

# Numerische Mathematik I

WS 2002/2003

Prof. Dr. E. P. Stephan

30. Juni 2003

# Contents

<b>1</b>	<b>Interpolation</b>	<b>3</b>
1.1	Interpolation und Approximation von Funktionen . . . . .	3
1.1.1	Interpolation durch Polynome und Dividierte Differenzen . . . . .	3
1.1.2	Interpolation bei äquidistanten Knoten, Tschebyscheff-Polynome . . . . .	7
1.1.3	Interpolationsfehler . . . . .	9
1.2	Hermite-Interpolation . . . . .	12
1.3	Stückweise polynomiale Interpolation . . . . .	14
1.4	Interpolation mit kubischen Splines . . . . .	16
<b>2</b>	<b>Lineare Gleichungssysteme, direkte Verfahren</b>	<b>25</b>
2.1	Vorbemerkungen . . . . .	25
2.2	Der Gaußsche Algorithmus . . . . .	26
2.2.1	Methoden zur Rechenstabilität . . . . .	27
2.2.2	Weitere Bemerkungen zu Gauß . . . . .	27
2.3	LGS mit positiv definiten Matrizen . . . . .	32
2.3.1	Die Cholesky-Zerlegung, Bandmatrizen . . . . .	32
<b>3</b>	<b>Vektor- und Matrixnormen, Fehlerabschätzungen...</b>	<b>36</b>
3.1	Vektor- und Matrixnormen . . . . .	36
3.1.1	Konvergenz von Folgen im $\mathbb{R}^n$ . . . . .	37
3.1.2	Matrix-Normen . . . . .	38
3.1.3	Einführung der Spektralnorm . . . . .	39
3.1.4	Beziehung zwischen Vektor- und Matrix-Normen . . . . .	40
3.1.5	Die Kondition . . . . .	42
3.2	Iterative Verfahren zur Lösung von LGS . . . . .	43
3.2.1	Die Verfahren . . . . .	43

---

3.2.2	Konvergenzuntersuchungen . . . . .	44
3.2.3	Bemerkungen zu Iterationsverfahren für LGS . . . . .	47
3.3	Nichtlineare Gleichungssysteme . . . . .	48
3.3.1	Das Banach'sche Verfahren . . . . .	48
3.3.2	Das Newton-Verfahren . . . . .	49
3.3.3	Nichtlineare GSV, ESV und SOR-Verfahren . . . . .	51
<b>4</b>	<b>Konjugierte Gradientenmethoden (CG)</b>	<b>52</b>
<b>5</b>	<b>Approximation</b>	<b>56</b>
5.1	Existenz und Eindeutigkeit der besten Approximation . . . . .	57
5.2	Approximation bei gegebener Orthonormalbasis . . . . .	59
5.3	Gram-Schmidt-Orthonormalisierung . . . . .	60
5.4	Diskrete Approximation im quadratischem Mittel . . . . .	62
<b>6</b>	<b>Numerische Quadratur</b>	<b>64</b>
6.1	Trapez- und Simpson-Regel . . . . .	64
6.2	Newton-Cotes Formeln . . . . .	66
6.3	Fehlerabschätzung mit Peano . . . . .	70
<b>7</b>	<b>Matrizen - Eigenwertaufgaben</b>	<b>73</b>
7.1	Vorbemerkungen, Eigenwertabschätzungen . . . . .	73
7.2	Die Verfahren von Wilkinson und von Householder . . . . .	75
7.3	Berechnung von EW und EV von Hessenberg-Matrizen . . . . .	79
7.4	von-Mises Verfahren . . . . .	82

# Chapter 1

## Interpolation

### 1.1 Interpolation und Approximation von Funktionen

Gegeben seien die Funktionswerte  $f(x_0), f(x_1), \dots, f(x_n)$  mit

$$f_\nu := f(x_\nu) \quad , \quad 0 \leq \nu \leq n$$

in paarweise verschiedenen Knoten  $x_0 < x_1 < \dots < x_\nu < \dots < x_n \in \mathbb{R}$ .

Gesucht ist eine Funktion  $p : [x_0, x_n] \rightarrow \mathbb{R}$  mit

$$(1.1) \quad p(x_\nu) = f_\nu \quad , \quad 0 \leq \nu \leq n$$

so daß der Fehler  $p - f$  minimal wird.

Diese Aufgabenstellung macht nur dann einen Sinn (im Bezug auf Existenz und Eindeutigkeit von  $p$ ), wenn nach einem  $p \in V$  gesucht wird, wobei  $V$  eine spezielle Menge von Funktionen ist.

Zu unterscheiden:

*Interpolation* : Finde  $p \in V$  mit (1.1) und Definitionsbereich  $[x_0, x_n]$ .

*Extrapolation* : Betrachte  $p$  ebenso in Punkten  $x \notin [x_0, x_n]$ .

Für gewöhnlich wird für  $V$  ein endlich dimensionaler, linearer Funktionenraum gewählt (Polynome, trigonometrische Funktionen, Finite Elemente). Ausnahmen bilden dabei die **rationale Interpolation**, z.B.

$$\frac{1}{1 + (e - 1)x} \text{ interpoliert } e^{-x} \text{ in } x_0 = 0, x_1 = 1$$

und die **exponentielle Interpolation**, z.B.

$$p(x) = a_1 e^{\lambda_1 x} + a_2 e^{\lambda_2 x} + \dots$$

#### 1.1.1 Interpolation durch Polynome und Dividierte Differenzen

Es sei

$$V := P_n = \{g : \mathbb{R} \rightarrow \mathbb{C}, g \text{ Polynom, Grad}(g) \leq n\}$$

**Satz 1.1.1**

Sei  $f_0, f_1, \dots, f_n \in \mathbb{R}$  und  $x_0, x_1, \dots, x_n \in \mathbb{R}$  gegeben mit  $x_i \neq x_j$  für  $i \neq j$ .  
 Dann existiert genau ein Polynom  $p_n \in \mathbb{P}_n$  mit  $p_n(x_\nu) = f_\nu, \nu = 0, \dots, n$ .

**Beweis:**

(a) *Existenz* :

Der Beweis wird über die Konstruktion der Polynome geführt. Es sei

$$l_\nu \in P_n \quad \text{mit} \quad l_\nu(x_i) = \delta_{\nu i} = \begin{cases} 1 & \text{,if } i = \nu \\ 0 & \text{,if } i \neq \nu \end{cases} \quad (\nu = 0, \dots, n)$$

wobei  $\delta_{\nu i}$  das **Kronecker-Symbol** ist.

$$(1.2) \quad \left\| \begin{aligned} l_\nu(x) &:= \prod_{\substack{i=0 \\ i \neq \nu}}^n \frac{x - x_i}{x_\nu - x_i} \\ &= \frac{(x - x_0)(x - x_1) \cdots (x - x_{\nu-1})(x - x_{\nu+1}) \cdots (x - x_n)}{(x_\nu - x_0)(x_\nu - x_1) \cdots (x_\nu - x_{\nu-1})(x_\nu - x_{\nu+1}) \cdots (x_\nu - x_n)} \\ p_n(x) &:= \sum_{\nu=0}^n f_\nu l_\nu(x) \quad \text{Lagrange-Interpolations Polynom} \end{aligned} \right.$$

$f_\nu$  wird durch  $p_n$  in  $x_\nu$  interpoliert:

$$p_n(x_k) = \sum_{\nu=0}^n f_\nu l_\nu(x_k) = f_k l_k(x_k) = f_1 \cdot 1$$

(b) *Eindeutigkeit* :

Seien  $p, q \in P_n$  zwei Interpolierende.

$\Rightarrow z := p - q \in P_n$  hat  $n + 1$  Nullstellen  $x_\nu$  ( $\nu = 0, \dots, n$ ).

$\Rightarrow z \equiv 0$  nach dem Fundamentalsatz der Algebra.

■

**Beispiel 1.1.2**

$$f(x) := x \sin \pi x, \quad x_\nu := -1 + \nu \frac{1}{2}, \quad \nu = 0, \dots, 4$$

$\nu$		0		1		2		3		4
$x_\nu$		-1		$-\frac{1}{2}$		0		$\frac{1}{2}$		1
$f_\nu$		0		$\frac{1}{2}$		0		$\frac{1}{2}$		0

$$\begin{aligned} \Rightarrow p_4(x) &= \frac{1}{2} \overbrace{\frac{(x+1) \cdot x \cdot (x - \frac{1}{2})(x-1)}{\frac{1}{2} \cdot (-\frac{1}{2}) \cdot (-1) \cdot (-\frac{3}{2})}}^{l_1(x)} + \frac{1}{2} \overbrace{\frac{(x+1)(x + \frac{1}{2})x(x-1)}{\frac{3}{2} \cdot 1 \cdot \frac{1}{2} \cdot (-\frac{1}{2})}}^{l_3(x)} \\ &= \frac{4}{3} x(x^2 - 1) \left[-x + \frac{1}{2} - x - \frac{1}{2}\right] \\ &= \frac{8}{3} x^2(1 - x^2) \end{aligned}$$

Eine andere mögliche Wahl für die Knoten ist  $x_\nu$  ( $\nu = 0, \dots, n$ ).

**Definition 1.1.3**

**Dividierte Differenzen** für  $0 \leq k, \nu, \nu + k \leq n$  :

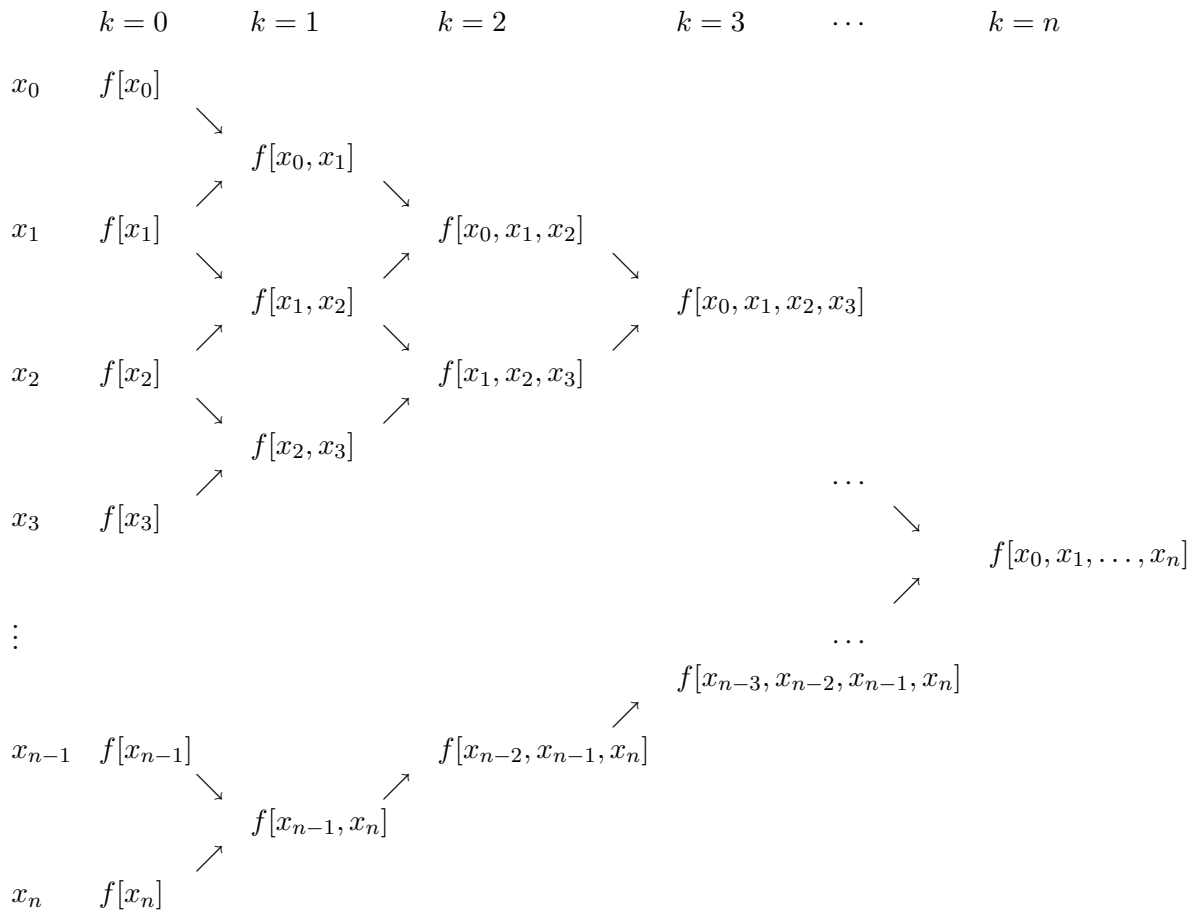
$$(1.3) \quad f[x_\nu, \dots, x_{\nu+k}] := \begin{cases} f[x_\nu] & (= f(x_\nu) = f_\nu) & , \quad k = 0 \\ \frac{f[x_{\nu+1}, \dots, x_{\nu+k}] - f[x_\nu, \dots, x_{\nu+k-1}]}{x_{\nu+k} - x_\nu} & , \quad k > 0 \end{cases}$$

Unter der Annahme, daß die Knoten  $x_\nu$  ( $\nu = 0, \dots, n$ ) paarweise verschieden sind, sind die  $f[x_\nu, \dots, x_{\nu+k}]$  für  $0 \leq k, \nu, \nu + k \leq n$  rekursiv durch (1.3) definiert.

Für  $k = 1$  ist zum Beispiel

$$f[x_\nu, x_{\nu+1}] = \frac{f[x_{\nu+1}] - f[x_\nu]}{x_{\nu+1} - x_\nu} = \frac{f(x_{\nu+1}) - f(x_\nu)}{x_{\nu+1} - x_\nu}$$

**Tabelle der Dividierten Differenzen :**



**Lemma 1.1.4**

Seien  $x_\nu$  ( $\nu = 0, \dots, n$ ) paarweise verschieden. Dann gilt ( $0 \leq k \leq n$ )

$$f[x_0, \dots, x_k] = \sum_{\nu=0}^k f(x_\nu) \left[ \prod_{\substack{j=0 \\ j \neq \nu}}^k (x_\nu - x_j) \right]^{-1}$$

**Beweis:**

Der Beweis wird durch Induktion über  $k$  geführt. Für  $k = 0$  gilt  $f_0 = f[x_0]$ . Angenommen, die Behauptung gilt bereits für  $k - 1$ . Es gilt

$$f[x_1, \dots, x_{1+(k-1)}] = \sum_{\nu=0}^{k-1} f(x_{\nu+1}) \left[ \prod_{\substack{j=0 \\ j \neq \nu}}^{k-1} (x_{\nu+1} - x_{j+1}) \right]^{-1}$$

$$f[x_0, \dots, x_k]$$

$$\begin{aligned} &= \frac{1}{x_k - x_0} \{ f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}] \} \\ &= \frac{1}{x_k - x_0} \left\{ \sum_{\nu=0}^{k-1} f(x_{\nu+1}) \prod_{\substack{j=0 \\ j \neq \nu}}^{k-1} (x_{\nu+1} - x_{j+1})^{-1} - \sum_{\nu=0}^{k-1} f(x_\nu) \prod_{\substack{j=0 \\ j \neq \nu}}^{k-1} (x_\nu - x_j)^{-1} \right\} \\ &= \frac{1}{x_k - x_0} \left\{ \frac{f(x_k)}{\prod_{\substack{j=1 \\ j \neq k}}^k (x_k - x_j)} + \sum_{\nu=1}^{k-1} f(x_\nu) \left[ \frac{1}{\prod_{\substack{j=1 \\ j \neq \nu}}^k (x_\nu - x_j)} - \frac{1}{\prod_{\substack{j=0 \\ j \neq \nu}}^{k-1} (x_\nu - x_j)} \right] - \frac{f(x_0)}{\prod_{\substack{j=0 \\ j \neq 0}}^{k-1} (x_0 - x_j)} \right\} \\ &= \frac{f_k}{\prod_{\substack{j=0 \\ j \neq k}}^k (x_k - x_j)} + \sum_{\nu=1}^{k-1} f_\nu \frac{1}{\prod_{\substack{j=0 \\ j \neq \nu}}^k (x_\nu - x_j)} \underbrace{\frac{x_\nu - x_0 - (x_\nu - x_k)}{x_k - x_0}}_{=1} + \frac{f_0}{\prod_{\substack{j=0 \\ j \neq 0}}^k (x_0 - x_j)} \end{aligned}$$

woraus schließlich die Behauptung folgt. ■

**Satz 1.1.5**

Seien  $x_0, x_1, \dots, x_n$  paarweise verschieden. Dann gilt  
(Newton-Interpolation Polynom)

$$(1.4) \quad p_n(x) = \sum_{k=0}^n f[x_0, \dots, x_k] \prod_{j=1}^k (x - x_{k-j}).$$

(Bezüglich der Basis

$$\prod_{j=1}^k (x - x_{k-j}), \quad k = 0, \dots, n$$

sieht man, daß die Dividierten Differenzen  $f[x_0, \dots, x_k]$  die Führungskoeffizienten des Interpolations-Polynoms sind.)

**Beweis:**

Der Beweis wird durch Induktion über  $n$  geführt. Für  $n = 0$  gilt  $p_0(x) = f[x_0] = f_0$ . Angenommen, die Behauptung gilt bereits für  $n - 1 \geq 0$ . Es gilt

$$p_{n-1}(x) = \sum_{k=0}^{n-1} f[x_0, \dots, x_k] \prod_{j=1}^k (x - x_{k-j})$$

$p_{n-1}(x)$  interpoliert  $f_0, f_1, \dots, f_{n-1}$  in den Knoten  $x_0, \dots, x_{n-1}$ . Dann kann das eindeutig bestimmte Interpolations-Polynom  $p_n$  geschrieben werden als

$$p_n(x) = p_{n-1}(x) + a \cdot \underbrace{\prod_{j=1}^n (x - x_{n-j})}_{\neq 0 \text{ for } x = x_n}$$

wobei  $a \in \mathbb{R}$  noch bestimmt werden muß. Wähle  $a$  so, daß  $p_n$  in  $x_n$  interpoliert. Unter Verwendung der Lagrange-Interpolations Polynome läßt sich  $a$  wie folgt berechnen:

$$\begin{aligned} a n! &= \frac{d^n}{dx^n} p_{n-1}(x) + a n! = \frac{d^n}{dx^n} p_n(x) \\ &\stackrel{(1.2)}{=} \sum_{j=0}^n f_j \frac{d^n}{dx^n} l_j(x) = n! \sum_{j=0}^n f_j \left( \prod_{\substack{i=0 \\ i \neq j}}^n (x_j - x_i) \right)^{-1} \stackrel{(1.1.4)}{=} n! f[x_0, \dots, x_n] \end{aligned}$$

■

**Beispiel 1.1.6** (Dividierte Differenzen)

$x_0 = 0$	<u>1</u>		
$x_1 = 1.5$	2	<u>0.6667</u>	
$x_2 = 2.5$	2	<u>-0.26667</u>	<u>0.0222222</u>
$x_3 = 4.5$	1	-0.5	-0.16667

⇒ *Newton-Interpolations Polynom* :

$$\begin{aligned} p_3(x) &= f[x_0] + (x - x_0)f[x_0, x_1] \\ &\quad + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &\quad + (x - x_0)(x - x_1)(x - x_2)f[x_0, x_1, x_2, x_3] \\ &= f[x_0] + (x - x_0) \{ f[x_0, x_1] + (x - x_1) \{ f[x_0, x_1, x_2] + (x - x_2) f[x_0, x_1, x_2, x_3] \} \} \\ &= 1 + x \{ 0.6667 + (x - 1.5) \{ -0.26667 + (x - 2.5) \cdot 0.02222 \} \} \end{aligned}$$

⇒  $p_3(2.1) \approx 2.0528$  .

**1.1.2 Interpolation bei äquidistanten Knoten, Tschebyscheff-Polynome**

Die Bezeichnung „ $\mathbf{f(x)}$  ist tabelliert für  $\mathbf{a(h)b}$ “ heißt, daß  $f$  an den Knoten  $x_i$  mit

$$x_i = a + ih \quad (i = 0, \dots, N, \quad N = \frac{b - a}{h})$$

zur Verfügung steht.

Variablentransformation:

$$\begin{aligned} s(x) &= \frac{x - x_0}{h} \quad \text{so daß } x = x(s) = x_0 + sh \\ f(x) &= f(x_0 + sh) =: f_s \end{aligned}$$

Gesucht ist das Interpolationspolynom  $p_n(x)$  zu  $f(x)$  an den äquidistanten Knoten  $x_k, \dots, x_{k+n}$ . Die Dividierten Differenzen sind hier nicht von Nöten, es reicht eine Differenzen-Tabelle zu erstellen.

### Definition 1.1.7

#### Vorwärtsgerichtete Differenzen

$$\Delta^i f_s = \begin{cases} f_s & , \quad i = 0 \\ \Delta(\Delta^{i-1} f_s) = \Delta^{i-1} f_{s+1} - \Delta^{i-1} f_s & , \quad i > 0. \end{cases}$$

**Lemma 1.1.8** (Zusammenhang zwischen vorw.Diff. und div.Diff.)

$$(1.5) \quad \text{Für alle } i \geq 0: \quad f[x_k, \dots, x_{k+i}] = \frac{1}{i! h^i} \Delta^i f_k$$

#### Beweis:

Der Beweis wird durch Induktion über  $i$  geführt.

Für  $i = 0$  gilt  $f[x_k] = f(x_k) = f_k = \Delta^0 f_k$

Angenommen (1.5) gilt für  $i = n \geq 0$ , dann

$$\begin{aligned} f[x_k, \dots, x_{k+n+1}] &= \frac{f[x_{k+1}, \dots, x_{k+n+1}] - f[x_k, \dots, x_{k+n}]}{x_{k+n+1} - x_k} \\ &= \frac{\left(\frac{1}{n! h^n} \Delta^n f_{k+1}\right) - \left(\frac{1}{n! h^n} \Delta^n f_k\right)}{(n+1)h} \\ &= \frac{1}{(n+1)! h^{n+1}} \Delta^{n+1} f_k \end{aligned}$$

■

Das Newton-Interpolations Polynom für  $f(x)$  in  $x_k, \dots, x_{k+n}$  läßt sich somit schreiben als

$$p_n(x) = \sum_{i=0}^n \underbrace{\frac{1}{i! h^i} \Delta^i f_k}_{f[x_k, \dots, x_{k+i}]} \prod_{j=0}^{i-1} (x - x_{k+j})$$

Für die Knoten ergibt sich

$$x - x_{k+j} = x_0 + sh - [x_0 + (k+j)h] = (s - k - j)h$$

Insgesamt erhalten wir die **Newton Vorwärts-Differenzen Formel**

$$(1.6) \quad p_n(x) = p_n(x_0 + sh) = \sum_{i=0}^n \Delta^i f_k \prod_{j=0}^{i-1} \frac{s - k - j}{j + 1} = \sum_{i=0}^n \Delta^i f_k \binom{s - k}{i}$$

mit der Binomialfunktion

$$\binom{y}{i} := \begin{cases} 1 & , \quad i = 0 \\ \frac{y(y-1) \cdots (y-i+1)}{1 \cdot 2 \cdots i} \left( = \prod_{j=0}^{i-1} \frac{y-j}{j+1} \right) & , \quad i > 0 \end{cases}$$

**Beispiel 1.1.9** (Vorwärts-Differenzen bei äquidistanten Knoten)

	$f_i$	$\Delta f_i$	$\Delta^2 f_i$	$\Delta^3 f_i$
$x_0 = 20$	<u>0.34202</u>			
$x_1 = 30$	0.5	<u>0.15798</u>		
$x_2 = 40$	0.64279	0.14279	<u>-0.1519</u>	
$x_3 = 50$	0.76604	0.12325	-0.1954	<u>-0.00435</u>

$$\begin{aligned}
 \Rightarrow p_3(x) &= \sum_{i=0}^3 \Delta^i f_0 \prod_{j=0}^{i-1} \frac{s-j}{j+1} \\
 &= f_0 + \Delta f_0 \frac{s}{1} + \Delta^2 f_0 \frac{s}{1} \frac{s-1}{2} + \Delta^3 f_0 \frac{s}{1} \frac{s-1}{2} \frac{s-2}{3} \\
 &= f_0 + s \left[ \Delta f_0 + \frac{s-1}{2} \left\{ \Delta^2 f_0 + \frac{s-2}{3} \Delta^3 f_0 \right\} \right] \\
 \Rightarrow p_3(36) &\approx 0.58778.
 \end{aligned}$$

### 1.1.3 Interpolationsfehler

Bevor wir zu den eigentlichen Betrachtungen kommen noch ein wichtiger Satz, den wir im weiteren benutzen werden.

**Satz 1.1.10** (Rolle's Theorem)

Sei  $f(x)$  stetig auf  $a \leq x \leq b$  und differenzierbar auf  $a < x < b$ .

Falls  $f(a) = f(b) = 0$  ist, dann existiert mindestens ein Punkt  $\xi$  zwischen  $a$  und  $b$  so daß  $f'(\xi) = 0$  ist.

**Satz 1.1.11**

Sei  $I := [a, b] \subset \mathbb{R}$ ,  $a < b$ ,  $x_0, x_1, \dots, x_n \in I$  mit  $x_j \neq x_i$  für alle  $i \neq j$  und das **Knotenpolynom**  $\omega(x)$  wie folgt definiert.

$$\omega(x) := \prod_{\nu=0}^n (x - x_\nu).$$

Sei  $f \in \mathbb{C}^{n+1}(I)$ . Dann gilt für alle  $x \in I$ , daß ein  $\xi(x) \in I$  existiert, so daß

$$(1.7) \quad f(x) - p_n(x) = \frac{\omega(x)}{(n+1)!} f^{(n+1)}(\xi(x))$$

wobei  $p_n$  das Interpolationspolynom ist.

**Beweis:**

Sei  $x \in I$ ,  $x \neq x_0, x_1, \dots, x_n$  ( andernfalls ist (1.7) trivial ),  $x$  fest und  $c := \frac{f - p_n}{\omega}(x)$ . Die Funktion

$$F(t) = f(t) - p_n(t) - c\omega(t) \quad (t \in I)$$

hat  $n+2$  Nullstellen in  $I$ , nämlich  $t = x_0, x_1, \dots, x_n$  und  $t = x$ . Nach dem Satz von Rolle (vgl. Analysis) hat  $\frac{d}{dt}F$   $n+1$  Nullstellen in den gegebenen Intervallen (zwischen den  $x'_s$  und  $x$ ). Das bedeutet, daß  $\frac{d^2}{dt^2}F$   $n$  Nullstellen besitzt. Führt man diese Überlegung weiter, wo erhält man schließlich, daß  $\frac{d^{n+1}}{dt^{n+1}}F$  eine Nullstelle  $\xi(x) \in I$  hat (sogar in dem Inneren von  $I$ ). Da außerdem noch gilt

$$\frac{d^{n+1}}{dt^{n+1}}p_n(t) = 0 \quad , \quad \frac{d^{n+1}}{dt^{n+1}}\omega(t) = (n+1)!$$

erhalten wir

$$0 = \left. \frac{d^{n+1}}{dt^{n+1}}(f(t) - p_n(t) - c\omega(t)) \right|_{t=\xi(x)} = f^{(n+1)}(\xi(x)) - c(n+1)!$$

Die Betrachtungen wurden für ein festes  $x \in I$  gemacht. Daher

$$\frac{f(x) - p_n(x)}{\omega(x)} = c = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}$$

■

Für  $f \in C^{n+1}(I)$  können wir Satz 1.1.11 benutzen, um den Interpolationsfehler abzuschätzen.

$$(1.8) \quad \|f - p_n\|_\infty := \max_{x \in I} |f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \|\omega\|_\infty \|f^{(n+1)}\|_\infty$$

### Beispiel 1.1.12

$$f(x) = \sin(x)$$

Interpoliere  $f(x)$  durch ein quadratisches Polynom  $p_2 \in \mathbb{P}_2$  in den Knoten  $x_0 = -h, x_1 = 0, x_2 = h$

$$\begin{aligned} f(x) - p_2(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{3!} f^{(3)}(\xi), \\ \omega(x) &= (x+h)x(x-h) = x^3 - h^2x \end{aligned}$$

Notwendige Bedingung für das Maximum von  $\omega(x)$  ist, daß  $\omega'(x) = 0$  ist. Damit folgt  $3x^2 - h^2 = 0$  und somit  $x = \pm \frac{h}{\sqrt{3}}$ .

$$\left| \omega\left(\pm \frac{h}{\sqrt{3}}\right) \right| = \left| \frac{h^3}{3\sqrt{3}} - \frac{h^3}{\sqrt{3}} \right| = \frac{2h^3}{3\sqrt{3}}$$

Da außerdem noch  $|f^{(3)}(\xi)| \leq 1$  gilt, folgt

$$\max_{-h \leq x \leq h} |f(x) - p_2(x)| \leq \frac{\sqrt{3}}{27} h^3$$

Um einen Interpolationsfehler kleiner als  $5 \cdot 10^{-8}$  zu erhalten, muß  $h$  wie folgt gewählt werden

$$\frac{\sqrt{3}}{27} h^3 < 5 \cdot 10^{-8} \quad \Rightarrow \quad h \approx 0.01$$

Es läßt sich nicht immer  $\lim_{n \rightarrow \infty} \|f - p_n\|_\infty = 0$  erreichen. Ein Gegenbeispiel dazu ist

$$f(x) = \frac{1}{1+x^2}, \quad -5 \leq x \leq 5$$

In (1.8) können wir im allgemeinen nicht die Konstante  $\|f^{(n+1)}\|_\infty$  verkleinern; aber durch eine spezielle Wahl der Knoten können wir  $\|\omega\|_\infty$  minimieren. Der Einfachheit halber betrachten wir nun die Interpolation im Intervall  $J := [-1, 1]$ . Der allgemeine Fall kann immer auf diesen zurückgeführt werden.

$$\varphi(x) = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)x$$

erfüllt  $I = \varphi(J)$ , wobei  $I = [a, b]$  gilt und ferner

$$\|f - p_n\|_{I, \infty} = \|f(\varphi(\cdot)) - p_n(\varphi(\cdot))\|_{J, \infty}$$

### Definition 1.1.13

Die Polynome  $T_n \in P_n$  mit

$$(1.9) \quad T_n(x) := \cos(n \arccos x)$$

heißen **Tschebyscheff-Polynomials** (1. Ordnung)

$T_n$  sind **Polynome**, da

$$(1.10) \quad \begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) + T_{n-1}(x) &= \cos(n+1)\phi + \cos(n-1)\phi \\ &= 2 \cos n\phi \cos \phi \\ &= 2T_n(x)x \quad \text{mit } \phi := \arccos x \end{aligned}$$

d.h.

$$(1.11) \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

### Nullstellen der $T_n$

Es gilt für ganzzahlige  $k$

$$n \arccos x_k = \left(k + \frac{1}{2}\right)\pi, \quad x_k \in J = [-1, 1]$$

und somit

$$(1.12) \quad x_k = \cos \frac{k + \frac{1}{2}}{n} \pi \quad (k = 0, 1, 2, \dots, n-1)$$

Aufgrund der Symmetrie von  $\cos$  und seiner Periode  $2\pi$  erhalten wir für alle anderen  $k$  keine weiteren  $x_k$

### Extrema von $T_n$

Aus (1.9) folgt  $|T_n(x)| \leq 1$ . Somit ist  $T_n(y_k)$  maximal oder minimal für  $y_k \in J$  mit  $|T_n(y_k)| = 1$ . Weiter ist  $n \arccos y_k = k\pi$  und so

$$(1.13) \quad y_k = \cos \frac{k\pi}{n} \quad (0 \leq k \leq n)$$

Dies sind schon  $n+1$  verschiedene  $y_k$  und mehr (zwei Randextrema und  $(n-1)$  innere Extrema) kann  $T_n$  als ein Polynom  $n$ -ten Grades nicht haben.

Bevor wir den nächsten Satz beweisen können, betrachten wir das folgende Lemma.

**Lemma 1.1.14**

Sei  $Q_n := \{q_n \in P_n; \frac{d^n}{dx^n} q_n(x) = n!\}$ . Dann gilt

$$(1.14) \quad \min_{q_n \in Q_n} \|q_n\|_{J,\infty} = \|2^{1-n}T_n\|_{J,\infty} = 2^{1-n}$$

**Beweis:**

Aus (1.10) und (1.11) folgt, daß  $T_n$  als führenden Koeffizienten  $2^{n-1}$  hat, d.h.  $2^{1-n}T_n \in Q_n$ . Dann ist das Minimum kleiner gleich  $2^{1-n}$ .

Angenommen, es existiert ein  $q_n \in Q_n$  mit  $\|q_n\| < 2^{1-n}$ . Setze  $q := q_n - 2^{1-n}T_n$ . Dann ist

$$q(y_k) = q_n(y_k) - 2^{1-n}T_n(y_k) \begin{cases} < 0 & , \text{ für gerade } k \quad (T(y_k) = 1) \\ > 0 & , \text{ für ungerade } k \quad (T(y_k) = -1) \end{cases}$$

Ist  $q \in C(J)$ , so wechselt  $q$  in  $J$   $n$ -mal das Vorzeichen, d.h. wenigstens  $n$  Nullstellen.

Ist  $q \in P_{n-1}$  so ist  $q = 0$  was ein Widerspruch zur Annahme ist. ■

**Satz 1.1.15**

Das Knotenpolynom  $\omega(x) = \prod_{\nu=0}^n (x - x_\nu) \in P_{n+1}$  mit der kleinsten  $\infty$ -Norm ist das Polynom

$$\omega_{n+1}(x) =: 2^{-n}T_{n+1}(x) \quad \text{mit} \quad \|\omega_{n+1}\|_{\infty,J} = 2^{-n}$$

**Beweis:**

$\omega_{n+1} \in Q_{n+1}$ . Somit folgt die Behauptung durch die Anwendung von (1.1.14). ■

Aus (1.8) folgt, daß wir in  $J$  in den **Tschebyscheff-Knotenpunkten**

$$x_k = \cos \frac{k + \frac{1}{2}}{n} \pi, \quad k = 0, \dots, n$$

für  $f \in C^{n+1}(J)$  die folgende Abschätzung haben:

$$(1.15) \quad \|f - p_n\|_{J,\infty} \leq \frac{1}{2^n(n+1)!} \|f^{(n+1)}\|_{J,\infty}$$

**1.2 Hermite-Interpolation**

Sei  $I := [a, b] \subset \mathbb{R}$ . Mit den Daten  $f(x_\nu) =: f_\nu$ ,  $f'(x_\nu) =: f'_\nu$ ,  $\nu = 0, \dots, n$  einer glatten Funktion  $f$  können wir auf eine „informationsbewahrende Art“ interpolieren.

**Satz 1.2.1**

Es existiert genau ein Hermite-Interpolations Polynom  $h_{2n+1} \in P_{2n+1}$  mit

$$(1.16) \quad h_{2n+1}(x_\nu) = f_\nu, \quad h'_{2n+1}(x_\nu) = f'_\nu \quad (0 \leq \nu \leq n)$$

für paarweise verschiedene Knotenpunkte  $x_\nu$ ,  $\nu = 0, \dots, n$ .

**Beweis:**(a) *Eindeutigkeit* :

Seien  $h_{2n+1}$  und  $g_{2n+1}$  zwei Polynome mit dieser Eigenschaft. Dann ist  $h_{2n+1} - g_{2n+1} \in P_{2n+1}$  und  $h_{2n+1} - g_{2n+1}$  hat  $n+1$  doppelte Nullstellen  $x_\nu$ , denn  $h_{2n+1}(x_\nu) = f_\nu = g_{2n+1}(x_\nu)$ . Dies wiederum bedeutet, daß  $h_{2n+1} - g_{2n+1} \equiv 0$  ist, woraus die Eindeutigkeit folgt.

(b) *Existenz* :

$$(1.17) \quad h_{2n+1}(x) := \sum_{\nu=0}^n f_\nu k_\nu(x) + \sum_{\nu=0}^n f'_\nu m_\nu(x)$$

mit

$$(1.18) \quad \begin{cases} k_\nu(x) & := [1 - 2l'_\nu(x_\nu)(x - x_\nu)] l_\nu^2(x) \quad (l_\nu^2(x) \text{ } 2n\text{-degree}) \\ m_\nu(x) & := (x - x_\nu) l_\nu^2(x) \end{cases}$$

wobei  $l_\nu$  das **Lagrange-Polynom** ist. Es ist

$$k_\nu(x_\mu) = m'_\nu(x_\mu) = \delta_{\nu\mu}, \quad k'_\nu(x_\mu) = m_\nu(x_\mu) = 0$$

Durch Einsetzen ergibt sich

$$h_{2n+1}(x_\mu) = f_\mu k_\mu(x_\mu) = f_\mu$$

und

$$h'_{2n+1}(x_\mu) = \sum_{\nu=0}^n f_\nu k'_\nu(x_\nu) + \sum_{\nu=0}^n f'_\nu m'_\nu(x_\nu) = f'_\mu$$

■

Analog dazu erhalten wir:

**Korollar 1.2.2**

Seien  $x_0, x_1, \dots, x_n \in I$  paarweise verschieden,  $f \in C^{2n+2}(I)$  und  $h_{2n+1}$  das Hermite-Interpolations Polynom zu  $f$  und  $x_0, x_1, \dots, x_n$ . Dann

$$(1.19) \quad \forall x \in I \exists \xi(x) \in I \quad f(x) - h_{2n+1}(x) = \frac{\omega(x)^2}{(2n+2)!} f^{(2n+2)}(\xi(x))$$

**Beispiel 1.2.3**

Gesucht ist das Hermite-Interpolations Polynom von

$$f(x) := \sin x \quad \text{in } x_0 = 0, \quad x_1 = \frac{\pi}{2}$$

$$\begin{array}{l} f(x) : \\ f'(x) : \end{array} \begin{array}{c|c|c} x_\nu & 0 & \frac{\pi}{2} \\ \hline \sin x & 0 & 1 \\ \hline \cos x & 1 & 0 \end{array}$$

Mit (1.17) ergibt sich

$$\begin{aligned}
 h_3(x) &= 1 \cdot k_1(x) + 1 \cdot m_0(x) \\
 \text{mit} \quad k_1(x) &= \left[ 1 - 2\left(x - \frac{\pi}{2}\right) \left(\frac{x-0}{\frac{\pi}{2}-0}\right)' \right] \left(\frac{x-0}{\frac{\pi}{2}-0}\right)^2 = \frac{4}{\pi^2} x^2 \left(3 - \frac{4}{\pi} x\right) \\
 m_0(x) &= (x-0) \left(\frac{x-\frac{\pi}{2}}{0-\frac{\pi}{2}}\right)^2 = \frac{4}{\pi^2} x \left(x - \frac{\pi}{2}\right)^2 \\
 \Rightarrow h_3(x) &= \frac{4}{\pi^2} x \left[ 3x - \frac{4}{\pi} x^2 + \left(x - \frac{\pi}{2}\right)^2 \right] \\
 \mathbf{NR} : \quad \frac{d}{dx} \{x(x - \frac{\pi}{2})\} &= (x - \frac{\pi}{2}) + x \stackrel{!}{=} 0 \quad \Rightarrow \quad x = \frac{\pi}{4}
 \end{aligned}$$

Wegen (1.19) gilt

$$\|f - h_3\|_\infty \leq \frac{\|\omega^2\|_\infty}{4!} \|f^{(4)}\|_\infty \leq \frac{1}{24} \left( \max_I x \left(x - \frac{\pi}{2}\right) \right)^2 \leq \frac{1}{24} \left( \frac{\pi^2}{16} \right)^2 = 0.016$$

als eine obere Schranke für den Fehler.

### 1.3 Stückweise polynomiale Interpolation

Es werde  $I = [a, b]$  in äquidistant Subintervalle

$$I_i := [x_{i-1}, x_i], \quad x_i = a + ih \quad (i = 0, \dots, n), \quad h = \frac{b-a}{n}$$

zerlegt und eine lineare Interpolation in jedem  $I_\nu$  durchgeführt. Dann kann die Lösung zu den Daten  $f_0, \dots, f_n$  sofort gegeben werden und zwar mit Hilfe sogenannter **Hut-Funktionen**  $\varphi_i$ .

**Stückweise linear Interpolierende :**

$$\begin{aligned}
 (1.20) \quad \varphi(x) &= \sum_{i=0}^n f_i \varphi_i(x) \\
 \text{mit} \quad \varphi_i(x) &:= \begin{cases} \frac{x - x_{i-1}}{h} & , x \in I_i \\ \frac{x_{i+1} - x}{h} & , x \in I_{i+1} \\ 0 & , \text{sonst} \end{cases}
 \end{aligned}$$

Schränkt man  $\varphi$  auf  $I_k$  ein, also  $\varphi|_{I_k} = f_{k-1} \varphi_{k-1}(x) + f_k \varphi_k(x)$ , so sieht man

$$\begin{aligned}
 \varphi(x_{k-1}) &= f_{k-1} \cdot 1 + f_k \cdot 0 = f_{k-1} \\
 \varphi(x_k) &= f_{k-1} \cdot 0 + f_k \cdot 1 = f_k
 \end{aligned}$$

#### Bemerkung 1.3.1

Somit ergibt sich für den Fehler die Abschätzung :

$$\|f - p_n\|_\infty = \max_{x \in I} |f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \|\omega\|_\infty \|f^{(n+1)}\|_\infty$$

und

$$\begin{aligned}\|f - \varphi\|_{I,\infty} &= \max_i \|f - \varphi\|_{I_i,\infty} \leq \max_i \frac{1}{2} \|\omega_i\|_{I_i,\infty} \|f''\|_{I_i,\infty} \\ &= \frac{h^2}{8} \|f''\|_{I_i,\infty} \quad \text{for } f \in C^2(I)\end{aligned}$$

*Bemerkung zu der zweiten Gleichung:*

$$\begin{aligned}\omega_i &= (x - x_{i-1})(x - x_i) \\ \frac{d\omega_i}{dx} &= 2x - x_i - x_{i-1} = 0 \quad \Rightarrow \quad x = \frac{x_i + x_{i-1}}{2} \\ \Rightarrow \omega_i &\leq \left(\frac{x_i - x_{i-1}}{2}\right) \left(\frac{x_{i-1} - x_i}{2}\right) = \frac{h^2}{4} \\ \Rightarrow \|\omega_i\|_{I_i,\infty} &= \frac{h^2}{4}\end{aligned}$$

### Satz 1.3.2

Sei  $f \in C^2(I)$ ,  $\varphi \in \mathbb{P}_1$  wie in (1.20) und  $x_i$  äquidistant gewählt. Dann gilt

$$(1.21) \quad \|f - \varphi\|_{I,\infty} \leq \frac{h^2}{8} \|f''\|_{I,\infty}$$

$$(1.22) \quad \|f' - \varphi'\|_{I,\infty} \leq \frac{h^2}{2} \|f''\|_{I,\infty}$$

**Beweis:**

(zu (1.21)) Mit Satz 1.1.11 ergibt sich

$$\|f - \varphi\|_{I-i,\infty} \leq \frac{1}{2} \|f''\|_{I,\infty} \|(x - x_i)(x - x_{i-1})\|_{I,\infty} = \frac{1}{2} \|f''\|_{I,\infty} \|\omega\|_{I,\infty}$$

Mit den obigen Bemerkungen zu  $\omega$  ergibt sich

$$\max_{i=0,\dots,n} \|f - \varphi\|_{I_i,\infty} \leq \frac{h^2}{8} \max_i \|f''\|_{I_i,\infty} = \frac{h^2}{8} \|f''\|_{I,\infty}$$

(zu (1.22))

$$\|f' - \varphi'\|_{I,\infty} = \max_i \|f' - \varphi'\|_{I_i,\infty}$$

In  $I_i$  gilt:

$$\begin{aligned}&h(f'(x) - \varphi'(x)) \\ &= hf'(x) - h \frac{f_i - f_{i-1}}{h} = hf'(x) - \int_{x_{i-1}}^{x_i} f'(t) dt \\ &= \underbrace{(x_i - x_{i-1})f'(x) - \left(x_i f'(x_i) - x_{i-1} f'(x_{i-1}) - \int_{x_{i-1}}^{x_i} f'(t) dt\right)}_{=0} \\ &= -x_i \int_x^{x_i} f''(t) dt - x_{i-1} \int_{x_{i-1}}^x f''(t) dt + \int_{x_{i-1}}^{x_i} t f''(t) dt \\ &= \int_{x_{i-1}}^x (t - x_{i-1}) f''(t) dt + \int_x^{x_i} (t - x_i) f''(t) dt\end{aligned}$$

Dann folgt

$$\begin{aligned} h|f'(x) - \varphi'(x)| &\leq \|f''\|_{I_i, \infty} \left\{ \int_{x_{i-1}}^x (t - x_{i-1}) dt + \int_x^{x_i} (t - x_i) dt \right\} \\ &= \|f''\|_{I_i, \infty} \frac{1}{2} \left( (x - x_{i-1})^2 + (x_i - x)^2 \right) \\ &\leq \frac{h^2}{2} \|f''\|_{I_i, \infty} \end{aligned}$$

und somit die Behauptung. ■

## 1.4 Interpolation mit kubischen Splines

### Definition 1.4.1

Sei  $I := [a, b]$ ,  $f \in C^1(I)$ ,  $x_i := a + ih$ ,  $I_i := [x_{i-1}, x_i]$ .

$\psi : I \rightarrow \mathbb{R}$  heißt (**vollständiger**) **kubischer Spline zu f**

- $\Leftrightarrow$  (i)  $\psi \in C^2(I)$  (iii)  $\psi(x_i) = f(x_i) =: f_i$  ( $0 \leq i \leq n$ )  
 (ii)  $\psi \in P_3(I_i)$  ( $i = 1, \dots, n$ ) (iv)  $\psi'(a) = f'(a)$ ,  $\psi'(b) = f'(b)$

### Bemerkung 1.4.2

Bedingung (iv) sichert die Eindeutigkeit. Wird stattdessen gefordert :

$$(iv) \quad \psi''(a) = \psi''(b) = 0$$

so spricht man von einem **natürlichen Spline** (mit weniger guten Approximationseigenschaften, da weniger Informationen über  $f$  vorliegen).

### Satz 1.4.3

Sei  $x_i, f_i$  wie in Definition 1.4.1. Dann existiert genau ein vollständiger kubischer Spline  $\psi$  zu  $f$ .

Im folgenden wird eines unserer Ziel der Beweis zu diesem Satz sein. Um den kubischen Spline bestimmen zu können, muß die Zahl der Freiheitsgrade der Gleichung in (ii) gleich der Anzahl der aus (1.4.1) folgenden Bedingungen sein.

(ii)	4	4	...	4	4	Freiheitsgrade = $4n$	
	$a = x_0$	$x_1$	$x_2$	...	$x_{n-2}$	$x_{n-1}$	$x_n = b$
(i)		3	3	...	3	3	Bedingungen :
(iii)	1	1	1	...	1	1	1
(iv)	1						1

Aus der Bedingung (ii) folgt:  $\psi(x)|_{I_i} = a_i + b_i x + c_i x^2 + d_i x^3$

So erhalten wir für jedes  $I_i$  4 Freiheitsgrade. Andererseits sind die Bedingungen in den Knotenpunkten gegeben. Daher ist (iv) notwendig für:  $\#$  Freiheitsgrade =  $\#$  Bedingungen.

**Fortsetzung des Gitters :**

$$x_i = a + ih, \quad i \in \mathbb{Z}, \quad I_i = [x_{i-1}, x_i], \quad i \in \mathbb{Z}$$

$S_h^2 := \{ \varphi \in C(\mathbb{R}), \varphi|_{I_i} \text{ linear } \forall i \in \mathbb{Z} \}$  **Raum der stückweise linearen und stetigen Funktionen**

$S_h^3 := \{ \psi \in C^2(\mathbb{R}), \psi|_{I_i} \in P_3 \forall i \in \mathbb{Z} \}$  **Raum der Spline-Funktionen**

$S_h^1 := \{ \varphi_i, i \in \mathbb{Z} \}$

mit 
$$\varphi_i(x) := \begin{cases} \frac{x - x_{i-1}}{h} & , x \in I_i \\ \frac{x_{i+1} - x}{h} & , x \in I_{i+1} \\ 0 & , \text{sonst} \end{cases}$$

**Bemerkung 1.4.4**

(1.23) 
$$\psi \in S_h^3 \Leftrightarrow \psi'' \in S_h^1$$

**Satz 1.4.5**

Sei  $f \in C^4(I)$  mit  $I = [a, b]$  und  $\psi$  der zugehörige Spline,  $x_i = a + ih$ ,  $h = |I_i|$ . Dann gilt: Es existiert eine Konstante  $c > 0$ , so daß

$$\|f^{(i)} - \psi^{(i)}\|_{I, \infty} = \max_k \|f^{(i)} - \psi^{(i)}\|_{I_k, \infty} \leq ch^{4-i} \quad (i = 0, 1, 2, 3)$$

Es stellt sich die Frage, ob sich mit  $\varphi_i$  eine Basis  $\psi_i$  von  $S_h^3$  bilden läßt. Dazu werden wir das Doppelintegral

$$\int_{-\infty}^x \int_{-\infty}^s \varphi_i(t) dt ds$$

betrachten. Als Einschränkung für unsere Basiselemente  $\psi_i$  soll der **kleinste kompakte Träger um  $\mathbf{x}_i$**  herum gewählt werden.

$$\begin{aligned} \chi_i(x) &:= \int_{-\infty}^x \int_{-\infty}^s \varphi_i(t) dt ds \\ &= \int_{-\infty}^x \int_{-\infty}^s \begin{cases} 0 \\ \frac{x - x_{i-1}}{h} \\ \frac{x_{i+1} - x}{h} \\ 0 \end{cases} = \int_{-\infty}^x \begin{cases} 0 \\ \frac{(x - x_{i-1})^2}{2h} \\ h - \frac{(x_{i+1} - x)^2}{2h} \\ h \end{cases} \\ &= \begin{cases} 0 & x \leq x_{i-1} \\ \frac{(x - x_{i-1})^3}{6h} & x_{i-1} \leq x \leq x_i \\ \frac{1}{6h}(x_{i+1} - x)^3 - h(x_{i+1} - x) + h^2 & x_i \leq x \leq x_{i+1} \\ h(x - x_{i+1}) + h^2 & x \geq x_{i+1} \end{cases} \end{aligned}$$

$\chi_i$  hat **keinen kompakten** Träger. Damit  $a_0\chi_i + a_\nu\chi_{i+\nu}$  einen kompakten Träger besitzt, muß  $a_0\chi_i + a_\nu\chi_{i+\nu}$  für nicht-triviale  $a_0$  und  $a_\nu$  verschwinden, also

$$\begin{aligned} 0 &\stackrel{!}{=} a_0\chi_i + a_\nu\chi_{i+\nu} \\ &= \underbrace{h(a_0 + a_\nu)(x + h)}_{=0} - h(a_0x_{i+1} + a_\nu x_{i+\nu+1}) \quad \text{für große } x \gg 1 \\ &= -h((a_0 + a_\nu)x_{i+1} + a_\nu(x_{i+\nu+1} - x_{i+1})) \end{aligned}$$

Da dies nur für  $a_0 = a_\nu = 0$  gilt, besitzt  $a_0\chi_i + a_\nu\chi_{i+\nu}$  ebenfalls keinen kompakten Träger. D.h. eine Linearkombination von drei verschiedenen  $\chi_i$ 's liefert den kleinsten Träger für

$$\psi_i^B = a_{-1}\chi_{i-1} + a_0\chi_i + a_1\chi_{i+1} \stackrel{!}{=} 0 \quad \text{für große } x \gg 1$$

also

$$0 = h \underbrace{(a_{-1} + a_0 + a_1)}_{=0} (x + h) - h [a_{-1}x_i + a_0x_{i+1} + a_1x_{i+2}] = -h [h(a_0 + 2a_1)]$$

mit  $x_i = 0$ ,  $x_{i+1} = 0 + h$  und  $x_{i+2} = 0 + 2h$ . Normalisierung:

$$a_i = \frac{1}{h^2} \Rightarrow a_0 = -\frac{2}{h^2}, \quad a_{-1} = \frac{1}{h^2}$$

liefert schließlich die **B-Splines (Basis Splines)**:

$$(1.24) \quad \psi_i^B(x) = \begin{cases} 0 & , \quad x \leq x_{i-2} \\ \frac{(x - x_{i-2})^3}{6h^3} & , \quad \text{in } I_{i-1} \\ \frac{(x_i - x)^3}{2h^3} - \frac{(x_i - x)^2}{h^2} + \frac{2}{3} & , \quad \text{in } I_i \\ \frac{(x - x_i)^3}{2h^3} - \frac{(x - x_i)^2}{h^2} + \frac{2}{3} & , \quad \text{in } I_{i+1} \\ \frac{(x_{i+2} - x)^3}{6h^3} & , \quad \text{in } I_{i+2} \\ 0 & , \quad x \geq x_{i+2} \end{cases}$$

Bevor wir diese Splines weiter betrachten, noch einige andere Aussagen.

**Satz 1.4.6**

Sei  $f \in \mathbb{C}^4[a, b]$  und  $\psi$  ein interpolierender kubischer Spline,  $a = x_0, \dots, x_n = b$  (äquidistante Knoten). Dann existiert eine Konstante  $c > 0$  unabhängig von  $h = x_i - x_{i-1}$ , so daß gilt

$$\begin{aligned} \|f - \psi\|_\infty &\leq ch^4 \|f^{(4)}\|_\infty \\ \|f'' - \psi''\|_\infty &\leq 2h^2 \|f^{(4)}\|_\infty . \end{aligned}$$

Außerdem

$$\max_{k=1, \dots, n} |f''(x_k) - \psi''(x_k)| \leq \frac{3}{4} f^2 \|f^{(4)}\|_\infty$$



Es ist  $\det L \neq 0$  und somit ist  $\psi(x) = \sum_{\nu=-1}^5 A_\nu \psi_\nu^B(x)$  mit

$$A_0 = A_4 = -\frac{1}{4} + \frac{7}{48}\pi, \quad A_1 = A_3 = \frac{1}{2} - \frac{\pi}{24}, \quad A_2 = \frac{\pi}{48} - \frac{7}{4}$$

$$A_{-1} = A_1 - 2hf'(0) = \frac{1}{2} - \frac{\pi}{24} - \frac{\pi}{2}, \quad A_5 = A_{n+1} = A_{n-1} + 2hf'(\pi) = \frac{1}{2} - \frac{\pi}{24} - \frac{\pi}{2}$$

(b) Bestimme

$$\begin{aligned} \psi''\left(\frac{\pi}{2}\right) &= \sum_{\nu=-1}^5 A_\nu \psi_\nu^{B''}\left(\frac{\pi}{2}\right) = \sum_{\nu=1}^3 A_\nu \psi_\nu^{B''}\left(\frac{\pi}{2}\right) = \frac{1}{h^2}(A_1 - 2A_2 + A_3) \\ &= \frac{2}{h^2} \left( \frac{1}{2} - \frac{\pi}{24} - \frac{\pi}{48} + \frac{7}{4} \right) = \frac{2}{h^2} \left( \frac{9}{4} - \frac{\pi}{16} \right) = \frac{72}{\pi^2} - \frac{2}{\pi} = 6.659 \end{aligned}$$

(c) Berechne  $f''\left(\frac{\pi}{2}\right)$  und  $|f''\left(\frac{\pi}{2}\right) - \psi''\left(\frac{\pi}{2}\right)|$ .

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

$$f'(x) = \frac{1}{2}(3\cos 3x - \cos x), \quad f''(x) = -\frac{1}{2}(9\sin 3x - \sin x)$$

Wir erhalten

$$\left| f''\left(\frac{\pi}{2}\right) - \psi''\left(\frac{\pi}{2}\right) \right| = 1.659 =: a \quad \text{mit} \quad f''\left(\frac{\pi}{2}\right) = -\frac{1}{2}(-9 - 1) = 5$$

Desweiteren

$$f^{(3)}(x) = -\frac{1}{2}(27\cos 3x - \cos x), \quad f^{(4)}(x) = \frac{1}{2}(81\sin 3x - \sin x)$$

Mit  $\|f^{(4)}\|_{I,\infty} \leq \frac{1}{2}(81 + 1) = 41$  folgt

$$\|f'' - \psi''\|_{I,\infty} \leq 2\|f^{(4)}\|_{I,\infty} \cdot h^2 \leq 2\frac{\pi^2}{16} \cdot 41 = 50.581723$$

und so  $2\frac{\pi^2}{16}41 \cdot \frac{1}{a} = 30.489284$ .

### Definition 1.4.9

$A \in \mathbb{R}^{n \times n}$  heißt **streng diagonaldominant** genau dann, wenn gilt

$$\sum_{\substack{l=1 \\ l \neq k}}^n |a_{kl}| < |a_{kk}| \quad \text{für alle } 1 \leq k \leq n$$

Ist  $A$  streng diagonaldominant, so existiert seine Inverse  $A^{-1}$ .

### Lemma 1.4.10 (Eindeutigkeit)

Sei  $\psi \in S_h^3(I)$ . Es existiert wenigstens ein  $A \in \mathbb{R}^{n+3}$ ,  $A^T = (A_{-1}, A_0, A_1, \dots, A_n, A_{n+1})$  mit

$$\psi = \sum_{\nu=-1}^{n+1} A_\nu \psi_\nu^B \Big|_I.$$





mit  $\chi$  analog zu  $\phi$  und  $l(x) := ax + b$ .

Mit Lemma 1.4.11 folgt die Existenz von  $B \in \mathbb{R}^{n+3}$  mit  $l = B^T \psi^B$  in  $I$ . Somit ist

$$\psi = (M^T T^{-1} + B^T) \psi^B =: N^T \psi^B \Rightarrow \psi \in \text{span}\{\psi_{-1}^B, \dots, \psi_{n+1}^B\}$$

Aufgrund von Lemma 1.4.11 ist  $N$  eindeutig. ■

### Bemerkung 1.4.12

Auf eine ähnliche Art und Weise wie in Lemma 1.4.10 erhalten wir:

$$(1.29) \quad \frac{1}{6} \begin{pmatrix} -1 & 0 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & 1 & 4 & 1 & \\ & & & \ddots & \ddots & \\ & & & & 1 & 4 & 1 \\ & & & & & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} A_{-1} \\ A_0 \\ A_1 \\ \vdots \\ A_n \\ A_{n+1} \end{pmatrix} = \begin{pmatrix} \frac{h}{3} f'(x_0) \\ f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \\ \frac{h}{3} f'(x_n) \end{pmatrix}.$$

Es folgt  $\psi = A^T \psi^B$  (gute Kondition) !!

Wir definieren

$$(1.30) \quad L \begin{pmatrix} A_0 \\ \vdots \\ A_n \end{pmatrix} := \frac{1}{6} \begin{pmatrix} 4 & 2 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 \\ & & & & 2 & 1 \end{pmatrix} \begin{pmatrix} A_0 \\ \vdots \\ A_n \end{pmatrix} = \begin{pmatrix} f(x_0) + \frac{h}{3} f'(x_0) \\ f(x_1) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) - \frac{h}{3} f'(x_n) \end{pmatrix} =: g$$

mit  $A_{-1} = A_1 - 2hf'(x_0)$  und  $A_{n+1} = A_{n-1} + 2hf'(x_n)$ .

### Lemma 1.4.13

Sei  $A, \psi, M$  wie im Beweis zu Satz 1.4.3. Dann ist

$$(1.31) \quad LM = \frac{1}{h^2} \begin{pmatrix} 2[(f_1 - f_0) - hf'(x_0)] \\ f_2 - 2f_1 + f_0 \\ \vdots \\ f_n - 1f_{n-1} + f_{n-2} \\ 2[hf'(x_n) - (f_n - f_{n-1})] \end{pmatrix} =: d \in \mathbb{R}^{n+1}$$

**Beweis:**

$$\begin{aligned} h^2 M^T \phi &= h^2 \psi'' = h^2 A^T (\psi^B)'' \\ &= \begin{pmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_{n-1} \\ A_n \end{pmatrix} + \begin{pmatrix} A_{-1} \\ 0 \\ \vdots \\ 0 \\ A_{n+1} \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
& \stackrel{(1.30)}{=} \begin{pmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 2 & -2 \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_{n-1} \\ A_n \end{pmatrix} + 2h \begin{pmatrix} -f'(x_0) \\ 0 \\ \vdots \\ 0 \\ f'(x_n) \end{pmatrix} \\
& = \begin{pmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 2 & -2 \end{pmatrix} \cdot L^{-1}g + 2h \begin{pmatrix} -f'(x_0) \\ 0 \\ \vdots \\ 0 \\ f'(x_n) \end{pmatrix} \\
& = 6(L - I)L^{-1}g + 2h \begin{pmatrix} -f'(x_0) \\ 0 \\ \vdots \\ 0 \\ f'(x_n) \end{pmatrix} .
\end{aligned}$$

Woraus die Behauptung leicht folgt. ■

Bedeutung von (1.31):

Linearkombination von  $M_i$ 's (in mittleren Zeilen) =  $2^{nd}$  (symmetrisch) Differenzenquotienten.

#### Lemma 1.4.14

$$\|L^{-1}\|_{\infty} \leq 3 \quad (L^{-1} \text{ existiert z.B. mit Gerschgorin})$$

#### Beweis:

Sei  $x, y \in \mathbb{R}^n$  und  $\|x\|_{\infty} = |x_j|$  mit  $Lx = y$ . Dann ist

$$6|y_j| = |4x_j + 2x_{j\pm 1}| \text{ or } |4x_j + x_{j-1} + x_{j+1}| \geq (4 - 2)|x_j| = 2|x_j|$$

woraus schließlich die Behauptung folgt. ■

## Chapter 2

# Lineare Gleichungssysteme, direkte Verfahren

### 2.1 Vorbemerkungen

Sei  $A$  eine  $(m, n)$ -Matrix.

$$A = (a_{ik})_{\substack{i=1,\dots,m \\ k=1,\dots,n}} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

mit  $a_{ik} \in \mathbb{R}$ . ( $A \in \mathbb{R}^{m \times n}$ ) (evtl. auch  $a_{ik} \in \mathbb{C}$ , d.h.  $A \in \mathbb{C}^{m \times n}$ , bei LGS selten (evtl. Aufteilen in Real- und Imaginär-Teil), aber bei Eigenwertaufgaben).

Der Vektor

$$\mathbf{r} = (r_i)_{i=1,\dots,m} = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} \in \mathbb{R}^m$$

sei die rechte Seite des folgenden Linearen Gleichungssystems bei dem der Vektor  $\mathbf{x}$  gesucht ist.

$$(2.1) \quad A \mathbf{x} = \mathbf{r}$$

Dies läßt sich schreiben als:

$$(2.2) \quad \begin{array}{rcl} a_{11}x_1 + \dots + a_{1n}x_n & = & r_1 \\ & & \vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n & = & r_m \end{array}$$

Das zugehörige homogene LGS ist

$$(2.3) \quad A \mathbf{x} = \mathbf{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^m$$

Aus der Linearen Algebra ist bekannt:

Hat (2.3) genau  $\rho$  linear unabhängige Lösungen, z.B.  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\rho)} \in \mathbb{R}^n$  und existiert eine spezielle Lösung  $\mathbf{x}^{(0)}$  von (2.1), so ergibt sich jede Lösung von (2.1) in der Form

$$\mathbf{x}^{(0)} + c_1 \mathbf{x}^{(1)} + \dots + c_\rho \mathbf{x}^{(\rho)}$$

mit  $c_i \in \mathbb{R}$  beliebig.

Ist  $\mathbf{m} = \mathbf{n}$ , also  $A \in \mathbb{R}^{n \times n}$ , so gilt folgendes ( $D := \det A$ ):

$$\begin{aligned} D \neq 0 &\Leftrightarrow (2.3) \text{ hat nur die triviale Lösung } \mathbf{x} = \mathbf{0} \\ &\Leftrightarrow (2.1) \text{ ist für jedes } r \in \mathbb{R}^n \text{ eindeutig lösbar.} \end{aligned}$$

$$\begin{aligned} D = 0 &\Leftrightarrow (2.3) \text{ hat } (0 <) \rho \leq n \text{ linear unabhängige Lösungen} \\ &\text{und (2.1) hat unendlich viele Lösungen oder ist unlösbar} \end{aligned}$$

Für den Fall  $D \neq 0$  gilt außerdem:  $\mathbf{x} = A^{-1} \mathbf{r}$  ist in geschlossener Form darstellbar.

Cramersche Regel:

$$x_k = \frac{D_k}{D}$$

dabei geht  $D_k$  aus  $D$  durch Ersetzen der  $k$ -ten Spalte durch  $\mathbf{r}$  hervor. Ungünstig für die Praxis.

## 2.2 Der Gaußsche Algorithmus

Zunächst betrachten wir das lineare Gleichungssystem (2.2) mit  $m = n$  und  $\det A \neq 0$ .

Sei  $a_{11} \neq 0$ . Dann gilt für  $i = 2, \dots, n$

$$\text{neue } i\text{-te Zeile} = \text{alte } i\text{-te Zeile} - \frac{a_{i1}}{a_{11}} \cdot 1.\text{Zeile}$$

Somit

$$(2.4) \quad \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = r_1 \\ 0 + a'_{22}x_2 + \dots + a'_{2n}x_n = r'_2 \\ \vdots \\ 0 + a'_{n2}x_2 + \dots + a'_{nn}x_n = r'_n \end{array}$$

Das Verfahren wird auf das  $(n-1) \times (n-1)$  LGS fortgesetzt. Voraussetzung dafür ist wiederum  $a'_{22}$ . Ist dies nicht der Fall, so wird eine Zeilenvertauschung durchgeführt, um diesen Zustand zu erreichen.

Schließlich erhält man ein LGS mit  $\Delta$ -Matrix:

$$(2.5) \quad \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = r_1 \\ a'_{22}x_2 + \dots + a'_{2n}x_n = r'_2 \\ \vdots \\ a^{(n-1)}_{nn}x_n = r^{(n-1)}_n \end{array}$$

Jetzt werden die  $x_i$  in der Reihenfolge  $x_n, x_{n-1}, \dots, x_1$  rückwärts bestimmt.

### Bemerkung 2.2.1

Der Gauß-Algorithmus hat in dieser Form  $\approx \frac{n^3}{3}$  Multiplikations und Divisions Operationen. Die Cramersche Regel inklusive der Entwicklung der Determinanten  $D_0, D_1, \dots, D_n$  benötigt  $\approx \frac{n^4}{3}$  Operationen.

### 2.2.1 Methoden zur Rechenstabilität

#### Skalierung

Wenn die Elemente von  $A$  nicht die gleiche Größenordnung haben, dann werden Zeilen und Spalten von  $A$  mit geeigneten Faktoren multipliziert (bei Spalten:  $x_i \rightarrow \alpha x_i = \hat{x}_i$ ). D.h. Es werden zwei Diagonalmatrizen  $D_1, D_2$  bestimmt mit  $A' = D_1 A D_2$ . Angestrebt wird dabei für alle  $i, l = 1, \dots, n$ :

$$\sum_{k=1}^n |a'_{ik}| \approx \sum_{j=1}^n |a'_{jl}| \quad .$$

$A'$  heißt dann **äquilibriert**. Die Bestimmung von  $D_1$  und  $D_2$  ist allerdings im allgemeinen schwierig. Daher wählt man der Einfachheit halber  $D_2 = E$  und

$$(2.6) \quad D_1 = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{pmatrix} \quad \text{mit} \quad d_i = \frac{1}{\sum_{k=1}^n |a_{ik}|}$$

#### Pivotwahl

Elemente von  $A$  haben bereits gleiche Größenordnung. Sei zum Beispiel

$$0 < |a_{11}| \ll |a_{i1}| \quad (i = 2, \dots, n)$$

Dann ist Gauß in der herkömmlichen Form (2.4)/(2.5) möglich (**natürliche Pivotwahl**). Da sich jedoch bei der Division eine große Fehlerfortpflanzung ergibt, verwendet man häufiger die folgenden Methoden.

- **Partielle Pivotwahl**

In der 1.Spalte von  $A$  (bzw. in den weiteren Schritten im jeweiligen "Rest" des LGS) wird das betragsgrößte Element (**Pivotelement**) gesucht und (jeweils) die 1.Zeile mit der Zeile des Pivotelements vertauscht.

- **Totale Pivotwahl**

In  $A$  (bzw. in den weiteren Schritten im jeweiligen "Rest" des LGS) wird das betragsgrößte Element gesucht und die jeweilige 1.Spalte und 1.Zeile mit Spalte und Zeile des Pivotelements vertauscht. (bei Spaltenvertauschung auf Gesamtmatrix ausdehnen)

#### Bemerkung 2.2.2

*Die totale Pivotwahl ist im allgemeinen stabiler, aber aufwendiger. Daher wird meistens nur die partielle Pivotwahl verwendet.*

### 2.2.2 Weitere Bemerkungen zu Gauß

#### 1) Berechnung der Determinante

Ist die Determinante von  $A$  ungleich Null, so sind bei Gauß zwei Operationen möglich.

- 1) Addition und Subtraktion von Vielfachen von Zeilen zu anderen Zeilen verändern die Determinante nicht.
- 2) Jede Zeilen- bzw. Spaltenvertauschung ändert die Determinante um den Faktor  $-1$ .

Aus (2.5) ergibt sich mit zusätzlichen Vertauschungen:

$$(2.7) \quad \det A = (-1)^\delta a_{11} a'_{22} \cdots a_{nn}^{(n-1)}$$

wobei  $\delta$  die Gesamtzahl der Vertauschungen ist.

## 2) Mehrere rechte Seiten, Inversenbestimmung

Schema für Gauß:

$$\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & r_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & r_n \end{array} \left\| \begin{array}{c} s_1 \\ \vdots \\ s_n \end{array} \right.$$

z.B. mit partieller Pivotwahl führt über (2.4) schließlich auf

$$\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & r_1 \\ a'_{22} & \cdots & a'_{2n} & r'_2 \\ & \ddots & \vdots & \vdots \\ & & a_{nn}^{(n-1)} & r_n^{(n-1)} \end{array} \left\| \begin{array}{c} s_1 \\ s'_2 \\ \vdots \\ s_n^{(n-1)} \end{array} \right.$$

So ist es möglich, eine neue rechte Seite  $\mathbf{s}$  in das Schema einzufügen, so daß die Dreieckszerlegung von  $A$  nur einmal durchgeführt werden muß.

### Beispiel 2.2.3

Es soll  $A^{-1}(=: X)$  durch Lösen des Matrix-LGS  $A \cdot X = E$  bestimmt werden. D.h.

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

wobei die Spalten von  $X$  die Vektoren  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  bilden. Somit ist zu lösen:

$$A\mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

d.h.

$$\begin{pmatrix} a_{11}x_{11} + \cdots + a_{1n}x_{n1} = 1 \\ a_{21}x_{11} + \cdots + a_{2n}x_{n1} = 0 \\ \vdots \\ a_{n1}x_{11} + \cdots + a_{nn}x_{n1} = 0 \end{pmatrix}$$

Entsprechend für die Vektoren  $\mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ .

Es werden hierbei nur  $\approx 4/3n^3$  Operationen benötigt anstelle von  $\frac{n^4}{3}$ .



Insbesondere gilt für  $A = C_p$ :

$$C_p^2 = C_p \cdots C_p = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & 2c_{p+1,p} & 1 & & \\ & & \vdots & & \ddots & \\ & & 2c_{np} & & & 1 \end{pmatrix}$$

Somit  $E + C_p^2 = 2C_p$ , bzw.:

$$C_p^{-1} = 2E - C_p = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -c_{p+1,p} & 1 & & \\ & & \vdots & & \ddots & \\ & & -c_{np} & & & 1 \end{pmatrix}$$

Bevor wir den nächsten Satz beweisen noch ein paar Definitionen. Eine Matrix  $P$ , die in jeder Zeile und in jeder Spalte genau eine "1" und sonst Nullen besitzt, heißt **Permutationsmatrix**. Die Zerlegung  $PA = LR$  heißt **verallgemeinerte LR-Zerlegung** von  $A$ .

#### Satz 2.2.4

Jede reguläre Matrix  $A$  besitzt eine verallgemeinerte LR-Zerlegung, d.h. zu  $A$  existieren eine Permutationsmatrix  $P$ , eine untere Dreiecksmatrix  $L$  und eine obere Dreiecksmatrix  $R$ , so daß gilt:  $PA = LR$ .

#### Beweis:

Mit den Frobenius-Matrizen läßt sich der Beweis nun wie folgt führen.

Die nach dem ersten Gauß-Schritt erhaltene Matrix bezeichnen wir mit  $B$ , also

$$1.\text{Schritt : } B := E_{1p}A = (b_{ik})$$

Für  $c_{i1} = \frac{b_{i1}}{b_{11}}$ , ( $i = 2, \dots, n$ ) sieht nun der weitere Schritt so aus:

$$2.\text{Schritt : } C_1^{-1}B = C_1^{-1}E_{1p}A$$

Dieses Vorgehen setzt man weiter fort, also

$$1.\text{Schritt : } E_{2q}(C_1^{-1}E_{1p}A)$$

$$2.\text{Schritt : } C_2^{-1}E_{2q}(C_1^{-1}E_{1p}A)$$

und so weiter.

Theoretisch ist es möglich alle notwendigen Zeilenvertauschungen vor sämtlichen Eliminationen durchzuführen. Das heißt, zunächst bildet man

$$E_{n-1,s}E_{n-2,t} \cdots E_{2q}E_{1p} \cdot A = PA$$

Dann wird der Gauß-Algorithmus mit der natürlichen Pivotwahl durchgeführt.

$$C_{n-1}^{-1}C_{n-2}^{-1}\cdots C_1^{-1}PA = \begin{pmatrix} * & \cdots & * \\ 0 & & \vdots \\ \vdots & \ddots & \ddots \\ 0 & \cdots & 0 & * \end{pmatrix} = R$$

(Wegen  $\det A \neq 0$  muß in der Restmatrix in der ersten Spalte ein Element  $\neq 0$  sein!) Somit ergibt sich:

$$PA = C_1C_2\cdots C_{n-1}R =: LR$$

wobei

$$C_1C_2 = \begin{pmatrix} 1 & & & & \\ c_{21} & 1 & & & \\ \vdots & & \ddots & & \\ c_{n1} & & & \ddots & \\ & & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ \vdots & c_{32} & 1 & & \\ \vdots & & & \ddots & \\ 0 & c_{n2} & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ c_{21} & 1 & & & \\ \vdots & c_{32} & 1 & & \\ \vdots & & & \ddots & \\ c_{n1} & c_{n2} & & & 1 \end{pmatrix}$$

Führt man diese Matrixmultiplikationen für alle  $C_i, i = 1, \dots, n$  durch, so erhält man

$$L = \begin{pmatrix} 1 & & & & \\ c_{21} & 1 & & & \\ \vdots & c_{32} & 1 & & \\ \vdots & & \ddots & \ddots & \\ c_{n1} & c_{n2} & \cdots & c_{n,n-1} & 1 \end{pmatrix}$$

d.h.  $L$  enthält gerade die Eliminationsfaktoren. ■

#### 4) Der Fall $\det A = 0$ , rechteckige Matrizen und Rangbestimmung

In diesem Fall wird der Gauß-Algorithmus entsprechend, notfalls mit der totalen Pivotwahl, durchgeführt bis nur noch Nullzeilen übrig bleiben.

##### Beispiel 2.2.5

$$A = \begin{pmatrix} 1 & 1 & 2 & 0 \\ 1 & 2 & 3 & \boxed{6} \\ 2 & 3 & 5 & 6 \end{pmatrix}$$

Gesucht ist der **Rang** der Matrix  $A$ , wobei dieser die maximale Zahl der linear unabhängigen Zeilen bzw. Spalten ist. Der Rang einer Matrix ist invariant gegenüber den Gauß-Operationen. Hier wenden wir Gauß mit totaler Pivotwahl an. Das Pivotelement ist in der obigen Matrix schon markiert. Es sind also die Zeilen 1. und 2. sowie die Spalten 1. und 4. gegeneinander

zu vertauschen.

$$\begin{array}{cccc|l}
 6 & 2 & 3 & 1 & \cdot(-1) \\
 0 & 1 & 2 & 1 & \downarrow + \\
 6 & 3 & 5 & 2 & \cdot(1) \\
 \hline
 6 & 2 & 3 & 1 & \\
 0 & \boxed{1} & 2 & 1 & 2. \text{ und } 3. \text{ Spalte vertauschen} \\
 0 & \boxed{1} & 2 & 1 & \\
 \hline
 6 & 3 & 2 & 1 & \\
 0 & 2 & 1 & 1 & \cdot(-1) \\
 0 & 2 & 1 & 1 & \leftrightarrow + \\
 \hline
 6 & \boxed{3} & 2 & 1 & \\
 0 & \boxed{2} & 1 & 1 & \\
 0 & 0 & 0 & 0 & 
 \end{array}$$

Wir erhalten zwei linear unabhängige Zeilen, folglich ist  $\text{Rang}(A) = 2$ , d.h. der Rang einer Matrix ist jeweils die Reihenzahl der in der Nord-West-Ecke stehenden regulären Matrix.

### 2.3 LGS mit positiv definiten Matrizen

#### Definition 2.3.1

Sei  $A \in \mathbb{R}^{n \times n}$ .  $A$  heißt **symmetrisch** genau dann, wenn für alle  $i, k$  gilt:  $a_{ik} = a_{ki}$  ( $\Leftrightarrow A = A^T$ ). Die symmetrische Matrix  $A$  ist **positiv definit** genau dann, wenn für alle  $\mathbf{x} \in \mathbb{R}^n$  mit  $\mathbf{x} \neq 0$  gilt:  $\mathbf{x}^T(A\mathbf{x}) > 0$ . Die symmetrische Matrix  $A$  ist **positiv semidefinit** genau dann, wenn für alle  $\mathbf{x} \in \mathbb{R}^n$  mit  $\mathbf{x} \neq 0$  gilt:  $\mathbf{x}^T(A\mathbf{x}) \geq 0$ .

#### Bemerkung 2.3.2

Für positiv definite Matrizen ist die Hauptdiagonale positiv:  $a_{ii} = \mathbf{e}_i^T A \mathbf{e}_i > 0$ , wobei  $\mathbf{e}_i$  der  $i$ -te Einheitsvektor ist.

Unsere erste Aussage ist eine Verallgemeinerung, die für Anwendungen jedoch unhandlich ist.

#### Lemma 2.3.3 (Kriterium von Schur)

Die symmetrische Matrix  $A$  ist genau dann positiv definit, wenn  $a_{11} > 0$  und alle Unterdeterminanten größer Null sind, d.h.

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & \cdots & a_{13} \\ \vdots & & \vdots \\ a_{31} & \cdots & a_{33} \end{vmatrix} > 0, \quad \dots, \quad \det A > 0$$

#### Bemerkung 2.3.4

$A$  positiv definit  $\Leftrightarrow$  alle EW von  $A$  sind größer Null.

#### 2.3.1 Die Cholesky-Zerlegung, Bandmatrizen

Es sei das lineare Gleichungssystem  $A\mathbf{x} = \mathbf{r}$  gegeben mit  $A$  positiv definit. Hier ist eine LR-Zerlegung mit  $P = E$  möglich:  $A = LR$ . Allerdings würde dabei die Symmetrie zerstört

werden. Die folgende Zerlegung dagegen bewahrt diese Eigenschaft.

$$(2.10) \quad A = S^T S \quad \text{mit} \quad S = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ & \ddots & \vdots \\ & & s_{nn} \end{pmatrix}, \quad s_{ii} > 0 \quad (i = 1, \dots, n)$$

Diese Zerlegung heißt **Cholesky-Zerlegung**.

Um diese Zerlegung zu erhalten, führen wir zuerst einen Gauß-Schritt durch:

$$C_1^{-1}A = \left( \begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & & \\ \vdots & & * & \\ 0 & & & \end{array} \right)$$

Nun multiplizieren wir diesen Ausdruck mit  $(C_1^{-1})^T$ . Da  $a_{ik} = a_{ki}$  folgt

$$B := C_1^{-1}A(C_1^{-1})^T = \left( \begin{array}{c|ccc} a_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & A^{(1)} & \\ 0 & & & \end{array} \right)$$

Aufgrund der Symmetrie von  $A$  ist auch  $B$  symmetrisch. Außerdem gilt:

**Beh.:**  $B$  und  $A^{(1)}$  sind wiederum positiv definit.

**Beweis:**

Sei  $\mathbf{x} \neq 0$ . Dann folgt mit  $\det(C_1^{-1})^T = 1$  aus

$$\mathbf{x}^T B \mathbf{x} = \mathbf{y}^T A \mathbf{y}$$

daß  $\mathbf{y} := (C_1^{-1})^T \mathbf{x} \neq 0$ . Da  $A$  positiv definit ist folgt  $\mathbf{y}^T A \mathbf{y} > 0$  und somit die Behauptung. Analog für  $A^{(1)}$ . ■

Setzen wir das Verfahren entsprechend fort, so erhalten wir

$$\dots C_2^{-1} C_1^{-1} A (C_1^{-1})^T (C_2^{-1})^T \dots = D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \quad \text{mit} \quad d_i > 0$$

Definieren wir  $C := C_1 C_2 \cdots C_{n-1}$ , so folgt

$$C^{-1} A (C^{-1})^T = D \quad \Rightarrow \quad A = C D C^T$$

Setzen wir

$$D^{1/2} = \begin{pmatrix} +\sqrt{d_1} & & \\ & \ddots & \\ & & +\sqrt{d_n} \end{pmatrix}$$

so ist

$$(2.11) \quad A = C D^{1/2} D^{1/2} C^T = (C D^{1/2})(C D^{1/2})^T =: S^T S$$

die **Cholesky-Zerlegung von  $A$** .

Ist umgekehrt  $A = S^T S$  die Cholesky-Zerlegung von  $A (= A^T)$ , so ist  $A$  positiv definit.

**Beweis:**

Es ist

$$\mathbf{x}^T A \mathbf{x} = \mathbf{x}^T S^T S \mathbf{x} =: \mathbf{y}^T \mathbf{y} \geq 0$$

Im Falle der Gleichheit ist dies äquivalent zu  $\mathbf{y} = 0$ . Daraus folgt  $\mathbf{x} = S^{-1} \mathbf{y} = 0$  und somit die Behauptung. ■

Festgehalten ist dieser Zusammenhang in folgendem Satz.

**Satz 2.3.5**

Die symmetrische Matrix  $A$  ist genau dann positiv definit, wenn Sie eine Cholesky-Zerlegung besitzt.

Eine Formel für die Cholesky-Zerlegung erhält man entweder über (2.11) oder direkt:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} s_{11} & & \\ \vdots & \ddots & \\ s_{1n} & \cdots & s_{nn} \end{pmatrix} \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ & \ddots & \vdots \\ & & s_{nn} \end{pmatrix}$$

Mit den Überlegungen

$$\begin{aligned} (0 <) a_{11} &= s_{11}^2 &\Rightarrow s_{11} &= +\sqrt{a_{11}} > 0 \\ a_{12} &= s_{12} \cdot s_{11} &\Rightarrow s_{12} &= \frac{a_{12}}{s_{11}} \end{aligned}$$

und so weiter, ergibt sich folgendes Iterationsschema

$$\begin{aligned} s_{ii}^2 &= a_{ii} - s_{1i}^2 - \dots - s_{i-1,i}^2 && (i = 1, \dots, n) \\ (2.12) \quad s_{ik} &= \frac{(a_{ik} - s_{1i}s_{1k} - \dots - s_{i-1,i}s_{i-1,k})}{s_{ii}} && (k > i; i = 1, \dots, n-1) \end{aligned}$$

Nun will man das Verfahren auf die rechte Seite des LGS übertragen, d.h.  $A\mathbf{x} = \mathbf{r}$  soll äquivalent sein zu  $S\mathbf{x} = \boldsymbol{\rho} = (\rho_i)$ . Multipliziert man diese LGS mit  $S^T$ , so ergibt sich

$$A\mathbf{x} = S^T S \mathbf{x} = S^T \boldsymbol{\rho} = \mathbf{r}.$$

Somit gilt für die rechte Seite in diesem Verfahren:

$$(2.13) \quad \rho_i = \frac{(r_i - s_{1i}\rho_1 - \dots - s_{i-1,i}\rho_{i-1})}{s_{ii}} \quad (\text{wie } s_{ik} (k > i) \text{ ohne Index } k)$$

Die  $x_i$  werden dann aus  $S\mathbf{x} = \boldsymbol{\rho}$  rekursiv wie in (2.13) berechnet.

**Bemerkung 2.3.6**

Hier noch ein paar Bemerkungen zu Cholesky.

1. Das Verfahren besitzt im allgemeinen eine gute numerische Stabilität. Unter anderem auch wegen der Quadratwurzeln bei  $s_{ii}$ .
2. Im allgemeinen ist eine Pivotwahl nicht nötig. Diese könnte, wegen der Symmetriehaltung, auch nur auf der Diagonalen durchgeführt werden (Zeilen und gleiche Spalten vertauschen).

3. Der Arbeitsaufwand ist bei Cholesky nur halb so groß wie bei Gauß.
4. Die Ausdehnung auf symmetrische, nicht positiv definite Matrizen ist unter Umständen möglich:  
Ist  $s_{ii}^2 < 0$ , so ist  $s_{ii}$  rein imaginär. Somit sind alle  $s_{ik}$  der  $i$ -ten Zeile und  $\rho_i$  imaginär. Daraus folgt  $x_i \in \mathbb{R}$ .
5. Auch im unsymmetrischen Fall existieren Ansätze der Form  $A = LR$ , woraus dann entsprechend (???)  $L$  und  $R$  rekursiv berechnet werden. (vgl. Banachiewicz, Chout). Der Arbeitsaufwand ist allerdings der gleiche wie bei Gauß. Eventuell ist dieses Verfahren für die Handrechnung geeignet.

### Bandmatrizen (Anwendungen!)

Die folgende Matrix wird als **Tridiagonalmatrix** bezeichnet.

$$\begin{pmatrix} * & * & & & \\ * & \ddots & \ddots & & \\ & \ddots & & * & \\ & & & * & * \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Entsprechend werden Diagonalmatrizen mit weiteren Nebendiagonalen bezeichnet. D.h. die Diagonale und jeweils die gleiche Anzahl von Nebendiagonalen über und unter der Hauptdiagonalen sind besetzt, der Rest ist Null.

Ist speziell  $A = A^T$  eine positiv definite Bandmatrix, so besitzt die Cholesky-Zerlegung dieselbe Struktur.

Zum Beispiel:

$$A := \begin{pmatrix} * & * & * & & \\ * & * & \ddots & \ddots & \\ * & \ddots & \ddots & & * \\ & \ddots & & * & \\ & & * & * & * \end{pmatrix} = \begin{pmatrix} * & & & & \\ * & \ddots & & & \\ * & \ddots & & & \\ & \ddots & & & \\ & & * & * & * \end{pmatrix} \begin{pmatrix} * & * & * & & \\ & \ddots & \ddots & \ddots & \\ & & & & * \\ & & & & * \\ & & & & * \end{pmatrix} =: S^T S$$

Entsprechend auch im unsymmetrischen Fall. (vgl.(5))

$$A = LR = \begin{pmatrix} 1 & & & & \\ l_1 & \ddots & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & l_{n-1} & 1 \end{pmatrix} \begin{pmatrix} m_1 & r_1 & & & \\ & \ddots & \ddots & & \\ & & & r_{n-1} & \\ & & & & m_n \end{pmatrix}$$

Wird die LR-Zerlegung ohne Pivotwahl durchgeführt, so ist der Arbeitsaufwand für das LGS  $\mathcal{O}(n^3)$ ! Mit Pivotwahl ist dieses Verfahren etwas aufwendiger. Auch in Spezialfällen wie z.B. in dem Fall, in dem die Elemente der Nebendiagonalen als Pivotelemente verwendet werden, ist der Arbeitsaufwand höher (siehe Literatur).

## Chapter 3

# Vektor- und Matrixnormen, Fehlerabschätzungen für LGS, iterative Verfahren, nicht lineare Gleichungssysteme

### 3.1 Vektor- und Matrixnormen

Der Raum  $\mathbb{R}^n$  ist mit der **Euklidischen Vektor-Norm**

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \quad \text{für } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

normiert. Dadurch ist ein **Abstandsbegriff** gegeben:

$$\|\mathbf{x} - \mathbf{y}\|_2$$

Die allgemeine Definition der Norm auf dem  $\mathbb{R}^n$  lautet wie folgt.

#### **Definition 3.1.1**

Die Funktion  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  heißt **Norm** auf  $\mathbb{R}^n$ , wenn gilt

(N1) **Definitheit** :

$$\forall \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \geq 0 \text{ und } \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$$

(N2) **positive Homogenität** :

$$\forall \mathbf{x} \in \mathbb{R}^n \forall \alpha \in \mathbb{R} : \|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$$

(N3) **Dreieck-Ungleichung** :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

**Beispiel 3.1.2**

$$(3.1) \quad \|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \quad (p \geq 1 \text{ reell})$$

Die wichtigsten Fälle sind

$$(3.2) \quad \mathbf{p} = \mathbf{1} : \quad \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

$$(3.3) \quad \mathbf{p} = \mathbf{2} : \quad \|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

$$(3.4) \quad \mathbf{p} = \infty : \quad \|\mathbf{v}\|_\infty = \max_{i=1}^n |x_i|$$

Läßt man in (3.1)  $p$  gegen unendlich laufen, so erhält man (3.4). Für die Numerik ist die **Maximums-Norm** (3.4) wichtiger als die **euklidische Norm** (3.3). Aber für Konvergenzaussagen werden auch andere Normen betrachtet.

Um zu beweisen, daß (3.1) eine Norm ist, sind die Eigenschaften (N1), (N2) und (N3) nachzurechnen. Dies geschieht hier mit Hilfe der Hölder-Ungleichung. Diese entspricht im Fall  $p = 2$  der Cauchy-Schwarz Ungleichung.

**Beispiel 3.1.3**

Es sei das LGS  $A\mathbf{x} = \mathbf{r}$  gegeben. Die Lösung soll näherungsweise bestimmt werden, d.h. es treten Rundungsfehler auf. Die Näherungslösung wird mit  $\tilde{\mathbf{x}}$  bezeichnet.

Man sagt, daß  $\tilde{\mathbf{x}}$  mit einem Fehler  $\leq \varepsilon (\varepsilon \in \mathbb{R}, > 0)$  behaftet ist.

$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty \leq \varepsilon$  : Jede Komponente hat einen Fehler vom Betrag  $\leq \varepsilon$ ,  
d.h. für alle  $i$  :  $|x_i - \tilde{x}_i| \leq \varepsilon$

$\|\mathbf{x} - \tilde{\mathbf{x}}\|_p \leq \varepsilon$  : Die Einzelfehler werden hier "gemischt". Dadurch werden im allgemeinen die Einzelfehler überschätzt:

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty \leq \|\mathbf{x} - \tilde{\mathbf{x}}\|_p$$

**3.1.1 Konvergenz von Folgen im  $\mathbb{R}^n$** 

Die Folge  $(\mathbf{x}^{(k)})_{k \in \mathbb{N}}$  konvergiert koordinatenweise gegen  $\mathbf{x} = (x_i) \in \mathbb{R}^n$ , wenn für alle  $i$  gilt

$$x_i^{(k)} \rightarrow x_i$$

Von einer **Konvergenz bzgl  $\|\cdot\|$**  spricht man, wenn zu jedem  $\varepsilon > 0$  ein  $M \in \mathbb{N}$  existiert, so daß für alle  $k \geq M$  gilt:

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| < \varepsilon$$

**Lemma 3.1.4**

Die Konvergenz bezüglich einer beliebigen Vektornorm und die koordinatenweise Konvergenz sind (im  $\mathbb{R}^n$ ) gleichwertig, d.h. keine neue Geometrie.

Ein allgemeiner Beweis für die Gleichwertigkeit der Konvergenz bzgl. (3.1), (3.4) aus

$$\|\mathbf{x}\|_{\text{inf ty}} \leq \|\mathbf{x}\|_p \leq n^{1/p} \|\mathbf{x}\|_{\text{inf ty}}$$

ist etwas aufwendig aufgrund der folgenden Überlegungen.

Ist  $\mathbf{x} = 0$ , so ist die Aussage trivial. Sei also  $\mathbf{x} \neq 0$ . Dann gilt

$$\|\mathbf{x}\|_\infty = \max_i (|x_i|) \neq 0 \quad (1 \leq i \leq n).$$

Somit

$$\|\mathbf{x}\|_p = |x_i| \left( \sum_{j=1}^n \underbrace{\left| \frac{x_j}{x_i} \right|^p}_{\leq 1} \right)^{1/p} \leq n^{1/p} \|x_i\|_\infty$$

wobei

$$1 \leq \sum_{j=1}^n \left| \frac{x_j}{x_i} \right|^p \leq n$$

**Bemerkung 3.1.5**

Die Vektor-Normen lassen sich entsprechend auch in  $\mathbb{C}^n$  definieren. Bei der Eigenschaft (N2) ist dann  $\alpha \in \mathbb{C}$ . Die  $p$ -Normen werden entsprechend übertragen.

**3.1.2 Matrix-Normen**

**Definition 3.1.6**

Sei  $A \in \mathbb{R}^{n \times n}$ . Eine **Matrix-Norm** ist eine Vektor-Norm auf dem  $\mathbb{R}^{n \times n}$  mit der zusätzlichen Forderung:

$$(3.5) \quad \|AB\| \leq \|A\| \cdots \|B\| \quad \forall A, B \in \mathbb{R}^{n \times n}$$

**Bemerkung 3.1.7**

Für Matrizen  $A, B \in \mathbb{C}^{n \times n}$  läßt sich die Definition analog formulieren. (Wichtig insbesondere im Zusammenhang mit EW und auch bei reellen Problemen)

**Beispiel 3.1.8**

(1) Wird der Operator  $\|\cdot\|_\infty$  auf  $\mathbb{R}^{n \times n}$  übertragen, so liefert er keine Matrix-Norm:

$$A \cdot A := \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

Es ist  $\|A\|_\infty = 1$  und  $\|A^2\|_\infty = 2$ . Somit ist (3.5) nicht erfüllt.

(2) Überträgt man  $\|\cdot\|_2$  auf  $\mathbb{R}^{n \times n}$  so erhält man die sogenannte **Frobenius-Norm** (auch **E-Schicht-Norm** genannt).

$$\|A\|_{ES} = \sqrt{\sum_{i,k} a_{ik}^2}$$

Der Operator  $\|\cdot\|_2$  liefert zwar eine Matrix-Norm, ist allerdings aufgrund von  $\|E\|_{ES} = \sqrt{n}$  ungünstig. Erstrebenswerter ist dagegen  $\|E\| = 1$ .

Beispiele von Matrix-Normen:

$$(3.6) \quad \|A\|_Z = \max_{i=1}^n \sum_{k=1}^n |a_{ik}| \quad \text{Zeilensummennorm}$$

$$(3.7) \quad \|A\|_S = \max_{k=1}^n \sum_{i=1}^n |a_{ik}| \quad \text{Spaltensummennorm}$$

Die Eigenschaften (N1), (N2) und (N3) sowie (3.5) sind über die Dreiecksungleichung nachzuweisen.

### Definition 3.1.9

Seien  $A = (a_{ik}), B = (b_{ik}) \in \mathbb{R}^{n \times n}$  beliebige Matrizen. Gilt  $|a_{ik}| \leq |b_{ik}|$ , so heißt die Matrix-Norm  $\|\cdot\|_M$  **monoton**.

### Beispiel 3.1.10

$\|\cdot\|_Z$  und  $\|\cdot\|_S$  sind monoton. Die Spektralnorm ist es nicht.

## 3.1.3 Einführung der Spektralnorm

### Definition 3.1.11

Ist  $A \in \mathbb{R}^{n \times n}$  (bzw.  $\mathbb{C}^{n \times n}$ ) und sind  $\lambda_1, \dots, \lambda_n \in \mathbb{C}$  die Eigenwerte von  $A$  (entsprechend ihrer Vielfachheit), so heißt

$$(3.8) \quad \rho(A) = \max_{i=1}^n |\lambda_i| \in \mathbb{R}$$

der **Spektralradius** von  $A$  (keine Norm).

### Bemerkung 3.1.12

Ist  $\|\cdot\|$  eine beliebige Matrix-Norm, so gilt für jede Matrix  $A \in \mathbb{R}^{n \times n}$  ( $\in \mathbb{C}^{n \times n}$ )

$$(3.9) \quad \rho(A) \leq \|A\|$$

### Beweis:

Sei  $A\mathbf{x} = \lambda\mathbf{x}$  mit  $\mathbf{x} \neq 0$  als Eigenvektor zu  $\lambda$ . Sei weiter  $\lambda \in \mathbb{R}$ ,  $\mathbf{x} = (x_1, \dots, x_n)^T$  und

$$\begin{pmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ x_n & 0 & \cdots & 0 \end{pmatrix} =: X \in \mathbb{R}^{n \times n}$$

mit  $X \neq 0$ . Dann gilt  $A\mathbf{x} = \lambda\mathbf{x} \Leftrightarrow AX = \lambda X$  woraus folgt

$$\|\lambda X\| \stackrel{(N2)}{=} |\lambda| \|X\| = \|AX\| \stackrel{(3.5)}{\leq} \|A\| \|X\|.$$

Erweitert man diese Ungleichung so erhält man

$$\frac{1}{\|X\|} \cdot |\lambda| \|X\| = |\lambda| \leq \|A\|$$

und somit  $\rho(A) \leq \|A\|$ . ■

Läßt sich in (3.9) auch die Gleichheit erreichen?

Sei  $A \in \mathbb{R}^{n \times n}$  beliebig. Dann ist  $B := A^T A$  positiv semidefinit, da gilt

$$\mathbf{x}^T B \mathbf{x} = \mathbf{x}^T A^T A \mathbf{x} = (A \mathbf{x})^T (A \mathbf{x}) \geq 0$$

Folglich sind alle Eigenwerte von  $B$  aus  $\mathbb{R}$  und positiv (oder Null) und wir können definieren

### Definition 3.1.13

Sei  $A \in \mathbb{R}^{n \times n}$  und  $B := A^T A$ .

$$(3.10) \quad \|A\|_{\text{Spek}} = \sqrt{\rho(B)} = \max_{\lambda \text{ EW von } B} \sqrt{\lambda}$$

$\|A\|_{\text{Spek}}$  heißt die **Spektralnorm** von  $A$ .

### Bemerkung 3.1.14

- (1) Zum Nachweis, daß die Spektralnorm eine Norm ist, wird der Rayleigh-Quotient verwendet.
- (2) Ist  $A = A^T \in \mathbb{R}^{n \times n}$ , dann folgt  $\rho(A) = \|A\|_{\text{Spek}}$ . Somit hat  $\|A\|_{\text{Spek}}$  für symmetrische Matrizen Minimaleigenschaft, d.h.  $\|A\|_{\text{Spek}} \leq \|A\|$  für alle Matrix-Normen  $\|\cdot\|$ .
- (3)  $\|\cdot\|_{\text{Spek}}$  ist im allgemeinen aufwendig zu bestimmen, hat mehr Bedeutung bei theoretischen Überlegungen zur Konvergenz.

### 3.1.4 Beziehung zwischen Vektor- und Matrix-Normen

Sei das LGS  $A \mathbf{x} = \mathbf{r}$  gegeben.  $A^{-1}$  existiere und  $\tilde{\mathbf{x}}$  sei die Näherung für  $\mathbf{x}$ . Dann ist  $\mathbf{d} = A \tilde{\mathbf{x}} - \mathbf{r}$  der **Defekt** der Näherung  $\tilde{\mathbf{x}}$ . Im allgemeinen ist der Defekt von Null verschieden. Es ist

$$\begin{aligned} (A \tilde{\mathbf{x}}) - (A \mathbf{x}) &= (\mathbf{r} + \mathbf{d}) - (\mathbf{r}) \\ \Rightarrow A(\tilde{\mathbf{x}} - \mathbf{x}) &= \mathbf{d} \\ \Rightarrow \tilde{\mathbf{x}} - \mathbf{x} &= A^{-1} \mathbf{d} \\ \Rightarrow \|\tilde{\mathbf{x}} - \mathbf{x}\| &= \|A^{-1} \mathbf{d}\| \end{aligned}$$

Es soll gelten

$$\|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|A^{-1}\|_M \cdot \|\mathbf{d}\|$$

Dabei ist  $\|\mathbf{d}\|$  bekannt und  $\|A^{-1}\|_M$  läßt sich eventuell abschätzen. Es wird natürlich eine möglichst gute Abschätzung bevorzugt.

### Definition 3.1.15

Sei  $\|\cdot\|$  eine Vektor-Norm auf  $\mathbb{R}^n$  und  $\|\cdot\|_M$  eine Matrix-Norm auf  $\mathbb{R}^{n \times n}$ .  $\|\cdot\|$  und  $\|\cdot\|_M$  heißen miteinander **verträglich (passend)**, wenn gilt

$$\|A \mathbf{x}\| \leq \|A\|_M \cdot \|\mathbf{x}\| \quad \forall \mathbf{x} \in \mathbb{R}^n \quad \forall A \in \mathbb{R}^{n \times n}$$

Darüber hinaus heißt  $\|\cdot\|_M$  der Vektor-Norm  $\|\cdot\|$  **zugeordnet**, wenn es ein  $\mathbf{x} \neq 0$  mit  $\|A \mathbf{x}\| = \|A\|_M \cdot \|\mathbf{x}\|$  gibt ( $\|\cdot\|_M$  ist dann eindeutig bestimmt).

**Beispiel 3.1.16**

Die Frobenius-Norm ist für  $n > 1$  keiner Vektor-Norm zugeordnet. Dies sieht man wie folgt. Sei  $E \in \mathbb{R}^{n \times n}$  die Einheitsmatrix und  $\|\cdot\|_M$  eine Matrix-Norm, die der Vektor-Norm  $\|\cdot\|$  zugeordnet ist. Dann existiert nach Definition 3.1.15 ein  $0 \neq \mathbf{x} \in \mathbb{R}^n$ , so daß gilt

$$\|E\mathbf{x}\| = \|E\|_M \|\mathbf{x}\|$$

Also folgt  $\|E\|_M = 1$ . Da  $\|E\|_{Frob} = \sqrt{\sum_{i,k=1}^n a_{ik}^2} = \sqrt{n} \neq 1$  falls  $n > 1$ , ist für  $n > 1$  der Frobenius-Norm keine Vektor-Norm zugeordnet.

Einige wichtige Paare von zugeordneten Vektor- und Matrix-Normen:

	V-Norm		M-Norm
(3.11) <b>Maximumsnorm</b>	$\max_{i=1, \dots, n}  x_i  =: \ \mathbf{x}\ _\infty$		$\ \cdot\ _Z$ <b>Zeilensummennorm</b>
	$\sum_{i=1}^n  x_i  =: \ \mathbf{x}\ _1$		$\ \cdot\ _S$ <b>Spaltensummennorm</b>
<b>Euklidische Norm</b>	$\ \cdot\ _2$		$\ \cdot\ _{Spek}$ <b>Spektralnrm</b>

Die Beweise sind für  $\|\cdot\|_\infty$  und  $\|\cdot\|_1$  relativ elementar. Für  $\|\cdot\|_2$  wird der Beweis über den Rayleigh-Quotienten geführt.

**Satz 3.1.17** (Matrixpotenzen)

Sei  $A \in \mathbb{R}^{n \times n}$  und  $m \in \mathbb{N}$ . Es gilt für  $m \rightarrow \infty$   $A^m \rightarrow 0$  komponentenweise genau dann, wenn  $\rho(A) < 1$  ist. In diesem Fall ist  $(E - A)$  regulär. Ist für eine Matrix-Norm  $\|\cdot\|_M$  mit  $\|E\|_M = 1$  auch  $\|A\| < 1$  (hinreichend wegen (3.9)), so gilt ferner

$$(3.12) \quad \|(E - A)^{-1}\|_M \leq \frac{1}{1 - \|A\|_M}$$

**Beweis:**

Der allgemeine Beweis für die erste Aussage ist sehr aufwendig (Jordansche Normalform). Daher wird er hier nur für diagonalähnliche Matrizen geführt.

Sei  $A := T^{-1}DT$  mit

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

Dann gilt

$$A^m = (T^{-1}DT)(T^{-1}DT)(T^{-1}DT) \dots = T^{-1}D^mT$$

mit  $D^m = (\lambda_i^m)$ , wobei  $T^{-1}$  und  $T$  unabhängig von  $m$  sind. Also folgt  $A^m$  geht für  $m \rightarrow \infty$  gegen Null genau dann, wenn  $D^m$  gegen Null geht. Dies ist genau dann der Fall, wenn für alle  $i$  gilt:  $|\lambda_i| < 1$ , was äquivalent ist zu  $\rho(A) < 1$ . q.e.d. Zu der Regularität von  $E - A$  läßt sich sagen, daß  $A\mathbf{x} = \mathbf{x}$  nur für  $\mathbf{x} = 0$  lösbar ist, da  $\lambda = 1$  kein EW ist ( $\rho(A) = \max_{i=1}^n |\lambda_i| < 1$  ( $\lambda_i$  EW)). Also

$$(E - A)\mathbf{x} = 0 \iff \mathbf{x} = 0$$

((3.12) läßt sich über Abschätzungen mit der geometrischen Reihe beweisen.) ■

### 3.1.5 Die Kondition

Sei  $A\mathbf{x} = \mathbf{r}$  ein LGS,  $A \in \mathbb{R}^{n \times n}$ ,  $\det A \neq 0$  und  $\mathbf{r} \neq \mathbf{0}$  (folglich  $\mathbf{x} \neq \mathbf{0}$ ). Seien  $\|\cdot\|$  und  $\|\cdot\|_M$  einander zugeordnete Vektor- und Matrix-Normen.  $A$  und  $\mathbf{r}$  seien mit „Eingangsfehlern“  $\delta A \in \mathbb{R}^{n \times n}$  und  $\delta \mathbf{r} \in \mathbb{R}^n$  behaftet (folglich  $\mathbf{x} \rightarrow \mathbf{x} + \delta \mathbf{x}$ ).

Somit ist

$$(A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{r} + \delta \mathbf{r}$$

Der Fehler  $\delta A$  sei so klein, daß gilt

$$\|A^{-1}\|_M \|\delta A\|_M \stackrel{!}{=} 1$$

wobei  $\|A^{-1}\|_M$  im allgemeinen nur abschätzbar ist.

#### Definition 3.1.18

Die Größe

$$(3.13) \quad \kappa(A) = \text{cond}(A) = \|A\|_M \|A^{-1}\|_M$$

heißt die **Kondition** von  $A$ .

Es gilt (durch geeignete Abschätzungen)

$$(3.14) \quad \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|_M}{\|A\|_M}} \left( \frac{\|\delta A\|_M}{\|A\|_M} + \frac{\|\delta \mathbf{r}\|}{\|\mathbf{r}\|} \right)$$

wobei  $\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|}$  der relative Fehler von  $\mathbf{x}$  ist,  $\frac{\|\delta A\|_M}{\|A\|_M}$  der relative Fehler von  $A$  und entsprechend  $\frac{\|\delta \mathbf{r}\|}{\|\mathbf{r}\|}$  der relative Fehler von  $\mathbf{r}$  ist. Außerdem gilt

$$\kappa(A) \frac{\|\delta A\|_M}{\|A\|_M} = \|A^{-1}\|_M \|\delta A\|_M < 1.$$

$\kappa(A)$  kann im allgemeinen nur abgeschätzt werden.

$$1 = \|E\|_M = \|AA^{-1}\|_M \stackrel{(3.5)}{\leq} \|A\|_M \|A^{-1}\|_M = \kappa(A)$$

Für  $\kappa(A) \gg 1$  ist die Fehlerfortpflanzung schlecht.  $A$  heißt dann **schlecht konditioniert**. Weitere Abschätzungen erhält man in der Literatur über den Satz von Prager und Oettli.

Im folgenden bezeichnen  $|A|$  nicht die Determinante von  $A$ , sondern  $|A| := (|a_{ik}|)_{i,k}$ . Entsprechend soll gelten:  $|\mathbf{x}| = (|x_1|, \dots, |x_n|)^T$ . Gegeben sei das LGS  $A\mathbf{x} = \mathbf{r}$  mit  $A \in \mathbb{R}^{n \times n}$  und  $A^{-1}$  existiere. Die Näherungslösung  $\tilde{\mathbf{x}}$  heißt **akzeptable Lösung** des LGS, wenn sie exakte Lösung eine „benachbarten“ (gestörten) LGS  $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{r}}$  ist, für dessen „Störungen“ komponentenweise gilt:

$$|A - \tilde{A}| \leq \Delta A \quad , \quad |\mathbf{r} - \tilde{\mathbf{r}}| \leq \Delta \mathbf{r}$$

wobei  $\Delta A$  und  $\Delta \mathbf{r}$  Matrix bzw. Vektor mit nichtnegativen Komponenten sind, d.h.  $\Delta A = |\Delta A|$  und  $\Delta \mathbf{r} = |\Delta \mathbf{r}|$ .

#### Satz 3.1.19 (Prager und Oettli (1964))

Eine Näherungslösung  $\tilde{\mathbf{x}}$  ist genau dann akzeptable Lösung von  $A\mathbf{x} = \mathbf{r}$ , wenn der Defektvektor  $\mathbf{d} = A\tilde{\mathbf{x}} - \mathbf{r}$  komponentenweise in der Form

$$|\mathbf{d}| \leq \Delta A |\tilde{\mathbf{x}}| + \Delta \mathbf{r}$$

abgeschätzt werden kann.

## 3.2 Iterative Verfahren zur Lösung von LGS

### 3.2.1 Die Verfahren

Der Vorteil von iterativen gegenüber direkten Verfahren ist die Selbstkorrektur von Rundungsfehlern. Der Nachteil liegt darin, daß sich die iterativen Verfahren nur für Spezialfälle, nämlich schwach besetzte Matrizen, eignen. Diese kommen in den Anwendungen jedoch häufiger vor.

Sei  $A\mathbf{x} = \mathbf{r}$  ein LGS mit  $A = (a_{ik})_{i,k} \in \mathbb{R}^{n \times n}$  und  $\mathbf{x}, \mathbf{r} \in \mathbb{R}^n$ . Die Matrix  $A$  wird in drei Matrizen zerlegt  $A = A_L + D + A_R$  mit

$$A_L := \begin{pmatrix} 0 & & & \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{pmatrix} \quad D := \begin{pmatrix} d_1 & & & \\ & \ddots & & \\ & & d_n & \end{pmatrix} \quad A_R := \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{pmatrix}$$

#### Gesamtschrittverfahren (Jacobi / GSV)

Dann folgt aus

$$D\mathbf{x} = \mathbf{r} - A_L\mathbf{x} - A_R\mathbf{x}$$

die **Iteration in Gesamtschritten (Jacobi / GSV)**:

$$(3.15) \quad D\mathbf{x}^{(k+1)} = \mathbf{r} - A_L\mathbf{x}^{(k)} - A_R\mathbf{x}^{(k)} \quad k = 0, 1, \dots$$

wobei  $\mathbf{x}^{(0)}$  vorgegeben wird, z.B.  $\mathbf{x}^{(0)} = \mathbf{0}$  oder eine Näherung. Die Mindestvoraussetzung für dieses Verfahren ist  $a_{ii} \neq 0$  für alle  $i$ , damit  $D^{-1}$  existiert.

$$(3.16) \quad \boxed{x_i^{(k+1)} = \frac{1}{a_{ii}} \left( r_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^{(k)} \right)}$$

für  $i = 1, \dots, n$  und  $k = 0, 1, \dots$ .

#### Einzelschrittverfahren (Gauß-Seidel / ESV)

Desweiteren existiert die **Iteration in Einzelschritten (Gauß-Seidel / ESV)**. Bei der Berechnung von  $x_i^{(k+1)}$  werden dabei schon die vorherigen Werte  $x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}$  eingesetzt. Es folgt aus

$$(A_L + D)\mathbf{x} = \mathbf{r} - A_R\mathbf{x}$$

die Iteration

$$(3.17) \quad (A_L + D)\mathbf{x}^{(k+1)} = \mathbf{r} - A_R\mathbf{x}^{(k)} \quad k = 0, 1, \dots$$

zu vorgegebenem  $\mathbf{x}^{(0)}$ .

$$(3.18) \quad \boxed{x_i^{(k+1)} = \frac{1}{a_{ii}} \left( r_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right)}$$

für  $i = 1, \dots, n$  und  $k = 0, 1, \dots$

### SOR-Verfahren (Young / Relaxationsverfahren)

Die Verallgemeinerung des ESV führt auf das **SOR-Verfahren (Young)**. SOR steht dabei für „successive-overrelaxation“. Ausgehend von

$$A_L \mathbf{x} = -D\mathbf{x} - A_R \mathbf{x} + \mathbf{r}$$

erhält man durch Multiplikation von  $\omega \in \mathbb{R}$  und Addition von  $D\mathbf{x}$

$$(D + \omega A_L) \mathbf{x} = [(1 - \omega)D - \omega A_R] \mathbf{x} + \omega \mathbf{r}$$

und somit die Iteration

$$(3.19) \quad (D + \omega A_L) \mathbf{x}^{(k+1)} = [(1 - \omega)D - \omega A_R] \mathbf{x}^{(k)} + \omega \mathbf{r} \quad k = 0, 1, \dots$$

zu vorgegebenem  $\mathbf{x}^{(0)}$ .

$$(3.20) \quad x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( r_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right)$$

für  $i = 1, \dots, n$  und  $k = 0, 1, \dots$ . Für  $\omega = 1$  ergibt sich das ESV.

### 3.2.2 Konvergenzuntersuchungen

Es werden im folgenden Verfahren der Art

$$(3.21) \quad \mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{s}$$

mit vorgegebenem  $\mathbf{x}^{(0)}$  untersucht. Dabei sei das  $M$  bei den einzelnen Verfahren folgendermaßen gesetzt:

$$\begin{aligned} \text{GSV : } M &= -D^{-1}(A_L + A_R) \\ \text{SOR : } M &= (D + \omega A_L)^{-1}[(1 - \omega)D - \omega A_R] \end{aligned}$$

#### Satz 3.2.1

Das Iterationsverfahren (3.21) konvergiert genau dann für beliebige  $\mathbf{x}^{(0)}$  gegen die Lösung des LGS  $(E - M)\mathbf{x} = \mathbf{s}$  (äquivalent zu  $A\mathbf{x} = \mathbf{r}$ ), wenn  $\rho(M) < 1$  ist.

#### Beweis:

Der Beweis wird mit Satz 3.1.17 geführt.

Ist  $E - M$  regulär, so läßt sich  $(E - M)\mathbf{x} = \mathbf{s}$  eindeutig auflösen.

$$\begin{aligned} \mathbf{x}^{(k+1)} - \mathbf{x} &= (M\mathbf{x}^{(k)} + \mathbf{s}) - (M\mathbf{x} + \mathbf{s}) \\ \Leftrightarrow \mathbf{x}^{(k+1)} - \mathbf{x} &= M(\mathbf{x}^{(k)} - \mathbf{x}) = M^2(\mathbf{x}^{(k-1)} - \mathbf{x}) = \dots = M^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}) \end{aligned}$$

Dabei geht  $M^{k+1}$  gegen Null und mit Satz 3.1.17 folgt dann die Behauptung. ■

Nun folgen einige weitere Aussagen über Satz 3.2.1 ohne Beweis. Dabei sei stets  $a_{ii} \neq 0$  für alle  $i$  angenommen.

- (1) Das SOR-Verfahren kann nur dann konvergieren für beliebige  $\mathbf{x}^{(0)}$ , wenn  $0 < \omega < 2$  gilt. Für Anwendungen wird im allgemeinen  $\omega \in [1, 2)$  gewählt.
- (2) Ist  $A = A^T$  und  $A$  positiv definit, so konvergiert das SOR-Verfahren für alle  $\omega \in (0, 2)$ .

### Beispiel 3.2.2

Die Hilbertmatrix  $H_2$  ist positiv definit.

$$H_2 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

Das LGS  $H_2 \mathbf{x} = (1/2, 1/3)^T$  hat die Lösung  $\mathbf{x} = (0, 1)^T$ . Mit dem ESV und  $\mathbf{x}^{(0)} = 0$  ergibt sich

$$x_1^{(30)} = 0.000028 \quad , \quad x_2^{(30)} = 0.000058$$

Offensichtlich sind diese Ergebnisse schlecht. Gute Werte sind nur bei Diagonaldominanz zu erwarten!

- (3) Es sei  $\omega \in (0, 2)$  und  $A = (a_{ik})_{i,k} \in \mathbb{R}^{n \times n}$ .

$$\begin{aligned} \beta_1(\omega) &= |1 - \omega| + \frac{\omega}{|a_{11}|} \sum_{j=2}^n |a_{1j}| \\ \beta_2(\omega) &= |1 - \omega| + \frac{\omega}{|a_{22}|} \left( |a_{21}| \beta_1(\omega) + \sum_{j=3}^n |a_{2j}| \right) \end{aligned}$$

Die allgemeine Formel für  $i = 2, \dots, n$  lautet

$$\beta_i(\omega) := |1 - \omega| + \frac{\omega}{|a_{ii}|} \left( \sum_{j=1}^{i-1} |a_{ij}| \beta_j(\omega) + \sum_{j=i+1}^n |a_{ij}| \right)$$

Als **Sassenfeldzahl** (zu  $\omega$ ) bezeichnet man

$$\beta(\omega) := \max_{i=1}^n |\beta_i(\omega)|$$

Ist  $\beta(\omega) < 1$ , so gilt

$$\|M\|_Z = \left\| (D + \omega A_L)^{-1} [(1 - \omega)D - \omega A_R] \right\|_Z \leq \beta(\omega) < 1$$

d.h. das SOR-Verfahren ist konvergent. (Diese Aussage ist auch als Sassenfeld-Kriterium bekannt.)

Die Sassenfeldzahl liefert auch eine Fehlerabschätzung. Der Beweis dazu verläuft ähnlich wie der Beweis zu Satz 3.2.1. Dazu folgende Betrachtungen:

Es ist

$$\begin{aligned} \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} &= (M \mathbf{x}^{(k)} + \mathbf{s}) - (M \mathbf{x}^{(k-1)} + \mathbf{s}) \\ (3.22) \quad &= M (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = M^2 (\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}) = \dots \end{aligned}$$

Betrachtet man nun

$$\begin{aligned} \mathbf{x} - \mathbf{x}^{(k)} &= \mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \\ &= \mathbf{x} - \mathbf{x}^{(k+m)} + \mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)} + \dots + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \end{aligned}$$

und somit

$$\|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \leq \|\mathbf{x} - \mathbf{x}^{(k+m)}\|_\infty + \dots + \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty,$$

so ergibt sich mit der vorangegangenen Überlegung für  $m \rightarrow \infty$

$$(3.23) \quad \|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \leq \sum_{j=0}^{\infty} \|\mathbf{x}^{(k+j+1)} - \mathbf{x}^{(k+j)}\|_\infty = \sum_{j=0}^{\infty} \|M^{j+1} (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\|_\infty$$

Da außerdem

$$\|M^{j+1} (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\|_\infty \leq \|M^{j+1}\|_Z \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$$

und durch rekursives Anwenden von (3.5)

$$\|M^{j+1}\|_Z \leq \|M\|_Z^{j+1} \leq \beta^{j+1}(\omega)$$

folgt aus (3.23)

$$\|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \leq [\beta(\omega) + \beta^2(\omega) + \dots] \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$$

und somit die **aposteriori Fehlerabschätzung** (Prinzip des Banachschen Fixpunktsatzes):

$$(3.24) \quad \|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \leq \frac{\beta(\omega)}{1 - \beta(\omega)} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$$

Baut man die Exponenten entsprechend (3.23) weiter ab, so erhält man die **apriori Fehlerabschätzung**:

$$(3.25) \quad \|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \leq \frac{\beta^k(\omega)}{1 - \beta(\omega)} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty$$

- (4) In Anwendungsfällen gilt oft für  $A$  das **starke Zeilensummenkriterium** (Diagonaldominanz):

$$(3.26) \quad \forall i = 1, \dots, n \quad \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}| < |a_{ii}|$$

bzw. das **schwache Zeilensummenkriterium**

$$(3.27) \quad \begin{aligned} \forall i = 1, \dots, n \quad \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}| &\leq |a_{ii}| \\ \exists i = 1, \dots, n \quad \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}| &< |a_{ii}| \end{aligned}$$

Diese Situation ist häufig bei Randwertaufgaben und partiellen Differentialgleichungen gegeben.  $A \in \mathbb{R}^{n \times n}$  heißt **zerfallend** (zerlegbar, reduzibel), wenn  $A$  durch Vertauschen von Zeilen und gleichnummerierten Spalten auf die Form

$$\tilde{A} = \left( \begin{array}{c|c} A_1 & A_2 \\ \hline 0 & A_3 \end{array} \right)$$

wobei  $A_1$  und  $A_2$  quadratische Matrizen sind, gebracht werden kann. Mit anderen Worten, es gibt eine Permutationsmatrix  $P$  mit  $\tilde{A} = PAP^T$ .

**Bemerkung 3.2.3**

Zerfällt das LGS  $A\mathbf{x} = \mathbf{r}$ , so ist  $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{r}}$  nach Umbenennung der Variablen mit  $\tilde{\mathbf{x}} = (\mathbf{z}, \mathbf{y})^T$  und  $\tilde{\mathbf{r}} = (\mathbf{t}, \mathbf{s})^T$  zu lösen. D.h. man löst erst  $A_3\mathbf{y} = \mathbf{s}$  und dann  $A_1\mathbf{z} = \mathbf{t} - A_2\mathbf{y}$ .

**Lemma 3.2.4**

Genügt  $A$  dem starken Zeilensummenkriterium oder ist  $A$  nicht zerfallen und genügt dem schwachen Zeilensummenkriterium, dann konvergiert das ESV und für jedes  $\omega \in (0, 1)$  das SOR-Verfahren.

**Lemma 3.2.5**

Ist  $\|\cdot\|_M$  eine monotone Matrix-Norm und gilt

$$\left\| D^{-1}(A_L + A_R) \right\|_M < 1$$

so konvergiert für  $A\mathbf{x} = \mathbf{r}$  für beliebige  $\mathbf{x}^{(0)}$  sowohl das ESV als auch das GSV.

(5) Welches ist der optimale  $\omega$ -Wert ?

Für spezielle Matrizen gibt es Aussagen mit Hilfe der EW der Iterationsmatrix des GSV. Diese sind aber im allgemeinen sehr aufwendig. Wegen (3.22) folgt im Regelfall

$$\frac{(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})_i}{(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})_i} \longrightarrow \rho(M_\omega) \quad (i = 1, \dots, n)$$

Das kann in manchen Anwendungsfällen zur Bestimmung von  $\omega_{opt}$  verwendet werden.

**3.2.3 Bemerkungen zu Iterationsverfahren für LGS****Iteration mit Elimination**

Sei  $A\mathbf{x} = \mathbf{r}$  mit der Zerlegung  $A = B + C$  gegeben. Das LGS  $B\mathbf{y} = \mathbf{s}$  sei für alle  $\mathbf{s}$  leicht auflösbar. (z.B.  $B^{-1}$  bekannt und  $A$  ist durch eine Störung  $C$  aus  $B$  hervorgegangen.) Mit dem Iterations-Verfahren

$$(3.28) \quad B\mathbf{x}^{(k+1)} = -C\mathbf{x}^{(k)} + \mathbf{r}$$

und gegebenem  $\mathbf{x}^{(0)}$  ist nach Satz 3.2.1 die Konvergenz für  $\rho(B^{-1}C) < 1$  gegeben.

**Nachiteration**

Es sei das LGS  $A\mathbf{x} = \mathbf{r}$  gegeben. Mit z.B. Gauß sei die Näherung  $\mathbf{x}^{(0)}$  für die Lösung bestimmt worden. Der Defekt (Residuum)  $\mathbf{d} = A\mathbf{x}^{(0)} - \mathbf{r}$  ist im allgemeinen ungleich Null. Es wird der Ansatz  $\mathbf{x} = \mathbf{x}^{(0)} + \mathbf{z}$  gemacht. Somit ist

$$0 = A\mathbf{x} - \mathbf{r} = A\mathbf{x}^{(0)} + A\mathbf{z} - \mathbf{r} = A\mathbf{z} + \mathbf{d}$$

Mit demselben (Rundungsfehler-behafteten) Eliminationsverfahren wird dann  $A\mathbf{z} = -\mathbf{d}$  gelöst, d.h. durch Hinzufügen einer neuen rechten Seite. Man erhält  $\mathbf{z}^{(0)}$  und definiert  $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{z}^{(0)}$ . Dies setzt man unter Umständen fort.

Eine Rundungsfehler-behaftete LR-Zerlegung ergibt im Prinzip eine Näherung  $B$  für  $A^{-1}$ . Das Verfahren konvergiert, wenn  $B \approx A^{-1}$  ist, d.h. in diesem Fall  $\rho(E - BA) < 1$ .

## Block-Iteration

Bei Anwendungen ist die Matrix des LGS  $A\mathbf{x} = \mathbf{r}$  häufig in Blöcke aufgeteilt. z.B.  $A \in \mathbb{R}^{m \cdot n \times m \cdot n}$ :

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix} \quad \text{mit } A_{ij} \in \mathbb{R}^{m \times m}$$

und entsprechend

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_n \end{pmatrix} \quad \text{mit } \mathbf{x}_i, \mathbf{r}_i \in \mathbb{R}^m$$

$A$  wird zerlegt in  $A = \hat{D} + \hat{A}_L + \hat{A}_R$  mit

$$\hat{D} = \begin{pmatrix} A_{11} & & \\ & \ddots & \\ & & A_{nn} \end{pmatrix} \quad \text{Blockdiagonalmatrix}$$

und  $\hat{A}_L$  und  $\hat{A}_R$  entsprechend. Sind alle  $A_{ii}$  regulär, dann läßt sich z.B. „Block ESV“ durchführen (vgl. (3.18)):

$$A_{ii}\mathbf{x}_i^{(k+1)} = \mathbf{r}_i - \sum_{j=1}^{i-1} A_{ij}\mathbf{x}_j^{(k+1)} - \sum_{j=i+1}^n A_{ij}\mathbf{x}_j^{(k)}$$

(jeweils LGS lösen, z.B. mit Gauß).

## 3.3 Nichtlineare Gleichungssysteme

### 3.3.1 Das Banach'sche Verfahren

In dem Abschnitt 3.2.2 wurden folgende, zu dem LGS  $A\mathbf{x} = \mathbf{r}$  äquivalente LGS betrachtet:

$$\mathbf{x} = M\mathbf{x} + \mathbf{s}$$

(vgl. (3.21)). Das zugehörige Iterationsverfahren  $\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{s}$  konvergiert, wenn  $\rho(M) < 1$  ist. Entsprechend wird bei einem NLGS  $F(\mathbf{x}) = 0$  mit  $F : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$  verfahren. Dabei ist  $F(\mathbf{x}) = 0$  äquivalent in  $\mathbf{x} = \varphi(\mathbf{x})$  umzuformen, so daß der Einfluß des nichtlinearen Anteils  $\varphi$  möglichst klein wird. Wir definieren das Iterationsverfahren

$$(3.29) \quad \mathbf{x}^{(k+1)} = \varphi(\mathbf{x}^{(k)})$$

zu gegebenem Startwert  $\mathbf{x}^{(0)}$ .

Das Verfahren konvergiert, wenn es einen Bereich  $B \subset \mathbb{R}^n$  ( $B$  abgeschlossen) gibt mit  $\varphi : B \rightarrow B$ , so daß  $\varphi$  kontrahierend ist.

#### Definition 3.3.1

$\varphi : B \rightarrow B$  ist kontrahierend, wenn ein  $\alpha < 1$  existiert, so daß für alle  $\mathbf{x}, \mathbf{y} \in B$  gilt

$$\|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| \leq \alpha \|\mathbf{x} - \mathbf{y}\|.$$

wobei  $\|\cdot\|$  eine beliebige Norm auf  $\mathbb{R}^n$  ist.

Fehlerabschätzungen lassen sich dann entsprechend wie (3.24) und (3.25) herleiten, z.B.

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\alpha}{1 - \alpha} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \frac{\alpha^k}{1 - \alpha} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

Dies läßt sich anwenden, wenn  $\alpha \ll 1$  gilt.

### Beispiel 3.3.2

*Gesucht ist die Lösung des NLGS*

$$x = \exp -x =: \varphi x \quad x \in \mathbb{R}$$

Setze  $B = [0.5, 0.69]$ . Man prüfe leicht nach, daß  $\varphi : B \rightarrow B$  (z.B. weil  $\varphi$  monoton und  $\varphi(0.69) \in B$ ).

$$\alpha := \max_{x \in B} \|\varphi'(x)\| = \max_{x \in B} \|-\exp -x\| = \exp -0.5 = 0.606531 < 1$$

$\varphi$  ist kontrahierend. Somit folgt mit dem Banachschen Fixpunktsatz:  $\varphi$  besitzt in  $B$  einen Fixpunkt. Durch Iteration erhält man

$k$	$\mathbf{x}^{(k)}$
0	0.55
1	0.5764981
2	0.56160877
3	0.57029086
4	0.56536097

### 3.3.2 Das Newton-Verfahren

#### Newton-Verfahren für Funktionen mit einer Veränderlichen

Das Newton Verfahren zur Lösung einer nichtlinearen Gleichung  $f(x) = 0$  beruht im eindimensionalen auf der Idee, die Funktion  $f$  durch ihre lineare Tangente in einem Startpunkt  $x^{(0)}$  zu approximieren und die Nullstelle der Tangente als Näherung für die Nullstelle von  $f$  zu betrachten. Als Tangente im Punkt  $x^{(0)}$  an  $f$  ergibt sich

$$p(x) := f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$$

Die Nullstelle von  $p$  läßt sich für den Fall, daß  $f'(x^{(0)}) \neq 0$  ist durch

$$x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}$$

berechnen. Es ist klar, daß die wiederholte Anwendung der Approximation und Nullstellenbestimmung der Tangente den Fehler minimiert. Diese wiederholte Anwendung wird als **Newton-Verfahren** (im eindimensionalen) bezeichnet, d.h.

(3.30)

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

wobei  $x^{(0)}$  wie üblich als Startvektor bezeichnet wird.

Eine zweite Möglichkeit das Newton-Verfahren herzuleiten ergibt sich aus dem **Taylor'schen Satz**:

$$0 = f(\xi) = f(x^{(0)}) + \frac{f'(x^{(0)})}{1!}(\xi - x^{(0)}) + \frac{f''(\zeta)}{2!}(\xi - x^{(0)})^2$$

wobei  $\frac{f''(\zeta)}{2!}(\xi - x^{(0)})^2$  das Restglied für ein  $\zeta \in [\xi, x^{(0)}]$  ist. Für  $|\xi - x^{(0)}| \ll 1$  d.h.  $(\xi - x^{(0)})^2 \ll |\xi - x^{(0)}|$  wird der letzte Term vernachlässigt, also

$$f(x^{(0)}) + \frac{f'(x^{(0)})}{1!}(\xi - x^{(0)}) \approx 0$$

Die Ersetzung von „ $\approx$ “ durch „ $=$ “ und  $\xi$  durch  $x^{(1)}$  liefert nach der Auflösung nach  $x^{(1)}$  das Newton-Verfahren (3.30).

Für die Konvergenz des Verfahrens gibt es viele Bedingungen. Deshalb ist das „drauf-losrechnen“ sinnvoller. Wenn, dann liegt quadratische Konvergenz vor.

### Newton-Verfahren für Funktionen mit mehreren Veränderlichen

Mittels des Satzes von Taylor ist es möglich, das Newton-Verfahren entsprechend für Funktionen von mehreren Veränderlichen herzuleiten. Aus Übersichtsgründen beschränken wir uns hier auf zwei Variablen.

Gesucht sei die Lösung eines nichtlinearen Gleichungssystems:

$$0 = F(\mathbf{x}) := \begin{pmatrix} f(\xi, \eta) \\ g(\xi, \eta) \end{pmatrix} \quad \text{mit} \quad \mathbf{x} = \begin{pmatrix} \xi \\ \nu \end{pmatrix} \quad \text{und} \quad f, g : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$$

Sei  $\mathbf{x}^{(0)} = (\xi_0, \eta_0)^T$  eine Näherung. Dann folgt mit Taylor

$$0 = F(\mathbf{x}) = F(\mathbf{x}^{(0)}) + F'(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) + R = F(\mathbf{x}^{(0)}) + \begin{pmatrix} (\xi - \xi_0)f_\xi(\xi_0, \eta_0) + (\eta - \eta_0)f_\eta(\xi_0, \eta_0) \\ (\xi - \xi_0)g_\xi(\xi_0, \eta_0) + (\eta - \eta_0)g_\eta(\xi_0, \eta_0) \end{pmatrix} + R$$

Der Rest  $R$  enthält partielle Ableitungen 2.Ordnung an einer Zwischenstelle. Parallel zum Eindimensionalen ist  $\mathbf{x}$  durch  $\mathbf{x}^{(1)}$  zu ersetzen und das Restglied zu vernachlässigen. Wir erhalten

$$(3.31) \quad \left( \begin{matrix} f_\xi & f_\eta \\ g_\xi & g_\eta \end{matrix} \right) \Big|_{\mathbf{x}^{(0)} = \begin{pmatrix} \xi_0 \\ \eta_0 \end{pmatrix}} (\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) =: F(\mathbf{x}^{(0)}) + \Phi(\xi_0, \eta_0)(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = 0$$

wobei  $\Phi$  die Funktional- bzw. Jacobi-Matrix ist. Ist  $\Phi$  regulär, so läßt sich (3.31) nach  $\mathbf{x}^{(1)}$  auflösen. Weiterführen dieses Verfahrens und auflösen nach  $\mathbf{x}^{(k+1)}$  liefert schließlich

$$(3.32) \quad \boxed{\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \Phi(\mathbf{x}^{(k)})^{-1} F(\mathbf{x}^{(k)})}$$

Entsprechend erhält man die Formeln für  $(n \times n)$ -nichtlineare Gleichungssysteme.

### Bemerkung 3.3.3

(1) Das Newton-Verfahren (3.31) konvergiert wesentlich schneller als Banach, wenn man bereits in der Nähe der Nullstelle ist. Oft muß man allerdings „hin- und herpendeln“, wenn z.B. mehrere Nullstellen vorliegen und  $\mathbf{x}^{(0)}$  nicht sehr nah an einer der Nullstellen liegt.

(2) **Vereinfachtes Newton-Verfahren**

Da  $\Phi^{-1}(\mathbf{x}^{(k)})$  oft mühsam zu berechnen ist, kann man in manchen Fällen stets  $\Phi^{-1}(\mathbf{x}^{(0)})$  verwenden, also

$$(3.33) \quad F(\mathbf{x}^{(k)}) + \Phi(\mathbf{x}^{(0)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = 0$$

Dies konvergiert natürlich langsamer.

### 3.3.3 Nichtlineare GSV, ESV und SOR-Verfahren

Es sei das NLGS

$$\begin{aligned} f_1(\xi_1, \dots, \xi_n) &= 0 \\ \dots & \\ f_n(\xi_1, \dots, \xi_n) &= 0 \end{aligned} \quad \text{mit } f_i : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$$

und  $\mathbf{x}^{(0)} = (\xi_1^{(0)}, \dots, \xi_n^{(0)})^T \in D$  gegeben.

Für  $i = 1, \dots, n$

$$f_i(\xi_1^{(0)}, \dots, \xi_{i-1}^{(0)}, \xi_i, \xi_{i+1}^{(0)}, \dots, \xi_1^{(n)}) = 0$$

nach  $\xi_i$  auflösen z.B. mit dem vereinfachten Newton-Verfahren oder ähnlichem. Dann wird  $\xi_i^{(1)} = \xi_i$  gesetzt und  $\mathbf{x}^{(1)} = (\xi_1, \dots, \xi_n)^T$  und weiter iteriert.

Das nichtlineare GSV hat den Vorteil, jeweils nur eine Gleichung lösen zu müssen. Das nichtlineare ESV erhält man entsprechend mit  $\xi_i$  aus

$$(3.34) \quad f_i(\xi_1^{(2)}, \dots, \xi_{i-1}^{(1)}, \xi_i, \xi_{i+1}^{(0)}, \dots, \xi_1^{(n)}) = 0$$

Das nichtlineare SOR-Verfahren erhält man, indem man (3.34) mit

$$\xi_i^{(1)} = (1 - \omega)\xi_1^{(0)} + \omega\xi_i$$

löst (entsprechend weiter iterieren).

## Chapter 4

# Konjugierte Gradientenmethoden (CG)

Diese Methode ist auch als **konjugierte Richtungsmethode** bekannt. Sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit,  $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$  und  $c \in \mathbb{R}$ . Sei weiter  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  definiert durch

$$(4.1) \quad f(\mathbf{x}) = \frac{1}{2} \langle A\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c$$

wobei

$$\langle \mathbf{b}, \mathbf{x} \rangle = \mathbf{b}^T \cdot \mathbf{x} = b_1x_1 + b_2x_2 + \dots + b_nx_n$$

das **euklidische Skalarprodukt** ist. Weiter ist

$$\nabla f = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$$

der **Gradient** von  $f$ .

Ziel der CG-Methode ist es, das Gleichungssystem  $A\mathbf{x} = \mathbf{b}$  zu lösen. Setzt man in (4.1)  $\mathbf{x} := A^{-1}\mathbf{b}$  ein, so ergibt sich

$$f(A^{-1}\mathbf{b}) = \frac{1}{2} \langle \mathbf{b}, A^{-1}\mathbf{b} \rangle - \langle \mathbf{b}, A^{-1}\mathbf{b} \rangle + c = -\frac{1}{2} \langle \mathbf{b}, A^{-1}\mathbf{b} \rangle + c$$

welches das Minimum von  $f$  mit  $c = 0$  ist. Die Aufgabe  $f$  zu minimieren ist also äquivalent dazu  $A\mathbf{x} - \mathbf{b} = 0$  zu lösen. Das Minimum von  $f$  ist charakterisiert durch  $\nabla f(\mathbf{x}) = 0$ .

Die Behauptung, daß  $\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$  gilt, zeigen wir hier exemplarisch für  $n = 2$ . Es ergibt sich  $\nabla \langle \mathbf{b}, \mathbf{x} \rangle = (b_1, b_2)^T$  und

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x} \rangle &= (x_1, x_2) \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_1x_2 + a_{22}x_2^2 =: g(\mathbf{x}) \end{aligned}$$

Aufgrund der Symmetrie von  $A$  folgt

$$\frac{1}{2} \nabla \langle \mathbf{x}, A\mathbf{x} \rangle = \frac{1}{2} \begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2a_{11}x_1 + (a_{12} + a_{21})x_2 \\ (a_{12} + a_{21})x_1 + 2a_{22}x_2 \end{pmatrix} = A\mathbf{x}$$

Und somit

$$\nabla f(\mathbf{x}) = \nabla \left[ \frac{1}{2} \langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c \right] = A\mathbf{x} - \mathbf{b}$$

**Definition 4.0.4**

Seien  $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$  mit  $\mathbf{p} \neq 0 \neq \mathbf{q}$ . Dann heißen  $\mathbf{p}$  und  $\mathbf{q}$  **konjugiert** bezüglich  $A$  genau dann, wenn  $\langle A\mathbf{p}, \mathbf{q} \rangle = 0$  für  $\mathbf{p} \neq \mathbf{q}$  ist.

**Bemerkung 4.0.5**

Sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit. Dann hat  $A$  genau  $n$  orthogonale Eigenvektoren  $\{\mathbf{x}_j\}$  mit zugehörigen Eigenwerten  $\lambda_j$ . Sei  $X := (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$  und  $D \in \mathbb{R}^{n \times n}$  die Diagonalmatrix, die auf der Diagonalen die Eigenwerte  $\lambda_j$ ,  $j = 1, \dots, n$  enthält und sonst Nullen. Dann läßt sich schreiben

$$AX = DX \quad \Rightarrow \quad X^T AX = D \quad \Rightarrow \quad A = XDX^T$$

und somit

$$\langle A\mathbf{u}, \mathbf{u} \rangle = \langle XDX^T \mathbf{u}, \mathbf{u} \rangle = \langle DX^T \mathbf{u}, X^T \mathbf{u} \rangle = \langle \sqrt{D}X^T \mathbf{u}, \sqrt{D}X^T \mathbf{u} \rangle.$$

**Methode des steilsten Abstiegs:**

Sei  $\mathbf{p}^0, \mathbf{p}^1, \dots, \mathbf{p}^{n-1}$  eine konjugierte Basis in  $\mathbb{R}^n$  bezüglich  $A$ . Dann sei

$$(4.2) \quad \mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \mathbf{p}^k$$

mit  $\alpha^k \in \mathbb{R}$  so gewählt, daß

$$f(\mathbf{x}^k - \alpha^k \mathbf{p}^k) \leq f(\mathbf{x}^k - \alpha \mathbf{p}^k) \quad \forall \alpha \in \mathbb{R}$$

Das führt auf

$$(4.3) \quad \alpha^k = \frac{\langle A\mathbf{x}^k - \mathbf{b}, \mathbf{p}^k \rangle}{\langle A\mathbf{p}^k, \mathbf{p}^k \rangle}$$

für  $k = 0, \dots, n-1$ , da

$$\begin{aligned} f(\mathbf{x}^k - \alpha \mathbf{p}^k) &= \frac{1}{2} \langle A(\mathbf{x}^k - \alpha \mathbf{p}^k), \mathbf{x}^k - \alpha \mathbf{p}^k \rangle - \langle \mathbf{b}, \mathbf{x}^k - \alpha \mathbf{p}^k \rangle + c \\ \Rightarrow \frac{df}{d\alpha} &= \frac{1}{2} \langle A(-\mathbf{p}^k), \mathbf{x}^k - \alpha \mathbf{p}^k \rangle + \frac{1}{2} \langle A(\mathbf{x}^k - \alpha \mathbf{p}^k), -\mathbf{p}^k \rangle - \langle \mathbf{b}, -\mathbf{p}^k \rangle = 0 \\ \Rightarrow &-\langle A\mathbf{p}^k, \mathbf{x}^k \rangle + \alpha \langle A\mathbf{p}^k, \mathbf{p}^k \rangle + \langle \mathbf{b}, \mathbf{p}^k \rangle = 0 \\ \Rightarrow &(4.3) \end{aligned}$$

**Satz 4.0.6** (Konvergenz)

Sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit und  $\{\mathbf{p}^k\}_{k=0}^{n-1}$  eine konjugierte Basis bezüglich  $A$ . Mit (4.2) und (4.3) folgt dann, daß ein  $m \in \mathbb{N}$  existiert mit  $m \leq n$ , so daß gilt:

$$(4.4) \quad \mathbf{x}^m = A^{-1} \mathbf{b}$$

**Bemerkung 4.0.7**

$\mathbf{x}^m = \mathbf{x}$  ist die exakte Lösung von  $A\mathbf{x} = \mathbf{b}$  d.h. das Verfahren endet nach endlich vielen Schritten, falls exakte Arithmetik verwendet wird.

**Beweis:**

$$\begin{aligned} \langle A\mathbf{x}^{k+1} - \mathbf{b}, \mathbf{p}^j \rangle &= \langle A\mathbf{x}^k - \mathbf{b} - \alpha^k A\mathbf{p}^k, \mathbf{p}^j \rangle \\ &= \langle A\mathbf{x}^k - \mathbf{b}, \mathbf{p}^j \rangle - \frac{\langle A\mathbf{x}^k - \mathbf{b}, \mathbf{p}^k \rangle}{\langle A\mathbf{p}^k, \mathbf{p}^k \rangle} \langle A\mathbf{p}^k, \mathbf{p}^j \rangle \\ &= \begin{cases} \langle A\mathbf{x}^k - \mathbf{b}, \mathbf{p}^j \rangle & , \quad j \neq k \\ 0 & , \quad j = k \end{cases} \end{aligned}$$

Somit

$$\langle A\mathbf{x}^n - \mathbf{b}, \mathbf{p}^j \rangle = \langle A\mathbf{x}^{n-1} - \mathbf{b}, \mathbf{p}^j \rangle = \dots = \langle A\mathbf{x}^{j+1} - \mathbf{b}, \mathbf{p}^j \rangle$$

für  $j = 0, \dots, n-1$ . Da die  $\mathbf{p}^j$  linear unabhängig sind folgt  $A\mathbf{x}^n = \mathbf{b}$  oder  $\mathbf{x}^n = A^{-1}\mathbf{b}$ . Ist  $A\mathbf{x}^m = \mathbf{b}$  für ein  $m < n$ , wähle  $\alpha^m = 0$ . Damit folgt  $\mathbf{x}^m = \mathbf{x}^{m+1} = \dots = \mathbf{x}^n$ . ■

### Bemerkung 4.0.8

Wesentlich für den CG-Algorithmus ist das Finden von konjugierten Richtungen  $\mathbf{p}^0, \dots, \mathbf{p}^{n-1}$ . Eine direkte Methode dafür ist die Gram-Schmidt-Orthogonalisierung. In jedem Fall ist dies eine sehr ineffiziente Prozedur.

Falls keine konjugierte Basis des  $\mathbb{R}^n$  bekannt ist, so muß auch  $\mathbf{p}^k$  iterativ bestimmt werden. Seien also nur die Startwerte  $\mathbf{x}^0, \mathbf{p}^0$  gegeben. Die weiteren Werte  $\mathbf{x}^{k+1}$  werden mit (4.2) berechnet. Für die  $\mathbf{p}^{k+1}$  gilt

$$\mathbf{p}^{k+1} = (A\mathbf{x}^{k+1} - \mathbf{b}) - \beta^k \mathbf{p}^k =: \mathbf{r}^{k+1} - \beta^k \mathbf{p}^k$$

wobei  $\beta^k$  so gewählt ist, daß

$$\langle A\mathbf{p}^{k+1}, \mathbf{p}^k \rangle = 0.$$

Das führt auf

$$(4.5) \quad \beta^k = \frac{\langle A\mathbf{x}^{k+1} - \mathbf{b}, A\mathbf{p}^k \rangle}{\langle A\mathbf{p}^k, \mathbf{p}^k \rangle}$$

da

$$0 = \langle A\mathbf{p}^{k+1}, \mathbf{p}^k \rangle = \langle \mathbf{p}^{k+1}, A\mathbf{p}^k \rangle = \langle A\mathbf{x}^{k+1} - \mathbf{b} - \beta^k \mathbf{p}^k, A\mathbf{p}^k \rangle$$

### Satz 4.0.9

Seien die Voraussetzungen wie in Satz 4.0.6 gegeben. Dann gilt

$$\langle A\mathbf{p}^j, \mathbf{p}^j \rangle = 0 \quad \text{für } i \neq j$$

**Beweis:**

Die Werte  $\alpha^k, \beta^k$  sind wohldefiniert, da  $\mathbf{p}^k \neq 0$ . Für den Fall  $\mathbf{p}^0 = 0$  ist die Aussage trivial. Sei daher  $\mathbf{p}^0 \neq 0$ . Wir nehmen an, daß  $\mathbf{p}^k \neq 0$  für  $k \leq m-1$  gilt und setzen  $\mathbf{r}^k := A\mathbf{x}^k - \mathbf{b}$  für  $k \leq m-1$ .

Man bemerke, daß  $\mathbf{r}^k \neq 0$  ist, da sonst  $\beta^{k-1} = 0$  und somit  $\mathbf{p}^k = \mathbf{r}^k - \beta^{k-1}\mathbf{p}^{k-1}$  wäre, was zu einem Widerspruch führt.

Wir schreiben nun  $\mathbf{x}^{j+1} = \mathbf{x}^j - \alpha^j \mathbf{p}^j$  als

$$\mathbf{r}^{j+1} = \mathbf{r}^j - \alpha^j A\mathbf{p}^j$$

Dann gilt

$$\mathbf{r}^{j+1} = -\mathbf{b} + A\mathbf{x}^{j+1} = A\mathbf{x}^j - \alpha_j A\mathbf{p}^j - \mathbf{b} = \mathbf{r}^j + \mathbf{b} - \alpha_j A\mathbf{p}^j - \mathbf{b}$$

und nach Definition der  $\alpha^j$ :

$$\langle \mathbf{r}^{j+1}, \mathbf{p}^j \rangle = 0 \quad \left( = \langle \mathbf{r}^j, \mathbf{p}^j \rangle - \underbrace{\frac{\langle \mathbf{r}^j, \mathbf{p}^j \rangle}{\langle A\mathbf{p}^j, \mathbf{p}^j \rangle}}_{=\alpha_j} \langle A\mathbf{p}^j, \mathbf{p}^j \rangle \right), \quad j = 0, \dots, m-1$$

Weiterhin gilt nach Definition der  $\beta^j$ :

$$\langle A\mathbf{p}^{j+1}, \mathbf{p}^j \rangle = 0.$$

Nehmen wir nun an, daß gilt

$$\begin{aligned} \langle A\mathbf{p}^k, \mathbf{p}^j \rangle &= 0, \quad j = 0, \dots, k-1 \\ \langle \mathbf{r}^k, \mathbf{r}^j \rangle &= 0, \quad j = 0, \dots, k-1 \end{aligned}$$

Der Beweis hierfür wird durch Induktion geführt. Für  $k = 1$  ist die Behauptung klar. Für die Induktionsvoraussetzung gilt

$$\langle \mathbf{r}^{k+1}, \mathbf{r}^j \rangle = \langle \mathbf{r}^k - \alpha^k A\mathbf{p}^k, \mathbf{r}^j \rangle = \langle \mathbf{r}^k, \mathbf{r}^j \rangle - \alpha^k \langle A\mathbf{p}^k, \mathbf{r}^j \rangle + \beta^{j-1} \langle \mathbf{p}^{j-1}, \mathbf{r}^j \rangle = 0$$

und

$$\langle \mathbf{r}^{k+1}, \mathbf{r}^k \rangle = \langle \mathbf{r}^{k+1}, \mathbf{p}^k + \beta^{k-1} \mathbf{p}^{k-1} \rangle = 0$$

Somit folgt  $\langle \mathbf{r}^{k+1}, \mathbf{r}^j \rangle = 0$ . Ebenso gilt

$$\begin{aligned} \langle A\mathbf{p}^{k+1}, \mathbf{p}^j \rangle &= \langle \mathbf{p}^{k+1}, A\mathbf{p}^j \rangle = \langle \mathbf{r}^{k+1} - \beta^k \mathbf{p}^k, A\mathbf{p}^j \rangle \\ &= \langle \mathbf{r}^{k+1}, A\mathbf{p}^j \rangle = \langle \mathbf{r}^{k+1}, \frac{\mathbf{r}^j - \mathbf{r}^{j+1}}{\alpha^j} \rangle = 0 \end{aligned}$$

und  $\langle A\mathbf{p}^{k+1}, \mathbf{p}^k \rangle = 0$  (nach Konstruktion)

Nehmen wir nun  $\mathbf{p}^m = 0$  für  $m < n$  an. Dann folgt

$$0 = \langle \mathbf{p}^m, \mathbf{p}^m \rangle = \langle \mathbf{r}^m - \beta^{m-1} \mathbf{r}^{m-1}, \mathbf{r}^m - \beta^{m-1} \mathbf{r}^{m-1} \rangle \geq \langle \mathbf{r}^m, \mathbf{r}^m \rangle$$

woraus dann  $\mathbf{r}^m = 0$  und schließlich  $\mathbf{x}^m = A^{-1}\mathbf{b}$  folgt. ■

#### Satz 4.0.10

Unter den Voraussetzungen wie in Satz 4.0.6 gilt die folgende Fehlerabschätzung

$$\|\mathbf{x} - \mathbf{x}^k\| \leq 2 \left( \frac{\sqrt{\chi} - 1}{\sqrt{\chi} + 1} \right)^k \|\mathbf{x} - \mathbf{x}^0\|$$

## Chapter 5

# Approximation

Die Interpolation lieferte zu einer gegebenen, hinreichend glatten Funktion  $f : I \rightarrow \mathbb{R}$  ein Polynom  $p_n \in P_n$ , welches  $f$  zu gegebenen Knoten  $x_1, \dots, x_n$  interpolierte. Dabei galt  $g(x_k) = f_k$  für  $k = 1, \dots, n$  und  $\|f - p_n\|_{I, \infty}$  sollte möglichst klein sein. Nun ist ein  $p_n \in P_n$  unabhängig von der Interpolation gesucht, welches die Forderung erfüllt, daß  $\|f - p_n\|_{I, \infty}$  möglichst klein ist. Ein solches  $p$  heißt **Approximation** von  $f$ . Existiert eine Approximation  $p$  von  $f$  mit minimalem  $\|f - p\|_{I, \infty}$ , so spricht man von einer **besten Approximation**. Wird als Norm die Norm  $\|\cdot\|_{I, \infty}$  verwendet, so heißt das entsprechende Polynom **Tschebyscheff-Approximation**. Es stellt sich die Frage nach dem Fehler in anderen Normen. Hier soll die  $L_2$ -Norm betrachtet werden

$$(5.1) \quad \|p - u\|_{L^2(0,1)}^2 := \int_0^1 (p(x) - u(x))^2 dx$$

### Definition 5.0.11

Ein Vektorraum  $U$  heißt ein **reeller unitärer Raum**, falls ein **Skalarprodukt**  $\langle \cdot, \cdot \rangle : U \times U \rightarrow \mathbb{R}$  existiert, so daß folgende Eigenschaften erfüllt sind.

$$(5.2) \quad \begin{aligned} \forall f, g, h \in U ; \lambda, \mu \in \mathbb{R} : \quad & 0 \neq f \Rightarrow \langle f, f \rangle > 0 \\ & \langle f, g \rangle = \langle g, f \rangle \\ & \langle f, \lambda g + \mu h \rangle = \lambda \langle f, g \rangle + \mu \langle f, h \rangle \end{aligned}$$

In einem unitären Raum können wir eine **Norm** definieren:

$$(5.3) \quad \forall f \in U \quad \|f\| := \langle f, f \rangle^{\frac{1}{2}}$$

Es gilt die *Cauchy-Schwarz'sche Ungleichung*

$$(5.4) \quad \forall f, g \in U \quad |\langle f, g \rangle| \leq \|f\| \cdot \|g\|$$

### Beispiel 5.0.12

Beide Räume sind sogar vollständig, d.h. *Hilbert-Räume*

$$(a) \quad \mathbb{R}^n : \quad \langle x, y \rangle := x^T y \quad \text{für } x, y \in \mathbb{R}^n$$

(b)  $w \in C(I)$  mit  $w > 0 \Big|_I$

$$L_2(I, w) := \left\{ f : I \rightarrow \mathbb{R} \mid \int_I f^2 w \, dx < \infty \right\}$$

$$\langle f, g \rangle := \int_I f(x)g(x)w \, dx$$

## 5.1 Existenz und Eindeutigkeit der besten Approximation

### Definition 5.1.1

Sei  $V \subset U$  ein linearer Unterraum von  $U$ . Zu  $u \in U$  heißt  $v \in V$  **beste Approximation** genau dann, wenn gilt

$$\forall w \in V \quad \|u - v\| \leq \|u - w\|$$

**Satz 5.1.2** (Existenz und Eindeutigkeit der besten Approximation (konstruktiv))

Seien  $v_1, \dots, v_n \in U$  linear unabhängig und  $V := \text{span}\{v_1, \dots, v_n\}$ . Dann gelten folgende Aussagen:

- Für alle  $u \in U$  existiert genau eine beste Approximation  $v \in V$ .
- Für alle  $w \in V$  gilt  $\langle u - v, w \rangle = 0$ , d.h.  $u - v \perp V$ . ( $v$  ist beste Approximation von  $u$ )
- Sei  $A \in \mathbb{R}^{n \times n}$  mit  $A = (\langle v_i, v_j \rangle)_{i,j}$  und  $\mathbf{b} \in \mathbb{R}^n$  mit  $b_j := \langle u, v_j \rangle$ ,  $j = 1, \dots, n$ . Dann ist  $\det A \neq 0$ .

Sei  $\mathbf{x} \in \mathbb{R}^n$  die Lösung der sogenannten **Normalengleichung**

$$(5.5) \quad Ax = \mathbf{b}$$

Dann ist die beste Approximation  $v \in V$  von  $u \in U$  eindeutig bestimmt durch

$$(5.6) \quad v = \sum_{j=1}^n x_j v_j$$

$A$  heißt **Gram'sche Matrix** und  $\det A$  die **Gram-Determinante**.

### Beweis:

Zuerst zeigen wir  $\det A \neq 0$ .

Angenommen, die Spaltenvektoren von  $A$  sind linear abhängig. Dann existiert ein  $\mathbf{a} \in \mathbb{R}^n$  mit  $a_j \neq 0$  ( $1 \leq j \leq n$ ), so daß gilt

$$\sum_{j=1}^n a_j \langle v_i, v_j \rangle = 0 \quad \Rightarrow \quad \langle v_i, \sum_{j=1}^n a_j v_j \rangle = 0 \quad \Rightarrow \quad \langle \sum_{j=1}^n a_j v_i, \sum_{j=1}^n a_j v_j \rangle = 0.$$

Aufgrund von (5.2) folgt dann  $\sum_{j=1}^n a_j v_j = 0$ , was ein Widerspruch zu der linearen Unabhängigkeit der  $v_j$  ist. Somit gilt  $\det A \neq 0$ . Folglich ist  $\mathbf{x} := A^{-1}\mathbf{b}$  wohldefinierte Lösung von (5.5). Sei  $y := \sum_{j=1}^n x_j v_j \in V$ . Dann gilt

$$(5.7) \quad \langle u - y, v_i \rangle = \langle u, v_i \rangle - \langle y, v_i \rangle = b_i - \sum_j x_j \langle v_j, v_i \rangle = b_i (Ax)_i = 0$$

d.h.  $u - y \perp V$ . Sei  $w \in V$ ,  $y \in V$  und  $u \in U$ . Dann gilt

$$\begin{aligned} \|u - w\|^2 &= \langle u - w, u - w \rangle \\ &= \langle u - y - (w - y), u - y - (w - y) \rangle \\ &\stackrel{(5.2)}{=} \langle u - y, u - y \rangle - 2\langle u - y, w - y \rangle + \langle w - y, w - y \rangle \\ &= \|u - y\|^2 + \|w - y\|^2 > \|u - y\|^2 \quad \text{für } w \neq y. \end{aligned}$$

wobei  $\langle u - y, w - y \rangle$  wegen (5.7) verschwindet. Insgesamt folgt, daß  $\|u - w\|$  minimal ist für genau ein  $w \in V$ . Daraus folgt a), wo  $w = y$  ist und somit (5.5) mit  $y = \sum x_j v_j = v$ , also c). Da weiter (5.7) gilt, ist schließlich auch b) bewiesen. ■

### Beispiel 5.1.3

Approximation von

$$u(x) := \begin{cases} 1 & , \quad 0 \leq x \leq \frac{1}{2} \\ 0 & , \quad \text{sonst} \end{cases}$$

in  $L_2(I)$ ,  $I = [0, 1]$  durch Polynome  $p(x) = \sum_{k=0}^{n-1} a_k x^k$ . D.h. es soll gelten

$$\int_0^1 \left( \sum_{k=0}^{n-1} a_k x^k - u(x) \right)^2 dx \stackrel{!}{=} \min$$

Die partielle Ableitung von (5.1) liefert als notwendige Bedingung für die beste Approximation ( $0 \leq \nu \leq n - 1$ )

$$\frac{\partial}{\partial a_\nu} \left( \sum_{k=0}^{n-1} a_k x^k - u(x) \right)^2 dx = \int_0^1 2 \left( \sum_{k=0}^{n-1} a_k x^k - u(x) \right) x^\nu dx \stackrel{!}{=} 0$$

was äquivalent ist zu

$$(5.8) \quad \sum_{k=0}^{n-1} \int_0^1 a_k x^k x^\nu dx = \int_0^1 u(x) x^\nu dx.$$

Sei  $V := \text{span}\{v_1, v_2, v_3, \dots, v_n\} = \text{span}\{1, x, x^2, \dots, x^{n-1}\} =: P_{n-1}$  und  $v_j := x^{j-1}$ . Sei  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$  definiert als

$$\langle f, g \rangle := \int_0^1 f \cdot g dx.$$

Dann ist  $\langle \cdot, \cdot \rangle$  ein Skalarprodukt. Mit dieser Notation ist (5.8) äquivalent zu folgendem linearen Gleichungssystem:

$$(5.9) \quad \begin{pmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots & \langle v_1, v_n \rangle \\ \vdots & \vdots & & \vdots \\ \langle v_n, v_1 \rangle & \langle v_n, v_2 \rangle & \cdots & \langle v_n, v_n \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} \langle u, v_1 \rangle \\ \vdots \\ \langle u, v_n \rangle \end{pmatrix}$$

mit

$$\begin{aligned} \langle v_i, v_j \rangle &= \int_0^1 x^{i-1} x^{j-1} dx = \frac{1}{i + j - 1} \\ \langle u, v_j \rangle &= \int_0^{1/2} x^{j-1} dx = \frac{1}{j} 2^{-j} \end{aligned}$$

Damit ergibt sich dann allgemein

$$H_n x := \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & & & \frac{1}{n+1} \\ \frac{1}{3} & & \ddots & & \\ \vdots & & & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & & \cdots & \frac{1}{2n-1} \end{pmatrix} x = \begin{pmatrix} 2^{-1} \\ \frac{1}{2} 2^{-2} \\ \frac{1}{3} 2^{-3} \\ \vdots \\ \frac{1}{n} 2^{-n} \end{pmatrix}$$

Im speziellen

$$\begin{aligned} n=2 & : & x^T & = \left( \frac{5}{4}, -\frac{3}{2} \right) \\ n=3 & : & x^T & = \left( \frac{5}{4}, -\frac{3}{2}, 0 \right) \\ n=4 & : & x^T & = (0.8125, 3.75, -13.125, 8.75) \end{aligned}$$

Folglich erhalten wir als Approximationsfunktionen zu  $u(x)$

$$\begin{aligned} v^{(2)} & = & \frac{5}{4} - \frac{3}{2}x & \in P_1 \\ v^{(3)} & = & v^{(2)} & \in P_2 \\ v^{(4)} & = & x_0 + x_1x + x_2x^2 + x_3x^3 & \in P_3 \end{aligned}$$

Der Vorteil dieser Approximationsmethode ist die Möglichkeit viele (nämlich alle  $L^2$ -) Funktionen approximieren zu können. Der Nachteil ist, daß für nicht-triviale  $n$  die Koeffizientenmatrix  $H_n$ , die sogenannte **Hilbert-Matrix**, sehr schlecht konditioniert ist.

## 5.2 Approximation bei gegebener Orthonormalbasis

Mit einem ONS  $\{e_1, \dots, e_n\}$  als Basis von  $V$  wird die Approximationsaufgabe sehr einfach, denn die Normalengleichung (5.5) lautet dann  $Ix = b$  d.h.  $x = b$ . Somit ergibt sich für die beste Approximation bei gegebener ONB:

$$v = \sum_{j=1}^n \langle u, e_j \rangle e_j$$

Der Approximationsfehler läßt sich wie folgt bestimmen. Sei  $u \in U$  und  $v \in V$  die beste Approximation von  $u$ . Dann folgt mit Pythagoras und der Bessel'schen Ungleichung

$$\|u - v\|^2 = \langle u, u - v \rangle - \langle v, u - v \rangle = \|u\|^2 - \|v\|^2 = \|u\|^2 - \sum_{j=1}^n \langle u, e_j \rangle^2 \geq 0$$

wobei  $\langle v, u - v \rangle$  wegen (5.1.2) verschwindet. Außerdem ist

$$\begin{aligned} \|v\|^2 &= \langle v, v \rangle = \left\langle \sum_{k=1}^n \langle u, e_k \rangle e_k, \sum_{j=1}^n \langle u, e_j \rangle e_j \right\rangle \\ &= \sum_{k=1}^n \langle u, e_k \rangle \left\langle e_k, \sum_{j=1}^n \langle u, e_j \rangle e_j \right\rangle = \sum_{k=1}^n \langle u, e_k \rangle \langle u, e_k \rangle \end{aligned}$$

**Bemerkung 5.2.1**

In  $L_2[-1, 1]$  erhalten wir durch Orthonormalisierung von  $1, x, x^2, \dots$  die **Legendre Polynome**:

$$(5.10) \quad l_1(x) = \frac{1}{\sqrt{2}} \quad ; \quad l_2(x) = \sqrt{\frac{3}{2}} x \quad ; \quad l_3(x) = \frac{1}{2} \sqrt{\frac{5}{2}} (3x^2 - 1) \quad ; \quad \dots$$

wobei  $u \in L_2[-1, 1]$  genau dann der Fall ist, wenn  $\int_{-1}^1 u^2(x) dx < \infty$  ist. Für die Legendre-Polynome folgt also

$$\int_{-1}^1 l_j(x) l_k(x) dx = \begin{cases} 0 & , j \neq k \\ 1 & , j = k \end{cases}$$

### 5.3 Gram-Schmidt-Orthonormalisierung

**Satz 5.3.1 (Gram-Schmidt-Orthonormalisierung)**

Seien  $v_1, \dots, v_n \in U$  linear unabhängig. Definiere rekursiv

$$e_1 = \frac{v_1}{\|v_1\|}$$

Angenommen, es sind bereits  $k$  orthonormale Vektoren  $e_1, \dots, e_k, k < n$  bestimmt.

$$(5.11) \quad \begin{cases} d_{k+1} := v_{k+1} - \sum_{\nu=1}^k \langle v_{k+1}, e_\nu \rangle e_\nu \\ e_{k+1} := \frac{d_{k+1}}{\|d_{k+1}\|} \end{cases}$$

Diese Rekursion kann immer durchgeführt werden und liefert  $n$  orthonormale Vektoren  $e_\nu, \nu = 1, \dots, n$ .

**Beweis:**

(Induktion über  $m$ )

Der Fall  $m = 1$  ist trivial.

Angenommen die Behauptung gilt für  $m < n$ . Dann sind  $e_1, \dots, e_m$  orthonormal. Aus (5.11) folgt

$$e_\nu = \sum_{\mu=1}^m a_{\nu\mu} v_\mu$$

Angenommen es ist  $d_{m+1} = 0$ . Dann sind  $v_1, \dots, v_{m+1}$  linear abhängig ( $v_{m+1}$  hat den Koeffizienten 1) was einen Widerspruch liefert. Somit ist  $d_{m+1} \neq 0$  d.h.  $e_{m+1}$  ist wohldefiniert.

Sei  $\nu \leq m$ :

$$\langle d_{m+1}, e_\nu \rangle = \langle v_{m+1}, e_\nu \rangle - \left\langle \sum_{\mu=1}^m \langle v_{m+1}, e_\mu \rangle e_\mu, e_\nu \right\rangle = 0$$

Folglich ist  $\langle e_{m+1}, e_\nu \rangle = 0$  für alle  $\nu \leq m$  und somit  $e_1, \dots, e_{m+1}$  orthonormal. ■

**Beispiel 5.3.2**

(a) Normiere  $v_1 = 1$ ,  $v_2 = x$ ,  $v_3 = x^2$  bezüglich  $U := L^2[0, 1]$ , d.h.

$$\langle u, v \rangle := \int_0^1 u(x)v(x) dx$$

Es ist

$$\begin{aligned} e_1(x) &= \frac{1}{\left(\int_0^1 1 dx\right)^{\frac{1}{2}}} = 1 \\ d_2(x) &= x - \langle x, 1 \rangle \cdot 1 = x - \frac{1}{2} \\ \|d_2(x)\|^2 &= \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12} \\ e_2(x) &= \frac{d_2(x)}{\|d_2(x)\|} = 2\sqrt{3}\left(x - \frac{1}{2}\right) \end{aligned}$$

und so weiter.

(b) Normiere  $\{1, x, x^2, x^3, \dots\}$  bezüglich  $L^2(-1, 1)$  (vergleiche (5.10)).

$$\begin{aligned} l_1(x) &= \frac{x^0}{\|x^0\|_{L^2[-1,1]}} = \frac{1}{\sqrt{2}} \\ d_2(x) &= x - \left\langle x, \frac{1}{\sqrt{2}} \right\rangle \frac{1}{\sqrt{2}} = x - \frac{1}{2} \int_{-1}^1 x dx = x - \frac{1}{2} \frac{x^2}{2} \Big|_{-1}^1 = x \\ \|d_2(x)\|^2 &= \int_{-1}^1 x^2 dx = \frac{2}{3} \end{aligned}$$

Ebenso

$$\begin{aligned} l_2(x) &= \frac{d_2(x)}{\|d_2(x)\|} = \sqrt{\frac{3}{2}} x \\ d_3(x) &= x^2 - \langle x^2, l_1(x) \rangle l_1(x) - \langle x^2, l_2(x) \rangle l_2(x) \\ &= x^2 - \left( \int_{-1}^1 \frac{x^2}{\sqrt{2}} dx \right) \frac{1}{\sqrt{2}} - \left( \int_{-1}^1 x^2 \sqrt{\frac{3}{2}} x dx \right) \sqrt{\frac{3}{2}} x \\ &= x^2 - \left( \frac{1}{\sqrt{2}} \frac{2}{3} \right) \frac{1}{\sqrt{2}} - 0 \sqrt{\frac{3}{2}} x = x^2 - \frac{1}{3} \\ \|d_3(x)\|^2 &= \int_{-1}^1 \left(x^2 - \frac{1}{3}\right)^2 dx = \int_{-1}^1 \left(x^4 - \frac{2x^2}{3} + \frac{1}{9}\right) dx = \frac{x^5}{5} - \frac{2}{3} \frac{x^3}{3} + \frac{x}{9} \Big|_{-1}^1 = \frac{8}{45} \end{aligned}$$

Weiter Schritte

$$l_3(x) = \frac{1}{2} \sqrt{\frac{5}{2}} (3x^2 - 1)$$

## 5.4 Diskrete Approximation im quadratischem Mittel

Sei  $I \subset \mathbb{R}$ . Gegeben seien die Knoten  $x_1, x_2, \dots, x_m \in I$  mit  $m \geq n$  und  $v_1, v_2, \dots, v_n : I \rightarrow \mathbb{R}$ . Von  $u : I \rightarrow \mathbb{R}$  seien nur die Funktionswerte an den Knoten bekannt  $u(x_1), \dots, u(x_m)$ . Gesucht ist  $v : I \rightarrow \mathbb{R}$  mit  $I = \text{span}\{v_1, \dots, v_n\}$ , so daß  $v$  die beste Approximation von  $u$  bezüglich der Fehlerquadrate ist. Zur Überprüfung der Fehlerquadrate wird die Euklidische Norm auf unser Problem übertragen.

$$\|u\|_{Euklid} := \sqrt{\sum_{i=1}^m (u(x_i))^2}$$

$\|\cdot\|_{Euklid}$  ist keine Norm, da aus  $\|u\|_{Euklid} = 0$  nicht  $u(x) = 0$  für alle  $x \in I$  folgt. Insgesamt ist  $v$  die beste Approximation von  $u$  genau dann, wenn gilt

$$\sum_{i=1}^m (u(x_i) - v(x_i))^2 = \|u - v\|_2^2 \leq \|u - w\|_2^2 \quad \forall w \in \text{span}\{v_1, \dots, v_n\}$$

d.h.

$$\sum_{i=1}^m (u(x_i) - v(x_i))^2 \stackrel{!}{=} \min .$$

### Beispiel 5.4.1

Gegeben seien die sieben Punkte  $(x_i, y_i)$  mit  $i = 1, \dots, 7$ . Passe eine Gerade  $Y = k_0 + k_1x$  so an, daß

$$S = \sum_{i=1}^n (Y_i - y_i)^2 = \min$$

$Y_i - y_i$  heißt Residuum ( $Y_i = Y(x_i)$ ).  $S$  wird minimal, wenn

$$\begin{aligned} \frac{\delta S}{\delta k_0} &= \sum_{i=1}^n \frac{\delta}{\delta k_0} (k_0 + k_1 x_i - y_i)^2 = \sum_{i=1}^n 2(k_0 + k_1 x_i - y_i) \cdot 1 = 0 \\ \frac{\delta S}{\delta k_1} &= \sum_{i=1}^n \frac{\delta}{\delta k_1} (k_0 + k_1 x_i - y_i)^2 = \sum_{i=1}^n 2(k_0 + k_1 x_i - y_i) \cdot x_i = 0 \end{aligned}$$

Dies ist äquivalent zu

$$\begin{aligned} \sum_{i=1}^n y_i &= \sum_{i=1}^n k_0 + k_1 x_i = k_0 n + k_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= k_0 \sum_{i=1}^n x_i + k_1 \sum_{i=1}^n x_i^2 \end{aligned}$$

zum Beispiel:

$i$	$x_i$	$y_i$	$x_i y_i$	$(x_i)^2$
1	0	2	0	0
2	1	3	3	1
3	2	5	10	4
4	3	5	15	9
5	4	9	36	16
6	5	8	40	25
7	6	10	60	36
Summe	21	42	64	91

Damit erhalten wir die beiden linearen Gleichungen  $42 = 7k_0 + 21k_1$  und  $64 = 21k_0 + 91k_1$  und weiter  $k_0 = 1.928571$  und  $k_1 = 1.357143$ .

In dem Beispiel wurde eine Funktion  $u$  von der nur sieben Funktionswerte bekannt waren durch  $v$  approximiert, wobei der Grad von  $v$  gleich eins war. Selbstverständlich kann man mit diesem Verfahren ebenso ein  $v$  mit höherem Grad, sagen wir allgemein vom Grad  $m$ , finden. Gegeben seien  $n$  Punkte  $(x_i, y_i)$  mit  $i = 1, \dots, n$ . Dann sei  $S$  äquivalent zum speziellen Fall oben definiert durch:

$$S := \sum_{i=1}^n (Y_i - y_i)^2 = \sum_{i=1}^n (k_0 + k_1 x_i + k_2 x_i^2 + \dots + k_m x_i^m - y_i)^2$$

Wenn  $Y$  beste Approximation ist, folgt, daß  $S$  minimal ist. Also fordern wir

$$\frac{\delta S}{\delta k_j} = 0 \quad \text{für } 0 \leq j \leq m$$

was genau dann gilt, wenn

$$\begin{aligned} \frac{\delta S}{\delta k_0} &= \sum_{i=1}^n 2(k_0 + k_1 x_i + k_2 x_i^2 + \dots + k_m x_i^m - y_i) \cdot 1 = 0 \\ \frac{\delta S}{\delta k_1} &= \sum_{i=1}^n 2(k_0 + k_1 x_i + k_2 x_i^2 + \dots + k_m x_i^m - y_i) \cdot x_i = 0 \\ &\vdots \\ \frac{\delta S}{\delta k_m} &= \sum_{i=1}^n 2(k_0 + k_1 x_i + k_2 x_i^2 + \dots + k_m x_i^m - y_i) \cdot x_m = 0 \end{aligned}$$

Dies ist äquivalent zu

$$\begin{aligned} k_0 n + k_1 \sum_{i=1}^n x_i + k_2 \sum_{i=1}^n x_i^2 + \dots + k_m \sum_{i=1}^n x_i^m &= \sum_{i=1}^n y_i \\ k_0 \sum_{i=1}^n x_i + k_1 \sum_{i=1}^n x_i^2 + k_2 \sum_{i=1}^n x_i^3 + \dots + k_m \sum_{i=1}^n x_i^{m+1} &= \sum_{i=1}^n x_i y_i \\ &\vdots \\ k_0 \sum_{i=1}^n x_i^m + k_1 \sum_{i=1}^n x_i^{m+1} + k_2 \sum_{i=1}^n x_i^{m+2} + \dots + k_m \sum_{i=1}^n x_i^{2m} &= \sum_{i=1}^n x_i^m y_i \end{aligned}$$

Dies sind die sogenannten **Normalgleichungen**. Es sind  $m+1$  Gleichungen für die  $m+1$  Unbekannten  $k_0, \dots, k_m$ , die sich ebenso als lineares Gleichungssystem schreiben lassen:

$$\sum_{i=1}^n \begin{pmatrix} n & x_i & x_i^2 & \dots & x_i^m \\ x_i & x_i^2 & \dots & & x_i^{m+1} \\ x_i^2 & \dots & & & x_i^{m+2} \\ \vdots & & & & \vdots \\ x_i^m & x_i^{m+1} & \dots & & x_i^{2m} \end{pmatrix} \begin{pmatrix} k_0 \\ k_1 \\ k_2 \\ \vdots \\ k_m \end{pmatrix} = \sum_{i=1}^n \begin{pmatrix} y_i \\ x_i y_i \\ x_i^2 y_i \\ \vdots \\ x_i^m y_i \end{pmatrix},$$

Es ist  $A = A^T$ , also symmetrisch. Im allgemeinen ist  $A$  hier schlecht konditioniert.

## Chapter 6

# Numerische Quadratur

### 6.1 Trapez- und Simpson-Regel

Seien  $x_0, x_1$  Knoten und  $f_0, f_1$  die zugehörigen Funktionswerte. Wir definieren

$$l_1 = f_0 + \frac{h}{2}f'_0 \quad , \quad l_2 = f_1 + \frac{h}{2}f'_1$$

Für die Flächen der Trapeze  $T_1$  und  $T_2$  ergeben sich die folgenden Berechnungen.

$$\begin{aligned} \text{area}(T_1) &= \frac{f_0 + l_1}{2} \frac{h}{2} = \frac{h}{4} \left[ f_0 + \left( f_0 + \frac{h}{2}f'_0 \right) \right] = \frac{h}{2}f'_0 + \frac{h^2}{8}f''_0 \\ \text{area}(T_2) &= \frac{f_1 + l_2}{2} \frac{h}{2} = \left( 2f_1 - \frac{h}{2}f'_1 \right) \frac{h}{4} = \frac{h}{2}f'_1 + \frac{h^2}{8}f''_1 \\ \text{area}(T_1) + \text{area}(T_2) &= \frac{h}{2}(f_0 + f_1) + \frac{h^2}{8}(f''_0 - f''_1) \end{aligned}$$

Für den **Diskretisierungsfehler** erhalten wir

$$|E(f)| := \left| \int_{x_0}^{x_1} f(x) dx - \frac{h}{2}(f_1 + f_0) \right| \Rightarrow |E(f)| \leq \frac{h^2}{8} |f''_1 - f''_0|$$

Dies ist ebenfalls in dem Fall  $f''(x) < 0$  für alle  $x \in (x_0, x_1)$  gültig.

Allgemein bezeichnet man als **Quadraturformel** die Formel

$$(6.1) \quad \sum_{i=0}^n \omega_i f_i$$

wobei  $f_i = f(x_i)$  gilt und  $\omega_i$  die **Gewichte** heißen. Der **Quadraturfehler** ergibt sich somit als

$$(6.2) \quad E(f) := \int_a^b f(x) dx - \sum_{i=0}^n \omega_i f_i$$

$\int_a^b f(x) dx$  kann im allgemeinen nur näherungsweise berechnet werden, z.B.

$$f(x) = \exp(-x^2) \quad \text{oder} \quad f(x) = \sin(x^2)$$

Ziel ist es also die Knoten  $x_0, \dots, x_n$  und die Gewichte  $\omega_0, \dots, \omega_n$  so zu bestimmen, daß  $|E(f)|$  möglichst klein wird.

Eine Zweite Möglichkeit, die Quadraturformel herzuleiten, ergibt sich durch die Substitution des Integranden  $f(x)$  durch das **Lagrange-Interpolationspolynom**

$$(6.3) \quad p_n(x) = \sum_{\nu=0}^n f_\nu l_\nu(x)$$

wobei die **Lagrange-Polynome**  $n$ -ten Grades  $l_0, \dots, l_n$  zu  $n+1$  paarweise verschiedenen Zahlen  $x_0, \dots, x_n \in \mathbb{R}$  wie folgt definiert sind. Für  $n=0$  ist

$$l_0(x) = 1 \quad x \in \mathbb{R}$$

und für  $n \geq 1$

$$l_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^n \frac{x - x_i}{x_k - x_i} \quad x \in \mathbb{R} \quad k = 0, \dots, n$$

Aus dieser Definition folgt sofort die charakteristische Eigenschaft der Lagrange-Polynome

$$l_k(x_j) = \delta_{jk} = \begin{cases} 1 & , \quad j = k \\ 0 & , \quad j \neq k \end{cases} \quad j, k = 0, \dots, n$$

Mit der so eingeführten Darstellung lassen sich nun die Trapez-Regel und eine weiter herleiten.

### Trapez-Regel (n=1)

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b \{l_0(x)f_0 + l_1(x)f_1\} dx \\ &= \int_a^b \left\{ \frac{(x-x_1)}{(x_0-x_1)}f_0 + \frac{(x-x_0)}{(x_1-x_0)}f_1 \right\} dx \quad , \quad h := (x_1 - x_0) \\ &= \frac{1}{2h} \left[ -(x-x_1)^2 f_0 + (x-x_0)^2 f_1 \right]_a^b \quad (\text{mit } a := x_0, b := x_1) \\ &= \frac{h}{2}(f_0 + f_1) \end{aligned}$$

### Simpson-Regel (n=2)

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b \{l_0(x)f_0 + l_1(x)f_1 + l_2(x)f_2\} dx \\ &= \int_a^b \left\{ \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f_2 \right\} dx \\ &= \frac{1}{2h^2} \left[ \frac{2h^3}{3}f_0 + \frac{8h^3}{3}f_1 + \frac{2h^3}{3}f_2 \right] \quad (\text{mit } a := x_0, b := x_2) \\ &= \frac{h}{3}(f_0 + 4f_1 + f_2) \quad , \quad h := (x_{i+1} - x_i) \end{aligned}$$

Für den Quadraturfehler benutzen wir den „Interpolationsfehler“-Term (wird später noch weiter erläutert)

$$f(x) - p_n(x) = \frac{\omega(x)}{(n+1)!} f^{(n+1)}(\xi)$$

wobei  $f \in C^{(n-1)}[a, b]$ ,  $\xi \in (a, b) = (x_0, x_n)$  und  $\omega(x) = \prod_{\nu=0}^n (x - x_\nu)$ . Für  $E(f)$  ergibt sich somit

$$(6.4) \quad E(f) = \int_a^b f(x) dx - \int_a^b p_n(x) dx = \int_a^b f^{(n+1)}(\xi) \frac{\omega(x)}{(n+1)!} dx$$

Bevor wir zu dem Quadraturfehler der Trapezregel kommen, vorweg noch ein Satz.

**Satz 6.1.1** (2.Mittelwertsatz der Differentialrechnung)

Seien  $f(x)$  und  $g(x)$  stetig auf  $a \leq x \leq b$  und  $g(x)$  wechsele nicht das Vorzeichen auf  $[a, b]$ . Dann existiert  $\eta \in (a, b)$  so daß gilt

$$\int_a^b f(x)g(x) dx = f(\eta) \int_a^b g(x) dx$$

Für den **Fehler bei der Trapez-Regel (n=1)** ergibt sich nun

$$E^T(f) = \int_{a=x_0}^{b=x_1} f^{(2)}(\xi) \frac{(x-x_0)(x-x_1)}{2!} dx$$

Ist  $x \in (a, b)$  so besitzt  $(x-x_0)(x-x_1)$  auf  $[a, b]$  konstantes Vorzeichen und es läßt sich Satz 6.1.1 anwenden.

$$E(f) = f^{(2)}(\xi) \int_a^b \frac{(x-x_0)(x-x_1)}{2} dx$$

und mit partieller Integration

$$\int_a^b \frac{(x-a)(x-b)}{2} dx = \underbrace{\frac{(x-a)^2(x-b)}{2}}_{=0} \Big|_a^b - \int_a^b \frac{(x-a)^2}{4} dx = -\frac{(x-a)^3}{3 \cdot 4} \Big|_a^b$$

folgt schließlich

$$E(f) = f^{(2)}(\xi) \frac{h^3}{12}, \quad a < \xi < b$$

## 6.2 Newton-Cotes Formeln

Seien  $x_i$  mit  $i = 0, 1, \dots, n$  äquidistante Knoten, d.h.  $x_j = x_0 + jh$ , und die Grenzen  $a = x_0 + ph$  und  $b = x_0 + qh$  mit  $p \geq 0, q \leq n, p, q \in \mathbb{R}$ . Mit der Substitution  $x = x_0 + sh$  ergibt sich als

**Newton-Cotes-Formel**

$$\begin{aligned}
I_{(n+1)}(f) &:= \int_a^b \left[ \sum_{j=0}^n l_j(x) f_j \right] dx \\
&= \sum_{j=0}^n \int_p^q \prod_{\substack{k=0 \\ k \neq j}}^n \frac{(x_0 + sh) - (x_0 + kh)}{(x_0 + jh) - (x_0 + kh)} f_j h ds \\
&= h \sum_{j=0}^n \int_p^q \prod_{\substack{k=0 \\ k \neq j}}^n \frac{(s-k)}{(j-k)} f_j ds \\
&= \sum_{j=0}^n \alpha_j f_j \quad \text{with} \quad \alpha_j = h \int_p^q \prod_{\substack{k=0 \\ k \neq j}}^n \frac{(s-k)}{(j-k)} ds
\end{aligned}$$

Man unterscheidet verschiedene Arten dieser Formeln.

**(a) Geschlossene Newton-Cotes Formeln**

Es gelte  $b - a = nh$ ,  $x_0 = a$ ,  $x_n = b$ ,  $p = 0$  und  $q = n$ . Für  $n = 1$  erhält man die Trapez-Regel und für  $n = 2$  erhält man die Simpson-Regel.

$$\alpha_j = h \int_0^{n-2} \prod_{\substack{k=0 \\ k \neq j}}^2 \frac{(s-k)}{(j-k)} ds$$

$$\begin{aligned}
\alpha_0 &= h \int_0^2 \frac{(s-1)(s-2)}{(-1)(-2)} ds = \frac{h}{2} \int_0^2 (s^2 - 3s + 2) ds = \frac{h}{2} \left( \frac{s^3}{3} - \frac{3s^2}{2} + 2s \right) \Big|_0^2 = \frac{h}{3} \\
\alpha_1 &= h \int_0^2 \frac{s(s-2)}{1(1-2)} ds = h \int_0^2 (-s^2 + 2s) ds = h \left( -\frac{s^3}{3} + s^2 \right) \Big|_0^2 = \frac{4}{3} h \\
\alpha_2 &= h \int_0^2 \frac{s(s-1)}{2 \cdot 1} ds = \frac{h}{2} \left( \frac{s^3}{3} - \frac{s^2}{2} \right) \Big|_0^2 = \frac{h}{2} \left( \frac{8}{3} - 2 \right) = \frac{h}{3}
\end{aligned}$$

**(b) Offene Newton-Cotes Formeln**

Sei wieder  $b - a = nh$ ,  $x_0 = a$ ,  $x_n = b$ ,  $p = 0$  und  $q = n$ . In die offene Newton-Cotes-Formeln gehen nur die Knoten  $x_1, \dots, x_{n-1}$  ein:

$$I(f) = \sum_{j=1}^{n-1} \beta_j f_j \quad \text{mit} \quad \beta_j = h \int_p^q \prod_{\substack{k=1 \\ k \neq j}}^{n-1} \frac{(s-k)}{(j-k)} ds$$

Im Fall  $n = 2$  ergibt sich

$$\beta_1 = h \int_0^2 ds = 2h$$

und im Fall  $n = 3$

$$\beta_1 = h \int_0^2 \frac{(s-2)}{(1-2)} ds = -h \left( \frac{s^2}{2} - 2s \right) \Big|_0^2 = ?$$

Als Fehler der Newton-Cotes-Formeln ergibt sich

$$E(f) := \int_a^b f(x) dx - I(f) = \int_a^b f(x) dx - \sum_{i=0}^n \omega_i f_i$$

Für den folgenden Satz definieren wir

$$a_+ := \begin{cases} |a| & \text{for } a \geq 0 \\ 0 & \text{for } a \leq 0 \end{cases} = \frac{1}{2}(|a| + a) \quad \text{for } a \in \mathbb{R}$$

**Satz 6.2.1 (Peano)**

Sei  $f \in C^{m+1}(I)$ ,  $I = [a, b]$  und für alle  $p \in P_m$  gelte mit dem Funktional  $E$

$$(6.5) \quad E(p) := \int_a^b p(x) dx - \sum_{i=0}^n \omega_i p_i = 0$$

Dann gilt

$$(6.6a) \quad E(f) = \int_a^b f^{(m+1)}(t) K(t) dt$$

$$(6.6b) \quad K(t) = \frac{1}{m!} E_x([(x-t)_+]^m), \quad t \in \mathbb{R}$$

wobei der Index die Integration über  $x$  anzeigt und  $K(t)$  heißt der **Peano-Kern**

**Beweis:**

Der Beweis wird mit Hilfe von Taylor geführt. Sei  $x \in I := [a, b]$ . Dann ist die Taylor-Reihe

$$(6.7) \quad f(x) = f(a) + (x-a)f'(a) + \dots + \frac{(x-a)^m}{m!} f^{(m)}(a) + R_{m+1}(x)$$

$$R_{m+1}(x) = \frac{1}{m!} \int_a^x f^{(m+1)}(t)(x-t)^m dt = \frac{1}{m!} \int_a^b f^{(m+1)}(t) [(x-t)_+]^m dt$$

wobei  $[(x-t)_+]^m = 0$  für  $(x-t \leq 0 \Leftrightarrow t \geq x)$ .

Da  $E(p)$  für alle  $p \in \mathbb{P}_m$  verschwindet, liefert die Anwendung des Funktionals  $E$  auf (6.7)

$$\begin{aligned} E(f) &= \frac{1}{m!} E_x \left( \int_a^b f^{(m+1)}(t) [(x-t)_+]^m dt \right) \\ &= \frac{1}{m!} \left\{ \int_a^b \left( \int_a^b f^{(m+1)}(t) [(x-t)_+]^m dt \right) dx - \sum_{i=0}^n \omega_i \left( \int_a^b f^{(m+1)}(t) [(x-t)_+]^m dt \right)_i \right\} \\ &= \frac{1}{m!} \int_a^b f^{(m+1)}(t) \underbrace{\left\{ \int_a^b [(x-t)_+]^m dx - \sum_{i=0}^n \omega_i [(x-t)_+]_i^m \right\}}_{E_x([(x-t)_+]^m)} dt \end{aligned}$$

■

Somit ist das weitere Vorgehen klar. Für die Konstruktion der Integrations-Formeln müssen  $x_i$  und  $\omega_i$  so bestimmt werden, daß  $E(p) = 0$  für  $p \in P$  ist. Der Quadraturfehler wird dann mit (6.6a) bestimmt. Um die Formeln zu vereinfachen sei  $J := [-s, s]$  mit  $s = \frac{b-a}{2}$  und  $g(\tilde{x}) := f(\tilde{x} + \frac{a+b}{2})$ . Dann folgt

$$\int_a^b f(x) dx = \int_{-s}^s g(\tilde{x}) d\tilde{x}$$

mit

$$\begin{aligned} \tilde{x} &= x - \frac{a+b}{2} \\ x_i &= a + \frac{b-a}{n} i = a + h i \\ \tilde{x}_i &= x_i - \frac{a+b}{2} = \frac{a-b}{2} + \frac{b-a}{n} i = -s + \frac{2s}{n} i \end{aligned}$$

Bislang waren die Knoten als  $x_i = a + \frac{b-a}{n} i$  gewählt. Von nun an seien die **Knotenpunkte der Newton-Cotes-Formeln**

$$x_i = -s + \frac{2s}{n} i$$

Dabei sind die Indexbereich bei den verschiedenen Verfahren wie folgt

- (1) Geschlossene Newton-Cotes-Formeln :  $i = 0, 1, \dots, n-1, n$
- (2) Offene Newton-Cotes-Formeln :  $i = 1, \dots, n-1$

Nun sind noch die Gewichte  $\omega_i$  wie erwähnt zu bestimmen. Sei dazu  $v_i(x) := x^i, i = 0, 1, \dots, m$ .  $\{v_0, \dots, v_m\}$  ist eine Basis von  $\mathbb{P}_m$ , d.h.  $E(p) = 0$  für alle  $p \in \mathbb{P}_m$  ist äquivalent zu

$$(6.8) \quad E(v_i) = \int_{-s}^s x^i dx - \sum_{\nu=0}^n \omega_\nu x_\nu^i = 0 \quad i = 0, 1, \dots, m$$

d.h.

$$(6.9) \quad \begin{cases} \omega_0 + \omega_1 + \dots + \omega_n = 2s & i = 0 \\ x_0\omega_0 + x_1\omega_1 + \dots + x_n\omega_n = 0 & i = 1 \\ x_0^2\omega_0 + x_1^2\omega_1 + \dots + x_n^2\omega_n = 2\frac{s^3}{3} & i = 2 \\ \vdots & \vdots \\ x_0^m\omega_0 + x_1^m\omega_1 + \dots + x_n^m\omega_n = (1 - (-1)^{m+1}) \frac{s^{m+1}}{m+1} & i = m \end{cases}$$

Ist nun  $m = n$  so ist die Koeffizientenmatrix der linken Seite von (6.9) eine Vandermonde-Matrix und somit sind die Gewichte eindeutig. Man bemerke, daß wir bis zu diesem Zeitpunkt die speziellen Newton-Cotes-Knoten nicht benötigt haben.

**Beispiel 6.2.2** (Trapez-Regel)

Die Trapez-Regel entspricht der geschlossenen NCF mit  $n = 1, m = 1, x_0 = -s, x_1 = s$ . Das lineare Gleichungssystem lautet demnach

$$\left. \begin{aligned} \omega_0 + \omega_1 &= 2s \\ -s\omega_0 + s\omega_1 &= 0 \end{aligned} \right\} \Rightarrow \omega_0 = \omega_1 = s = \frac{b-a}{2}$$

Für den Fehler der Trapez-Regel ergibt sich

$$(6.10) \quad E^T(f) = \int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b))$$

Es ist noch zu bemerken, daß  $E^T(x^2)$  nicht verschwindet, da für  $s \neq 0$  gilt

$$\int_{-s}^s x^2 ds = \frac{2}{3}s^3 \neq s \cdot 2s^2$$

### Beispiel 6.2.3 (Simpson-Regel)

Die Simpson-Regel entspricht der geschlossenen NCF mit  $n = 2$ ,  $m = 2$ ,  $x_0 = -s$ ,  $x_1 = 0$ ,  $x_2 = s$ . Das lineare Gleichungssystem lautet demnach

$$\left. \begin{array}{l} \omega_0 + \omega_1 + \omega_2 = 2s \\ -s\omega_0 + 0\omega_1 + s\omega_2 = 0 \\ s^2\omega_0 + 0\omega_1 + s^2\omega_2 = \frac{2}{3}s^3 \end{array} \right\} \Rightarrow \begin{array}{l} \omega_1 = \frac{4}{3}s \\ \omega_0 = \omega_2 = \frac{1}{3}s \end{array}$$

Für den Fehler der Simpson-Regel ergibt sich

$$(6.11) \quad \begin{aligned} E^S(f) &= \int_a^b f(x) dx - \frac{b-a}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \\ &= \int_{-s}^s g(x) dx - \sum_{i=0}^2 \omega_i g_i \end{aligned}$$

wobei

$$g_i := g(x_i) = g\left(-s + \frac{2s}{n}i\right)$$

In diesem Fall verschwindet  $E^S(x^3)$  aber  $E^S(x^4)$  nicht. (Begründung siehe Übung)

### Beispiel 6.2.4 (Mittelpunkt-Regel)

Die Mittelpunkt-Regel entspricht der offenen NCF mit  $n = 2$ ,  $m = 0$ ,  $\omega_0 = 2s$ . Für den Fehler der Mittelpunkt-Regel ergibt sich

$$E^M(f) = \int_a^b f(x) dx - (b-a)f\left(\frac{a+b}{2}\right)$$

## 6.3 Fehlerabschätzung mit Peano

Eine weitere Möglichkeit den Quadratur-Fehler zu bestimmen ergibt sich mit Peano (6.6a). Dafür wird zuerst (6.6b) für  $t \in J := [-s, s]$ , also  $t \geq -s$  bestimmt.

**Trapez-Regel (m=n=1)**

Es ist

$$\begin{aligned}
 E_x^T[(x-t)_+]^1 &= \int_{-s}^s (x-t)_+ dx - (\omega_0(x_0-t)_+ + \omega_1(x_1-t)_+) \\
 &= \int_{-s}^s (x-t)_+ dx - s[(-s-t)_+ + (s-t)_+] \\
 &= \int_t^s (x-t) dx - s(s-t) \\
 &= \frac{s^2}{2} - ts + \frac{t^2}{2} - s(s-t) = -\frac{1}{2}(s-t)(s+t) = -\frac{s^2-t^2}{2}
 \end{aligned}$$

Damit läßt sich nun mit Hilfe von Satz 6.1.1 berechnen

$$E^T(g) = -\frac{1}{2} \int_{-s}^s g''(t)(s^2-t^2) dt = -\frac{1}{2} g''(\eta) \int_{-s}^s (s^2-t^2) dt = -g''(\eta) \frac{2}{3} s^3$$

mit  $\eta \in (-s, s)$ . Für  $\xi \in (a, b)$  ergibt sich nun

$$(6.12) \quad E^T(f) = -\frac{1}{12} f''(\xi)(b-a)^3$$

**Simpson-Regel (n=m=2)**

Analog zur Trapez-Regel erhält man

$$\begin{aligned}
 E_x^S[(x-t)_+]^3 &= \int_{-s}^s [(x-t)_+]^3 dx - \sum_{i=0}^2 \omega_i [(x_i-t)_+]^3 \\
 &= \int_{-s}^s [(x-t)_+]^3 dx - \frac{s}{3} \{ [(-s-t)_+]^3 + 4(-t)_+^3 + (s-t)_+^3 \} \\
 &= \int_t^s (x-t)^3 dx - \frac{s}{3} \{ 2(|t|^3 - t^3) + (s-t)^3 \}
 \end{aligned}$$

wobei  $[(-s-t)_+]^3 = 0$  und  $4(-t)_+^3 = 2(|t|^3 - t^3)$  ist. Mit

$$\int_t^s (x-t)^3 dx = \int_0^{s-t} u^3 du = \frac{(s-t)^4}{4}, \quad u = x-t$$

folgt

$$E_x^S[(x-t)_+]^3 = \frac{(s-t)^4}{4} - \frac{s}{3}(s-t)^3 - \frac{2s}{3}(|t|^3 - t^3) = 3!K(t)$$

Dabei ist zu beachten, daß

$$\frac{(s-t)^4}{4} - \frac{s}{3}(s-t)^3 \leq 0 \leq |t|^3 - t^3$$

ist. Wiederum gilt  $K(t) \leq 0$  für  $t \in J$  und mit dem zweiten MWS 6.1.1

$$(6.13) \quad E^S(g) = \int_{-s}^s g^{(4)}(t)K(t) dt = g^{(4)}(\eta) \int_{-s}^s K(t) dt$$

Das verbleibende Integral wird mit (6.6a) für  $g(x) = x^4$  bestimmt.

$$E^S(x^4) = 4! \int_{-s}^s K(t) dt = \int_{-s}^s x^4 dx - \frac{s}{3} [(-s)^4 + 0 + s^4] = \frac{2}{5}s^5 - \frac{2}{3}s^5 = -\frac{4}{15}s^5$$

Mit

$$\int_{-s}^s K(t) dt = -\frac{4}{15 \cdot 4!} s^5 = -\frac{s^5}{90}$$

folgt schließlich

$$(6.14) \quad E^S(f) = -\frac{1}{90} f^{(4)}(\xi) \left( \frac{b-a}{2} \right)^5$$

für  $\xi \in I = [a, b]$ .

### Mittelpunkt-Regel

Ebenso läßt sich zeigen

$$(6.15) \quad E^M(f) = f^{(2)}(\xi) \frac{(b-a)^3}{3}$$

# Chapter 7

## Matrizen - Eigenwertaufgaben

### 7.1 Vorbemerkungen, Eigenwertabschätzungen

$A \in \mathbb{R}^{n \times n}$  (oder  $A \in \mathbb{C}^{n \times n}$ , i.d. Praxis selten)

$x (\neq 0) \in \mathbb{R}^n$  (oder  $\mathbb{C}^n$  zugelassen, auch für  $A \in \mathbb{R}^{n \times n}$ ) ist Eigenvektor (EV) zum Eigenwert (EW)  $\lambda \in \mathbb{C}$  der Matrix  $A$ , wenn gilt

$$(7.1) \quad Ax = \lambda x$$

EWAn z.B. bei Schwingungsproblemen, Dgl.-EWA numerische Lösungsansätze  $\implies$  Matrizen-EWA

(7.1)  $\implies$

$$(7.2) \quad \det(A - \lambda E) = \underbrace{\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}}_{(-1)^n \lambda^n + \text{Spur}(A)(-1)^{n-1} \lambda^{n-1} + \cdots + \det A} = 0$$

Spektralradius

$$\rho(A) = \max_{i=1}^n |\lambda_i|$$

wenn  $\lambda_1, \dots, \lambda_n$  die EW von  $A$  sind (gezählt entsprechend ihrer Vielfachheit).

#### Definition 7.1.1

Zwei Matrizen  $A, B \in \mathbb{R}^{n \times n}$  sind **ähnlich**, wenn es eine Matrix  $T \in \mathbb{R}^{n \times n}$  (bzw.  $\mathbb{C}^{n \times n}$ ) gibt, mit  $A = T^{-1}BT$ .

Ähnliche Matrizen haben dieselben EW:

$$\begin{aligned} \det(A - \lambda E) &= \det(T^{-1}BT - \lambda E) = \det(T^{-1}BT - \lambda T^{-1}T) = \det(T^{-1}(B - \lambda E)T) \\ &= \det T^{-1} \det(B - \lambda E) \det T = \det(B - \lambda E). \end{aligned}$$

Sei  $Ax = \lambda x$ , d.h.  $T^{-1}BTx = \lambda x \implies B \underbrace{Tx}_y = \lambda \underbrace{Tx}_y \implies y$  ist EV von  $B$  zum EW  $\lambda$  (und umgekehrt).

Ist  $A = A^T$  positiv definit, so gilt: alle EW von  $A$  sind positiv:

$$\begin{aligned} Ax &= \lambda x & x &\in \mathbb{R}^n \\ \underbrace{x^T Ax}_{>0, \text{ da } x \neq 0} &= \lambda \underbrace{x^T x}_{>0} & \implies \lambda &> 0. \end{aligned}$$

Entsprechend gilt Umkehrung.

Ist  $\lambda \in \mathbb{C}$  EW von  $A \in \mathbb{R}^{n \times n}$ , so ist auch  $\bar{\lambda}$  EW von  $A$ :

$$Ax = \lambda x \implies A\bar{x} = \bar{\lambda}\bar{x}, \quad \text{da } A \in \mathbb{R}^{n \times n} \implies \bar{\lambda} \text{ EW zum EV } \bar{x}.$$

Ist  $A = A^T$ , so sind alle EW von  $A$  reell:

$$\left. \begin{aligned} \bar{x}^T \cdot |Ax = \lambda x & \quad \bar{x}^T Ax = \lambda \bar{x}^T x \\ x^T \cdot |A\bar{x} = \bar{\lambda}\bar{x} & \quad x^T A\bar{x} = \bar{\lambda} x^T \bar{x} \end{aligned} \right\} (\lambda - \bar{\lambda}) \underbrace{\bar{x}^T x}_{>0} = 0 \implies \lambda = \bar{\lambda} \iff \lambda \in \mathbb{R}$$

$A = A^T$  besitzt  $n$  linear unabhängige EV, die als Orthonormal-Basis (ONB) gewählt werden können.

$$B = (\mathbf{b}_1, \dots, \mathbf{b}_n) \quad B^{-1}AB = D \quad D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad \lambda_i \text{ EW zu } \mathbf{b}_i.$$

EW-Abschätzungen

(i)  $\|\cdot\|_M$  M-Norm  $\implies |\lambda| \leq \|A\|_M$  für alle EW von  $A$

(ii) Abschätzung mit den **Gerschgorin-Kreisen**

Sei  $A = (a_{ik})_{i,k=1,\dots,n}$ , die Mengen

$$(7.3) \quad K_\mu = \left\{ z \in \mathbb{C} : |z - a_{\mu\mu}| \leq \sum_{\substack{\nu=1 \\ \nu \neq \mu}}^n |a_{\mu\nu}| \right\} \quad (\mu = 1, \dots, n)$$

heißen **Gerschgorin-Zeilenkreise**.

Entsprechend

$$(7.4) \quad K'_\nu = \left\{ z \in \mathbb{C} : |z - a_{\nu\nu}| \leq \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^n |a_{\mu\nu}| \right\}$$

**Gerschgorin-Spaltenkreise**.

**Satz 7.1.2**

a) Seien  $\lambda_1, \dots, \lambda_n \in \mathbb{C}$  die EW von  $A$  (gezählt entsprechend der Vielfachheit), so gilt für alle  $k$

$$\lambda_k \in \bigcup_{\mu=1}^n K_\mu, \quad \lambda_k \in \bigcup_{\nu=1}^n K'_\nu \implies \lambda_k \in \left( \bigcup_{\mu=1}^n K_\mu \right) \cap \left( \bigcup_{\nu=1}^n K'_\nu \right)$$

b) Ist die Vereinigung von  $m$  ( $\leq n$ ) Gerschgorin-Zeilenkreisen bzw. Spaltenkreisen disjunkt zu den übrigen (d.h. Durchschnitt leer), so enthält sie genau  $m$  EW (entsprechend ihrer Vielfachheit).

**Bemerkung 7.1.3**

Numerisch brauchbare Werte i.a. nur bei diagonaldominanten Matrizen zu erwarten.

**Beispiel 7.1.4**

$$A = \begin{pmatrix} 1 & 2 & -1 \\ -2 & 3 & 1 \\ -3 & 8 & 1 \end{pmatrix} \quad EW: \lambda_{1,2} = 0, \lambda_3 = 5$$

Gerschgorin-Zeilenkreise  $|z - 1| \leq 3, |z - 3| \leq 3, |z - 1| \leq 11$

eventuell Verbesserung durch Ähnlichkeitstransformation mit Diagonalmatrizen; für obiges Beispiel 7.1.4

$$T = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad T^{-1}AT = \begin{pmatrix} 1 & 2 & -2 \\ -2 & 3 & 2 \\ -\frac{3}{2} & 4 & 1 \end{pmatrix} \text{ hat gleiche EW}$$

Gerschgorin-Zeilenkreise  $|z - 1| \leq 4, |z - 3| \leq 4, |z - 1| \leq 5.5$

**7.2 Die Verfahren von Wilkinson und von Householder**

Ziel:  $A$  durch Ähnlichkeitstransformation auf einfachere Form bringen, i.a. auf sog. Hessenberg-Form.

**Definition 7.2.1**

Eine Matrix  $H = (h_{ik})_{i,k=1,\dots,n}$  heißt **Hessenberg-Matrix**, wenn gilt  $h_{ik} = 0$  für  $k + 1 < i$  ( $i, k = 1, \dots, n$ )

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & & h_{2n} \\ 0 & h_{32} & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{nn} \end{pmatrix}$$

Wilkinson: schrittweites Überführen auf Hessenberg-Form

$$(A =) A^{(0)} \xrightarrow{\text{Ähnl.trafo}} A^{(1)} \xrightarrow{\text{Ähnl.trafo}} A^{(2)} \longrightarrow \dots \longrightarrow A^{(n-2)} = H$$

Die Matrizen  $A^{(j-1)}$  sind von der Form

$$A^{(j-1)} = \begin{pmatrix} * & & & & & & \\ & \ddots & & & & & \\ 0 & \ddots & * & & & & \\ \vdots & \ddots & * & * & & & \\ 0 & \cdots & 0 & * & \ddots & & \\ \vdots & & \vdots & \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & * & & \ddots & \ddots \end{pmatrix} = (a_{ik}^{(j-1)})_{i,k=1,\dots,n}$$

**Wilkinson-Algorithmus:**

$A^{(j-1)}$  sei bestimmt

- (i)  $q$  ( $j + 1 \leq q \leq n$ ) so bestimmen, dass gilt

$$(7.5) \quad |a_{qj}^{(j-1)}| = \max_{j+1 \leq \mu \leq n} |a_{\mu j}^{(j-1)}|$$

vgl. partielle Pivotwahl bei Gauß, d.h. es ist in der  $j$ -ten Spalte unterhalb der Hauptdiagonalen das betragsgrößte Element zu suchen (bei mehreren z.B. kleinster Zeilenindex)

- (ii) Für  $q \neq j + 1$  Vertauschen der  $(j + 1)$ -te Zeile und Spalte mit der  $q$ -ten Zeile und Spalte

$$(7.6) \quad A^{(j-1)} \longrightarrow E_{j+1,q} A^{(j-1)} E_{j+1,q} = B^{(j-1)} = (b_{ik}^{(j-1)})_{i,k}$$

wegen  $E_{j+1,q} = E_{j+1,q}^{-1}$  sind  $A^{(j-1)}$  und  $B^{(j-1)}$  ähnlich.

- (iii) Ist  $b_{j+1,j}^{(j-1)} = 0$  (d.h. wären unter der Hauptdiagonalen von  $A^{(j-1)}$  nur Nullen), dann nächster Schritt, sonst Matrix  $C_{j+1} = (c_{ik})_{i,k}$  bestimmen mit

$$(7.7) \quad c_{i,j+1} = \frac{b_{ij}^{(j-1)}}{b_{j+1,j}^{(j-1)}} \quad (i > j + 1)$$

$$(7.8) \quad A^{(j)} = C_{j+1}^{-1} B^{(j-1)} C_{j+1}$$

**Bemerkung 7.2.2**

zu (iii):

$$A^{(j)} = \underbrace{C_{j+1}^{-1} E_{j+1,q}}_{=F^{-1} \text{ wegen } E_{j+1,q} = E_{j+1,q}^{-1}} A^{(j-1)} \underbrace{E_{j+1,q} C_{j+1}}_{=F}$$

d.h.  $A^{(j)}$  ist ähnlich zu  $A^{(j-1)}$ .

Vertauschen von Zeilen und Spalten  $j - 1 \longleftrightarrow q$  bewirkt keine Veränderung der Nullstruktur

$$\text{von } \begin{pmatrix} 0 & & & & \\ \vdots & \ddots & & & \\ 0 & \cdots & 0 & & \\ \vdots & & & & \\ 0 & \cdots & 0 & & \end{pmatrix}, \text{ wegen } C_{j+1}^{-1} = \begin{pmatrix} 1 & & & & \\ & \vdots & & & 0 \\ & & 1 & & \\ & & -c_{j+2,j} & 1 & \\ 0 & & \vdots & & \ddots \\ & & -c_{n,j} & & 1 \end{pmatrix}.$$

(7.7) erzeugt  $\begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ .

**Beispiel 7.2.3**

$$A = A^{(0)} = \begin{pmatrix} 1 & -2 & 3 & -2 \\ 1 & 5 & -1 & -1 \\ 2 & 3 & 2 & -2 \\ 2 & -2 & 6 & -3 \end{pmatrix} \text{ vertausche 2. und 3. Spalte/Zeile}$$

$$B^{(0)} = \begin{pmatrix} 1 & 3 & -2 & -2 \\ 2 & 2 & 3 & -2 \\ 1 & -1 & 5 & -1 \\ 2 & 6 & -2 & -3 \end{pmatrix} \text{ Matrix } C_2 \text{ bilden}$$

$$C_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.5 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \implies C_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -0.5 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

$$C_2^{-1} \cdot B^{(0)} = \begin{pmatrix} 1 & & & \\ 2 & * & & \\ 0 & & & \\ 0 & & & \end{pmatrix}$$

$$A^{(1)} = C_2^{-1} B^{(0)} C_2 = \begin{pmatrix} 1 & 0 & -2 & -2 \\ 2 & 1.5 & 3 & -2 \\ 0 & -0.25 & 3.5 & 0 \\ 0 & 0.5 & -5 & -1 \end{pmatrix} \text{ Pivotelement 0.5} \implies \text{vertausche 3. und 4. Zeile}$$

und Spalte.

Schließlich  $C_3$  bestimmen

$$C_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -0.5 & 1 \end{pmatrix}$$

$$A^{(2)} = \underbrace{H}_{\text{Hessenbergmatrix}} = C_3^{-1} B^{(1)} C_3 = \begin{pmatrix} 1 & 0 & -1 & -2 \\ 2 & 1.5 & -3.5 & 3 \\ 0 & 0.5 & 1.5 & -5 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Bestimmung der EW von  $H$  s.u.

Sollen auch die EV bestimmt werden, dann evtl. Transformationsmatrizen  $T$  mit  $H = T^{-1}AT$  abspeichern, wobei  $T$  Produkt von  $E_{p,q}$  und  $C_p$ -Matrizen.

**Verfahren von Householder** (entsprechend wie Wilkinson)

$$A \longrightarrow A^{(1)} \longrightarrow A^{(2)} \longrightarrow \dots \longrightarrow A^{(n-2)} \quad (\text{Hessenberg - Matrix})$$

Transformationsmatrizen  $P = E - \frac{1}{c} w w^T$ ,  $w \in \mathbb{R}^n$

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \cdot (w_1, \dots, w_n) = \begin{pmatrix} w_1^2 & w_1 w_2 & \cdots & w_1 w_n \\ \vdots & \vdots & & \vdots \\ w_n w_1 & w_n w_2 & \cdots & w_n^2 \end{pmatrix} \quad \text{dyadisches Produkt}$$

$\implies$  für  $c \in \mathbb{R}$  ist  $P = P^T$

es sei  $c = \frac{1}{2}w^T w$ , dann ist

$$P^2 = (E - \frac{1}{c}ww^T)(E - \frac{1}{c}ww^T) = E - \frac{2}{c}ww^T + \frac{1}{c^2}w \underbrace{w^T w}_{2c} w^T = E$$

d.h.  $P^{-1} = P$  ( $P$  ist symmetrisch und orthogonal (involutorisch)).

Householder für 1. Schritt:

$$w = \begin{pmatrix} 0 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \quad A^{(1)} = PAP \stackrel{!}{=} \begin{pmatrix} * & & & \\ * & & & \\ 0 & & * & \\ \vdots & & & \\ 0 & & & \end{pmatrix}$$

$A^{(1)} = (a_{ik}^{(1)})$  es gilt  $a_{11}^{(1)} = a_{11}$

also, zu bestimmen:  $a_{21}^{(1)}, w_2, \dots, w_n$  führt auf NLGS, speziell gilt  $\frac{w_2^2}{2\|w\|_2^2} = 1 - \frac{a_{21}}{\pm \sqrt{\sum_{i=2}^n a_{i1}^2}}$  (\*).

1)  $\sum_{i=2}^n a_{i1}^2 = 0$ , dann  $A = \begin{pmatrix} * & & & \\ 0 & & * & \\ \vdots & & & \\ 0 & & & \end{pmatrix}$ , d. h.  $A$  hat schon gewünschte Form  $\implies$  1. Schritt

überschlagen

2)  $\sum_{i=2}^n a_{i1}^2 > 0$ , Vorzeichen in (\*) eigentlich beliebig, aber wegen Rechenstabilität so wählen, dass  $|w_2|$  möglichst groß wird, dann Householder i.a. sehr rechenstabil.

Verfahren auf Restmatrizen (vgl. Wilkinson)  $A^{(1)}, A^{(2)}, \dots$  fortsetzen

$\implies$  Algorithmus

$A^{(0)} = A$  für  $k = 1, \dots, n - 2$  sind folgende Schritte durchzuführen

$$s = \sqrt{\sum_{j=k+1}^n a_{jk}^{(k-1)2}}$$

$s = 0$ , dann nächsten Schritt ausführen

$$s \neq 0 : \quad t = a_{k+1,k}^{(k-1)} \quad c = s(s + |t|)$$

(7.9)  $w_j^{(k)} = 0 \quad (j = 1, \dots, k)$

$$w_{k+1}^{(k)} = t + s \cdot \text{sign}(t)$$

$$w_j^{(k)} = a_{jk}^{(k-1)} \quad (j = k + 2, \dots, n)$$

$$P^{(k)} = E - \frac{1}{c}w^{(k)}w^{(k)T}$$

$$A^{(k)} = P^{(k)} A^{(k-1)} P^{(k)}$$

$$w^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_{k+1}^{(k)} \\ \vdots \\ w_n^{(k)} \end{pmatrix}$$

Vergleich der Rechenoperationen (Div./Multi.)

Wilkinson:  $\frac{5}{6}n^3 + O(n^2)$

Householder:  $\frac{5}{3}n^3 + O(n^2)$

also: für allgemeine Matrizen Wilkinson günstiger, Householder allerdings noch rechenstabiler.

Aber:  $A = A^T$  s.o.  $A^{(1)T} = (PAP)^T = \underbrace{P^T}_P \underbrace{A^T}_A P = A^{(1)}$ , d.h. auch alle „Zwischenmatrizen“  $A^{(k)}$  sowie  $H = A^{(n-2)}$  sind symmetrisch  $\implies$

$$H = \begin{pmatrix} * & * & & 0 \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix} \quad \text{Tridiagonalmatrix.}$$

Für  $A = A^T$  (insbesondere) müssen die  $P^{(k)}$  nicht explizit berechnet werden:  $c, w^{(k)}$  wie berechnet

$$\begin{aligned} (7.10) \quad u^{(k)} &= \frac{1}{c} A^{(k-1)} w^{(k)} \\ v^{(k)} &= u^{(k)} - \frac{1}{2c} w^{(k)} \cdot \underbrace{w^{(k)T} w^{(k)}}_{\in \mathbb{R}} \\ \implies A^{(k)} &= A^{(k-1)} - v^{(k)} w^{(k)T} - w^{(k)} v^{(k)T} \end{aligned}$$

In diesem Fall werden nur  $\frac{2}{3}n^3 + O(n^2)$  Div./Mult.-Operationen benötigt, d.h hier stets Householder vorzuziehen.

### 7.3 Berechnung von EW und EV von Hessenberg-Matrizen

$H = (h_{ik}) \in \mathbb{R}^{n \times n}$

charakteristisches LGS

$$(7.11) \quad \begin{aligned} (h_{11} - \lambda)x_1 + h_{12}x_2 + \dots + h_{1n}x_n &= 0 \\ h_{21}x_1 + (h_{22} - \lambda)x_2 + \dots + h_{2n}x_n &= 0 \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ h_{n,n-1}x_{n-1} + (h_{nn} - \lambda)x_n &= 0 \end{aligned}$$

charakteristische Gleichung

$$\begin{vmatrix} h_{11} - \lambda & \dots & \dots & h_{1n} \\ & h_{22} - \lambda & \dots & h_{2n} \\ & \vdots & \ddots & \vdots \\ 0 & \dots & h_{n,n-1} & h_{nn} - \lambda \end{vmatrix} \stackrel{!}{=} 0$$

1.) es gibt ein  $k$  ( $1 \leq k \leq n - 1$ ) mit  $h_{k+1,k} = 0 \implies$  LGS (7.11) zerfällt in 2 getrennte LGS,  $\implies \det(H - \lambda E_n) = \det(H - \lambda E_\rho) \cdot \det(H - \lambda E_{n-\rho})$

D.h. weiter mit 2.) bei geringerer Reihenzahl

2.) für alle  $k$  ( $1 \leq k \leq n-1$ )  $h_{k+1,k} \neq 0$  (sei im folgenden stets erfüllt)

( $\alpha$ ) Entwicklung der  $k \times k$ -Unterdeterminanten in der linken oberen Ecke der char. Determinanten nach jeweiliger letzter Spalte

$$(f_0(\lambda) = 1) \quad f_1(\lambda) = (h_{11} - \lambda)f_0(\lambda)$$

$$f_2(\lambda) = \begin{vmatrix} h_{11} - \lambda & h_{12} \\ h_{21} & h_{22} - \lambda \end{vmatrix} = (h_{22} - \lambda)f_1(\lambda) - h_{12}h_{21}$$

$$f_3(\lambda) = \begin{vmatrix} h_{11} - \lambda & h_{12} & h_{13} \\ h_{21} & h_{22} - \lambda & h_{23} \\ 0 & h_{32} & h_{33} - \lambda \end{vmatrix} = (h_{33} - \lambda)f_2(\lambda) - h_{23}h_{32}f_1(\lambda) + h_{13}h_{32}h_{21}f_0(\lambda)$$

usw. (vollst. Induktion)

$$\begin{aligned} f_0(\lambda) &= 1 \\ f_j(\lambda) &= (\alpha_{jj} - \lambda)f_{j-1}(\lambda) + \alpha_{j-1,j}f_{j-2}(\lambda) + \cdots + \alpha_{2j}f_1(\lambda) + \alpha_{1j}f_0(\lambda) \end{aligned} \quad (j = 1, \dots, n)$$

(7.12)

$$\begin{aligned} \alpha_{jj} &= h_{jj} \\ \alpha_{kj} &= (-1)^{j+k} h_{kj} h_{j,j-1} h_{j-1,j-2} \cdots h_{k+1,k} \quad (k = 1, \dots, j-1) \end{aligned}$$

$$f_n(\lambda) = \det(H - \lambda E)$$

### Beispiel 7.3.1

s.o. *Wilkinson*

$$H = \begin{pmatrix} 1 & 0 & -1 & -2 \\ 2 & 1.5 & -3.5 & 3 \\ 0 & 0.5 & 1.5 & -5 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

$$\begin{aligned} (f_0(\lambda) &= 1) \quad f_1(\lambda) = 1 - \lambda \\ f_2(\lambda) &= (1.5 - \lambda)f_1(\lambda) - 0.2 = \lambda^2 - 2.5\lambda + 1.5 \\ f_3(\lambda) &= (1.5 - \lambda)f_2(\lambda) - (-3.5)0.5f_1(\lambda) + (-1)0.5 \cdot 2 \cdot f_0(\lambda) = -\lambda^3 + 4\lambda^2 - 7\lambda + 3 \\ f_4(\lambda) &= (1 - \lambda)f_3(\lambda) - (-5)(-1)f_2(\lambda) + 3(-1)0.5 \cdot f_1(\lambda) - (-2)(-1)0.5 \cdot 2f_0(\lambda) \\ &= \lambda^4 - 5\lambda^3 + 6\lambda^2 + 4\lambda - 8 \end{aligned}$$

Lösungen  $\lambda_{1,2,3} = 2, \quad \lambda_4 = -1$

( $\beta$ ) Bestimmung von EV von  $H$  (zum berechneten EW  $\lambda$ )  
charakteristisches LGS (7.11) hat eine nichttriviale Lösung.

1)  $x_n = 0$  setzen letzte Gleichung  $\implies x_{n-1} = 0$

vorletzte Gleichung  $\implies x_{n-2} = 0$

...

$$2. \text{ Gleichung} \implies x_1 = 0$$

d.h. kein EV mit  $x_n = 0$ . Daher

2)  $x_n \neq 0$  (o.B.d.A.)  $x_n = 1$  setzen  $\implies x_{n-1}, \dots, x_1$  eindeutig bestimmt  $\implies$  nur 1 EV bis auf lin. Unabh.

D.h. jede nichtzerfallende Hessenbergmatrix hat genauso viele linear unabhängige EV, wie sie verschiedene EW hat.

### Beispiel 7.3.2

obige Matrix  $H$ , (7.11) für  $\lambda = 2$

$x_4 = 1$  gewählt, 1. Gleichung wegen lin. Unabh. weglassen

$$\begin{aligned} 2x_1 - 0.5x_2 - 3.5x_3 &= -3 \\ 0.5x_2 - 0.5x_3 &= 5 \\ -x_3 &= 1 \end{aligned} \implies x = \begin{pmatrix} -1 \\ 9 \\ -1 \\ 1 \end{pmatrix}$$

LGS mit  $\Delta$ -Matrix

ist einziger EV von  $H$  zu  $\lambda = 2$

$$\text{Entsprechend } \lambda = -1, x_4 = 1 \implies \tilde{x} = \begin{pmatrix} 2 \\ 0 \\ 2 \\ 1 \end{pmatrix}$$

Transformationsmatrix war

$$\begin{aligned} T &= E_{23} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.5 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \cdot E_{34} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -0.5 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & -0.5 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \\ \implies T \begin{pmatrix} -1 \\ 9 \\ -1 \\ 1 \end{pmatrix} &= \begin{pmatrix} -1 \\ 6 \\ 9 \\ 8 \end{pmatrix} \text{ ist (bis lin. Unabh.) einziger EV von } A = \begin{pmatrix} 1 & -2 & 3 & -2 \\ 1 & 5 & -1 & -1 \\ 2 & 3 & 2 & -2 \\ 2 & -2 & 6 & 3 \end{pmatrix} \end{aligned}$$

$$\text{entsprechend } T \begin{pmatrix} 2 \\ 0 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 2 \end{pmatrix}.$$

### ( $\gamma$ ) Verfahren von Hyman

char. Polynom  $p(\lambda) = \det(H - \lambda E) = 0$  für EW

Aus (7.11) folgt (im Prinzip entsprechend der Bestimmung der EV)

$$(7.13) \quad p(\lambda) = (-1)^n h_{21} \cdot h_{32} \cdots h_{n,n-1} \cdot \rho(\lambda)$$

wobei für geg.  $\lambda \in \mathbb{R} (\in \mathbb{C})$  sich  $\rho(\lambda)$  rekursiv berechnen lässt:

$h_{10} = 0 \quad x_n = 1$  gesetzt

für  $i = 1, \dots, n$ :

$$(7.14) \quad x_{n-i} = \frac{1}{\underbrace{h_{n-i+1, n-i}}_{(\neq 0!)}} [\lambda x_{n-i+1} - \sum_{j=n-i+1}^n h_{n-i+1, j} x_j]$$

$x_0 = \rho(\lambda)$

ist  $\rho(\lambda) = 0$ , so ist  $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$  zugehöriger EV.

Entsprechend für die Ableitungen

$$(7.15) \quad \begin{aligned} y_n &= 0 \quad (\text{gesetzt}) \\ y_{n-i} &= \frac{1}{h_{n-i+1, n-i}} [x_{n-i+1} + \lambda y_{n-i+1} - \sum_{j=n-i+1}^n h_{n-i+1, j} y_j] \\ y_0 &= \rho'(\lambda) \end{aligned}$$

$\rho(\lambda)$ ,  $\rho'(\lambda)$  evtl. für Newton-Verfahren einsetzen (auch für komplexe EW üblich).

### Bemerkung 7.3.3

Methoden von 7.3 zur EW-Bestimmung (anders als Householder) nicht sehr stabil, daher für große  $n$  oft QR-Verfahren.

## 7.4 von-Mises Verfahren

**Ziel:** Bestimmung von gewissen EW, als erstes den betragsgrößten EW (den dominanten EW)

### Bemerkung 7.4.1

Sei  $A = A^T \in \mathbb{R}^{n \times n}$  ( $A$  symmetrisch). Es ist bekannt, dass

(a) alle EW reell sind. Wir sortieren  $\lambda_i$  so, dass

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$$

(mit Beachtung der Vielfachheiten);

(b) es gibt  $n$  linear unabhängige EV  $x^{(1)}, \dots, x^{(n)}$  mit  $Ax^{(i)} = \lambda_i x^{(i)}$  ( $A$  ist **diagonalisierbar**). Die EV können zur Bestimmung einer Orthonormalbasis benutzt werden

$$(7.16) \quad x^{(i)T} x^{(k)} = \begin{cases} 1 & , \quad i = k \\ 0 & , \quad i \neq k \end{cases} .$$

Wähle beliebiges  $z^{(0)} \in \mathbb{R}^n$ . Da die  $x^{(i)}$  eine Basis des  $\mathbb{R}^n$  bilden, können wir  $z^{(0)}$  so ausdrücken:

$$z^{(0)} = c_1 x^{(1)} + \dots + c_n x^{(n)}.$$

Sei  $c_1 \neq 0$  (dies kann oft nicht bewiesen werden, ist aber „für gewöhnlich“ erfüllt). Um dies zu prüfen, nehme man verschiedene  $z^{(0)}$ .

**von-Mises Verfahren :**

$$(7.17) \quad z^{(v)} = A z^{(v-1)} \quad (v = 1, 2, \dots)$$

Wir haben

$$(7.18) \quad \begin{aligned} z^{(1)} &= A z^{(0)} = c_1 A x^{(1)} + \dots + c_n A x^{(n)} = c_1 \lambda_1 x^{(1)} + \dots + c_n \lambda_n x^{(n)} \\ z^{(2)} &= A z^{(1)} = A^2 z^{(0)} = \dots = c_1 \lambda_1^2 x^{(1)} + \dots + c_n \lambda_n^2 x^{(n)} \\ &\vdots \\ z^{(v)} &= A z^{(v-1)} = A^v z^{(0)} = \dots = \underbrace{c_1 \lambda_1^v x^{(1)}}_{\star} + \dots + c_n \lambda_n^v x^{(n)}. \end{aligned}$$

★ überwiegt gegenüber den anderen Termen, falls  $c_1 \neq 0$ .

Für  $v \rightarrow \infty$  bekommen wir asymptotisches Verhalten

$$(7.19) \quad z^{(v)} \sim c_1 \lambda_1^v x^{(1)}, \quad z^{(v+1)} \sim c_1 \lambda_1^{v+1} z^{(v)}.$$

Falls  $z_i^{(v)} \neq 0$  folgt

$$q_i^{(v+1)} = \frac{z_i^{(v+1)}}{z_i^{(v)}} \rightarrow \lambda_1 \quad (i = 1, \dots, n).$$

Bei der Berechnung werden die Ergebnisse oft normiert, damit sie nicht zu groß oder zu klein werden:

$$\hat{z}^{(1)} = A z^{(0)}, \quad z^{(1)} = \frac{\hat{z}^{(1)}}{\|\hat{z}^{(1)}\|_\infty} \quad (\text{etc.})$$

**Beispiel 7.4.2**

$A (= A^T)$	$z^{(0)}$	$\hat{z}^{(1)}$	$z^{(1)}$	$\hat{z}^{(2)}$	$z^{(2)}$
5 -2 -4	1	5	1	9	1
-2 2 2	0	-2	-0.4	-4.4	-0.489
-4 2 5	0	-4	-0.8	-8.8	-0.978

$$z^{(v)} \rightarrow \begin{pmatrix} 1 \\ -0.5 \\ -1 \end{pmatrix} \quad \text{EV von } A$$

Durch die Normierung haben wir nun anstatt Asymptotik Konvergenz (7.19).

$$q_1^{(4)} = 9.9888, \quad q_2^{(4)} = 10.0086, \quad q_3^{(4)} = 10.0085.$$

Es gilt  $q_j^{(v+1)} \rightarrow 10 = \lambda_1$ . Wir haben gute Konvergenz, da  $\lambda_2 = \lambda_3 = 1 \ll 10$ .

Von (7.18) folgt

$$(7.20) \quad \left| \lambda_1 - q_j^{(v+1)} \right| = \left| \frac{\lambda_2}{\lambda_1} \right|^v \cdot O(1),$$

d.h. dass die Konvergenz für  $|\lambda_2| \approx |\lambda_1|$  nur mittelmäßig ist.

**Bemerkung 7.4.3**

Das von-Mises Verfahren kann auch auf nichtsymmetrische Matrizen ( $A \neq A^T$ ) angewendet werden, es ist aber notwendig, dass  $A$   $n$  linear unabhängige EV besitzt.

**Verbesserung für  $A = A^T$  mit Hilfe des Rayleigh-Quotienten**

Sei  $x \in \mathbb{R}^n$ ,  $x \neq 0$ .

Der Wert

$$(7.21) \quad R[x] = \frac{x^T A x}{x^T x}$$

heißt **Rayleigh-quotient** (von  $A$  für den Vektor  $x$ ).

**Satz 7.4.4**

Sei  $A = A^T$ . Der Rayleigh-Quotient erreicht sein Maximum / Minimum an dem EV, der zu dem größten EW gehört.

**Beweis:**

Sei  $0 \neq x \in \mathbb{R}^n$  mit der Darstellung

$$x = c_1 x^{(1)} + \dots + c_n x^{(n)},$$

wobei  $x^{(i)}$  die EV von  $A$  sind. Dann gilt

$$\begin{aligned} x^T A x &= \left( c_1 x^{(1)T} + \dots + c_n x^{(n)T} \right) \left( \lambda_1 c_1 x^{(1)} + \dots + \lambda_n c_n x^{(n)} \right) \\ &\stackrel{(7.16)}{=} \lambda_1 c_1^2 + \dots + \lambda_n c_n^2 \end{aligned}$$

und  $x^T x = c_1^2 + \dots + c_n^2$ .

Wir folgern

$$\lambda_{min} \leq R[x] = \frac{\lambda_1 c_1^2 + \dots + \lambda_n c_n^2}{c_1^2 + \dots + c_n^2} \leq \lambda_{max}.$$

■

Wir können den Rayleigh-Quotienten  $R[z^{(v)}]$  im von-Mises-Verfahren mit kleinem zusätzlichem Aufwand berechnen:

$$R[z^{(v)}] = \frac{z^{(v)T} A z^{(v)}}{z^{(v)T} z^{(v)}} = \frac{z^{(v)T} \hat{z}^{(v+1)}}{z^{(v)T} z^{(v)}}.$$

Aus (7.18) folgt

$$(7.22) \quad \left| \lambda_1 - R[z^{(v)}] \right| = \left( \frac{\lambda_2}{\lambda_1} \right)^{2v} \cdot O(1).$$

Dies ist eine erhebliche Verbesserung im Vergleich zu (7.20).

Beispiel 7.4.2 ergibt:

$$R[z^{(3)}] = \frac{z^{(3)T} \hat{z}^{(4)}}{z^{(3)T} z^{(3)}} = 9.9997 < 10 = \lambda_1 = \lambda_{max}.$$

Nun weitere Bemerkungen zum von-Mises-Verfahren.

**Bemerkung 7.4.5**

(a) Let  $\lambda_1 = \dots = \lambda_p$  and  $|\lambda_p| > |\lambda_{p+1}|$ . Dann folgt

$$z^{(0)} = \underbrace{c_1x^{(1)} + \dots + c_px^{(p)}}_{=: y} + \sum_{i=p+1}^n c_ix^{(i)},$$

wobei  $y$  EV zum EW  $\lambda_1$  ist. Die ergibt

$$z^{(v)} = \lambda_1^v y + \sum_{i=p+1}^n c_i \lambda_i^v x^{(i)},$$

und daher gilt

$$q_i^{(v)} \rightarrow \lambda_1, \quad z^{(v)} \sim \lambda_1^v y,$$

d.h. dass es kaum Veränderung im Verhalten von  $q_i$  gibt.

(b) Sei  $\lambda_1 = -\lambda_2$ . Dann ist  $q_i^{(v)}$  nutzlos durch die „Oszillation“ der Iteration. Daher nehme

$$\tilde{q}_i^{(v)} := \frac{z_i^{(v)}}{z_i^{(v-2)}}.$$

Denn es gilt  $\tilde{q}_i^{(v)} \rightarrow \lambda_1^2$ .

**(c) Inverse Iteration**

In vielen Fällen (Oszillation, Knicklasten, ...) ist der betragskleinste EW gesucht (i.a.  $\neq 0$ ).

$$Ax = \lambda_n x \Rightarrow \frac{1}{\lambda_n} x = A^{-1}x$$

d.h., dass wir den dominanten EW bestimmen müssen  $K_n (= \frac{1}{\lambda_n})$  of  $A^{-1}$ .

**von-Mises:**

$$z^{(v+1)} = A^{-1}z^{(v)} \Leftrightarrow Az^{(v+1)} = z^{(v)}.$$

Wir müssen also in jedem Iterationsschritt ein lineares Gleichungssystem lösen, wobei die Matrix in jeden Schritt gleich bleibt. Nur die rechte Seite ändert sich.

**(d) Wielandt Korrektur für EW**

Sei  $l$  eine Approximation von  $\lambda_j$  ( $1 \leq j \leq n$ ). Dann hat  $A-lE$  die EW  $\lambda_i-l$  ( $i = 1, \dots, n$ ):

$$Ax = \lambda_j x \Rightarrow (A-lE)x = (\lambda_j-l)x.$$

Falls  $l$  eine gute Approximation von  $\lambda_j$  ist, dann ist  $\lambda_j-l$  der betragskleinste EW von  $A-lE$ .

( $\Rightarrow$  Inverse Iteration:  $(A-lE)z^{(v+1)} = z^{(v)}$ .)

**Achtung:** Beachte, dass  $A-lE$  „fast singular“ ist und singularer wird, je näher  $l$  zu  $\lambda$  kommt. Aber es gibt eine „stabile Verbindung“ zum QR Algorithmus.

**(e) Bestimmung von höheren EW durch Matrix-Deflation**

(z.B.  $\lambda_2$  in Oszillations Problemen)

Bestimme  $x^{(1)}$ ,  $\lambda_1$  näherungsweise mit von-Mises und  $z^{(0)}$  so, dass  $z^{(0)T}x^{(1)} = 0$ .

$$z^{(0)} = \underbrace{c_1}_{=0} x^{(1)} + c_2x^{(2)} + \dots + c_nx^{(n)}, \quad x^{(1)T}x^{(1)} = 1,$$

dann wiederhole Anwendung des von-Mises-Verfahrens.

**Aber:** Leider ist  $c_1 \neq 0$  durch den Approximationsfehler (oder Rundungsfehler) und somit konvergiert das Verfahren gegen  $\lambda_1$ , falls man lang genug rechnet.

**Besser:** Der Einfluss von  $\lambda_1$  kann durch Modifikation der Matrix  $A$  reduziert werden:

$$\begin{aligned} B &= A - \lambda_1 x^{(1)} x^{(1)T} \\ Bx^{(1)} &= Ax^{(1)} - \lambda_1 x^{(1)} \left( x^{(1)T} x^{(1)} \right) = 0 \\ Bx^{(i)} &= Ax^{(i)} - \lambda_1 x^{(1)} \underbrace{\left( x^{(1)T} x^{(i)} \right)}_{= 0} = \lambda_i x^{(i)} \end{aligned}$$

d.h.  $B$  hat die EW  $0, \lambda_2, \dots, \lambda_n$  mit den EV  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ .

**in der Praxis:** Bestimme  $\lambda_1$ ,  $x^{(1)}$  näherungsweise.

$\Rightarrow$   $B$  hat evtl. nicht den exakten EW  $0$ , aber  $\lambda_2$  ist dominant.