Numerik I

Universität Siegen
Wintersemester 2004/05

Prof. Dr. Franz-Jürgen Delvos
Prof. Dr. Hans-Jürgen Reinhardt

Überarbeitet, ergänzt und in LaTeX gesetzt von
Uwe Nowak und Christian Schneider

Version: 10. Juli 2005

Für die Unterstützung bei der Erstellung dieses Scripts danken wir:

- Jana Peters für die Bereitstellung des LaTeX-Codes ihres Scripts zur Numerik
- Christoph Otto für technische Unterstützung
- Verschiedenen Kommilitonen/innen für das Finden und Melden diverser Fehler

Wir haben uns Mühe gegeben, Fehler zu finden und zu korrigieren. Dennoch wird dieses Script vermutlich Fehler enthalten. Falls ihr Fehler findet, meldet diese bitte Uwe Nowak (mail@uwenowak.de).

Inhaltsverzeichnis

1 Berechnung von Polynom- und Reihenwerten

1.1 Hornerschema

Das Hornerschema wird verwendet, um

- Funktionswerte eines Polynoms und seiner Ableitung an bestimmten Stellen auszuwerten,
- ein Polynom durch einen linearen Faktor zu dividieren,
- das Taylorpolynom eines Polynoms um einen beliebigen Entwicklungspunkt zu berechnen.

Der Vorteil des Hornerschemas ist, dass der Rechenaufwand linear mit dem Grad des Polynoms steigt, während bei naivem Ausrechnen der Aufwand quadratisch zunimmt.

Bemerkung 1.1 (Nicht-Definiertes ist null)
Um in diesem Skript überflüssige Fallunterscheidungen in Rekursionsformeln zu vermeiden, wird Nicht-Definiertes als 0 definiert. Insbesondere ist dieses der Fall, wenn Indizes zu Beginn oder Ende einer Rekursion den gültigen Bereich überschreiten.

1.1.1 Einfaches Hornerschema

Mittels einfachen Hornerschemas kann ein gegebenes Polynom an einer Stelle ξ ausgewertet werden.

Beispiel 1.2 (Idee des Hornerschemas)
Die Auswertung des Polynoms

$$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

an einer Stelle ξ ist

$$\begin{aligned} p(\xi) &= a_0 + a_1\xi + a_2\xi^2 + a_3\xi^3 \\ &= a_0 + a_1\xi + (a_2 + a_3\xi)\xi^2 \\ &= a_0 + (a_1 + (a_2 + a_3\xi)\xi)\xi. \end{aligned}$$

Proposition 1.3 (Rekursionsformel des Hornerschemas)
Sei p ein Polynom,

$$p(x) = \sum_{k=0}^{n} a_k^0 x^k \quad \textit{mit} \quad a_n^0 \neq 0$$

Zur Berechnung von $p(\xi)$ setze (sukzessive für a_n^1 bis a_0^1):

$$\begin{aligned} a_n^1 &= a_n^0 \\ a_j^1 &= a_j^0 + a_{j+1}^1 \xi \quad (\textit{für } j = n-1, \ldots, 0)\,. \end{aligned}$$

Dann gilt $p(\xi) = a_0^1$.

Beweis. Der Beweis folgt durch Verallgemeinerung des obigen Beispiels. □

Proposition 1.4 (Lineare Polynomdivision mittels Hornerschema)
Seien p_n, p_{n-1} Polynome,

$$p_n(x) = \sum_{j=0}^{n} a_j^0 x^j$$
$$p_{n-1}(x) = \sum_{j=1}^{n} a_j^1 x^{j-1}.$$

Dann gilt für beliebiges ξ und nach obigem Schema entwickelten a_0^1

$$p_n(x) = p_{n-1}(x)(x-\xi) + a_0^1.$$

Beweis.

$$\begin{aligned} p_n(x) &= \sum_{j=0}^{n} a_j^0 x^j = \sum_{j=0}^{n} (a_j^1 - a_{j+1}^1 \xi) x^j = \sum_{j=0}^{n} a_j^1 x^j - \xi \sum_{j=0}^{n} a_{j+1}^1 x^j \\ &= a_0^1 + \sum_{j=1}^{n} a_j^1 x^j - \xi \sum_{j=1}^{n+1} a_j^1 x^{j-1} \qquad (a_{n+1}^1 := 0) \\ &= a_0^1 + x \cdot p_{n-1}(x) - \xi \cdot p_{n-1}(x) \\ &= p_{n-1}(x) \cdot (x-\xi) + a_0^1 \end{aligned}$$

□

1.1.2 Vollständiges Hornerschema

Zur Verallgemeinerung des obigen Verfahrens definieren wir für $k = 0, \ldots, n$ und $j = k, \ldots, n$

$$a_j^{k+1} = a_j^k + a_{j+1}^{k+1} \xi.$$

Für das Polynom $p_{n-k}(x)$, definiert durch

$$p_{n-k}(x) = \sum_{j=k}^{n} a_j^k x^{j-k},$$

gilt analog zu oben

$$p_{n-k}(x) = p_{n-(k+1)}(x)(x-\xi) + a_k^{k+1}.$$

Proposition 1.5 (Taylorpolynom um den Entwicklungspunkt ξ)
Für ein Polynom $p(x)$ gilt

$$p(x) = \sum_{k=0}^{n} a_k^{k+1} (x-\xi)^k$$
$$p^{(k)}(\xi) = k! \cdot a_k^{k+1}.$$

Beweis. Folgendes Schema kann induktiv formalisiert werden:

$$\begin{aligned}
p(x) = p_n(x) &= p_{n-1}(x)(x-\xi) + a_0^1 \\
&= \left[p_{n-2}(x)(x-\xi) + a_1^2\right](x-\xi) + a_0^1 \\
&= p_{n-2}(x)(x-\xi)^2 + a_1^2(x-\xi) + a_0^1 \\
&= \left[p_{n-3}(x)(x-\xi) + a_2^3\right](x-\xi)^2 + a_1^2(x-\xi) + a_0^1 \\
&= p_{n-3}(x)(x-\xi)^3 + a_2^3(x-\xi)^2 + a_1^2(x-\xi) + a_0^1 \\
&= \dots \\
&= \underbrace{p_0(x)}_{=a_n^{n+1}}(x-\xi)^n + a_{n-1}^n(x-\xi)^{n-1} + \dots + a_1^2(x-\xi) + a_0^1
\end{aligned}$$

Da die Taylorreihe eines Polynoms endlich und exakt ist, folgt mittels Koeffizientenvergleiches zwischen Taylorpolynom und Polynom:

$$\begin{aligned}
&\sum_{k=0}^{n} \frac{p^{(k)}(\xi)}{k!}(x-\xi)^k = T_\xi p(x) = p(x) = \sum_{k=0}^{n} a_k^{k+1}(x-\xi)^k \\
\implies &\frac{p^{(k)}(\xi)}{k!} = a_k^{k+1} \\
\implies &p^{(k)}(\xi) = k! \cdot a_k^{k+1}
\end{aligned}$$

□

Bemerkung 1.6 (Händiges Berechnen des Hornerschemas)
Zum händigen Auswerten des Hornerschemas ist folgendes Schema praktisch:

	a_n^0	a_{n-1}^0	a_{n-2}^0	$\dots$	a_2^0	a_1^0	a_0^0
ξ	a_n^1	a_{n-1}^1	a_{n-2}^1	$\dots$	a_2^1	a_1^1	a_0^1
ξ	a_n^2	a_{n-1}^2	a_{n-2}^2	$\dots$	a_2^2	a_1^2	
$\vdots$	$\vdots$	$\vdots$	$\vdots$				
ξ	a_n^{n-1}	a_{n-1}^{n-1}	a_{n-2}^{n-1}				
ξ	a_n^n	a_{n-1}^n					
ξ	a_n^{n+1}						

Die Rekursionsformel $a_j^{k+1} = a_j^k + a_{j+1}^{k+1}\xi$ bedeutet nun, dass ein Element die Summe aus dem darüberstehenden Element und dem links danebenstehenden Element multipliziert mit ξ ist.

Beispiel 1.7 (Hornerschema für ein Polynom dritten Gerades)
Bestimme für $p(x) = x^3 - 5x + 1$ die Funktionswerte von p und dessen Ableitungen an der Stelle $\xi = 1$ sowie das Taylorpolynom um diese Stelle.

ξ	1	0	-5	1	
1	1	1	-4	-3	$\implies p(1) = -3$
1	1	2	-2		$\implies p'(1) = -2$
1	1	3			$\implies p''(1) = (2!) \cdot 3 = 6$
1	1				$\implies p'''(1) = (3!) \cdot 1 = 6$

Somit liest man für das Taylorpolynom um den Entwicklungspunkt $\xi = 1$ ab:

$$p(x) = -3 - 2(x-1) + 3(x-1)^2 + (x-1)^3$$

1.2 Bairstowschema

Das Hornerschema kann auch naiv auf komplexe Zahlen angewandt werden. Dabei muss man allerdings mit komplexen Zahlen rechnen, was aufwendig ist.

Beispiel 1.8 (Hornerschema mit komplexen Zahlen)
Bestimme für $p(x) = x^3 - 3x^2 + x - 2$ den Wert $p(z)$ mit $z = 2 + 3i$.

z	1	-3	1	-2
$2+3i$	1	$-1+3i$	$-10+3i$	$-31-24i$

Dafür liefert das Bairstowschema eine Vereinfachung. Sei $z = a + bi \in \mathbb{C}$ und p ein Polynom mit reellen Koeffizienten a_j. Gesucht ist $p(z)$. Dann ist

$$(x-z)(x-\bar{z}) = x^2 - x(z+\bar{z}) + |z|^2 = x^2 - 2ax + (a^2+b^2)$$

ein reelles quadratisches Polynom. Somit gibt es ein Polynom q und $c_0, c_1 \in \mathbb{R}$ mit

$$p(x) = q(x)(x-z)(x-\bar{z}) + c_1 x + c_0.$$

Insbesondere folgt

$$p(z) = c_1 z + c_0.$$

Proposition 1.9 (Bairstowschema)
Sei

$$p(x) = \sum_{k=0}^{n} a_k^0 x^k$$

ein Polynom mit reellen Koeffizienten. Seien $\xi, \eta \in \mathbb{R}$ fest. Gesucht ist ein Polynom

$$q(x) = \sum_{k=2}^{n} a_k^1 x^{k-2}$$

und reelle Zahlen a_1^1, a_0^1, so dass gilt:

$$p(x) = q(x)(x^2 - \xi x - \eta) + a_1^1 x + a_0^1$$

Diese Koeffizienten sind zu berechnen mittels folgender Rekursion:

$$\begin{aligned}
a_n^1 &= a_n^0 \\
a_{n-1}^1 &= a_{n-1}^0 + \xi a_n^1 \\
a_k^1 &= a_k^0 + \xi a_{k+1}^1 + \eta a_{k+2}^1 \quad (\textit{für } k = n-2, \ldots, 1) \\
a_0^1 &= a_0^0 + \eta a_2^1
\end{aligned}$$

Beweis. Durch Ausmultiplizieren ergibt sich

$$\begin{aligned}
&q(x)(x^2-\xi x-\eta)+a_1^1x+a_0^1\\
&=\sum_{k=2}^{n}a_k^1x^k-\sum_{k=2}^{n}\xi a_k^1x^{k-1}-\sum_{k=2}^{n}\eta a_k^1x^{k-2}+a_1^1x+a_0^1\\
&=\sum_{k=2}^{n}a_k^1x^k-\sum_{k=1}^{n-1}\xi a_{k+1}^1x^k-\sum_{k=0}^{n-2}\eta a_{k+2}^1x^k+a_1^1x+a_0^1\\
&=\sum_{k=2}^{n-2}(a_k^1-\xi a_{k+1}^1-\eta a_{k+2}^1)x^k+a_{n-1}^1x^{n-1}+a_n^1x^n-\xi a_2^1x-\xi a_n^1x^{n-1}-\eta a_2^1-\eta a_3^1x+a_1^1x+a_0^1\\
&=(a_0^1-\eta a_2^1)+(a_1^1-\xi a_2^1-\eta a_3^1)x+\sum_{k=2}^{n-2}(a_k^1-\xi a_{k+1}^1-\eta a_{k+2}^1)x^k+(a_{n-1}^1-\xi a_n^1)x^{n-1}+a_n^1x^n.
\end{aligned}$$

Aus Koeffizientenvergleich mit $p(x)$ folgt die Behauptung. □

Bemerkung 1.10 (Händiges Berechnen des Bairstowschemas)
Zum händigen Auswerten des Bairstowschemas ist folgendes Schema praktisch:

		a_n^0	a_{n-1}^0	a_{n-2}^0	$\dots$	a_2^0	a_1^0	a_0^0
η	ξ	a_n^1	a_{n-1}^1	a_{n-2}^1	$\dots$	a_2^1	a_1^1	a_0^1

Dabei entpricht die formale Rekursion $a_k^1=a_k^0+\xi a_{k+1}^1+\eta a_{k+2}^1$, $a_0^1=a_0^0+\eta a_2^1$ der Faustregel: Jedes Element ist die Summe aus dem darüberstehenden Element, ξ mal dem linken Nachbarn und η mal dem zweiten Nachbarelement von links. Nur in der letzten Spalte wird das ξ mal dem linken Nachbarn weggelassen.

Beispiel 1.11 (Bairstowschema für ein Polynom dritten Gerades)
Sei $z=2+3i$. Bestimme für $p(x)=x^3-3x^2+x-2$ den Wert $p(z)$. Dann ist

$$\begin{aligned}
&(x^2-\xi x-\eta):=(x-z)(x-\bar{z})=(x-2-3i)(x-2+3i)=x^2-4x+13\\
\Longrightarrow\ &\xi=4\quad\text{und}\quad\eta=-13.
\end{aligned}$$

Das Bairstowschema ergibt:

η	ξ	1	-3	1	-2
-13	4	1	1	-8	-15

Somit ist

$$p(z)=-8z-15=-8(2+3i)-15=-31-24i.$$

1.3 Unendliche Reihen

Definition 1.12 (Reihe, Partialsummen, Konvergenz)
Sei a_k eine Folge. Dann wird die Folge der Partialsummen definiert durch

$$S_n=\sum_{k=1}^{n}a_k.$$

Man sagt, die Reihe konvergiert, wenn die Folge der Partialsummen S_n gegen eine Zahl S konvergiert. In diesem Fall schreibt man auch

$$\sum_{k=1}^{\infty} a_k := \lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} S_n = S.$$

Falls dieser Grenzwert nicht explizit berechnet werden kann, stellt sich die Frage nach einem Abbruchfehler, wenn nur die n-te Partialsumme berechnet wird. Gesucht ist somit eine obere Schranke für die Differenz $|S - S_n|$.

Proposition 1.13 (Leibnizkriterium)
Sei a_k eine monotone Nullfolge. Dann konvergiert die Reihe

$$S = \sum_{k=1}^{\infty} (-1)^k a_k$$

und es ergibt sich die Fehlerabschätzung

$$|S - S_n| \leq |a_{n+1}|.$$

Beweis. Sei o.B.d.A. a_k monoton fallend (ansonsten setzte $b_k := -a_k$). Es gilt

$$S_k - S_{k-2} = (-1)^k \underbrace{(a_k - a_{k-1})}_{\leq 0} \begin{cases} \leq 0 & \text{wenn } k \text{ gerade ist,} \\ \geq 0 & \text{wenn } k \text{ ungerade ist.} \end{cases}$$

Somit folgt

$$\begin{aligned} S_1 \leq S_3 \leq S_5 \leq \cdots \leq S_{2k+1} \leq \cdots \\ \cdots \leq S_{2k} \leq \cdots \leq S_6 \leq S_4 \leq S_2. \end{aligned}$$

Weiter folgt

$$S_{2k} - S_{2k-1} = \sum_{j=1}^{2k} (-1)^j a_j - \sum_{j=1}^{2k-1} (-1)^j a_j = (-1)^{2k} a_{2k} = a_{2k} > 0.$$

Also gilt für alle k die Ungleichungskette

$$S_{2k-3} \leq S_{2k-1} \leq S_{2k} \leq S_{2k-2}.$$

Somit sind die Intervalle $[S_{2k-1}, S_{2k}]$ ineinader geschachtelt, d.h.

$$[S_1, S_2] \supset [S_3, S_4] \supset [S_5, S_6] \supset \cdots \supset [S_{2k-1}, S_{2k}] \supset [S_{2k+1}, S_{2k+2}] \supset \cdots .$$

Damit ist die oben definierte Folge von Intervallen eine Intervallschachtelung. Ist S die durch die Intervallschachtelung eindeutig bestimmte Zahl in allen Intervallen, so gilt

$$\begin{aligned} &\lim_{k\to\infty} S_{2k-1} = S = \lim_{k\to\infty} S_{2k} \\ \Longrightarrow\ &\lim_{k\to\infty} S_k = S. \end{aligned}$$

Somit konvergiert die Folge der Partialsummen, also die Reihe.

Durch

$$\begin{aligned} & S_{2k-1} \leq S_{2k+1} \leq S \leq S_{2k+2} \leq S_{2k} \\ \Longrightarrow & |S - S_{2k}| \leq |S_{2k+1} - S_{2k}| = a_{2k+1} \\ \Longrightarrow & |S - S_{2k-1}| \leq |S_{2k-1} - S_{2k}| = a_{2k} \end{aligned}$$

ergibt sich als Zusammenfassung der geraden und ungeraden Fälle

$$|S - S_n| \leq a_{n+1}.$$

□

Bemerkung 1.14 (Äquivalente Formulierung)
Eine äquivalente Formulierung des Leibnizkriteriums ist: Sei a_k eine betraglich monoton fallende alternierende Nullfolge, d.h. eine Folge mit den Eigenschaften

$$\begin{aligned} a_k a_{k+1} &\leq 0 \\ |a_k| &\geq |a_{k+1}|. \end{aligned}$$

Dann konvergiert die Reihe $\sum a_k$.

Beispiel 1.15 (Leibnizreihe)
Da $\frac{1}{k}$ eine monotone Nullfolge ist, konvergiert die Reihe

$$\sum_{k=1}^{\infty} \frac{(-1)^k}{k}.$$

Beispiel 1.16 (Leibnizreihe)
Der allgemeine Binomialkoeffizient ist definiert durch

$$\binom{y}{k} = \frac{y(y-1)\cdots(y-(k-1))}{k!}.$$

Sei $0 < w < 1$. Die Frage ist, ob die folgende Reihe konvergiert:

$$\sum_{k=0}^{\infty} a_k := \sum_{k=0}^{\infty} w^k \binom{y}{k}$$

Man sieht, dass für $k > y + 1$ das Vorzeichen von a_k alterniert, also $a_k a_{k+1} < 0$ ist. Weiter ist

$$\begin{aligned} \frac{|a_k|}{|a_{k+1}|} &= \left| \frac{w^k \binom{y}{k}}{w^{k+1} \binom{y}{k+1}} \right| = \left| \frac{w^k \frac{y(y-1)\cdots(y-(k-1))}{k!}}{w \cdot w^k \frac{y(y-1)\cdots(y-(k-1))(y-k)}{k!(k+1)}} \right| \\ &= \frac{k+1}{w|y-k|} = \frac{1}{w} \cdot \underbrace{\frac{k+1}{k-y}}_{\to 1 \ (k\to\infty)} > 1 \quad \text{(für } k \text{ geeignet groß)}. \end{aligned}$$

Somit konvergiert die Reihe nach dem Leibnizkriterium. Beachte, dass gilt:

$$(1+w)^y = \sum_{k=0}^{\infty} w^k \binom{y}{k}$$

Proposition 1.17 (Quotientenkriterium)
Sei a_k eine Folge und es gebe $N \in \mathbb{N}$ und $q < 1$ mit

$$\forall k \geq N : |a_k| \neq 0 \text{ und } \frac{|a_{k+1}|}{|a_k|} \leq q.$$

Dann konvergiert die Reihe $\sum a_k = S$ absolut mit der Fehlerabschätzung

$$\forall n \geq N : |S - S_n| \leq \frac{|a_{n+1}|}{1-q}.$$

Beweis. Die absolute Konvergenz folgt mit

$$|a_{N+k}| \leq q \cdot |a_{N+(k-1)}| \leq q^2 \cdot |a_{N+(k-2)}| \leq \cdots \leq q^k \cdot |a_N|$$

aus dem Majorantenkriterium und der Konvergenz der geometrischen Reihe, denn

$$\sum_{k=N}^{\infty} q^k \cdot |a_N| = |a_N| \sum_{k=N}^{\infty} q^k.$$

Die Fehlerabschätzung ergibt sich damit für $n \geq N$ aus

$$\begin{aligned} |S - S_n| &= \left| \sum_{k=n+1}^{\infty} a_k \right| \leq \sum_{k=n+1}^{\infty} |a_k| \leq \sum_{k=n+1}^{\infty} q^{k-n+1} |a_{n+1}| \\ &= |a_{n+1}| \sum_{k=n+1}^{\infty} q^{k-n+1} = |a_{n+1}| \sum_{k=0}^{\infty} q^k = |a_{n+1}| \frac{1}{1-q}. \end{aligned}$$

□

2 Nullstellenapproximation

Definition 2.1 (Funktionenräume)
Im Folgenden verwenden wir die Bezeichnungen

- $C[a,b]$ *ist der Raum der stetigen Funktionen auf* $[a,b]$.
- $C^n[a,b]$ *ist der Raum der n-mal stetig differenzierbaren Funktionen auf* $[a,b]$.
- $\mathrm{Lip}[a,b]$ *ist der Raum der lipschitzstetigen Funktionen auf* $[a,b]$.

2.1 Intervallschachtelung

Sei nun $f \in C[a,b]$ und $a', b' \in [a,b]$ mit $a' < b'$. Wenn $f(a')$ und $f(b')$ unterschiedliche Vorzeichen haben (d.h. $f(a')f(b') < 0$), dann besitzt f nach dem Zwischenwertsatz im Intervall (a,b) mindestens eine Nullstelle. Das folgende Verfahren dient nun zur beliebig genauen Annäherung dieser Nullstelle.

Proposition 2.2 (Intervallschachtelung der Nullstelle)
Sei $f \in C[a,b]$ *mit* $f(a)f(b) < 0$. *Definiere eine Folge von Intervallen* $[a_n, b_n]$ *durch*

$$[a_0, b_0] = [a,b],$$

$$[a_{n+1}, b_{n+1}] = \begin{cases} [a_n, c_n] & \text{wenn } f(a_n)f(c_n) < 0, \\ [c_n, b_n] & \text{wenn } f(c_n)f(b_n) < 0, \\ [c_n, c_n] & \text{wenn } f(c_n) = 0 \end{cases}$$

mit $c_n = \frac{1}{2}(a_n + b_n)$ *als Mittelpunkt des Intervalls* $[a_n, b_n]$.

Dann gilt $c_n \to \xi$ $(n \to \infty)$ *für eine Nullstelle* ξ *von* f *und*

$$|c_n - \xi| \leq \frac{b-a}{2^{n+1}}.$$

Beweis. Sollte für ein $n \in \mathbb{N}$ der Fall $f(c_n) = 0$ auftreten, so ist die Nullstelle exakt gefunden und damit die Rekursion für praktische Zwecke beendet. Dieser Fall soll von nun an nicht mehr gesondert betrachtet werden.

Für alle $n \in \mathbb{N}$ gilt nach Konstruktion:

$$a \leq a_n \leq a_{n+1} \leq b_{n+1} \leq b_n \leq b$$
$$b_n - a_n \leq \frac{b-a}{2^n}$$

Somit ist a_n eine monoton wachsende, nach oben beschränkte Folge, also konvergent. Analog konvergiert b_n als nach unten beschränkte, monoton fallende Folge.

Dann gilt

$$\xi := \lim_{n\to\infty} b_n = \lim_{n\to\infty} (a_n + (b_n - a_n)) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} (b_n - a_n) = \lim_{n\to\infty} a_n.$$

Wegen $a_n \leq c_n \leq b_n$ für alle n folgt $\lim_{n\to\infty} c_n = \xi$.

Aus der Stetigkeit von f folgt

$$\begin{aligned} &\forall n \in \mathbb{N} : f(a_n)f(b_n) \leq 0 \\ \Longrightarrow\ &f(\xi)^2 = \lim_{n\to\infty} f(a_n)f(b_n) \leq 0 \\ \Longrightarrow\ &f(\xi) = 0. \end{aligned}$$

Für die Fehlerabschätzung gilt $\forall n : a_n \leq \xi \leq b_n$ und

$$c_n - a_n = b_n - c_n = \frac{b-a}{2^{n+1}}.$$

Aus $\xi \in [a_n, c_n]$ oder $\xi \in [c_n, b_n]$ folgt damit die Behauptung

$$|c_n - \xi| \leq \frac{b-a}{2^{n+1}}.$$

□

Proposition 2.3 (Fehlerabschätzung für stetig differenzierbare Funktionen)
Sei $f \in C^1[a,b]$ mit $f(a)f(b) < 0$ und

$$\forall x \in [a,b] : 0 < m \leq f'(x) \leq M.$$

Für die Nullstelle ξ von f und die in Proposition 2.2 definierte Folge gilt dann die Fehlerabschätzung

$$\frac{|f(c_n)|}{M} \leq |c_n - \xi| \leq \frac{|f(c_n)|}{m}.$$

Beweis. Nach dem Mittelwertsatz existiert ein $\eta \in [a,b]$ mit

$$|f(c_n)| = |f(c_n) - f(\xi)| = |f'(\eta)||c_n - \xi|.$$

Somit folgt

$$\begin{aligned} &m|c_n - \xi| \leq |f(c_n)| \ \wedge\ |f(c_n)| \leq M|c_n - \xi| \\ \Longrightarrow\ &|c_n - \xi| \leq \frac{|f(c_n)|}{m} \ \wedge\ \frac{|f(c_n)|}{M} \leq |c_n - \xi| \\ \Longrightarrow\ &\frac{|f(c_n)|}{M} \leq |c_n - \xi| \leq \frac{|f(c_n)|}{m}. \end{aligned}$$

□

Beispiel 2.4 (Berechnung der Umkehrfunktion)
Sei $f \in C^1[a,b]$ mit $f' \geq m > 0$ auf $[a,b]$. Sei $f(a) = c$, $f(b) = d$. Wegen $f' > 0$ ist f streng monoton steigend, also injektiv. Dann gilt nach dem Zwischenwertsatz

$$\forall \eta \in [c,d] \exists! \xi : f(\xi) = \eta.$$

Somit existiert die Umkehrfunktion, sie ist definiert durch $f^{-1}(\eta) = \xi$.

Zur Bestimmung von $f^{-1}(\eta)$ für $\eta \in (c,d)$ setze

$$F(x) = f(x) - \eta.$$

Dann ist $F \in C^1[a,b]$, $F' = f' \geq m$ und $F(a)F(b) < 0$. Bestimme nach Proposition 2.3 ein ξ mit $F(\xi) = 0$. Dann gilt

$$f(\xi) = F(\xi) + \eta = \eta.$$

Proposition 2.5 (Fehlerabschätzung für Bestimmung der Umkehrfunktion)
Sei f wie in oberem Beispiel, gesucht ist $\xi = f^{-1}(\eta)$. Dann gilt für die zur Bestimmung der Nullstelle von F oben definierten Folge c_n

$$|c_n - \xi| \leq \frac{|f(c_n) - \eta|}{m}.$$

Beweis. Nach Proposition 2.3 für F gilt wegen $F' = f'$:

$$|c_n - \xi| \leq \frac{|F(c_n)|}{m} = \frac{|f(c_n) - \eta|}{m}$$

□

2.2 Sukzessive Approximation

Die Methode der sukzessiven Approximation ist ein wichtiges iteratives Verfahren zur Lösung von Gleichungen. Hier dient sie zur Bestimmung von Fixpunkten und Nullstellen reell- oder komplexwertiger Funktionen einer Veränderlichen. Mit der Bestimmung von Näherungslösungen sowie den a-priori- und a-posteriori- Fehlerabschätzungen kann man dabei gleichzeitig die Existenz und Eindeutigkeit von Nullstellen und Fixpunkten zeigen.

Definition 2.6 (Mehrdimensionales Intervall)
Sei $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$. Dann ist das mehrdimensionale Intervall $[a, b] \subset \mathbb{R}^n$ definiert als

$$[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n].$$

Definition 2.7 (Maximumnorm)
Auf $\mathbb{R}^n$ ist die Maximumnorm $\|\cdot\|_\infty$ definiert durch

$$\|(a_1, \ldots, a_n)\|_\infty := \max_{1 \leq i \leq n} |a_i|.$$

Im eindimensionalen Fall ist die Maximumnorm also gerade der gewöhnliche Betrag.

Bemerkung 2.8 (Funktionenräume)
Sei $[a, b] \subset \mathbb{R}^n$. Mit $C[a, b]$ bzw. $C^1[a, b]$ ist in diesem Kapitel der Raum der stetigen bzw. stetig differenzierbaren Funktionen $f : [a, b] \mapsto \mathbb{R}^n$ gemeint.

Definition 2.9 (Lipschitzstetig)
Eine Funktion $f : [a, b] \mapsto \mathbb{R}$ heißt lipschitzstetig, wenn gilt:

$$\exists q \in \mathbb{R} : \|f(x) - f(y)\|_\infty \leq q \|x - y\|_\infty$$

Falls $q < 1$ ist, so nennt man f eine Kontraktion.

Mit $\mathrm{Lip}[a, b] \subset C[a, b]$ wird der Raum der lipschitzstetigen Funktionen über $[a, b]$ bezeichnet.

Proposition 2.10 (Banachscher Fixpunktsatz)
Sei $g : [a, b] \mapsto [a, b]$ eine Kontraktion.

- *Dann gibt es genau ein $\xi \in [a, b]$ mit $g(\xi) = \xi$.*
- *Für beliebiges x_0 in $[a, b]$ konvergiert $x_{n+1} = g(x_n)$ gegen diesen Fixpunkt ξ.*

Als Fehlerabschätzung für die Folge x_n gilt:

$$\|x_n - \xi\|_\infty \leq \frac{q^n}{1-q} \|x_1 - x_0\|_\infty \qquad \text{(a priori)}$$

$$\|x_n - \xi\|_\infty \leq \frac{q}{1-q} \|x_n - x_{n-1}\|_\infty \qquad \text{(a posteriori)}$$

Beweis. Existenz des Fixpunkts und Konvergenz der Folge:

Für beliebiges $x_0 \in [a,b]$ definiere die Folge x_n in $[a,b]$ durch $x_{n+1} = g(x_n)$. Wir zeigen: x_n ist eine Cauchyfolge. Dazu zeigen wir

$$\|x_{n+k} - x_{n+k-1}\|_\infty \leq q^k \|x_n - x_{n-1}\|_\infty \tag{1}$$

$$\|x_{n+k} - x_n\|_\infty \leq \frac{q}{1-q} \|x_n - x_{n-1}\|_\infty \tag{2}$$

$$\|x_{n+k} - x_n\|_\infty \leq \frac{q^n}{1-q} \|x_1 - x_0\|_\infty . \tag{3}$$

Nun gilt mittels vollständiger Induktion (1), denn

$$\begin{aligned} k = 1: \quad & \|x_{n+1} - x_n\|_\infty = \|g(x_n) - g(x_{n-1})\|_\infty \leq q \|x_n - x_{n-1}\|_\infty \\ k \to k+1: \quad & \|x_{n+k+1} - x_{n+k}\|_\infty = \|g(x_{n+k}) - g(x_{n+k-1})\|_\infty \leq q \|x_{n+k} - x_{n+k-1}\|_\infty \\ & \overset{IV}{\leq} q \cdot q^k \|x_n - x_{n-1}\|_\infty = q^{k+1} \|x_n - x_{n-1}\|_\infty . \end{aligned}$$

Ungleichung (2) folgt unter der Verwendung des Teleskopprinzips aus

$$\begin{aligned} \|x_{n+k} - x_n\|_\infty = \left\| \sum_{j=1}^{k} (x_{n+j} - x_{n+j-1}) \right\|_\infty &\leq \sum_{j=1}^{k} \|x_{n+j} - x_{n+j-1}\|_\infty \leq \sum_{j=1}^{k} q^j \|x_n - x_{n-1}\|_\infty \\ &\leq \|x_n - x_{n-1}\|_\infty \sum_{j=1}^{\infty} q^j = \frac{q}{1-q} \|x_n - x_{n-1}\|_\infty . \end{aligned}$$

Ungleichung (3) gilt nach

$$\|x_{n+k} - x_n\|_\infty \overset{(2)}{\leq} \frac{q}{1-q} \|x_n - x_{n-1}\|_\infty \overset{(1)}{\leq} \frac{q}{1-q} q^{n-1} \|x_1 - x_0\|_\infty = \frac{q^n}{1-q} \|x_1 - x_0\|_\infty .$$

Somit ist x_n eine Cauchyfolge, denn nach (3) gilt $\forall \varepsilon > 0 \exists N : \forall n > N \forall k > 0 : \|x_{n+k} - x_n\|_\infty < \varepsilon$.

Da $\mathbb{R}^n$ vollständig und $[a,b]$ abgeschlossen ist, konvergiert x_n in $[a,b]$, d.h. es gilt

$$\xi := \lim_{n \to \infty} x_n \in [a,b].$$

Somit ist wegen der Stetigkeit von g

$$g(\xi) = \lim_{n \to \infty} g(x_n) = \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} x_n = \xi.$$

Eindeutigkeit des Fixpunktes:

Seien $\xi, \eta \in [a,b]$ mit $g(\xi) = \xi$ und $g(\eta) = \eta$. Dann ist

$$\begin{aligned} & \|\xi - \eta\|_\infty = \|g(\xi) - g(\eta)\|_\infty \leq q \|\xi - \eta\|_\infty \\ \Longrightarrow\ & \|\xi - \eta\|_\infty = 0 \\ \Longrightarrow\ & \xi = \eta. \end{aligned}$$

Zu den Fehlerabschätzungen:

Mit $k \to \infty$ in (3) folgt

$$\|x_n - \xi\|_\infty \leq \frac{q^n}{1-q} \|x_1 - x_0\|_\infty .$$

Mit $k \to \infty$ in (2) folgt

$$\|x_n - \xi\|_\infty \leq \frac{q}{1-q} \|x_n - x_{n-1}\|_\infty .$$

□

Proposition 2.11 (Kriterium zur Überprüfung auf Selbstabbildung)
Sei $g \in C[a,b]$ eine Kontraktion mit Lipschitzkonstante q. Wenn

$$\left\| g\left(\frac{a+b}{2}\right) - \frac{a+b}{2} \right\|_\infty \leq (1-q)\frac{b_i - a_i}{2} \quad (\textit{für } i = 1, \ldots, n),$$

dann gilt $g([a,b]) \subset [a,b]$.

Insbesondere folgt für $[a,b] \subset \mathbb{R}$, dass $g([a,b]) \subset [a,b]$, falls gilt:

$$\left| g\left(\frac{a+b}{2}\right) - \frac{a+b}{2} \right| \leq (1-q)\frac{b-a}{2}$$

Beweis. Zuerst der eindimensionale Fall: Sei $x \in [a,b] \subset \mathbb{R}$ beliebig. Dann ist

$$\begin{aligned} \left| g(x) - \frac{a+b}{2} \right| &\leq \left| g(x) - g\left(\frac{a+b}{2}\right) \right| + \left| g\left(\frac{a+b}{2}\right) - \frac{a+b}{2} \right| \\ &\leq q \left| x - \frac{a+b}{2} \right| + (1-q)\frac{b-a}{2} \\ &\leq q\frac{b-a}{2} + (1-q)\frac{b-a}{2} = \frac{b-a}{2}. \end{aligned}$$

Somit folgt $g(x) \in [a,b]$.

Sei nun $x \in [a,b] \subset \mathbb{R}^n$. Dann folgt für jede Komponentenfunktion g_i:

$$\left| g_i\left(\frac{a+b}{2}\right) - \frac{a_i + b_i}{2} \right| \leq \left\| g\left(\frac{a+b}{2}\right) - \frac{a+b}{2} \right\|_\infty \leq (1-q)\frac{b_i - a_i}{2} \Rightarrow \quad \left| g_i(x) - \frac{a_i + b_i}{2} \right| \leq \frac{b_i - a_i}{2}$$

und somit die Behauptung. □

Proposition 2.12 (Stetig differenzierbare Funktionen sind lipschitzstetig auf kompakten Intervallen) *Sei $g \in C^1[a,b]$. Dann ist g lipschitzstetig mit Lipschitzkonstante q,*

$$q := \max_{1\le i\le n} \sum_{k=1}^{n} \left\| \frac{\partial g_i}{\partial x_k} \right\|_\infty .$$

Insbesondere gilt für den eindimensionalen Fall $[a,b] \subset \mathbb{R}$

$$q = \max_{x\in[a,b]} |g'(x)|.$$

Beweis. Der eindimensionale Fall ist ein Spezialfall des mehrdimensionalen Falls. Aufgrund des einfachen Beweises soll er jedoch getrennt bewiesen werden.

Nach dem reellen Mittelwertsatz gibt es ein $z \in [a,b]$ so, dass

$$|g(x) - g(y)| = |g'(z)(x-y)| = |g'(z)||x-y| \le q|x-y|.$$

Mehrdimensionaler Fall: Seien zwei Punkte $x, y \in [a,b]$. Um den Mittelwertsatz anwenden zu können, betrachten wir die (reellwertigen) Komponentenfunktionen auf der Verbindungsstrecke zwischen a und b. Dazu definieren wir

$$\varphi_i : [0,1] \mapsto [a_i, b_i] \quad (\text{für } t \to g_i((1-t)x + ty)) .$$

Nach dem Zwischenwertsatz gibt es ein $\tau_i \in [0,1]$ mit

$$g_i(y) - g_i(x) = \varphi_i(1) - \varphi_i(0) = \varphi_i'(\tau_i).$$

Somit gibt es τ_i für $i = 1, \dots, n$, so dass

$$\begin{aligned} |g_i(y) - g_i(x)| = |\varphi_i'(\tau_i)| &= \left| \sum_{k=1}^{n} \frac{\partial g_i}{\partial x_k}((1-\tau_i)x + \tau_i y)(y_k - x_k) \right| \\ &\le \sum_{k=1}^{n} \left| \frac{\partial g_i}{\partial x_k}((1-\tau_i)x + \tau_i y) \right| |(y_k - x_k)| \\ &\le \|y-x\|_\infty \sum_{k=1}^{n} \left\| \frac{\partial g_i}{\partial x_k} \right\|_\infty \le q \|y-x\|_\infty . \end{aligned}$$

Aus der Definition der Maximumnorm folgt

$$\|g(y) - g(x)\|_\infty = \max_{i=1,\dots,n} |g_i(y) - g_i(x)| \le q \|y-x\|_\infty .$$

□

2.3 Nullstellenbestimmung durch sukzessive Approximation

Sei $f \in C^1[a,b]$. Gesucht sind die Nullstellen von f.

Bemerkung 2.13 (Verallgemeinerung auf mehrdimensionale Funktionen)
Wir betrachten im folgenden Abschnitt nur reellwertige Funktionen. Alle Ideen und Beweise lassen sich jedoch relativ leicht auf den mehrdimensionalen Fall verallgemeinern.

Definition 2.14 (Iterationsfunktion und Iterationsfolge)
Sei $f \in C^1[a,b]$ und $\omega > 0$. Dann definiert man die Iterationsfunktion *g zu f und ω durch*

$$g(x) = x - \frac{f(x)}{\omega}.$$

Zu einem Startwert x_0 definiert man die Iterationsfolge

$$x_{k+1} = x_k - \frac{f(x_k)}{\omega}.$$

Bemerkung 2.15 (Konvergenz der Iterationsfolge)
Mit f ist auch $g \in C^1[a,b]$ und die Fixpunkte von g sind genau die Nullstellen von f, d.h.

$$f(z) = 0 \iff g(z) = z.$$

Die Iterationsfolge konvergiert also genau dann gegen eine Nullstelle von f, wenn sie gegen einen Fixpunkt von g konvergiert. Kriterien für diese Konvergenz folgen aus dem vorherigen Kapitel. Es genügt, dass g eine kontrahierende Selbstabbildung ist.

Insbesondere hofft man, den Startwert x_0 und ω geeignet wählen zu können, so dass diese Bedingungen in einem gewissen Intervall erfüllt sind.

Proposition 2.16 (Kriterium für Kontraktion der Iterationsfunktion)
Sei $f \in C^1[a,b]$ und $\omega > 0$. Es existiere ein $q < 1$ mit

$$\sup_{x\in[a,b]} \left|1 - \frac{f'(x)}{\omega}\right| \leq q < 1.$$

Dann ist die Iterationsfunktion g zu f und ω eine Kontraktion.

Beweis. Aus $g(x) = x - \frac{f(x)}{\omega}$ folgt durch Differentiation

$$|g'(x)| = \left|1 - \frac{f'(x)}{\omega}\right| \leq q < 1,$$

also dass g eine Kontraktion ist. □

Proposition 2.17 (Iterationsfunktion ist für kleine Intervalle eine Kontraktion)
Es sei $f \in C^1[a,b]$. Sei $c \in [a,b]$ beliebig mit $\omega := f'(c) \neq 0$. Dann gibt es ein $q < 1$ und ein hinreichend kleines Intervall $I = [c-\delta, c+\delta]$ so, dass die Iterationsfunktion g bezüglich f und ω auf I eine Kontraktion mit der Lipschitzkonstanten q ist.

Beweis. Wegen der Stetigkeit von f' in c gibt es ein $\delta > 0$ so, dass für $x \in I := [c-\delta, c+\delta]$ gilt: $|f'(c) - f'(x)| < q \cdot |\omega|$. Dann gilt für $x \in I$

$$\left|1 - \frac{f'(x)}{\omega}\right| = \frac{|f'(c) - f'(x)|}{|\omega|} \leq q < 1$$
$$\implies \sup_{x\in I} \left|1 - \frac{f'(x)}{\omega}\right| \leq q < 1.$$

Nach Propositopn 2.16 ist g eine Kontraktion. □

Proposition 2.18 (Kriterium auf Selbstabbildung der Iterationsfunktion)
Gilt zusätzlich zu den in Proposition 2.17 geforderten Bedingungen noch

$$|f(c)| \leq (1-q)|w|\frac{|b-a|}{2},$$

so ist $g(I) \subset I$. Insbesondere ist g dann auf I eine kontrahierende Selbstabbildung, hat also genau einen Fixpunkt z.

Beweis. Dann ist

$$|g(c)-c| = \left|c - \frac{f(c)}{\omega} - c\right| = \frac{|f(c)|}{|\omega|} \leq (1-q)\frac{|b-a|}{2}$$

und somit g nach Proposition 2.11 eine Selbstabbildung. Darum ist g eine kontrahierende Selbstabbildung, hat nach dem Banachschen Fixpunktsatz also genau einen Fixpunkt. □

Proposition 2.19 (Vereinfachtes Newtonverfahren)
Sei $f \in C^1[a,b]$ und $c \in [a,b]$ beliebig mit

$$\omega := f'(c) \neq 0$$
$$|f(c)| \leq (1-q)|\omega|\frac{|b-a|}{2}.$$

Dann hat f eine Nullstelle in einer Umgebung von c und die Folge

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(c)}$$

konvergiert für Startwerte hinreichend dicht bei c gegen diese Nullstelle von f.

Beweis. Nach Proposition 2.17 ist die Iterationsfunktion

$$g(x) = x - \frac{f(x)}{f'(c)}$$

auf dem kleinen Intervall $I := [c-\delta, c+\delta]$ eine Kontrakion, nach Proposition 2.18 dort eine Selbstabbildung. Somit hat g dort genau einen Fixpunkt, also f genau eine Nullstelle. Nach dem Banachschen Fixpunktsatz konvergiert x_k gegen diesen Fixpunkt von g, also gegen die Nullstelle von f. □

Proposition 2.20 (Fehlerabschätzung für das vereinfachte Newtonverfahren)
Unter den Bedingungen von Proposition 2.19 ist

$$q := \sup_{x \in I}\left|1 - \frac{f'(x)}{f'(c)}\right| < 1.$$

Ist z die Nullstelle von f, so folgt für die a-priori Fehlerabschätzung

$$|x_k - z| \leq \frac{q^k}{1-q}\left|\frac{f(x_0)}{f'(c)}\right|.$$

Beweis. Aus dem Banachschen Fixpunktsatz folgt sofort

$$|x_k - z| \leq \frac{q^k}{1-q}|x_1 - x_0| = \frac{q^k}{1-q}\left|x_0 + \frac{f(x_0)}{f'(c)} - x_0\right| = \frac{q^k}{1-q}\left|\frac{f(x_0)}{f'(c)}\right|.$$

□

Proposition 2.21 (Stärkere Fehlerabschätzung für das vereinfachte Newtonverfahren)
Gilt zusätzlich noch $f'(x) \geq m > 0$ für $x \in I$, so folgen die Fehlerabschätzungen:

$$|x_k - z| \leq \frac{q^k}{m}|f(x_0)| \qquad \text{(a priori)}$$

$$|x_k - z| \leq \frac{|f(x_k)|}{m} \qquad \text{(a posteriori)}$$

Beweis. Zur a-posteriori-Abschätzung: Nach dem Mittelwertsatz gibt es ein y mit

$$\frac{|f(x_k)|}{|x_k - z|} = \frac{|f(x_k) - f(z)|}{|x_k - z|} = |f'(y)| \geq m.$$

Zur a-priori-Abschätzung: Aus der Rekursionsvorschrift folgt

$$\begin{aligned} & x_{k+1} = x_k - \frac{f(x_k)}{f'(c)} \\ \Longrightarrow\ & |f'(c)||x_{k+1} - x_k| = |f(x_k)| \\ \Longrightarrow\ & |f(x_k)| = |x_{k+1} - x_k||f'(c)| \leq q|x_k - x_{k-1}||f'(c)| \leq \cdots \leq q^k|x_1 - x_0| \cdot |f'(c)|. \end{aligned}$$

Weiter folgt

$$\begin{aligned} & x_1 = x_0 - \frac{f(x_0)}{f'(c)} \\ \Longrightarrow\ & |x_1 - x_0| = \left|\frac{f(x_0)}{f'(c)}\right| \\ \Longrightarrow\ & |f(x_k)| \leq q^k \left|\frac{f(x_0)}{f'(c)}\right| \cdot |f'(c)| = q^k|f(x_0)|. \end{aligned}$$

und somit gilt die a-priori-Abschätzung. □

2.4 Newton-Verfahren

Das Newton-Verfahren dient zur näherungsweisen Nullstellenbestimmung. Gegenüber der sukzessiven Approximation konvergiert das Newton-Verfahren wesentlich besser, allerdings müssen einige Voraussetzungen erfüllt sein.

Definition 2.22 (Iterationsfunktion)
Sei $f \in C^2([a,b])$. Die Bestimmung der Nullstelle $z \in [a,b]$ von f erfolgt unter Verwendung eines geeigneten Anfangswertes $x_0 \in [a,b]$ durch das Iterationsverfahren

$$x_{t+1} = g(x_t) = x_t - \frac{f(x_t)}{f'(x_t)},$$

wobei

$$g(x) := x - \frac{f(x)}{f'(x)}$$

die Iterationsfunktion ist.

Bemerkung 2.23 (Geometrische Interpretation)
Das Newton-Verfahren lässt sich geometrisch wie folgt deuten: Die Tangente an den Graphen von f im Punkt $(x_0, f(x_0))$ hat die Form $y = f'(x_0)(x - x_0) + f(x_0)$ mit der Nullstelle $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. Nehmen wir nun x_1 als neuen Ausgangspunkt dieser Betrachtung, so gelangen wir zur Iterationsvorschrift in Definition 2.22.

Im Folgenden wird die Konvergenz des Newtonverfahrens untersucht. Dazu sei festgelegt: Für die Funktion $f \in C^2([a,b])$ gelte $0 < m \leq |f'(x)| \leq M$ und $|f''(x)| \leq N$ für alle $x \in [a,b]$. Des weiteren habe die Funktion f an der Stelle $z \in [a,b]$ eine Nullstelle.

Proposition 2.24 (Lipschitzstetigkeit der Iterationsfunktion)
Die Iterationsfunktion $g(x)$ ist lipschitzstetig in $[z-r, z+r] \subseteq [a,b]$ mit der Lipschitzkonstanten $q_r = \frac{MN}{m^2} r$.

Beweis. Wir zeigen die zur Lipschitzstetigkeit notwendige und hinreichende Bedingung $|g'(x)| \leq q_r$ für alle $x \in [z-r, z+r]$:

$$|g'(x)| = \left| 1 - \frac{f'^2(x) - f(x)f''(x)}{f'^2(x)} \right| = \left| \frac{f(x)f''(x)}{f'^2(x)} \right| \leq \frac{|f(x)|N}{m^2}$$

Nun gilt nach dem Mittelwertsatz der Differentialrechnung, dass ein $\xi \in [z-r, z+r]$ existiert, mit

$$\frac{f(x) - f(z)}{x - z} = f'(\xi) \overset{f(z)=0}{\Longrightarrow} |f(x)| = \underbrace{|f'(\xi)|}_{\leq M} \cdot \underbrace{|x - z|}_{\leq r}.$$

Somit führt die obige Abschätzung auf

$$|g'(x)| \leq \frac{MN}{m^2} r =: q_r.$$

□

Proposition 2.25 (Selbstabbildung der Iterationsfunktion)
Ist $0 < r < \frac{2m}{N}$, dann gilt $g([z-r, z+r]) \subseteq [z-r, z+r]$, d.h. $g(x)$ ist eine Selbstabbildung.

Beweis. Wir müssen zeigen, dass für $x \in [z-r, z+r]$ gilt:

$$|x - z| \leq r \iff |g(x) - z| \leq r$$

Dazu verwenden wir die Taylor-Entwicklung von $f(z)$ um x in z:

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

Dabei ist $\frac{1}{2}f''(\zeta)(z-x)^2$ mit $\zeta \in [z-r, z+r]$ die Lagrange-Form des Restgliedes. Damit erhalten wir

$$\begin{aligned} |g(x)-z| &\overset{\text{Def.}}{=} \left| x - \frac{f(x)}{f'(x)} - z \right| \\ &= \frac{1}{|f'(x)|} \cdot |f(x) + f'(x)(z-x)| \\ &\overset{\text{Taylor}}{=} \frac{1}{|f'(x)|} \cdot \left| \frac{1}{2}(z-x)^2 f''(\zeta) \right| \\ &\leq \frac{1}{2}\frac{N}{m}r^2 \leq \frac{1}{2}\frac{N}{m}\frac{2m}{N}r = r \end{aligned}$$

□

Proposition 2.26 (Konvergenzsatz)
Es gelte $q_r = \frac{MN}{m^2}r < 1$ und $0 < r < \frac{2m}{N}$. Dann ist die Iterationsfunktion g nach Prop. 2.24 (Lipschitzstetigkeit) und 2.25 (Selbstabbildung) offenbar eine Kontraktion in $[z-r, z+r]$ und konvergiert somit bei beliebigem Anfangswert $x_0 \in [z-r, z+r]$ gegen z:

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} g(x_t) = z.$$

Proposition 2.27 (a-priori-Fehlerabschätzung)
Ist $\rho := \frac{N}{2m}|x_0 - z| < 1$, so gilt für das Newton'sche Iterationsverfahren die a-priori-Fehlerabschätzung

$$|x_k - z| \leq \frac{2m}{N}\rho^{2^k}.$$

Beweis. Wie im Beweis von Prop. 2.25 (Selbstabbildung) verwenden wir die Taylor-Entwicklung von f um z in x

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

mit $x, \zeta \in [z - |x_0 - z|, z + |x_0 - z|]$. Wir erhalten analog

$$\begin{aligned} |g(x)-z| &\overset{\text{Def.}}{=} \left| x - \frac{f(x)}{f'(x)} - z \right| \\ &= \frac{1}{|f'(x)|} \cdot |f(x) + f'(x)(z-x)| \\ &\overset{\text{Taylor}}{=} \frac{1}{|f'(x)|} \cdot \left| \frac{1}{2}(z-x)^2 f''(\zeta) \right| \\ &\leq \frac{N}{2m}(x-z)^2. \end{aligned}$$

Wir setzen nun $\rho_k := \frac{N}{2m}|x_k - z|$. Mit obiger Ungleichung gilt

$$\rho_k = \frac{N}{2m}|x_k - z| = \frac{N}{2m}|g(x_{k-1}) - z| \overset{\text{s.o.}}{\leq} \frac{N}{2m}\frac{N}{2m}(x_{k-1} - z)^2 = \rho_{k-1}^2.$$

Daraus folgt

$$\rho_k \leq \rho_{k-1}^2 \leq \rho_{k-2}^4 \leq \rho_{k-3}^8 \leq \cdots \leq \rho_0^{2^k} = \rho^{2^k}.$$

Schließlich ergibt sich daraus die Gleichung für die a-priori-Fehlerabschätzung zu

$$\rho_k = \frac{N}{2m}|x_k - z| \leq \rho^{2^k} \iff |x_k - z| \leq \frac{2m}{N}\rho^{2^k}.$$

□

Proposition 2.28 (a-posteriori-Fehlerabschätzung)
Für das Newton'sche Iterationsverfahren gilt die a-posteriori-Fehlerabschätzung

$$|x_k - z| \leq \frac{|f(x_k)|}{m} \leq \frac{N}{2m}|x_k - x_{k-1}|^2.$$

Beweis. Der linke Teil der Ungleichung wird mit Hilfe des Mittelwertsatzes der Differentialrechnung bewiesen, nach dem ein $\xi \in [a, b]$ existiert, so dass

$$\begin{aligned} |f(x_k)| = |f(x_k) - \underbrace{f(z)}_{=0}| &= |f'(\xi)(x_k - z)| \\ \Longrightarrow |x_k - z| = \frac{|f(x_k)|}{|f'(\xi)|} &\leq \frac{|f(x_k)|}{m} \end{aligned}$$

gilt. Für den rechten Teil benutzen wir einerseits die Iterationsvorschrift

$$\begin{aligned} x_k &= x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})} \\ \Longrightarrow \; -f(x_{k-1}) &= f'(x_{k-1})(x_k - x_{k-1}), \end{aligned}$$

andererseits liefert die Taylor-Entwicklung von f um x_{k-1} in x_k

$$f(x_k) = f(x_{k-1}) + \underbrace{f'(x_{k-1})(x_k - x_{k-1})}_{=-f(x_{k-1})} + \frac{f''(\zeta)}{2}(x_k - x_{k-1})^2 = \frac{f''(\zeta)}{2}(x_k - x_{k-1})^2$$

mit $\zeta \in [a, b]$. Den Betrag obiger Gleichung dividieren wir durch m, so dass wir schließlich erhalten:

$$\frac{|f(x_k)|}{m} = \frac{|f''(\zeta)|}{2m}|x_k - x_{k-1}|^2 \leq \frac{N}{2m}|x_k - x_{k-1}|^2$$

□

3 Polynominterpolation

3.1 Interpolationspolynome

Definition 3.1 (Vektorraum der Polynome)
Π_m *sei der Raum der Polynome vom Grad* $\leq m$,

$$\Pi_m := \langle x^j : j = 0, \ldots, m \rangle = \left\{ \sum_{j=0}^{m} a_j x^j : a_j \in \mathbb{R},\ j = 0, \ldots, m \right\}.$$

Gegeben seien $m+1$ paarweise verschiedene *Interpolationsstützstellen*

$$x_0, x_1, \ldots, x_m \in \mathbb{R}$$

und ebenso viele *Interpolationsdaten*

$$y_0, y_1, \ldots, y_m \in \mathbb{R}.$$

Die Grundaufgabe der Polynominterpolation ist es nun, ein Polynom $P \in \Pi_m$ zu finden, so dass für $0 \leq j \leq m$ gilt:

$$P(x_j) = y_j$$

Bemerkung 3.2 (Geometrische Interpretation)
Man kann die Polynominterpolation als Suche nach einem Polynom $P \in \Pi_m$ verstehen, dessen Graph $y = P(x)$ in der x-y-Ebene durch die Punkte $(x_0, y_0), \ldots, (x_m, y_m)$ geht.

3.2 Lagrange-Darstellung

Proposition 3.3 (Eindeutigkeit des Interpolationspolynoms)
Es gibt genau ein Interpolationspolynom $P \in \Pi_m$ *mit* $P(x_j) = y_j$ *für* $j = 0, \ldots, m$.

Beweis. Zunächst definieren wir das *Knotenpolynom*

$$w(x) := (x - x_0)(x - x_1) \cdots (x - x_m) \in \Pi_{m+1}$$

sowie die $m+1$ Polynome ($0 \leq j \leq m$)

$$w_j(x) := \frac{w(x)}{x - x_j} = (x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_m) = \prod_{k=0, k \neq j}^{m} (x - x_k) \in \Pi_m.$$

Für diese Polynome gilt

$$w(x_k) = 0 \quad \text{und} \quad w_j(x_k) = 0 \quad (\text{für } j \neq k)$$

und

$$w'(x_j) = \lim_{x \to x_j} \frac{w(x) - \overbrace{w(x_j)}^{=0}}{x - x_j} = \lim_{x \to x_j} \frac{w(x)}{x - x_j} = \lim_{x \to x_j} w_j(x) = w_j(x_j),$$

zusammengefasst also

$$w_j(x_k) = \delta_{jk} w'(x_k).$$

Wir definieren nun die *Lagrange-Grundpolynome* $(0 \leq j \leq m)$

$$l_j(x) := \frac{w(x)}{w'(x_j)(x - x_j)} = \frac{w_j(x)}{w'(x_j)} = \frac{w_j(x)}{w_j(x_j)} \in \Pi_m,$$

die die Eigenschaften

$$\begin{aligned} l_j(x_j) &= \frac{w_j(x_j)}{w'(x_j)} = \frac{w_j(x_j)}{w_j(x_j)} = 1 \\ l_j(x_k) &= \frac{w_j(x_k)}{w'(x_k)} = \frac{0}{w'(x_k)} = 0 \quad (\text{für } j \neq k)\,, \end{aligned}$$

haben. Zusammengefasst erhalten wir:

$$l_j(x_k) = \delta_{jk} \quad (\text{für } 0 \leq j, k \leq m)$$

Damit sind die Lagrange-Grundpolynome linear unabhängig. Wegen $\dim \Pi_m = m + 1$ bilden die Lagrange-Grundpolynome eine Basis von Π_m. Als nächstes beweisen wir die Existenz eines Interpolationspolynoms mit der *Lagrange-Form*:

$$\begin{aligned} P(x) &:= \sum_{j=0}^{m} y_j l_j(x) \in \Pi_m \\ \Longrightarrow\ P(x_k) &:= \sum_{j=0}^{m} y_j \underbrace{l_j(x_k)}_{=\delta_{jk}} = y_k \quad (\text{für } 0 \leq k \leq m) \end{aligned}$$

Schließlich ist noch die Eindeutigkeit zu zeigen. Dazu nehmen wir an, dass $Q \in \Pi_m$ ein weiteres Interpolationspolynom sei:

$$Q(x) = \sum_{j=0}^{m} c_j l_j(x) \quad \text{und} \quad Q(x_j) = y_j \quad (\text{für } 0 \leq j \leq m)$$

Dann gilt

$$\begin{aligned} y_k = Q(x_k) &= \sum_{j=0}^{m} c_j l_j(x_k) = c_k \\ \Longrightarrow\ & c_0 = y_0, \ldots, c_m = y_m \\ \Longrightarrow\ & P = Q. \end{aligned}$$

□

Beispiel 3.4 (Lineare Interpolation)
Gegeben seien die Punkte $(x_0, y_0), (x_1, y_1)$. Das zugehörige Interpolationspolynom lautet

$$P(x) = y_0 l_0(x) + y_1 l_1(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0} \in \Pi_1.$$

Beispiel 3.5 (Quadratische Interpolation)
Gegeben seien die Punkte $(x_0, y_0), (x_1, y_1), (x_2, y_2)$. Das zugehörige Interpolationspolynom lautet

$$\begin{aligned} P(x) &= y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x) \\ &= y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \in \Pi_2. \end{aligned}$$

3.3 Operatorschreibweise

Definition 3.6 (Operatorschreibweise)
Handelt es sich bei den Interpolationsdaten um Funktionswerte $y_j = f(x_j)$ für $j = 0, \ldots, m$, so benutzen wir die Operatorschreibweise $L_m : C(\mathbb{R}) \mapsto \Pi_m$ mit

$$L_m(f)(x) := \sum_{j=0}^{m} f(x_j) l_j(x).$$

Proposition 3.7 (Linearität und Idempotenz des Operators)
L_m ist ein linearer und idempotenter Operator (ein Projektor): Für $\alpha \in \mathbb{R}$ und $f, g \in C(\mathbb{R})$ gilt:

$$\begin{aligned} &(i) && L_m(f+g) = L_m(f) + L_m(g) \\ &(ii) && L_m(\alpha f) = \alpha L_m(f) \\ &(iii) && L_m(L_m(f)) = L_m(f) \end{aligned}$$

Beweis. Zu (i):

$$\begin{aligned} L_m(f+g)(x) &= \sum_{j=0}^{m} \underbrace{(f+g)(x_j)}_{=f(x_j)+g(x_j)} l_j(x) \\ &= \sum_{j=0}^{m} f(x_j) l_j(x) + \sum_{j=0}^{m} g(x_j) l_j(x) \\ &= L_m(f)(x) + L_m(g)(x) \end{aligned}$$

Zu (ii): Der Beweis wird analog dem Beweis zu (1) durchgeführt.

Zu (iii): $L_m(L_m(f))$ ist das eindeutige Polynom in Π_m mit

$$L_m(L_m(f))(x_j) = L_m(f)(x_j) = f(x_j) \quad (\text{für } 0 \le j \le m)$$

Da $L_m(f)$ dieselben Eigenschaften besitzt, folgt $L_m(L_m(f)) = L_m(f)$. □

Proposition 3.8 (Bild und Kern des Operators)
L_m hat als Projektor auf dem Raum der Polynome Π die Eigenschaften

$$\begin{aligned} &(i) && \text{Bild}(L_m) = \Pi_m \\ &(ii) && \text{Kern}(L_m) = w(x) \cdot \Pi = \{w(x) \cdot q(x) : q \in \Pi\}. \end{aligned}$$

Beweis. Zu (i): Die Inklusion $\text{Bild}(L_m) \subset \Pi_m$ ist klar. Umgekehrt ist $\Pi_m \subset C(\mathbb{R})$ und für $p \in \Pi_m$ gilt $L_m(p) = p$. Somit ist auch $\Pi_m \subset \text{Bild}(L_m)$.

Zu (ii): Sei einerseits $f \in w(x) \cdot \Pi$ beliebig, also $f(x) = (x - x_0) \cdots (x - x_m) q(x)$. Dann gilt

$$\begin{aligned} L_m(f)(x) &= \sum_{j=0}^{m} \underbrace{f(x_j)}_{=0} l_j(x) = 0 \\ \Longrightarrow w(x) \cdot \Pi &\subseteq \text{Kern}(L_m). \end{aligned}$$

Sei andererseits $f \in \mathrm{Kern}(L_m) \cap \Pi$ beliebig. Dann verschwindet $L_m(f) \in \Pi_m$ an jeder Stelle x_j:

$$f(x_j) = L_m(f)(x_j) = 0 \quad (\text{für } 0 \le j \le m)$$

Die Nullstellen x_j können wir nun abdividieren und erhalten

$$\begin{aligned} & f(x) = (x - x_0) \cdots (x - x_m) q(x) \in w(x) \cdot \Pi \\ \Longrightarrow\ & \mathrm{Kern}(L_m) \subseteq w(x) \cdot \Pi. \end{aligned}$$

Insgesamt ergibt sich somit $\mathrm{Kern}(L_m) = w(x) \cdot \Pi$. □

Proposition 3.9 (Vandermonde-Matrix)
Die Vandermonde-Matrix

$$(x_j^k)_{j,k=0,\ldots,m} = \begin{pmatrix} x_0^0 & x_0^1 & \ldots & x_0^m \\ x_1^0 & x_1^1 & \ldots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ x_m^0 & x_m^1 & \ldots & x_m^m \end{pmatrix}$$

ist für paarweise verschiedene x_j mit $0 \le j \le m$ regulär (dabei sind die unteren Zahlen als Index, die oberen als Exponent zu lesen).

Beweis. Setze $p(x) := \sum_{j=0}^m a_j x^j$. Das Interpolationsproblem $p(x_j) = y_j$ ist nach Proposition 3.3 für beliebige y_j mit $0 \le j \le m$ eindeutig lösbar. Äquivalent dazu ist das lineare Gleichungssystem

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} p(x_0) \\ p(x_1) \\ \vdots \\ p(x_m) \end{pmatrix} = \begin{pmatrix} x_0^0 & x_0^1 & \ldots & x_0^m \\ x_1^0 & x_1^1 & \ldots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ x_m^0 & x_m^1 & \ldots & x_m^m \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix},$$

das somit ebenfalls eindeutig lösbar ist. Daraus folgt, dass die Systemmatrix des linearen Gleichungssystems, die Vandermonde-Matrix, regulär ist. □

3.4 Dividierte Differenzen und Newton-Darstellung

Definition 3.10
Gegeben seien $m+1$ (paarweise verschiedene) Interpolationsstützstellen $x_0, \ldots, x_m \in \mathbb{R}$ und ebenso viele Interpolationsdaten $y_0, \ldots, y_m \in \mathbb{R}$. Ferner sei für ein l mit $0 \le l \le m$ die $(l+1)$-elementige Menge $\{i_0, \ldots, i_l\}$ gegeben durch

$$\{i_0, \ldots, i_l\} \subseteq \{0, \ldots, m\}.$$

Wir bezeichnen mit $L_{i_0,\ldots,i_l}(f) \in \Pi_l$ das eindeutig bestimmte Polynom l-ten Grades, das durch die Bedingung

$$L_{i_0,\ldots,i_l}(f)(x_{i_k}) = f(x_{i_k}) \quad (\text{für } 0 \le k \le l)$$

festgelegt ist. (Es ist somit das Interpolationspolynom zu den Interpolationsstützstellen $x_{i_0}, \ldots, x_{i_l}$ und den zugehörigen Interpolationsdaten $y_{i_0}, \ldots, y_{i_l}$.)

Beispiel 3.11
Gegeben sei $m = 5$, $l = 2$ sowie $i_0 = 0$, $i_1 = 2$, $i_2 = 5$. Es ist $\{0, 2, 5\} \subseteq \{0, 1, 2, 3, 4, 5\}$. Nun muss für $L_{0,2,5}(f)$ gelten:

$$\begin{aligned} L_{0,2,5}(f)(x_0) &= f(x_0) \\ L_{0,2,5}(f)(x_2) &= f(x_2) \\ L_{0,2,5}(f)(x_5) &= f(x_5) \end{aligned}$$

Wir leiten nun eine Rekursion zur Berechnung des Interpolationspolynoms her. Dazu betrachten wir zunächst die quadratische Interpolation von drei Punkten (x_0, y_0), (x_1, y_1) und (x_2, y_2). Die Gerade $L_{0,1}(f)$ (lineares Polynom!) durch (x_0, y_0) und (x_1, y_1) sowie die Gerade $L_{1,2}(f)$ durch (x_1, y_1) und (x_2, y_2) sind einfach zu bestimmen. Sei

$$P(x) := \frac{x - x_2}{x_0 - x_2} L_{0,1}(f)(x) + \frac{x - x_0}{x_2 - x_0} L_{1,2}(f)(x) \in \Pi_2.$$

Das so definierte Polynom hat die Eigenschaften

$$\begin{aligned} P(x_0) &= L_{0,1}(f)(x_0) + 0 = f(x_0), \\ P(x_2) &= 0 + L_{1,2}(f)(x_2) = f(x_2), \\ P(x_1) &= \frac{x_1 - x_2}{x_0 - x_2} \underbrace{L_{0,1}(f)(x_1)}_{=f(x_1)} + \frac{x_1 - x_0}{x_2 - x_0} \underbrace{L_{1,2}(f)(x_1)}_{=f(x_1)} \\ &= \underbrace{\frac{x_1 - x_2 + x_0 - x_1}{x_0 - x_2}}_{=1} f(x_1) = f(x_1). \end{aligned}$$

Daher gilt $P(x) \equiv L_{0,1,2}(f)(x)$. Allgemein gilt:

Proposition 3.12 (Aitken-Rekursion oder rekursive lineare Interpolation)
Es gilt die Dreiecksrekursion

$$L_{j,\dots,j+l}(f)(x) \cdot (x_j - x_{j+l}) = L_{j,\dots,j+l-1}(f)(x) \cdot (x - x_{j+l}) + L_{j+1,\dots,j+l}(f)(x) \cdot (x_j - x).$$

Beweis. Setze

$$Q(x) := \frac{1}{x_j - x_{j+l}} \left[L_{j,\dots,j+l-1}(f)(x) \cdot (x - x_{j+l}) + L_{j+1,\dots,j+l}(f)(x) \cdot (x_j - x) \right] \in \Pi_l.$$

Analog dem obigen quadratischen Fall berechnen wir die Werte von $Q(x_k)$. Für $j < k < j + l$ gilt

$$\begin{aligned} Q(x_k) &= \frac{1}{x_j - x_{j+l}} \left[L_{j,\dots,j+l-1}(f)(x_k) \cdot (x_k - x_{j+l}) + L_{j+1,\dots,j+l}(f)(x_k) \cdot (x_j - x_k) \right] \\ &= \frac{1}{x_j - x_{j+l}} \left[f(x_k) \cdot (x_k - x_{j+l}) + f(x_k) \cdot (x_j - x_k) \right] \\ &= f(x_k). \end{aligned}$$

Für die noch verbleibenden Fälle $k = j$ bzw. $k = j + l$ erhalten wir

$$\begin{aligned} Q(x_j) &= \frac{1}{x_j - x_{j+l}} \left[L_{j,\dots,j+l-1}(f)(x_j) \cdot (x_j - x_{j+l}) + L_{j+1,\dots,j+l}(f)(x_j) \cdot (x_j - x_j) \right] \\ &= f(x_j), \\ Q(x_{j+l}) &= f(x_{j+l}). \end{aligned}$$

Damit erhalten wir das Resultat $Q(x) \equiv L_{j,\dots,j+l}(f)(x)$. □

Bemerkung 3.13 (Rekursions-Schema zur Dreiecksrekursion)
Die Aitken-Rekursion erlaubt die Berechnung des Interpolationspolynoms nach dem Schema:

x_0		x_1		x_2		x_3	Knoten
y_0		y_1		y_2		y_3	Daten
$\downarrow$	$\swarrow$	$\downarrow$	$\swarrow$	$\downarrow$	$\swarrow$		
$L_{0,1}$		$L_{1,2}$		$L_{2,3}$			
$\downarrow$	$\swarrow$	$\downarrow$	$\swarrow$				
$L_{0,1,2}$		$L_{1,2,3}$					
$\downarrow$	$\swarrow$						
$L_{0,1,2,3}$							

Definition 3.14 (Dividierte Differenz)
Den höchsten Koeffizienten (d.h. den Koeffizienten der höchsten Potenz) von $L_{i_0,\ldots,i_l}(f)(x)$ nennt man dividierte Differenz *von f in $x_{i_0},\ldots,x_{i_l}$. Man bezeichnet ihn mit $f[x_{i_0},\ldots,x_{i_l}]$. Offensichtlich gilt:*

$$f[x_{i_0},\ldots,x_{i_l}] = \frac{1}{l!}\frac{\mathrm{d}^l}{\mathrm{d}x^l} L_{i_0,\ldots,i_l}(f)(x).$$

Proposition 3.15 (Unabhängigkeit von der Anordnung der Stützstellen)
Die dividierte Differenz ist von der Anordnung der Stützstellen unabhängig, d.h.

$$f[x_{i_0},\ldots,x_{i_l}] = f[x_{\varphi(i_0)},\ldots,x_{\varphi(i_l)}]$$

für jede Permutation φ von $\{i_0,\ldots,i_l\}$.

Beweis. Das Interpolationspolynom hängt nicht von der Reihenfolge der Knoten x_k ab. Es ist

$$L_{i_0,\ldots,i_l}(f)(x) \equiv L_{\varphi(i_0),\ldots,\varphi(i_l)}(f)(x).$$

Daher ist auch der höchste Koeffizient unabhängig von der Reihenfolge der Knoten, woraus die Behauptung folgt. □

Proposition 3.16 (Newton-Darstellung des Interpolationspolynoms)
Das Interpolationspolynom kann dargestellt werden durch

$$\begin{aligned} L_{0,\ldots,l}(f)(x) &= f[x_0] + f[x_0,x_1]v_0(x) + \ldots + f[x_0,\ldots,x_l]v_{l-1}(x) \\ &= \sum_{j=0}^{l} f[x_0,\ldots,x_j]v_{j-1}(x) \end{aligned}$$

mit $f[x_0] = f(x_0)$ sowie $v_{-1}(x) := 1$ und

$$v_j(x) := (x-x_j)\cdot v_{j-1}(x) = (x-x_j)(x-x_{j-1})\ldots(x-x_0) \quad (\textit{für } 0 \le j \le l)$$

Beweis. Wir setzen $Q_l(x) := L_{0,\ldots,l}(f)(x) - L_{0,\ldots,l-1}(f)(x) \in \Pi_l$. Dann gilt für $0 \le k \le l-1$

$$Q_l(x_k) = L_{0,\ldots,l}(f)(x_k) - L_{0,\ldots,l-1}(f)(x_k) = f(x_k) - f(x_k) = 0.$$

Wir können daher die Linearfaktoren $(x - x_k)$ mit $0 \leq k \leq l-1$ abdividieren und erhalten

$$Q_l(x) = \gamma \underbrace{(x - x_{l-1})(x - x_{l-2}) \dots (x - x_0)}_{=v_{l-1}(x)}$$

Hier ist γ der höchste Koeffizient von $Q_l(x)$ und somit auch der höchste Koeffizient von $L_{0,\dots,l}(f)(x)$, weil der Grad von $L_{0,\dots,l-1}(f)(x)$ um eins niedriger ist als der von $L_{0,\dots,l}(f)(x)$. Deshalb gilt $\gamma = f[x_0, \dots, x_l]$ und wir schreiben

$$Q_l(x) = f[x_0, \dots, x_l] v_{l-1}(x)$$

oder

$$L_{0,\dots,l}(f)(x) = L_{0,\dots,l-1}(f)(x) + f[x_0, \dots, x_l] v_{l-1}(x).$$

Führen wir diese Betrachtung nun auch für $L_{0,\dots,l-1}(f)(x)$ usw. durch, so gelangen wir zur Newton-Darstellung des Interpolationspolynoms. □

Proposition 3.17 (Rekursion für die dividierten Differenzen)
Die dividierten Differenzen können mit Hilfe der folgenden Rekursion gewonnen werden:

$$f[x_j, \dots x_{j+l}] = \frac{f[x_j, \dots, x_{j+l-1}] - f[x_{j+1}, \dots, x_{j+l}]}{x_j - x_{j+l}}.$$

Beweis. Nach der Aitken-Rekursion (siehe Proposition 3.12) gilt

$$L_{j,\dots,j+l}(f)(x) = \frac{1}{x_j - x_{j+l}} \left[L_{j,\dots,j+l-1}(f)(x) \cdot (x - x_{j+l}) + L_{j+1,\dots,j+l}(f)(x) \cdot (x_j - x) \right].$$

Dabei können wir schreiben:

$$\begin{aligned}
L_{j,\dots,j+l}(f)(x) &= f[x_j, \dots, x_{j+l}] x^l + \dots \\
L_{j,\dots,j+l-1}(f)(x) &= f[x_j, \dots, x_{j+l-1}] x^{l-1} + \dots \\
L_{j+1,\dots,j+l}(f)(x) &= f[x_{j+1}, \dots, x_{j+l}] x^{l-1} + \dots
\end{aligned}$$

Sortieren wir nun die rechte Seite der Aitken-Rekursions-Formel nach Potenzen von x, so erhalten wir durch Koeffizientenvergleich für den höchsten Koeffizienten

$$\frac{f[x_j, \dots, x_{j+l-1}] - f[x_{j+1}, \dots, x_{j+l}]}{x_j - x_{j+l}} = f[x_j, \dots, x_{j+l}].$$

□

Beispiel 3.18 (Dividierte Differenzen)
Wir erhalten für die dividierten Differenzen:

$$\begin{aligned}
f[x_0] &= f(x_0) \\
f[x_0, x_1] &= \frac{f[x_0] - f[x_1]}{x_0 - x_1} = \frac{f(x_0) - f(x_1)}{x_0 - x_1} \\
f[x_0, x_1, x_2] &= \frac{f[x_0, x_1] - f[x_1, x_2]}{x_0 - x_2} = \frac{\frac{f(x_0) - f(x_1)}{x_0 - x_1} - \frac{f(x_1) - f(x_2)}{x_1 - x_2}}{x_0 - x_2} \\
&\dots
\end{aligned}$$

Damit erklärt sich der Name *dividierte Differenzen*.

Bemerkung 3.19 (Rekursions-Schema für die dividierten Differenzen)
Für die dividierten Differenzen ergibt sich das folgende Schema:

x_0	x_1	x_2	x_3
$\mathbf{f[x_0] = y_0}$	$f[x_1] = y_1$	$f[x_2] = y_2$	$f[x_3] = y_3$
$\mathbf{f[x_0, x_1]}$	$f[x_1, x_2]$	$f[x_2, x_3]$	
$\mathbf{f[x_0, x_1, x_2]}$	$f[x_1, x_2, x_3]$		
$\mathbf{f[x_0, x_1, x_2, x_3]}$			

In der ersten Spalte des Schemas stehen gerade die Koeffizienten des Interpolationspolynoms in der Newton-Darstellung.

Beispiel 3.20 (Dividierten Differenzen und Newton-Darstellung)
Gegeben: $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 4$ und $y_0 = 1$, $y_1 = 0$, $y_2 = 3$, $y_3 = -1$. Die dividierten Differenzen sollen mit Hilfe des obigen Schemas berechnet werden:

0	1	2	4
$\mathbf{1}$	0	3	-1
$\frac{1-0}{0-1} = \mathbf{-1}$	$\frac{0-3}{1-2} = 3$	$\frac{3-(-1)}{2-4} = -2$	
$\frac{-1-3}{0-2} = \mathbf{2}$	$\frac{3-(-2)}{1-4} = -\frac{5}{3}$		
$\frac{2-(-5/3)}{0-4} = -\mathbf{\frac{11}{12}}$			

Damit erhalten wir für das Interpolationspolynom in der Newton-Darstellung

$$L_{0,1,2,3}(f)(x) = 1 + (-1)(x-0) + 2(x-0)(x-1) + \left(-\frac{11}{12}\right)(x-0)(x-1)(x-2).$$

3.5 Hermite-Interpolation

Mit den bisherigen Approximationsverfahren kann eine Funktion an bestimmten Stützstellen approximiert werden. Mittels Hermite-Interpolation können auch Ableitungen von Funktionen approximiert werden.

Integraldarstellung der dividierten Differenzen

Im Folgenden sei $f : [a, b] \mapsto \mathbb{R}$ eine $(m+1)$ mal differenzierbare Funktion. Seien weiter $x_0, \ldots, x_{m+1} \in [a, b]$, jedoch die Punkte nicht notwendigerweise verschieden.

Definition 3.21 (Notation für Integrale)
Als vereinfachte Schreibweise definieren wir

$$\int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{l-1}} dt_l f(t_1, \ldots, t_l) := \int_0^1 \left(\int_0^{t_1} \left(\cdots \left(\int_0^{t_{l-1}} f(t_1, \ldots, t_l) dt_l \right) \cdots \right) dt_2 \right) dt_1.$$

Definition 3.22 (Integraldarstellung der dividierten Differenzen)
Mit obiger Integralschreibweise definieren wir

$$f\langle x_0,\dots,x_l\rangle := \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{l-1}} dt_l f^{(l)}\left(x_0 + t_1(x_1-x_0) + t_2(x_2-x_1) + \cdots + t_l(x_l - x_{l-1})\right)$$

Proposition 3.23 (Relation von Hermite)
Für $x_0,\dots,x_{m+1} \in [a,b]$ nicht notwendigerweise verschieden gilt

$$f\langle x_0,\dots,x_{l-1},x\rangle = f\langle x_0,\dots,x_{l-1},x_l\rangle + f\langle x_0,\dots,x_{l-1},x_l,x\rangle\,(x-x_l).$$

Beweis. Für $x = x_l$ ist

$$\begin{aligned} &f\langle x_0,\dots,x_{l-1},x_l\rangle + f\langle x_0,\dots,x_{l-1},x_l,x\rangle\,(x-x_l)\\ &= f\langle x_0,\dots,x_{l-1},x_l\rangle + f\langle x_0,\dots,x_{l-1},x_l,x_l\rangle\,(x_l-x_l)\\ &= f\langle x_0,\dots,x_{l-1},x\rangle\,. \end{aligned}$$

Sei nun $x \neq x_l$. Dann definieren wir

$$\begin{aligned} & F(t_{l+1}) := f^{(l)}(x_0 + \cdots + t_l(x_l - x_{l-1}) + t_{l+1}(x - x_l))\\ \Longrightarrow\ & \frac{\mathrm{d}}{\mathrm{d}t_{l+1}} F(t_{l+1}) = f^{(l+1)}(x_0 + \cdots + t_l(x_l - x_{l-1}) + t_{l+1}(x - x_l))\cdot(x - x_l)\\ \Longrightarrow\ & \frac{1}{x-x_l}\frac{\mathrm{d}}{\mathrm{d}t_{l+1}} F(t_{l+1}) = f^{(l+1)}(x_0 + \cdots + t_l(x_l - x_{l-1}) + t_{l+1}(x - x_l)). \end{aligned}$$

Somit folgt

$$\begin{aligned} &(x-x_l) f\langle x_0,\dots,x_l,x\rangle\\ &= (x-x_l)\int_0^1 dt_1 \cdots \int_0^{t_{l-1}} dt_l \int_0^{t_l} dt_{l+1}\, f^{(l+1)}\left(x_0 + t_1(x_1-x_0) + \cdots + t_l(x_l-x_{l-1}) + t_{l+1}(x-x_l)\right)\\ &= (x-x_l)\frac{1}{x-x_l}\int_0^1 dt_1 \cdots \int_0^{t_{l-1}} dt_l \int_0^{t_l} dt_{l+1}\, \frac{\mathrm{d}}{\mathrm{d}t_{l+1}} F(t_{l+1})\\ &= \int_0^1 dt_1 \cdots \int_0^{t_{l-1}} dt_l\ F(t_{l+1})|_0^{t_l}\\ &= \int_0^1 dt_1 \cdots \int_0^{t_{l-1}} dt_l f^{(l)}(x_0 + \cdots + t_l(x - x_{l-1}))) - \int_0^1 dt_1 \cdots \int_0^{t_{l-1}} dt_l f^{(l)}(x_0 + \cdots + t_l(x_l - x_{l-1})))\\ &= f\langle x_0,\dots,x_{l-1},x\rangle - f\langle x_0,\dots,x_{l-1},x_l\rangle\,. \end{aligned}$$

□

Proposition 3.24 (Dividierte Differenzen und Integraldarstellung stimmen überein)
Sind $x_0,\dots,x_m$ paarweise verschieden, so ist

$$f[x_0,\dots,x_m] = f\langle x_0,\dots,x_m\rangle\,.$$

Beweis. Das Newton-Interpolationspolynom P zu f und paarweise verschiedenen $x_0, \dots, x_m$ hat die Darstellung

$$P(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \dots, x_m](x - x_0) \cdots (x - x_{m-1}).$$

Dann gilt $f(x_k) = P(x_k)$ für $k = 0, \dots, m$.

Andererseits folgt aus der Relation von Hermite

$$\begin{aligned} f(x) &= f\langle x_0 \rangle + (x - x_0) f\langle x_0, x \rangle \\ &= f\langle x_0 \rangle + (x - x_0)(f\langle x_0, x_1 \rangle + (x - x_1) f\langle x_0, x_1, x \rangle) \\ &= f\langle x_0 \rangle + (x - x_0) f\langle x_0, x_1 \rangle + (x - x_0)(x - x_1) f\langle x_0, x_1, x \rangle \\ &= f\langle x_0 \rangle + (x - x_0) f\langle x_0, x_1 \rangle + (x - x_0)(x - x_1)(f\langle x_0, x_1, x_2 \rangle + (x - x_2) f\langle x_0, x_1, x_2, x \rangle) \\ &= \dots \\ &= f\langle x_0 \rangle + (x - x_0) f\langle x_0, x_1 \rangle + \cdots + (x - x_0) \cdots (x - x_{m-1}) f\langle x_0, \dots, x_{m-1}, x \rangle . \end{aligned}$$

Somit gilt die Identität

$$f(x) = f\langle x_0 \rangle + (x - x_0) f\langle x_0, x_1 \rangle + \cdots + (x - x_0) \cdots (x - x_{m-1}) f\langle x_0, \dots, x_{m-1}, x \rangle \tag{4}$$

Nun gilt für $k = 0, \dots, m$ die Gleichheit $P(x_k) = f(x_k)$ und somit

$$f[x_0] = P[x_0] = f(x_0) = f\langle x_0 \rangle .$$

Weiter ist für x_1 wegen $x_1 - x_0 \neq 0$ nach Identität (4):

$$\begin{aligned} P(x_1) &= f[x_0] + f[x_0, x_1](x_1 - x_0) \\ f(x_1) &= f\langle x_0 \rangle + f\langle x_0, x_1 \rangle (x_1 - x_0) \\ \Longrightarrow f\langle x_0, x_1 \rangle &= f[x_0, x_1] \end{aligned}$$

Induktiv folgt für $k = 1, \dots, m$:

$$\begin{aligned} P(x_k) &= f[x_0] + f[x_0, x_1](x_1 - x_0) + \cdots + f[x_0, \dots, x_k](x_k - x_0) \cdots (x_k - x_{k-1}) \\ f(x_k) &= f\langle x_0 \rangle + f\langle x_0, x_1 \rangle (x_1 - x_0) + \cdots + f\langle x_0, \dots, x_k \rangle (x_k - x_0) \cdots (x_k - x_{k-1}) \\ \Longrightarrow f\langle x_0, \dots, x_k \rangle &= f[x_0, \dots, x_k] \end{aligned}$$

□

Bemerkung 3.25
Die Integraldarstellung ist somit eine Erweiterung der dividierten Differenzen. Beide Definitionen stimmen für paarweise verschiedene Stützstellen überein, ansonsten ist ausschließlich die Integraldarstellung definiert.

Somit ist es im Folgenden nicht nötig, zwischen den Schreibweisen zu unterscheiden. Sowohl für die Integraldarstellung als auch für die dividierten Differenzen verwenden wir die Schreibweise $f[x_0, \dots, x_m]$.

Lösung der Hermiteschen Interpolationsaufgabe

Proposition 3.26 (Restglied)
Sei nun f eine Funktion, p das Newtonsche Interpolationspolynom an den Stützstellen $x_0, \ldots, x_m$. Dann ist

$$f(x) = p(x) + R(x)$$

mit

$$\begin{aligned} R(x) &= (x - x_0) \cdots (x - x_m) f[x_0, \ldots, x_m, x] \\ &= 0 \quad \text{wenn} \quad x = x_0, \ldots, x_m \end{aligned}$$

Beweis. Die Behauptung folgt unmittelbar aus Gleichung (4) und Proposition (3.24). □

Die bisherigen Sätze erlauben die Lösung der Hermiteschen Interpolationsaufgabe. Dabei sucht man für eine hinreichend oft differenzierbare Funktion f ein Polynom, das an gegebenen Stellen mit dem Polynom und Ableitungen bis zu einer gewissen Ordnung übereinstimmt.

Proposition 3.27 (Hermite-Interpolationsaufgabe)
Sei f eine hinreichend oft differenzierbare Funktion und $m + 1 = r_0 + r_1 + \cdots + r_n$ Punkte

$$x_0, \ldots, x_m = \underbrace{z_0, \ldots, z_0}_{r_0}, \underbrace{z_1, \ldots, z_1}_{r_1}, \ldots, \underbrace{z_n, \ldots, z_n}_{r_n}$$

gegeben. Dann gibt es ein Polynom p vom Grad m, so dass gilt:

$$f^{(s)}(z_k) = p^{(s)}(z_k) \quad (\text{für } s = 0, \ldots, r_k - 1, \quad k = 0, \ldots, n)$$

Dieses Polynom hat die Form

$$p(x) = f[x_0] + (x - x_0) f[x_0, x_1] + \cdots + (x - x_0) \cdots (x - x_{m-1}) f[x_0, \ldots, x_m].$$

Beweis. Es ist $f(x) = p(x) + R(x)$ mit obigem Restglied

$$R(x) = (x - x_0) \cdots (x - x_m) f[x_0, \ldots, x_m, x] = (x - z_0)^{r_0} \cdots (x - z_n)^{r_n} f[x_0, \ldots, x_m, x].$$

Dann gilt

$$\frac{\mathrm{d}^s}{\mathrm{d}x^s} f(x)\Big|_{x=z_k} = \frac{\mathrm{d}^s}{\mathrm{d}x^s} p(x)\Big|_{x=z_k} + \underbrace{(x - z_k)^{r_k - s} g_{k,s}(x)\big|_{x=z_k}}_{=0 \quad (\text{für } s=0,\ldots,r_k-1)} = \frac{\mathrm{d}^s}{\mathrm{d}x^s} p(x)\Big|_{x=z_k}.$$

□

Proposition 3.28 (Abschätzung des Restgliedes)
Sei G konvex, $z_0, \ldots, z_n \in G$ und $f \in C^{m+1}(G)$. Dann gilt für obiges Restglied der Hermite-Interpolation die Abschätzung

$$|R(x)| \leq \frac{|(x - x_0) \cdots (x - x_m)|}{(m+1)!} \max_{z \in G} \left| f^{(m+1)}(z) \right|.$$

Beweis. Durch schlichtes Ausrechnen des Integrals erhält man

$$\begin{aligned}
|f[x_0,\dots,x_m,x]| &\le \max_{z\in G}|f^{(m+1)}(z)| \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{m-1}} dt_m \int_0^{t_m} dt1 \\
&= \max_{z\in G}|f^{(m+1)}(z)| \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{m-1}} dt_m(t_m) \\
&= \max_{z\in G}|f^{(m+1)}(z)| \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{m-2}} dt_{m-1}\frac{1}{2}t_{m-1}^2 \\
&= \dots \\
&= \max_{z\in G}|f^{(m+1)}(z)|\frac{1}{(m+1)!}
\end{aligned}$$

und somit

$$|R(x)| = |(x-x_0)\cdots(x-x_m)||f[x_0,\dots,x_m,x]| \le \frac{|(x-x_0)\cdots(x-x_m)|}{(m+1)!}\max_{z\in G}|f^{(m+1)}(z)|.$$

□

Proposition 3.29 (Darstellung des Restgliedes)
Sei $G \subset \mathbb{R}$. Dann gibt es ein ξ mit

$$\min\{x_0,\dots,x_m,x\} \le \xi \le \max\{x_0,\dots,x_m,x\},$$

so dass

$$R(x) = \frac{(x-x_0)\cdots(x-x_m)}{(m+1)!}f^{(m+1)}(\xi).$$

Beweis. Folgt aus dem reellen Zwischenwertsatz angewandt auf $f^{(m+1)}$. □

Beispiel 3.30 (Restgliedabschätzung)
Sei $f(x) = 2^x$ und $m = 4$, $x_j = \frac{j}{4}$, $j = 0,1,2,3,4$. Dann

$$\begin{aligned}
& f^{(l)}(x) = 2^x(\ln 2)^l \\
\Longrightarrow\ & |R(x)| \le \frac{1}{120}\left|x\left(x-\frac{1}{4}\right)\left(x-\frac{1}{2}\right)\left(x-\frac{3}{4}\right)(x-1)\right|(\ln 2)^5\cdot 2^z
\end{aligned}$$

für $z = \max\{x_0,x_1,x_2,x_3,x_4,x\}$

Hermite-Interpolation für identische Stützstellen

Seien $x_0 = x_1 = \cdots = x_m$. Somit folgt analog zur Abschätzung des Restgliedes:

$$f[\underbrace{x_0,\dots,x_0}_{l+1}] = \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{l-1}} dt_l f^{(l)}(x_0) = \frac{1}{l!}f^{(l)}(x_0)$$

Mittels des Hermiteschen Interpolationspolynoms gilt

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^m(x_0)}{m!}(x - x_0)^m + R_m(x),$$

wobei

$$|R_m(x)| \leq \frac{|x - x_0|^{(m+1)}}{(m+1)!} \max_{0 \leq t \leq 1} \left| f^{(m+1)}(x_0 + t(x - x_0)) \right|.$$

Dieses ist gerade die Taylor-Formel. Das Ergebnis ist plausibel, da die Hermitesche Interpolationsaufgabe für identische Stützstellen gerade dem Taylorproblem entspricht.

Als Darstellung des Restgliedes ergibt sich: Es existiert ein ξ mit $\min(x_0, x) \leq \xi \leq \max(x_0, x)$ mit

$$R_m(x) = \frac{(x - x_0)^{m+1}}{(m+1)!} f^{(m+1)}(\xi).$$

Hermite-Interpolation für äquidistante Stützstellen

Seien $m + 1$ verschiedene äquidistante Stützstellen $x_j = x_0 + j \cdot h, j = 0, \ldots, m$ mit Schrittweite h gegeben.

Definition 3.31 (vorwärtsgenommene Differenzen)
Die vorwärtsgenommen Differenzen sind definiert für $j = 0, \ldots, m - s$ und $s = 1, \ldots, m$ durch:

$$\begin{aligned} \Delta^0 y_j &= y_j \\ \Delta^s y_j &= \Delta^{s-1} y_{j+1} - \Delta^{s-1} y_j \end{aligned}$$

Proposition 3.32
Die vorwärtsgenommenen Differenzen lassen sich berechnen durch

$$\Delta^s y_j = \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} y_{j+k}.$$

Beweis. Für $s = 0$ gilt

$$\Delta^0 y_j = y_j = \binom{0}{0} y_j.$$

Induktionsschritt: $s \to s+1$

$$\begin{aligned}
\Delta^{s+1} y_j &= \Delta^s y_{j+1} - \Delta^s y_j \\
&\overset{IV}{=} \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} y_{j+k+1} - \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} y_{j+k} \\
&= \sum_{k=1}^{s+1} (-1)^{s-(k-1)} \binom{s}{k-1} y_{j+k} + \sum_{k=0}^{s} (-1)^{s+1-k} \binom{s}{k} y_{j+k} \\
&= (-1)^{s+1} y_j + y_{j+s+1} + \sum_{k=1}^{s} \left((-1)^{s+1-k} \binom{s}{k-1} y_{j+k} + (-1)^{s+1-k} \binom{s}{k} y_{j+k} \right) \\
&= (-1)^{s+1-0} y_j + (-1)^{s+1-(s+1)} y_{j+s+1} + \sum_{k=1}^{s} (-1)^{s+1-k} \binom{s+1}{k} y_{j+k} \\
&= \sum_{k=0}^{s+1} (-1)^{s+1-k} \binom{s+1}{k} y_{j+k}
\end{aligned}$$

□

Proposition 3.33 (Dividierte Differenzen im Falle äquidistanter Stützstellen)
Für die dividierten Differenzen ergibt sich

$$f[x_j, \ldots, x_{j+s}] = \frac{1}{s! h^s} \Delta^s y_j \quad (\textit{für } j = 0, \ldots, m-s, \quad s = 1, \ldots, m).$$

Beweis. Es ist

$$\begin{aligned}
f[x_j] &= y_j \\
f[x_j, x_{j+1}] &= \frac{y_j - y_{j+1}}{x_j - x_{j+1}} = \frac{1}{h}(y_{j+1} - y_j)
\end{aligned}$$

und somit per Induktion

$$\begin{aligned}
f[x_j, \ldots, x_{j+s}] &\overset{\text{Hermite}}{=} \frac{f[x_j, \ldots, x_{j+s-1}] - f[x_{j+1}, \ldots, x_{j+s}]}{x_j - x_{j+s}} \\
&\overset{IV}{=} \frac{1}{x_j - x_{j+s}} \left(\frac{1}{(s-1)! h^{s-1}} \Delta^{s-1} y_j - \frac{1}{(s-1)! h^{s-1}} \Delta^{s-1} y_{j+1} \right) \\
&= -\frac{1}{sh} \left(-\frac{1}{(s-1)! h^{s-1}} (\Delta^{s-1} y_{j+1} - \Delta^{s-1} y_j) \right) \\
&= \frac{1}{s! h^s} \Delta^s y_j.
\end{aligned}$$

□

Proposition 3.34 (Newtondarstellung des Interpolationspolynoms mit vorwärtsgenommenen Differenzen)
Im Falle äquidistanter Stützstellen ergibt sich die Newtonsche Darstellung des Interpolationspolynoms zu

$$\begin{aligned} p(x) &= \sum_{k=0}^{m} \frac{(x-x_0)\cdots(x-x_{k-1})}{k!h^k}\Delta^k y_0 \\ &= y_0 + \frac{x-x_0}{h}\Delta y_0 + \frac{(x-x_0)(x-x_1)}{2!h^2}\Delta^2 y_0 + \cdots + \frac{(x-x_0)\cdots(x-x_{m-1})}{m!h^m}\Delta^m y_0. \end{aligned}$$

Als Restglied verbleibt

$$R(x) = (x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x].$$

Definition 3.35 (rückwärtsgenommene Differenzen)
Die rückwärtsgenommen Differenzen sind definiert für $j = s,\ldots,m$ und $s = 1,\ldots,m$ durch

$$\begin{aligned} \nabla^0 y_j &= y_j \\ \nabla^s y_j &= \nabla^{s-1} y_j - \nabla^{s-1} y_{j-1}. \end{aligned}$$

Proposition 3.36 (Dividierte Differenzen im Falle äquidistanter Stützstellen)
Für die dividierten Differenzen ergibt sich

$$f[x_j,\ldots,x_{j-s}] = \frac{1}{s!h^s}\nabla^s y_j \quad (\text{für } j = s,\ldots,m, \quad s = 1,\ldots,m).$$

Beweis. Der Beweis ist analog zum Fall der vorwärtsgenommenen Differenzen. □

Proposition 3.37 (Newtondarstellung des Interpolationspolynoms mit rückwärtsgenommenen Differenzen)
Im Falle äquidistanter Stützstellen ergibt sich die Newtonsche Darstellung des Interpolationspolynoms zu

$$\begin{aligned} p(x) &= \sum_{k=0}^{m} \frac{(x-x_m)\cdots(x-x_{m-k+1})}{k!h^k}\nabla^k y_m \\ &= y_m + \frac{x-x_m}{h}\nabla y_m + \frac{(x-x_m)(x-x_{m-1})}{2!h^2}\nabla^2 y_m + \cdots + \frac{(x-x_m)\cdots(x-x_1)}{m!h^m}\nabla^m y_m. \end{aligned}$$

Als Restglied verbleibt

$$R(x) = (x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x].$$

4 Numerische Differentiation

Im folgenden Kapitel sei immer $G \subset \mathbb{R}$.

Proposition 4.1 (Taylorformel)
Sei $f \in C^{m+1}(G)$. Die Taylorentwicklung von f um $c \in G$ ist

$$f(x) = f(c) + f'(c)(x-c) + \cdots + \frac{f^{(m)}(c)}{m!}(x-c)^m + R(x).$$

Es gibt ein ξ mit $\min(x,c) \leq \xi \leq \max(x,c)$ so dass das Restglied die Darstellung

$$R(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!}(x-c)^{m+1}$$

hat. Der Abbruchfehler lässt sich abschätzen durch

$$|R(x)| \leq \frac{|(x-c)|^{m+1}}{(m+1)!} \cdot \max_{z \in G} \left| f^{(m+1)}(z) \right|.$$

Beweis. Die Existenz der Taylorentwicklung und die Lagrange-Darstellung des Restgliedes sind aus Analysis bekannt. Die Abschätzung folgt trivial. □

4.1 Differenzenquotienten erster Ordnung

Mit Differenzenquotienten erster Ordnung lässt sich die erste Ableitung einer Funktion approximieren.

Proposition 4.2 (Einseitige Differenzenquotienten erster Ordnung)
Für $f \in C^2(G)$ und $c \in G$ gelten die Approximationen:

$$f'(c) = \frac{f(c+h) - f(c)}{h} + \mathcal{O}(h) \qquad \text{(Vorwärtsgenommener Differenzenquotient)}$$

$$f'(c) = \frac{f(c) - f(c-h)}{h} + \mathcal{O}(h) \qquad \text{(Rückwärtsgenommener Differenzenquotient)}$$

Beweis. Nach der Taylorentwicklung um c gilt

$$f(x) = f(c) + f'(c)(x-c) + R(x).$$

Somit folgt:

$$f(c+h) = f(c) + hf'(c) + R(c+h) \tag{5}$$

$$f(c-h) = f(c) - hf'(c) + R(c-h) \tag{6}$$

Mit der Lagrange-Darstellung des Restgliedes gibt es ein $\xi_1 \in [c, c+h]$ und $\xi_2 \in [c-h, c]$, so dass:

$$R(c+h) = \frac{(c+h-c)^2}{2} \cdot f''(\xi_1) = \frac{h^2}{2} \cdot f''(\xi_1)$$

$$R(c-h) = \frac{(c-h-c)^2}{2} \cdot f''(\xi_2) = \frac{h^2}{2} \cdot f''(\xi_2)$$

Zum vorwärtsgenommenen Differenzenquotient: Dieser ergibt sich nach (5) durch

$$\begin{aligned} f'(c) &= \frac{f(c+h)-f(c)}{h} - \frac{1}{h}\cdot R(c+h) \\ &= \frac{f(c+h)-f(c)}{h} - \frac{1}{h}\cdot\frac{h^2}{2}\cdot f''(\xi_1) \\ &= \frac{f(c+h)-f(c)}{h} - h\cdot\frac{f''(\xi_1)}{2} \\ &= \frac{f(c+h)-f(c)}{h} + \mathcal{O}(h). \end{aligned}$$

Zum rückwärtsgenommenen Differenzenquotient: Dieser ergibt sich nach (6) durch

$$\begin{aligned} f'(c) &= \frac{f(c)-f(c-h)}{h} + \frac{1}{h}\cdot R(c-h) \\ &= \frac{f(c)-f(c-h)}{h} + \frac{1}{h}\cdot\frac{h^2}{2}\cdot f''(\xi_2) \\ &= \frac{f(c)-f(c-h)}{h} + h\cdot\frac{f''(\xi_1)}{2} \\ &= \frac{f(c)-f(c-h)}{h} + \mathcal{O}(h). \end{aligned}$$

□

Proposition 4.3 (Zentraler Differenzenquotient erster Ordnung)
Sei $c \in G$.

1. *Für $f \in C^2(G)$ gilt die Approximation*

$$f'(c) = \frac{f(c+h)-f(c-h)}{2h} + \mathcal{O}(h).$$

2. *Für $f \in C^3(G)$ gilt die Approximation*

$$f'(c) = \frac{f(c+h)-f(c-h)}{2h} + \mathcal{O}(h^2).$$

Beweis. Zu 1. Nach (5) und (6) aus dem vorherigen Beweis gilt mit geeigneten ξ_1, ξ_2:

$$\begin{aligned} f(c+h)-f(c-h) &= (f(c)+hf'(c)+R(c+h)) - (f(c)-hf'(c)+R(c-h)) \\ &= 2h\cdot f'(c) + R(c+h) - R(c-h) \\ &= 2h\cdot f'(c) + \frac{h^2}{2}\left(f''(\xi_1) - f''(\xi_2)\right) \end{aligned}$$

Somit folgt

$$\begin{aligned} f'(c) &= \frac{f(c+h)-f(c-h)}{2h} - \frac{1}{2h}\cdot\frac{h^2}{2}\cdot\left(f''(\xi_1)-f''(\xi_2)\right) \\ &= \frac{f(c+h)-f(c-h)}{2h} - h\cdot\frac{f''(\xi_1)-f''(\xi_2)}{4} \\ &= \frac{f(c+h)-f(c-h)}{2h} + \mathcal{O}(h). \end{aligned}$$

Zu 2. Nach der Taylorformel und der Lagrangedarstellung des Restgliedes gilt mit geeigneten ξ_1, ξ_2:

$$\begin{aligned} f(c+h) &= f(c) + hf'(c) + \frac{h^2}{2}f''(c) + \frac{h^3}{6}f'''(\xi_1) \\ f(c-h) &= f(c) - hf'(c) + \frac{h^2}{2}f''(c) - \frac{h^3}{6}f'''(\xi_2) \\ \implies f(c+h) - f(c-h) &= 2hf'(c) + \frac{h^3}{6}(f'''(\xi_1) + f'''(\xi_2)) \end{aligned}$$

Nach dem Zwischenwertsatz angewandt auf f''' gibt es ein ξ mit $\min(\xi_1, \xi_2) \leq \xi \leq \max(\xi_1, \xi_2)$, so dass

$$2f'''(\xi) = f'''(\xi_1) + f'''(\xi_2)$$

ist. Damit ergibt sich

$$\begin{aligned} f'(c) &= \frac{f(c+h) - f(c-h)}{2h} - \frac{1}{2h} \cdot \frac{h^3}{6}(f'''(\xi_1) + f'''(\xi_2)) \\ &= \frac{f(c+h) - f(c-h)}{2h} - h^2\frac{f'''(\xi)}{6} \\ &= \frac{f(c+h) - f(c-h)}{2h} + \mathcal{O}(h^2) \end{aligned}$$

□

4.2 Differenzenquotienten zweiter Ordnung

Mit Differenzenquotienten zweiter Ordnung lässt sich die zweite Ableitung einer Funktion approximieren.

Proposition 4.4 (Zentraler Differenzenquotient zweiter Ordnung)
Sei $c \in G$.

1. *Für $f \in C^3(G)$ gilt die Approximation*

$$f''(c) = \frac{f(c-h) - 2f(c) + f(c+h)}{h^2} + \mathcal{O}(h).$$

2. *Für $f \in C^4(G)$ gilt die Approximation*

$$f''(c) = \frac{f(c-h) - 2f(c) + f(c+h)}{h^2} + \mathcal{O}(h^2).$$

Beweis. Zu 1. Nach der Taylorformel und der Lagrangedarstellung des Restgliedes ist mit geeigneten ξ_1, ξ_2:

$$\begin{aligned} f(c+h) &= f(c) + hf'(c) + \frac{h^2}{2}f''(c) + \frac{h^3}{6}f'''(\xi_1) \\ f(c-h) &= f(c) - hf'(c) + \frac{h^2}{2}f''(c) - \frac{h^3}{6}f'''(\xi_2) \\ \implies f(c+h) + f(c-h) &= 2f(c) + h^2f''(c) + \frac{h^3}{6}(f'''(\xi_1) - f'''(\xi_2)) \end{aligned}$$

Damit ergibt sich

$$\begin{aligned} f''(c) &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} - \frac{1}{h^2}\cdot\frac{h^3}{6}\left(f'''(\xi_1)-f'''(\xi_2)\right) \\ &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} - h\frac{f'''(\xi_1)-f'''(\xi_2)}{6} \\ &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} + \mathcal{O}(h). \end{aligned}$$

Zu 2. Nach der Taylorformel und der Lagrangedarstellung des Restgliedes ist mit geeigneten ξ_1, ξ_2:

$$\begin{aligned} f(c+h) &= f(c) + hf'(c) + \frac{h^2}{2}f''(c) + \frac{h^3}{6}f'''(c) + \frac{h^4}{24}f''''(\xi_1) \\ f(c-h) &= f(c) - hf'(c) + \frac{h^2}{2}f''(c) - \frac{h^3}{6}f'''(c) + \frac{h^4}{24}f''''(\xi_2) \\ \implies f(c+h)+f(c-h) &= 2f(c) + h^2f''(c) + \frac{h^4}{24}(f''''(\xi_1)+f''''(\xi_2)) \end{aligned}$$

Nach dem Zwischenwertsatz angewandt auf f'''' gibt es ein ξ mit $\min(\xi_1,\xi_2) \le \xi \le \max(\xi_1,\xi_2)$, so dass

$$2f''''(\xi) = f''''(\xi_1) + f''''(\xi_2)$$

ist. Damit ergibt sich

$$\begin{aligned} f''(c) &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} - \frac{1}{h^2}\cdot\frac{h^4}{24}\left(f''''(\xi_1)+f''''(\xi_2)\right) \\ &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} - h^2\frac{f''''(\xi)}{12} \\ &= \frac{f(c-h)-2f(c)+f(c+h)}{h^2} + \mathcal{O}(h^2). \end{aligned}$$

□

4.3 Differenzenquotienten beliebiger Ordnung

Proposition 4.5 (Zentraler Differenzenquotient beliebiger Ordnung)
Sei $c \in G$.

1. *Sei $f \in C^{s+1}(G)$. Dann gilt*

$$f^{(s)}(c) = \frac{f^{(s-1)}(c+\frac{h}{2}) - f^{(s-1)}(c-\frac{h}{2})}{h} + \mathcal{O}(h).$$

2. *Sei $f \in C^{s+2}(G)$. Dann gilt*

$$f^{(s)}(c) = \frac{f^{(s-1)}(c+\frac{h}{2}) - f^{(s-1)}(c-\frac{h}{2})}{h} + \mathcal{O}(h^2).$$

Beweis. Beide Aussagen folgen durch Anwendung der entsprechenden Aussage von Proposition 4.3 auf $f^{(s-1)}$. □

Proposition 4.6 (Rekursionsformel zur Berechnung beliebiger Ableitungen)
Sei $f \in C^{m+1}$. Definiere $x_l = c + lh$ und $f_l := f(x_l)$ für $l \in \{0, \pm\frac{1}{2}, \pm 1, \dots\}$. Dann lässt sich die s-te Ableitung rekursiv approximieren durch die Rekursion:

$$\begin{aligned} D^0 f_j &= f_j \\ D^s f_j &= \frac{1}{h}\left(D^{s-1} f_{j+\frac{1}{2}} - D^{s-1} f_{j-\frac{1}{2}}\right) \end{aligned}$$

Beweis. Klar aus der vorherigen Proposition 4.5. □

Proposition 4.7 (Approximation beliebiger Ableitungen durch Ableitungen des Interpolationspolynoms)
Seien f, f_l und x_l wie in der vorherigen Proposition. Sei $p = p_{j-\frac{s}{2}, j-\frac{s}{2}+1, \dots, j+\frac{s}{2}}$ das Interpolation von f zu den Stützstellen $x_{j-\frac{s}{2}}, \dots, x_{j+\frac{s}{2}}$. Dann gilt

$$D^s f_j = \frac{\mathrm{d}^s}{\mathrm{d}x^s} p_{j-\frac{s}{2}, \dots, j+\frac{s}{2}}.$$

Beweis. Beweis durch Induktion: Induktionsverankerung $s = 1$: Es ist

$$D^1 f_j \overset{(4.6)}{=} \frac{1}{h}\left(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}\right) = \frac{\mathrm{d}}{\mathrm{d}x} p_{j-\frac{1}{2}, j+\frac{1}{2}},$$

denn $p_{j-\frac{1}{2}, j+\frac{1}{2}}$ ist das Interpolationspolynom zu den Stützstellen $x_{j-\frac{1}{2}} = c + (j - \frac{1}{2})h$ und $x_{j+\frac{1}{2}} = c + (j + \frac{1}{2})h$ und als solches eine Gerade der Steigung

$$\frac{f\left(c + \left(j + \frac{1}{2}\right)h\right) - f\left(c + \left(j - \frac{1}{2}\right)h\right)}{\left(c + \left(j + \frac{1}{2}\right)h\right) - \left(c + \left(j - \frac{1}{2}\right)h\right)} = \frac{1}{h}\left(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}\right).$$

Induktionsschritt $s \to s + 1$: Mit $l := j - \frac{s}{2}$ ist

$$\begin{aligned} \frac{\mathrm{d}^s}{\mathrm{d}x^s} p_{j-\frac{s}{2}, \dots, j+\frac{s}{2}} = \frac{\mathrm{d}^s}{\mathrm{d}x^s} p_{l, \dots, l+s} &\overset{\text{Def.}}{=} s! f[x_l, \dots, x_{l+s}] \\ &\overset{(3.17)}{=} s! \frac{1}{x_l - x_{l+s}} \left(f[x_l, \dots, x_{l+s-1}] - f[x_{l+1}, \dots, x_{l+s}]\right) \\ &= s! \frac{1}{-sh} \left(\frac{1}{(s-1)!} \frac{\mathrm{d}^{s-1}}{\mathrm{d}x^{s-1}} p_{l, \dots, l+s-1} - \frac{1}{(s-1)!} \frac{\mathrm{d}^{s-1}}{\mathrm{d}x^{s-1}} p_{l+1, \dots, l+s}\right) \\ &\overset{IV}{=} -\frac{1}{h}\left(D^{s-1} f_{j-\frac{1}{2}} - D^{s-1} f_{j+\frac{1}{2}}\right) \overset{(4.6)}{=} D^s f_j. \end{aligned}$$

□

Proposition 4.8 (Nullstellen der Ableitungen des Restgliedes)
Zu $f \in C^{m+1}[a, b]$ und paarweise verschiedenen $x_0, \dots, x_m \in [a, b]$ sei p das zugehörige Interpolationspolynom und $f(x) = p(x) + R(x)$. Dann ist für $k = 0, \dots, m$ und $c \in [a, b]$ beliebig

$$f^{(k)}(c) = p^{(k)}(c) + R^{(k)}(c)$$

mit

$$R^{(k)}(c) = \frac{d^k}{dx^k}\left((x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x]\right)(c).$$

Somit besitzt für jedes $k = 0,\ldots,m$ die Ableitung $R^{(k)}$ mindestens $m+1-k$ Nullstellen $x_j^{(k)}$, $j = 0,\ldots,m-k$ mit der Eigenschaft

$$\min_{j=0,\ldots,m} x_j \le x_0^{(k)} \le x_1^{(k)} \le \cdots \le x_{m-k}^{(k)} \le \max_{j=0,\ldots,m} x_j.$$

Für jede Stelle $c \in G$ und jedes $k = 0,\ldots,m$ gibt es eine Zahl ξ, so dass die Darstellung gilt:

$$R^{(k)}(c) = \frac{1}{(m+1-k)!}(c-x_0^{(k)})\cdots(c-x_{m-k}^{(k)})f^{(m+1)}(\xi)$$

Beweis. Zu den Nullstellen von $R^{(k)}$: R ist laut Voraussetzung $(m+1)$-mal stetig differenzierbar und besitzt als Restglied des Interpolationspolynoms p die $m+1$ verschiedenen Nullstellen $x_0,\ldots,x_m$. Dabei sei o.B.d.A. $x_0^{(0)} < \cdots < x_m^{(0)}$. Nach dem Satz von Rolle existiert für $j = 0,\ldots,m-1$ jeweils ein $x_j^{(1)}$ mit $R'(x_j^{(1)}) = 0$ und

$$\min_{j=0,\ldots,m} x_j = x_0^{(0)} < x_0^{(1)} < x_1^{(0)} < x_1^{(1)} < \cdots x_{m-1}^{(1)} < x_m^{(0)} = \max_{j=0,\ldots,m} x_j.$$

Durch vollständige Induktion folgt die allgemeine Behauptung für $R^{(k)}$.

Zur Darstellung von $R^{(k)}(c)$: Für $c = x_j^{(k)}$ mit $j = 0,\ldots,m-k$ sind beide Seiten der Gleichung Null. Andernfalls definiere

$$\gamma := \left(c-x_0^{(k)}\right)\cdots\left(c-x_{m-k}^{(k)}\right) \neq 0$$
$$\varphi(x) := R^{(k)}(x) - \frac{1}{\gamma}\left(x-x_0^{(k)}\right)\cdots\left(x-x_{m-k}^{(k)}\right)R^{(k)}(c).$$

Offensichtlich hat φ die Nullstellen $x_0^{(k)},\ldots,x_{m-k}^{(k)},c$. Wiederholte Anwendung des Satzes von Rolle ergibt: $\varphi^{(m+1-k)}$ hat eine Nullstelle ξ mit

$$\min(c,x_0^{(k)},\ldots,x_{m-k}^{(k)}) \le \xi \le \max(c,x_0^{(k)},\ldots,x_{m-k}^{(k)}),$$

d.h. es gilt

$$0 = \varphi^{(m+1-k)}(\xi) = R^{(m+1)}(\xi) - \frac{1}{\gamma}(m+1-k)!R^{(k)}(c)$$
$$\Longrightarrow R^{(k)}(c) = \frac{\gamma}{(m+1-k)!}R^{(m+1)}(\xi).$$

Mit $R^{(m+1)} = f^{(m+1)} - p^{(m+1)} = f^{(m+1)}$ folgt die Behauptung. □

Proposition 4.9 (Restgliedabschätzung)
Mit $c \in [a,b]$ und $\rho = \max_{j=0,\ldots,m}|x_j - c|$ ergibt sich für die Restgliedabschätzung

$$\left|R^{(k)}(c)\right| \le \frac{\rho^{m+1-k}}{(m+1-k)!}\cdot\max_{a\le x\le b}\left|f^{(m+1)}(x)\right| \quad (\textit{für } k = 0,\ldots,m).$$

Beweis. Folgt unmittelbar aus dem vorherigen Satz. □

5 Numerische Integration

Definition 5.1 (Quadraturformel)
Sei die Funktion $f : I \mapsto \mathbb{R}$ mit $f \in C^{m+1}(I)$ gegeben, wobei $[a, b] \subset I \subset \mathbb{K}$ ist (wobei $\mathbb{K}$ für $\mathbb{R}$ oder $\mathbb{C}$ steht). Wir wählen Stützstellen $x_0, \ldots, x_m \in I$ *und* Gewichte $\alpha_0, \ldots, \alpha_m$ *so, dass*

$$Q(f) = (b-a)\sum_{j=0}^{m} \alpha_j f(x_j)$$

eine gute Approximation von

$$J(f) = \int_a^b f(x)\mathrm{d}x$$

ist. Dabei nennt man $Q(f)$ eine Quadraturformel. *Es gilt dann*

$$J(f) = Q(f) + E(f),$$

wobei $E(f)$ das zugehörige Restglied ist.

5.1 Interpolatorische Quadraturformeln

Proposition 5.2 (Lagrange-Darstellung)
Gegeben seien $m+1$ paarweise verschiedene Stützstellen $x_0, \ldots, x_m \in I$. Für die Funktion f lautet die interpolatorische Quadraturformel in der Lagrange-Darstellung

$$Q(f) = (b-a)\sum_{j=0}^{m} \alpha_j f(x_j) \quad \textit{mit} \quad \alpha_j := \frac{1}{b-a}\int_a^b l_j(x)\mathrm{d}x$$

und dem Restglied

$$E(f) = \int_a^b (x-x_0)\cdots(x-x_m) f[x_0, \ldots, x_m, x]\mathrm{d}x.$$

Beweis. Für f existiert genau ein Interpolationspolynom $p_{0,\ldots,m}(x)$ in der Lagrange-Darstellung mit dem zugehörigen Restglied $R(x)$. Somit kann f durch $f(x) = p_{0,\ldots,m}(x) + R(x)$ dargestellt werden. Wird über diese Darstellung integriert, so erhalten wir

$$\begin{aligned}
\int_a^b f(x)\mathrm{d}x &= \underbrace{\int_a^b p_{0,\ldots,m}(x)\mathrm{d}x}_{=:Q(f)} + \underbrace{\int_a^b R(x)\mathrm{d}x}_{=:E(f)} \\
&= \int_a^b \sum_{j=0}^{m} f(x_j) l_j(x)\mathrm{d}x + E(f) \\
&= \sum_{j=0}^{m} f(x_j)\int_a^b l_j(x)\mathrm{d}x + E(f) \\
&= (b-a)\sum_{j=0}^{m} f(x_j)\underbrace{\frac{1}{b-a}\int_a^b l_j(x)\mathrm{d}x}_{=:\alpha_j} + E(f).
\end{aligned}$$

Wir benutzen für $R(x)$ die Darstellung

$$\begin{aligned} R(x) &= (x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x] \\ \Longrightarrow\ E(f) &= \int_a^b (x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x]\mathrm{d}x. \end{aligned}$$

□

Proposition 5.3 (Newton-/Hermite-Darstellung)
Gegeben seien wieder $m+1$ (nicht notwendigerweise verschiedene) Stützstellen $x_0,\ldots,x_m \in I$. Für die Funktion f lautet die interpolatorische Quadraturformel in der Newton- bzw. Hermite-Darstellung

$$\begin{aligned} Q(f) &= \sum_{j=0}^{m} \tilde{\beta}_j f[x_0,\ldots,x_j] \\ &= \sum_{j=0}^{m} \beta_j (b-a)^{j+1} f[x_0,\ldots,x_j] \end{aligned}$$

mit $\tilde{\beta}_0 = b-a$, $\beta_0 = 1$ und

$$\tilde{\beta}_j := \int_a^b (x-x_0)\cdots(x-x_{j-1})\mathrm{d}x, \qquad \beta_j := \int_0^1 (z-z_0)\cdots(z-z_{j-1})\mathrm{d}z, \qquad z := \frac{x-a}{b-a}.$$

Das Restglied ist das gleiche wie bei der Lagrange-Darstellung:

$$E(f) = \int_a^b (x-x_0)\cdots(x-x_m)f[x_0,\ldots,x_m,x]\mathrm{d}x.$$

Beweis. Das Interpolationspolynom von f in der Newton- bzw. Hermite-Darstellung hat die Gestalt

$$p_{0,\ldots,m}(x) = f[x_0] + (x-x_0)f[x_0,x_1] + \cdots + (x-x_0)\cdots(x-x_{m-1})f[x_0,\ldots,x_m]$$

Integriert man analog dem Beweis zur Lagrange-Darstellung über das Interpolationspolynom, so kann man eine Quadraturformel angeben:

$$\begin{aligned} Q(f) &= \int_a^b p_{0,\ldots,m}(x)\mathrm{d}x \\ &= \int_a^b \sum_{j=0}^{m} (x-x_0)\cdots(x-x_{j-1})f[x_0,\ldots,x_j]\mathrm{d}x \\ &= \sum_{j=0}^{m} \underbrace{\int_a^b (x-x_0)\cdots(x-x_{j-1})\mathrm{d}x}_{=:\tilde{\beta}_j} f[x_0,\ldots,x_j] \\ &= \sum_{j=0}^{m} \tilde{\beta}_j f[x_0,\ldots,x_j] \\ &= \sum_{j=0}^{m} \underbrace{\beta_j (b-a)^{j+1}}_{:=\tilde{\beta}_j} f[x_0,\ldots,x_j]. \end{aligned}$$

Somit sind die Koeffizienten $\tilde{\beta}_0 = b - a$ bzw. $\beta_0 = 1$ und

$$\begin{aligned}\tilde{\beta}_j &= \int_a^b (x - x_0) \cdots (x - x_{j-1}) \mathrm{d}x \\ &= (b-a)^j \int_a^b \frac{x - x_0}{b-a} \cdots \frac{x - x_{j-1}}{b-a} \mathrm{d}x \\ &= (b-a)^{j+1} \underbrace{\int_0^1 (z - z_0) \cdots (z - z_{j-1}) \mathrm{d}z}_{=:\beta_j}.\end{aligned}$$

Der letzte Schritt ergibt sich durch Substitution

$$z = \frac{x-a}{b-a} \implies \frac{x - x_k}{b-a} = z - z_k \quad \text{und} \quad \mathrm{d}x = (b-a)\mathrm{d}z.$$

Zum Beweis für die Darstellung des Fehlers siehe Prop. 5.2. □

Proposition 5.4 (Fehlerabschätzung)
Man erhält für den Betrag des Restglieds $E(f)$ die Abschätzung

$$\begin{aligned}|E(f)| &\leq \frac{1}{(m+1)!} \cdot \int_a^b |(x - x_0) \cdots (x - x_m)| \mathrm{d}x \cdot \max_{s \in I} \left| f^{(m+1)}(s) \right| \\ &\leq \frac{1}{(m+1)!} \cdot (b-a)^{m+2} \cdot \int_0^1 |(z - z_0) \cdots (z - z_m)| \mathrm{d}z \cdot \max_{s \in I} \left| f^{(m+1)}(s) \right|.\end{aligned}$$

Sofern $(x - x_0) \cdots (x - x_m) \leq 0$ oder $(x - x_0) \cdots (x - x_m) \geq 0$ für alle $x \in [a, b]$ gilt (also kein Vorzeichenwechsel auftritt), kann man die Betragsstriche der Integranden alternativ auch aus dem Integral herausziehen und die Integrale analog Prop. 5.3 durch $\tilde{\beta}_{m+1}$ bzw. β_{m+1} ersetzen:

$$\begin{aligned}|E(f)| &\leq \frac{|\tilde{\beta}_{m+1}|}{(m+1)!} \cdot \max_{s \in I} \left| f^{(m+1)}(s) \right| \\ &\leq \frac{|\beta_{m+1}|}{(m+1)!} \cdot (b-a)^{m+2} \max_{s \in I} \left| f^{(m+1)}(s) \right|.\end{aligned}$$

Beweis. Die erste Form der obigen Fehlerabschätzung ergibt sich sofort durch Integration über die Restgliedabschätzung des Interpolationspolynoms (vgl. Kapitel zur Hermite-Interpolation)

$$\begin{aligned}|R(x)| &\leq \frac{1}{(m+1)!} \cdot |(x - x_0) \cdots (x - x_m)| \cdot \max_{s \in I} \left| f^{(m+1)}(s) \right| \\ \implies |E(f)| &\leq \frac{1}{(m+1)!} \cdot \int_a^b |(x - x_0) \cdots (x - x_m)| \mathrm{d}x \cdot \max_{s \in I} \left| f^{(m+1)}(s) \right|\end{aligned}$$

und die zweite Form durch anschließendes Ersetzen von x durch z. □

Bemerkung 5.5 (Spezialfall Polynome)
Für ein Polynom f vom Grad $\leq m$ ist f identisch mit dem Interpolationspolynom:

$$f = p_{0,\ldots,m} \implies R = 0 \implies E(f) = 0.$$

Daher haben die Quadraturformeln nach Lagrange, Newton und Hermite die wichtige Eigenschaft, dass sie für Polynome vom Grad $\leq m$ exakt sind, also gilt

$$\int_a^b f(x)dx = \sum_{j=0}^{m} \beta_j (b-a)^{j+1} f[x_0, \ldots, x_j].$$

Es ist möglich, dass eine solche Beziehung auch noch für Polynome höheren Grades $n \geq m$ gilt. Die größte ganze Zahl n mit dieser Eigenschaft nennt man den Genauigkeitsgrad der Quadraturformel.

Definition 5.6 (Genauigkeitsgrad)
$n \in \mathbb{N}$ *heißt* Genauigkeitsgrad *von* Q*, wenn für alle Polynome* p *mit* $\operatorname{grad}(p) \leq n$ *gilt*

$$E(p) = 0 \iff E(1) = E(x) = \cdots = E(x^n) = 0 \quad \textit{und} \quad E(x^{n+1}) \neq 0.$$

Bemerkung 5.7
Für die Lagrange-Darstellung gilt immer die Beziehung

$$\sum_{j=0}^{m} \alpha_j = 1 \iff E(1) = 0$$

wegen

$$J(1) = Q(1) \implies (b-a) = (b-a) \sum_{j=0}^{m} \alpha_j$$

5.2 Spezielle Quadraturformeln

Proposition 5.8 (Sehnentrapezformel)
Man ersetzt die Kurve $f(x)$ *durch die Verbindungsgerade zwischen den Punkten* $(a, f(a))$ *und* $(b, f(b))$ *- also durch die Sehne - und erhält somit ein Trapez. Die Quadraturformel lautet*

$$Q(f) = \frac{b-a}{2} \left(f(a) + f(b)\right)$$

mit der Fehlerabschätzung

$$|E(f)| \leq \frac{1}{12}(b-a)^3 \max_{x \in [a,b]} |f''(x)|\,.$$

Beweis. Wir benutzen die interpolatorische Quadraturformel in der Hermite-Darstellung. Mit $x_0 = a$, $x_1 = b$, $m = 1$ werden $z_0 = 0$, $z_1 = 1$ und $\beta_j = \frac{\tilde{\beta}_j}{(b-a)^{j+1}}$. Man erhält

$$\beta_0 = 1, \qquad \beta_1 = \int_0^1 (z-0)\mathrm{d}z = \frac{1}{2}, \qquad \beta_2 = \int_0^1 (z-0)(z-1)\mathrm{d}z = -\frac{1}{6}.$$

Weiter ist

$$f[a] = f(a) \quad \text{und} \qquad f[a,b] = \frac{f(b) - f(a)}{(b-a)}.$$

Damit erhalten wir für die Quadraturformel

$$\begin{aligned} Q(f) &= \beta_0(b-a)f[a] + \beta_1(b-a)^2 f[a,b] \\ &= (b-a)f[a] + \frac{1}{2}(b-a)^2 f[a,b] \\ &= (b-a)\left(f(a) + \frac{1}{2}\left(f(b) - f(a)\right)\right) \\ &= (b-a)f(a) + \frac{1}{2}(b-a)f(b) - \frac{1}{2}(b-a)f(a) \\ Q(f) &= \frac{b-a}{2}\left(f(a) + f(b)\right). \end{aligned}$$

Für die Fehlerabschätzung erhalten wir mit Proposition 5.4 für jedes $f \in C^2[a,b]$

$$\begin{aligned} \int_0^1 |(z-z_0)(z-z_1)|dz = \left|\int_0^1 (z-z_0)(z-z_1)dz\right| = |\beta_2| = \frac{1}{6} \\ \implies |E(f)| \leq \frac{1}{2\cdot 6}(b-a)^3 \max_{x\in[a,b]} |f''(x)|. \end{aligned}$$

Der Genauigkeitsgrad ist $n = 1$, da $f'' = 0$ für Polynome mit $\operatorname{grad} f \leq 1$ gilt. □

Proposition 5.9 (Tangententrapezformel)
Man legt an die Kurve $f(x)$ im Punkt c in der Mitte des Intervalls $[a,b]$ die Tangente und erhält so wieder ein Trapez. Die Quadraturformel lautet

$$Q(f) = (b-a)f(c) \quad \left(\text{für } c = \frac{a+b}{2}\right)$$

mit der Fehlerabschätzung

$$|E(f)| \leq \frac{1}{24}(b-a)^3 \max_{x\in[a,b]} |f''(x)|.$$

Beweis. Wir benutzen wieder die interpolatorische Quadraturformel in der Hermite-Darstellung. Mit $x_0 = x_1 = c$, $m = 1$ werden $z_0 = z_1 = 1/2$ und

$$\beta_0 = 1, \qquad \beta_1 = \int_0^1 \left(z - \frac{1}{2}\right) \mathrm{d}z = 0, \qquad \beta_2 = \int_0^1 \left(z - \frac{1}{2}\right)^2 \mathrm{d}z = \frac{1}{3}\left(z - \frac{1}{2}\right)^3 \Bigg|_0^1 = \frac{1}{12}.$$

Weiter ist

$$f[c] = f(c), \qquad f[c,c] = f'(c) \qquad \text{(Spezialfall: } x_0 = x_1 = c\text{)}.$$

Die Quadraturformel ist somit

$$\begin{aligned} Q(f) &= \beta_0(b-a)f[c] + \beta_1(b-a)^2 f[c,c] \\ &= (b-a)f(c) + 0\cdot(b-a)^2 f'(c) \\ &= (b-a)f(c). \end{aligned}$$

Für jedes $f \in C^2[a,b]$ gilt wie oben die Abschätzung

$$|E(f)| \leq \frac{1}{12 \cdot 2}(b-a)^3 \max_{x\in[a,b]} |f''(x)|$$

Der Genauigkeitsgrad ist wie oben $n = 1$. □

Proposition 5.10 (Simpsonformel)
Interpoliert man die Funktion $f(x)$ mittels eines quadratischen Polynoms in den Punkten a, b, c (c liegt in der Mitte von $[a,b]$), dann erhält man die Simpsonsche Formel

$$Q(f) = \frac{b-a}{6}\left(f(a) + 4f(c) + f(b)\right) \quad \left(\text{für } c = \frac{a+b}{2}\right)$$

mit der Fehlerabschätzung

$$|E(f)| \leq \frac{1}{2880}(b-a)^5 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|.$$

Beweis. Erneut benutzen wir die interpolatorische Quadraturformel in der Hermite-Darstellung. Mit $x_0 = a$, $x_1 = b$, $x_2 = x_3 = c$, $m = 3$ werden $z_0 = 0$, $z_1 = 1$, $z_2 = z_3 = 1/2$ und

$$\beta_0 = 1, \qquad \beta_1 = \int_0^1 z\mathrm{d}z = \frac{1}{2}, \qquad \beta_2 = \int_0^1 z(z-1)\mathrm{d}z = -\frac{1}{6}$$
$$\beta_3 = \int_0^1 z(z-1)\left(z-\frac{1}{2}\right)\mathrm{d}z = 0, \qquad \beta_4 = \int_0^1 z(z-1)\left(z-\frac{1}{2}\right)^2\mathrm{d}z = -\frac{1}{120}.$$

Weiter ist

$$f[a] = f(a), \qquad f[a,b] = \frac{f(b)-f(a)}{(b-a)}, \qquad f[a,b,c] = \frac{2(f(a)+f(b)-2f(c))}{(b-a)^2}.$$

Nebenrechnung zur dritten Formel:

$$\begin{aligned}
f[a,b,c] &= \frac{f[a,b]-f[b,c]}{a-c} \\
&= \frac{1}{a-c}\left(\frac{f(a)-f(b)}{a-b} - \frac{f(b)-f(c)}{b-c}\right) \quad \text{mit} \quad c = \frac{a+b}{2} \\
&= \frac{1}{a-\frac{a+b}{2}}\left(\frac{f(a)-f(b)}{a-b} - \frac{f(b)-f(c)}{b-\frac{a+b}{2}}\right) \\
&= \frac{2}{a-b}\left(\frac{f(a)-f(b)}{a-b} - \frac{2f(b)-2f(c)}{b-a}\right) \\
&= \frac{2}{a-b}\left(\frac{f(a)-f(b)}{a-b} - \frac{2f(c)-2f(b)}{a-b}\right) \\
&= \frac{2}{a-b}\left(\frac{f(a)-f(b)-2f(c)+2f(b)}{a-b}\right) \\
f[a,b,c] &= \frac{2}{(b-a)^2}\left(f(a)-2f(c)+f(b)\right)
\end{aligned}$$

Damit wird die Quadraturformel zu

$$\begin{aligned}
Q(f) &= \beta_0(b-a)f[a] + \beta_1(b-a)^2 f[a,b] + \beta_2(b-a)^3 f[a,b,c] + \beta_3(b-a)^4 f[a,b,c,c] \\
&= (b-a)f(a) + \frac{1}{2}(b-a)^2\frac{f(b)-f(a)}{b-a} - \frac{1}{6}(b-a)^3\frac{2(f(a)-2f(c)+f(b))}{(b-a)^2} + 0 \\
&= (b-a)f(a) + \frac{1}{2}(b-a)(f(b)-f(a)) - \frac{2}{6}(b-a)(f(a)-2f(c)+f(b)) \\
&= (b-a)\left(f(a) + \frac{3}{6}(f(b)-f(a)) - \frac{2}{6}(f(a)-2f(c)+f(b))\right) \\
&= \frac{b-a}{6}(6f(a)+3f(b)-3f(a)-2f(a)-2f(b)+4f(c)) \\
Q(f) &= \frac{b-a}{6}(f(a)+f(b)+4f(c)).
\end{aligned}$$

Für $f \in C^4[a,b]$ folgt dann die Restgliedabschätzung

$$\int_0^1 \left|z(z-1)\left(z-\frac{1}{2}\right)^2\right| \mathrm{d}z = \left|\int_0^1 z(1-z)\left(z-\frac{1}{2}\right)^2 \mathrm{d}z\right| = |\beta_4| = \frac{1}{120}$$
$$\implies |E(f)| \le \frac{1}{4!\cdot 120}(b-a)^5 \max_{x\in[a,b]}\left|f^{(4)}(x)\right|.$$

Somit ist der Genauigkeitsgrad $n = 3$. □

Proposition 5.11 (Bessel-Formel)
Diese Formel zeigt, dass man auch Punkte außerhalb des Integrationsintervalls $[a,b]$ heranziehen kann. Die Quadraturformel lautet

$$Q(f) = \frac{b-a}{24}\left(-f(a-h) + 13f(a) + 13f(b) - f(b+h)\right) \quad (\textit{für } h = b-a)$$

mit der Fehlerabschätzung

$$|E(f)| \le \frac{11}{720}(b-a)^5 \max_{x\in[a-h,b+h]}\left|f^{(4)}(x)\right|.$$

Beweis. Mit $x_0 = a$, $x_1 = b$, $x_2 = a-h$, $x_3 = b+h$, $m = 3$ werden

$$z_j = \frac{x_j - a}{b-a} \implies z_0 = 0, \qquad z_1 = 1, \qquad z_2 = \frac{-h}{b-a} = -1, \qquad z_3 = \frac{b+h-a}{b-a} = \frac{2(b-a)}{b-a} = 2$$

und

$$\beta_0 = 1, \qquad \beta_j = \int_0^1 (z-z_0)\cdots(z-z_{j-1})\mathrm{d}z$$
$$\implies \beta_0 = 1, \qquad \beta_1 = \frac{1}{2}, \qquad \beta_2 = -\frac{1}{6}, \qquad \beta_3 = -\frac{1}{4}, \qquad \beta_4 = \frac{11}{30}.$$

Die Rechnung für die Quadraturformel ist recht aufwändig wird hier nicht durchgeführt. Für Funktionen $f \in C^4[a-h,b+h]$ gilt die Restgliedabschätzung

$$|E(f)| \le \frac{11}{4!\cdot 30}(b-a)^5 \max_{x\in[a-h,b+h]}\left|f^{(4)}(x)\right|.$$

Der Genauigkeitsgrad ist $n = 3$. □

Proposition 5.12 (Hermitesche Formel)
In der Hermitschen Formel tritt neben der Funktion f auch ihre erste Ableitung auf. Die Quadraturformel lautet

$$Q(f) = \frac{b-a}{2}\left(f(a)+f(b)\right) + \frac{(b-a)^2}{12}\left(f'(a)-f'(b)\right)$$

und hat die Fehlerabschätzung

$$|E(f)| \le \frac{1}{720}(b-a)^5 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|.$$

Beweis. Mit $x_0 = x_2 = a$, $x_1 = x_3 = b$, $m = 3$ werden $z_0 = z_2 = 0$, $z_1 = z_3 = 1$ und

$$\beta_0 = 1, \qquad \beta_1 = \frac{1}{2}, \qquad \beta_2 = -\frac{1}{6}, \qquad \beta_3 = -\frac{1}{12}, \qquad \beta_4 = \frac{1}{30}.$$

Weiter ist nach der Hermite-Relation 3.23

$$\begin{aligned} & f[a,b,b] = f[a,b,a] + f[a,b,a,b](b-a) \\ \Longrightarrow\ & f[a,b,a,b] && = \frac{1}{h}\left(f[a,b,b]-f[a,b,a]\right). \end{aligned}$$

Außerdem rechnet man leicht nach, dass gilt:

$$f[a,b,a] + f[a,b,b] = \frac{1}{h}\left(f'(b)-f'(a)\right)$$

Somit folgt für die Quadraturformel

$$\begin{aligned} Q(f) &= (b-a)\left(f(a) + \frac{1}{2}(b-a)f[a,b] - \frac{1}{6}(b-a)^2 f[a,b,a] - \frac{1}{12}(b-a)^3 f[a,b,a,b]\right) \\ &= (b-a)\left(f(a) + \frac{1}{2}(b-a)f[a,b] - \frac{1}{6}(b-a)^2 f[a,b,a] - \frac{1}{12}(b-a)^3 \frac{1}{h}\left(f[a,b,b]-f[a,b,a]\right)\right) \\ &= (b-a)\left(f(a) + \frac{1}{2}(b-a)f[a,b] - \frac{1}{12}(b-a)^2\left(2f[a,b,a] + f[a,b,b] - f[a,b,a]\right)\right) \\ &= (b-a)\left(f(a) + \frac{1}{2}(b-a)f[a,b] - \frac{1}{12}(b-a)^2\left(f[a,b,a] + f[a,b,b]\right)\right) \\ &= (b-a)\left(f(a) + \frac{1}{2}(b-a)f[a,b] - \frac{1}{12}(b-a)^2 \frac{1}{h}\left(f'(b)-f'(a)\right)\right) \\ &= (b-a)\left(f(a) + \frac{1}{2}\left(f(b)-f(a)\right) - \frac{1}{12}(b-a)\left(f'(b)-f'(a)\right)\right) \\ &= (b-a)\left(\frac{2f(a)+f(b)-f(a)}{2} - \frac{(b-a)}{12}\left(f'(b)-f'(a)\right)\right) \\ Q(f) &= \frac{(b-a)}{2}\left(f(a)+f(b)\right) + \frac{(b-a)^2}{12}\left(f'(a)-f'(b)\right). \end{aligned}$$

Für Funktionen $f \in C^4[a-h, b+h]$ gilt die Restgliedabschätzung

$$|E(f)| \le \frac{1}{4!\cdot 30}(b-a)^5 \max_{x\in[a-h,b+h]} \left|f^{(4)}(x)\right| \quad \text{und} \qquad n = 3.$$

□

Proposition 5.13 (Formel von Euler-MacLaurin)
Für eine Funktion $f \in C^{2m}(I)$ mit $m \geq 2$ kann man die Quadraturformel

$$Q(f) = \frac{(b-a)}{2}\left(f(a)+f(b)\right) + \sum_{l=1}^{m-1} \frac{B_{2l}}{(2l)!}(b-a)^{2l}\left(f^{(2l-1)}(a) - f^{(2l-1)}(b)\right)$$

angeben, wobei B_j die Bernoulli-Zahlen *sind. Diese sind über eine Rekursion oder (alternativ) über die* Bernoulli-Polynome *$B_j(x)$ wie folgt definiert sind:*

$$B_0 := 1, \qquad B_j := -\frac{1}{j+1}\sum_{k=0}^{j-1}\binom{j+1}{k}B_k \qquad \textit{bzw.} \qquad B_j := B_j(0).$$

Die ersten Bernoulli-Zahlen sind

$$B_1 = -\frac{1}{2}, \qquad B_2 = \frac{1}{6}, \qquad B_3 = 0, \qquad B_4 = -\frac{1}{30}, \qquad B_5 = 0, \qquad B_6 = \frac{1}{42}, \qquad B_{2j+1} = 0 \;\forall\, j \geq 1.$$

Für die Fehlerabschätzung der Euler-MacLaurin-Formel gilt:

$$|E(f)| \leq \frac{(b-a)^{2m+1}}{(2m)!} \cdot C_m \max_{x\in[a,b]}\left|f^{(2m)}(x)\right| \qquad \textit{wobei} \qquad C_m : \|B_{2m}\|_\infty \leq C_m.$$

Die Euler-MacLaurin-Formel hat den Genauigkeitsgrad $n = 2m-1$.

Bemerkung 5.14 (Spezialfälle von Euler-MacLaurin)
Für $m = 1$ erhält man die Sehnentrapezformel und für $m = 2$ die Hermiteformel.

5.3 Summierte Quadraturformeln

Summierte Quadraturformeln erhält man, wenn man das betrachtete Intervall in Teilintervalle der Länge h zerlegt und dann in jedem dieser Teilintervalle die Quadraturformeln anwendet.

$$[a,b] = \bigcup_{r=1}^{N} \underbrace{[a_r, b_r]}_{:=I_r}, \qquad a_1 = a, \qquad b_N = b, \qquad a_{r+1} = b_r, \qquad r = 1,\ldots,N-1$$

Mit Quadraturformeln Q_r und zugehörigen Restgliedern E_r für Integrale über diese Teilintervalle

$$\int_{a_r}^{b_r} f(x)\mathrm{d}x = Q_r(f) + E_r(f) \quad (\text{für } r = 1,\ldots,N)$$

erhält man

$$\int_a^b f(x)\mathrm{d}x = \sum_{r=1}^{N}\int_{a_r}^{b_r} f(x)\mathrm{d}x = Q(f) + E(f)$$

mit den Abkürzungen

$$Q(f) = \sum_{r=1}^{N} Q_r(f), \qquad E(f) = \sum_{r=1}^{N} E_r(f).$$

Sei im Folgenden $f \in C^{m+1}(I)$ und die Teilintervalle I_r haben alle die gleiche Länge h.

$$h = b_r - a_r = \frac{b-a}{N} \quad (\text{für } r = 1, \ldots, N)$$

Damit folgt die Restgliedabschätzung

$$|E_r(f)| \leq \frac{C}{(m+1)!} h^{m+2} \max_{x \in I_r} \left| f^{(m+1)}(x) \right|,$$

wobei

$$C = \begin{cases} \int_0^1 |(z-z_0)\cdots(z-z_m)| \mathrm{d}z & \text{immer} \\ |\beta_{m+1}| & \text{ohne Vorzeichenwechsel.} \end{cases}$$

Damit ergibt sich unter der Verwendung von

$$\sum_{r=1}^{N} h^{m+2} = h^{m+1} \sum_{r=1}^{N} h = h^{m+1}(b-a)$$

für die Fehlerabschätzung über das gesamte Intervall

$$|E(f)| \leq \sum_{r=1}^{N} |E_r(f)| \leq (b-a) \frac{C}{(m+1)!} h^{m+1} \max_{x \in I} \left| f^{(m+1)}(x) \right|.$$

Bemerkung 5.15 (Notation)
Im Folgenden sei $x_r = a + rh$ mit $r = 0, \ldots, N$ und $h = \frac{b-a}{N}$.

Proposition 5.16 (Summierte Sehnentrapezformel)
Die summierte Sehnentrapezformel

$$Q(f) = \sum_{r=1}^{N} \frac{b_r - a_r}{2} \left(f(a_r) + f(b_r) \right) = h \left(\frac{1}{2} f(\underbrace{x_0}_{=a}) + \sum_{r=1}^{N-1} f(x_r) + \frac{1}{2} f(\underbrace{x_N}_{=b}) \right)$$

hat für $f \in C^2[a,b]$ Restgliedabschätzung

$$|E(f)| \leq (b-a) \frac{h^2}{12} \max_{x \in [a,b]} |f''(x)|.$$

Proposition 5.17 (Summierte Tangententrapezformel)
Die summierte Tangententrapezformel der Funktion f lautet

$$Q(f) = \sum_{r=1}^{N} (b_r - a_r) f\left(\frac{a_r + b_r}{2} \right) = h \sum_{r=1}^{N} f\left(\frac{x_{r-1} + x_r}{2} \right)$$

und hat für $f \in C^2[a,b]$ die Restgliedabschätzung

$$|E(f)| \leq (b-a) \frac{h^2}{24} \max_{x \in [a,b]} |f''(x)|.$$

Proposition 5.18 (Summierte Simpsonformel)
Die summierte Simpsonformel lautet

$$\begin{aligned} Q(f) &= \sum_{r=1}^{N} \frac{b_r - a_r}{6} \left(f(a_r) + 4f\left(\frac{a_r + b_r}{2}\right) + f(b_r) \right) \\ &= \frac{h}{3} \left(\frac{1}{2} f(x_0) + \sum_{r=1}^{N-1} f(x_r) + 2 \sum_{r=1}^{N} f\left(\frac{x_{r-1} + x_r}{2}\right) + \frac{1}{2} f(x_N) \right). \end{aligned}$$

Damit ist $Q(f)$ das gewichtete Mittel aus der summierten Sehnentrapezformel $Q_S(f)$ und der summierten Tangententrapezformel $Q_T(f)$ in der Form

$$Q(f) = \frac{1}{3} \left(Q_S(f) + 2Q_T(f) \right).$$

Die Restgliedabschätzung für $f \in C^4[a,b]$ lautet

$$|E(f)| \leq (b-a) \frac{h^4}{2880} \max_{x \in [a,b]} \left| f^{(4)}(x) \right|.$$

Proposition 5.19 (Summierte Hermitesche Formel)
Die summierte Hermitesche Formel lautet

$$\begin{aligned} Q(f) &= \sum_{r=1}^{N} \left(\frac{b_r - a_r}{2} \left(f(a_r) + f(b_r) \right) + \frac{(b_r - a_r)^2}{12} \left(f'(a_r) - f'(b_r) \right) \right) \\ &= h \left(\frac{1}{2} f(x_0) + \sum_{r=1}^{N-1} f(x_r) + \frac{1}{2} f(x_N) \right) + \frac{h^2}{12} \left(f'(x_0) - f'(x_N) \right) \end{aligned}$$

und hat für $f \in C^4[a,b]$ die Restgliedabschätzung

$$|E(f)| \leq (b-a) \frac{h^4}{720} \max_{x \in [a,b]} \left| f^{(4)}(x) \right|.$$

Proposition 5.20 (Summierte Formel von Euler-MacLaurin)
Die summierte Euler-MacLaurin-Formel ist

$$\begin{aligned} Q(f) &= \sum_{r=1}^{N} \left(\frac{b_r - a_r}{2} \left(f(a_r) + f(b_r) \right) + \sum_{l=1}^{m-1} \frac{B_{2l}}{(2l)!} (b_r - a_r)^{2l} \left(f^{(2l-1)}(a_r) - f^{(2l-1)}(b_r) \right) \right) \\ &= \underbrace{h \left(\frac{1}{2} f(x_0) + \sum_{r=1}^{N-1} f(x_r) + \frac{1}{2} f(x_N) \right)}_{= \textit{Summierte Sehnentrapezformel}} + \sum_{l=1}^{m-1} h^{2l} \underbrace{\frac{B_{2l}}{(2l)!} \left(f^{(2l-1)}(x_0) - f^{(2l-1)}(x_N) \right)}_{=: c_l}. \end{aligned}$$

Diese Formel geht durch ein Korrekturglied aus der summierten Sehnentrapezformel hervor. Die zugehörige Restgliedabschätzung für $f \in C^{2m}[a,b]$ lautet

$$|E(f)| \leq (b-a) \frac{C_m}{(2m)!} h^{2m} \max_{x \in [a,b]} \left| f^{(2m)}(x) \right|.$$

5.4 Übersicht

Spezielle Quadraturformeln

Es sei $c = \frac{a+b}{2}$, $h = b - a$ und C_m so, dass $\|B_{2m}\|_\infty \leq C_m$.

	Quadraturformel	Fehlerabschätzung	Genauigkeitsgrad
Sehnentrapezformel	$Q(f) = \frac{b-a}{2}\left(f(a) + f(b)\right)$	$\lvert E(f)\rvert \leq \frac{1}{12}(b-a)^3 \max\limits_{x\in[a,b]} \lvert f''(x)\rvert$	$n = 1$
Tangententrapezformel	$Q(f) = (b-a)f(c)$	$\lvert E(f)\rvert \leq \frac{1}{24}(b-a)^3 \max\limits_{x\in[a,b]} \lvert f''(x)\rvert$	$n = 1$
Simpson-Formel	$Q(f) = \frac{b-a}{6}\left(f(a) + 4f(c) + f(b)\right)$	$\lvert E(f)\rvert \leq \frac{1}{2880}(b-a)^5 \max\limits_{x\in[a,b]} \left\lvert f^{(4)}(x)\right\rvert$	$n = 3$
Bessel-Formel	$Q(f) = \frac{b-a}{24}\left(-f(a-h) + 13f(a) + 13f(b) - f(b+h)\right)$	$\lvert E(f)\rvert \leq \frac{11}{720}(b-a)^5 \max\limits_{x\in[a,b]} \left\lvert f^{(4)}(x)\right\rvert$	$n = 3$
Hermite-Formel	$Q(f) = \frac{b-a}{2}\left(f(a) + f(b)\right) + \frac{(b-a)^2}{12}\left(f'(a) - f'(b)\right)$	$\lvert E(f)\rvert \leq \frac{1}{720}(b-a)^5 \max\limits_{x\in[a,b]} \left\lvert f^{(4)}(x)\right\rvert$	$n = 3$
Formel von Euler-MacLaurin	$Q(f) = \frac{(b-a)}{2}\left(f(a) + f(b)\right) + \sum\limits_{l=1}^{m-1} \frac{B_{2l}}{(2l)!}(b-a)^{2l}\left(f^{(2l-1)}(a) - f^{(2l-1)}(b)\right)$	$\lvert E(f)\rvert \leq \frac{(b-a)^{2m+1}}{(2m)!} C_m \max\limits_{x\in[a,b]} \left\lvert f^{(2m)}(x)\right\rvert$	$n = 2m - 1$

Summierte Quadraturformeln

Setze $x_r = a + rh$ für $r = 0, \ldots, N$ und $h = b_r - a_r = \frac{b-a}{N}$ für $r = 1, \ldots, N$.

	Quadraturformel	Fehlerabschätzung
Summ. Sehnentrapezf.	$Q(f) = \sum\limits_{r=1}^{N} \frac{b_r - a_r}{2}\left(f(a_r) + f(b_r)\right) = h\left(\frac{1}{2}f(\underbrace{x_0}_{=a}) + \sum\limits_{r=1}^{N-1} f(x_r) + \frac{1}{2}f(\underbrace{x_N}_{=b})\right)$	$\lvert E(f)\rvert \leq (b-a)\frac{h^2}{12} \max\limits_{x\in[a,b]} \lvert f''(x)\rvert$
Summ. Tangententrapezf.	$Q(f) = \sum\limits_{r=1}^{N}(b_r - a_r) f\left(\frac{a_r+b_r}{2}\right) = h\sum\limits_{r=1}^{N} f\left(\frac{x_{r-1}+x_r}{2}\right)$	$\lvert E(f)\rvert \leq (b-a)\frac{h^2}{24} \max\limits_{x\in[a,b]} \lvert f''(x)\rvert$
Summ. Simpson-F.	$Q(f) = \frac{h}{3}\left(\frac{1}{2}f(x_0) + \sum\limits_{r=1}^{N-1} f(x_r) + 2\sum\limits_{r=1}^{N} f\left(\frac{x_{r-1}+x_r}{2}\right) + \frac{1}{2}f(x_N)\right) = \frac{1}{3}\left(Q_S(f) + 2Q_T(f)\right)$	$\lvert E(f)\rvert \leq (b-a)\frac{h^4}{2880} \max\limits_{x\in[a,b]} \left\lvert f^{(4)}(x)\right\rvert$
Summ. Hermite-F.	$Q(f) = h\left(\frac{1}{2}f(x_0) + \sum\limits_{r=1}^{N-1} f(x_r) + \frac{1}{2}f(x_N)\right) + \frac{h^2}{12}\left(f'(x_0) - f'(x_N)\right)$	$\lvert E(f)\rvert \leq (b-a)\frac{h^4}{720} \max\limits_{x\in[a,b]} \left\lvert f^{(4)}(x)\right\rvert$
Summ. F. v. Euler-MacLaurin	$Q(f) = \underbrace{h\left(\frac{1}{2}f(x_0) + \sum\limits_{r=1}^{N-1} f(x_r) + \frac{1}{2}f(x_N)\right)}_{=\text{Summierte Sehnentrapezformel}} + \sum\limits_{l=1}^{m-1} h^{2l} \underbrace{\frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(x_0) - f^{(2l-1)}(x_N)\right)}_{=:c_l}$	$\lvert E(f)\rvert \leq (b-a)\frac{C_m}{(2m)!}h^{2m} \max\limits_{x\in[a,b]} \left\lvert f^{(2m)}(x)\right\rvert$

5.5 Romberg-Integration

Die Summenformel nach Euler-MacLaurin gibt eine Entwicklung nach Potenzen der Schrittweite h für den Fehler bei der Approximation des Integrals durch die Sehnentrapezregel. Durch fortgesetzte Halbierung der Schrittweite und geeignete Linearkombination der zugehörigen Näherungen kann man jeweils den führenden Fehlerterm eliminieren und Näherungen höherer Genauigkeit erzielen.

Proposition 5.21 (Algorithmus der Romberg-Integration)
Für eine Funktion $f \in C^{2n}[a,b]$ bezeichne $S(f)$ die summierte Sehnentrapezformel mit Schrittweite $h = (b-a)/N$ und Anzahl N an Stützstellen. Mit $S_j(f)$ wird die summierte Sehnentrapezformel zur Schrittweite $h_j = h/2^j$ und der Anzahl $N_j = 2^j N$ an Stützstellen bezeichnet. Die Romberg-Integration erfolgt durch die Rekursion

$$S_j^{(0)} := S_j(f), \qquad S_j^{(k+1)}(f) := \frac{1}{4^{k+1}-1}\left\{4^{k+1} S_{j+1}^{(k)}(f) - S_j^{(k)}(f)\right\}$$

für $k = 0, \ldots, n-2$ und $j = 0, 1, 2, \ldots$. Dabei sind die Summen $S_j^{(k)}(f)$ eine Näherung für das Integral

$$\int_a^b f(x)\mathrm{d}x = S_j^{(k)}(f) + \sum_{l=k+1}^{n-1} c_l^{(k)} h_j^{2l} + E_j^{(k)}(f) = S_j^{(k)}(f) + \mathcal{O}\left(h_j^{2(k+1)}\right)$$

mit $k = 0, \ldots, n-1$ und $j = 0, 1, 2, \ldots$ und geeigneten $c_l^{(k)}$ sowie dem Restglied $E_j^{(k)}(f)$.

Beweis. Wir verwenden die summierten Euler-MacLaurinschen Quadraturformeln zu den Schrittweiten h_j und h_{j+1}:

$$\begin{aligned}
\int_a^b f(x)\mathrm{d}x &= \underbrace{h_j\left(\frac{1}{2}f(a) + \sum_{r=1}^{N_j-1} f(a+rh_j) + \frac{1}{2}f(b)\right)}_{=S_j(f)} + \sum_{l=1}^{n-1} h_j^{2l} \underbrace{\frac{B_{2l}}{(2l)!}\left(f^{(2l-1)}(a) - f^{(2l-1)}(b)\right)}_{=:c_l} + E_j(f) \\
&= S_j(f) + \sum_{l=1}^{n-1} h_j^{2l} c_l + E_j(f) = S_j(f) + \mathcal{O}\left(h_j^2\right), \\
\int_a^b f(x)\mathrm{d}x &= S_{j+1}(f) + \sum_{l=1}^{n-1} h_{j+1}^{2l} c_l + E_{j+1}(f) = S_{j+1}(f) + \mathcal{O}\left(h_{j+1}^2\right)
\end{aligned}$$

Dabei wurde bedacht, dass $E_j(f) = \mathcal{O}\left(h_j^{2n}\right)$ gilt. Nun multiplizieren wir die untere der beiden Gleichungen mit 4, ziehen die obere von ihr ab und dividieren schließlich durch 3. Somit erhalten wir

$$\begin{aligned}
\int_a^b f(x)\mathrm{d}x &= \underbrace{\frac{1}{3}\left(4S_{j+1}(f) - S_j(f)\right)}_{=:S_j^{(1)}(f)} + \sum_{l=1}^{n-1} c_l \frac{1}{3} \underbrace{\left(4h_{j+1}^{2l} - h_j^{2l}\right)}_{=\left(4\left(\frac{h_j}{2}\right)^{2l} - h_j^{2l}\right)} + \underbrace{\frac{1}{3}\left(4E_{j+1}(f) - E_j(f)\right)}_{=:E_j^{(1)}(f)} \\
&= S_j^{(1)}(f) + \sum_{l=1}^{n-1} \underbrace{c_l \frac{1}{3}\left(\frac{1}{2^{2l-2}} - 1\right)}_{=:c_l^{(1)}} h_j^{2l} + E_j^{(1)}(f) \\
&= S_j^{(1)}(f) + \sum_{l=1}^{n-1} c_l^{(1)} h_j^{2l} + E_j^{(1)}(f).
\end{aligned}$$

Man beachte, dass $c_1^{(1)} = 0$. Somit haben wir erreicht, dass in der Summe die niedrigste Potenz von h_j mit $l = 2$ auftritt:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &= S_j^{(1)}(f) + \sum_{l=2}^{n-1} c_l^{(1)} h_j^{2l} + E_j^{(1)}(f) \\ &= S_j^{(1)}(f) + \mathcal{O}\left(h_j^4\right).\end{aligned}$$

Nun fehlt noch die Induktion $k \to k+1$. Wir gehen analog dem Fall $k = 0$ oben vor und verwenden die beiden Formeln zu den Schrittweiten h_j und h_{j+1}

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x =& S_j^{(k)}(f) + \sum_{l=k+1}^{n-1} c_l^{(k)} h_j^{2l} + E_j^{(k)}(f), \\ \int_a^b f(x)\mathrm{d}x =& S_{j+1}^{(k)}(f) + \sum_{l=k+1}^{n-1} c_l^{(k)} h_{j+1}^{2l} + E_{j+1}^{(k)}(f).\end{aligned}$$

Wir multiplizieren die untere der beiden Gleichungen mit 4^{k+1}, ziehen die obere von ihr ab und dividieren schließlich durch $4^{k+1} - 1$:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x =& \underbrace{\frac{1}{4^{k+1}-1}\left(4^{k+1} S_{j+1}^{(k)}(f) - S_j^{(k)}(f)\right)}_{=:S_j^{(k+1)}(f)} + \sum_{l=k+1}^{n-1} c_l^{(k)} \frac{1}{4^{k+1}-1} \underbrace{\left(4^{k+1} h_{j+1}^{2l} - h_j^{2l}\right)}_{=\left(4^{k+1}\left(\frac{h_j}{2}\right)^{2l} - h_j^{2l}\right)} \\ &+ \underbrace{\frac{1}{4^{k+1}-1}\left(4^{k+1} E_{j+1}^{(k)}(f) - E_j^{(k)}(f)\right)}_{=:E_j^{(k+1)}(f)} \\ =& S_j^{(k+1)}(f) + \sum_{l=k+1}^{n-1} \underbrace{c_l^{(k)} \frac{1}{4^{k+1}-1}\left(\frac{1}{4^{l-(k+1)}} - 1\right)}_{=:c_l^{(k+1)}} h_j^{2l} + E_j^{(k+1)}(f) \\ =& S_j^{(k+1)}(f) + \sum_{l=k+1}^{n-1} c_l^{(k+1)} h_j^{2l} + E_j^{(k+1)}(f)\end{aligned}$$

Erneut beachte man, dass $c_{k+1}^{(k+1)} = 0$, wodurch der Term mit $h_j^{2(k+1)}$ wegfällt:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x =& S_j^{(k+1)}(f) + \sum_{l=k+2}^{n-1} c_l^{(k+1)} h_j^{2l} + E_j^{(k+1)}(f) \\ =& S_j^{(k+1)}(f) + \mathcal{O}\left(h_j^{2(k+2)}\right)\end{aligned}$$

□

Bemerkung 5.22 (Schema für die Romberg-Integration)
Die Näherungen $S_j^{(k)} = S_j^{(k)}(f)$ können mit Hilfe dieses Schemas berechnet werden:

$$
\begin{array}{ccccccccccc}
S_0^{(0)} & & & & & & & & & & \\
 & \searrow & & & & & & & & & \\
S_1^{(0)} & \to & S_0^{(1)} & & & & & & & & \\
 & \searrow & & \searrow & & & & & & & \\
S_2^{(0)} & \to & S_1^{(1)} & \to & S_0^{(2)} & & & & & & \\
 & \searrow & & \searrow & & \searrow & & & & & \\
\vdots & & \vdots & & \vdots & & \ddots & & & & \\
 & \searrow & & \searrow & & \searrow & & \searrow & & & \\
S_{j-1}^{(0)} & \to & S_{j-2}^{(1)} & \to & S_{j-3}^{(2)} & \to & \cdots & \to & S_0^{(j-1)} & & \\
 & \searrow & & \searrow & & \searrow & & \searrow & & \searrow & \\
S_j^{(0)} & \to & S_{j-1}^{(1)} & \to & S_{j-2}^{(2)} & \to & \cdots & \to & S_1^{(j-1)} & \to & S_0^{(j)}
\end{array}
$$

Hier ist anzumerken, dass bei der Berechnung des ersten Elements der nächsten Zeile ggfs. die Beziehung

$$S_{j+1}(f) = S_{j+1}^{(0)}(f) = \frac{1}{2}\left(S_j(f) + T_j(f)\right), \qquad T_j(f) = h_j \sum_{l=0}^{N_j-1} f\left(a + (l+1/2)\, h_j\right)$$

mit der summierten Tangententrapezformel $T_j(f)$ ausgenutzt werden kann.

Bemerkung 5.23 (Konvergenz)
Auch für stetige Funktionen f auf $[a, b]$ konvergieren die Spalten des obigen Schemas gegen das Integral,

$$S_j^{(k)}(f) \to \int_a^b f(x)\mathrm{d}x \quad (\text{für } j \to \infty, \quad k = 0, 1, 2, \dots)$$

und die Diagonalen des Schemas ebenso

$$S_j^{(k)}(f) \to \int_a^b f(x)\mathrm{d}x \quad (\text{für } k \to \infty, \quad j = 0, 1, 2, \dots).$$

5.6 Gaußsche Quadraturformeln

Die interpolatorischen Quadraturformeln haben die besondere Eigenschaft, dass sie für $n+1$ gegebene Stützstellen $x_0, \dots, x_n$ für Polynome n-ten Grades exakt sind.

Unsere bisherige Aufgabenstellung lautete wie folgt: Gegeben seien die $n+1$ Stützstellen x_j, wir suchen $n+1$ Gewichte α_j, so dass

$$Q(f) = (b-a) \sum_{j=0}^{n} \alpha_j f(x_j)$$

eine Quadraturformel ist.

Frage: Kann man bei diesen Quadraturformeln zusammen mit den $n+1$ Koeffizienten $\alpha_0, \ldots, \alpha_n$ bzw. $\beta_0, \ldots, \beta_n$ auch noch die $n+1$ Stützstellen $x_0, \ldots, x_n$ optimal wählen, so dass die Quadraturformeln für Polynome bis zum $(2n+1)$-ten Grade exakt werden?

Dies führt uns auf die erweiterte Aufgabenstellung: Wir suchen $\alpha_0, \ldots, \alpha_n$ und $x_0, \ldots, x_n$, sodass der Genauigkeitsgrad von Q mindestens $2n+1$ ist.

Definition 5.24 (Gaußsche Quadraturformel)
Die Funktion ϱ sei eine reelle, stückweise stetige, positive Gewichtsfunktion *auf (a,b) mit $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ und die Integrale*

$$\int_a^b x^j \varrho(x) \mathrm{d}x \quad (\text{für } j = 0, 1, 2, \ldots)$$

existieren und seien absolut konvergent[1]*. Außerdem existiere das Integral*

$$J(f) = \int_a^b f(x)\varrho(x)\mathrm{d}x.$$

Eine Quadraturformel

$$Q(f) = (b-a)\sum_{k=0}^{n} \alpha_k f(x_k)$$

mit $x_0, \ldots, x_n \in (a,b)$ und $\alpha_0, \ldots, \alpha_n \in \mathbb{K}$ heißt eine Gaußsche Quadraturformel*, wenn gilt*

$$\int_a^b x^j \varrho(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} \alpha_k x_k^j \iff E(x^j) = 0 \quad (\text{für } j = 0, \ldots, 2n+1)$$

wobei $E(f) = J(f) - Q(f)$ ist. Die Quadraturformel hat also mindestens den Genauigkeitsgrad $2n+1$ und ist damit für Polynome $(2n+1)$-ten Grades exakt.

Definition 5.25 (Skalarprodukt von Polynomen)
Wir führen auf dem Vektorraum

$$\Pi = \left\{ p : (a,b) \mapsto \mathbb{K} \;\middle|\; \exists j, \gamma_0, \ldots, \gamma_j \in \mathbb{K} : p(x) = \sum_{l=0}^{j} \gamma_l x^l \right\}$$

der Polynome über dem Körper $\mathbb{K}$ das Skalarprodukt wie folgt ein:

$$(p,q) := \int_a^b p(x)\overline{q(x)}\varrho(x)\mathrm{d}x \quad (\text{für } p, q \in \Pi)$$

Des weiteren schreiben wir für das Skalarprodukt eines Polynoms p mit sich selbst

$$\|p\|^2 := (p,p).$$

Proposition 5.26 (Orthogonale Polynome)
Durch Orthogonalisierung der Basis $\{1, x, x^2, \ldots, x^n\}$ des Π_n nach dem Gram-Schmidtschen Orthogonalisierungs-Verfahren erhalten wir orthogonale Polynome p_j vom Grad j:

$$p_0(x) := 1$$
$$p_j(x) := x^j - \sum_{k=0}^{j-1} \frac{(x^j, p_k(x))}{\|p_k\|^2} p_k(x)$$

[1] z.B.: $\varrho(x) = e^{-x}$, e^{-x^2}, $\ln(1/x)$. Im Folgenden beschränken wir uns jedoch im wesentlichen auf $\varrho(x) = 1$.

Diese Polynome $p_0, p_1, \ldots, p_n$ sind offensichtlich selbst auch wieder eine Basis des Π_n, d.h.

$$\Pi_n = \langle 1, x, \ldots, x^n \rangle = \langle p_0, p_1, \ldots, p_n \rangle .$$

Proposition 5.27 (Wurzeln des Polynoms $\mathbf{p_{n+1}}$)
Das Polynom p_{n+1} (mit $\operatorname{grad}(p_{n+1}) = n+1$*) hat im Körper $\mathbb{K} = \mathbb{C}$ genau $n+1$ reelle, paarweise verschiedene Nullstellen oder* Wurzeln*. Diese Wurzeln $x_0, \ldots, x_n$ haben mit den zugehörigen Lagrange-Polynomen $l_j(x)$ die* Rayleight-Quotienten-Darstellung

$$x_j = \frac{(xl_j, l_j)}{\|l_j\|^2} = \frac{\int_a^b x|l_j(x)|^2 \varrho(x)\mathrm{d}x}{\int_a^b |l_j(x)|^2 \varrho(x)\mathrm{d}x} \quad \textit{und} \quad l_j(x) = \prod_{k=0, k\neq j}^{n} \frac{x - x_k}{x_j - x_k} \quad (\textit{für } j = 0, \ldots, n) .$$

Die Wurzeln $x_0, \ldots, x_n$ liegen im Intervall (a, b).

Beweis. Der Beweis erfolgt später im Kapitel über orthogonale Polynome. □

Proposition 5.28 (Bedingung für Gaußsche Quadraturformel)
Die Quadraturformel Q ist genau dann eine Gaußsche Quadraturformel, wenn die Stützstellen $x_0, \ldots, x_n$ gleich den Wurzeln des orthogonalen Polynoms p_{n+1} sind und mit den zugehörigen Lagrange-Polynomen $l_0, \ldots, l_n$ die Koeffizienten $\alpha_0, \ldots, \alpha_n$ die Darstellung

$$\alpha_j = \frac{1}{b-a}\int_a^b l_j(x)\varrho(x)\mathrm{d}x \overset{!}{=} \frac{1}{b-a}\int_a^b l_j(x)^2\varrho(x)\mathrm{d}x > 0 \quad (\textit{für } j = 0, \ldots, n)$$

haben. Das Restglied der Gaußschen Quadraturformel Q hat für $f \in C^{2n+2}(a, b)$ die Gestalt

$$E(f) = \int_a^b f[x_0, \ldots, x_n, x_0, \ldots, x_n, x] p_{n+1}^2(x)\varrho(x)\mathrm{d}x$$

und gestattet für reellwertige Funktionen f mit $\xi \in (a, b)$ die Darstellung

$$E(f) = \frac{1}{(2n+2)!} \underbrace{\int_a^b p_{n+1}^2(x)\varrho(x)\mathrm{d}x}_{=\|p_{n+1}\|^2} \cdot f^{(2n+2)}(\xi). \tag{7}$$

Beweis. Im Kapitel über orthogonale Polynome werden wir sehen, dass die Quadraturformel genau dann exakt ist für Polynome $2n+1$-ten Grades, wenn die Koeffizienten und Stützstellen die genannten Eigenschaften haben.

Zum Restglied: Mit dem Interpolationspolynom $p_{0,\ldots,n,0,\ldots,n} \in \Pi_{2n+1}$ von f in der Hermite-Darstellung mit den $2n+2$ Stützstellen $x_0, \ldots, x_n, x_0, \ldots, x_n$ und dem Restglied R kann man f wie folgt darstellen:

$$\begin{aligned} f(x) &= p_{0,\ldots,n,0,\ldots,n}(x) + R(x) \\ R(x) &= (x - x_0)^2 \cdots (x - x_n)^2 f[x_0, \ldots, x_n, x_0, \ldots, x_n, x]. \end{aligned}$$

Sind nun die Stützstellen x_j gleich den Wurzeln (Nullstellen) des orthogonalen Polynoms $p_{n+1} \in \Pi_{n+1}$, so muss p_{n+1} die Darstellung

$$p_{n+1}(x) \overset{\text{Def.}}{=} x^{n+1} - \sum_{k=0}^{n} \underbrace{\frac{(x^{n+1}, p_k(x))}{\|p_k\|^2}}_{\in\mathbb{K}} \underbrace{p_k(x)}_{\in\Pi_k,\, k\leq n} \overset{!}{=} \gamma(x - x_0)\cdots(x - x_n) \quad (\text{für } \gamma \in \mathbb{K})$$

haben. Jeder Summand in der Summe ist nur ein Polynom vom Grad $k \leq n$, womit der Koeffizient der höchsten Potenz x^{n+1} von p_{n+1} als 1 identifiziert wird. Damit erhalten wir $\gamma = 1$ und können schreiben

$$p_{n+1}(x) = (x - x_0) \cdots (x - x_n) \implies R(x) = p_{n+1}^2(x) f[x_0, \ldots, x_n, x_0, \ldots, x_n, x].$$

Da die Quadraturformel exakt ist für Polynome vom Grad $\leq 2n + 1$, gilt (vgl. auch Prop. 5.24)

$$J(p_{0,\ldots,n,0,\ldots,n}) = Q(p_{0,\ldots,n,0,\ldots,n}) = (b - a) \sum_{k=0}^{n} \alpha_k f(x_k) = Q(f).$$

Wir können $J(f)$ also darstellen als

$$J(f) = J(p_{0,\ldots,n,0,\ldots,n}) + J(R) = Q(f) + E(f) \implies E(f) = J(R).$$

Also erhalten wir für die Restgliedabschätzung

$$E(f) = J(R) \overset{\text{Def.}}{=} \int_a^b R(x)\varrho(x)\mathrm{d}x = \int_a^b p_{n+1}^2(x) f[x_0, \ldots, x_n, x_0, \ldots, x_n, x]\varrho(x)\mathrm{d}x.$$

Mit dem Mittelwertsatz der Integralrechnung erhalten wir schließlich für reellwertige Funktionen $f \in C^{2n+2}(a, b)$ auch die zweite Darstellung von $E(f)$. □

Beispiel 5.29 (Legendre-Polynome)
Für $\varrho = 1$, $a = -1$, $b = 1$ erhält man die orthogonalen *Legendre-Polynome* in der Form

$$p_0 = 1, \qquad p_1 = x, \qquad p_2 = x^2 - \frac{1}{3}, \qquad p_3 = x^3 - \frac{3}{5}x, \qquad p_4 = x^4 - \frac{6}{7}x^2 + \frac{3}{35}, \qquad \ldots$$

Die zugehörigen Normierungsintegrale lauten

$$\|p_j\|^2 = \int_{-1}^{1} p_j^2(x)dx = \frac{2^{2j+1}}{2j+1}\binom{2j}{j}^{-2} \quad (\text{für } j = 0, 1, 2, \ldots)$$
$$\implies \|p_0\|^2 = 2, \qquad \|p_1\|^2 = \frac{2}{3}, \qquad \|p_2\|^2 = \frac{8}{45}, \qquad \|p_3\|^2 = \frac{8}{175}, \qquad \ldots$$

Die Wurzeln (Nullstellen) auf $[-1, 1]$ ergeben sich zu

$$p_2: \qquad x_0 = -\frac{1}{\sqrt{3}}, \qquad x_1 = \frac{1}{\sqrt{3}};$$
$$p_3: \qquad x_0 = -\sqrt{\frac{3}{5}}, \qquad x_1 = 0, \qquad x_2 = \sqrt{\frac{3}{5}};$$
$$\ldots$$

Proposition 5.30 (Restglied für $\varrho(\mathbf{x}) = \mathbf{1}$)
Für eine reellwertige Funktion $f \in C^{(2n+2)}(a, b)$ kann das Restglied in der Form

$$E(f) = \frac{(b-a)^{2n+3}}{2^{2n+3}(2n+2)!} \cdot \underbrace{\int_{-1}^{1} (z - z_0)^2 \cdots (z - z_n)^2 \mathrm{d}z}_{= \|p_{n+1}\|^2_{[-1,1]}} f^{(2n+2)}(\xi), \qquad z = 2\frac{x-c}{b-a}, \qquad c = \frac{a+b}{2}$$

angegeben werden.

Beweis. Wir substituieren in dem Integral in Gleichung (7)

$$z := 2\frac{x-c}{b-a}, \qquad c := \frac{b+a}{2} \implies \frac{x-x_j}{b-a} = \frac{1}{2}(z-z_j), \qquad \mathrm{d}x = \frac{1}{2}(b-a)\mathrm{d}z.$$

Damit erhalten wir

$$\begin{aligned}\|p_{n+1}\|^2 &= \int_a^b (x-x_0)^2 \cdots (x-x_n)^2 \mathrm{d}x \\ &= (b-a)^{2n+2} \int_a^b \left(\frac{x-x_0}{b-a}\right)^2 \cdots \left(\frac{x-x_n}{b-a}\right)^2 \mathrm{d}x \\ &= \frac{(b-a)^{2n+3}}{2^{2n+3}} \int_{-1}^1 (z-z_0)^2 \cdots (z-z_n)^2 \mathrm{d}z\end{aligned}$$

Setzen wir das Integral in Gleichung (7) ein, so erhalten wir das Ergebnis. □

Proposition 5.31 (Gaußsche Quadraturformel mit zwei Punkten)
Die Gaußsche Quadraturformel mit zwei Punkten lautet

$$Q_{G_2}(f) = \frac{h}{2}\left(f\left(c - \frac{h}{2\sqrt{3}}\right) + f\left(c + \frac{h}{2\sqrt{3}}\right)\right), \qquad h := b-a, \qquad c := \frac{a+b}{2}$$

und hat das Restglied

$$E(f) = \frac{h^5}{4320} f^{(4)}(\xi), \qquad \xi \in (a,b).$$

Die Genauigkeit ist somit $n = 3$ (wie bei Simpson oder Hermite, allerdings wird ein Punkt weniger benötigt).

Beweis. Die Formel ergibt sich aus dem Fall $n = 1$: $p_{n+1} = p_2 = z^2 - \frac{1}{3}$, $h = b - a$. Wir erhalten

$$z_{0,1} = \pm\frac{1}{\sqrt{3}} \implies x_0 = c - \frac{h}{2\sqrt{3}}, \qquad x_1 = c + \frac{h}{2\sqrt{3}}, \qquad \alpha_{0,1} = \frac{1}{b-a}\int_a^b l_1(x)dx = \frac{1}{2}.$$

□

Proposition 5.32 (Gaußsche Quadraturformel mit drei Punkten)
Die Gaußsche Quadraturformel mit drei Punkten lautet

$$Q_{G_3}(f) = \frac{h}{18}\left(5f\left(c - h\frac{\sqrt{15}}{10}\right) + 8f(c) + 5f\left(c + h\frac{\sqrt{15}}{10}\right)\right), \qquad h := b-a, \qquad c := \frac{a+b}{2}$$

und hat das Restglied

$$E(f) = \frac{h^7}{2016000} f^{(6)}(\xi), \qquad \xi \in (a,b).$$

Der Genauigkeitsgrad ist $n = 5$.

Beweis. Der Beweis ergibt sich aus dem Fall $n = 2$ und soll hier nicht geführt werden. □

5.7 Summierte Gaußsche Quadraturformel

Proposition 5.33 (Summierte Gaußsche Quadraturformel)
Die summierte Gaußsche Quadraturformel lautet

$$Q_{G_2}(f) = \frac{h}{2} \sum_{r=1}^{N} \left(f\left(c_r - \frac{h}{2\sqrt{3}}\right) + f\left(c_r + \frac{h}{2\sqrt{3}}\right)\right),$$

wobei die Abkürzungen

$$\forall r : h := b_r - a_r, \qquad c_r := x_{r-\frac{1}{2}} = \frac{x_\nu + x_{\nu-1}}{2}$$

verwendet wurden. Die Fehlerabschätzung lautet

$$|E(f)| = \frac{b-a}{4320} h^4 \max_{a \le x \le b} \left| f^{(4)}(x) \right|.$$

6 Wiederholung zur Linearen Algebra

Im Folgenden sei $\mathbb{K}$ ein Körper, $\mathbb{K} \in \{\mathbb{C}, \mathbb{R}\}$.

Definition 6.1 (Vektorraum $\mathbb{K}^n$)
Der n-dimensionale Vektorraum $V = \mathbb{K}^n$ ist die Menge $V = \{(x_1, \ldots, x_n) : x_i \in \mathbb{K}, i = 1, \ldots, n\}$. Addition von Vektoren und Multiplikation mit Zahlen sind erklärt durch:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$
$$\lambda(x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n)$$

Definition 6.2 (Standardbasis des $\mathbb{K}^n$)
Die Standardbasis des $\mathbb{K}^n$ bilden die Vektoren

$$e_l = (\delta_{1l}, \ldots, \delta_{nl}) \quad (\textit{für } l = 1, \ldots, n)\,.$$

6.1 Skalarprodukt

Definition 6.3 (Skalarprodukt)
Sei E ein Vektorraum über $\mathbb{K}$. Ein Skalarprodukt ist eine Abbildung $(\cdot, \cdot) : E^2 \mapsto \mathbb{K}$ mit den Eigenschaften:

(S_1) $(u, u) \geq 0$

(S_2) $(u, u) = 0 \iff u = 0$

(S_3) $(\lambda u + \mu v, w) = \lambda\,(u, w) + \mu\,(v, w)$

(S_4) $(u, v) = \overline{(v, u)}$

Proposition 6.4 (Cauchy-Schwarz-Ungleichung)
Für ein Skalarprodukt gilt die Ungleichung

$$|\,(u, v)\,|^2 \leq (u, u) \cdot (v, v)\,.$$

Insbesondere gilt die Gleichheit genau dann, wenn u und v linear abhängig sind.

Beweis. Der technische, aber nicht schwierige Beweis findet sich in jedem Buch zur linearen Algebra. Er soll hier nicht geführt werden. □

Definition 6.5 (Standardskalarprodukt auf dem $\mathbb{K}^n$)
Das Standardskalarprodukt auf $\mathbb{C}^n$ ist definiert durch

$$(x, y)_2 = \sum_{j=1}^{n} x_j \overline{y_j}.$$

Für den $\mathbb{R}^n$ ergibt sich

$$(x, y)_2 = \sum_{j=1}^{n} x_j y_j.$$

Beweis. Alle Eigenschaften (S_1) bis (S_4) ergeben sich unmittelbar durch Nachrechnen. □

6.2 Normen

Definition 6.6 (Norm, normierter Raum)
Sei V ein Vektorraum. Die Abbildung $\|\cdot\| : E \mapsto \mathbb{R}$ ist eine Norm, wenn gilt:

(N_1) $\|u\| \geq 0$ *Semidefinitheit*

(N_2) $\|u\| = 0 \iff u = 0$ *Definitheit*

(N_3) $\|\lambda u\| = |\lambda| \, \|u\|$ *Homogenität*

(N_4) $\|u + v\| \leq \|u\| + \|v\|$ *Dreiecksungleichung*

Ein Vektorraum V mit einer Norm ist ein normierter Raum.

Bemerkung 6.7 (Dreiecksungleichung nach unten)
Aus (N_4) folgt die Ungleichung $|\, \|u\| - \|v\| \,| \leq \|u + v\|$, denn

$$\begin{aligned} \|u\| = \|(u+v) - v\| &\leq \|u+v\| + \|v\| \\ \implies \|u\| - \|v\| &\leq \|u+v\| \,. \end{aligned}$$

Analog gilt

$$\begin{aligned} \|v\| = \|(v+u) - u\| &\leq \|u+v\| + \|u\| \\ \implies \|v\| - \|u\| &\leq \|u+v\| \,. \end{aligned}$$

Proposition 6.8 (Vom Skalarprodukt induzierte Norm)
Sei E ein Vektorraum mit Skalarprodukt. Dann wird durch

$$\|u\| := \sqrt{(u,u)}$$

eine Norm auf E definiert.

Beweis. (N_1) und (N_2) sind klar. (N_3) folgt mittels

$$\|\lambda u\| = \sqrt{(\lambda u, \lambda u)} = \sqrt{\lambda\overline{\lambda}\,(u,u)} = \sqrt{|\lambda|^2\,(u,u)} = |\lambda|\sqrt{(u,u)} = |\lambda| \, \|u\| \,.$$

Die Eigenschaft (N_4) ist gerade die Cauchy-Schwarz-Ungleichung. □

Bemerkung 6.9 (Umformulierung der Cauchy-Schwarz-Ungleichung)
Die Cauchy-Schwarz-Ungleichung besagt somit

$$|\,(u,v)\,| \leq \|u\| \cdot \|v\| \,.$$

Insbesondere gilt für das Standardskalarprodukt auf dem $\mathbb{K}^n$

$$\left| \sum_{j=1}^{n} x_j \overline{y_j} \right| \leq \sqrt{\sum_{j=1}^{n} |x_j|^2} \sqrt{\sum_{j=1}^{n} |y_j|^2}.$$

Definition 6.10 (Normen auf dem $\mathbb{K}^n$)
Folgende Abbildungen definieren Normen auf $\mathbb{K}^n$. Mit $x = (x_1, \ldots, x_n) \in \mathbb{K}^n$ definiert man:

- *Die Maximumnorm:*
$$\|x\|_\infty = \max_{j=1,\ldots,n} |x_j|$$
- *Die euklidische bzw. unitäre Norm*
$$\|x\|_2 = \left(\sum_{j=1}^{n} |x_j|^2 \right)^{\frac{1}{2}}$$
- *Die 1-Norm*
$$\|x\|_1 = \sum_{j=1}^{n} |x_j|$$

Beweis. Dass die Maximumnorm und die 1-Norm Normen sind, ist unmittelbar klar. Die euklidische Norm ist die von dem euklidischen Skalarprodukt induzierte Norm. □

Definition 6.11 (Geometrische Objekte in normierten Räumen)
Sei E ein normierter Raum. Dann definiert man:

Punkte	*Vektoren $x, y \in \mathbb{K}^n$*
Abstand der Punkte x, y	$\|x - y\|$
Offene Kugel mit Mittelpunkt c und Radius ϱ	$K_\varrho(c) = \{x \in \mathbb{K}^n \mid \|x - c\| < \varrho\}$
Abgeschlossene Kugel mit Mittelpunkt c und Radius ϱ	$\overline{K_\varrho(c)} = \{x \in \mathbb{K}^n \mid \|x - c\| \leq \varrho\}$
Sphäre mit Mittelpunkt c und Radius ϱ	$S_\varrho(c) = \{x \in \mathbb{K}^n \mid \|x - c\| = \varrho\}$

Beachte, dass in normierten Räumen die abgeschlossene Kugel tatsächlich der Abschluss der offenen Kugel ist (dieses gilt i.A. in beliebigen metrischen Räumen nicht).

Beispiel 6.12
Die Kugeln um $c = (c_1, \ldots, c_n)$ mit Radius ϱ bezüglich den oben definierten Normen sind:

$$\begin{aligned}
\|\cdot\|_\infty : &\quad \overline{K_\varrho(c)} = \{(x_1, \ldots, x_n) : |x_i - c_i| \leq \varrho\} \\
\|\cdot\|_2 : &\quad \overline{K_\varrho(c)} = \left\{ (x_1, \ldots, x_n) : \sum_{i=1}^{n} |x_i - c_i|^2 \leq \varrho^2 \right\} \\
\|\cdot\|_1 : &\quad \overline{K_\varrho(c)} = \left\{ (x_1, \ldots, x_n) : \sum_{i=1}^{n} |x_i - c_i| \leq \varrho \right\}
\end{aligned}$$

Definition 6.13 (Äquivalenz von Normen)
Zwei Normen $\|\cdot\|$ und $\|\cdot\|'$ heißen äquivalent, wenn es $\gamma_0, \gamma_1 > 0$ gibt, mit

$$\forall x : \gamma_0 \|x\| \leq \|x\|' \leq \gamma_1 \|x\| .$$

Proposition 6.14 (Alle Normen auf dem $\mathbb{K}^n$ sind äquivalent)
Für je zwei Normen $\|\cdot\|$ und $\|\cdot\|'$ auf $\mathbb{K}^n$ gilt, dass $\|\cdot\|$ und $\|\cdot\|'$ äquivalent sind.

Beweis. Zuerst zeigen wir, dass alle Normen äquivalent zur euklidischen Norm sind.

Sei $e_1, \dots, e_n$ die kanonische Basis des $\mathbb{K}^n$. Dann gilt mittels Cauchy-Schwarz die Abschätzung

$$\|x\| = \left\| \sum_{k=1}^{n} x_k e_k \right\| \leq \sum_{k=1}^{n} |x_k| \, \|e_k\| \overset{CS}{\leq} \sqrt{\sum_{k=1}^{n} \|e_k\|^2} \cdot \|x\|_2 =: \gamma_1 \|x\|_2 .$$

Weiter ist jede Norm lipschitzstetig, denn

$$| \, \|x\| - \|y\| \, | \leq \|x - y\| \leq \gamma_1 \|x - y\|_2 .$$

Die Einheitssphäre $S := \{x \in \mathbb{K}^n : \|x\|_2 = 1\}$ ist kompakt. Dann besitzt die Norm $\|\cdot\|$ als stetige Funktion auf einer kompakten Menge ein Minimum γ_0 an der Stelle $z \in S$, d.h.

$$\|z\| = \gamma_0 = \min_{\|u\|_2 = 1} \|u\| .$$

Insbesondere ist $\gamma_0 > 0$, denn sonst wäre $z = 0 \notin S$. Nun gilt

$$\gamma_0 \|x\|_2 \leq \left\| \frac{x}{\|x\|_2} \right\| \|x\|_2 = \frac{1}{\|x\|_2} \cdot \|x\|_2 \cdot \|x\| = \|x\| .$$

Insgesamt folgt $\gamma_0 \|x\|_2 \leq \|x\| \leq \gamma_1 \|x\|_2$.

Nun zeigen wir, dass je zwei beliebige Normen äquivalent sind.

Denn sind $\|\cdot\|, \|\cdot\|'$ zwei Normen, dann gibt es $\gamma_0, \gamma_1, \gamma_0', \gamma_1'$ so, dass

$$\begin{aligned} &\gamma_0 \|x\|_2 \leq \|x\| \leq \gamma_1 \|x\|_2 \\ &\gamma_0' \|x\|_2 \leq \|x\|' \leq \gamma_1' \|x\|_2 \\ \Longrightarrow \; &\frac{\gamma_0}{\gamma_1'} \|x\|' \leq \gamma_0 \|x\|_2 \leq \|x\| \leq \gamma_1 \|x\|_2 \leq \frac{\gamma_1}{\gamma_0'} \|x\|' . \end{aligned}$$

□

Definition 6.15 (Konvergenz)
Sei E ein normierter Raum mit Norm $\|\cdot\|$. Eine beliebige Folge von Vektoren $x^{(t)} = \left(x_1^{(t)}, \dots, x_n^{(t)}\right) \in E$ heißt konvergent gegen einen Vektor $x \in E$, wenn gilt

$$\lim_{t \to \infty} \left\| x^{(t)} - x \right\| = 0.$$

In diesem Fall schreibt man

$$\lim_{t \to \infty} x^{(t)} = x.$$

Proposition 6.16 (Konvergenz im $\mathbb{K}^n$ ist äquivalent zu elementweiser Konvergenz)
Eine Folge von Vektoren $x^{(t)} = \left(x_1^{(t)}, \dots, x_n^{(t)}\right) \in \mathbb{K}^n$ konvergiert genau dann gegen einen Vektor x, wenn alle Koordinatenfolgen konvergieren, d.h.

$$\lim_{t \to \infty} x^{(t)} = x \iff \forall i = 1, \dots, n : \lim_{t \to \infty} x_i^{(t)} = x_i .$$

Insbesondere hängen Konvergenz und Grenzwert nicht von der Wahl der jeweiligen Norm ab.

Beweis. Je zwei Normen sind äquivalent, d.h. es gibt $\gamma_1, \gamma_2 > 0$ mit $\forall x \in \mathbb{K}^n : \gamma_1 \|x\|_\infty \leq \|x\| \leq \gamma_2 \|x\|_\infty$. Nun gilt

$$\begin{aligned}
&\lim_{t\to\infty} x^{(t)} = x \\
\iff &\lim_{t\to\infty} \left\|x^{(t)} - x\right\| = 0 \\
\iff &\lim_{t\to\infty} \gamma \left\|x^{(t)} - x\right\|_\infty = 0 \\
\iff &\lim_{t\to\infty} \left\|x^{(t)} - x\right\|_\infty = 0 \\
\iff &\forall j = 1, \ldots, n : \lim_{t\to\infty} |x_j^{(t)} - x_j| = 0 \\
\iff &\forall j = 1, \ldots, n : \lim_{t\to\infty} x_j^{(t)} = x_j.
\end{aligned}$$

□

6.3 Matrizen

Als Notation für eine $m \times n$ - Matrix mit m Zeilen und n Spalten verwenden wir

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} = (a_{jk})_{\substack{j=1,\ldots,m \\ k=1,\ldots,n}}.$$

Definition 6.17 (Raum der Matrizen)
Die $m \times n$ - Matrizen mit Elementen aus $\mathbb{K}$ bilden einen $(n \cdot m)$-dimensionalen Vektorraum $\mathrm{Mat}_\mathbb{K}(m, n)$ über $\mathbb{K}$ mit den Verknüpfungen

$$\begin{aligned}
C = A + B \quad &\textit{mit} \quad c_{jk} := a_{jk} + b_{jk}, \\
C = \lambda A \quad &\textit{mit} \quad c_{jk} = \lambda a_{jk}.
\end{aligned}$$

Beachte, dass somit eine triviale Isomorphie zwischen $\mathrm{Mat}_\mathbb{K}(m, n)$ und $\mathbb{K}^{mn}$ besteht.

Ist $m = n$, so ist die Matrix quadratisch. Der Raum aller $n \times n$-Matrizen wird mit $\mathrm{Mat}_\mathbb{K}(n) := \mathrm{Mat}_\mathbb{K}(n, n)$ bezeichnet. Dieser Raum ist zusätzlich abgeschlossen unter der Matrizenmultiplikation, d.h. für $A, B \in \mathrm{Mat}_\mathbb{K}(n)$ ist das Produkt AB definiert und $AB \in \mathrm{Mat}_\mathbb{K}(n)$.

Proposition 6.18 (Alle Normen auf $\mathrm{Mat}_\mathbb{K}(m, n)$ sind äquivalent)
Für je zwei Normen $\|\cdot\|$ und $\|\cdot\|'$ auf $\mathrm{Mat}_\mathbb{K}(m, n)$ gilt, dass $\|\cdot\|$ und $\|\cdot\|'$ äquivalent sind.

Beweis. Dieses folgt unmittelbar aus der Isomorphie von $\mathrm{Mat}_\mathbb{K}(m, n)$ und $\mathbb{K}^{mn}$ sowie Proposition 6.14.
□

Definition 6.19 (Konvergenz von Matrizenfolgen)
Für eine beliebige Norm $\|\cdot\|$ auf $\mathrm{Mat}_\mathbb{K}(m, n)$ definiert man: Eine Folge von Matrizen $A^{(t)} = \left(a_{jk}^{(t)}\right)$ konvergiert gegen eine Matrix $A = (a_{jk})$ wenn gilt:

$$\lim_{t\to\infty} \left\|A^{(t)} - A\right\| = 0$$

Bemerkung 6.20 (Wohldefiniertheit und Äquivalenz zur elementweisen Konvergenz)
Obige Definition ist unabhängig von der Wahl der Norm, da je zwei Normen äquivalent sind. Außerdem gilt durch die Isomorphie von $\mathrm{Mat}_{\mathbb{K}}(m,n)$ und $\mathbb{K}^{mn}$, dass eine Folge von Matrizen $A^{(t)}$ genau dann gegen eine Matrix A konvergiert, wenn die Folge elementweise gegen A konvergiert.

Proposition 6.21 (Konvergenz von Matrizenfolgen)
Eine Folge von Matrizen $A^{(t)} = \left(a_{jk}^{(t)}\right)$ *konvergiert in* $\mathrm{Mat}_{\mathbb{K}}(m,n)$ *genau dann gegen eine Matrix* A*, wenn gilt:*

$$\forall x \in \mathbb{K}^n : \lim_{t\to\infty} A^{(t)}x = Ax$$

Beweis. Es konvergiere $A^{(t)}$ gegen A. Sei $x = (x_1, \ldots, x_n) \in \mathbb{K}^n$. Dann gilt für $j = 1, \ldots, m$:

$$\lim_{t\to\infty} (A^{(t)}x)_j = \lim_{t\to\infty} \left(\sum_{k=1}^{n} a_{jk}^{(t)} x_k \right) = \sum_{k=1}^{n} \left(\lim_{t\to\infty} a_{jk}^{(t)} \right) x_k = \sum_{k=1}^{n} a_{jk}x_k = (Ax)_j$$

Gelte umgekehrt für jeden Vektor x die Identität $\lim_{t\to\infty} A^{(t)}x = Ax$. Dann gilt dieses insbesondere für die Einheitsvektoren e_l für $l = 1, \ldots, n$. Damit erhält man die Konvergenz der Spalten von $A^{(t)}$ gegen die Spalten von A und hieraus wiederum die elementweise Konvergenz der Matrix $A^{(t)}$ gegen A für $t \to \infty$. □

Definition 6.22 (Matrizen definieren eine Abbildung)
Sei $A = (a_{jk})$ *eine* $m \times n$*-Matrix. Dann ist eine lineare Abbildung* L_A *definiert durch*

$$L_A : x \mapsto Ax \quad \textit{mit} \quad (Ax)_j = \sum_{k=1}^{n} a_{jk}x_k.$$

Beweis. Die Linearität folgt unmittelbar durch Nachrechnen:

$$(A(\lambda x + \mu y))_j = \sum_{k=1}^{n} a_{jk}(\lambda x + \mu y)_k = \lambda \sum_{k=1}^{n} a_{jk}x_k + \mu \sum_{k=1}^{n} a_{jk}y_k = (\lambda Ax)_j + (\mu Ay)_j$$

□

Definition 6.23 (Verträglichkeit)
Eine Norm $\|\cdot\|$ *auf* $\mathrm{Mat}_{\mathbb{K}}(m,n)$ *heißt verträglich mit* $\|\cdot\|_{\mathbb{K}^n}, \|\cdot\|_{\mathbb{K}^m}$ *wenn für jedes* $A \in \mathrm{Mat}_{\mathbb{K}}(m,n)$ *gilt:*

$$\|Ax\|_{\mathbb{K}^m} \leq \|A\| \, \|x\|_{\mathbb{K}^n}$$

Definition 6.24 (Matrizennorm auf dem $\mathrm{Mat}_{\mathbb{K}}(n)$)
Sei $\|\cdot\|$ *eine Norm auf* $\mathrm{Mat}_{\mathbb{K}}(n)$*. Man nennt* $\|\cdot\|$ *eine Matrizennorm, wenn sie die folgende Eigenschaft erfüllt:*

(N_5) $\forall A, B \in \mathrm{Mat}_{\mathbb{K}}(n) : \|AB\| \leq \|A\| \, \|B\|$ *Submultiplikativität*

Definition 6.25 (natürliche Norm)
Die natürliche Norm $\|\cdot\|_{nat}$ *auf* $\mathrm{Mat}(m,n)$ *bezüglich der Normen* $\|\cdot\|_{\mathbb{K}^n}, \|\cdot\|_{\mathbb{K}^m}$ *ist definiert durch*

$$\|A\|_{nat} = \sup_{0\neq x\in\mathbb{K}^n} \frac{\|Ax\|_{\mathbb{K}^m}}{\|x\|_{\mathbb{K}^n}} = \sup_{\substack{y\in\mathbb{K}^n \\ \|y\|_{\mathbb{K}^n}=1}} \|Ay\|_{\mathbb{K}^m}.$$

Bemerkung 6.26 (Eigenschaften der natürlichen Norm)
Sei $\|\cdot\|_{\text{nat}}$ die natürliche Norm bezüglich $\|\cdot\|_{\mathbb{K}^n}$ und $\|\cdot\|_{\mathbb{K}^m}$. Man rechnet leicht nach, dass die beiden gegebenen Definitionen übereinstimmen und wirklich eine Norm definieren. Des weiteren gilt

1. Die natürliche Norm $\|\cdot\|_{\text{nat}}$ ist verträglich mit $\|\cdot\|_{\mathbb{K}^n}, \|\cdot\|_{\mathbb{K}^m}$.
2. $\|A\|_{\text{nat}} = \min\{c \geq 0 | \|Ax\|_{\mathbb{K}^m} \leq c\|x\|_{\mathbb{K}^n}, x \in \mathbb{K}^n\}$, d.h. die natürliche Norm ist die kleinste verträgliche Norm.

Für den Fall $m = n$ gilt zusätzlich

1. $\|E\|_{\text{nat}} = 1$, wobei E die Einheitsmatrix ist.
2. $\|\cdot\|_{\text{nat}}$ ist eine Matrizennorm.

6.4 Besondere Matrizennormen

Definition 6.27
Als besondere Normen auf $\text{Mat}_{\mathbb{K}}(m,n)$ *definieren wir:*

- *Die maximale Zeilensumme:*
$$\|A\|_\infty := \max_{j=1,\ldots,m} \sum_{k=1}^{n} |a_{jk}|$$
- *Die maximale Spaltensumme:*
$$\|A\|_1 := \max_{k=1,\ldots,n} \sum_{j=1}^{m} |a_{jk}|$$
- *Die Quadratsummennorm (sie entspricht der euklidischen bzw. unitären Norm auf dem* $\mathbb{K}^{mn}$*):*
$$\|A\|_2 := \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{n} |a_{jk}|^2}$$

Proposition 6.28 (Verträglichkeit und Natürlichkeit der maximalen Zeilensumme)
Für die maximale Zeilensumme gilt:

(i) $\|\cdot\|_\infty$ *auf* $\text{Mat}_{\mathbb{K}}(m,n)$ *ist verträglich bezüglich den Maximumnormen* $\|\cdot\|_\infty$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

(ii) Ist $m = n$*, so ist* $\|\cdot\|_\infty$ *auf* $\text{Mat}_{\mathbb{K}}(m,n)$ *sogar die natürliche Norm bezüglich den Maximumnormen* $\|\cdot\|_\infty$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

Beweis. Zu (i): Dieses folgt aus der für alle $x \in \mathbb{K}^n$ geltenden Ungleichungskette

$$\|Ax\|_\infty = \max_{j=1,\ldots,m} \left| \sum_{k=1}^{n} a_{jk}x_k \right| \leq \max_{j=1,\ldots,m} \sum_{k=1}^{n} |a_{jk}| \max_{k=1,\ldots,n} |x_k| = \|A\|_\infty \|x\|_\infty .$$

Zu (ii): Für $A = 0$ gilt $\|A\|_\infty = 0 = \|A\|_{\text{nat}}$. Sei $A \neq 0$.

Für alle x mit $\|x\|_\infty = 1$ gilt $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty = \|A\|_\infty$. Insbesondere gilt

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty \leq \|A\|_\infty .$$

Nun gibt es ein $l \in \{1, \ldots, n\}$ so, dass

$$\|A\|_\infty = \max_{j=1,\ldots,n} \sum_{k=1}^{n} |a_{jk}| = \sum_{k=1}^{n} |a_{lk}|.$$

Setze nun $z = (z_1, \ldots, z_n)$ mit

$$z_k = \begin{cases} \frac{\overline{a_{lk}}}{|a_{lk}|} & \text{wenn } a_{lk} \neq 0, \\ 0 & \text{sonst.} \end{cases}$$

Wegen $A \neq 0$ ist $\|z\|_\infty = 1$. Weiter ist

$$\|Az\|_\infty = \max_{j=1,\ldots,n} \left| \sum_{k=1}^{n} a_{jk} z_k \right| \geq \left| \sum_{k=1}^{n} a_{lk} z_k \right| = \sum_{k=1}^{n} |a_{lk}| = \|A\|_\infty .$$

Aus der Ungleichungskette

$$\|A\|_{\text{nat}} = \sup_{\|x\|_\infty=1} \|Ax\|_\infty \leq \|A\|_\infty \leq \|Az\|_\infty \leq \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \|A\|_{\text{nat}} .$$

folgt die Identität $\|A\|_\infty = \|A\|_{\text{nat}}$. □

Proposition 6.29 (Verträglichkeit und Natürlichkeit der maximalen Spaltensumme)
Für die maximale Spaltensumme gilt:

(i) $\|\cdot\|_1$ *auf* $\text{Mat}_{\mathbb{K}}(m,n)$ *ist verträglich bezüglich den 1-Normen* $\|\cdot\|_1$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

(ii) Ist $m = n$, *so ist* $\|\cdot\|_1$ *auf* $\text{Mat}_{\mathbb{K}}(m,n)$ *sogar die natürliche Norm bezüglich den 1-Normen* $\|\cdot\|_1$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

Beweis. Der Beweis verläuft analog zum obigen Beweis für die maximale Zeilensumme. □

Proposition 6.30 (Verträglichkeit und Natürlichkeit der Quadratsummennorm)
Für die Quadratsummennorm gilt:

(i) $\|\cdot\|_2$ *auf* $\text{Mat}_{\mathbb{K}}(m,n)$ *ist verträglich bezüglich den euklidischen bzw. unitären Normen* $\|\cdot\|_2$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

(ii) Jedoch ist sie auf $\text{Mat}_{\mathbb{K}}(n)$ *für* $n > 1$ *keine natürliche Norm bezüglich den unitären bzw. euklidischen Normen* $\|\cdot\|_2$ *auf* $\mathbb{K}^m$ *und* $\mathbb{K}^n$.

Beweis. Zu (i): Aus der Cauchy-Schwarz-Ungleichung folgt

$$\|Ax\|_2^2 = \sum_{j=1}^{m} \left| \sum_{k=1}^{n} a_{jk} x_k \right|^2 \overset{CS}{\leq} \sum_{j=1}^{m} \left(\sum_{k=1}^{n} |a_{jk}|^2 \sum_{k=1}^{n} |x_k|^2 \right) = \left(\sum_{j=1}^{m} \sum_{k=1}^{n} |a_{jk}|^2 \right) \sum_{k=1}^{n} |x_k|^2 = \|A\|_2^2 \|x\|_2^2$$

und somit $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$.

Zu (ii): Für $n \neq 1$ berechnet man $\|E\|_2 = \sqrt{n} \neq 1 = \|E\|_{\text{nat}}$. □

6.5 Eigenwerte und Eigenvektoren

Im Folgenden betrachten wir ausschließlich quadratische Matrizen $A \in \mathrm{Mat}_{\mathbb{K}}(n)$.

Definition 6.31 (Eigenwert, Eigenvektor)
Der Skalar $\lambda \in \mathbb{K}$ ist Eigenwert von A zu dem zugehörigen Eigenvektor $w \in \mathbb{K}^n \setminus \{0\}$, wenn $Aw = \lambda w$ gilt.

Definition 6.32 (charakteristisches Polynom)
Das charakteristische Polynom der Matrix A ist definiert als

$$p_A(\lambda) = \det(\lambda E - A) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0.$$

Proposition 6.33 (Eigenwerte sind Nullstellen des charakteristischen Polynoms)
Ein Skalar λ ist genau dann ein Eigenwert von A, wenn $\det(\lambda E - A) = 0$ gilt. Insbesondere sind die Eigenwerte genau die Nullstellen des charakteristischen Polynoms.

Beweis. Es gilt die Äquivalenzkette

$$\begin{aligned}
&\det(\lambda E - A) = 0 \\
\iff &\lambda E - A \text{ ist nicht injektiv} \\
\iff &\exists b \in \mathbb{K}^n \setminus \{0\} : (\lambda E - A)b = 0 \\
\iff &\exists b \in \mathbb{K}^n \setminus \{0\} : Ab = \lambda E b = \lambda b \\
\iff &\lambda \text{ ist Eigenwert zum Eigenvektor } b.
\end{aligned}$$

□

Definition 6.34 (Spektrum einer Matrix)
Das Spektrum der Matrix A ist die Menge aller Eigenwerte von A.

Definition 6.35 (Vielfachheit von λ)
Die Vielfacheit eines Eigenwerts λ ist die eindeutig bestimmte Zahl k, so dass

$$p_A(\lambda) = p_A'(\lambda) = \cdots = p_A^{(k-1)}(\lambda) = 0 \quad \textit{und} \quad p_A^{(k)}(\lambda) \neq 0.$$

Proposition 6.36 (Eigenwerte liegen in einer Kreisscheibe)
Jeder Eigenwert λ einer quadratischen Matrix A genügt für jede beliebige, mit einer Vektornorm verträglichen Matrizennorm $\|\cdot\|$ der Ungleichung $|\lambda| \leq \|A\|$, d.h. die Eigenwerte von A liegen in einer Kreisscheibe um 0 mit Radius $\|A\|$.

Beweis. Für einen Eigenwert λ gilt die Ungleichungskette

$$|\lambda| = \frac{1}{\|x\|}|\lambda|\,\|x\| = \frac{1}{\|x\|}\,\|\lambda x\| = \frac{1}{\|x\|}\,\|Ax\| \leq \frac{1}{\|x\|}\,\|A\|\,\|x\| = \|A\|$$

□

Proposition 6.37 (Eigenvektoren bilden eine Orthonormalbasis)
Ist A eine symmetrische bzw. hermitische Matrix, so hat A genau n verschiedene Eigenwerte $\lambda_1, \ldots, \lambda_n$. Die zugehörigen Eigenvektoren $w_1, \ldots, w_n$ bilden eine Orthonormalbasis des $\mathbb{K}^n$.

Beweis. Diese Eigenschaften sind bekannt aus der linearen Algebra und sollen hier nicht bewiesen werden. □

Bemerkung 6.38 (Eigenschaften der Eigenvektoren)
Seien A und $\lambda_1, \dots, \lambda_n$ und $w_1, \dots, w_n$ wie in der vorherigen Proposition. Sei $x \in \mathbb{K}^n$ beliebig. Dann lassen sich die folgenden Gleichungen leicht nachrechnen:

$$\begin{aligned} x &= \sum_{j=1}^{n} (x, w_j)\, w_j \\ \|x\|_2^2 &= \sum_{j=1}^{n} |\,(x, w_j)\,|^2 \\ Ax &= \sum_{j=1}^{n} \lambda_j\, (x, w_j)\, w_j \\ \|Ax\|_2^2 &= \sum_{j=1}^{n} \lambda_j^2 |\,(x, w_j)\,|^2 \\ (Ax, x) &= \sum_{j=1}^{n} \lambda_j |\,(x, w_j)\,|^2 \end{aligned}$$

Definition 6.39 (Spektralnorm)
Für eine symmetrische bzw. hermitische Matrix A ist die Spektralnorm definiert durch

$$\|A\|_{spec} = \max_{j=1,\dots,m} |\lambda_j|,$$

wobei $\lambda_1, \dots, \lambda_m$ die Eigenwerte von A sind.

Dass dieses eine Norm ist, folgt aus der folgenden Proposition:

Proposition 6.40
Für eine symmetrische bzw. hermitische Matrix ist die Spektralnorm $\|\cdot\|_{spec}$ auf $\mathrm{Mat}_{\mathbb{K}}(n)$ gerade die natürliche Norm bezüglich der euklidischen bzw. unitären Norm $\|\cdot\|_2$ auf $\mathbb{K}^n$, d.h. für alle symmetrischen bzw. hermitischen Matrizen A gilt

$$\|A\|_{nat} = \|A\|_{spec}\,.$$

Beweis. Wir zeigen zuerst die Ungleichung $\|A\|_{\mathrm{nat}} \le \|A\|_{\mathrm{spec}}$. (Dieses ist keineswegs trivial, denn wir haben noch nicht gezeigt, dass $\|\cdot\|_{\mathrm{spec}}$ eine Norm ist.) Es gilt

$$\begin{aligned} \|Ax\|_2^2 = \sum_{j=1}^{n} \lambda_j^2 |\,(x, w_j)\,|^2 &\le \max_{j=1,\dots,n} |\lambda_j|^2 \sum_{j=1}^{n} |\,(x, w_j)\,|^2 = \|A\|_{\mathrm{spec}}^2 \|x\|_2^2 \\ \implies \|Ax\|_2 &\le \|A\|_{\mathrm{spec}} \|x\|_2 \\ \implies \|A\|_{\mathrm{nat}} = \sup_{0 \neq x \in \mathbb{K}} \frac{\|Ax\|_2}{\|x\|_2} &\le \frac{\|A\|_{\mathrm{spec}} \|x\|_2}{\|x\|_2} = \|A\|_{\mathrm{spec}}\,. \end{aligned}$$

Weiter gilt auch $\|A\|_{\mathrm{spec}} \le \|A\|_{\mathrm{nat}}$, denn für alle Eigenwerte gilt nach Proposition 6.36 $|\lambda_i| \le \|A\|_{\mathrm{nat}}$.

Somit folgt insgesamt $\|A\|_{\mathrm{nat}} = \|A\|_{\mathrm{spec}}$. Insbesondere ist die Spektralnorm wirklich eine Norm. □

7 Gauß-Elimination

Bemerkung 7.1 (Notationen)
Für eine Matrix $A \in \mathrm{Mat}_{\mathbb{K}}(n)$ verwenden wir die Schreibweisen:

- $A = (a^1|\cdots|a^n)$, wobei $a^k = (a_{1k}, \ldots, a_{nk})^{\mathrm{T}}$ die k-te Spalte der Matrix A ist.
- $A = (a_1|\cdots|a_n)^{\mathrm{T}}$, wobei $a_k = (a_{j1}, \ldots, a_{jn})$ die j-te Zeile der Matrix A ist.

Proposition 7.2 (Eindeutige Lösbarkeit eines Gleichungssystems)
Sei A eine reguläre $n \times n$-Matrix. Dann ist für jedes $f \in \mathbb{K}^n$ das Gleichungssystem $Ax = f$ eindeutig lösbar.

Beweis. A ist regulär, d.h. $A \in \mathrm{Mat}_{\mathbb{K}}(n)$ und $\det(A) \neq 0$. Dann sind die n Spalten von A linear unabhänging, bilden also eine Basis des $\mathbb{K}^n$. Somit lässt sich jedes $f \in \mathbb{K}^n$ als eindeutige Linearkombination dieser Spalten schreiben. Zu $f = (f_1, \ldots, f_n)$ gibt es daher ein eindeutig bestimmtes $x = (x_1, \ldots, x_n)$, so dass die folgenden äquivalenten Aussagen gelten:

$$f = \sum_{k=1}^{n} x_k a^k \iff f_j = \sum_{k=1}^{n} x_k a_{jk} \iff f = Ax$$

□

7.1 Gauß-Algorithmus zum Lösen eines Gleichungssystems

Der Gauß-Algorithmus dient zur Lösung eines solchen Gleichungssystems. Er ist ein Verfahren, wie ein n-reihiges Gleichungssystem auf ein $(n-1)$-reihiges zurückgeführt werden kann. Durch Rekursion erhält man nach $n-1$ solchen Schritten dieser *Vorwärtseliminaton* ein lineares Gleichungssystem mit nur einer Unbekannten. Daraufhin lassen sich durch rückwärtiges Einsetzen rekursiv alle Variablen bestimmen.

Sei $A = (a_{jk})$ eine reguläre Matrix, $f = (f_1, \ldots, f_n) \in \mathbb{K}^n$ ein Vektor. Gesucht ist die eindeutig bestimmte Lösung des Gleichungssystems $Ax = f$.

Erste Stufe der Vorwärtselimination

Aus Gründen der einheitlichen Notation setzen wir:

$$\begin{aligned} \widetilde{A}^1 &:= A \quad \text{mit } \widetilde{a}^1_{jk} := a_{jk} \\ \widetilde{f}^1 &:= f \quad \text{mit } \widetilde{f}^1_j := f_j \\ x^1 &:= (x_1, \ldots, x_n) \end{aligned}$$

Das Gleichungsystem ist somit

$$\widetilde{A}^1 x^1 = \widetilde{f}^1 \iff \begin{pmatrix} \widetilde{a}^1_{11} & \ldots & \widetilde{a}^1_{1n} \\ \vdots & \ddots & \vdots \\ \widetilde{a}^1_{n1} & \ldots & \widetilde{a}^1_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \widetilde{f}^1_1 \\ \vdots \\ \widetilde{f}^1_n \end{pmatrix}.$$

Schritt 1.1: Erste Zeilenvertauschung: Da $\widetilde{A}^1$ eine reguläre Matrix ist, besteht die erste Spalte nicht nur aus Nullen. Da die Lösung des Gleichungssystems unabhängig von der Reihenfolge der Gleichungen ist, kann man gegebenenfalls eine Zeilenvertauschung so vornehmen, dass ein von Null verschiedenes Element der ersten Spalte von $\widetilde{A}^1$ in der ersten Zeile steht. Dieses ist möglich, indem man die Zeile, deren erstes Element von Null verschieden ist, mit der ersten Zeile vertauscht.

Durch diese Vertauschung entsteht ein modifiziertes Gleichungsystem. Sei $A^1 = (a^1_{jk})_{1\le j,k\le n}$ die so entstehende Koeffizientenmatrix, $f^1 = (f^1_1, \ldots, f^1_n)$ der durch die Zeilenvertauschung entstehende Vektor.

Das neue Gleichungssystem ist

$$A^1 x^1 = f^1 \iff \begin{pmatrix} a^1_{11} & \ldots & a^1_{1n} \\ \vdots & \ddots & \vdots \\ a^1_{n1} & \ldots & a^1_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} f^1_1 \\ \vdots \\ f^1_n \end{pmatrix}.$$

Da die erste Gleichung nicht mehr verändert wird, definiert man:

$$\begin{aligned} b_{1k} &:= a^1_{1k} \quad (\text{für } k = 1, \ldots, n) \\ g_1 &:= f^1_1 \end{aligned}$$

Schritt 1.2: Erste Elimination: Für die Zeilen $j = 2, \ldots, n$ eliminiert man durch Subtraktion eines Vielfachen der ersten Zeile nun das Element der ersten Spalte, d.h. den Koeffizienten von x_1. Für die j-te Zeile setzt man

$$m_{j1} = \frac{a^1_{j1}}{b_{11}} \quad (\text{für } j = 2, \ldots, n).$$

Subtrahiert man nun für $j = 2, \ldots, n$ von der j-ten Zeile das m_{j1}-fache der ersten Zeile, so ergibt sich ein neues Gleichungssystem

$$\begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ 0 & a^1_{22} - m_{21}b_{12} & \ldots & a^1_{2n} - m_{21}b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a^1_{n2} - m_{n1}b_{12} & \ldots & a^1_{nn} - m_{n1}b_{1n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} g_1 \\ f^1_2 - m_{21}g_1 \\ \vdots \\ f^1_n - m_{n1}g_1 \end{pmatrix}.$$

Mit den Bezeichnungen

$$\begin{aligned} \widetilde{A}^2 := (\widetilde{a}^2_{jk})_{2\le j,k\le n} \quad &\text{mit} \quad \widetilde{a}^2_{jk} = a^1_{jk} - m_{j1}b_{1k} \\ \widetilde{f}^2 = (\widetilde{f}^2_2, \ldots, \widetilde{f}^2_n) \quad &\text{mit} \quad \widetilde{f}^2_j = f^1_j - m_{j1}g_1 \end{aligned}$$

wird dieses Gleichungssystem zu

$$\begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ 0 & \widetilde{a}^2_{22} & \ldots & \widetilde{a}^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{a}^2_{n2} & \ldots & \widetilde{a}^2_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} g_1 \\ \widetilde{f}^2_2 \\ \vdots \\ \widetilde{f}^2_n \end{pmatrix}.$$

Die Variable x_1 ist damit aus der zweiten bis n-ten Gleichung eliminiert worden, so dass die Unbekannten $x^2 := (x_2, \ldots, x_n)$ sich als Lösungen des $(n-1)$-reihigen Gleichungssystems der zweiten bis n-ten Zeile bestimmen lassen.

Die Matrix $\widetilde{A}^2$ ist regulär, so dass sich das so entstandene Gleichungssystem

$$\widetilde{A}^2 x^2 = \widetilde{f}^2 \iff \begin{pmatrix} \widetilde{a}^2_{22} & \ldots & \widetilde{a}^2_{2n} \\ \vdots & \ddots & \vdots \\ \widetilde{a}^2_{n2} & \ldots & \widetilde{a}^2_{nn} \end{pmatrix} \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \widetilde{f}^2_2 \\ \vdots \\ \widetilde{f}^2_n \end{pmatrix}$$

analog reduzieren lässt.

Zusätzlich speichert man die Gleichung

$$b_{11}x_1 + \cdots + b_{1n}x_n = g_1.$$

Allgemeine Rekursion der Vorwärtselimination

Vor dem t-ten Schritt sind die folgenden Gleichungen gespeichert:

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + \cdots + b_{1n}x_n &= g_1 \\ b_{22}x_2 + \cdots + b_{2n}x_n &= g_2 \\ \ddots \qquad \vdots \qquad &\vdots \\ b_{t-1,t-1}x_{t-1} + \cdots + b_{t-1,n}x_n &= g_{t-1} \end{aligned}$$

Solange $t < n$ ist, wird im t-ten Schritt das Gleichungssystem

$$\widetilde{A}^t x^t = \widetilde{f}^t \iff \begin{pmatrix} \widetilde{a}^t_{tt} & \ldots & \widetilde{a}^t_{tn} \\ \vdots & \ddots & \vdots \\ \widetilde{a}^t_{nt} & \ldots & \widetilde{a}^t_{nn} \end{pmatrix} \begin{pmatrix} x_t \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \widetilde{f}^t_t \\ \vdots \\ \widetilde{f}^t_n \end{pmatrix}$$

auf ein $(n-t)$-reihiges Gleichungssystem reduziert.

Schritt t.1: t-te Zeilenvertauschung: Da $\widetilde{A}^t$ eine reguläre Matrix ist, besteht die erste Spalte nicht nur aus Nullen. Da die Lösung unabhängig von der Reihenfolge der Gleichungen ist, kann man gegebenenfalls eine Zeilenvertauschung so vornehmen, dass ein von Null verschiedenes Element der ersten Spalte von $\widetilde{A}^t$ in die erste Zeile kommt. Dieses ist möglich, indem man die Zeile, deren erstes Elememt von Null verschieden ist, mit der ersten Zeile vertauscht.

Durch diese Vertauschung entsteht ein modifiziertes Gleichungsystem. Sei $A^t = (a_{jk})_{t \le j,k \le n}$ die so entstehende Koeffizientenmatrix, $f^t = (f^t_t, \ldots, f^t_n)$ der durch die Zeilenvertauschung entstehende Vektor.

Das neue Gleichungssystem lässt sich schreiben als

$$A^t x^t = f^t \iff \begin{pmatrix} a_{tt}^t & \dots & a_{tn}^t \\ \vdots & \ddots & \vdots \\ a_{nt}^t & \dots & a_{nn}^t \end{pmatrix} \begin{pmatrix} x_t \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} f_t^t \\ \vdots \\ f_n^t \end{pmatrix}.$$

Da die erste Gleichung nicht mehr verändert wird, definiert man:

$$\begin{aligned} b_{tk} &:= a_{tk}^t \quad (\text{für } k = t, \dots, n) \\ g_t &:= f_t^t \end{aligned}$$

Schritt t.2: t-te Elimination: Für die Zeilen $j = t+1, \dots, n$ eliminiert man durch Subtraktion eines Vielfachen der Zeile t nun das Element in Spalte t. Für die Zeile j setzt man

$$m_{jt} = \frac{a_{jt}^t}{b_{tt}} \quad (\text{für } j = t+1, \dots, n).$$

Subtrahiert man nun für $j = t+1, \dots, n$ von der Zeile j das m_{jt}-fache der Zeile t, so ergibt sich ein neues Gleichungssystem:

$$\begin{pmatrix} b_{tt} & b_{t,t+1} & \dots & b_{tn} \\ 0 & a_{t+1,t+1}^t - m_{t+1,t} b_{t,t+1} & \dots & a_{t+1,n}^t - m_{t+1,t} b_{tn} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n,t+1}^t - m_{n,t} b_{t,t+1} & \dots & a_{nn}^t - m_{nt} b_{tn} \end{pmatrix} \begin{pmatrix} x_t \\ x_{t+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} g_t \\ f_{t+1}^t - m_{t+1,t} g_t \\ \vdots \\ f_n^t - m_{nt} g_t \end{pmatrix}.$$

Mit den Bezeichnungen

$$\begin{aligned} \widetilde{A}^{t+1} &:= (\widetilde{a}_{jk}^{t+1})_{t+1 \le j,k \le n} \quad \text{mit} \quad \widetilde{a}_{jk}^{t+1} = a_{jk}^t - m_{jt} b_{tk} \\ \widetilde{f}^{t+1} &= (\widetilde{f}_{t+1}^{t+1}, \dots, \widetilde{f}_n^{t+1}) \quad \text{mit} \quad \widetilde{f}_j^{t+1} = f_j^t - m_{jt} g_t \end{aligned}$$

wird dieses Gleichungssystem zu

$$\begin{pmatrix} b_{tt} & b_{t,t+1} & \dots & b_{tn} \\ 0 & \widetilde{a}_{t+1,t+1}^{t+1} & \dots & \widetilde{a}_{t+1,n}^{t+1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{a}_{n,t+1}^{t+1} & \dots & \widetilde{a}_{nn}^{t+1} \end{pmatrix} \begin{pmatrix} x_t \\ x_{t+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} g_t \\ \widetilde{f}_{t+1}^{t+1} \\ \vdots \\ \widetilde{f}_n^{t+1} \end{pmatrix}.$$

Die Variable x_t ist damit aus den unteren Gleichung eliminiert worden, so dass die Unbekannten $x^{t+1} := (x_{t+1}, \dots, x_n)$ sich als Lösungen des $(n-t)$- reihigen Gleichungssystems der Zeilen $t+1$ bis n bestimmen lassen. Wende die Reduzierung erneut auf das folgende Gleichungssystem an:

$$\begin{pmatrix} \widetilde{a}_{t+1,t+1}^{t+1} & \dots & \widetilde{a}_{t+1,n}^{t+1} \\ \vdots & \ddots & \vdots \\ \widetilde{a}_{n,t+1}^{t+1} & \dots & \widetilde{a}_{nn}^{t+1} \end{pmatrix} \begin{pmatrix} x_{t+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \widetilde{f}_t^{t+1} \\ \vdots \\ \widetilde{f}_n^{t+1} \end{pmatrix}$$

Rückwärtseinsetzen

Nach Beendigung obiger Rekursion erhält man das gestaffelte Gleichungssystem

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + \dots + b_{1n}x_n &= g_1 \\ b_{22}x_2 + \dots + b_{2n}x_n &= g_2 \\ \ddots \qquad \vdots \qquad &\vdots \\ b_{nn}x_n &= g_n \end{aligned}$$

Dieses lässt sich nun rückwärts für $t = n, \dots, 1$ lösen. Berechne x_t durch

$$x_t = \frac{1}{b_{tt}} \left(g_t - \sum_{k=t+1}^{n} b_{tk} x_k \right).$$

Bemerkung 7.3 (Spaltenpivotsuche)
In der Praxis ist es oft sinnvoll, die Faktoren m_{jt} klein zu halten. Dieses geschieht mittels Spaltenpivotsuche. Dabei führt man die Zeilenvertauschungen jeweils so aus, dass das betraglich größte Element der ersten Spalte in der ersten Zeile steht, d.h. so, dass

$$a_{tt}^t = \max_{j=t,\dots,n} |a_{jt}^t| = \max_{j=t,\dots,n} |\widetilde{a}_{jt}^t|.$$

Mit dieser Pivotisierung ist für alle j, t: $|m_{jk}| \leq 1$. Man nennt die erste Gleichung des Systems $A^t x^t = f^t$ die *t-te Pivotgleichung*, $b_{tt} = a_{tt}^t$ das *t-te Pivotelement*.

7.2 Praktische Durchführung des Gauß-Algorithmus

Für die praktische Durchführung schreibt man das Gleichungssystem $Ax = f$ in der Form $(A|f)$.

Die Ausgangsform des Gauß-Algorithmus schreibt sich in der Form

$$(\widetilde{A}^1|\widetilde{f}^1) \iff \left(\begin{array}{cccc|c} \widetilde{a}_{11}^1 & \widetilde{a}_{12}^1 & \dots & \widetilde{a}_{1n}^1 & \widetilde{f}_1^1 \\ \widetilde{a}_{21}^1 & \widetilde{a}_{22}^1 & \dots & \widetilde{a}_{2n}^1 & \widetilde{f}_2^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \widetilde{a}_{n1}^1 & \widetilde{a}_{n2}^1 & \dots & \widetilde{a}_{nn}^1 & \widetilde{f}_n^1 \end{array} \right).$$

Nach der ersten Elimination stehen in der ersten Spalte unterhalb der ersten Zeile nur Nullen. An deren Stelle speichert man in der ersten Spalte die Koeffizienten $m_{j1}^1 := m_{j1}$.

Somit ergibt sich nach der ersten Elimination das System

$$\left(\begin{array}{c|ccc|c} b_{11} & b_{12} & \dots & b_{1n} & g_1 \\ \hline m_{21}^1 & & & & \\ \vdots & & \widetilde{A}^2 & & \widetilde{f}^2 \\ m_{n1}^1 & & & & \end{array}\right) \Longleftrightarrow \left(\begin{array}{c|ccc|c} b_{11} & b_{12} & \dots & b_{1n} & g_1 \\ \hline m_{21}^1 & \widetilde{a}_{22}^2 & \dots & \widetilde{a}_{2n}^2 & \widetilde{f}_2^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n1}^1 & \widetilde{a}_{n2}^2 & \dots & \widetilde{a}_{nn}^2 & \widetilde{f}_n^2 \end{array}\right).$$

In der nun folgenden Zeilenvertauschung bringt man ein von Null verschiedenes Element in die zweite Spalte der ersten Zeile. Dabei müssen die Koeffizienten m_{j1}^1 der entsprechenden Zeilen mitgetauscht werden, die durch Vertauschung entstehenden Koeffizienten werden mit m_{j1}^2 bezeichnet. Wir erhalten das Gleichungssystem

$$\left(\begin{array}{c|ccc|c} b_{11} & b_{12} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & & & & \\ \vdots & & A^2 & & f^2 \\ m_{n1}^2 & & & & \end{array}\right) \Longleftrightarrow \left(\begin{array}{c|ccc|c} b_{11} & b_{12} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & a_{22}^2 & \dots & a_{2n}^2 & f_2^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n1}^2 & a_{n2}^2 & \dots & a_{nn}^2 & f_n^2 \end{array}\right).$$

Bei der Elimination werden die Koeffizienten m_{j1}^2 der entsprechenden Zeilen nicht verändert. Wir erhalten mit den neuen Koeffizienten $m_{j2}^2 := m_{j2}$ für $j = 3, \dots, n$ die Matrix

$$\left(\begin{array}{c|c|ccc|c} b_{11} & b_{12} & b_{13} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & b_{22} & b_{23} & \dots & b_{2n} & g_2 \\ \hline m_{31}^2 & m_{32}^2 & & & & \\ \vdots & \vdots & & \widetilde{A}^3 & & \widetilde{f}^3 \\ m_{n1}^2 & m_{n2}^2 & & & & \end{array}\right) \Longleftrightarrow \left(\begin{array}{c|c|ccc|c} b_{11} & b_{12} & b_{13} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & b_{22} & b_{23} & \dots & b_{2n} & g_2 \\ \hline m_{31}^2 & m_{32}^2 & \widetilde{a}_{33}^3 & \dots & \widetilde{a}_{3n}^3 & \widetilde{f}_3^3 \\ \vdots & \vdots & & \ddots & & \vdots \\ m_{n1}^2 & m_{n2}^2 & \widetilde{a}_{n3}^3 & \dots & \widetilde{a}_{nn}^3 & \widetilde{f}_n^3 \end{array}\right).$$

Bei der nun folgenden Zeilenvertauschung bleiben die ersten beiden Zeilen fest. Daher wird der Koeffizient m_{21}^2 nicht mitgetauscht, bleibt also fest. Die anderen Koeffizienten m_{jk}^2 müssen entsprechend den Zeilen vertauscht werden, die neuen Koeffizienten bezeichnen wir mit m_{jk}^3. Man erhält nach der Vertauschung

$$\left(\begin{array}{c|c|ccc|c} b_{11} & b_{12} & b_{13} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & b_{22} & b_{23} & \dots & b_{2n} & g_2 \\ \hline m_{31}^3 & m_{32}^3 & & & & \\ \vdots & \vdots & & A^3 & & f^3 \\ m_{n1}^3 & m_{n2}^3 & & & & \end{array}\right) \Longleftrightarrow \left(\begin{array}{c|c|ccc|c} b_{11} & b_{12} & b_{13} & \dots & b_{1n} & g_1 \\ \hline m_{21}^2 & b_{22} & b_{23} & \dots & b_{2n} & g_2 \\ \hline m_{31}^3 & m_{32}^3 & a_{33}^3 & \dots & a_{3n}^3 & f_3^3 \\ \vdots & \vdots & & \ddots & & \vdots \\ m_{n1}^3 & m_{n2}^3 & a_{n3}^3 & \dots & a_{nn}^3 & f_n^3 \end{array}\right).$$

Führt man dieses Schema bis zum Ende durch, erhält man die Matrix

$$\left(\begin{array}{ccccccc} b_{11} & b_{12} & b_{13} & b_{14} & \dots & b_{1n} & g_1 \\ m_{21}^2 & b_{22} & b_{23} & b_{24} & \dots & b_{2n} & g_2 \\ m_{31}^3 & m_{32}^3 & b_{33} & b_{34} & \dots & b_{3n} & g_3 \\ m_{41}^4 & m_{42}^4 & m_{43}^4 & b_{44} & \dots & b_{4n} & g_4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n1}^n & m_{n2}^n & m_{n3}^n & \dots & m_{n,n-1}^n & b_{nn} & g_n \end{array}\right).$$

7.3 Dreieckszerlegung

Bemerkung 7.4 (Notation)
Um nicht noch weitere Symbole einzuführen, fassen wir außerdem von nun die Vektoren f^t und $\widetilde{f}^t$ als Vektoren des $\mathbb{K}^n$ auf und füllen sie dazu von vorne mit Nullen auf, d.h. von nun an ist:

$$f^t = (0, \dots, 0, f_t^t, f_{t+1}^t, \dots f_n^t)$$
$$\widetilde{f}^t = (0, \dots, 0, \widetilde{f}_t^t, \widetilde{f}_{t+1}^t, \dots \widetilde{f}_n^t)$$

Außerdem setzt man zur Vereinfachung der Schreibweise noch:

$$\begin{aligned} j_t &:= \text{die Nummer der Zeile, mit der im } t\text{-ten Schritt vertauscht wurde} \\ l_{jk} &:= m_{jk}^j \quad (\text{für } 1 \le k < j \le n) \\ m^t &:= (0, \dots, 0, 1, m_{t+1,t}, \dots, m_{nt}) \\ l^t &:= (0, \dots, 0, 1, l_{t+1,t}, \dots, l_{nt}) \end{aligned}$$

Sei nun P_t die Matrix, die die t-te Zeile eines Vektors mit der j_t-ten vertauscht, d.h.

$$P_t(x_1, \dots, x_t, \dots, x_{j_t}, \dots, x_n)^{\mathrm{T}} = (x_1, \dots, x_{j_t}, \dots, x_t, \dots, x_n)^{\mathrm{T}}$$

Die im Gauß-Algorithmus durchgeführten Schritte lassen sich für den Vektor f wiefolgt beschreiben:

- Vertauschung: $f^t = P_t \widetilde{f}^t$
- Elimination: $\widetilde{f}^{t+1} = f^t - g_t m^t$

Bemerkung 7.5 (Form der Permutationsmatrix)
Die Permutationsmatrix, die die Zeilen i und j vertauscht, hat die Form

$$\begin{pmatrix} 1 & 0 & & & & & 0 \\ 0 & \ddots & & & & & \\ & & 0 & \dots & 1 & & \\ & & \vdots & \ddots & \vdots & & \\ & & 1 & \dots & 0 & & \\ & & & & & \ddots & 0 \\ 0 & & & & & 0 & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow j\text{-te Zeile} \\ \\ \leftarrow i\text{-te Zeile} \\ \\ \\ \end{matrix}$$

Es steht in jeder Zeile genau eine Eins. Diese stehen überall auf der Hauptdiagonalen, nur in der i-ten Zeile steht die Eins in der j-ten Spalte und in der j-ten Zeile steht die Eins in der i-ten Spalte.

Proposition 7.6
Aus dem Gauß-Algorithmus lassen sich folgende Formeln herleiten:

(i) Es gilt für $t = 1, \dots, n$:

$$f^t = P_t \cdots P_1 f - \sum_{k=1}^{t-1} g_k P_t \cdots P_{k+1} m^k$$

(ii) Mit $P := P_n \cdots P_1 = \prod_{i=1}^{n} P_i$ gilt insbesondere

$$Pf = \sum_{k=1}^{n} g_k l^k$$

Beweis. Zu (i): Induktionsverankerung: $t = 1$: Es gilt

$$f^1 = P_1 \widetilde{f}^1 = P_1 f.$$

Induktionsschritt: $t \to t+1$: Für $t < n$ gilt

$$\begin{aligned} f^{t+1} &= P_{t+1} \widetilde{f}^{t+1} = P_{t+1}(f^t - g_t m^t) = P_{t+1} f^t - g_t P_{t+1} m^t \\ &\overset{IV}{=} P_{t+1} \left(P_t \cdots P_1 f - \sum_{k=1}^{t-1} g_k P_t \cdots P_{k+1} m^k \right) - g_t P_{t+1} m^t \\ &= P_{t+1} \cdot P_t \cdots P_1 f - \sum_{k=1}^{t-1} g_k P_{t+1} \cdots P_{k+1} m^k - g_t P_{t+1} m^t \\ &= P_{t+1} \cdot P_t \cdots P_1 f - \sum_{k=1}^{t} g_k P_{t+1} \cdots P_{k+1} m^k. \end{aligned}$$

Damit folgt die erste Behauptung.

Zu (ii): Man beachte noch, dass l^k aus m^k gerade durch Anwendung aller noch folgenden Zeilenvertauschungen $P_n, \dots, P_{k+1}$ entsteht. Somit gilt

$$P_n \cdots P_{k+1} m^k = l^k.$$

Mit $f^n = g_n m^n = (0, \dots, 0, g_n) = g_n l^n$ und $t = n$ folgt

$$\begin{aligned} f^n &= P_n \cdots P_1 f - \sum_{k=1}^{n-1} g_k P_n \cdots P_{k+1} m^k \\ \Longrightarrow g_n l^n &= Pf - \sum_{k=1}^{n-1} g_k l^k \\ \Longrightarrow Pf &= \sum_{k=1}^{n} g_k l^k. \end{aligned}$$

□

Wir setzen nun

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & 1 \end{pmatrix} \quad \text{und} \quad U = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & b_{22} & b_{23} & \dots & b_{2n} \\ 0 & 0 & b_{33} & \dots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & b_{nn} \end{pmatrix}.$$

Dann besagt Proposition 7.6.(ii) gerade die Identität $Pf = Lg$. Das aufgelöste Gleichungssystem $Ax = f$ ist $g = Ux$. Da P, L, U nur von A abhängig sind, gilt für jedes f und somit für jedes $x \in \mathbb{K}^n$:

$$PAx = Pf = Lg = LUx \implies PA = LU$$

Somit haben wir eine Dreieckszerlegung für PA gefunden.

Ist während des Gauß-Algorithmus keine Zeilenvertauschung notwendig, so ist P die Einheitsmatrix und wir erhalten für A die Dreieckszerlegung $A = LU$.

Proposition 7.7 (Eindeutigkeit der Dreieckszerlegung)
Gibt es eine Dreieckszerlegung $A = LU$, so sind die Faktoren L und U eindeutig bestimmt.

Beweis. Skizzenhaft: Wegen $\det L = 1$ existiert L^{-1}. Da A regulär ist, existiert A^{-1}. Somit existiert wegen $U = L^{-1}A$ auch $U^{-1} = A^{-1}L$.

Existieren nun zwei Zerlegungen $A = L_1U_1$ und $A = L_2U_2$. Dann

$$\begin{aligned} L_1U_1 &= A = L_2U_2 \\ \Longrightarrow L_1^{-1}L_2 &= U_1U_2^{-1} \end{aligned}$$

Nun ist auch U_2^{-1} eine rechte obere Dreiecksmatrix, somit auch $U_1U_2^{-1}$. Außerdem ist L_1^{-1} eine linke untere Dreiecksmatrix, somit auch $L_1^{-1}L_2$.

Wegen $L_1^{-1}L_2 = U_1U_2^{-1}$ sind beide Matrizen Diagonalmatrizen, da bei $L_1^{-1}L_2$ die Diagonale nur aus Einsen besteht, ist $U_1U_2^{-1} = L_1^{-1}L_2 = E$. Somit folgt $L_1 = L_2$ und $U_1 = U_2$, also die Eindeutigkeit der Zerlegung. □

Bemerkung 7.8 (Methode zur Bestimmung der Permutationsmatrix P)
Zur Bestimmung der Permutationsmatrix P ergänzt man das Schema $(A|f)$ um eine $(n+2)$-te Spalte $(1, 2, \ldots, n)^T$. Die Elemente dieser Spalte werden bei jeder Zeilenvertauschung mit vertauscht, bei der Elimination bleiben sie unverändert. Nach Beendigung der Vorwärtselimination steht in der $(n+2)$-ten Spalte eine Permutation der ersten n natürlichen Zahlen. Die Spalte ist

$$P \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} = P_{n-1} \cdots P_1 \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix},$$

d.h. die Permutationsmatrix ist $P = (\delta_{p_j k})$.

7.4 Simultane Lösung und Inversion von Matrizen

Bei einigen Anwendungen des Gaußschen Eliminationsverfahrens stellt sich die Aufgabe zu einer gegebenen regulären Matrix A und mehreren rechten Seiten $f^{(l)}$ die Lösung $x^{(l)}$ der zugehörigen linearen Gleichungssysteme $Ax^{(l)} = f^{(l)}$ für $l = 1, \ldots, m$ zu bestimmen.

Dazu wendet man den Gauß-Algorithmus gleichzeitig auf die m rechten Seiten an und transformiert dadurch die erweiterte Matrix $(A|F)$ auf die Gestalt $(U|G)$.

$$\underbrace{\left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1n} & f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} & f_{n1} & \cdots & f_{nm} \end{array}\right)}_{\begin{matrix} A \end{matrix} \quad \begin{matrix} F \end{matrix}} \xrightarrow[\text{Gauß}]{} \underbrace{\left(\begin{array}{ccc|ccc} b_{11} & \cdots & b_{1n} & g_{11} & \cdots & g_{1m} \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ * & & b_{nn} & g_{n1} & \cdots & g_{nm} \end{array}\right)}_{\begin{matrix} U \end{matrix} \quad \begin{matrix} G \end{matrix}}$$

Die zugehörigen gestaffelten Gleichungssysteme lassen sich dann rekursiv nach x auflösen.

Das simultane Eliminationsverfahren gestattet weiter die Berechnung der Inversen A^{-1} der Matrix A. Die Lösung der inhomogenen Gleichung $Ax = f$ hat die Darstellung $x = A^{-1}f$. Wählt man als rechte Seite die n Einheitsvektoren $e_l = (\delta_{1l}, \ldots, \delta_{nl})$ für $l = 1, \ldots, n$, also $F = E$, dann werden die zugehörigen Lösungen $x^{(l)}$ gerade die Spaltenvektoren der inversen Matrix.

7.5 Berechnung von Determinanten

Proposition 7.9 (Verhalten von Determinanten unter elementaren Zeilenumformungen)
Sei $A = (a_1|\cdots|a_n)^{\mathrm{T}}$ eine Matrix mit den Zeilen $a_1, \ldots, a_n$. Dann gilt:

(i) Für $l \neq k$ ist $det(A) = det(a_1|\cdots|a_k + \lambda a_l|\cdots|a_n)^{\mathrm{T}}$, d.h. durch Addition eines Vielfachen einer anderen Zeile ändert sich die Determinante einer Matrix nicht.

(ii) Für $l \neq k$ ist $det(A) = det(a_1|\cdots|a_k|\cdots|a_l|\cdots|a_n)^{\mathrm{T}} = -det(a_1|\cdots|a_l|\cdots|a_k|\cdots|a_n)^{\mathrm{T}}$, d.h. bei Vertauschung zweier Zeilen ändert sich das Vorzeichen der Determinante.

Nun lässt sich nach obigem Gauß-Algorithmus die Determinante einer Matrix A bestimmen. Dazu ist der Vektor f zu vernachlässigen.

Für die Determinante von A gilt nach dem Laplaceschen Entwicklungssatz nach der ersten Spalte

$$\det A = \det \widetilde{A}^1 = \sigma_1 \det A^1 = \sigma_1 \begin{vmatrix} a_{11}^1 & \dots & a_{1n}^1 \\ \vdots & \ddots & \vdots \\ a_{n1}^1 & \dots & a_{nn}^1 \end{vmatrix} = \sigma_1 \begin{vmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ 0 & a_{22}^1 - m_{21}b_{12} & \dots & a_{2n}^1 - m_{21}b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^1 - m_{n1}b_{12} & \dots & a_{nn}^1 - m_{n1}b_{1n} \end{vmatrix}$$

$$= \sigma_1 \begin{vmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ 0 & \widetilde{a}_{22}^2 & \dots & \widetilde{a}_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{a}_{n2}^2 & \dots & \widetilde{a}_{nn}^2 \end{vmatrix} = \sigma_1 b_{11} \begin{vmatrix} \widetilde{a}_{11}^2 & \dots & \widetilde{a}_{1n}^2 \\ \vdots & \ddots & \vdots \\ \widetilde{a}_{n1}^2 & \dots & \widetilde{a}_{nn}^2 \end{vmatrix} = \sigma_1 b_{11} \widetilde{A}^2.$$

Dabei ist $\sigma_1 = -1$, wenn eine Zeilenvertauschung vorgenommen wurde, $\sigma_1 = +1$ sonst.

Durch wiederholte Anwendung folgt rekursiv

$$\det A = \det \widetilde{A}^1 = \sigma_1 \cdot b_{11} \det \widetilde{A}^2 = \sigma_1\sigma_2 \cdot b_{11}b_{22} \det \widetilde{A}^3 = \dots = \sigma_1\sigma_2 \cdots \sigma_{n-1} \cdot b_{11}b_{22} \cdots b_{nn}.$$

7.6 Nachiteration

Computer machen Rundungsfehler. Daher ist die numerische Lösung eines linearen Gleichungssystems meist nicht exakt und kann durch Nachiteration verbessert werden.

Definition 7.10 (Defekt, Fehler)
Sei $Ax = f$ ein Gleichungssystem und $\widetilde{x}$ eine Näherungslösung. Dann definiert man:

- *Defekt:* $d := A\widetilde{x} - f$
- *Fehler:* $r := \widetilde{x} - x$

Fehler und Defekt erfüllen das Gleichungssysstem $Ar = d$.

Erste Näherungslösung bestimmen:

Bestimme eine Näherungslösung $\widetilde{x}^0$ des Gleichungssystems $Ax = f$. Bestimme den Defekt $d^0 := A\widetilde{x}^0 - f$.

Erste Nachiteration:

Löse für den ersten Fehler das Gleichungssystems $Ar^0 = d^0$. Wegen

$$A(\widetilde{x}^0 - r^0) = A\widetilde{x}^0 - d^0 = f$$

setzt man als neue Näherung für x den Wert $\widetilde{x}^1 = \widetilde{x}^0 - r^0$ an. Bestimme dazu den Defekt $d^1 = A\widetilde{x}^1 - f$.

Vergrößert sich der Defekt, d.h. ist $\left\|d^1\right\| > \left\|d^0\right\|$, so ist keine Verbesserung durch Nachiteration mehr zu erzielen und man bricht die Rekursion ab. Ansonsten wird sie fortgesetzt.

Allgemeine Nachiteration:

Bestimmt wurde die Näherung $\widetilde{x}^t$ mit dem Defekt $d^t := A\widetilde{x}^t - f$. Löse für den Fehler r^t das Gleichungssystem $Ar^t = d^t$. Setze dann als neue Näherung für x den Wert $\widetilde{x}^{t+1} = \widetilde{x}^t - r^t$ an und bestimme dazu den Defekt $d^{t+1} = A\widetilde{x}^{t+1} - f$.

Falls $\left\|d^{t+1}\right\| > \|d^t\|$, so bricht man die Rekursion ab. Ansonsten fährt man rekursiv fort.

Definition 7.11 (Konditionszahl einer Matrix)
Für eine Matrix A ist die Konditionszahl $\kappa(a)$ bezüglich einer verträglichen Norm definiert duch

$$\kappa(A) = \|A\| \left\|A^{-1}\right\|.$$

Matrizen mit großem κ nennt man schlecht konditioniert.

Proposition 7.12 (Fehlerabschätzung für Näherungslösungen)
Für eine Näherungslösung $\widetilde{x}$ gelten für den Fehler $r := \widetilde{x} - x$ die Fehlerabschätzungen:

$$\frac{\|d\|}{\|A\|} \leq \|r\| \leq \left\|A^{-1}\right\| \|d\| \qquad \text{(Absoluter Fehler)}$$

$$\frac{1}{\kappa(A)} \frac{\|d\|}{\|f\|} \leq \frac{\|r\|}{\|x\|} \leq \kappa(A) \frac{\|d\|}{\|f\|} \qquad \text{(Relativer Fehler)}$$

Beweis. Es gelten für verträgliche Normen und $Ax = f$ die Abschätzungen:

$$\|f\| = \|Ax\| \leq \|A\| \|x\|$$
$$\|x\| = \left\|A^{-1}f\right\| \leq \left\|A^{-1}\right\| \|f\|$$

Somit gilt wegen $Ar = d$ die Abschätzung des absoluten Fehlers die Ungleichung

$$\frac{\|d\|}{\|A\|} \leq \|r\| \leq \left\|A^{-1}\right\| \|d\|.$$

Weiter gilt der relative Fehler wegen

$$\frac{1}{\kappa(A)} \frac{\|d\|}{\|f\|} = \frac{\|d\|}{\|A\|} \frac{1}{\|A^{-1}\| \|f\|} \leq \frac{\|r\|}{\|x\|} \leq \left\|A^{-1}\right\| \|d\| \frac{\|A\|}{\|f\|} = \kappa(A) \frac{\|d\|}{\|f\|}.$$

□

7.7 Dreieckszerlegung für symmetrische, positiv definite Matrizen

Dieses Kapitel ist teilweise redundant zu Kapitel 7.3 (Dreieckszerlegung). Da zwischenzeitlich die Dozenten gewechselt haben, weicht die Notation und Methodik dieses Kapitels von Kapitel 7.3 ab.

Proposition 7.13
Sei $A \in \mathbb{R}^{n \times n}$ eine symmetrische und positiv definite Matrix, also:

$$A = A^{\mathrm{T}} \quad \textit{und} \quad (Ax, x) = x^{\mathrm{T}} Ax > 0 \quad (\textit{für } x \in \mathbb{R}^n \setminus \{0\})$$

Dann gilt:

$$a_{ii} > 0 \qquad (\textit{für } i = 1, \ldots, n)$$
$$a_{ii}^{(i-1)} > 0 \quad (\textit{für } i = 2, \ldots, n)$$

Beweis. Der Beweis soll hier nicht aufgeführt werden. □

Wir werden für eine symmetrische und positiv definite Matrix $A = (a_{ij})$ eine Dreieckszerlegung durchführen $A = LU$ durchführen, wobei $L = (l_{ij})$ eine untere Dreiecksmatrix mit $l_{ii} = 1$ für alle i und U eine obere Dreiecksmatrix ist (vgl. Kapitel 7.3). Nach Proposition 7.13 ist dieses ohne Zeilenvertauschung möglich (wenn wir auf Pivotisierung verzichten), weil für alle Matrix-Elemente, durch die dividiert wird, gilt $a_{ii}^{(i-1)} \neq 0$. Zur Einführung der *Frobenius-Matrizen* $L^{(1)}, L^{(2)}, \ldots, L^{(n-1)}$ berechnen wir L hier mit deren Hilfe. Es gilt

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ -l_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -l_{n1} & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} \end{pmatrix},$$

wobei

$$L^{(1)} := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -l_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -l_{n1} & 0 & \ldots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -a_{21}/a_{11} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1}/a_{11} & 0 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

und

$$A^{(1)} := \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} \end{pmatrix} = \left(\begin{array}{c|ccc} a_{11} & a_{12} & \ldots & a_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A}^{(1)} & \\ 0 & & & \end{array}\right) \in \mathbb{R}^{n \times n}.$$

(Hinweis: Die Matrixelemente l_{ij} von $L^{(i)}$ werden hier mit dem umgekehrten Vorzeichen wie in der Vorlesung definiert, damit später die Definition von L mit der in den vorhergehenden Kapiteln übereinstimmt.) Im nächsten Schritt erhalten wir für die Untermatrix $\tilde{A}^{(1)} \in \mathbb{R}^{(n-1)\times(n-1)}$:

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ -l_{32} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -l_{n2} & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \ldots & a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & \ldots & a_{3n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n2}^{(1)} & a_{n3}^{(1)} & \ldots & a_{nn}^{(1)} \end{pmatrix} = \begin{pmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \ldots & a_{2n}^{(1)} \\ 0 & a_{33}^{(2)} & \ldots & a_{3n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n3}^{(2)} & \ldots & a_{nn}^{(2)} \end{pmatrix},$$

wobei

$$\tilde{L}^{(2)} := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -l_{32} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -l_{n2} & 0 & \ldots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -a_{32}^{(1)}/a_{22}^{(1)} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n2}^{(1)}/a_{22}^{(1)} & 0 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}$$

und

$$A^{(2)} := \begin{pmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix} = \left(\begin{array}{c|ccc} a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ \hline 0 & & & \\ \vdots & & \tilde{A}^{(2)} & \\ 0 & & & \end{array}\right) \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Dies führen wir durch, bis wir bei $\tilde{L}^{(n-1)}$ angelangen (die darauffolgende Matrix $\tilde{A}^{(n-1)}$ nur noch eine 1×1-Matrix ist). Wir setzen für $i = 2, \dots, n-1$

$$L^{(i)} := \left(\begin{array}{c|c} I_{i-1} & 0 \\ \hline 0 & \tilde{L}^{(i)} \end{array}\right) \in \mathbb{R}^{n\times n}$$

mit der Einheitsmatrix $I_{i-1} \in \mathbb{R}^{(i-1)\times(i-1)}$. Damit erhalten wir für A

$$\begin{aligned} & L^{(n-1)} \cdots L^{(2)} L^{(1)} A = U \\ \Longrightarrow\ & A = \left(L^{(n-1)} \cdots L^{(2)} L^{(1)}\right)^{-1} U = LU \end{aligned}$$

mit der oberen Dreiecksmatrix

$$U = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & a_{nn}^{(n-1)} \end{pmatrix}$$

und der Matrix

$$L := \left(L^{(n-1)} \cdots L^{(2)} L^{(1)}\right)^{-1} = \left(L^{(1)}\right)^{-1} \left(L^{(2)}\right)^{-1} \cdots \left(L^{(n-1)}\right)^{-1}.$$

Die Inversen der Frobenius-Matrizen erhält man durch Umkehrung des Vorzeichens der Nichtdiagonal-Elemente. Um dies zu zeigen, stellen wir $L^{(k)}$ und $\left(L^{(k)}\right)^{-1}$ dar durch

$$\begin{aligned} L^{(k)} &= I_n - \sum_{i=k+1}^{n} l_{ik} e_i e_k^{\mathrm{T}}, \\ \left(L^{(k)}\right)^{-1} &= I_n + \sum_{i=k+1}^{n} l_{ik} e_i e_k^{\mathrm{T}} \\ \Longrightarrow\ L^{(k)} \left(L^{(k)}\right)^{-1} &= I_n - \sum_{i=k+1}^{n} l_{ik} e_i e_k^{\mathrm{T}} + \sum_{i=k+1}^{n} l_{ik} e_i e_k^{\mathrm{T}} - \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} l_{ik} l_{jk} e_i \underbrace{e_k^{\mathrm{T}} e_j}_{=0} e_k^{\mathrm{T}} = I_n. \end{aligned}$$

Außerdem gilt für das Produkt L der inversen Frobenius-Matrizen:

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn-1} & 1 \end{pmatrix}$$

Denn für das folgende Teilprodukt gilt

$$\begin{aligned} \left(L^{(1)}\right)^{-1} \left(L^{(2)}\right)^{-1} &= \left(I_n + \sum_{i=2}^{n} l_{i1} e_i e_1^{\mathrm{T}}\right) \left(I_n + \sum_{i=3}^{n} l_{i2} e_i e_2^{\mathrm{T}}\right) \\ &= I_n + \sum_{i=2}^{n} l_{i1} e_i e_1^{\mathrm{T}} + \sum_{i=3}^{n} l_{i2} e_i e_2^{\mathrm{T}} + \sum_{i=2}^{n} \sum_{j=3}^{n} l_{i1} l_{j2} e_i \underbrace{e_1^{\mathrm{T}} e_j}_{=0} e_2^{\mathrm{T}}. \end{aligned}$$

Mittels Induktion erhält man den allgemeinen Fall. Somit haben wir mit L und U eine eindeutige Dreieckszerlegung von A gefunden (vgl. vorheriges Kapitel zur Eindeutigkeit), wobei L eine untere Dreiecksmatrix mit $l_{ii} = 1 \ \forall \ i$ und U eine obere Dreiecksmatrix mit Hauptdiagonal-Elementen > 0 ist (vgl. Proposition 7.13). Damit ist $\det(L) = 1$ und $\det(U) > 0$, womit beide Matrizen regulär sind (Anmerkung: U muss auch ohne Proposition 7.13 regulär sein, da sowohl A als auch L regulär sind).

Beispiel 7.14 (Dreieckszerlegung ohne Zeilenvertauschung)
Ein Beispiel zur Dreieckszerlegung ist im Folgenden angegeben:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 1 & 2 & 3 \\ & 2 & 5 & 9 \\ & 3 & 9 & 19 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 1 & 2 & 3 \\ & & 1 & 3 \\ & & 3 & 10 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 1 & 2 & 3 \\ & & 1 & 3 \\ & & & 1 \end{pmatrix} = U$$

Es ist

$$L = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 2 & 1 & \\ 1 & 3 & 3 & 1 \end{pmatrix}.$$

7.8 Cholesky-Zerlegung

Proposition 7.15 (Cholesky-Zerlegung I)
Sei $A \in \mathbb{R}^{n \times n}$ eine symmetrische und positiv definite Matrix. Dann existiert die Darstellung

$$A = LDL^{\mathrm{T}},$$

wobei $L = (l_{ij}) \in \mathbb{R}^{n \times n}$ eine untere Dreiecksmatrix mit $l_{ii} = 1 \ \forall \ i$ und $D = \mathrm{diag}(d_{11}, \dots, d_{nn}) \in \mathbb{R}^{n \times n}$ eine Diagonalmatrix mit $d_{ii} > 0 \ \forall \ i$ ist.

Beweis. Wir führen zuerst eine Dreiecks-Zerlegung von A gemäß dem vorherigen Abschnitt durch: $A = LU$. Es ist jetzt zu zeigen, dass eine Zerlegung $U = DL^{\mathrm{T}}$ existiert mit $D = \mathrm{diag}(d_{11}, \ldots, d_{nn})$ und $d_{ii} > 0 \; \forall \; i$. Dazu zerlegen wir $U = \tilde{D}L^{\mathrm{T}}$ mit einer nicht unbedingt diagonalen Matrix $\tilde{D}$. Diese Zerlegung existiert wegen der Regularität von L und somit L^{T}. Aus der Symmetrie von A folgt

$$\left.\begin{aligned} A &= LU = L\tilde{D}L^{\mathrm{T}} \\ A^{\mathrm{T}} &= U^{\mathrm{T}}L^{\mathrm{T}} = L\tilde{D}^{\mathrm{T}}L^{\mathrm{T}} \end{aligned}\right\} \quad \wedge \quad A = A^{\mathrm{T}} \implies \tilde{D} = \tilde{D}^{\mathrm{T}}.$$

Nun ist sowohl U als auch L^{T} eine obere Dreiecksmatrix:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \ldots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \ldots & a_{3n}^{(2)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & a_{nn}^{(n-1)} \end{pmatrix} = \begin{pmatrix} \tilde{d}_{11} & \tilde{d}_{12} & \tilde{d}_{13} & \ldots & \tilde{d}_{1n} \\ \tilde{d}_{21} & \tilde{d}_{22} & \tilde{d}_{23} & \ldots & \tilde{d}_{2n} \\ \tilde{d}_{31} & \tilde{d}_{32} & \tilde{d}_{33} & \ldots & \tilde{d}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{d}_{n1} & \tilde{d}_{n2} & \tilde{d}_{n3} & \ldots & \tilde{d}_{nn} \end{pmatrix} \begin{pmatrix} 1 & l_{21} & l_{31} & \ldots & l_{n1} \\ 0 & 1 & l_{32} & \ldots & l_{n2} \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & l_{nn-1} \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

Aus den Hauptdiagonal- und den Null-Elementen von U ergibt sich sofort:

$$\begin{aligned} a_{11} &= 1 \cdot \tilde{d}_{11} && \implies && \tilde{d}_{11} = a_{11} \\ 0 &= 1 \cdot \tilde{d}_{j1} && \implies && \tilde{d}_{j1} = \tilde{d}_{1j} = 0 \quad (\text{für } 2 \leq j \leq n) \\ a_{22}^{(1)} &= 1 \cdot \tilde{d}_{22} && \implies && \tilde{d}_{22} = a_{22}^{(1)} \\ 0 &= l_{21} \cdot \underbrace{\tilde{d}_{j1}}_{=0} + 1 \cdot \tilde{d}_{j2} && \implies && \tilde{d}_{j2} = \tilde{d}_{2j} = 0 \quad (\text{für } 3 \leq j \leq n) \\ a_{jj}^{(j-1)} &= 1 \cdot \tilde{d}_{jj} && \implies && \tilde{d}_{jj} = a_{jj}^{(j-1)} \quad (\text{für } 2 \leq j \leq n) \\ 0 &= 1 \cdot \tilde{d}_{jk} && \implies && \tilde{d}_{jk} = \tilde{d}_{kj} = 0 \quad (\text{für } 1 \leq j < k \leq n) \end{aligned}$$

Damit ist gezeigt, dass $\tilde{D}$ diagonal ist und wir setzen

$$D := \tilde{D} = \mathrm{diag}\left(a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}\right) =: \mathrm{diag}(d_{11}, \ldots, d_{nn})$$

mit $a_{11} > 0$ und $a_{jj}^{(j-1)} > 0$ für $2 \leq j \leq n$. □

Proposition 7.16 (Cholesky-Zerlegung II)
Sei $A \in \mathbb{R}^{n \times n}$ eine symmetrische und positiv definite Matrix. Dann existiert die Darstellung

$$A = CC^{\mathrm{T}},$$

wobei $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ eine untere Dreiecksmatrix mit $c_{ii} > 0 \; \forall \; i$ ist.

Beweis. Nach Proposition 7.15 existiert eine Zerlegung $A = LDL^{\mathrm{T}}$. Wir definieren die Dreiecksmatrix

$$C := LW \qquad \text{mit} \qquad W := \sqrt{D} = \mathrm{diag}(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}), \qquad W^2 = D.$$

Dann gilt

$$CC^{\mathrm{T}} = (LW)\,(LW)^{\mathrm{T}} = LWW^{\mathrm{T}}L^{\mathrm{T}} = LW^2L^{\mathrm{T}} = LDL^{\mathrm{T}} = A.$$

Da die Hauptdiagonal-Elemente von L Einsen sind und $\sqrt{d_{ii}} > 0 \; \forall \; i$, gilt auch für die Hauptdiagonal-Elemente von C wieder $c_{ii} > 0 \; \forall \; i$. □

Bemerkung 7.17 (Cholesky-Zerlegung II)
Natürlich kann die Cholesky-Zerlegung auch durch direktes Lösen des Gleichungssystems $CC^{\mathrm{T}} = A$ erfolgen. Dies ist allerdings recht aufwendig, da das Gleichungssystem nicht linear ist. Im Folgenden ist ein Beispiel für den Fall $n = 3$ angegeben.

Beispiel 7.18 (Direkte Cholesky-Zerlegung)
Es gilt das folgende Gleichungssystem zu lösen:

$$\begin{pmatrix} c_{11} & 0 & 0 \\ c_{21} & c_{22} & 0 \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \begin{pmatrix} c_{11} & c_{21} & c_{31} \\ 0 & c_{22} & c_{32} \\ 0 & 0 & c_{33} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Man erhält den Satz von Gleichungen (die mit der jeweiligen Gleichung zu bestimmende Unbekannte ist unterstrichen):

$$\begin{aligned} \underline{c_{11}^2} &= a_{11} & & & & \\ \underline{c_{21}} c_{11} &= a_{21} & c_{21}^2 + \underline{c_{22}^2} &= a_{22} & & \\ \underline{c_{31}} c_{11} &= a_{31} & c_{31} c_{21} + \underline{c_{32}} c_{22} &= a_{32} & c_{31}^2 + c_{32}^2 + \underline{c_{33}^2} &= a_{33}. \end{aligned}$$

8 Orthogonalisierungsverfahren

Ziel der Orthogonalisierungsverfahren ist es, eine Matrix $A \in \mathbb{R}^{n\times n}$ in $A = QR$ zu zerlegen. Dabei ist Q eine orthogonale Matrix ($Q^{\mathrm{T}} = Q^{-1}$) und R eine obere Dreiecksmatrix. Diese Art zur Transformation von A in Dreiecksform ist numerisch stabiler als die Methode nach Gauß mit Frobeniusmatrizen.

Bemerkung 8.1 (Schreibweise)
Gegeben sei eine Matrix $A = (a_{ij}) \in \mathbb{K}^{n\times m}$. Dann bezeichnen wir mit a_j die j-te Spalte von A. Damit lässt sich A schreiben als

$$A = (a_1 a_2 \dots a_m).$$

8.1 Gram-Schmidt-Verfahren

Proposition 8.2 (Orthogonalisierung nach Gram-Schmidt)
Gegeben sei die reguläre Matrix $A \in \mathbb{R}^{n\times n}$. Mit Hilfe des Gram-Schmidtschen Orthogonalisierungs-Verfahrens kann man eine orthogonale Form Q von A mit

$$\langle q_1, \dots, q_j\rangle = \langle a_1, \dots, a_j\rangle \quad \textit{und} \quad (q_i, q_j) = \delta_{ij} \quad (\textit{für } i, j = 1, \dots, n)$$

zu

$$q'_k = a_k - \sum_{i=1}^{k-1} (a_k, q_i) q_i, \qquad q_k := \frac{1}{\sqrt{(q'_k, q'_k)}} q'_k \quad (\textit{für } k = 1, \dots, n)$$

bestimmen.

Proposition 8.3 (Bestimmung der Dreiecksmatrix)
Für die Elemente der Dreiecksmatrix R gilt dann

$$r_{ik} = (q_i, a_k) \quad \textit{und} \quad r_{ik} = \sqrt{(a_k, a_k) - \sum_{j=1, j\neq i}^{k} r_{jk}^2} \quad (\textit{für } i, k = 1, \dots, n)\,.$$

Beweis. Wir schreiben $QR = A$ in der Form

$$(q_1 q_2 \dots q_n) \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots \\ 0 & r_{22} & r_{23} & \dots \\ 0 & 0 & r_{33} & \dots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix} = (a_1 a_2 \dots a_n)$$

und lesen für $1 \le k \le n$ unmittelbar ab:

$$a_k = \sum_{j=1}^{k} r_{jk} q_j \tag{8}$$

Multiplizieren wir Gleichung (8) mit q_i (Skalarprodukt), so erhalten wir

$$(q_i, a_k) = \left(q_i, \sum_{j=1}^{k} r_{jk} q_j\right) = \sum_{j=1}^{k} r_{jk} \underbrace{(q_i, q_j)}_{=\delta_{ij}} = r_{ik}.$$

Mutliplizieren wir Gleichung (8) mit sich selbst, so erhalten wir

$$(a_k, a_k) = \left(\sum_{j=1}^{k} r_{jk} q_j, \sum_{j=1}^{k} r_{jk} q_j \right) = \sum_{j=1}^{k} \sum_{l=1}^{k} r_{jk} r_{lk} \underbrace{(q_j, q_l)}_{=\delta_{jl}} = \sum_{j=1}^{k} r_{jk}^2 = \sum_{j=1, j\neq i}^{k} r_{jk}^2 + r_{ik}^2$$

$$\iff r_{ik} = \sqrt{(a_k, a_k) - \sum_{j=1, j\neq i}^{k} r_{jk}^2}.$$

Anmerkung: Die Aussage der Proposition ist äquivalent zu $R = Q^{-1}A = Q^{\mathrm{T}}A$. □

8.2 Givens-Rotation

Bei der Orthogonalisierung nach Givens nutzt man eine Rotationsmatrix, um ein Element der Matrix A durch Drehung zu Null werden zu lassen. Durch Wiederholung kann die Ausgangsmatrix auf Dreiecksform transformiert werden.

Definition 8.4 (Jakobi-Rotationen)
Eine Matrix $\Omega_{j,k} \in \mathbb{R}^{n\times n}$ mit der Gestalt

$$\Omega_{j,k} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & & s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -s & & & c & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow j\text{-te Zeile} \\ \\ \\ \\ \leftarrow k\text{-te Zeile} \\ \\ \\ \\ \end{matrix}$$

und $c^2 + s^2 = 1$ heißt Jakobi-Rotation. *Jakobi-Rotationen sind orthogonal.*

Proposition 8.5 (Eliminination eines Matrix-Elements)
Die Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ mit $a_{kl} \neq 0$ und $k > l$ kann durch Multiplikation mit der Jakobi-Rotation

$$\Omega_{k-1,k}^{(l)} \in \mathbb{R}^{n\times n}: \qquad s = \frac{a_{kl}}{\sqrt{a_{(k-1)l}^2 + a_{kl}^2}}, \qquad c = \frac{a_{(k-1)l}}{\sqrt{a_{(k-1)l}^2 + a_{kl}^2}}$$

zu

$$A^{(1)} = \left(a_{ij}^{(1)}\right) := \Omega_{k-1,k}^{(l)} A \qquad \text{mit} \qquad a_{kl}^{(1)} = 0$$

transformiert werden. Der hochgestellte Index bei $\Omega_{k-1,k}^{(l)}$ soll andeuten, dass das Matrixelement in der l-ten Spalte eliminiert wurde.

Beweis. Wir schreiben die Matrizen aus:

$$\begin{pmatrix} \ddots & & & \\ & c & s & \\ & -s & c & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} * & * & * & * \\ * & a_{(k-1)l} & * & * \\ * & a_{kl} & * & * \\ * & * & * & * \end{pmatrix} = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & a_{kl}^{(1)} & * & * \\ * & * & * & * \end{pmatrix}$$

Es soll nun gelten:

$$\begin{aligned} a_{kl}^{(1)} &= -s \cdot a_{(k-1)l} + c \cdot a_{kl} \stackrel{!}{=} 0 \\ \Longrightarrow s &= \kappa \cdot a_{kl}, \qquad c = \kappa \cdot a_{(k-1)l} \end{aligned}$$

Die Konstante κ wird aus der Nebenbedingung $c^2 + s^2 = 1$ zu

$$\kappa = \frac{1}{\sqrt{a_{(k-1)l}^2 + a_{kl}^2}}$$

bestimmt. Damit erhalten c und s die geforderte Form. □

Proposition 8.6 (Givens-Rotation)
Durch Multiplikation der Matrix A mit den Jakobi-Rotationen

$$\begin{matrix} \Omega_{n-1,n}^{(1)}, & \dots, & \Omega_{2,3}^{(1)}, & \Omega_{1,2}^{(1)}, \\ \Omega_{n-1,n}^{(2)}, & \dots, & \Omega_{2,3}^{(2)}, & \\ \vdots & & & \\ \Omega_{n-1,n}^{(n-1)} & & & \end{matrix}$$

in dieser Reihenfolge können die Matrix-Elemente unterhalb der Hauptdiagonalen sukzessive eliminiert und A auf die Dreiecksform R transformiert werden. (Eine Elimininationist natürlich nur dann durchzuführen, wenn das jeweilige Matrix-Element nicht ohnehin schon Null ist.) Damit ergibt sich

$$A = QR \qquad \textit{mit} \qquad Q := \left(\Omega_{n-1,n}^{(1)}\right)^{\mathrm{T}} \cdots \left(\Omega_{n-1,n}^{(n-1)}\right)^{\mathrm{T}}.$$

Beweis. Durch Multiplikation der Matrix A mit $\Omega_{n-1,n}^{(1)}$ im ersten Schritt erhalten wir nach Proposition 8.5 eine Matrix $A^{(1)}$ mit $a_{n1}^{(1)} = 0$. Multiplizieren wir $A^{(1)}$ nun mit $\Omega_{n-2,n-1}^{(1)}$, so erhalten wir eine Matrix $A^{(2)}$ mit $a_{n1}^{(2)} = a_{(n-1)1}^{(2)} = 0$ usw. Die Reihenfolge der Multiplikation stellt dabei sicher, dass ein eliminiertes Matrix-Element nicht in einem späteren Schritt wieder ungleich Null wird. (Das soll hier nicht ausführlich gezeigt werden.) Dies führt schließlich zu

$$\Omega_{n-1,n}^{(n-1)} \cdots \Omega_{n-1,n}^{(1)} A = R.$$

Wegen der Orthogonalität der Jakobi-Rotationen gilt

$$\begin{aligned} A &= \left(\Omega_{n-1,n}^{(n-1)} \cdots \Omega_{n-1,n}^{(1)}\right)^{-1} R \\ &= \left(\Omega_{n-1,n}^{(1)}\right)^{-1} \cdots \left(\Omega_{n-1,n}^{(n-1)}\right)^{-1} R \\ &= \underbrace{\left(\Omega_{n-1,n}^{(1)}\right)^{\mathrm{T}} \cdots \left(\Omega_{n-1,n}^{(n-1)}\right)^{\mathrm{T}}}_{=:Q} R. \end{aligned}$$

□

Bemerkung 8.7 (Konditionszahlen der Matrizen)
Für die Konditionszahl $\kappa(A) = \|A\| \left\|A^{-1}\right\|$ gilt

$$\|A\| = \max_i |\lambda_i| \quad \wedge \quad \left\|A^{-1}\right\| = \left(\min_i |\lambda_i|\right)^{-1}$$
$$\implies \kappa(A) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|},$$

wobei λ_i die Eigenwerte von A sind. Es gilt

$$Q^{-1} = Q^{\mathrm{T}} \implies \left\|Q^{\mathrm{T}} A x\right\|_2 = \|Ax\|_2 .$$

Damit gilt

$$\kappa(Q^{\mathrm{T}} A) = \kappa(A) \quad \wedge \quad A = QR \implies \kappa(A) = \kappa(R).$$

8.3 Householder-Spiegelung

In den folgenden Definitionen werden wir uns auf die Matrix-Darstellung sowie meistens auf den reellen Fall beschränken, wie es zu unseren Zwecken am günstigsten ist.

Definition 8.8 (Projektor)
Eine lineare Abbildung P auf einem vollständigen unitären Raum ist genau dann ein Projektor, wenn gilt:

$$(i) \qquad P^2 = P$$
$$(ii) \qquad P^* = P$$

Offenbar gilt dann auch $(Pu, v) = (u, Pv)$.

Proposition 8.9 (Komplementärer Projektor)
Der zu P komplementäre Projektor ist $(I - P)$.

Beweis. Zu zeigen: (i) und (ii) aus obiger Definition und $(I - P)P = 0$.
zu (i):

$$(I - P)^2 = I^2 - IP - PI + \underbrace{P^2}_{=P} = I - P$$

zu (ii):

$$(I - P)^* = I^* - P^* = I - P$$

Somit ist $(I - P)$ ein Projektor. Zu dem letzten Punkt:

$$(I - P)P = IP - P^2 = P - P = 0$$

□

Definition 8.10 (Spiegelung)
Eine Spiegelung S ist definiert durch die Differenz eines Projektors P und seines komplementären Projektors $(I - P)$:

$$S := (I - P) - P = I - 2P$$

Proposition 8.11 (Eigenschaften von Spiegelungen)
Spiegelungen haben die Eigenschaften:

$$\begin{aligned} &(i) && S^2 = I \\ &(ii) && S = S^* \end{aligned}$$

Beweis. zu (i):

$$S^2 = (I - 2P)(I - 2P) = I - 4P + 4P^2 = I - 4P + 4P = I$$

zu (ii):

$$S^* = (I - 2P)^* = I^* - 2P^* = I - 2P = S$$

□

Da in den betrachteten Fällen die Matrizen stets reell sind, wird von nun an statt A^* nur noch A^{T} geschrieben.

Proposition 8.12 (Householder-Matrix)
Eine Spiegel-Matrix $H_u \in \mathbb{R}^{n\times n}$ zum Orthogonal-Projektor P

$$Px = \frac{(x,u)}{(u,u)}u = \frac{uu^{\mathrm{T}}}{u^{\mathrm{T}}u}x \quad (\textit{für } x \in \mathbb{R}^n)$$

hat die Gestalt

$$H_u = I - 2\frac{uu^{\mathrm{T}}}{u^{\mathrm{T}}u}.$$

Sie wird Householder-Matrix *genannt. Es gilt:*

$$\begin{aligned} &(i) && H_u u = -u \\ &(ii) && H_u v = v, \qquad v \in \langle u\rangle^{\perp} \end{aligned}$$

Bemerkung 8.13 (Geometrische Betrachtung)
Die Householder-Matrix H_u bewirkt eine Spiegelung an der Ebene $\langle u\rangle^{\perp}$.

Proposition 8.14 (Householder-Spiegelung)
Die Matrix-Elemente einer Matrix $A \in \mathbb{R}^{n\times n}$ unterhalb der Hauptdiagonalen in der ersten Spalte können durch Multiplikation mit der Householder-Matrix $H^{(1)} = \tilde{H}^{(1)} := H_u \in \mathbb{R}^{n\times n}$ mit $u := a_1 - \|a_1\|_2 e_1$ annulliert werden:

$$H^{(1)}A = A^{(1)} = \left(\begin{array}{c|c} \|a_1\|_2 & * \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right) \qquad \textit{mit} \qquad a_1^{(1)} = \|a_1\|_2 e_1$$

Durch Anwendung dieses Verfahrens auf den Block $\tilde{A}^{(1)} \in \mathbb{R}^{(n-1)\times(n-1)}$ usw. kann A auf Dreiecksform R transformiert werden

$$A = QR, \qquad Q = H^{(1)}H^{(2)}\cdots H^{(n-1)},$$

wobei $H^{(k)}$ gegeben ist durch

$$H^{(k)} := \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & \tilde{H}^{(k)} \end{array}\right) \in R^{n\times n}$$

mit der im k-ten Schritt erzeugten Householder-Matrix $\tilde{H}^{(k)} \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$.

Beweis. Für zwei beliebige Vektoren $x, y \in \mathbb{R}^n$ mit $\|x\|_2 = \|y\|_2 \neq 0$ können wir die Householder-Matrix $H_u \in \mathbb{R}^{n\times n}$ mit $u = x - y$ bilden. Es gilt

$$\begin{aligned} H_u x &= x - 2\frac{(x,u)}{(u,u)}u \\ &= x - 2\frac{(x,x-y)}{\|x-y\|_2^2}(x-y) \\ &= \frac{1}{\|x-y\|_2^2}\left(\|x-y\|_2^2\, x - 2(x,x-y)(x-y)\right). \end{aligned}$$

Mit den Beziehungen

$$\|x-y\|_2^2 = (x,x) - 2(x,y) + (y,y)$$

und

$$(x,x-y)(x-y) = (x,x)x - (x,y)x - (x,x)y + (x,y)y$$

können wir dies schreiben als

$$\begin{aligned} H_u x &= \frac{1}{\|x-y\|_2^2}\left\{2(x,x)x - 2(x,y)x - 2(x,x)x - 2(x,y)y + 2(x,y)x + 2(x,x)y\right\} \\ &= \frac{1}{\|x-y\|_2^2}\underbrace{\{2(x,x) - 2(x,y)\}}_{\|x-y\|_2^2} y \\ &= y. \end{aligned}$$

Setzen wir nun $x = a_1$ und $y = \|a_1\|_2 e_1$ sowie $H^{(1)} := H_u$, so erhalten wir für die erste Spalte von $A^{(1)}$

$$\begin{aligned} H^{(1)}A = H^{(1)}\left(a_1 \dots a_n\right) &= \left(a_1^{(1)} \dots a_n^{(1)}\right) \\ \Longrightarrow a_1^{(1)} = H^{(1)}a_1 &= \|a_1\|_2\, e_1, \end{aligned}$$

womit alle Elemente in der ersten Spalte von $A^{(1)}$ unterhalb der Hauptdiagonalen annuliert sind. Völlig analog können wir dies für den Block $\tilde{A}^{(1)}$ durchführen usw. Damit haben wir A auf Dreiecksform gebracht:

$$H^{(n-1)} \cdots H^{(1)} A = R$$

Da die Housholder-Matrizen als spezielle Spiegel-Matrizen orthogonal sind und die Eigenschaft (ii) aus Proposition 8.11 besitzen, gilt dies auch für die $H^{(k)}$ und wir erhalten schließlich

$$\begin{aligned} A &= \left(H^{(n-1)} \cdots H^{(1)}\right)^{-1} R \\ &= \left(H^{(1)}\right)^{-1} \cdots \left(H^{(n-1)}\right)^{-1} R \\ &= \left(H^{(1)}\right)^{\mathrm{T}} \cdots \left(H^{(n-1)}\right)^{\mathrm{T}} R \\ &= H^{(1)}H^{(2)} \cdots H^{(n-1)} R. \end{aligned}$$

□

8.4 Modifiziertes Gram-Schmidt-Verfahren

Gegeben sei ein überbestimmtes Gleichungssystem

$$Ax = b \qquad \text{bzw.} \qquad \sum_{i=1}^{n} a_i x_i = b$$

mit $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $m \geq n$ und $\dim\langle a_1, \ldots, a_n\rangle = n$. Unsere Aufgabe soll es sein, die „beste Lösung“ $\xi \in \mathbb{R}^n$ dieses Gleichungssystems zu finden:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\| = \|A\xi - b\| .$$

Dies ist geometrisch (nach dem Satz von Pythagoras) gleichbedeutend mit $(A\xi - b)\perp Ax \;\forall\; x \in \mathbb{R}^n$. Somit erhalten wir

$$\begin{aligned} &\forall x \in \mathbb{R}^n : (A\xi - b, Ax) = (A^{\mathrm{T}}A\xi - A^{\mathrm{T}}b, x) = 0 \\ \iff &A^{\mathrm{T}}A\xi - A^{\mathrm{T}}b = 0 \\ \iff &A^{\mathrm{T}}A\xi = A^{\mathrm{T}}b. \end{aligned}$$

Der letzte Ausdruck sind die *Normalgleichungen*

$$(a_k, b) = \sum_{j=1}^{n} (a_k, a_j)\xi_j \quad (\text{für } k = 1, \ldots, n) .$$

Proposition 8.15
Sei $a_k^{(1)} := a_k$ und $b^{(1)} := b$. Wir erhalten für $k, l = 1, \ldots, n$ mit $k \geq l$ die Vektoren

$$b^{(l)} := b^{(l-1)} - a_{l-1}^{(l-1)} \frac{\left(a_{l-1}^{(l-1)}, b^{(l-1)}\right)}{\left(a_{l-1}^{(l-1)}, a_{l-1}^{(l-1)}\right)}, \qquad a_k^{(l)} := a_k^{(l-1)} - a_{l-1}^{(l-1)} \frac{\left(a_{l-1}^{(l-1)}, a_k^{(l-1)}\right)}{\left(a_{l-1}^{(l-1)}, a_{l-1}^{(l-1)}\right)}$$

und die $(n - l + 1)$ neuen Normalgleichungen

$$\left(a_k^{(l)}, b^{(l)}\right) = \sum_{j=l}^{n} \left(a_k^{(l)}, a_j^{(l)}\right) \xi_j \quad (\textit{für } k = l, \ldots, n) .$$

Beweis. Mit der Austauschmethode führen wir den Austausch $\xi_1 \leftrightarrow (a_1, b)$ durch (hier für den Fall $n = 3$ angegeben):

	ξ_1	ξ_2	ξ_3
(a_1, b)	(a_1, a_1)	(a_1, a_2)	(a_1, a_3)
(a_2, b)	(a_2, a_1)	(a_2, a_2)	(a_2, a_3)
(a_3, b)	(a_3, a_1)	(a_3, a_2)	(a_3, a_3)

$$\downarrow$$

	(a_1, b)	ξ_2	ξ_3
ξ_1	$\frac{1}{(a_1,a_1)}$	$-\frac{(a_1,a_2)}{(a_1,a_1)}$	$-\frac{(a_1,a_3)}{(a_1,a_1)}$
(a_2, b)	$\frac{(a_2,a_1)}{(a_1,a_1)}$	$(a_2, a_2) - \frac{(a_1,a_2)(a_2,a_1)}{(a_1,a_1)}$	$(a_2, a_3) - \frac{(a_1,a_3)(a_2,a_1)}{(a_1,a_1)}$
(a_3, b)	$\frac{(a_3,a_1)}{(a_1,a_1)}$	$(a_3, a_2) - \frac{(a_1,a_2)(a_3,a_1)}{(a_1,a_1)}$	$(a_3, a_3) - \frac{(a_1,a_3)(a_3,a_1)}{(a_1,a_1)}$

Wir lesen an der k-ten Zeile ab, dass allgemein für $k > 1$ gilt

$$\begin{aligned}
(a_k, b) &= \frac{(a_k, a_1)}{(a_1, a_1)}(a_1, b) + \sum_{j=2}^{n} \left((a_k, a_j) - \frac{(a_1, a_j)(a_k, a_1)}{(a_1, a_1)} \right) \xi_j \\
\iff (a_k, b) - \frac{(a_k, a_1)}{(a_1, a_1)}(a_1, b) &= \sum_{j=2}^{n} \left((a_k, a_j) - \frac{(a_1, a_j)(a_k, a_1)}{(a_1, a_1)} \right) \xi_j .
\end{aligned}$$

Den Term auf der linken Seite der Gleichung formen wir um zu

$$\begin{aligned}
(a_k, b) - \frac{(a_k, a_1)}{(a_1, a_1)}(a_1, b) &= (a_k, b) - (a_k, a_1)\frac{(a_1, b)}{(a_1, a_1)} \\
&= \left(a_k, b - a_1 \frac{(a_1, b)}{(a_1, a_1)} \right) \\
&= \left(a_k, b^{(2)} \right), \qquad b^{(2)} := b - a_1 \frac{(a_1, b)}{(a_1, a_1)}.
\end{aligned}$$

Aus dem Ausdruck in der Klammer auf der rechten Seite der Gleichung wird völlig analog

$$(a_k, a_j) - \frac{(a_1, a_j)(a_k, a_1)}{(a_1, a_1)} = \left(a_k, a_j^{(2)} \right), \qquad a_j^{(2)} := a_j - a_1 \frac{(a_1, a_j)}{(a_1, a_1)}.$$

Nun gilt

$$\begin{aligned}
\left(a_1, b^{(2)} \right) &= \left(a_1, b - a_1 \frac{(a_1, b)}{(a_1, a_1)} \right) \\
&= (a_1, b) - \left(a_1, a_1 \frac{(a_1, b)}{(a_1, a_1)} \right) \\
&= (a_1, b) - \frac{(a_1, b)}{(a_1, a_1)} (a_1, a_1) \\
&= 0, \\
\left(a_1, a_k^{(2)} \right) &= 0
\end{aligned}$$

und daher ist

$$\begin{aligned}\left(a_k, b^{(2)}\right) &= \left(a_k, b^{(2)}\right) - \underbrace{\left(a_1 \frac{(a_1, a_k)}{(a_1, a_1)}, b^{(2)}\right)}_{=0} \\ &= \left(a_k - a_1 \frac{(a_1, a_k)}{(a_1, a_1)}, b^{(2)}\right) \\ &= \left(a_k^{(2)}, b^{(2)}\right), \\ \left(a_k, a_j^{(2)}\right) &= \cdots = \left(a_k^{(2)}, a_j^{(2)}\right).\end{aligned}$$

Wir können somit ein Schema der Austauschmethode für die verbleibenden $\xi_2, \ldots, \xi_n$ wie folgt schreiben:

	ξ_2	ξ_3
$\left(a_2^{(2)}, b^{(2)}\right)$	$\left(a_2^{(2)}, a_2^{(2)}\right)$	$\left(a_2^{(2)}, a_3^{(2)}\right)$
$\left(a_3^{(2)}, b^{(2)}\right)$	$\left(a_3^{(2)}, a_2^{(2)}\right)$	$\left(a_3^{(2)}, a_3^{(2)}\right)$

Durch vollständige Induktion ergibt sich die allgemeine Form der Proposition. □

Proposition 8.16 (Modifiziertes Gram-Schmidtsches-Verfahren)
Für die Vektoren $a_i^{(i)}$ gilt:

$$\left\langle a_1^{(1)}, a_2^{(2)}, \ldots, a_j^{(j)} \right\rangle = \langle a_1, a_2, \ldots, a_j \rangle \quad (\textit{für } j = 1, \ldots, n)$$

und

$$\left(a_i^{(i)}, a_j^{(j)}\right) = 0 \quad (\textit{für } i, j = 1, \ldots, n \quad i \neq j)$$

Somit kann mit $\left(a_1^{(1)} a_2^{(2)} \ldots a_n^{(n)}\right)$ eine orthogonale Form der Matrix A angegeben werden, falls $A \in \mathbb{R}^{n \times n}$.

Beweis. Zum zweiten Teil (Orthogonalitäten): Allgemein gilt (vgl. Beweis zur vorherigen Proposition)

$$\begin{aligned}\left(a_{j-1}^{(j-1)}, a_k^{(j)}\right) &= \left(a_{j-1}^{(j-1)}, a_k^{(j-1)} - a_{j-1}^{(j-1)} \frac{\left(a_{j-1}^{(j-1)}, a_k^{(j-1)}\right)}{\left(a_{j-1}^{(j-1)}, a_{j-1}^{(j-1)}\right)}\right) \\ &= \left(a_{j-1}^{(j-1)}, a_k^{(j-1)}\right) - \left(a_{j-1}^{(j-1)}, a_{j-1}^{(j-1)} \frac{\left(a_{j-1}^{(j-1)}, a_k^{(j-1)}\right)}{\left(a_{j-1}^{(j-1)}, a_{j-1}^{(j-1)}\right)}\right) \\ &= \left(a_{j-1}^{(j-1)}, a_k^{(j-1)}\right) - \left(a_{j-1}^{(j-1)}, a_{j-1}^{(j-1)}\right) \frac{\left(a_{j-1}^{(j-1)}, a_k^{(j-1)}\right)}{\left(a_{j-1}^{(j-1)}, a_{j-1}^{(j-1)}\right)} \\ &= 0.\end{aligned} \tag{9}$$

Daraus folgt (zur besseren Lesbarkeit wird der skalare Bruch durch Pünktchen abgekürzt)

$$\begin{aligned}\left(a_{j-2}^{(j-2)}, a_k^{(j)}\right) &= \left(a_{j-2}^{(j-2)}, a_k^{(j-1)} - a_{j-1}^{(j-1)} \cdots\right) \\ &= \left(a_{j-2}^{(j-2)}, a_k^{(j-1)}\right) - \left(a_{j-2}^{(j-2)}, a_{j-1}^{(j-1)} \cdots\right) \\ &= \underbrace{\left(a_{j-2}^{(j-2)}, a_k^{(j-1)}\right)}_{=0, \text{ vgl. Gl. (9)}} - \underbrace{\left(a_{j-2}^{(j-2)}, a_{j-1}^{(j-1)}\right)}_{=0, \text{ vgl. Gl. (9)}} \cdots \\ &= 0\end{aligned} \tag{10}$$

und

$$\begin{aligned}\left(a_{j-3}^{(j-3)}, a_k^{(j)}\right) &= \left(a_{j-3}^{(j-3)}, a_k^{(j-1)} - a_{j-1}^{(j-1)} \cdots\right) \\ &= \left(a_{j-3}^{(j-3)}, a_k^{(j-1)}\right) - \left(a_{j-3}^{(j-3)}, a_{j-1}^{(j-1)} \cdots\right) \\ &= \underbrace{\left(a_{j-3}^{(j-3)}, a_k^{(j-1)}\right)}_{=0, \text{ vgl. Gl. (10)}} - \underbrace{\left(a_{j-3}^{(j-3)}, a_{j-1}^{(j-1)}\right)}_{=0, \text{ vgl. Gl. (10)}} \cdots \\ &= 0.\end{aligned}$$

Wir können dies analog für alle weiteren Skalarprodukte $\left(a_i^{(i)}, a_k^{(j)}\right)$ mit $i = j-4, \ldots, 1$ unter Benutzung der jeweils zuvor gezeigten Orthogonalität weiterführen und erhalten so allgemein

$$\left(a_i^{(i)}, a_k^{(j)}\right) = 0 \quad (\text{für } i, j = 1, \ldots, n \quad i < j)$$

Die Aussage in der Proposition ergibt sich für den Spezialfall $k = j$ aus obiger Gleichung.

Zum ersten Teil (Spans): Nach Definition ist $a_k^{(j)}$ eine Linearkombination aus $a_k^{(j-1)}$ und $a_{j-1}^{(j-1)}$:

$$a_k^{(j)} \in \left\langle a_{j-1}^{(j-1)}, a_k^{(j-1)} \right\rangle .$$

Speziell für $a_j^{(j)}$ und $a_j^{(j-1)}$ gilt damit

$$\begin{aligned} a_j^{(j)} \in \left\langle a_{j-1}^{(j-1)}, a_j^{(j-1)} \right\rangle \qquad \wedge \qquad a_j^{(j-1)} \in \left\langle a_{j-2}^{(j-2)}, a_j^{(j-2)} \right\rangle \\ \Longrightarrow a_j^{(j)} \in \left\langle a_{j-1}^{(j-1)}, a_j^{(j-1)} \right\rangle \subseteq \left\langle a_{j-1}^{(j-1)}, a_{j-2}^{(j-2)}, a_j^{(j-2)} \right\rangle . \end{aligned}$$

Daher gilt, wenn wir die Ersetzung im Span für $a_j^{(j-2)}$ usw. weiterführen, allgemein

$$a_j^{(j)} \in \left\langle a_{j-1}^{(j-1)}, a_{j-2}^{(j-2)}, \ldots, a_1^{(1)}, a_j^{(1)} \right\rangle \quad (\text{für } j = 1, \ldots, n) .$$

Tauscht man nun $a_{j-1}^{(j-1)}$ im Span von $a_j^{(j)}$ wie folgt aus

$$\begin{aligned} a_{j-1}^{(j-1)} \in \left\langle a_{j-2}^{(j-2)}, \ldots, a_1^{(1)}, a_{j-1}^{(1)} \right\rangle \\ \Longrightarrow a_j^{(j)} \in \left\langle a_{j-2}^{(j-2)}, \ldots, a_1^{(1)}, a_j^{(1)}, a_{j-1}^{(1)} \right\rangle \end{aligned}$$

und führt dies sukzessive für die weiteren $a_i^{(i)}$ mit $i = j-2, \ldots, 2$ durch, so erhält man schließlich

$$a_j^{(j)} \in \left\langle a_j^{(1)}, a_{j-1}^{(1)}, \ldots, a_2^{(1)}, a_1^{(1)} \right\rangle \equiv \langle a_j, a_{j-1}, \ldots, a_2, a_1 \rangle \quad (\text{für } j = 1, \ldots, n)\,.$$

Da auch alle $a_i^{(i)}$ mit $i < j$ in dem obigen Span enthalten sind, können wir auch schreiben

$$\left\langle a_j^{(j)}, a_{j-1}^{(j-1)}, \ldots, a_1^{(1)} \right\rangle \subseteq \langle a_j, a_{j-1}, \ldots, a_2, a_1 \rangle \quad (\text{für } j = 1, \ldots, n)\,.$$

Die $a_i^{(i)}$ sind, wie oben gezeigt wurde, orthogonal und damit linear unabhängig (ebenso wie die a_i), woraus schließlich die Gleichheit der Spans folgt. □