $0 < \gamma < \frac{1}{2}$, x_j, z_j

Berechne: $\alpha := 1$. Halbiere $\alpha > 0$ solange $\varphi(x_j + \alpha z_j) > \varphi(x_j) + \gamma \alpha \varphi'(x_j) z_j$.

Ausgabe: $\alpha > 0$ mit $\varphi(x_j + \alpha z_j) \leq \varphi(x_j) + \gamma \alpha \varphi'(x_j) z_j$

15 **Algorithmus: Gedämpftes Gauß-Newton-Verfahren**

Eingabe: $f \in C^1(D, \mathbb{R}^m)$, $\varphi(x) := \frac{1}{2} f(x) \cdot f(x)$, $x_0 \in D \subset \mathbb{R}^n$, $n \leq m$, $\gamma > 0$

Berechne: Für $j = 0, 1, 2, \cdots$: $z_j := -f'(x_j)^\dagger f(x_j)$; $\alpha_j :=$ Schrittweite nach Armijo; $x_{j+1} := x_j + \alpha_j z_j$.

Ausgabe: Folge (x_j).

16 **Satz: Konvergenzsatz**

Unter den obigen Voraussetzungen sei f' Lipschitz-stetig und habe vollen Spaltenrang auf $N_0 := \left\{ x \in \mathbb{R}^n \;\middle|\; |f(x)| \leq |f(x_0)| \right\}$ und $\exists \alpha > 0 \forall x \in N_0 \forall z \in \mathbb{R}^n \alpha |z| \leq |f'(x)z|$, dann gilt:

1. Das gedämpfte Gauß-Newton-Verfahren ist durchführbar mit x_0 und $(\varphi(x_j))_{j\in\mathbb{N}}$ ist monoton fallend wobei $\lim_{j\to\infty} |f'(x_j)^T f(x_j)| = 0$.
2. Sofern N_0 kompakt und φ_0 auf N_0 genau einen stationären Punkt x^* hat, dann bricht das Verfahren entweder beim Index k ab und $f'(x_n)^T f(x_n) = 0$ sowie $x_k = x^*$, oder $\lim_{j\to\infty} x_j = x^*$.
3. Ist weiterhin $f(x^*) = 0$, dann existiert ein k_0 sodass für alle $j \geq k$: $\alpha_j = 1$. Das heißt das gedämpfte Gauß-Newton-Verfahren geht ab dem Index k_0 in das quadratisch konvergente Newton-Verfahren über.

2 Interpolation

Grundproblem Darstellung und Auswertung von Funktionen

- Beispielsweise ist $f(x)$ ist nur für diskrete Punkte $x_0, \cdots, x_n$ bekannt und man möchte f rekonstruieren.
- Es kann auch vorkommen, dass f durch einen komplizierten Funktionsausdruck dargestellt wird, der nicht leicht auszuwerten ist, sodass man diesen durch eine Approximation ersetzen will, die einfacher auszuwerten ist.

Ziel: Rekonstruktion bzw. Darstellung mit einfach strukturierten Funktionen einer bestimmten Klasse. (Z.B. Polynome, rationale Funktionen, trigonometrische Funktionen)

Beispiel Ist f an den Stellen $x_1, \cdots, x_4$ gegeben, so können wir mittels Lagrange-Basispolynome interpolieren:

$$g(x) = \sum_{j=1}^{4} f(x_j) \prod_{\substack{k=1 \\ k\neq j}}^{4} \frac{(x - x_k)}{(x_j - x_k)}$$

2.1 Polynominterpolation

Hier wählen wir als Interpolationsfunktionenklasse den Vektorraum der Polynomfunktionen.

Lagrangesche Interpolationsaufgabe Es sei $x_0, \cdots, x_n \in \mathbb{R}$, $i \neq j \Rightarrow x_i \neq x_j$ Stützstellen und $y_0, \cdots, y_n \in \mathbb{R}$ Knotenwerte. Dann suchen wir eine Polynomfunktion p des Grades n, das Interpolationspolynom so dass $p(x_i) = y_i$ für $i = 0, \cdots, n$.

17 **Satz: eindeutige Lösbarkeit**
Die Lagrangesche Interpolationsaufgabe besitzt eine eindeutig bestimmte Lösung.

18 **Definition: „Lagrangesche Basispolynome“**

$$L_i^{(n)}(x) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

heißt Lagrangesches Basispolynom.

19 **Satz: Lagrangesche Darstellung der Lösung**
Die Lösung der Lagrangeschen Interpolationsaufgabe ist

$$p(x) = \sum_{i=0}^{n} y_i L_i^{(n)}$$

Dies ist die *Lagrangesche Darstellung*.

Das Problem bei der Lagrangeschen Darstellung ist, dass ein Hinzufügen weiterer Stützstellen alle bisherige Basispolynome verändert. Eine Alternative stellen die Newtonschen Basispolynome dar:

$$N_0(x) = 1$$
$$N_i(x) = \prod_{j=0}^{i-1} (x - x_j)$$

Dann können die Koeffizienten ebenfalls iterativ bestimmt werden.

20 **Bemerkung:**

1. Die Hinzunahme einer neuen Basisfunktion N_{n+1} ist leicht durchführbar.
2. Mit Hilfe der „dividierten Differenzen“ können die Koeffizienten numerisch stabiler bestimmt werden.

21 **Definition: „dividierte Differenzen“**
Die rekursiv definierten Werte

$$y\left[x_i\right] = y_i, i = 0, \cdots, n$$
$$y\left[x_i, \cdots, x_{i+k}\right] = \frac{y\left[x_{i+1}, \cdots, x_{i+k}\right] - y\left[x_i, \cdots, x_{i+k-1}\right]}{x_{i+k} - x_i}$$

heißen *dividierte Differenzen*.

22 **Satz: Newtonsche Lösung**

Es sei $p_{i,i+k}$ die Interpolationspolynomfunktion k-ten Grades zu den Knoten $x_i, \cdots, x_{i+k}$ und Knotenwerten $y_i, \cdots, y_{i+k}$. Dann gilt

$$p_{i,i+k} = y\left[x_i\right] + y\left[x_i, x_{i+1}\right](x - x_i) + \cdots + y\left[x_i, \cdots, x_{i+k}\right](x - x_i)\cdots(x - x_{i+k-1})$$

Beweis (22) Der Beweis erfolgt per Induktion über k.

Induktionsanfang $k = 0$: $p_{i,i} = y_i = y\left[x_i\right]$

Induktionsschritt Wir nehmen an, dass die Behauptung für $k - 1 \geq 0$ bereits bekannt ist. Dann definieren wir

$$q(x) := \frac{(x - x_i)p_{i+1,i+k}(x) - (x - x_{i+k})p_{i,i+k-1}(x)}{x_{i+k} - x_i}$$

$$q(x_i) = \frac{-(x_i - x_{i+k})p_{i,i+k-1}(x_i)}{x_{i+k} - x_i} = y_i$$

$$q(x_{i+k}) = \frac{(x_{i+k} - x_i)p_{i+1,i+k}(x_{i+k})}{x_{i+k} - x_i} = y_{i+k}$$

$$q(x_j) = \frac{(x_j - x_i)p_{i+1,i+k}(x_j) - (x_j - x_{i+k})p_{i,i+k-1}(x_j)}{x_{i+k} - x_i} = y_j$$

$$\Rightarrow p_{i,i+k}(x) = q(x) = p_{i,i+k-1}(x) + (x - x_i)\frac{p_{i+1,i+k}(x) - p_{i,i+k-1}(x)}{x_{i+k} - x_i}$$

$$= p_{i,i+k-1}(x) + y\left[x_i, \cdots, x_{i+k}\right](x - x_i)\cdots(x - x_{i+k-1})$$

23 **Korollar: Lösungsdarstellungen**

Die Lösung der Lagrangeschen Interpolationsaufgabe ist

1. die Newtonsche Darstellung:

$$p(x) = \sum_{i=0}^{n} y\left[x_0, \cdots, x_i\right] N_i(x)$$

2. und die Nevillesche Darstellung:

$$p(x) = p_{0,n-1}(x) + (x - x_0)\frac{p_{1,n}(x) - p_{0,n-1}(x)}{x_n - x_0}$$

Diese ist besonders für die Auswertung des Polynoms an einzelnen Stellen sinnvoll.

Berechnung einzelner Funktionswerte Ist ein Polynom in Monomdarstellung gegeben, so verwendet man das Hornerschema zur Auswertung einzelner Funktionswerte. Dies kann auf die

Newtonsche Darstellung übertragen werden wenn man die dividierten Differenzen für die Koeffizienten einmal berechnet hat:

$$p(x) = \sum_{i=0}^{n} y\left[x_0, \cdots, x_i\right] N_i(x) = \sum_{i=0}^{n} y[x_0, \cdots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$
$$= \left(\cdots (y\left[x_0, \cdots, x_n\right] (x - x_{n-1}) + y\left[x_0, \cdots, x_{n-1}\right])(x - x_{n-2}) + \cdots\right) + y\left[x_0\right]$$
$$\Rightarrow b_n = y\left[x_0, \cdots, x_n\right]$$
$$b_k = y\left[x_0, \cdots, x_k\right] + (\xi - x_k) b_{k+1}$$
$$p(\xi) = b_0$$

2.1.1 Interpolation von Funktionen

Seien die Knotenwerte y_i nun durch eine Funktion f gegeben: $y_i := f(x_i)$. Wie gut wird dann f durch das Interpolationspolynom approximiert?

Im folgenden schreiben wir $\left[x_0, \cdots, x_n\right]$ für die konvexe Hülle der x_i.

24 **Satz: Qualität der Interpolation von hinreichend glatten Funktionen**

Es sei $f \in C^{n+1}([a, b])$. Dann existiert zu jedem $x \in [a, b]$ ein $\xi_x \in \left[x_0, \cdots, x_n\right]$, sodass für den Interpolationsfehler gilt

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$$

Beweis (24) Für $x \in \left\{x_0, \cdots, x_n\right\}$ folgt die Behauptung unmittelbar, da p das Interpolationsproblem löst. Sei nun $x \in [a, b] \setminus \left\{x_0, \cdots, x_n\right\}$. Dann definieren wir

$$l(t) := \prod_{j=0}^{n} (t - x_j), \qquad c(x) := \frac{f(x) - p(x)}{l(x)}$$
$$F(t) := f(t) - p(t) - c(x) l(t)$$

Dabei hat F mindestens die $n+2$ Nullstellen $x_0, \cdots, x_n$ und x in $[a, b]$. Nach dem Satz von Rolle hat dann $F^{(n+1)}$ mindestens eine Nullstelle in $[a, b]$. Diese nennen wir ξ_x. Dann gilt

$$0 = F^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - p^{(n+1)}(\xi_x) - c(x) l^{(n+1)}(\xi_x)$$
$$= f^{(n+1)}(\xi_x) - c(x)(n+1)$$

25 **Satz: Fehlerdarstellung**
Es sei $f \in C^{n+1}([a,b])$. Dann gilt für $x \in [a,b] \setminus \{x_0, \cdots, x_n\}$:

$$f(x) - p(x) = f\left[x_0, \cdots, x_n, x\right] \prod_{j=0}^{n} (x - x_j)$$

Ferner gilt

$$f[x_0, \cdots, x_n, x] = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} f^{(n+1)}(x_0 + t_1(x_1 - x_0) + \cdots + t_n(x_n - x_{n-1}) + t(x - x_n))\mathrm{d}t\,\mathrm{d}t_n \cdots \mathrm{d}t_1\,\mathrm{d}t_0$$

Beweis (25) per vollständiger Induktion

Hermite Interpolation Es sind paarweise verschieden Stützstellen x_i mit den Knotenwerten $y_{i,k}$, $i = 0, \cdots, m$, $k = 0, \cdots, \mu_i$ ($n = m + \sum_{i=0}^{m} \mu_i$). Dann ist das Polynom n-ten Grades mit $p^{(k)}(x_i) = y_{i,k}$.

26 **Satz: eindeutige Lösbarkeit**
Die Hermitesche Interpolationsaufgabe besitzt eine eindeutige Lösung.

27 **Satz: Fehlerdarstellung**
Es sei $f \in C^{n+1}([a,b])$ und $y_{i,k} = f^{(k)}(x_i)$ sowie $p \in P_n$ die Lösung der Hermiteschen Interpolationsaufgabe. Dann existiert zu jedem $x \in [a,b]$ ein $\xi_x \in \left[x_0, \cdots, x_m, x\right]$ sodass

$$f(x) - p(x) = f[\underbrace{x_0, \cdots, x_0}_{\mu_0\text{-mal}}, \cdots, \underbrace{x_m, \cdots, x_m}_{\mu_m\text{-mal}}, x] \prod_{i=0}^{m} (x - x_i)^{\mu_i + 1}$$

28 **Definition: „Dividierte Differenzen mit gleichen Stützstellen“**
Um auch gleiche Stützstellen zuzulassen definieren wir $y[\underbrace{x_i, \cdots, x_i}_{k\text{-mal}}] = \frac{1}{(k-1)!} y_i^{(k-1)}$.

2.2 Extrapolation

Es ist eine Funktion $a(h)$ im Positiven gegeben und soll auf $h = 0$ fortgesetzt werden. Das Beispiel $a(h) = \frac{\cos(h)-1}{\sin(h)}$ zeigt uns, dass das Interpolationspolynom für Stützstellen $h_i > 0$ eine deutlich bessere Näherung für $\lim_{h\to 0} a(h) = 0$ erzeugt als die Funktionswerte an den Stützstellen selber.

Eine weiter Anwendung ist das numerische Differenzieren. Wählen wir $a(h) = \frac{f(x+h)-f(x)}{h}$ für f analytisch, so ist $f'(x) = \lim_{h\to 0} a(h)$. Mit der Taylerentwicklung erhalten wir dann $a(h) = f'(x) + \sum_{i=1}^{n} \frac{1}{(i+1)!} f^{(i+1)}(x)h^i + \frac{1}{(n+2)!} f^{(n+2)}(\xi_{x,h})h^{n+1} = a_0 + \sum_{i=1}^{n} a_i h^i + a_{n+1}(h)h^{n+1}$.

Betrachtet man auf der anderen Seite $a(h) = \frac{f(x+h)-f(x-h)}{2h} = f'(x) + \sum_{i=1}^{n} a_i(x)h^{2i} + a_{n+1}^{(n)} h^{2(n+1)}$. Dann ist $a(h)$ eine gerade Funktion, also sollte man auch gerade Polynome zur Interpolation verwenden.

29 **Satz: Extrapolationssatz**

Die Funktion $a(h)$ habe die asymptotische Entwicklung $a(h) = a_0 + \sum_{i=1}^{n} a_1 h^{iq} + a_{n+1}(h)h^{(i+1)q}$ mit $q > 0$ und $a_{n+1}(h) = a_{n+1} + o(1)$ für $h \to 0$. Ferner sei (h_k) eine monoton fallende Folge positiver Zahlen mit der Eigenschaft $0 < \frac{h_{k+1}}{h_k} \leq \rho < 1$. Für das Interpolationspolynom $p_n^{(k)} \in P_n$ durch $(h_k^q, a(h_k)), \cdots, (h_{k+n}^q, a(h_{k+n}))$ gilt dann $a(0) - p_n^{(k)}(0) = O(h_0^{(i+1)q})$ für $h \to \infty$.

Beweis (29) O.B.d.A. $k = 0$ und wir setzen $z := h^q$, $z_i = h_i^q$. Dann betrachten wir das Interpolationspolynom zu den Stützpunkten $(z_i, a(h_i))$

$$p_n(z) = \sum_{i=0}^{n} a(h_i) L_i^{(n)}(z), \quad L_i^{(n)} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{z - z_j}{z_i - z_j}$$

Aus der Fehlerdarstellung

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{j=0}^{n} (x - x_j)$$

folgt mit $f = 1$ und $f(x) = x^k$

$$\sum_{i=0}^{n} z_i^k L_i^{(n)}(0) = \begin{cases} 1, k = 0 \\ 0, k = 1, \cdots, n \\ (-1)^n \prod_{i=0}^{n} z_i, k = n+1 \end{cases}$$

Daraus folgt

$$\begin{aligned} p_k(0) &= \sum_{i=0}^{n} \left(\sum_{j=0}^{n} a_j z_i^j + a_{n+1}(h_i) z_j^{n+1} \right) L_i^{(n)}(0) \\ &= a_0 \sum_{i=0}^{n} L_i^{(n)}(0) + \sum_{j=1}^{n} a_j \left(\sum_{i=0}^{n} z_i^j L_i^{(n)}(0) \right) + a_{n+1} \sum_{i=0}^{n} z_i^{n+1} L_i^{(n)}(0) + \sum_{i=0}^{n} o(1) z_i^{n+1} L_i^{(n)}(0) \\ &= a_0 + a_{n+1}(-1)^n \prod_{i=0}^{n} h_i^q + o(h_0^{(k+1)q}) \end{aligned}$$

wegen $|L_i^{(n)}(0)| = \prod_{\substack{j=0 \\ j \neq i}}^{n} |\frac{z_j}{z_i - z_j}| = \prod_{\substack{j=0 \\ j \neq i}}^{n} |\frac{1}{\frac{z_i}{z_j} - 1}|$. Ferner ist

$$\prod_{i=0}^{n} h_i^q = O(h_0^{(n+1)q})$$

Neville-Schema zur Auswertung der Extrapolation Es sei $a_{i,k} := p_{i-k,i}$. Dann ist $a_{i,0} = a(h_i)$ und $a_{i,k} = a_{i,k-1} + \frac{a_{i,k-1} - a_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^q - 1}$.

Schrittweitenfolgen Aus Satz 29 (Extrapolationssatz) folgt für festes k:

$$a(0) - a_{i,k} = O(h_i^{(k+1)q}), i \to \infty$$

für eine Schrittweitenfolge $(h_i)_{i\in\mathbb{N}}$ mit $\limsup_{i\to\infty} \frac{h_{i+1}}{h_i} \leq \rho < 1$. Typischerweise verwendet man $h_i = \frac{h_0}{n_i}$ mit $n_i = 2^i$ oder $n_i = 2i$. $n_i = i$ funktioniert nicht.

Abbruchkriterium Nach dem Beweis von Satz 29 (Extrapolationssatz) haben wir die Fehlerdarstellung:

$$a_{i,k} = a_0 + a_{n+1}(-1)^k \prod_{j=0}^{k} h_{i-k+j}^q + o(h_{i-k}^{(n+1)q})$$

Der Fehler $a_{i,k} - a(0)$ konvergiert also monoton gegen 0 für genügend große i falls $a_{n+1} \neq 0$.

Es sei $b_{i,k} := 2a_{i+1,k} - a_{i,k}$ und $q \geq 1$. Dann ist

$$\begin{aligned} b_{i,k} - a(0) &= 2(a_{i+1,k} - a(0)) - (a_{i,k} - a(0)) \\ &= 2a_{k+1}(-1)^k \prod_{j=0}^{k} h_{i+1-k+j}^q + o(h_{i+1-k}^{(k+1)q}) - a_{k+1}(-1)^k \prod_{j=0}^{k} h_{i-k+j}^q + o(h_{i-k}^{(k+1)q}) \\ &= (-a_{k+1}(-1)^k + 2(\frac{h_{i+1}}{h_{i-k}})^q a_{k+1(-1)^k}) \prod_{j=0}^{k} h_{i-k+j}^q + o(h_{i-k}^{(k+1)q}) \end{aligned}$$

Wegen $h_i^q \ll h_{i-k}^q$ ist $b_{i,k} - a(0) \approx (-1)^{k+1} a_{k+1} \prod_{j=0}^{k} h_{i-k+j}^q$ und somit ist $a_{i,k} - a(0) \approx -(b_{i,k} - a(0))$ für festes k und hinreichend großes i. Damit gilt asymptotisch entweder $a_{i,k} \leq a(0) \leq b_{i,k}$ oder $a_{i,k} \geq a(0) \geq b_{i,k}$ und beide Seiten konvergieren monoton gegen $a(0)$. Als Abbruchbedingung eignet sich also $|a_{i,k} - b_{i,k}| < \varepsilon$.

30 **Bemerkung: Konvergenzgeschwindigkeiten im Neville-Schema**

Die Diagonalfolge $(a_{i,i})_{i\in\mathbb{N}}$ konvergiert schneller gegen $a(0)$ als jede Spaltenfolge $(a_{i,k})_{i\in\mathbb{N}}$, $k \geq 0$, falls $\sup \frac{h_{i+1}}{h_i} < 1, \inf \frac{h_{i+1}}{h_i} > 0$.

2.3 Spline-Interpolation

Das Problem bei der Lagrangeschen Interpolation ist, dass bei Vermehrung der Stützstellen zwischen diesen die Polynome stark oszillieren können. Dabei ist die C^∞-Anforderung an den Stützstellen selber meist nicht notwendig. Wir wollen dieses Problem angehen, indem wir nur eine Stückweise polynomiale Interpolation verwenden und an den Stützstellen einfache Differenzierbarkeit fordern.

Beispiel Deformation eines Biegestabs

Wir haben Stützstellen x_i mit Werten y_i gegeben. Die Aufgabe aus der Mechanik sucht nun die Krümmung mit minimaler Biegeenergie

$$E(y) = \int_{x_0}^{x^n} \frac{|y''(x)|^2}{1 + |y'(x)|^2} \mathrm{d}x$$

Bei Vernachlässigung von $|y'(x)|^2$ müssen wir also noch $\int_{x_0}^{x_n} |y''(x)|^2 \mathrm{d}x$ minimieren für die möglichen Interpolationsfunktionen hinreichender Glattheit.

Ist y nun eine Lösung dieses Minimierungsproblems, dann wählen wir eine Störfunktion φ mit $\varphi(x_i) = 0$. Dies ist dann der Fall, wenn $y \int_{x_0}^{x_n} |y''(x) + \varepsilon\varphi''(x)|^2 \mathrm{d}x$ minimiert. Dann folgt weiter

$$\begin{aligned} &\frac{d}{d\varepsilon} \int_{x_0}^{x_n} |y''(x) + \varepsilon\varphi''(x)|^2 \Big|_{\varepsilon=0} = 0 \\ &\Rightarrow \int_{x_0}^{x_n} y''(x)\varphi''(x)\mathrm{d}x = 0 \end{aligned}$$

Für $\varphi = \varphi' = 0$ auf $\left[x_0, x_n\right] \setminus \left[x_i, x_{i+1}\right]$ folgt

$$\begin{aligned} 0 &= \int_{x_i}^{x_{i+1}} y''(x)\varphi''(x)\mathrm{d}x = \int_{x_i}^{x_{i+1}} y^{(4)}\varphi(x)\mathrm{d}x \\ &\Rightarrow y^{(4)} = 0 \Rightarrow y \in P_3(x_i, x_{i,1}) \end{aligned}$$

Es lässt sich zudem zeigen, dass y in x_i zweimal stetig differenzierbar ist.

Funktionen dieser Art heißen kubische Splines. Falls $y''(x_0) = y''(x_n) = 0$ gilt, heißen sie natürliche kubische Splines.

31 **Definition: „Splinefunktion“**

Es sei $a = x_0 < \cdots < x_n = b$. Eine Funktion S heißt *Splinefunktion* vom Grad k, falls

1. $\forall i = 1, \cdots, nS|_{(x_{i-1}, x_i)} \in P_k$
2. $S \in C^{n-1}([a, b])$

Dann bezeichnen wir den Raum der Splinefunktionen mit $S_{k,n}$.

32 **Definition: „abgeschnittene Potente“**

Die Funktion $\mathrm{Pot}_n := \max(0, x - x_i)^n$ heißt *abgeschnitten Potente* bzgl. der Zerlegung $x_0 < \cdots < x_n$.

33 **Satz: Basis des Spline-Raumes**

Es sei $s \in S_{k,n}$. Dann existieren Koeffizienten $a_j, b_j \in \mathbb{R}$, sodass $s(x) = \sum_{j=0}^{k} a_j x^j + \sum_{i=1}^{n-1} b_i \mathrm{Pot}_k$ gilt. Ferner ist $\dim S_{k,n} = \dim P_k + (n-1) = n + k$.

34 **Bemerkung: Freiheitsgrade**

1. Im Fall $k = 1$ legen die Funktionswerte in den $n + 1$ Stützstellen den linearen Spline eindeutig fest.

2. Für $k > 1$ sind dagegen neben den Stützstellen noch $k - 1$ viele Bedingungen zur eindeutigen Bestimmung des Splines notwendig.

   Für $k = 2$ können wir zum Beispiel noch die Ableitung an einem Intervallende festlegen. Dass der Spline dann eindeutig bestimmt ist, ist klar. Für $k = 3$, wenn wir die zweiten Ableitungen an den beiden Intervallenden festlegen ist dies jedoch nicht mehr offensichtlich.

35 **Satz: Abschätzung des kubischen Splines**

Es sei s_n der natürliche kubische Spline (das heißt $k = 3$ und $s_n''(a) = s_n''(b) = 0$). Dann ist

$$\int_a^b |s_n''(x)|^2 \mathrm{d}x \leq \int_a^b |f'(x)|^2 \mathrm{d}x$$

für alle $f \in C^2([a,b])$ mit $s_n(x_i) = f(x_i)$, $i = 0, \cdots, n$.

Beweis (35) Wir definieren zuerst $N := \left\{ w \in C^2([a,b]) \;\middle|\; w(x_i) = 0, i = 0, \cdots, n \right\}$. Ist $f \in C^2([a,b])$ mit $s_n(x_i) = f(x_i)$. Dann ist $f = s_n + w$ für ein $w \in N$. Mit partieller Integration folgt nun

$$\int_a^b s_n'' w'' \mathrm{d}x = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s_n'' w'' \mathrm{d}x = \sum_{i=0}^{n-1} \left(s_n'' w' |_{x_i}^{x_{i+1}} - s_n''' w |_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} s_n^{(4)} w \mathrm{d}x \right)$$

$$= \sum_{i=0}^{n-1} s_n'' w' |_{x_i}^{x_{i+1}} = s_n''(b) w'(a) - s_n''(a) w'(a) = 0$$

$$\int_a^b |f''|^2 \mathrm{d}x = \int_a^b |s_n''|^2 \mathrm{d}x + 2 \int_a^b s_n' w'' \mathrm{d}x + \int_a^b |w''|^2 \mathrm{d}x \geq \int_a^b |s_n'|^2 \mathrm{d}x$$

36 **Satz: Existenz und Eindeutigkeit des natürlichen kubischen Splines**

Zu den Knotenwerten $y_0, \cdots, y_n$ existiert genau ein natürlicher kubischer Spline $s_n \in S_{3,n}$ mit $s_n(x_i) = y_i$, $i = 0, \cdots, n$ und $s_n''(a) = s_n''(b) = 0$.

Beweis (36) Wegen $\dim S_{3,n} = n + 3$ reicht es, die Eindeutigkeit nachzuweisen. Angenommen es gibt einen weiteren solchen natürlichen kubischen Spline $\tilde{s}_n$. Für $s := s_n - \tilde{s}_n$ gilt dann $s(x_1) = 0$ sowie $s''(a) = s''(b) = 0$. Für $f := 0 \in C^2([a,b])$ und $f(x_i) = s(x_i)$ folgt dann aus Satz 35 (Abschätzung des kubischen Splines):

$$\int_a^b |s''|^2 \mathrm{d}x \leq \int_a^b |f|^2 \mathrm{d}x = 0 \Rightarrow s'' = 0 \Rightarrow s \in P_1 \Rightarrow s = 0$$

2.3.1 B-Splines

37 **Definition: „B-Spline“**
Sei $a = x_0 < \cdots < x_n = b$ eine Zerlegung. Dann heißt B_i mit

$$B_i(x) = B_{ik}(x) = (x_{i+k} - x_i) \max(0, \cdot - x)^{k-1} \left[x_i, \cdots, x_{i+k}\right]$$

der i-te normierte B-Spline vom Grad $k-1$.

38 **Satz: Eigenschaften des B-Spline**

1. $B_{ik} \in S_{k-1,n}$
2. $\left\{B_{ik}\right\}_{-k+1 \le i \le n-1}$ ist eine Basis von $S_{k-1,n}$.
(bei Hinzufügen der Knoten $x_{-k+1}, \cdots, x_{-1}$ und $x_{n+1}, \cdots, x_{n+k-1}$)
3. $B_{ik}(x) = 0$ für $x \le x_i$ und $x \ge x_{i+k}$
4. $B_{ik}(x) \ge 0$
5. $\sum_{i=1}^{n} B_{ik}(x) = 1$

Um einen Spline $s \in S_{nk,n}$ mit $s = \sum_{i=-k+1}^{n-1} \lambda_i B_i$ zu bestimmen müssen wir ein lineares Gleichungssystem der Form $K\lambda = F$ lösen mit $K \in \mathbb{R}^{(n+k-1)\times(n+k-1)}$ wobei $K_{ij} := B_i(x_j)$, $F_i = y_i$ ist. (mit $k-2$ zusätzlichen Bedingungen.) Aufgrund des kleinen Trägers der B-Spline-Basisfunktionen besitzt die Matrix K Bandgestalt.

39 **Lemma: Multplikation von dividierten Differenzen**
Es seien g und h zwei Funktionen. Dann gilt für $f = gh$

$$f[x_0, \cdots, x_k] = \sum_{i=0}^{k} g[x_0, \cdots, x_i] h[x_i, \cdots, x_k]$$

40 **Satz: rekursive Berechnung des B-Spline**
Es sei $N_{i,k} = \frac{B_{ik}}{x_{i+k} - x_i}$. Für $k \ge 2$ gilt

$$N_{ik}(x) = w_i(x) N_{i,k-1}(x) + \tilde{w}_i(x) N_{i+1,k-1}(x)$$

mit den Gewichtsfunktionen

$$w_i(x) = \frac{x - x_i}{x_{i+k} - x_i}, \quad \tilde{w}_i(x) = \frac{x_{i+k} - x}{x_{i+k} - x_i}$$

Beweis (40) Es ist $(\cdot - x)\left[x_i\right] = (x_i - x)$, $(\cdot - x)\left[x_i, x_{i+1}\right] = 1$ und $(\cdot - x)\left[x_i, x_{i+1}, x_{i+2}\right] = 0$. Wenden wir nun das obige Lemma auf $\max(0, t-x)^{k-1} = (t-x) \cdot \max(0, t-x)^{k-2}$ an, so liefert

es:

$$\max(\cdot - x)^{k-1}\left[x_i, \cdots, x_{i+k-1}\right] = \sum_{j=0}^{k-1} (\cdot - x)\left[x_i, \cdots, x_{i+j}\right] \max(0, \cdot - x)^{k-2}\left[x_{i+j,\cdots,x_{i+k-1}}\right]$$

$$= (x_i - x)\max(0, \cdot - x)^{k-2}\left[x_i, \cdots, x_{i+k-1}\right] + \max(0, \cdot - x)^{k-2}\left[x_{i+1}, \cdots, x_{i+k-1}\right]$$

Analog folgt

$$\max(0, \cdot - x)^{k-1}\left[x_{i+1}, \cdots, x_{i+k}\right] = \max(0, \cdot - x)^{k-1}\left[x_{i+k,\cdots,x_{i+1}}\right]$$

$$= (x_{i+k} - x)\max(0, \cdot - x)^{k-2}\left[x_{i+1}, \cdots, x_{i+k}\right] + \max(0, \cdot - x)^{k-2}\left[x_{i+1,\cdots,x_{i+k-1}}\right]$$

$$\max(0, \cdot - x)^{k-1}\left[x_{i+1}, \cdots, x_{i+k}\right] - \max(0, \cdot - x)^{k-1}\left[x_i, \cdots, x_{i+k-1}\right]$$

$$= (x_{i+1} - x)\max(0, \cdot - x)^{k-2}\left[x_{i+1,\cdots,x_{i+k}}\right] - (x_i - x)\max(0, \cdot - x)^{k-2}\left[x_{i,\cdots,x_{i+k-1}}\right]$$

$$B_{ik}(x) = (x_{i+1} - x)\frac{B_{i+1,k-1}(x)}{x_{i+k} - x_i} - (x_i - x)\frac{B_{i,k-1}(x)}{x_{i+k-1} - x_i} = (x_{i+1} - x)N_{i+1,k-1}(x) - (x_i - x)N_{i,k-1}(x)$$

41 **Bemerkung: numerische Güte der rekursiven Formel**

Es ist $w_i \geq 0$ und $\tilde{w}_i \geq 0$ sowie $N_{i,k} \geq 0$ in $[x_i, x_{i+k}]$. Das heißt es tritt keine Stellenauslöschung auf.

42 **Satz: Interpolationsfehler**

Es sei $f \in C^4([a,b])$ und $s \in S_{3,n}$ mit $s(x_i) = f(x_i), i = 0, \cdots, n$ und $s''(a) = f''(a)$ sowie $s''(b) = f''(b)$. Dann gilt $h = \max h_i$ und $h_i := x_{i+1} - x_i$:

$$\max_{a \leq x \leq b} |f(x) - s(x)| \leq \frac{3}{8} h^4 \max_{a \leq x \leq b} |f^{(4)}(x)|$$

Beweis (42) (siehe SW92, S. 159 f.)

2.4 Trigonometrische Interpolation

43 **Definition: „periodische Funktionen“**

Eine Funktion $f: \mathbb{R} \to \mathbb{R}$ heißt *periodisch* mit der *Periode* $\omega \in \mathbb{R}$, falls gilt

$$\forall x \in \mathbb{R}\, f(x + \omega) = f(x)$$

44 **Bemerkung: trigonometrische Funktionen**

1. Für periodische Funktionen bietet sich offenbar eine Interpolation mit trigonometrischen Funktionen an:

$$t_{2m}(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \left(a_k \cos(k\frac{2\pi x}{\omega}) + b_k \sin(k\frac{2\pi x}{\omega}) \right)$$

Mit $a_k, b_k \in \mathbb{R}$.

2. Im Folgenden wählen wir o.B.d.A. $\omega = 2\pi$.

Interpolationsaufgabe Wir betrachten das Interpolationsintervall $[0, 2\pi]$, die Stützstellen gegeben durch die äquidistante Zerlegung $x_j = j\frac{2\pi}{n+1}$, $j = 0, \cdots, n$ mit den Stützwerten $y_0, \cdots, y_n \in \mathbb{C}$. Gesucht ist eine Funktion $t_n\colon [0, 2\pi] \to \mathbb{C}$, so dass $t_n(x_j) = y_j$ für $j = 0, \cdots, n$ gilt.

45 **Satz: Lösung der Interpolationsaufgabe**

Die Interpolationsaufgabe ist mit

$$t_n(x) = \sum_{k=0}^{n} c_k e^{ikx}$$

gelöst. Die Koeffizienten sind bestimmt durch

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}$$

Beweis (45) Es sei $w = e^{ix}$, $w_k = e^{ix_k} = e^{\frac{2k\pi i}{n+1}}$. Insbesondere sind die w_k paarweise verschieden. Dann ist die trigonometrische Interpolation isomorph zur polynomialen Interpolationsaufgabe bei der wir $\tilde{t}_n \in P_n$ suchen mit $\tilde{t}_n(w) = \sum_{k=0}^{n} c_k w^k$ und $\tilde{t}_n(w_k) = y_k$. Dafür haben wir die eindeutige Lösbarkeit aber bereits gezeigt.

Nun müssen wir noch die Koeffizienten c_j berechnen:

$$\sum_{k=0}^{n} y_k w_k^{-j} = \sum_{k=0}^{n} \tilde{t}_k(w_k) w_k^{-j} = \sum_{k=0}^{n} w_k^{-j} \sum_{l=0}^{n} c_l w_k^l = \sum_{l=0}^{n} c_l \sum_{k=0}^{n} w_k^{l-j}$$

Die w_k sind Wurzeln von $w^{n+1} - 1 = (w-1)(w^n + w^{n-1} + \cdots + 1)$. Dann gilt für $k = \pm 1, \cdots, \pm n$

$$w_k \neq 1 \Rightarrow \sum_{j=0}^{n} w_k^j = 0$$

$$\sum_{k=0}^{n} w_k^{l-j} = \sum_{k=0}^{n} \left(e^{\frac{2k\pi i}{n+1}}\right)^{l-j} = \sum_{k=0}^{n} w_{l-j}^k = (n+1)\delta_{l,j}$$

$$c_j = \frac{1}{n+1}\left(\sum_{k=0}^{n} y_k w_k^{-j}\right) = \frac{1}{n+1} \sum_{k=0}^{n} y_k e^{\frac{-2\pi jk}{n+1}}$$

46 **Satz: diskrete Fourier-Analyse**

Die Stützwerte $y_0, \cdots, y_n$ seien reell, dann existiert genau ein „trigonometrisches Polynom“ der Form

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \big(a_k \cos(kx) + b_k \sin(kx)\big) + \frac{\theta}{2} a_{m+1} \cos((m+1)x)$$

das die Interpolationsaufgabe löst. Hierbei ist $\theta = 0$ und $m = \frac{n}{2}$ falls n gerade ist sowie $\theta = 1$ und $m = \frac{n-1}{2}$ falls n ungerade ist. Die Koeffizienten sind bestimmt durch

$$a_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \cos(jx_k)$$

$$b_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \sin(jx_k)$$

Beweis (46) Wir wissen, dass $t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}$ mit $c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}$ die Interpolationsaufgabe löst. Da e^{-ix} periodisch ist wissen wir

$$e^{-ijx_{n+1-k}} = e^{-ij\frac{2(n+1-k)\pi}{n+1}} = e^{ij2\pi + ijx_k} = e^{ijx_k}$$

Daraus folgt

$$c_{n+1-k} = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_{n+1-k}} = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{ijx_k} =: c_{-k}$$

Dann setzen wir $a_k := c_k + c_{-k}$ und $b_k := i(c_k - c_{-k})$ für $k = 1, \cdots, m$ sowie $a_0 = 2c_0$, $a_{m+1} = 2c_{m+1}$, falls $n = 2m + 1$ ungerade ist. Damit können wir zeigen, dass $t_n(x_j) = y_j$ für $j = 0, \cdots, n$ ist.

$$t_n(x_j) = c_0 + \sum_{k=1}^{m} \left((c_k + c_{-k}) \cos(kx_j) + i(c_k - c_{-k}) \sin(kx_j) + \theta c_{m+1} \cos((m+1)x_j) \right)$$

$$= c_0 + \sum_{k=1}^{m} c_k \left(\cos(kx_j) + i \sin(kx_j) \right) + \sum_{k=1}^{m} c_{-k} \left(\cos(kx_j) - i \sin(kx_j) \right) + \theta c_{m+1}((m+1)x_j)$$

Wegen $(m+1)x_j = \left(\frac{n-1}{2} + 1 \right) \frac{2j\pi}{n+1} = j\pi$ ist $\sin((m+1)x_j) = 0$. Damit gilt weiter

$$t_n(x_j) = c_0 + \sum_{k=1}^{m} c_k e^{ikx_j} + \sum_{k=1}^{m} c_{-k} e^{-ikx_j} + \theta c_{m+1} e^{i(m+1)x_j}$$

Es ist $\sum_{k=1}^{m} c_{-k} e^{-ikx_j} = \sum_{k=1}^{m} c_{n+1-k} e^{i(n+1-k)x_j}$, also

$$t_n(x_j) = \sum_{k=0}^{n} c_k e^{ikx_j} = y_j$$

Es bleibt noch zu zeigen, dass die Koeffizienten übereinstimmen:

$$a_k = c_k + c_{-k} = \frac{1}{n+1} \sum_{j=0}^{n} y_j (e^{-ijx_k} + e^{ijx_k}) = \frac{2}{n+1} \sum_{j=0}^{n} y_j \cos(jx_k)$$

$$b_k = i(c_k - c_{-k}) = \frac{i}{n+1} \sum_{j=0}^{n} y_j (e^{-ijx_k} - e^{ijx_k}) = \frac{2}{n+1} \sum_{j=0}^{n} y_j \sin(jx_k)$$

□

47 **Bemerkung:**

1. Eine derartige Interpolation mit trigonometrischen Funktionen heißt *trigonometrische* Interpolation.
2. Für eine 2π-periodische Funktion f sind die Stützwerte durch $y_j = f(x_j)$ gegeben.
3. Die trig. Interpolation wie wir sie oben durchgeführt haben wird *diskrete Fourier-Analyse* genannt. Die Abbildung $(y_j)_{j=0,\cdots,n} \mapsto (a_k, b_k)_{k=0,\cdots,m+1}$ heißt *diskrete Fourier Transformation*.

Effiziente Berechnung des trigonometrischen Interpolationspolynom Im Beweis der diskreten Fourier Analyse haben wir die Koeffizienten bestimmt über

$$a_k = c_k + c_{-k}, \quad b_k = i(c_k - c_{-k})$$
$$c_k = \sum_{j=0}^{n} y_i w^{jk}, \quad w = e^{-i\frac{2\pi}{n+1}}, \quad w^{n+1} = 1$$

Bestimmen wir die c_k über das Horner-Schema, so brauchen wir n Operationen (wobei eine Operation 1 komplexe Multiplikation und 1 komplexe Addition enthält), insgesamt also $n^2 + n$ komplexe Operationen. Dieser Aufwand ist zu hoch.

schnelle Fourier Transformation (FFT = Fast Fourier Transformation) Wir betrachten zuerst den Spezialfall $n = 2^p - 1$, $p \in \mathbb{N}$. Dann ist

$$c_k = \sum_{j=0}^{n} y_j w^{kj} = \sum_{j=0}^{\frac{n-1}{2}} y_{2j} (w^2)^{kj} + \sum_{j=0}^{\frac{n-1}{2}} y_{2j+1} (w^2)^{kj} w^k$$

Sei k_1 der ganzzahlige Rest bei Division von k durch $\frac{n+1}{2}$, das heißt: $k = N\frac{n+1}{2} + k_1$. Dann gilt

$$(w^2)^{kj} = (w^2)^{N\frac{n+1}{2}} (w^2)^{k_1 j} = (w^{n+1})^{Nj} (w^2)^{k_1 j} = (w^2)^{k_1 j}$$
$$\Rightarrow c_k = \sum_{j=0}^{\frac{n-1}{2}} y_{2j} (w^2)^{k_1 j} + w^k \sum_{j=0}^{\frac{n-1}{2}} y_{2j+1} (w^2)^{k_1 j} =: c'_{k_1} + w^k c''_{k_1}$$

Zur Bestimmung der c'_{k_1} und c''_{k_1} kann wieder eine Aufspaltung durchgeführt werden. Wir ersetzen also die Bestimmung der c_k mit $n + 1 = 2^p$ Termen durch 2 Bestimmungen c'_{k_1} und c''_{k_1} mit $\frac{n+1}{2} = 2^{p-1}$ Termen und so weiter, bis wir 2^p Bestimmungen mit je einem Term haben.

48 **Satz: Komplexität der FFT**

Im Fall $n = 2^p - 1$ bestimmt die FFT die Koeffizienten c_k mit $2(n + 1)\,\mathrm{ld}\,(n + 1)$ komplexen Operationen.

49 **Bemerkung: Verallgemeinerung**

Der Algorithmus lässt sich für beliebige $n \in \mathbb{N}$ verallgemeinern. Die FFT benötigt dann höchstens $(n+1)(q_1 + \cdots + q_p)$ Operationen, wenn $n+1 = q_1 \cdots q_p$ die Primfaktorzerlegung von $n+1$ ist.

2.5 Allgemeine Interpolationsaufgabe

50 **Definition: „Vandermonde-Matrix“**

Es sei M eine Menge, $u = (u_1, \cdots, u_m)\colon M \to \mathbb{K}^m$ für $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ und $z = (z_1, \cdots, z_n) \in M^n$, dann heißt

$$V(z,u) = \begin{pmatrix} u_1(z_1) & \cdots & u_1(z_n) \\ \vdots & \ddots & \vdots \\ u_m(z_1) & \cdots & u_m(z_n) \end{pmatrix}$$

Vandermonde-Matrix zu u und z. Für $m = n$ heißt $|V(u,z)|$ *Vandermonde-Determinante*.

Beispiel Es seien $p_j\colon \mathbb{K} \to \mathbb{K}$ mit $p_j(x) := x^j$ und $z_0, \cdots, z_n \in \mathbb{K}$ paarweise verschiedene Zahlen. Dann erhalten wir die Matrix

$$V(z,p) = \begin{pmatrix} 1 & \cdots & 1 \\ z_0 & \cdots & z_n \\ \vdots & \ddots & \vdots \\ z_0^n & \cdots & z_n^n \end{pmatrix}$$

$$|V(z,p)| = \prod_{0 \le j < k \le n} (z_k - z_j)$$

51 **Satz: Čebyšëv-Bedingungen**

Es sei M eine Menge mit Kardinalität größer gleich n und $u_1, \cdots, u_n\colon M \to \mathbb{K}$ Funktionen. Dann sind die folgenden Aussagen äquivalent:

1. Für jede n-elementige Teilmenge $\{z_1, \cdots, z_n\} \subset M$ ist $|V(z,u)| \neq 0$.
2. Jede nicht-triviale Linearkombination von $u_1, \cdots, u_n$ hat höchstens $n-1$ Nullstellen in M.
3. Für jede n-elementige Teilmenge $\{z_1, \cdots, z_n\} \subset M$ sind $u_1|_z, \cdots, u_n|_z$ linear unabhängig.

52 **Definition: „Čebyšëv-System“**

Ein System von Funktionen $u_1, \cdots, u_n$ das diese Aussagen erfüllt heißt $\mathbb{K}$-*Čebyšëv-System* (kurz $\mathbb{K}$-ČS)

Beispiel für $\mathbb{K}$-ČSe

- Die Monome bilden ein $\mathbb{K}$-ČS.
- Die trigonometrischen Funktionen $\sin_j, \cos_j :]-\pi, \pi] \to \mathbb{R}$ mit $\sin_j(x) := \sin(jx)$ und $\cos_j(x) := \cos(jx)$ bilden als $(1, \cos_1, \sin_1, \cos_2, \sin_2, \cdots, \cos_n, \sin_n)$ ein $\mathbb{R}$-ČS.
- Die Rationalen Funktionen $r_b(z) := \frac{1}{z-b}$ bilden auf ihrem gemeinsamen Definitionsbereich $\mathbb{C} \setminus \{b_1, \cdots, b_n\}$ ein $\mathbb{C}$-ČS.
- Die Exponentialpolynome $exp_b := e^{bx}$ bilden $(p_0 exp_{b_1}, \cdots, p_m exp_{b_1}, \cdots, p_0 exp_{b_n}, \cdots, p_m exp_{b_n})$ ein $\mathbb{R}$-ČS.

53 **Definition: „allgemeine Interpolationsaufgabe“**

Es sei $u_1, \cdots, u_n$ ein $\mathbb{K}$-ČS auf M und $z_1, \cdots, z_n \in M$ paarweise verschieden sowie $y_1, \cdots, y_n \in \mathbb{K}$. Dann heißt $u \in span\{u_1, \cdots, u_n\}$ Lösung der Interpolationsaufgabe zu $(u_1, \cdots, u_n)$, Knoten $(z_1, \cdots, z_n)$ und Knotenwerten $(y_1, \cdots, y_n)$, wenn $\forall j = 1, \cdots, nu(z_j) = y_j$.

54 **Satz: eindeutige Lösbarkeit**

Die allgemeine Interpolationsaufgabe ist eindeutig lösbar durch $u = x_1 u_1 + \cdots + x_n u_n$ wobei

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \left(V(z,u)^T\right)^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

die Koeffizienten liefert.

Beweis (54)

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = V(z,u)^T \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} u_1(z_1) & \cdots & u_n(z_1) \\ \vdots & \ddots & \vdots \\ u_1(z_n) & \cdots & u_n(z_n) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} u(z_1) \\ \vdots \\ u(z_n) \end{pmatrix}$$

3 Numerische Integration

Ziel ist die näherungsweise Bestimmung von bestimmten Integralen mit Hilfe von sogenannten Quadraturformeln. Es sei $f \in C([a,b])$. Dann verwenden wir

$$I(f) := \int_a^b f(x)\mathrm{d}x \approx I_n(f) := \sum_{i=0}^{n} \alpha_i f(x_i)$$

mit Stützstellen $a = x_0 < \cdots < x_n = b$ und passenden Gewichten $\alpha_i \in \mathbb{R}$.

Beispiel: Rechteckregel Wir approximieren die Funktion f durch Rechtecke (vergleiche Riemannsche Zwischensummen). Dann erhalten wir

$$I_n(f) = \sum_{i=0}^{n} (x_{i+1} - x_i) f(x_i)$$

3.1 Interpolatorische Quadraturformeln

Idee Wir konstruieren die Quadraturformeln über Polynomfunktionen. Das heißt

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

55 **Bemerkung: Abhängigkeit**
Die Gewichte hängen nur von $[a, b]$ und den Stützstellen $x_0, \cdots, x_n$ ab.

56 **Satz: Fehler der interpolatorischen Quadraturformel**
Der Fehler der interpolatorischen Quadraturformeln ist

$$I(f) - I(f_n) = \int_a^b f\left[x_0, \cdots, x_n, x\right] \prod_{j=0}^{n} (x - x_j) \mathrm{d}x$$

Beweis (56) Ergibt sich unmittelbar aus der Restglieddarstellung bei der Polynominterpolation.

57 **Definition: „exakte Quadraturformeln“**
Eine Quadraturformel I_n heißt *(exakt) von der Ordnung n*, falls $I(p) = I_n(p)$ für alle Polynome p vom Grad $n - 1$ gilt.

58 **Satz: minimale Exaktheit der interpolatorischen Quadraturformeln**
Interpolatorische Quadraturformeln sind bei beliebiger Wahl der $n+1$ Stützstellen mindestens von der Ordnung $n + 1$.

3.1.1 Newton-Cotes-Quadraturformeln

Die Konstruktion erfolgt mit äquidistant verteilten Stützstellen.

- abgeschlossene Newton-Cotes-Formeln: $x_i := a + ih$, $i = 0, \cdots, n$, $h := \frac{b-a}{n}$ (das heißt a und b sind ebenfalls Stützstellen)
- offene Newton-Cotes-Formeln: $x_i := a + (i + 1)h$, $i = 0, \cdots, n$, $h := \frac{b-a}{n+2}$ (das heißt a und b sind keine Stützstellen)

Betrachten wir die abgeschlossenen Newton-Cotes-Formeln und wenden die Koordinatentransformation $x \mapsto t := \frac{x-a}{h}$ an:

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{a + th - a - jh}{a + ih - a - jh} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - j}{i - j}$$

$$\Rightarrow \alpha_i = \int_a^b L_i^{(n)} \mathrm{d}x = h \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - i}{i - j} dt$$

Beispiel: $n = 2$

$$\alpha_0 = h \int_0^n \frac{t-1}{0-1} \cdot \frac{t-2}{0-2} = \frac{1}{3}h$$
$$\alpha_1 = \frac{4}{3}h, \alpha_2 = \frac{1}{3}h$$
$$\Rightarrow I_2(t) = \frac{h}{3}(f(a) + 4f(\frac{a+b}{2}) + f(b))$$

einige Newton-Cotes-Formeln mit Restglied

- Trapezregel: $I(f) = \frac{b-a}{2}(f(a) + f(b)) - \frac{(b-a)^3}{12} f''(\zeta)$
- Simpsonregel: $I(f) = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b)) - \frac{(b-a)^5}{2880} f^{(4)}(\zeta)$
- $\frac{3}{8}$-Regel: $I(f) = \frac{b-a}{8}(f(a) + 3f(a+h) + 3f(b-h) + f(b)) - \frac{(b-a)^5}{6480} f^{(4)}(\zeta)$ mit $h := \frac{b-a}{3}$
- Mittelpunktregel: $I(f) = (b-a)f(\frac{a+b}{2}) + \frac{(b-a)^3}{24} f''(\zeta)$
- (ohne Bezeichnung): $I(f) = \frac{b-a}{2}(f(a+h) + f(b-h)) + \frac{(b-a)^3}{108} f''(\zeta)$

Einschließung des Integralwertes Es habe $f^{(k)}$ konstantes Vorzeichen auf $[a, b]$. Z.B. $k = 2$ und f konvex / $f'' \geq 0$, so erhalten wir mit Mittelpunkt- und Trapezregel:

$$(b-a)f(\frac{a+b}{2}) \leq I(f) \leq \frac{b-a}{2}(f(a) + f(b))$$

Nachweis der Restglieddarstellung

Trapezregel Wegen $(x-a)(x-b) \leq 0$ in $[a,b]$ gilt nach der bekannten Darstellung des Fehlers der Lagrange-Interpolation:

$$I(f)-I_1(f) = \int_a^b \frac{f''(\zeta_x)}{2}(x-a)(x-b)\mathrm{d}x \; = f''(\zeta)\int_a^b (x-a)(x-b)\mathrm{d}x \; = \frac{-(b-a)^3}{12}f''(\zeta)$$

Simpson-Regel

$$\begin{aligned} I(f) - I_2(f) &= \int_a^b f\left[a, \frac{a+b}{2}, b, x\right](x-a)(x-\frac{a+b}{2})(x-b)\mathrm{d}x \\ &= \int_a^b \frac{f\left[a,\frac{a+b}{2},b,x\right] - f\left[a,\frac{a+b}{2},b,\frac{a+b}{2}\right]}{x-\frac{a+b}{2}}(x-a)(x-\frac{a+b}{2})^2(x-b)\mathrm{d}x \\ &+ f\left[a,\frac{a+b}{2},b,\frac{a+b}{2}\right]\underbrace{\int_a^b (x-a)(x-\frac{a+b}{2})(x-b)\mathrm{d}x}_{=0} \end{aligned}$$

Dabei ist

$$\frac{f\left[a,\frac{a+b}{2},b,x\right] - f\left[a,\frac{a+b}{2},b,\frac{a+b}{2}\right]}{x-\frac{a+b}{2}} = f\left[a,\frac{a+b}{2},b,\frac{a+b}{2},x\right] = \frac{f^{(4)}(\zeta_x)}{4!}$$

und somit können wir den Fehler weiter umformen:

$$\begin{aligned} I(f) - I_2(f) &= \int_a^b \frac{f^{(4)}(\zeta_x)}{4!}(x-a)(x-\frac{a+b}{2})^2(x-b)\mathrm{d}x \\ &= f^{(4)}(\zeta)\int_a^b (x-a)(x-\frac{a+b}{2})^2(x-b)\mathrm{d}x \; = -\frac{(b-a)^5}{2880}f^{(4)}(\zeta) \end{aligned}$$

Besselsche Formeln erhalten wir indem wir Stützstellen außerhalb von $[a,b]$ wählen:

$$I(f) = \frac{b-a}{24}(-f(2a-b) + Bf(a) + Bf(b) - f(2b-a)) + \frac{11}{720}(b-a)^5 f^{(4)}(\zeta)$$

Hermitesche Formeln Beachten wir auch noch die Ableitungswerte, so erhalten wir Beispielsweise:

$$I(f) = \frac{b-a}{2}(f(a)+f(b)) + \frac{(b-a)^2}{12}(f'(a)-f'(b)) + \frac{(b-a)^5}{720}f^{(4)}(\zeta)$$

59 **Bemerkung: Auslöschung, Konvergenz**

1. Bei abgeschlossenen Newton-Cotes-Formeln treten ab $n = 7$ und bei offenen bereits ab $n = 2$ Stützstellen negative Gewichte auf, sodass es zu Rundungsfehlern und Stellenauslöschungen kommen kann.

2. Es ist im allgemeinen keine Konvergenz der Form $I_n(f) \xrightarrow{n\to\infty} I(f)$ zu erwarten.

3.1.2 Summierte Quadraturformeln

Um dem Problem der immer stärkeren Oszillation zu entgehen, erhöhen wir nicht immer weiter den Grad des Interpolationspolynoms, sondern zerteilen wir das Integrationsintervall $[a, b]$ durch $a = x_0 < \cdots < x_N = b$ und wenden Quadraturformeln $I^{(n)}_{[x_i,x_{i+1}]}$ an:

$$I^{\Sigma}_n(f) := \sum_{i=0}^{N-1} I^{(n)}_{[x_i,x_{i+1}]}(f)$$

60 **Satz: Gesamtfehler**

Es sei der Fehler auf den einzelnen Teilstücken $I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f) = w_n h^{m+2} f^{(m+1)}(\xi_i)$, $h = \frac{b-a}{4}$ mit $m \geq n$ und $\xi \in [x_i, x_{i+1}]$. (f entsprechend oft differenzierbar) Dann ist der gesamte Fehler:

$$I(f) - I^{\Sigma}_n(f) = w_n(b-a)h^{m+1} f^{(m+1)}(\zeta), \quad \zeta \in [a, b]$$

Beweis (60)

$$g(x) := w_n \frac{b-a}{N} h^{m+1} f^{(m+1)}(x)$$
$$g(\zeta_{\min}) = \min_i g(\xi_i), \quad g(\zeta_{\max}) = \max_i g(\xi_i)$$
$$\Rightarrow I(f) - I^{\Sigma}_n(f) = \sum_{i=0}^{N-1} w_n h^{m+2} f^{(m+1)}(\xi_i) \geq N g(\zeta_{\min})$$

Nach dem Zwischenwertsatz existiert dann ein $\zeta \in [a, b]$ mit $I(f) - I^{\Sigma}_n(f) = Ng(\zeta) = w_n(b - a)h^{m+1} f^{(m+1)}(\zeta)$

Beispiele

1. *Summierte Trapezregel*

$$I^{\Sigma}_1(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2}(f(x_i) + f(x_{i+1})) = \frac{h}{2}(f(a) + 2\sum_{i=1}^{N-1} f(x_i) + f(b))$$
$$I(f) - I^{\Sigma}_1(f) = -\frac{b-a}{12} h^2 f''(\zeta), \quad \zeta \in [a, b]$$

2. *Summierte Simpson-Regel*

$$I_2^{\Sigma}(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{6}(f(x_i) + 4f(\frac{x_i + x_{i+1}}{2} + f(x_{i+1}))$$

$$= \frac{h}{6}(f(a) + 2\sum_{i=1}^{N-1} f(x_i) + 4\sum_{i=0}^{N-1} f(\frac{x_i + x_{i+1}}{2}) + f(b))$$

$$I(f) - I_2^{\Sigma}(f) = -\frac{b-a}{2880}h^4 f^{(4)}(\zeta), \quad \zeta \in [a, b]$$

3. *Summierte Mittelpunktregel*

$$I_0^{\Sigma}(f) = \sum_{i=0}^{N-1} (x_{i+1} - x_i) f(\frac{x_i + x_{i+1}}{2}) = h\sum_{i=0}^{N-1} f(\frac{x_i + x_{i+1}}{2})$$

3.2 Gauß-Quadraturformeln

Gesucht sind Stützstellen $x_0, \cdots, x_n$ und Gewichte $\alpha_0, \cdots, \alpha_n$ sodass die Ordnung der Quadraturformel möglichst hoch ist. Das heißt, dass Polynome möglichst hohen grades exakt integriert werden können.

61 **Satz: maximale Exaktheit**

Die Ordnung einer Quadraturformel I_n ist höchstens $2n + 2$.

Beweis (61) Angenommen I_n ist von höherer Ordnung als $2n+2$. Dann gilt für $p(x) = \prod_{i=0}^{n} (x - x_i)^2 \in P_{2n+2}$:

$$0 < \int_a^b p(x)\mathrm{d}x \; = I_n(p) = \sum_{i=0}^{n} \alpha_i p(x_i) = 0 \quad \lightning$$

62 **Satz: hinreichendes Kriterium für maximale Exaktheit**

Die interpolatorische Quadraturformel I_n mit Stützstellen $x_0, \cdots, x_n$ ist von der Ordnung $2n+2$, falls

$$\forall q \in P_n \int_a^b \prod_{j=0}^{n} (x - x_j) q(x)\mathrm{d}x \; = 0$$

Beweis (62) Es sei $f \in P_{2n+1}$ und I_{2n+1} eine interpolatorische Quadraturformel mit den Stützstellen $x_0, \cdots, x_{2n+1}$. Nach (*Satz* 58 (*minimale Exaktheit der interpolatorischen Quadratur formeln*))

ist dann in der Newtonschen Darstellung:

$$0 = I(f) - I_{2n+1}(f) = I(f) - \sum_{i=0}^{2n+1} f[x_0, \cdots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j) \mathrm{d}x$$
$$= I(f) - \sum_{i=0}^{n} f[x_0, \cdots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j) - \sum_{i=n+1}^{2n+1} f[x_0, \cdots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j)$$
$$= I(f) - I_n(f) - \sum_{i=n+1}^{2n+1} f[x_0, \cdots, x_i] \int_a^b \prod_{j=0}^{i-1} (x - x_j)$$

Es bleibt also noch zu zeigen, dass der letzte Summand 0 wird. Für $i \geq n + 1$ gilt:

$$\int_a^b \prod_{j=0}^{i-1} (x - x_j) \mathrm{d}x = \int_a^b \prod_{j=0}^{n} (x - x_j) \underbrace{\prod_{j=n+1}^{i-1} (x - x_j)}_{\in P_n} \mathrm{d}x = 0$$

63 **Bemerkung: Deutung der hinreichenden Bedingung**

Die obige hinreichende Voraussetzung besagt gerade, dass das Polynom $p(x) = \prod_{j=0}^{n} (x - x_j) = x^{n+1} + r(x), r \in P_n$ orthogonal zu $P_n \subset C([a, b])$ bezüglich des Skalarprodukts $(f, g) = \int_a^b f(x)g(x)\mathrm{d}x$ ist.

64 **Satz: Schmidtsches Orthogonalisierungsverfahren**

Definiert man

$$p_0(x) = 1, \tilde{p}_0(x) = \|1\|^{-1} p_0(x)$$
$$p_k(x) = x^k - \sum_{j=0}^{k-1} \langle x^k, \tilde{p}_j \rangle \tilde{p}_j(x), \quad \tilde{p}_k(x) = \|p_k\|^{-1} p_k(x)$$

Dann bilden $\{p_0, \cdots, p_{n+1}\}$ ein Orthogonalsystem und $\{\tilde{p}_0, \cdots, \tilde{p}_{n+1}\}$ ein Orthonormalsystem in P_{n+1}.

65 **Bemerkung: Stützstellen**

Setzt man $p(x) = p_{n+1}(x)$ ergeben sich die gesuchten Stützstellen als Nullstellen von p_{n+1}. (sofern diese reell sind)

66 **Satz: Eigenschaften der Nullstellen**

Die Nullstellen der orthogonalen Polynome sind reell, einfach und liegen alle in $[a, b]$.

Beweis (66) Sei N_n die Menge der Nullstellen von p_n ungerader Vielfachheit und $q(x) := 1$ für $N_n = \varnothing$ sowie $q(x) = \prod_{i=1}^{m}(x - \lambda_i)$ für $N_n = \{\lambda_1, \cdots, \lambda_m\}$. Dann ist das Polynom $p_n q \in P_{n+m}$ reell, hat nur Nullstellen gerader Vielfachheit, also keinen Vorzeichenwechsel.

Angenommen $m < n$. Dann gilt $\langle p_n, q\rangle = 0$ wegen $p_n \perp P_{n-1}$. Jedoch ist $\langle p_n, q\rangle = \int_a^b p_n q \mathrm{d}x \neq 0$. Also ist $m \geq n$.

67 **Definition: „Legendre Polynome“**

Die orthogonale Polynome heißen Legendre Polynome auf $[a, b]$ und werden auf dem Referenzintervall $[-1, 1]$ in der Regel mit L_n bezeichnet.

68 **Bemerkung: Eigenschaften der Legendre Polynome**

Die Legendre-Polynome mit Leitkoeffizienten 1 sind eindeutig bestimmt und es gilt:

1. $L_n(x) = \frac{n!}{(2n!)}\frac{d^n}{dx^n}(x^2 - 1)^n$ (explizite Darstellung)
2. $L_0(x) = 1$, $L_1(x) = x$, $L_{n+1}(x) = xL_n(x) - \frac{n^2}{4n^2-1}L_{n-1}(x)$ (rekursive Darstellung)

69 **Definition: „Gaußsche Quadraturformel“**

Die interpolatorische Quadraturformel auf dem Intervall $[-1, 1]$ mit den Nullstellen des $(n+1)$-ten Legrende Polynoms L_{n+1} als Stützstellen heißt *Gaußsche Quadraturformel*.

70 **Bemerkung: Bestimmung der Stützstellen**

Die Stützstellen (das heißt die Nullstellen von L_{n+1}) werden numerisch bestimmt und tabelliert. Es ergibt sich jedoch noch exakt:

$$n = 1\colon \lambda_0 = -\sqrt{\frac{1}{3}}, \lambda_1 = \sqrt{\frac{1}{3}}$$
$$n = 2\colon \lambda_0 = -\sqrt{\frac{3}{5}}, \lambda_1 = 0, \lambda_2 = \sqrt{\frac{3}{5}}$$

Die Gewichte der Quadraturformel können folgendermaßen bestimmt werden:

$$\alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - \lambda_j}{\lambda_i - \lambda_j}\mathrm{d}x = \frac{1}{L'_{n+1}(x_i)L_n(x_i)} \cdot \frac{(n!)^4 2^{n+1}}{((2n)!)^2(2n+1)}$$

71 **Satz: Eindeutigkeit**

Es existiert genau eine interpolatorische Quadraturformel auf $[-1, 1]$ mit der Ordnung $2n + 2$. Diese ist die Gaußsche Quadraturformel. Für die Gewichte α_i gilt

$$\alpha_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - \lambda_j}{\lambda_i - \lambda_j}\mathrm{d}x > 0$$

Sofern $f \in C^{2n+2}([-1,1])$ ist, gilt für das Restglied

$$R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{j=0}^{n} (x - x_j)^2 \mathrm{d}x$$

Beweis (71) L_{n+1} ist orthogonal zu P_n und besitzt nach (Satz 66 (Eigenschaften der Nullstellen)) die Darstellung

$$L_{n+1}(x) = \prod_{i=0}^{n} (x - \lambda_i)$$

mit reellen, paarweise verschiedenen Nullstellen λ_0, λ_n. Nach (Satz 62 (hinreichendes Kriterium für maximale Exaktheit)) ist die hiermit definierte Quadraturformel von der Ordnung $2n+2$. Die Gewichte ergeben sich über die Lagrangeschen Basispolynome

$$l_i := L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - \lambda_i}{\lambda_i - \lambda_j}$$

$$l_i^2 \in P_{2n} \Rightarrow 0 < \int_{-1}^{1} l_i(x)^2 \mathrm{d}x = \sum_{j=0}^{n} \alpha_i l_i(\lambda_j)^2 = \alpha_i$$

Eindeutigkeit: Angenommen, es gibt eine weitere interpolatorische Quadraturformel $I_n^*(f)$ von der Ordnung $2n+2$. Dann folgt analog, dass $\alpha_i^* = \int_{-1}^{1} l_i^*(x)^2 \mathrm{d}x$ mit $l_i^*(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - \lambda_j^*}{\lambda_i^* - \lambda_j^*} \in P_n$

Es ist

$$0 = \int_{-1}^{1} \frac{1}{\alpha_i^*} l_i^*(x) L_{n+1}(x) \mathrm{d}x = \sum_{j=0}^{n} \frac{\alpha_j^*}{\alpha_i^*} l_i(\lambda_j^*) L_{n+1}(\lambda_j^*) = L_{n+1}(\lambda_i^*)$$

$\lambda_i = \lambda_i^*$, da die Nullstellen von L_{n+1} eindeutig bestimmt sind.

Restglieddarstellung: Es ist bekannt, dass die Hermitesche Interpolationsaufgabe $h(\lambda_i) = f(\lambda_i)$, $h'(\lambda_i) = f'(\lambda_i)$ eindeutig lösbar ist mit der Restglieddarstellung:

$$f(x) - h(x) = \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!} \prod_{i=0}^{n} (x - \lambda_i)^2$$

Aus $I(h) = I_n(h)$ folgt

$$I(f) - I_n(f) = I(f-h) - I_n(f-h) = \int_{-1}^{1} \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!} \prod_{i=0}^{n} (x - \lambda_i)^2 \mathrm{d}x - \sum_{i=0}^{n} \alpha_i \underbrace{(f(\lambda_i) - h(\lambda_i))}_{=0}$$

Beachte $\frac{f^{(2n+2)(\xi_x)}}{(2n+2)!} = \frac{f(x)-h(x)}{\prod_{i=0}^{n}(x-\lambda_i)^2}$ ist definiert auf $(-1,1)$ und stetig. Sodann können wir den Mittelwertsatz anwenden:

$$I(f) - I_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{1} \prod_{i=0}^{n} (x - \lambda_i)^2$$

Beispiele

1. $n = 1$ (Ordnung 4): $I_1(f) = f(-\sqrt{\frac{1}{3}}) + f(\sqrt{\frac{1}{3}}) = \int_{-1}^{1} f(x)\mathrm{d}x \; - \frac{1}{135} f^{(iv)}(\xi)$
2. $n = 2$ (Ordnung 6): $I_2(f) = \frac{1}{9}(5f(-\sqrt{\frac{3}{5}}) + 8f(0) + 5f(\sqrt{\frac{3}{5}})) = \int_{-1}^{1} f(x)\mathrm{d}x \; - \frac{1}{15750} f^{(vi)}(\zeta)$
3. Gaußsche Quadraturformeln über einem beliebigen Intervall $[a, b]$: Wir benutzen die Koordinatentransformation $\varphi\colon [-1, 1] \to [a, b], x \mapsto \frac{b-a}{2}x + \frac{b-a}{2}$:

$$\int_a^b f(x)\mathrm{d}x \; = \int_{-1}^{1} f(\varphi(x))\varphi'(x)\mathrm{d}x \; = \frac{b-a}{2} \sum_{i=0}^{n} \alpha_i f(\varphi(x_i)) + \frac{b-a}{2} R_n(f \circ \varphi)$$

$$= \sum_{i=0}^{n} \underbrace{\tilde{\alpha}_i}_{\frac{b-a}{2}\alpha_i} f(\underbrace{\tilde{\lambda}_i}_{\frac{b-a}{2}\lambda_i}) + \frac{b-a}{2} \cdot \frac{d^{2n+2}}{dx^{2n+2}}(f \circ \varphi)(\xi) \int_{-1}^{1} \prod_{j=0}^{n} (x - \lambda_j)^2 \mathrm{d}x$$

72 **Satz: Konvergenz der Gaußschen Quadraturformeln**

Es sei $f \in C([-1, 1])$. Dann gilt für die Gaußschen Quadraturformeln I_n

$$\lim_{n\to\infty} I_n(f) = \int_{-1}^{1} f(x)\mathrm{d}x$$

Beweis (72) Die Gewichte $\alpha_i^{(n)}$ der n-ten Gaußschen Quadraturformel sind nicht-negativ. Außerdem ist

$$\sum_{i=0}^{n} \alpha_i^{(n)} = \sum_{i=0}^{n} \int_{-1}^{1} L_i(x)\mathrm{d}x \; = \int_{-1}^{1} \sum_{i=0}^{n} L_i(x)\mathrm{d}x \; = \int_{-1}^{1} 1 = 2$$

da die Lagrangeschen Basispolynome eine Zerlegung der 1 bilden. Es sei nun $\varepsilon > 0$ beliebig fest vorgegeben. Dann existiert nach dem Weierstraßschen Approximationssatz ein $p_\varepsilon \in P_m$ mit hinreichend großem m sodass über $[-1, 1]$ gilt

$$\|f(x) - p(x)\|_\infty \leq \frac{\varepsilon}{4}$$

Es ist $I(p_\varepsilon) = I_n(p_\varepsilon)$ für $2n + 2 \geq m$ und damit

$$|I(f) - I_n(f)| = |I(f) - I(p_\varepsilon) + I(p_\varepsilon) - I_n(p_\varepsilon) + I_n(p_\varepsilon) - I_n(f)| \leq |I(f - p_\varepsilon)| + |I_n(p_\varepsilon - f)|$$

$$\leq \int_{-1}^{1} |f(x) - p_\varepsilon(x)|\mathrm{d}x \; + \sum_{i=0}^{n} |\alpha_i^{(n)}| p_\varepsilon(\lambda_i) - f(\lambda_i)||$$

$$\leq 2 \max_{-1\leq x\leq 1} |f(x) - p_\varepsilon(x)| + \max_{-1\leq x\leq 1} |f(x) - p_\varepsilon(x)| \underbrace{\sum_{i=0}^{n} \alpha_i^{(n)}}_{=2} \leq 2\frac{\varepsilon}{4} + 2\frac{\varepsilon}{4} = \varepsilon$$

73 **Bemerkung: gewichtete Integrale**

Zur Herstellung von Quadraturformeln für Integrale der Form $\int_a^b f(x)\omega(x)\mathrm{d}x$ mit einer integrablen Gewichtsfunktion $\omega(x) > 0$ kann analog vorgegangen werden. Die Stützstellen sind dann die Nullstellen der Polynome, welche mit dem Skalarprodukt $(p,q)_\omega = \int_a^b p(x)q(x)\omega(x)\mathrm{d}x$ orthogonal sind.

Zum Beispiel ergeben sich für $\omega(x) = (1-x^2)^{-\frac{1}{2}}$ auf $[-1,1]$ die orthogonalen Čebyšëv-Polynome:

$$T_0(x) = 1, T_1(x) = x, T_{n+1}(x) = 2xT_n(x) - T_{n+1}(x)$$

Mit den Stützstellen $\lambda_i = \cos\left(\frac{\pi}{2} \cdot \frac{2i+1}{n+1}\right)$, $\alpha_i = \frac{\pi}{n+1}$.

3.3 Rombergsches Interpolationsverfahren

Idee Man verwendet summierte Quadraturformeln zur Extrapolation.

Beispiel zusammengesetzte Trapezregel

$$\int_a^b f(x)\mathrm{d}x = \underbrace{h\left(\frac{1}{2}f(a) + \sum_{j=1}^{n-1} f(x_j) + \frac{1}{2}f(b)\right)}_{=:a(h)} - h^2\frac{b-a}{12}f''(\zeta)$$

für ein $f \in C([a,b])$.

74 **Satz: Euler-Maclaurinsche Summenformel**

Für $f \in C^{2m+2}([a,b])$ gilt

$$a(h) = \int_a^b f(x)\mathrm{d}x + \sum_{k=1}^{m} h^{2k}\frac{1}{(2k)!}B_{2k}(f^{(2k-1)}(b) - f^{(2k-1)}(a)) + h^{2m+2}\frac{b-a}{(2m+2)!}B_{2m+2}f^{(2m+2)}(\zeta)$$

Beweis (74) (siehe Sto02, S. 163-165)

75 **Bemerkung: Bernoulli-Zahlen**

Die B_i genügen der Rekursionsvorschift

$$B_k = -\sum_{j=0}^{k-1}\frac{k!}{j!\,(k-j+1)!}B_j$$

mit $B_0 := 1$.

Rombergsches Integrationsverfahren

1. Für eine Schrittweitenfolge $h_0 > h_1 > \cdots > h_m$ wird $a(h_k)$ berechnet. Z.B. verwendet man oft die *Romberg-Folge* $h_k = \frac{h}{2^k}$, da dann Funktionswerte immer wiederverwendet werden können.
2. Dann betrachtet man das Interpolationspolynom p_i durch $(h_i^2, a(h_i)), i = 0, \cdots, m$ (Satz 29 (Extrapolationssatz), Satz 74 (Euler-Maclaurinsche Summenformel)).
3. Da wir die p_i nur an einer Stelle $h = 0$ auswerten wollen, verwenden wir den Neville Algorithmus.

76 **Satz: Konvergenz des Rombergverfahren**
Es sei $f \in C^{2m+2}([a,b])$ und $h_k := \frac{h}{2^k}$. Dann gilt für $a_{m,m}$ des Nevilleschen Algorithmus:

$$\int_a^b f(x)\mathrm{d}x - a_{m,m} = O(h^{2m+2})$$

Beweis (76) Dies gilt nach (Satz 29 (Extrapolationssatz)), (Satz 74 (Euler-Maclaurinsche Summenformel)), (Bemerkung 30 (Konvergenzgeschwindigkeiten im Neville-Schema)).

4 Numerik gewöhnlicher Differentialgleichungen

Beispiel Astrophysik
Wir schauen uns zwei astronomische Körper im gegenseitigen Schwerefeld an. Dabei betrachten wir nur den 2-dimensionalen Raum ($\mathbb{R}^2$) und die Körper als Punktmassen. O.B.d.A liegt ein Körper bei $(0,0)$ und die Position des anderen ist dargestellt als $(x(t), y(t))$. Dann gilt das *Newtonsche Gesetz*:

$$x''(t) = -\frac{\gamma}{r(t)^3}x(t), \quad y''(t) = -\frac{\gamma}{r(t)^3}y(t)$$
$$r(t) := \sqrt{x(t)^2 + y(t)^2}$$

mit den Anfangsbedingungen $x(0) = 1 - \varepsilon$, $x'(0) = 0$ sowie $y(0) = 0$, $y'(0) = \sqrt{\gamma\frac{1+\varepsilon}{1-\varepsilon}}$ für ein $\varepsilon \in (0,1)$.
Durch die Modellbildung gibt es bereits einen Modellfehler. Löst man dieses Anfangswertproblem nun numerisch kommt noch ein Diskretisierungsfehler hinzu.

4.1 Grundlagen

Notation Es sind $x, y \in \mathbb{R}^d$, $d \in \mathbb{N}$. Dann verwenden wir das euklidische Skalarprodukt $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$ wobei x_i, y_i die Komponenten der Vektoren sind. Dazu ist $\|x\| = \langle x, x\rangle^{\frac{1}{2}}$ die euklidische Norm. Für eine Matrix $A \in \mathbb{R}^{d\times d}$ wählen wir die induzierte Matrixnorm $\|A\| = \sup_{x\in\mathbb{R}^d} \frac{\|Ax\|}{|x|}$. Ableitungen bezeichnen wir wie folgt: $u'(t) = \frac{du}{dt}$, $f_t(t,x) := \frac{\partial f(t,x)}{\partial t}$, $\partial_i f(t,x) := \frac{\partial f(t,x)}{\partial x_i}$.

77 **Definition: „Anfangswertproblem“**

Es sei $I := [t_0 - T, t_0 + T]$, $\Omega \subset \mathbb{R}^d$ abgeschlossen und $(t_0, u_0) \in D := I \times \Omega$ sowie $f: D \to \mathbb{R}^d$ stetig. Dann heißt eine stetig differenzierbare Funktion $u: I \to \mathbb{R}^d$ mit

1. $\text{Graph}(u) := \{ (t, u(t)) \mid t \in I \} \subset D$
2. $u'(t) = f(t, u(t)), t \in I$
3. $u(t_0) = u_0$

Lösung der Anfangswertaufgabe

78 **Satz: Existenzsatz von Peano**

Es sei $D := \left\{ (t, x) \in \mathbb{R} \times \mathbb{R}^d \;\middle|\; |t - t_0| \leq \alpha, \|x - x_0\| \leq \beta \right\}$ und $f: D \to \mathbb{R}^d$ stetig. Dann existiert eine Lösung auf dem Intervall $I := [t_0 - T, t_0 + T]$ mit $T := \min(\alpha, \frac{\beta}{M})$ wobei $M := \max_{(t,x) \in D} \|f(t, x)\|$.

79 **Satz: Fortsetzungssatz**

Es sei $D \subset \mathbb{R} \times \mathbb{R}^d$ abgeschlossen mit $(t_0, u_0) \in D$ und $f: D \to \mathbb{R}^d$ stetig sowie u die Lösung der Anfangswertaufgabe auf einem Intervall $I = [t_0 - T, t_0 + T]$. Dann ist die lokale Lösung u nach rechts und links auf ein maximales Existenzintervall $I_{\max} = [t_0 - T^*, t_0 + T^* j]$ stetig differenzierbar fortsetzbar, solange der Graph nicht an den Rand von D stößt.

80 **Definition: „Lipschitz-Bedingung“**

Es sei $D \subset \mathbb{R} \times \mathbb{R}^d$ und $f: D \to \mathbb{R}^d$.

1. f erfüllt eine (gleichmäßige) *Lipschitz-Bedingung*, falls mit einer stetigen Funktion $L(t)$ gilt:

$$\forall (t, x), (t, x') \in D \| f(t, x) - f(t, x') \| \leq L(t) \| x - x' \|$$

2. f erfüllt in D eine *lokale Lipschitz-Bedingung*, wenn auf jeder beschränkten Teilmenge von D eine gleichmäßige Lipschitz-Bedingung erfüllt ist.

81 **Satz: Lokaler Stabilitätssatz**

Es sei $D = I \times \Omega \subset \mathbb{R} \times \mathbb{R}^d$ und $f, g: D \to \mathbb{R}^d$. Ferner erfüllen u, v auf $I = [t_0, t_0 + T]$ die Anfangswertaufgabe:

$$u'(t) = f(t, u(t)), t \in I, u(t_0) = u_0$$
$$v'(t) = g(t, u(t)), t \in I, v(t_0) = v_0$$

Erfüllt f eine Lipschitz-Bedingung, so gilt

$$\|u(t) - v(t)\| \leq e^{\tilde{L}(t - t_0)} (\|u_0 - v_0\| + \int_{t_0}^{t} \varepsilon(s) \mathrm{d}s)$$

mit $\tilde{L} := \max_{t \in I} L(t)$ und $\varepsilon(s) := \sup_{x \in \Omega} \|f(s, x) - g(s, x)\|$.

82 **Korollar: Eindeutigkeit**

Es sei $D \subset \mathbb{R} \times \mathbb{R}^d$ und $f\colon D \to \mathbb{R}^d$ erfülle eine Lipschitz-Bedingung, dann existiert höchstens eine Lösung der Anfangswertaufgabe. Insbesondere ist jede lokale Lösung eindeutig fortsetzbar.

83 **Bemerkung: Satz von Picard-Lindelöf**

(Satz 78 (Existenzsatz von Peano)) und (Korollar 82 (Eindeutigkeit)) ergeben zusammen die Aussage des klassischen Existenzsatzes von Picard-Lindelöf.

4.2 Einschrittmethoden

Wir betrachten die Anfangswertaufgabe $u'(t) = f(u(t)), t \in [t_0, t_0 + T]; u(t_0) = u_0$ mit einer stetigen Funktion $f\colon I \times \mathbb{R}^d \times \mathbb{R}^d$, welche die globale Lipschitzbedingen

$$\forall (t,x),(t,y) \in I \times \mathbb{R}^d \, \|f(t,x) - f(t,y)\| \leq L\|x - y\|$$

erfüllt. Nach *Korollar 82 (Eindeutigkeit)* ist Existenz und Eindeutigkeit einer Lösung auf dem Existenzintervall $[t_0, t_0 + T]$ bekannt.

4.2.1 Explizites Euler-Verfahren

Wir betrachten eine Zerlegung $t_0 < t_1 < \cdots < t_N = t_0 + T$ von I in $I_n := [t_{n-1}, t_n]$ und setzen $h_n := t_n - t_{n-1}$ und $h := \max_{1 \leq n \leq N} h_n$ die Feinheit der Zerlegung. Dann nähern wir die Anfangswertaufgabe mittels des Differenzenquotienten an, mit Lösung y_n^h:

$$\frac{y_n^h - y_{n-1}^h}{h_n} = f(t, y_{n-1}^h) \Rightarrow y_n^h = y_{n-1}^h + h_n f(t_{n-1}, y_{n-1}^h)$$

welche also rekursiv berechnet werden kann. Man kann es auch als $(L_h y^h)_n = 0, n = 1, 2, \cdots$ schreiben mit dem „Differenzenoperator“ $(L_h y^h)_n := h_n^{-1}(y_n^h - y_{n-1}^h) - f(t_{n-1}, y_{n-1}^h)$. Nach dem Satz von Peano konvergieren diese Approximationen $y_n^h \to u(t_n)$ für $h \to 0$.

84 **Definition: „Abschneidefehler“**

Der Ausdruck $\tau_n^h := (L_h u^h)_n$ mit Differenzenoperator L_h und $u_n^h := u(t_n)$ heißt *Abschneidefehler*.

85 **Satz: Abschätzung des Abschneidefehlers**

Für den Abschneidefehler

$$\tau_n^h = (L_h u^h)_n = h_n^{-1}(u_n - u_{n-1}) - f(t_{n-1}, u_{n-1})$$

des expliziten Euler-Verfahrens gilt

$$\|\tau_n^h\| \leq \frac{1}{2} h_n \max_{t \in I_n} \|u''(t)\|$$

Beweis (85)

$$\tau_n^h = h_n^{-1} \int_{I_n} u'(t)dt - u'(t_{n-1}) = h_n^{-1} \left[(t-t_n)u'(t)\right]_{t_{n-1}}^{t_n} - \int_{t_{n-1}}^{t_n} (t-t_n)u''(t)dt - u'(t_{n-1})$$
$$= -h_n^{-1} \int_{t_{n-1}}^{t_n} (t-t_n)u''(t)dt$$
$$\|\tau_n^h\| \leq h_n^{-1} \int_{t_{n-1}}^{t_n} (t-t_n)\|u''(t)\|dt = h_n^{-1} \max_{t\in I_n} \|u''(t)\| \int_{t_{n-1}}^{t_n} (t-t_n)dt$$
$$= h_n^{-1} \max_{t\in I_n} \|u''(t)\| \frac{1}{2}(t_n - t_{n-1})^2 = \frac{1}{2} h_n \max_{t\in I_n} \|u''(t)\|$$

86 **Lemma: Diskretes Gronwallsches Lemma**

Es seien $(a_i)_{i\in\mathbb{N}}, (b_i)_{i\in\mathbb{N}}, (w_i)_{i\in\mathbb{N}}$ Folgen nicht-negativer Zahlen und $(b_i)_{i\in\mathbb{N}}$ nicht fallend (d.h. monoton wachsend ?!?). Dann gilt

1. Ist $\forall i \in \mathbb{N} a_i < 1$ sowie $w_n \leq \sum_{i=0}^{n} a_i w_i + b_n, n \geq 1$, so ist mit $\sigma_i := (1-a_i)^{-1}$:

$$w_n \leq \exp\left(\sum_{i=0}^{n} \sigma_i a_i\right) b_n, n \geq 1$$

2. Gilt $w_n \leq \sum_{i=0}^{n-1} a_i w_i + b_n, n \geq 0$, so ist mit $\sigma_i = 1$:

$$w_n \leq \exp\left(\sum_{i=1}^{n-1} \sigma_i a_i\right) b_n, n \geq 1$$

87 **Satz: Abschätzung des Diskretisierungsfehlers**

Für den Diskretisierungsfehler $e_n^h := y_n^h - u_n$ gilt

$$\max_{1\leq n\leq N} \|e_n^h\| \leq e^{LT}(\|e_0^h\| + T \max_{1\leq n\leq N} \|\tau_n^h\|)$$
$$\max_{1\leq n\leq N} \|e_n^h\| \leq e^{LT}(\|e_0^h\| + \frac{1}{2} T \max_{1\leq n\leq N} (h_n \max_{t\in I_n} \|u''(t)\|))$$

Beweis (87)

$$u_n = u_{n-1} + h_n f(t_{n-1}, u_{n-1}) + h_n(h_n^{-1}(u_n - u_{n-m}) - f(t_{n-1}, u_{n-1})) = u_{n-1} + h_n f(t_{n-1}, u_{n-1}) + h_n \tau_n^h$$
$$e_n^h = y_n^h - u_n = y_{n-1}^h + h_n f(t_{n-1}, y_{n-1}) - (u_{n-1} + h_n f(t_{n-1}, u_{n-1}) + h_n \tau_n^h)$$
$$= e_{n-1}^h + h_n(f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})) - h_n \tau_n^h$$
$$\Rightarrow \|e_n^h\| \leq \|e_{n-1}^h\| + h_n L \|e_{n-1}^h\| + h_n \|\tau_n^h\|$$
$$\Rightarrow \|e_n^h\| \leq \|e_0\| + L \sum_{i=0}^{n-1} h_{i+1} \|e_i\| + \sum_{i=1}^{n} h_i \|\tau_i^h\|$$

Dann setzen wir in das Gronwallsche Lemma ein: $w_n = \|e_n^h\|$, $a_n = h_{n+1}$, $b_n = \sum_{i=1}^n h_i \|\tau_i^h\|$. Damit folgt aus (Lemma 86 (Diskretes Gronwallsches Lemma)) und (Satz 85 (Abschätzung des Abschneidefehlers)) die Behauptung.

4.2.2 Differenzenformeln höherer Ordnung über Taylor-Entwicklung

Wir betrachten nur den skalaren Fall. Dann ist

$$u(t) = \sum_{k=0}^{n} \frac{h^k}{k!} u^{(k)}(t-h) + \frac{h^{n+1}}{(n+1)!} u^{(n+1)}(\xi), \xi \in [t-h, t]$$

Aus $u'(t) = f(t, u(t))$ folgt

$$u^{(r)}(t) = \left(\frac{d}{dt}\right)^{r-1} f(t, u(t)) =: f^{(r-1)}(t, u(t))$$

Dann erhalten wir die „R-stufige Taylor-Methode“

$$y_n = y_{n-1} + h_n \sum_{r=1}^{R} \frac{h_n^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}) =: y_{n-1} + h_n F(h_n, t_{n-1}, y_n, y_{n-1})$$

mit der Verfahrensfunktion

$$F(h, t, y, x) = \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t, x)$$

bzw. mit Differenzenoperator

$$(L_h y^h)_n = h_n^{-1}(y_n - y_{n-1}) - F(h_n, t_{n-1}, y_n, y_{n-1})$$

88 **Bemerkung: Bezeichnungen**

1. Verfahren mit der Verfahrensvorschrift $F(h_n, t_{n-1}, y_n, y_{n-1})$ heißen Einschrittverfahren.
2. Hängt f nicht von y_n ab, so handelt es sich um ein explizites Verfahren, das lediglich eine Funktionsauswertung pro Schritt erfordert. Sonst heißt das Verfahren implizit. Bei impliziten Verfahren müssen im Allgemeinen nichtlineare Gleichungssysteme pro Schritt gelöst werden. (Implizites Euler-Verfahren: $h_n^{-1}(y_n - y_{n-1}) = f(t_n, y_n), n \geq 1, y_0 = u_0$)
3. Der Abschneidefehler für allgemeine Einschrittverfahren ist definiert als

$$\tau_n^h := (L^h u^h)_n = h_n^{-1}(u_n - u_{n-1}) - F(h_n, t_{n-1}, u_n, u_{n-1})$$

89 **Definition: „Konsistenz“**

Ein Einschrittverfahren heißt *konstistent* (mit der *Konstistenzordnung m*) falls

$$\max_{t_n \in I} \|\tau_n^h\| = O(h^m), h \to 0$$

90 **Bemerkung: R-stufige Taylor-Formel**

1. Wegen $\tau_n^h = \frac{h^R}{(R+1)!}u^{(R+1)}(\xi), \xi \in [t_{n-1}, t_n]$ hat die R-stufige Taylor-Formel (für skalare Anfangswertaufgaben) die Konstistenzordnung $m = R$.

2. Für die Auswertung der R-stufigen Taylor-Formel müssen Ableitungen von $f(t, x)$ bestimmt werden, was in der Regel unpraktikabel. Stattdessen werden Ableitungen durch Differenzenquotienten ersetzt, so dass nur Auswertungen von $f(t, u(t))$ auftreten.

Beispiel: 2-stufige Taylor-Formel

$$f'(t,u(t)) \approx h^{-1}(f(t+h,u(t+h)) - f(t,u(t))) \approx h^{-1}(f(t+h,u(t+h)) - f(t,u(t)))$$
$$y_n = y_{n-1} + h_n f(t_{n-1}, y_{n-1}) + \frac{1}{2}h_n(f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1})) - f(t_{n-1}, y_{n-1}))$$
$$= y_{n-1} + h_n(\frac{1}{2}f(t_{n-1}, y_{n-1}) + \frac{1}{2}f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1})))$$

Bei Beachtung aller Restglieder kann gezeigt werden, dass diese Differenzenformel die Konstistenzordnung 2 hat.

4.2.3 Explizite Runge-Kutta-Formeln

Die Verfahrensfunktion der expliziten Runge-Kutta-Verfahren ist:

$$F(h,t,y) := \sum_{r=1}^{R} c_r k_r(h,t,y)$$
$$k_1(h,t,y) = f(t,y), k_r = f(t + ha_r, y + h\sum_{s=1}^{r-1} b_{rs}k_s(h,t,y))$$

mit $r = 2, \cdots, R$ und geeigneten Konstanten a_r, c_r, b_{rs}.

Ziel Die Konstanten sollen so bestimmt werden, dass mit möglichst großem m (idealerweise $m = R$) gilt

$$\sum_{r=1}^{R} c_r k_r(h,t,u(t)) = \sum_{r=1}^{R} \frac{h^{r-1}}{r!} \underbrace{\frac{d^{r-1}f(t,u(t))}{dt^{r-1}}}_{=u^{(r)}(t)} + O(h^m)$$

Dann ist die Konstistenzordnung der Runge-Kutta-Formeln gleich m.

Beispiel

$R = 1$: Explizites Euler-Verfahren

$R = 2$**:** Die Taylorentwicklung liefert mit anschließendem Koeffizientenvergleich:

$$\begin{aligned} & c_1 f + c_2 f(t + ha_2, u + hb_{2,1} f) = c_1 f + c_2 (f + ha_2 f_t + hb_{2,1} f f_x + O(h^2)) \\ & \stackrel{!}{=} f + \frac{1}{2} h (f_t + f_x f) + O(h^2) \\ & \Rightarrow c_1 + c_2 = 1, c_2 a_2 = c_2 b_{2,1} = \frac{1}{2} \end{aligned}$$

Mögliche Lösungen:

1. $c_1 = c_2 = \frac{1}{2}$, $a_2 = b_{2,1} = 1$ (*Heunsche Formel* zweiter Ordnung)

$$y_n = y_{n-1} + \frac{1}{2} h_n (f(t_{n-1}, y_{n-1}) + f(t_n, y_{n-1} + h_n f(t_{n-1}, y_{n-1})))$$

2. $c_1 = 0, c_2 = 1, a_2 = b_{2,1} = \frac{1}{2}$ (*modifizierte Euler-Formel*)

$$y_n = y_{n-1} + h_n f(t_{n-1} + \frac{1}{2} h_n, y_{n-1} + \frac{1}{2} h_n f(t_{n-1}, y_{n-1}))$$

91 **Bemerkung: Übertragbarkeit auf Systeme**

R-stufige Taylor-Formeln bzw. Runge-Kutta-Verfahren sind auf Systeme übertragbar. Jedoch kann die Anzahl zur Verfügung stehender Parameter für den Abgleich zu klein sein.

92 **Definition: „Lipschitzstetigkeit“**

Ein Einschrittverfahren heißt *Lipschitzstetig*, wenn seine Verfahrensfunktion eine Lipschitzbedingung erfüllt:

$$\forall t \in I, x, \tilde{x}, y, \tilde{y} \in \mathbb{R}^d \| F(h, t, x, y) - F(h, t, \tilde{x}, \tilde{y}) \| \leq L(\|x - \tilde{x}\| + \|y - \tilde{y}\|)$$

93 **Satz: Stabilität von Einschrittverfahren**

Ein Lipschitzstetiges Einschrittverfahren ist im Folgenden Sinn stabil: Für hinreichend kleines $h, 0 < h < \frac{1}{4} L^{-1}$ gilt

$$\|y_n - z_n\| \leq e^{4L(t_n - t_0)} (\|y_0 - z_0\| + \sum_{r=1}^{n} h_r \| L_h y^h - L_h z^h \|)$$

für beliebige *Gitterfunktionen*

$$y^h = \{y_n\}_{n=0,\cdots,N}, z_n = \{z_n\}_{n=0,\cdots,N}$$

Für explizite Einschrittverfahren kann die Schrittweitenbedingung entfallen.

Beweis (93)

$$y_n - z_n = y_{n-1} - z_{n-1} + h_n(F(h,t_n,y_n,y_{n-1}) - F(h,t_n,z_n,z_{n-1}) + (L_h y^h - L_h z^h)_n)$$
$$\Rightarrow \|y_n - z_n\| \le \|y_{n-1} - z_{n-1}\| + h_n L(\|y_n - z_n\| + \|y_{n-1} - z_{n-1}\|) + h_n\|(L_h y^h - L_h z^h)_n\|$$
$$\le \|y_0 - z_0\| + L\sum_{r=0}^{n} h_r(\|y_r - z_r\| + \|y_{r-1} - z_{r-1}\|) + h_n \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|$$

Nach dem diskreten Gronwall-Lemma Teil 1

$$\le \exp(\sum_{r=0}^{n} \frac{L(h_r + h_{r+1})}{1 - L(h_r + h_{r+1})})(\|y_0 - z_0\| + \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|)$$
$$\le \exp(2L\sum_{r=0}^{n} h_r + h_{r+1})(|y_0 - z_0| + \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|)$$
$$\le \exp(4L(t_n - t_0))(\|y_0 - z_0\| + \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|)$$

Explizites Verfahren:

$$\|y_n - z_n\| \le \|y_{n-1} - z_{n-1}\| + h_n L\|y_{n-1} - z_{n-1}\| + h_n\|(L_h y^h - L_h z^h)_n\|$$
$$\le \|y_0 - z_0\| + L\sum_{r=0}^{n-1} h_{r+1}\|y_r - z_r\| + \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|$$

Nach dem diskreten Gronwall-Lemma Teil 2

$$\|y_n - z_n\| \le \exp(\sum_{r=0}^{n} h_r)(\|y_0 - z_0\| + \sum_{r=0}^{n} h_r\|(L_h y^h - L_h z^h)_r\|)$$

94 **Satz: Konvergenz von Einschrittverfahren**

Es sei ein lipschitzstetiges Einschrittverfahren gegeben. Dann gilt für den Diskretisierungsfehler für eine hinreichend kleine Schrittweite h die a priori Fehlerabschätzung

$$\|y_n^h - u(t_n)\| \le e^{4L(t_n - t_0)}(\|y_0 - u_0\| + \sum_{\nu=1}^{n} h_\nu\|\tau_\nu^h\|)$$

Für explizite Einschrittverfahren kann die Schrittweitenbedingung wiederum entfallen.

Bei konsistenten Einschrittverfahren mit $y_0 = u_0$ gilt demnach

$$\lim_{h\to 0}\max_{t_n\in I}\|y_n - u(t_n)\| = 0$$

Die Konvergenzordnung ist mindestens so groß wie die Konsistenzordnung.

Beweis (94) Wir wissen $L_n y_n^h = 0$ und $L_n u^h = \tau^h$, sodass Einsetzen in (Satz 93 (Stabilität von Einschrittverfahren)) die Behauptung gibt.

95 **Bemerkung: Konsistenz von speziellen Einschrittverfahren**

Taylor-Formeln und Runge-Kutta-Verfahren sind konsistent und Lipschitz-stetig für Lipschitz-stetige f. Bei Runge-Kutta-Verfahren braucht man für die Konsistenz allerdings $\sum_{r=1}^{R} c_r = 1$.

4.2.4 Globale Konvergenzaussagen

96 **Definition: „Monotoniebedingung“**

Es sei $D = \mathbb{R} \times \mathbb{R}^d$. Die Funktion $f\colon D \to \mathbb{R}^d$ erfüllt eine Monotoniebedingung, falls mit $\lambda(t) > 0$ gilt

$$-\langle f(t,x) - f(t,y), x - y\rangle \geq \lambda(t)\|x - y\|^2$$

97 **Satz: globale Fehlerabschätzung des impliziten Eulerverfahrens**

Die Funktion $f\colon \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ erfülle die Lipschitzbedingung und sei monoton mit konstantem $\lambda(t) =: \lambda$. Dann sind die Lösungen des impliziten Eulerverfahrens $y_n = y_{n-1} + h_n f(t_n, y_n), n \geq 1, y_0 = u_0$ wohldefiniert und es gilt die globale Fehlerabschätzung

$$\lambda > 0\colon \|y_n - u(t_n)\| \leq \frac{1}{2} \min\left\{t_n - t_0, \lambda^{-1}\right\} \max_{\nu\in\mathbb{N}} \left\{h_\nu \|u''\|\right\}$$

$$\lambda \geq 0\colon \|y_n - u(t_n)\| \leq \frac{1}{2}(t_n - t_0) \max_{\nu\in\mathbb{N}} \left\{h_\nu \|u''\|\right\}$$

Beweis (97) (Ran)

98 **Bemerkung: Beschränktheit**

1. Für homogene ($f(t,0) = 0$) Lipschitz-stetiges und monotone Anfangswertaufgaben kann für das explizite Euler-Verfahren gezeigt werden

   $$\|y_n\| \leq (1 + h_n^2 L(t_{n-1})^2 - 2\lambda(t_{n-1})h_n)\|y_{n-1}\|$$

   Das heißt y_n ist beschränkt, falls $h_n \leq \frac{2\lambda(t_{n-1})}{L(t_{n-1})^2}$.

2. Falls $h_n < \frac{2\lambda(t_{n-1})}{L(t_{n-1})^2}$ erfüllt ist, gilt insbesondere

   $$\|y_n - u(t_n)\| \leq c \max_{1\leq\nu\leq n} \left\{ h_\nu \max_{I_\nu} \|u''\| \right\}$$

3. Für Einschrittverfahren höherer Ordnung können die Aussagen im Allgemeinen nicht übertragen werden.

Numerische Stabilität (Modellbeispiel) Betrachtet wird das Modellbeispiel

$$u'(t) = \lambda u(t), \lambda \in \mathbb{C}, u(t_0) = u_0$$

Die Lösung ist bekanntlich $u(t) = u_0 e^{\lambda(t-t_0)}$. Für das Verhalten bei $t \to \infty$ müssen wir dann drei Fälle unterscheiden:

$\mathrm{Re}(x) < 0$: $|u(t)| \to 0$
$\mathrm{Re}(x) = 0$: $|u(t)| \to |u_0|$
$\mathrm{Re}(x) > 0$: $|u(t)| \to \infty$

99 **Definition: „numerisch stabil“**
Eine Einschrittmethode heißt numerisch stabil für $\lambda h \neq 0$, wenn sie für das obige skalare Modellbeispiel im Falle $\mathrm{Re}(\lambda) \leq 0$ beschränkte Näherungen liefert: $\sup_{n \geq 0} |y_n| < \infty$.

Explizites Euler-Verfahren

$$y_n = (1 + \lambda h) y_{n-1} = \cdots = (1 + \lambda h)^n y_0$$

Es ergibt sich also ein sogenannter Verstärkungsfaktor $\omega(\lambda h) = 1 + \lambda h$ und das Verfahren ist numerisch stabil, wenn der Verstärkungsfaktor ω die Bedingung $|\omega| \leq 1$ erfüllt. Die Menge $\mathrm{SG} := \{ z = \lambda h \in \mathbb{C} \mid \omega(z) \leq 1 \} = B_1(-1)$ heißt *Stabilitätsgebiet*. Für ein festes λ muss die Schrittweite h so gewählt werden, dass $\lambda h \in \mathrm{SG}$ ist.

Taylor-Methode

$$y_n = y_{n-1} + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t_{n-1}, y_{n-1}) = y_{n-1} + h \sum_{r=1}^{R} \frac{h^{r-1}}{r!} \lambda^r y_{n-1} = (\sum_{r=0}^{R} \frac{(\lambda h)^r}{r!})^n y_0$$

Der Verstärkungsfaktor ist hier als $\omega(\lambda h) = \sum_{r=0}^{R} \frac{(\lambda h)^r}{r!}$ und das Stabilitätsgebiet wird ein Zwiebelgebilde um $B_1(-1)$. Die ersten Schnitte mit der reellen Achse sind zum Beispiel: $[-l_R, 0]$ mit $l_1 = -2$, $l_2 = -2$, $l_3 = -2,51 \cdots$ und $l_4 = -2,78 \cdots$.

Runge-Kutta-Formeln

$$F(h,t,u) = \sum_{r=1}^{R} c_r k_r \qquad F(h,t,u) = \sum_{r=1}^{R} \frac{h^{r-1}}{r!} f^{(r-1)}(t,u) + O(h^R)$$

Für das Modellproblem ist $F(h,t,u)$ ein Polynom in h der Ordnung $R-1$. Das heißt die Runge-Kutta-Formeln weisen den gleichen Verstärkungsfaktor auf wie die Taylorformeln und somit auch das gleiche Stabilitätsgebiet.

Implizites Euler-Verfahren

$$y_n = y_{n-1} + h f(t_n, y_n) = y_{n-1} + h\lambda y_n \Rightarrow y_n = (1 - \lambda h)^{-1} y_{n-1} = (1 - \lambda h)^{-n} y_0$$

Der Verstärkungsfaktor ist also $\omega(\lambda h) = (1 - \lambda h)^{-1}$ und das Stabilitätsgebiet $\mathrm{SG} = \mathbb{C} \setminus B_1(1)$.

100 **Definition: „A-Stabilität“**

Ein Einschrittverfahren heißt *A-stabil*, wenn das Stabilitätsgebiet die gesamte linke komplexe Halbebene umfasst.

101 **Bemerkung: A-Stabile Verfahren**

1. Man kann zeigen, dass explizite Methoden nicht A-stabil sein können.
2. Das implizite Euler-Verfahren ist A-stabil.
3. Zu beachten ist, dass das implizite Eulerverfahren für $\mathrm{Re}(\lambda) > 0$ und $|1 - \lambda h| \geq 1$ beschränkte Näherungen liefert, obwohl die exakte Lösung exponentiell wächst. Hierfür ist allerdings $(1 - \lambda h)^{-1} < 0$, das heißt die Näherungswerte oszillieren.

5 Abstiegsverfahren

102 **Satz: Minimierungskriterium**

Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Ferner sei $Q(y) := \frac{1}{2}\langle Ay, y\rangle - \langle b, y\rangle$. Dann ist $x \in \mathbb{R}^n$ genau dann Lösung des LGS $Ax = b$ mit $b \in \mathbb{R}^n$, wenn

$$\forall y \in \mathbb{R}^n \setminus \{x\} Q(x) < Q(y)$$

Beweis (102**)** Es sei $Ax = b$. Für $y \neq x$ gilt dann

$$Q(y) - Q(x) = \frac{1}{2}(\langle Ay, y\rangle - 2\langle b, y\rangle - \langle Ax, x\rangle + 2\langle b, x\rangle)$$
$$= \frac{1}{2}(\langle Ay, y\rangle - 2\langle Ax, y\rangle + \langle Ax, x\rangle) = \frac{1}{2}\langle A(x - y), x - y\rangle > 0$$

Sei auf der anderen Seite $Q(x) < Q(y)$ für $x \neq y$, dann gilt

$$\nabla Q(x) = 0 \qquad (\nabla Q(x)_i = \frac{\partial Q}{\partial x_i}(x))$$
$$\Rightarrow 0 = \frac{\partial Q}{\partial x_i}(x) = \frac{1}{2}\frac{\partial}{\partial x_i}\sum_{j,k=1}^{n} A_{jk}x_j x_k - \frac{\partial}{\partial x_i}\sum_{k=1}^{n} A_k x_k = \sum_{k=1}^{n} A_{ik}x_k - b_i = (Ax - b)_i$$

Idee Wir bestimmen eine Folge von Iterierten $x^t, t \in \mathbb{N}$ durch

$$x^{t+1} = x^t + \alpha_t r^t, x^0 \in \mathbb{R}^n$$
$$Q(x^{t+1}) = \min_{\alpha\in\mathbb{R}} Q(x^t + \alpha r^t)$$

Die Frage ist also welche *Abstiegsrichtung* r^t und *Schrittweite* α_t wir wählen.

103 **Lemma: Schrittweite**
Ist Q wie in Satz 102 (Minimierungskriterium), dann gilt

$$\alpha_t = -\frac{\langle g^t, r^t\rangle}{\langle Ar^t, r^t\rangle}$$

Mit $g^t = Ax^t - b = \nabla Q(x^t)$.

Beweis (103)

$$0 = \frac{\partial}{\partial\alpha} Q(x^t + \alpha r^t)|_{\alpha=\alpha_t} = \langle \nabla Q(x^t + \alpha_t r^t), r^t\rangle$$
$$= \langle Ax^t + \alpha_t Ar^t - b, r^t\rangle = \langle g^t, r^t\rangle + \alpha_t \langle Ar^t, r^t\rangle$$

Wahl der Abstiegsrichtung

- Man kann zum Beispiel die kartesischen Koordinaten zyklisch durchlaufen. (*Koordinatenrelaxation*, äquivalent zum *Gauß-Seidel-Verfahren*)
- Besser ist es allerdings eine Richtung zu wählen, in der wir eine möglichst große Verbesserung erwarten. Deshalb nimmt man zum Beispiel die Richtung des steilsten Abstiegs. O.B.d.A $\|r^t\| = 1$:

$$\frac{\partial}{\partial\alpha} Q(x^t + \alpha r^t)|_{\alpha=0} = \langle \nabla Q(x^t), r^t\rangle = \cos\measuredangle(\nabla Q(x^t), r^t)\|\nabla Q(x^t)\| = \begin{cases} 1 & , r^t = \nabla Q(x^t) \\ -1 & , r^t = -\nabla Q(x^t) \end{cases}$$

Das heißt aus lokaler Sicht ist die Reduktion von $Q(x^t+\alpha r^t)|_{\alpha=0}$ in Richtung $r^t = -\nabla Q(x^t) = -g^t$ am größten.

5.1 Gradientenverfahren

Man wählt $x^0 \in \mathbb{R}^n$, der erste Gradient ist dann $g^0 = Ax^0 - b$ und iterieren dann über t mittels

$$\alpha_t = \frac{\|g^t\|^2}{\langle Ag^t, g^t\rangle}$$
$$x^{t+1} = x^t - \alpha_t g^t$$
$$g^{t+1} = g^t - \alpha_t Ag^t$$

104 **Bemerkung: Optimierung**

1. Beachte, dass man zwar $g^{t+1} = Ax^{t+1} - b$ hat, aber dies gleich $A(x^t - \alpha^t g^t) - b = g^t - \alpha_t g^t$ ist.
2. Die letztere Berechnung ist günstiger, da wir Ag^t bereits für α_t berechnet haben und nicht noch einmal Ax^{t+1} berechnen müssen.

105 **Satz: Konvergenz des Gradientenverfahrens**
Für das Gradientenverfahren gilt die folgende Fehlerabschätzung

$$\|x^t - x\|_A \leq \left(\frac{1-\frac{1}{\kappa}}{1+\frac{1}{\kappa}}\right)^t \|x^0 - x\|_A$$

wobei $\|y\|_A := \langle Ay, y\rangle^{\frac{1}{2}}$ und $\kappa = \text{cond}_2(A) = \frac{1}{\lambda}$ die Spektralnorm ist.

Beweis (105) Mit dem Lemma von Kantorowitsch gilt

$$\begin{aligned}
&E(y) := \langle y - x, A(y-x)\rangle \\
&E(x^{t+1}) = \left(1 - \frac{\|g^t\|^4}{\langle g^t, Ag^t\rangle\langle g^t, A^{-1}g^t\rangle}\right) E(x^t) \\
&\leq \left(1 - 4\frac{\lambda\Lambda}{(\lambda+\Lambda)^2}\right) E(x^t) \\
&\|x^t - x\|_A^2 \leq \left(\frac{\lambda-\Lambda}{\lambda+\Lambda}\right)^{2t} \|x^0 - x\|_A^2
\end{aligned}$$

Wobei λ und Λ der kleinste bzw. größte Eigenwert von A ist.

106 **Bemerkung: Abstiegsrichtungen**
Die Abstiegsrichtungen des Gradientenverfahrens zu aufeinanderfolgenden Schritten sind orthogonal

$$\langle g^{t+1}, g^t\rangle = \langle g^t - \alpha_t Ag^t, g^t\rangle \quad = \|g^t\|^2 - \alpha_t\langle Ag^t, g^t\rangle = 0$$

5.2 Verfahren der konjugierten Gradienten (CG-Verfahren)

Idee Wir bestimmen nun Abstiegsrichtungen, die alle paarweise orthogonal zueinander sind. (Nicht nur bezüglich des euklidischen Skalarprodukts sondern bezüglich des von A induzierten Skalarprodukts.)
Wir beginnen also wieder mit $x^0 \in \mathbb{R}^n$, $d^0 = -g^0 = b - Ax^0$ und iterieren dann

$$\begin{aligned}
x^{t+1} &= x^t + \alpha_t d^t, \quad \alpha_t = -\frac{\langle g^t, d^t\rangle}{\langle d^t, Ad^t\rangle} \\
g^{t+1} &= g^t + \alpha_t Ad^t \qquad (= Ax^{t+1} - b) \\
d^{t+1} &= -g^{t+1} + \beta_t d^t := -g^{t+1} + \frac{\langle g^{t+1}, Ad^t\rangle}{\langle d^t, Ad^t\rangle} d^t
\end{aligned}$$

107 **Lemma: Orthogonalität**
Es seien $g^i \neq 0, i \leq t \in \mathbb{N}$, dann gilt

1. $\text{span}(g^0, \cdots, g^t) = \text{span}(A^0 g^0, \cdots, A^t g^0)$

2. $\operatorname{span}(g^0, \cdots, g^t) = \operatorname{span}(d^0, \cdots, d^t)$
3. $i < t \Rightarrow \langle g^t, g^i \rangle = 0$
4. $i < t \Rightarrow \langle d^t, Ad^i \rangle = 0$

Beweis (107) Per Induktion über t.

108 **Bemerkung: Abstiegsrichtungen**

1. Nach Lemma 107 (Orthogonalität) sind die Abstiegsrichtungen paarweise A-orthogonal (A-konjugiert).
2. Die Abstiegsrichtungen sind linear unabhängig und es gilt falls $g^t \neq 0$: $\operatorname{span}(d^0, \cdots, d^{n-1}) = \mathbb{R}^n$.
3. $$\alpha_t = \frac{\|g^t\|^2}{\langle d^t, Ad^t \rangle}, \qquad \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}$$

109 **Satz: Terminierung**

Bei Ausführung des CG-Verfahrens gilt nach spätestens n Schritten $x^t = x$. Dabei gilt in jedem Schritt

$$Q(x^t) = \min_{\alpha \in \mathbb{R}} Q(x^{t-1} + \alpha d^{t-1}) = \min_{y \in \operatorname{span}(d^0, \cdots, d^{t-1})} Q(x^0 + y)$$

Beweis (109) Es sei $g^t \neq 0$ für $t < n$ und x die Lösung des LGS, dann gilt, da die d^t den gesamten Raum aufspannen:

$$x - x_0 = \sum_{t=0}^{n-1} \gamma_t d^t \qquad \gamma_t = \frac{\langle d^t, A(x - x_0) \rangle}{\langle d^t, Ad^t \rangle}$$

$$x^t - x_0 = \sum_{i=0}^{t-1} \alpha_i d^i$$

$$\Rightarrow \langle d^t, A(x^t - x_0) \rangle = \sum_{i=0}^{n-1} \alpha_i \langle d^t, Ad^i \rangle = 0$$

$$\Rightarrow \gamma_t = \frac{\langle d^t, A(x - x_0) \rangle - \langle d^t, A(x^t - x^0) \rangle}{\langle d^t, Ad^t \rangle} = \frac{\langle d^t, A(x - x^t) \rangle}{\langle d^t, Ad^t \rangle} = -\frac{\langle d^t, g^t \rangle}{\langle d^t, Ad^t \rangle} = \alpha_t$$

$$\Rightarrow x^n = x_0 + \sum_{i=0}^{n-1} \alpha_i d^t = x^0 + \sum_{i=0}^{n-1} \gamma_i d^i = x$$

Damit ist die erste Behauptung gezeigt, sei nun $y = \sum_{i=0}^{t-1} \varepsilon_i d^i \in \operatorname{span}(d^0, \cdots, d^{t-1})$ mit $\varepsilon_i \in \mathbb{R}$, dann gilt

$$\frac{\partial}{\partial \varepsilon_j} Q(x^0 + y) = \langle \nabla Q(x^0 + y), d^j \rangle = \langle A(x^0 + y) - b, d^j \rangle = \langle A(x^0 - x), d^j \rangle + \varepsilon_j \langle A d^j, d^j \rangle \stackrel{!}{=} 0$$

$$\varepsilon_j = -\frac{\langle A(x^0 - x), d^j \rangle}{\langle A d^j, d^j \rangle}$$

$$x^0 - x = -\sum_{i=0}^{n-1} \alpha_i d^i$$

$$\Rightarrow -\langle A(x^0 - x), d^j \rangle = \sum_{i=0}^{n-1} \alpha_i \langle A d^i, d^j \rangle = \alpha_j \langle A d^j, d^j \rangle$$

$$\Rightarrow \alpha_j = \varepsilon_j$$

110 **Bemerkung: direkt/iterativ**

1. Gemäß Satz 109 (Terminierung) gehört das CG-Verfahren zur Klasse der direkten Verfahren.
2. In der Praxis wird es jedoch wie ein iteratives Verfahren verwendet, da
   - wegen Rundungsfehlern die Richtungen d^t nicht genau orthogonal sind und
   - bei großen Matrizen mit deutlich weniger als n Iterationsschritten brauchbare Näherungen gewonnen werden.

111 **Lemma: Fehlerabschätzung**

Es sei $S \subset \mathbb{R}$ mit $\{ \lambda \in \mathbb{R} \mid \lambda \text{ ist Eigenwert von } A \} \subset S$. Ferner gebe es ein Polynom $\tilde{p} \in P_t$ mit $\tilde{p}(0) = 1$, so dass $\sup_{\mu \in S} |\tilde{p}(\mu)| \leq M$. Dann gilt $\|x^t - x\|_A \leq M \|x^0 - x\|_A$ für die Iterierten $\{x^t\}$ des CG-Verfahrens.

Beweis (111) Es gilt $\|x^t - x\|_A = \min_{y \in B_t} \|x^0 - x + y\|_A$ wegen $Q(y) - Q(x) = \frac{1}{2}\|y - x\|_A^2$ mit $B_t = \operatorname{span}(d^0, \cdots, d^{t-1}) = \operatorname{span}(A^0 g^0, \cdots, A^{t-1} g^0)$. Dann folgt

$$\|x^t - t\|_A = \min_{p \in P_{t-1}} \|x^0 - x + p(A) g^0\|_A$$

Weiter ist $g^0 = Ax^0 - b = A(x^0 - x)$, das heißt

$$\|x^t - x\|_A = \min_{p \in P_{t-1}} \|(I + Ap(A))(x^0 - x)\|_A \leq \min_{p \in P_{t-1}} \|I + Ap(A)\|_A \|x^0 - x\|_A$$

$$= \min_{\substack{p \in P_t \\ p(0)=1}} \|p(A)\|_A \|x^0 - x\|_A \leq \|\tilde{p}(A)\|_A \|x^0 - x\|_A$$

Es sei nun $\{w_1, \cdots, w_n\}$ eine Orthonormalbasis aus Eigenvektoren von A. Für $y \in \mathbb{R}^n$ gilt dann $\gamma = \sum_{i=1}^n \gamma_i w_i, \gamma_i = \langle y, w_i \rangle$, also

$$\|\tilde{p}(A)y\|_A^2 = \|\sum_{i=1}^n \gamma_i \tilde{p}(A)w_i\|_A^2 = \|\sum_{i=1}^n \gamma_i \tilde{p}(\lambda_i)w_i\|_A^2$$

$$= \langle \sum_{i=1}^n \gamma_i \tilde{p}(\lambda_i)Aw_i, \sum_{i=1}^n \gamma_i \tilde{p}(\lambda_i)w_i \rangle = \sum_{i=1}^n \gamma_i^2 \lambda_i \tilde{p}(\lambda_i)^2 \leq M^2 \sum_{i=1}^n \lambda_i \gamma_i^2 = M^2 \|y\|_A^2$$

$$\Rightarrow \|\tilde{p}(A)\|_A = \sup_{\|y\|_A=1} \|\tilde{p}(A)y\| \leq M$$

112 **Definition: „Čebyšëv-Polynome“**
Die Polynome T_m: $T_m(x) := \cos(m \arccos(x))$, $m \in \mathbb{N}$, $|x| \leq 1$ heißen *Čebyšëv-Polynome*.

113 **Lemma: Eigenschaften der Čebyšëv-Polynome**

1. Die Čebyšëv-Polynome genügen der Rekursion $T_0(x) = 1, T_1(x) = x, T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x)$.
2. Für $|x| \geq 1$ gilt $T_m(x) = \cosh(m \operatorname{arcosh}(x))$.
3. $T_m(x) = \frac{1}{2}((x + \sqrt{x^2-1})^m + (x - \sqrt{x^2-1})^m)$
4. Die $\{T_m\}$ bilden ein Orthogonalsystem bezüglich des Skalarprodukts

$$\langle v, w \rangle := \int_{[-1,1]} (1-x^2)^{-\frac{1}{2}} vw \mathrm{d}x$$

114 **Lemma: Beschränkung**
Es sei $0 < \lambda < \Lambda$ und H: $P_m \to \mathbb{R}$ mit $H(p) = \sup_{\lambda \leq \zeta \leq \Lambda} |p(\zeta)|$. Ferner sei

$$p_m(\zeta) = T_m\left(\frac{1 + \lambda - 2\zeta}{\Lambda - \lambda}\right) C_m^{-1} \qquad C_m := T_m\left(\frac{1+\lambda}{1-\lambda}\right)$$

Dann gilt $H(p_m) = \min_{\substack{p \in P_m \\ p(0)=1}} H(p)$ sowie $\max_{\lambda \leq \zeta \leq \Lambda} |p_m(\zeta)| = C_m^{-1}$.

Beweis (114) Es ist $\frac{\Lambda+\lambda}{\Lambda-\lambda} > 1$, also nach Lemma Lemma 113 (Eigenschaften der Čebyšëv-Polynome) $C_m \neq 0$. Ferner ist $p_m(0) = 1$ und $p_m \in P_m$. Für $\lambda \leq \zeta \leq \Lambda$ ist

$$x = \frac{1 + \lambda - 2\zeta}{\Lambda - \lambda} \in [-1, 1] \Rightarrow |T_m(x)| \leq 1 \quad \wedge \quad T_m(x_\nu) = (-1)^m$$

für $x_\nu = \cos(\nu \frac{\pi}{m})$, $\nu = -m, 1-m, \cdots, 0$ paarweise verschieden.

$$\Rightarrow |T_m(x_\nu)| = 1 \Rightarrow \max_{\lambda \leq \zeta \leq \Lambda} |p_m|(\zeta) = C_m^{-1}$$

Minimaleigenschaft: Es sei $q_m \in P_m$ mit $q_m(0) = 1$ und $\max_{\lambda \le \zeta \le \Lambda} |q_m|(\zeta) \le C_m^{-1}$ sowie $\zeta(x) = \frac{1}{2}(\Lambda + \lambda - x(\Lambda - \lambda))$

$$\Rightarrow p_m(\zeta(x_\nu)) = T_m(x_\nu) C_m^{-1} = (-1)^\nu C_m^{-1}$$
$$\Rightarrow |q_m(\zeta(x_\nu))| \le \frac{1}{C_m} = |p_m(\zeta(x_\nu))|$$

Es sei $r := p_m - q_m \Rightarrow r(\zeta(x_\nu)) \ge 0$ für ν gerade und $r(\zeta(x_\nu)) \le 0$ für ν ungerade. Nach dem Zwischenwertsatz gibt es dann ein $\xi_\nu \in [\zeta(x_{\nu-1}), \zeta(x_\nu)]$ mit $r(\xi_\nu) = 0$. Ist $r(\zeta(x_{\nu-1})) = 0$ und $r(\zeta) \ne 0$ für $\zeta \in (\zeta(x_{\nu-1}), \zeta(x_\nu))$. Also ist $\zeta(x_{\nu-1})$ eine doppelte Nullstelle und somit hat r mindestens m Nullstellen (inklusive Vielfachheit).

115 **Satz: weitere Fehlerabschätzung**

1. Für das CG-Verfahren gilt die Fehlerabschätzung

$$\|x^t - x\|_A \le 2 \left(\frac{1 - \frac{1}{\sqrt{\kappa}}}{1 + \frac{1}{\sqrt{\kappa}}} \right)^t \|x^0 - x\|_A$$

   mit der Spektralkondition $\kappa = \mathrm{cond}_2(A) = \frac{1}{\lambda}$ von A.

2. Zur Reduzierung des Anfangsfehlers um den Faktor ε sind höchstens $t(\varepsilon) < \frac{1}{2}\sqrt{\kappa}\ln(\frac{2}{\varepsilon})$ Iterationsschritte erforderlich.

Beweis (115) Es sei λ der kleinste und Λ der größte Eigenwert von A. Nach Lemma 114 (Beschränkung) folgt damit $\|x^t - x\|_A \le C_t^{-1} \|x^0 - x\|_A$

$$\frac{\Lambda - \lambda}{\Lambda - \lambda} = \frac{\kappa + 1}{\kappa - 1}, \qquad \frac{\kappa + 1}{\kappa - 1} \pm \sqrt{\left(\frac{\kappa + 1}{\kappa - 1}\right)^2 - 1} = \cdots = \frac{\sqrt{\kappa} \pm 1}{\sqrt{\kappa} \mp 1}$$
$$C_t = T_t(\frac{\kappa + 1}{\kappa - 1}) = \frac{1}{2}\left(\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^t + \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^t \right) \ge \frac{1}{2}\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^t$$

und somit haben wir die erste Behauptung gezeigt. Der zweite folgt wegen:

$$2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{t(\varepsilon)} < \varepsilon$$
$$\Rightarrow t(\varepsilon) > \ln(\frac{2}{3}) \ln\left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^{-1}$$
$$\ln\left(\frac{x + 1}{x - 1} \right) = 2\left(\frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \cdots \right) \ge \frac{2}{x}$$
$$\Rightarrow t(\varepsilon) < \frac{1}{2}\sqrt{\kappa}\ln(\frac{2}{\varepsilon})$$

116 **Bemerkung: Vergleich zum Gradientenverfahren**
Wegen $\kappa = \text{cond}_2(A) > 1$ ist $\sqrt{\kappa} < \kappa$. Ferner ist

$$f(\lambda) = (1 - \lambda^{-1})(1 + \lambda^{-1})$$

streng monoton wachsend für $\lambda > 0$. Also gilt

$$f(\sqrt{\kappa}) < f(\kappa) \Rightarrow \frac{1 - \frac{1}{\sqrt{\kappa}}}{1 + \frac{1}{\sqrt{\kappa}}} < \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}}$$

Also sollte das CG-Verfahren schneller konvergieren als das Gradienten-Verfahren.

5.3 CG-Verfahren für allgemeine Gleichungssysteme

Betrachtet wird das Gleichungssystem $Ax = b$ wobei $A \in \mathbb{R}^{n \times n}$ regulär aber nicht notwendigerweise positiv definit bzw. symmetrisch ist. Multiplikation mit A^T liefert

$$A^T Ax = A^T b$$

mit positiv definiter und symmetrischer Matrix $A^T A$. Das CG-Verfahren hierauf angewendet liefert ausgehend von einem $x^0 \in \mathbb{R}^n$ mit $d^0 = A^T(b - Ax^0)$ die Iterationen

$$\alpha_t = \frac{\|g^t\|^2}{\|Adt\|^2}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A^T A d^t$$
$$\beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t$$

und die *Konvergenzgeschwindigkeit* ist gegeben durch $\kappa = \text{cond}_2(A^T A)$.

5.4 Vorkonditionierung

Ziel ist die Reduktion der Konditionszahl κ.

Idee: Wir formen äquivalent in ein anderes LGS $\tilde{A}\tilde{x} = \tilde{b}$ um, sodass $\text{cond}_2(\tilde{A})$ möglichst klein ist. Es sei $C = KK^T$ (positiv definit und symmetrisch) mit regulärem $K \in \mathbb{R}^{n \times n}$. Damit formen wir dann das LGS um zu:

$$\underbrace{K^{-1}A(K^T)^{-1}}_{:=\tilde{A}} \underbrace{K^T x}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}$$

Daraus erhält man das CG-Verfahren mit Startvektor $\tilde{x}^0 \in \mathbb{R}^n$, $\tilde{d}^0 = -\tilde{g}^0 = \tilde{b} - \tilde{A}\tilde{x}^0$ und Iterationen

$$\alpha_t = \frac{\|\tilde{g}\|^2}{\langle \tilde{d}^t, \tilde{A}\tilde{d}^t \rangle}, \quad \tilde{x}^{t+1} = \tilde{x}^t + \alpha_t \tilde{d}^t, \quad \tilde{g}^{t+1} = \tilde{g}^t + \alpha_t \tilde{A}\tilde{d}^t$$
$$\beta_t = \frac{\|\tilde{g}^{t+1}\|^2}{\|\tilde{g}^t\|2}, \quad \tilde{d} = -\tilde{g}^{t+1} + \beta_t \tilde{d}^t$$

Bezogen auf x^t erhält man das *PCG-Verfahren*:

$$x^0 \in \mathbb{R}^n, \quad g^0 = Ax^0 - b, \quad Cp^0 = g^0, \quad d^0 = -p^0$$
$$\alpha_t = \frac{\langle g^t, p^t g^t \rangle}{\langle d^t, Ad^t \rangle}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A d^t$$
$$Cp^{t+1} = g^{t+1}, \quad \beta_t = \frac{\langle g^{t+1}, p^{t+1} \rangle}{\langle g^t, p^t \rangle}, \quad d^{t+1} = p^{t+1} + \beta_t d^t$$

117 **Bemerkung: Aufwand**

1. In jedem Schritt ist ein Gleichungssystem mit der Matrix $C = KK^T$ zu lösen.
2. K sollte „möglichst einfach“ sein, z.B. eine Diagonalmatrix oder mit einer Dreiecksmatrix K. (In diesem Fall löst man durch Vorwärts- und Rückwärtseinsetzen)

Beispiele für Vorkonditionierung

Skalierung $A = D + L + R, R = L^T \Rightarrow C = D, K = D^{\frac{1}{2}} \Rightarrow \tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

SSOR $C = (D + \omega L)D^{-1}(D + \omega R) = \underbrace{(D^{\frac{1}{2}} + \omega L D^{-\frac{1}{2}})}_{K} \underbrace{(D^{\frac{1}{2}} + \omega D^{-\frac{1}{2}} R)}_{K^T}$

ICCG (Incomplete Cholesky Conjugate Gradient)

Cholesky-Zerlegung: $A = LL^T$ mit unterer Dreiecksmatrix

$$L_{ii} = \left(A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 \right)^{\frac{1}{2}}$$
$$L_{ji} = \frac{\left(A_{ji} - \sum_{k=1}^{i-1} L_{jk} L_{ik} \right)}{L_{ii}}$$

Typischerweise ist A schwach besetzt.

Idee: Setze $\tilde{L}_{ji} = 0$ falls $A_{ji} = 0$ und somit ist $A = \tilde{L}\tilde{L}^T + E$ und die Vorkonditionierung $C = KK^T = \tilde{L}\tilde{L}^T$.

6 Finite Elements Method

6.1 Triangulations and Finite Element Spaces

Let Ω a polygonal domain, e.g. $(0, 1)^2$ or an L-shaped domain 

Ω has to be open and connected — not necessarily simply connected, but the border $\partial\Omega$ has to be the union of finite many closed polygons.

Let the triangle $T := \operatorname{conv}\{A, B, C\}$ be the convex closure of the *vertices* A, B and C with the *edges* $a := \operatorname{conv}\{B, C\}$, $b := \operatorname{conv}\{A, C\}$ and $c := \operatorname{conv}\{A, B\}$. We assume that T is *nondegenerate*. That means its area $|T|$ shall not be 0.

Given a triangle T, $\mathcal{N}(T) := \{A, B, C\}$ is the set of nodes and $\mathcal{E}(T) := \{a, b, c\}$ is the set of edges.

118 **Definition: „Triangulation“**

$\mathcal{T}$ is called regular *triangulation* of Ω (polygonal domain as above) if $\mathcal{T}$ is a set of (closed & nondegenerate) triangles such that

$$\Omega = \bigcup \mathcal{T}$$

and for any two triangles of the triangulation $T_1, T_2 \in \mathcal{T}$ exactly one of the following cases holds:

1. They are disjoint: $T_1 \cap T_2 = \varnothing$
2. They are one and the same: $T_1 = T_2$
3. They have (a) common edge(s): $T_1 \cap T_2 = \mathcal{E}(T_1) \cap \mathcal{E}(T_2)$.
4. They have a common node: $T_1 \cap T_2 = \mathcal{N}(T_1) \cap \mathcal{N}(T_2)$

119 **Definition: „nodes and edges of a triangulation“**

For triangulations $\mathcal{T}$ we use the same symbols for nodes and edges as for triangles:

$$\mathcal{N}(\mathcal{T}) := \bigcup_{T \in \mathcal{T}} \mathcal{N}(T)$$

$$\mathcal{E}(\mathcal{T}) := \bigcup_{T \in \mathcal{T}} \mathcal{E}(T)$$

6.1.1 Data Structures

We need at least two files c4n (coordinates for nodes) and n4e (node for elements) to define $\mathcal{N} = \{Z_1, \cdots, Z_n\}$ and $\mathcal{T}$. We choose c4n $\in \mathbb{R}^{n \times 2}$ with each row describing one point of the euclidean plane and n4e $\in \{1, \cdots, n\}^{m \times 3}$ describing the triangles of the triangulation in counter clockwise order. We call the edge between the nodes in the first and second column of n4e refinement edge which is often but not necessarily the longest of the three edges.

Example of a triangulation We could describe and triangulate a L-shaped domain with:

$$\text{c4n} = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \\ 1 & 0 \\ 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} \\ 0 & 1 \\ \frac{1}{2} & 1 \end{pmatrix} \qquad \text{n4e} = \begin{pmatrix} 5 & 1 & 2 \\ 4 & 2 & 3 \\ 2 & 4 & 5 \\ 1 & 5 & 6 \\ 8 & 6 & 5 \\ 6 & 8 & 7 \end{pmatrix}$$

Exercise: Given the matrices c4n and n4e write a Matlab program to check whether or not this defines a regular triangulation and if so, compute the outer boundary $\partial\Omega$.

Remark / Fun-Exercise: A cube $(0,1)^3$ can be refined into 5 tetrahedra by planes through the cube vertices that share no common edge.

120 **Definition: „nodal basis functions“**

For a triangulation $\mathcal{T}$ we define a *nodal basis function* for each node $z \in \mathcal{N}(\mathcal{T})$ as the unique continuous function which is Dirac δ_z on $\mathcal{N}(\mathcal{T})$ and affine on each $T \in \mathcal{T}$. We note it as φ_z. The support $\operatorname{supp}\varphi_z = \operatorname{cl}\left\{ x \in \Omega \;\middle|\; \varphi_z(x) > 0 \right\}$ is called the *patch of z*.

121 **Bemerkung: barycentric coordinates**

For a triangle $T = \operatorname{conv}\left\{z_\alpha, z_\beta, z_\gamma\right\}$ we know

$$\forall x \in T \exists!\, (\lambda_\alpha, \lambda_\beta, \lambda_\gamma) \in [0,1]^3 \lambda_\alpha + \lambda_\beta + \lambda_\gamma = 1 \wedge x = \lambda_\alpha z_\alpha + \lambda_\beta z_\beta + \lambda_\gamma z_\gamma$$

then $(\lambda_\alpha, \lambda_\beta, \lambda_\gamma)$ are called the *barycentric coordinates* of x.

Indeed as functions of x: $\lambda_\alpha = \varphi_{z_\alpha}$, $\lambda_\beta = \varphi_{z_\beta}$ and $\lambda_\gamma = \varphi_{z_\gamma}$ — the barycentric coordinates describe the nodal basis functions on T.

122 **Lemma: direct form of the nodal basis functions**

We can calculate φ_α for $x \in \operatorname{conv}\{\alpha, \beta, \gamma\}$ directly as

$$\varphi_\alpha(x) = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_1 & (z_\beta)_1 & (z_\gamma)_1 \\ x_2 & (z_\beta)_2 & (z_\gamma)_2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ (z_\alpha)_1 & (z_\beta)_1 & (z_\gamma)_1 \\ (z_\alpha)_2 & (z_\beta)_2 & (z_\gamma)_2 \end{vmatrix}} \overset{\text{write as}}{=:} \frac{\begin{vmatrix} 1 & 1 & 1 \\ x & z_\beta & z_\gamma \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ z_\alpha & z_\beta & z_\gamma \end{vmatrix}}$$

(this formula can be generalized for any finite dimension)

Beweis (122**)** Because both sides of the above equation are affine on T it is sufficient to show that they are equal at the vertices. That is obviously true. The denominator is not 0 because T is not degenerated and it is

$$\begin{vmatrix} 1 & 1 & 1 \\ z_\alpha & z_\beta & z_\gamma \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ z_\alpha & z_\beta - z_\alpha & z_\gamma - z_\alpha \end{vmatrix} = \begin{vmatrix} z_\beta - z_\alpha & z_\gamma - z_\alpha \end{vmatrix} = 2|T|$$

because α, β and γ are in counterclockwise order.

123 **Lemma: derivative of node basis function**

$$\nabla\varphi_\alpha(x) = \frac{1}{2|T|}\begin{pmatrix} \begin{vmatrix} 0 & 1 & 1 \\ 1 & & \\ 0 & z_\beta & z_\gamma \end{vmatrix} & \begin{vmatrix} 0 & 1 & 1 \\ 0 & & \\ 1 & z_\beta & z_\gamma \end{vmatrix} \end{pmatrix}$$

124 **Definition: „stiffness matrix“**

For a triangle $T = \operatorname{conv}\{P_1, P_2, P_3\}$ with nodal basis functions $\varphi_k := \varphi_{P_k}, k = 1,2,3$ we define the *stiffness matrix* $\mathrm{STIMA}(T) \in \mathbb{R}^{3\times 3}$ entry wise:

$$\mathrm{STIMA}(T)_{k,l} := \int_T \nabla\varphi_k(x)\nabla\varphi_l(x)^T \,\mathrm{d}\lambda_2(x)$$

125 **Lemma: representation of the stiffness matrix**

We can represent the local stiffness matrix as

$$\mathrm{STIMA}(T) = |T|QQ^T, \quad Q = \nabla\varphi := \begin{pmatrix}\nabla\varphi_1 \\ \nabla\varphi_2 \\ \nabla\varphi_3\end{pmatrix} \overset{!}{=} \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ P_1 & P_2 & P_3 \end{pmatrix}^{-1}}_{\in\mathbb{R}^{3\times 3}} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2}$$

126 **Definition: „global stiffness matrix“**

Just as we defined the local stiffness matrix for the three nodes of a triangle we can now define the *global stiffness matrix* for all nodes of a triangulation $\mathcal{T}$.

$$\mathrm{STIMA}(\mathcal{T}) := \left(\int_\Omega \nabla\varphi_\alpha(x)\nabla\varphi_\beta(x)^T \,\mathrm{d}\lambda_2(x)\right)_{\alpha,\beta=1,\cdots,N}$$

if we enumerate the nodes like in c4n from 1 to $N = |\mathcal{N}(\mathcal{T})|$.

Bemerkung Entries, where α and β share no common edge, are 0.

Because of the integrals set additivity we can compute a very sparse global stiffness matrix $A(T)$ for each triangle $T \in \mathcal{T}$ by only integrating over $T \subset \Omega$. (For α, β nodes of T we have $A(T)_{\alpha,\beta} = \mathrm{STIMA}(T)_{\alpha',\beta'}$ with some $\alpha', \beta' \in \{1,2,3\}$ while the other entries are 0) We get the complete matrix as

$$\mathrm{STIMA}(\mathcal{T}) = \sum_{T\in\mathcal{T}} A(T)$$

MATLAB-realization

```
STIMA = sparse(N, N);
for T = n4e
  STIMA += A(T)
end
```

6.2 FEM for Poisson-Model-Problem

$$-\Delta u = f \text{ in } \Omega \quad \wedge \quad u = 0 \text{ on } \partial\Omega \tag{6-1}$$

Given Ω polygonal domain as above. $f \in L^2(\Omega)$, $\Delta u := \operatorname{div}\operatorname{grad} u = \left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\operatorname{grad} u$ and $\operatorname{grad} u := \nabla u$. We seek a solution $u: \Omega \to \mathbb{R}$ of 6-1. If u is smooth enough, then

$$\Delta u(x) = \frac{\partial^2 u(x)}{\partial x_1^2} + \frac{\partial^2 u(x)}{\partial x_2^2}$$

One can show that 6-1 is equivalent to the *weak formulation*

$$\forall v \in C^1(\operatorname{cl}\Omega) \cap C_0(\Omega) a(u,v) = F(v)$$

with the following definitions:

$$a(u,v) := \int_\Omega \langle \nabla u(x), \nabla v(x)\rangle \mathrm{d}x$$
$$F(x) := \int_\Omega f(x)v(x)\mathrm{d}x$$
$$C_0(A) := \left\{ f \in C(A) \;\middle|\; \partial\Omega \subset f^{-1}(0) \right\}$$

Since the scalar product a on $C^1(\operatorname{cl}\Omega) \cap C_0(\Omega)$ does not lead to a Hilbert space. PDE theory considers completion of the space, i.e. $V = \operatorname{cl}_{|||\cdot|||}\left(C^1(\operatorname{cl}\Omega) \cap C_0(\Omega)\right)$ with $|||\cdot||| := \sqrt{a(\cdot,\cdot)}$.

a is a symmetric positive-semidefinite bilinear form and $a(u,u) = 0 \Rightarrow \int_\Omega |\nabla u|^2 \mathrm{d}x = 0 \Rightarrow \Delta u = 0$ because Δu is continuous. That means such u are constant and with $u \in C_0(\Omega)$ we know $u = 0$.

Indeed: Friedrichs inequality says:

$$\forall v \in C^1(\operatorname{cl}\Omega) \cap C_0(\Omega) \|v\|_{L^2(\Omega)} := \sqrt{\int_\Omega v(x)^2 \mathrm{d}x} \le \frac{\operatorname{diam}(\Omega)}{\pi} |||v|||$$

127 **Definition: „Sobolev space“**

$$H_0^1(\Omega) := V = \operatorname{cl}_{|||\cdot|||}\left(\mathcal{L}^1(\Omega)\right)$$

is a *sobolev space*.

128 **Satz: Riesz representation theorem**

$$\forall F \in V^* \exists! \, u \in V \, a(u,\cdot) = F \text{ in } V$$

In 6-1, $F \in V^*$ because Friedrichs inequality holds in V. u is called a *weak solution*.

129 **Satz: Existence, uniqueness, Galerkin orthogonality, best approximation**

In above notation (V, a) is a Hilbert space. Let $V_l \subset V$ be a closed linear subspace. Given $F \in V^*$

1. - There exists a *weak solution* of 6-1 u i.e. $\exists! u \in V \forall v \in V\, a(u,v) = F(v)$
   - There exists a *discrete solution* of 6-1 u i.e. $\exists! u_l \in V_l \forall v_l \in V_l a(u_l, v_l) = F(v_l)$
2. $e := u - u_l$ satisfies *Galerkin orthogonality*, i.e. $e \perp_a V_l \Leftrightarrow \forall v_l \in V_l a(v_l, e) = 0$
3. $|||u - u_l||| = \min_{v_l \in V_l} |||u - v_l||| =: \operatorname{dist}(u, V_l)$

Beweis (129)

1. The first existence follows from Satz 128 (Riesz representation theorem) where u is the Riesz representation of F. $\Lambda: V \to V^*, v \mapsto a(v, \cdot)$ is a topological and algebraical isomorphism because (V, a) is a Hilbert space.

   The second existence follows with $(V_l, a/V_l \times V_l)$ (is Hilbert space because V_l is closed) and $F|_{V_l} \in V_l^*$.
2. $\forall v_l \in V_l \subset V\, a(e, v_l) = \underbrace{a(u, v_l)}_{F(v_l)} - \underbrace{a(u_l, v_l)}_{F|_{V_l}(v_l)} = 0$
3. $|||e|||^2 = a(e, u - u_l) = a(e, u) - a(e, u_l) \overset{2.}{=} a(e, u - v_l) \leq |||e||| \cdot |||u - v_l|||$
   $\Rightarrow \forall v_l \in V_l |||e||| \leq |||u - v_l|||$

130 **Definition: „vector space for triangulation“**

We define $V(\mathcal{T}) := V_l = P_1(\mathcal{T}) \cap C_0(\Omega)$ for a regular triangulation $\mathcal{T}$ of Ω.

PDE theory shows $V(\mathcal{T})$ is a Lipschitz continuous function in H^1 and so $V_l \subset V$.

It is $m := \dim V_l = |\mathcal{N}(\Omega)| < \infty$, that means $V_l \subset V$ is a closed subspace and we can write the discrete solution as

$$u_l = \sum_{z \in \mathcal{N}(\Omega)} u_l(z) \varphi_z$$

with some coefficients $u_l(z)$ and the nodal basis functions φ_z. Let's suppose, that $\mathcal{N} = \{z_1, \cdots, z_m\}$ and $\varphi_i := \varphi_{z_i}$ as well as $x_i := u_l(z_i)$. Then we can see:

$$\begin{aligned}
& \forall v_l \in V_l && a(u_l, v_l) = F(v_l) && \\
\Leftrightarrow\ & \forall k = 1, .., m && a(u_l, \varphi_k) = F(\varphi_k) && \text{because } (\varphi_k) \text{ is a basis of } V_l \\
\Leftrightarrow\ & \forall k = 1, \cdots, m && \sum_{j=1}^{m} x_j \underbrace{a(\varphi_j, \varphi_k)}_{=: A_{jk}} = \underbrace{F(\varphi_k)}_{=: b_k} && \\
\Leftrightarrow\ & && Ax = b \quad \text{for } x = (x_1, \cdots, x_m)^T &
\end{aligned}$$

$$b_k = \int_\Omega \varphi_k \mathrm{d}x = \sum_{\substack{T\in\mathcal{T}\\ T\subset\operatorname{supp}(\varphi_k)}} \int_T \varphi_k \mathrm{d}x = \sum_{\substack{T\in\mathcal{T}\\ T\subset\operatorname{supp}(\varphi_k)}} \frac{|T|}{3}$$

131 **Bemerkung: regularity**

A is symmetric and positive-definite and hence the coefficients vector $x = A^{-1}b$ defines the discrete solution $u_l = \sum_{j=1}^m x_j \varphi_j$.

6.3 A Priori Error Analysis

Integration by parts in $\mathbb{R}^1$:

$$[f(x)g(x)]_a^b = \int_a^b (f(x)g(x))' \mathrm{d}x = \int_a^b f(x)g'(x) + f'(x)g(x) \mathrm{d}x$$

Let Ω be a domain in $\mathbb{R}^2$ with $\partial\Omega$ piecewise smooth with outer unit normal $\nu\colon \mathbb{R}^2 \to \mathbb{R}^2$, $f, g \in C^1(\mathbb{R}^2)$

$$\int_\Omega \frac{\partial f}{\partial x_j} g + f \frac{\partial g}{\partial x_j} \mathrm{d}\lambda_2 = \int_{\partial\Omega} (fg)(x)\nu_j(x) \mathrm{d}\lambda_1(x)$$

132 **Lemma: Trace-Identity**

Let $T = \operatorname{conv} P, E_1, E_2$ be a triangle with edge E and vertex P and let $f \in C^1(T)$.

$$⨍_E f \mathrm{d}s = ⨍_T f \mathrm{d}x + \frac{1}{2} \int_T (x - P) \cdot \Delta f(x) \mathrm{d}x$$

where $⨍_M g(x)\mathrm{d}\mu(x) = \frac{1}{\mu(E)} \int_E g(x)\mathrm{d}\mu(x)$ is an *integral mean*.

133 **Bemerkung: bounding**

Since $F\colon C^1(\mathbb{R}^2) \to C^1, f \mapsto f|_E$ is bounded w.r.t.

$$\|\cdot\|_{H(T)} := \sqrt{\|\cdot\|^2_{L^2(T)} + \|\nabla\cdot\|^2_{L^2(T)}}$$

and $L^2(E)$ the existence of traces for the Sobolow functions can be proven (PDE theory).

134 **Definition: „nodal interpolant“**

Given $v \in C_0(\Omega)$ we define the *nodal interpolant* as

$$Iv := \sum_{z\in\mathcal{N}(\mathcal{T})} f(z)\varphi_z$$

in $V(\mathcal{T})$ with

$$\forall z \in \mathcal{N}\, v(z) = (Iv)(z)$$

135 **Bemerkung:**

The Galerkin orthogonality shows $\|u-u_l\| = \operatorname{dist}(u, V(\mathcal{T})) \leq |||u-Iu|||$ in case $u \in C_0(\Omega) \cap V$. For $u|T \in C^2(T) \subset H^2(T)$ we will prove $|\nabla(u-Iu)|_{L^2(T)} \leq c(T) h_T |D^2 u|_{L^2(T)}$ for some $c(T) \leq \frac{\sqrt{\frac{1}{2}+\frac{2}{\pi^2}}}{\sqrt{1-|\cos(\omega)|}}$ for some angle ω in T and $h_T = \operatorname{diam}(T)$. (where D^2 is the hessian matrix operator)

136 **Lemma: A**

$T = \operatorname{conv} p_1, p_2, p_3$, $u \in C^2(T)$, $Iu \in P_1(T)$ the nodal interpolant. Let $h_j := |p_{j+2} - p_{j+1}|$, $\mathcal{T}_j := \frac{p_{j+2}-p_{j+1}}{h_j}$ and $f_j := \mathcal{T}_j \nabla(u - Iu)$ then

$$\|f_j\|_{L^2(T)} \leq h_T \sqrt{\frac{1}{8} + \frac{1}{\pi^2}} \|\nabla f_j\|_{L^2(T)}$$

Beweis (136) Let $\bar{f}_j := \fint_T f_j(x)\mathrm{d}x$ then

$$\|f_j\|^2_{L^2(T)} = \|f_j - \bar{f}_j\|^2_{L^2(T)} + \|\bar{f}_j\|^2_{L^2(T)} + 2\int_T (f_j - \bar{f}_j)\mathrm{d}x \ = 0$$

because

$$\int_T f_j - \bar{f}_j \mathrm{d}x \ = \int_T f_j \mathrm{d}x \ - |T| \bar{f}_j = \int_T f_j \mathrm{d}x \ - |T| \fint_T f_j \mathrm{d}x \ = 0$$

Poincaré inequality with Payne-Weinberger-constant

$$\|f_j - \bar{f}_j\|_{L^2(T)} = \frac{h_T}{\pi} \|\nabla f_j\|_{L^2(T)}$$

since

$$\int_{E_j} f_j \mathrm{d}s \ = \int_{E_j} \tau_j \cdot \nabla(u - Iu)\mathrm{d}s \ = (u - Iu)\frac{P_{j+2}}{n_j} - (u - Iu)\frac{P_{j+1}}{n_j} = 0$$

The trace identity shows

$$|\bar{f}_j| = \frac{1}{2} \frac{|\int_T (x - z_j)\nabla f_j(x)\mathrm{d}x \,|}{|T|} \leq \frac{1}{2} \frac{\|(\cdot - p_j)\|_{L^2(T)} \|\nabla(\tau_j \nabla(u - Iu))\|_{L^2(I)}}{|T|}$$
$$\leq \frac{\sqrt{|T|} h_T}{2\sqrt{2}} \cdot \frac{\|D^2(u - Iu)\|_{L^2(T)}}{|T|} = |T|^{-\frac{1}{2}} \frac{h_T}{\sqrt{8}} \|D^2 u\|_{L^2(T)}$$

($D^2 Iu = 0$ because Iu is affine.)

137 **Lemma: B**

$$\forall a \in \mathbb{R}^2 \forall \nu, \mu \in \mathbb{R}^2, |\mu| = |\mu| = 1, -1 < \nu \cdot \mu < 1 |a|^2 \leq \frac{(a \cdot \nu)^2 + (a \cdot \mu)^2}{1 - |\nu \cdot \mu|}$$

(the constant $\frac{1}{1-|\nu\cdot\mu|}$ is optimal.)

Beweis (137) Let $a = \alpha\nu + \beta\mu$ with $\alpha, \beta \in \mathbb{R}$ and $\alpha^2 + \beta^2 = 1$. Set $\gamma := \mu \cdot \nu$, then we know

$$
\begin{aligned}
&-1 \le 2\alpha\beta \le 1 \\
&\Rightarrow 0 \le (1 + |\gamma|)(|\gamma| + 2\alpha\beta\gamma) \\
&\Leftrightarrow -(1 + |\gamma|)2\alpha\beta\gamma + 1 - |\gamma| + 4\alpha\beta\gamma \le |\gamma| + \gamma^2 + 1 - |\gamma| + 4\alpha\beta\gamma \\
&\Leftrightarrow (1 - |\gamma|)(1 + 2\alpha\beta\gamma) \le \alpha^2\gamma^2 + \beta^2\gamma^2 + \alpha^2 + \beta^2 + 4\alpha\beta\gamma \\
&\Leftrightarrow (1 - |\gamma|)|a|^2 \le (\alpha + \beta\gamma)^2 + (\beta + \alpha\gamma)^2 = (a \cdot \nu)^2 + (a \cdot \mu)^2
\end{aligned}
$$

138 **Satz: local interpolation error estimate**

$$\|\nabla(u - Iu)\|_{L^2(T)} \le \frac{\sqrt{\frac{1}{2} + \frac{2}{\pi^2}}}{\sqrt{1 - |\cos\omega|}} h_T \|D^2 u\|_{L^2(T)}$$

Beweis (138) $\omega = \measuredangle(E_j, E_{j+1})$. Lemma 137 (B) with $a := \nabla(u - Iu)(x)$, $\nu := \tau_j$, $\mu := \tau_{j+1}$ and $|\tau_j \cdot \tau_{j+1}| = |\cos\omega|$ gives

$$
\begin{aligned}
(1 - |\cos\omega|) \int_T |\nabla(u - Iu)(x)|^2 \mathrm{d}x &\le \int_T |\tau_j \cdot \nabla(u - Iu)|^2 \mathrm{d}x + \int_T |\tau_{j+1} \cdot \nabla(u - Iu)|^2 \mathrm{d}x \\
&\overset{\text{Lemma } 136\,(A)}{\le} h_T^2 \left(\frac{1}{8} + \frac{1}{\pi^2}\right) \cdot 2\|D^2 u\|^2_{L^2(T)}
\end{aligned}
$$

6.4 Min vs. Max angle condition

$T_1(\delta) := \mathrm{conv}\,\{(-1, 0), (1, 0), (0, \delta)\}$, $T_2(\delta) := \mathrm{conv}\,\{(0, 0), (1, 0), (0, \delta)\}$ with $0 < \delta \ll 1$
maxangle $\le \pi - c_1 < \pi$
minangle $\ge c_0 > 0$
$u(x, y) := 1 - x^2 \Rightarrow (Iu)(x, y) = \frac{y}{\delta}$ on $T_1(\delta)$ (strange: u depends only on x, interpolant depends only on y)

$$\frac{1}{\delta} \le \|\nabla(u - Iu)\|^2_{L^2(T_1)}$$
$$\|u\|^2_{H^2(T_1)} = 4\delta$$
$$\Rightarrow \frac{1}{2\delta} \le \frac{\|\nabla(u - Iu)\|_{L^2(T_1)}}{\|D^2 u\|_{L^2(T_1)}}$$
$$\cos\omega = \frac{1}{\sqrt{1+\delta^2}}$$
$$\frac{1}{\sqrt{1-|\cos\omega|}} = \frac{\sqrt[4]{1+\delta^2}}{\sqrt{\sqrt{1+\delta^2}-1}}$$

on $T_2(\delta)$:

$$Iu = 1 - x$$
$$\frac{\|\nabla(u - Iu)\|}{\|D^2 u\|} \qquad \text{bounded as } \delta \searrow 0$$

$$|||u - u_l||| \le C(\measuredangle\mathcal{T})\|h_l D^2 u\|_{L^2(\Omega)}$$

Main result with $C(\measuredangle\mathcal{T})$ depending on *max angle* in $\mathcal{T}$ and $h_l \in P_0(\mathcal{T})$ with $h_l|_T = h_t = \operatorname{diam}(T)$. That means if $D^2 u$ is big we have to reduce h_l.

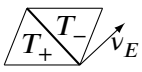
6.5 AFEM (adaptive FEM)

Input $\mathcal{T}_0$ coarse regular triangulation, $0 < \theta \approx \frac{1}{10} < 1$
loop $l = 0, 1, \cdots$ (until some termination)

SOLVE compute *FE* solution u_l w.r.t. mesh $\mathcal{T}_l$.

ESTIMATE compute $\eta^2(T) := |T|\|f\|^2_{L^2(T)} + \sqrt{|T|}\left\|\left[\frac{\delta u_l}{\delta \nu_E}\right]_E\right\|^2_{L^2(\delta T)}$ where

$$\left[\frac{\delta u_l}{\delta \nu_E}\right]_E := \begin{cases} 0 & \text{if } E \subset \delta\Omega \\ (\nabla u_l|_{T_+} - \nabla u_l|_{T_-}) \cdot \nu_E & \text{otherwise} \end{cases}$$

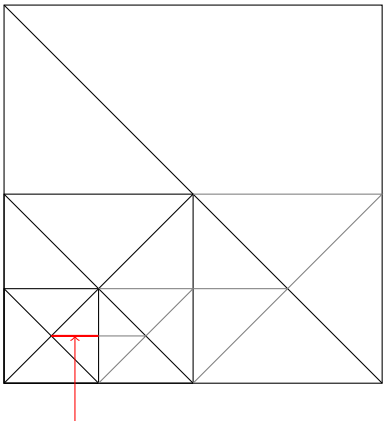


MARK choose $\mathcal{M}_l \subset \mathcal{T}_l$ of minimal cardinality such that $\theta \sum_{T\in\mathcal{T}_l} \eta_l^2(T) \le \sum_{M\in\mathcal{M}_l} \eta_l^2(M)$

REFINE bisect all triangles in $\mathcal{M}_l$ and avoid hanging nodes to define $\mathcal{T}_{l+1}$

end loop

example



allow only right-angled triangles
→ must bisect many other triangles to avoid hanging node!
$|\mathcal{T}_{l+1}| - |\mathcal{T}_l| \not\leq |M_l|$

new bisection
the bisected triangle is the only element in $\mathcal{M}_l$

139 **Satz: optimality**

AFEM is optimal in the sense that if data and solution can be approximated with a certain rate in terms of *ndof*, then AFEM produces the same convergence speed.

7 Eigenwertaufgabe

7.1 Grundlagen

Es sei $A \in \mathbb{K}^{n\times n}, \mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

- *Eigenwert*: $\lambda \in \mathbb{C}$ sodass $\exists \omega \in \mathbb{C}^n, \omega \neq 0 A\omega = \lambda\omega$ (ω heißt dann *Eigenvektor*)
- *charakteristisches Polynom*: $p_A(z) := \det(A - zI_n)$ insbesondere $p_A(\lambda) = 0$ für Eigenwerte λ.
- Bestimmung der Eigenvektoren: Ist der Eigenwert λ bekannt, so können die Eigenvektoren als Lösung der Gleichung $(A - \lambda I)\omega = 0$ bestimmt werden.
- Rayleigh-Quotient: $\lambda = r(\omega) = \frac{\langle A\omega,\omega\rangle}{\|\omega\|_2^2}$ falls der Eigenvektor ω bekannt ist.
- geometrische Vielfachheit: $\varrho_i := \dim \mathrm{Ker}(A - \lambda_i I)$
- *geometrischer Eigenraum* zu einem Eigenwert λ: $\mathrm{Ker}(A - \lambda I)$.
- algebraische Vielfachheit von: σ_i, sodass $p_A(z) = (z - \lambda)^{\sigma_i} q(z)$, $q(\lambda) \neq 0$
- Es gilt stets $\varrho_i \leq \sigma_i$.

7.2 Verteilung der Eigenwerte

140 **Lemma: Notwendige Bedingung für Eigenwerte**

Es seien $A, B \in \mathbb{K}^{n\times n}$, $\|\cdot\|$ natürlich Matrixnormen (also induzierte Operatornormen). Ferner sei λ Eigenwert von A, aber nicht Eigenwert von B. Dann gilt

$$\|(\lambda I - B)^{-1}(A - B)\| \geq 1$$

Beweis (140) Es ist $(A-B)\omega = (\lambda I - B)\omega$ mit ω Eigenvektor von A zu λ. Da λ kein Eigenwert von B ist gilt $\det(\lambda I - B) \neq 0 \Leftrightarrow \lambda I - B$ ist regulär.

$$\Rightarrow \omega = (\lambda I - B)^{-1}(A - B)\omega$$

$$\Rightarrow 1 = \frac{\|\omega\|}{\|\omega\|} = \frac{\|(\lambda I - B)^{-1}(A - B)\omega\|}{\|\omega\|} \leq \sup_{x\in\mathbb{C}\setminus\{0\}} \frac{\|(\lambda I - B)^{-1}\|(A - B)x}{\|x\|} = \|(\lambda I - B)^{-1}(A - B)\|$$

141 **Satz: Gerschgorin**

Es sei $\lambda \in \mathbb{C}$ ein Eigenwert von $A \in \mathbb{K}^{n\times n}$. Dann gilt

$$\lambda \in \bigcup_{j=1}^{n} K_j, \qquad K_j := \left\{ z \in \mathbb{C} \;\middle|\; |z - A_{jj}| \leq \sum_{\substack{k=1 \\ k\neq j}}^{n} |A_{jk}| \right\}$$

Beweis (141) Es sei $D := \operatorname{diag}(A_{11}, \cdots, A_{nn})$ und o.B.d.A. $\forall j = 1, \cdots, n \lambda \neq A_{jj}$ dann gilt nach Lemma 140 (Notwendige Bedingung für Eigenwerte)

$$1 \leq \|(\lambda I - D)^{-1}(A - D)\|_\infty = \max_{j=1,\cdots,n} \left(\frac{1}{|\lambda - A_{jj}|} \sum_{\substack{k=1 \\ k\neq j}}^{n} |A_{jk}| \right)$$

Ist j' der Index an dem das Maximum angenommen wird, so gilt

$$|\lambda - A_{j'j'}| \leq \sum_{\substack{k=1 \\ k\neq j'}}^{n} |A_{j'k}|$$

142 **Satz: Konditionierung**

Es seien $A, B \in \mathbb{K}^{n\times n}$ und $\{\omega_1, \cdots, \omega_n\}$ linear unabhängige Eigenvektoren von A. Dann existiert zu jedem Eigenwert λ_B von B ein Eigenwert λ_A von A, sodass $|\lambda_A - \lambda_B| \leq \operatorname{cond}_2(W)\|A - B\|_2$ mit $W := (\omega_1, \cdots, \omega_n)$

Beweis (142) Es sei λ_i Eigenwert von A zum Eigenvektor ω_i, dann

$$AW = W \operatorname{diag}(\lambda_1, \cdots, \lambda_n) \Rightarrow A = W \operatorname{diag}(\lambda_1, \cdots, \lambda_n) W^{-1}$$

Es sei λ_B Eigenwert von B, aber nicht von A, dann folgt aus Lemma 140 (Notwendige Bedingung für Eigenwerte)

$$\begin{aligned}&\|(\lambda_B I - A)^{-1}\|_2 = \|W(\lambda_B I - \operatorname{diag}(\lambda_1, \cdots, \lambda_n))^1 W^{-1}\|_2 \\ &\leq \|W\|_2 \|W^{-1}\|_2 \|(\lambda_B I - \operatorname{diag}(\lambda_1, \cdots, \lambda_n))^{-1}\|_2 = \operatorname{cond}_2(W) \max_{j=1,\cdots,n} |\lambda_B - \lambda_i|^{-1}\end{aligned}$$

Sei i' der Index, an dem das Maximum angenommen wird, dann gilt

$$\begin{aligned}|\lambda_B - \lambda_{i'}| &\leq \operatorname{cond}_2(W) \frac{1}{\|(\lambda_B I - A)^{-1}\|_2} \\ &\overset{\text{Lemma 140 (Notwendige Bedingung für Eigenwerte)}}{\leq} \operatorname{cond}_2(W) \frac{\|(\lambda_B I - A)^{-1}(A - B)\|_2}{\|(\lambda_B I - A)^{-1}\|_2} \leq \operatorname{cond}_2(W) \|A - B\|_2\end{aligned}$$

7.3 Iterative Verfahren

Potenzmethode (von Mises) Startvektor $z^0 \in \mathbb{C}^n$, $\|z^0\| = 1$ mit beliebiger Norm
Iterationen $t = 1, 2, \cdots$:

$$\tilde{z}^t := A z^{t-1}, \quad z^t := \frac{\tilde{z}^t}{\|\tilde{z}^t\|}$$
$$\lambda^t := \frac{(A z^t)_k}{z_k^t}$$

mit $k \in \{1, \cdots, n\}$ beliebig, z.B. jenes wo z_k maximal ist.

143 **Satz:**
Es sei $\{\omega_1, \cdots, \omega_n\}$ eine Basis aus Eigenvektoren von A mit $\|\omega_n\| = 1$ sowie $\lambda_i \in \mathbb{C}$ die zugehörigen Eigenwerte von A mit $|\lambda_n| > |\lambda_i|$ für $i = 1, \cdots, n-1$. Ferner sei $z^0 = \sum_{i=1}^n \alpha_i \omega \neq 0$.

1. Dann existiert eine Folge $(\sigma_t)_{t \in \mathbb{N}}$, $\sigma_t \in \mathbb{C}$ mit $|\sigma_t| = 1$ sodass $\lim_{t \to \infty} \|z^t - \sigma_t \omega_n\| = 0$.
2. Es gilt

$$|\lambda^t - \lambda_n| = \mathcal{O}_{t \to \infty}\left(\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^t\right)$$

wobei λ_{n-1} betragsmäßig der zweitgrößte Eigenwert ist.

Beweis (143)

$$z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|} = \frac{Az^{t-1}}{\|Az^{t-1}\|} = \frac{A\tilde{z}^{t-1}}{\|\tilde{z}^{t-1}\|} \cdot \frac{\|\tilde{z}^{t-1}\|}{\|Az^{t-1}\|} = \cdots = \frac{(A)^t z^0}{\|(A)^t z^0\|}$$

$$A^t z^0 = \sum_{i=1}^{n} \alpha_i \lambda_j^t \omega_i = \lambda_n^t \alpha_n \left(\omega_n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n} \left(\frac{\lambda_i}{\lambda_n} \right)^t \omega_i \right)$$

Wegen $|\frac{\lambda_i}{\lambda_n}| < 1, i = 1, \cdots, n-1$ erhält man

$$A^t z_0 = \lambda_n^t \alpha_n (\omega_n + \mathcal{O}_{t\to\infty}(1))$$

$$\Rightarrow z^t = \frac{\lambda_n^t \alpha_n (\omega_n + O(1))}{|\alpha_n||\lambda_n^t|\|\omega_n + \mathcal{O}_{t\to\infty}(1)\|} = \frac{\lambda_n^t \alpha_n}{|\lambda_n^t \alpha_n|} \omega_n + \mathcal{O}_{t\to\infty}(1)$$

Dabei ist $\sigma_t = |\lambda_n^t \alpha_n|$ und der erste Teil wurde gezeigt. Zum zweiten:

$$\lambda^t = \frac{(Az^t)_k}{z_k^t} = \frac{A^{t+1} z_0}{\|A^t z^0\|} \cdot \frac{\|A^t z^0\|}{(A^t z^0)_k} = \frac{\lambda_n^{t+1} \left(\alpha_n \omega_{n,k} + \sum_{i=1}^{n-1} \alpha_i (\frac{\lambda_i}{\lambda_n})^{t+1} \omega_{i,k} \right)}{\lambda_n^t \left(\alpha_n \omega_{n,k} + \sum_{i=1}^{n-1} \alpha_i (\frac{\lambda_i}{\lambda_n})^t \omega_{i,k} \right)} = \lambda_n + \mathcal{O}_{t\to\infty} \left(|\frac{\lambda_{n-1}}{\lambda_n}|^t \right)$$

Literatur

[Ran] RANNACHER, Prof. Dr. R.: *Numerik 0/1*

[Sch79] SCHWETLICK, Horst: *Numerische Lösung nichtlinearer Gleichungen*. Deutscher Verlag der Wissenschaft, 1979. – 142 ff. S.

[Sto02] STOER, Prof. Dr. J.: *Numerische Mathematik*. 8. Springer, 2002. – 163–165 S.

[SW92] SCHABACK, Prof. Dr. R. ; WERNER, Prof. Dr. H.: *Numerische Mathematik*. 4. Springer, 1992. – 159 f. S.