Kapitel 1

Approximation von Funktionen

Die polynomiale Approximation von Funktionen $f(x)$ einer Veränderlichen $x \in \mathbb{R}$ dient einerseits der Verarbeitung von experimentellen Daten, d.h. für eine gegebene Punktmenge $\{(x_i, f_i)\}_{i=0}^n$ ist eine geeignete funktionale Darstellung $f_n(x)$ zu finden, andererseits ermöglicht die Approximation einer gegebenen Funktion $f(x)$ durch ein Polynom eine einfache Realisierung von Differentiation und Integration. Später werden diese Konzepte zur Approximation von Funktionen auch zur näherungsweisen Lösung von Anfangs– und Randwertproblemen gewöhnlicher und partieller Differentialgleichungen bzw. von Integralgleichungen eingesetzt.

1.1 Interpolation

Gegeben seien $n+1$ Paare $\{(x_i, f_i)\}_{i=0}^n$ von paarweise verschiedenen Stützstellen $x_i \neq x_j$ für $i \neq j$ und zugehörigen Funktionswerten f_i; für eine gegebene Funktion $f(x)$ sei $f_i = f(x_i)$. Gesucht ist das Interpolationspolynom

$$f_n(x) = \sum_{k=0}^{n} a_k x^k \tag{1.1}$$

mit noch zu bestimmenden Zerlegungskoeffizienten $a_0, \ldots, a_n$, welches in den Stützstellen x_i die Interpolationsgleichungen $f_n(x_i) = f_i$ erfüllt, d.h.

$$f_n(x_i) = \sum_{k=0}^{n} a_k x_i^k = f_i \quad \text{für } i = 0, \ldots, n. \tag{1.2}$$

Diese $n+1$ Gleichungen entsprechen dem linearen Gleichungssystem

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix} \tag{1.3}$$

bzw.

$$A_n \underline{a} = \underline{f} \tag{1.4}$$

mit der durch die Einträge

$$A_n[i,k] = x_i^k \quad \text{für } i,k = 0,\ldots,n \tag{1.5}$$

definierten Systemmatrix

$$A_n \in \mathbb{R}^{(n+1)\times(n+1)}.$$

Die Zerlegungskoeffizienten a_k des Interpolationspolynoms $f_n(x)$ sind genau dann eindeutig bestimmt, wenn das lineare Gleichungssystem (1.4) eindeutig lösbar ist. Zu untersuchen ist deshalb die Invertierbarkeit der Systemmatrizen A_n.

Lemma 1.1. *Die durch*

$$A_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix} \tag{1.6}$$

gegebenen Systemmatrizen $A_n \in \mathbb{R}^{(n+1)\times(n+1)}$ sind Vandermonde–Matrizen. Für diese gilt

$$\det A_n = \prod_{0\le i<j\le n} (x_j - x_i). \tag{1.7}$$

Beweis: Der Beweis von (1.7) erfolgt durch vollständige Induktion nach n. Für $n = 1$ ist

$$A_1 = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \end{pmatrix}$$

und somit folgt die Induktionsverankerung

$$\det A_1 = \det \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \end{pmatrix} = x_1 - x_0\,.$$

Für Matrizen der Gestalt

$$A_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}$$

gelte also die Induktionsvoraussetzung

$$\det A_n = \prod_{0\le i<j\le n} (x_j - x_i)\,.$$

Sei nun

$$A_{n+1} = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n & x_0^{n+1} \\ 1 & x_1 & x_1^2 & \dots & x_1^n & x_1^{n+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n & x_n^{n+1} \\ 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n & x_{n+1}^{n+1} \end{pmatrix}.$$

Für die Berechnung der Determinante von A_{n+1} ergibt die Subtraktion des x_0–fachen der $(k-1)$–ten Spalte von der k–ten Spalte für $k = n+1, n, \dots, 1$

$$\det A_{n+1} = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & x_1 - x_0 & x_1^2 - x_0 x_1 & \dots & x_1^n - x_0 x_1^{n-1} & x_1^{n+1} - x_0 x_1^n \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n - x_0 & x_n^2 - x_0 x_n & \dots & x_n^n - x_0 x_n^{n-1} & x_n^{n+1} - x_0 x_n^n \\ 1 & x_{n+1} - x_0 & x_{n+1}^2 - x_0 x_{n+1} & \dots & x_{n+1}^n - x_0 x_{n+1}^{n-1} & x_{n+1}^{n+1} - x_0 x_{n+1}^n \end{vmatrix}.$$

Durch Entwicklung nach der ersten Zeile folgt dann

$$\begin{aligned} \det A_{n+1} &= \begin{vmatrix} x_1 - x_0 & (x_1 - x_0)x_1 & \dots & (x_1 - x_0)x_1^{n-1} & (x_1 - x_0)x_1^n \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n+1} - x_0 & (x_{n+1} - x_0)x_{n+1} & \dots & (x_{n+1} - x_0)x_{n+1}^{n-1} & (x_{n+1} - x_0)x_{n+1}^n \end{vmatrix} \\ &= (x_1 - x_0) \cdots (x_{n+1} - x_0) \begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1} & x_1^n \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{n+1} & \dots & x_{n+1}^{n-1} & x_{n+1}^n \end{vmatrix}. \end{aligned}$$

Mit der Induktionsvoraussetzung ist

$$\begin{vmatrix} 1 & x_1 & \dots & x_1^{n-1} & x_1^n \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{n+1} & \dots & x_{n+1}^{n-1} & x_{n+1}^n \end{vmatrix} = \prod_{1 \le i < j \le n+1} (x_j - x_i)$$

und somit gilt

$$\det A_{n+1} = (x_1 - x_0) \dots (x_{n+1} - x_0) \prod_{1 \le i < j \le n+1} (x_j - x_i) = \prod_{0 \le i < j \le n+1} (x_j - x_i).$$

■

Für paarweise verschiedene Stützstellen $x_i \neq x_j$ für alle $i \neq j$ folgt somit $\det A_n \neq 0$ und damit die Invertierbarkeit der Systemmatrizen A_n. Damit ist das lineare Gleichungssystem (1.4) und somit die Interpolationsaufgabe (1.2) eindeutig lösbar.

Beispiel 1.1. *Für $n = 2$ und $x \in [-5, 5]$ führt die Interpolation der Funktion*

$$f(x) \;=\; \frac{1}{1+x^2}$$

in den gleichmäßig verteilten Stützstellen

$$x_0 = -5, \quad x_1 = 0, \quad x_2 = 5$$

auf das lineare Gleichungssystem

$$\begin{pmatrix} 1 & -5 & 25 \\ 1 & 0 & 0 \\ 1 & 5 & 25 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \;=\; \frac{1}{26} \begin{pmatrix} 1 \\ 26 \\ 1 \end{pmatrix}$$

mit der Lösung

$$a_0 \;=\; 1, \quad a_1 \;=\; 0, \quad a_2 \;=\; -\frac{1}{26}\,.$$

Damit lautet das zugehörige Interpolationspolynom

$$f_2(x) \;=\; 1 - \frac{1}{26}x^2\,.$$

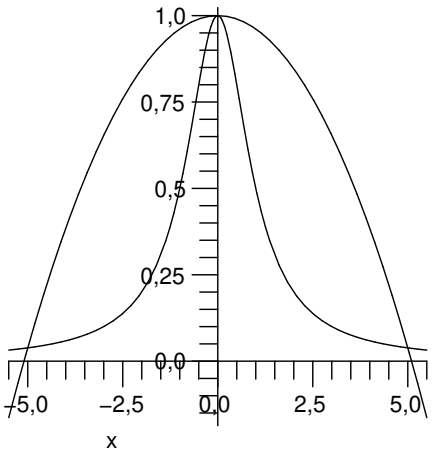


Abbildung 1.1: Interpolationspolynom $f_2(x)$ von $f(x) = \dfrac{1}{1+x^2}$.

Allgemein kann das Interpolationspolynom (1.1) geschrieben werden als

$$f_n(x) \;=\; \sum_{k=0}^{n} a_k x^k \;=\; \sum_{k=0}^{n} a_k \varphi_k(x)$$

mit den Basisfunktionen

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x, \quad \varphi_2(x) = x^2, \quad \ldots, \quad \varphi_n(x) = x^n. \tag{1.8}$$

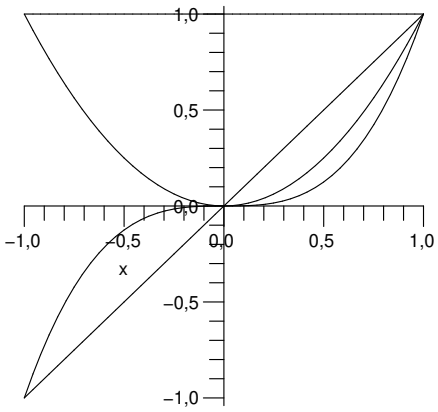


Abbildung 1.2: Monome $\varphi_k(x) = x^k$ für $k = 0, 1, 2, 3$.

Die Verwendung von Monomen $\varphi_k(x) = x^k$ als Basisfunktionen erfordert die Lösung der linearen Gleichunssysteme (1.4), deren Systemmatrizen durch die Vandermonde–Matrizen (1.6) gegeben sind. Es stellt sich die Frage, wie durch eine geeignete Wahl von Basisfunktionen $\varphi_k(x)$ die Lösung des resultierenden linearen Gleichungssystems möglichst einfach gestaltet werden kann. Dies motiviert die folgende Definition der Lagrange–Polynome.
Die Monome $\varphi_k(x) = x^k$ bilden eine Basis des linearen Raumes der Polynome vom Grad n,

$$\Pi_n \;=\; \mathrm{span}\Big\{\varphi_k(x)\Big\}_{k=0}^{n} \;=\; \mathrm{span}\Big\{x^k\Big\}_{k=0}^{n}.$$

Jedes Polynom $f_n \in \Pi_n$ mit maximalen Grad n kann also als Linearkombination der Basisfunktionen $\varphi_k(x)$ dargestellt werden,

$$f_n(x) \;=\; \sum_{k=0}^{n} a_k \varphi_k(x) \;=\; \sum_{k=0}^{n} a_k x^k.$$

Der Übergang zu einer anderen Basis

$$\Pi_n \;=\; \mathrm{span}\Big\{x^k\Big\}_{k=0}^{n} \;=\; \mathrm{span}\Big\{\psi_k(x)\Big\}_{k=0}^{n}$$

mit Basisfunktionen $\psi_k(x)$ ermöglicht für das Interpolationspolynom den Ansatz

$$f_n(x) \;=\; \sum_{k=0}^{n} b_k \psi_k(x)$$

mit noch zu bestimmenden Zerlegungskoeffizienten b_k für $k = 0, \ldots, n$. Die zugehörigen Interpolationsgleichungen lauten dann

$$f_n(x_i) \;=\; \sum_{k=0}^{n} b_k \psi_k(x_i) \;=\; f_i \quad \text{für } i = 0, \ldots, n$$

bzw.

$$\begin{pmatrix} \psi_0(x_0) & \dots & \psi_n(x_0) \\ \vdots & & \vdots \\ \psi_0(x_n) & \dots & \psi_n(x_n) \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}. \tag{1.9}$$

Die Basisfunktionen $\psi_k(x)$ sollen nun derart gewählt werden, so daß das lineare Gleichungssystem (1.9) besonders einfach zu lösen ist. Insbesondere aus der Forderung

$$\psi_k(x_i) = \begin{cases} 1 & \text{für } i = k, \\ 0 & \text{für } i \neq k \end{cases} \tag{1.10}$$

folgt

$$b_k = f_k \quad \text{für alle } k = 0, \dots, n$$

und somit

$$f_n(x) = \sum_{k=0}^{n} f_k \psi_k(x).$$

Die Forderung (1.10) motiviert die Definition der Lagrange–Polynome

$$L_k^n(x) \;=\; \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} \quad \text{für } k = 0, \dots, n. \tag{1.11}$$

Die in (1.11) definierten Lagrange–Polynome $\{L_k^n(x)\}_{k=0}^n$ bilden eine Basis im Raum Π_n der Polynome vom Grad n. Für das Interpolationspolynom ergibt sich dann die Darstellung

$$f_n(x) \;=\; \sum_{k=0}^{n} f(x_k) L_k^n(x)\,. \tag{1.12}$$

Beispiel 1.2. *Für die Stützstellen*

$$x_0 = -5, \quad x_1 = 0, \quad x_2 = 5$$

ergeben sich die in Abbildung 1.3 *dargestellten Lagrange–Polynome*

$$L_0^2(x) = \frac{1}{50}x^2 - \frac{1}{10}x, \quad L_1^2(x) = -\frac{1}{25}x^2 + 1, \quad L_2^2(x) = \frac{1}{50}x^2 + \frac{1}{10}x\,.$$

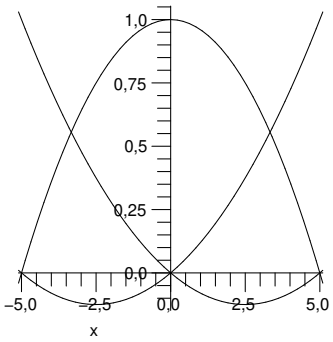


Abbildung 1.3: Lagrange–Polynome $L_k^2(x)$, $k = 0, 1, 2$.

Für das Interpolationspolynom $f_2(x)$ der Funktion

$$f(x) \;=\; \frac{1}{1+x^2}$$

folgt dann

$$f_2(x) \;=\; f(-5)L_0^2(x) + f(0)L_1^2(x) + f(5)L_2^2(x) \;=\; 1 - \frac{1}{26}x^2\,.$$

1.2 Abschätzung des Interpolationsfehlers

Für eine gegebene Funktion f bezeichne $f_n \in \Pi_n$ das Interpolationspolynom mit

$$f_n(x_i) \;=\; f(x_i) \quad \text{für } i = 0, \ldots, n. \tag{1.13}$$

Dabei seien die $n+1$ paarweise verschiedenen Stützstellen $x_i \in [a,b]$ in einem beschränkten Intervall $[a,b]$ gegeben. Abzuschätzen bleibt der Fehler

$$e_n(x) \;:=\; f(x) - f_n(x) \quad \text{für } x \in [a,b].$$

Satz 1.1. *Sei $f(x)$ im Intervall $[a,b]$ $n+1$–mal stetig differenzierbar, und sei $f_n(x)$ das Interpolationspolynom vom Grad n mit $f_n(x_i) = f(x_i)$ in den $n+1$ paarweise verschiedenen Stützstellen $x_i \in [a,b]$. Für den Interpolationsfehler in $x \in [a,b]$ gilt dann die Darstellung*

$$f(x) - f_n(x) \;=\; \frac{1}{(n+1)!} f^{(n+1)}(\xi(x)) \prod_{j=0}^{n} (x - x_j) \tag{1.14}$$

mit einer geeigneten Zwischenwertstelle $\xi(x) \in [a,b]$.

Der Beweis von Satz 1.1 beruht auf einer Anwendung des Satzes von Rolle:

Satz 1.2 (Satz von Rolle)**.** *Sei $f(x)$ für $x \in [a,b]$ stetig, differenzierbar in (a,b), und es gelte $f(a) = f(b)$. Dann gibt es wenigstens eine Stelle $\xi \in (a,b)$ mit*

$$f'(\xi) = 0.$$

Beweis: Eine in $[a,b]$ stetige Funktion nimmt dort ihr Maximum und ihr Minimum an. Folglich existieren $\alpha, \beta \in [a,b]$ mit

$$f(\alpha) = \min_{x\in[a,b]} \le \max_{x\in[a,b]} f(x) = f(\beta)\,.$$

Im Fall $f(\alpha) = f(\beta)$ ist $f(x)$ in $[a,b]$ konstant und somit gilt $f'(x) = 0$ für alle $x \in (a,b)$.
Im Fall $f(\alpha) < f(\beta)$ sind wieder zwei Fälle zu unterscheiden: Für $f(a) = f(b) < f(\beta)$ folgt $\beta \in (a,b)$, d.h. $f(x)$ hat in β ein lokales Maximum und es folgt $f'(\beta) = 0$.
Für $f(a) = f(b) > f(\alpha)$ folgt entsprechend, dass $f(x)$ in $\alpha \in (a,b)$ ein lokales Minimum hat, d.h. es gilt $f'(\alpha) = 0$. ∎

Beweis von Satz 1.1: In den Stützstellen $x = x_i$ folgt aus den Interpolationsgleichungen (1.13)

$$e_n(x_i) \;=\; f(x_i) - f_n(x_i) \;=\; 0 \quad \text{für } i = 0, \ldots, n.$$

Für die von einem reellen Parameter $\alpha \in \mathbb{R}$ abhängige Funktion

$$g_\alpha(x) \;:=\; e_n(x) - \alpha \prod_{j=0}^{n} (x - x_j)$$

folgt dann

$$g_\alpha(x_i) = 0 \quad \text{für } i = 0, \ldots, n.$$

Für ein beliebiges $\overline{x} \in [a, b]$ mit $\bar{x} \neq x_i$ für $i = 0, \ldots, n$ ist

$$\overline{\alpha} \;=\; \frac{e_n(\overline{x})}{\prod\limits_{j=0}^{n} (\overline{x} - x_j)}$$

wohldefiniert und es folgt

$$g_{\overline{\alpha}}(\overline{x}) \;=\; 0.$$

Damit hat die Funktion $g_{\overline{\alpha}}(x)$ im Intervall $[a, b]$ $n + 2$ paarweise verschiedene Nullstellen $x_0, \ldots, x_n$ und $\overline{x}$. Nach dem Satz von Rolle besitzt dann $g'_{\overline{\alpha}}(x)$ in $[a, b]$ $n + 1$ paarweise verschiedene Nullstellen und durch rekursives Anwenden folgt, daß $g^{(n+1)}_{\overline{\alpha}}(x)$ in $[a, b]$ eine Nullstelle $\xi(\overline{x}) \in [a, b]$ besitzt. Für diese ist

$$\begin{aligned} 0 = g^{(n+1)}_{\overline{\alpha}}(\xi(\overline{x})) \;&=\; \frac{d^{n+1}}{dx^{n+1}} \left[f(x) - f_n(x) - \overline{\alpha} \prod_{j=0}^{n} (x - x_j) \right]_{x=\xi(\overline{x})} \\ &=\; f^{(n+1)}(\xi(\overline{x})) - \overline{\alpha}\,(n+1)! \end{aligned}$$

und somit

$$\overline{\alpha} \;=\; \frac{1}{(n+1)!} f^{(n+1)}(\xi(\overline{x}))\,.$$

Dann folgt

$$0 \;=\; g_{\overline{\alpha}}(\overline{x}) \;=\; f(\overline{x}) - f_n(\overline{x}) - \frac{1}{(n+1)!} f^{(n+1)}(\xi(\overline{x})) \prod_{j=0}^{n} (\overline{x} - x_j)$$

und somit gilt

$$f(\overline{x}) - f_n(\overline{x}) \;=\; \frac{1}{(n+1)!} f^{(n+1)}(\xi(\overline{x})) \prod_{j=0}^{n} (\overline{x} - x_j)$$

für alle $\overline{x} \in [a, b]$ mit $\overline{x} \neq x_i$, $i = 0, \ldots, n$. Offensichtlich bleibt dies auch für alle Stützstellen $\overline{x} = x_i$, $i = 0, \ldots, n$, richtig, da dann beide Seiten Null sind. ∎

Folgerung 1.1. *Aus der punktweisen Fehlerdarstellung* (1.14) *folgt auch eine Fehlerabschätzung in der Maximum–Norm,*

$$\begin{aligned}\max_{x\in[a,b]} |f(x)-f_n(x)| &= \frac{1}{(n+1)!}\max_{x\in[a,b]}\left|f^{(n+1)}(\xi(x))\prod_{j=0}^{n}(x-x_j)\right| \\ &\le \frac{1}{(n+1)!}\max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|\max_{x\in[a,b]}\left|\prod_{j=0}^{n}(x-x_j)\right|. \qquad (1.15)\end{aligned}$$

Beispiel 1.3. *Betrachtet wird die Interpolationsaufgabe zur Approximation der Funktion* $f(x)=\sin x$ *für* $x\in[0,\frac{\pi}{2}]$. *Für* $n=1$ *ist*

$$x_0=0,\quad x_1=\frac{\pi}{2},\quad f_1(x)=\frac{2}{\pi}\,x$$

und es gilt die Fehlerabschätzung

$$\max_{x\in[0,\frac{\pi}{2}]}|f(x)-f_1(x)| \le \frac{1}{2}\max_{x\in[0,\frac{\pi}{2}]} x\Big(\frac{\pi}{2}-x\Big) = \frac{\pi^2}{32}\approx 0.3084\,.$$

Dabei wird das Maximum für $x=\frac{\pi}{4}$ *angenommen. Andererseits ist*

$$\max_{x\in[0,\frac{\pi}{2}]}|f(x)-f_1(x)| = \max_{x\in[0,\frac{\pi}{2}]}\Big|\sin x-\frac{2}{\pi}x\Big| \approx 0.2105\,.$$

Der tatsächliche Interpolationsfehler wird in diesem Beispiel um einen Faktor von ca. 1.5 *überschätzt.*

Die Abschätzung (1.15) des Interpolationsfehlers $f(x)-f_n(x)$ zeigt, daß neben der Differenzierbarkeit der zu approximierenden Funktion $f(x)$ auch die Wahl der Stützstellen x_i wesentlich für die Güte der Approximation f_n ist, siehe hierzu auch die beiden folgenden Beispiele.

Beispiel 1.4. *Für die lineare Interpolierende* $f_1(x)=x$ *der Funktion* $f(x)=\sqrt{x}$ *für* $x\in[0,1]$ *mit den Stützstellen* $x_0=0$ *und* $x_1=1$ *folgt für den Fehler die Darstellung*

$$f(x)-f_1(x)=\frac{1}{8}[\xi(x)]^{-3/2}x(1-x)$$

für $x\in(0,1)$ *mit einer geeigneten (unbekannten) Zwischenwertstelle* $\xi(x)\in(0,1)$. *Offenbar kann in diesem Beispiel nicht auf die Fehlerabschätzung* (1.15) *geschlossen werden, da das Maximum*

$$\max_{x\in[0,1]}|f''(x)| = \frac{1}{4}\max_{x\in[0,1]} x^{-3/2}$$

nicht existiert.

Beispiel 1.5. *Die Interpolation der Funktion*

$$f(x) = \frac{1}{1+x^2} \quad \textit{für } x \in [-5,+5]$$

in den gleichmäßig verteilten Stützstellen

$$x_i = -5 + \frac{10i}{n} \quad \textit{für } i = 0,\ldots,n$$

ergibt die in Abbildung 1.4 *für* $n = 5$ *und* $n = 10$ *dargestellten Interpolationspolynome mit den in der Nähe der Randpunkte* ± 5 *auftretenden Oszillationen.*

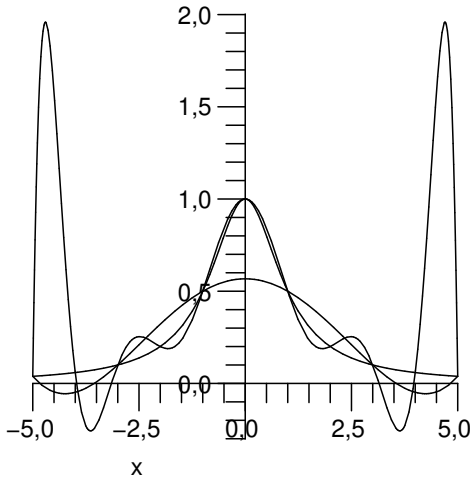


Abbildung 1.4: Interpolationspolynome $f_5(x)$ und $f_{10}(x)$ der Funktion $f(x) = \dfrac{1}{1+x^2}$.

Die in Beispiel 1.5 betrachtete Interpolationsaufgabe motiviert die Wahl der Stützstellen x_j in einer solchen Weise, so daß

$$\max_{x\in[a,b]} \left| \prod_{j=0}^{n} (x - x_j) \right|$$

minimal wird. Zu lösen ist also das Minimierungsproblem

$$\min_{x_0,\ldots,x_n} \max_{x\in[a,b]} \left| \prod_{j=0}^{n} (x - x_j) \right| .$$

Die Lösung dieser Min–Max–Aufgabe beruht auf den im folgenden Abschnitt behandelten Tschebyscheff–Polynomen.

1.3 Tschebyscheff–Polynome

Der Raum Π_n der Polynome vom Grad n kann neben der Beschreibung durch die Monome x^k bzw. durch die Lagrange–Polynome $L_k^n(x)$ auch durch die Tschebyscheff–Polynome $T_k(x)$ charakterisiert werden. Diese werden rekursiv definiert durch

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \quad \text{für } k = 1, 2, \dots . \end{aligned} \tag{1.16}$$

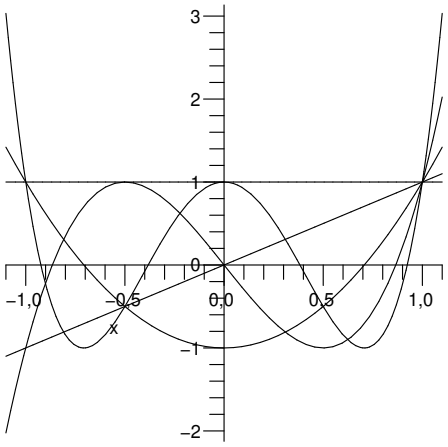


Abbildung 1.5: Tschebyscheff–Polynome $T_k(x)$ für $k = 0, \dots, 4$.

Lemma 1.2. *Für $x \in [-1, +1]$ und $k = 0, 1, 2, \dots$ gilt für die in (1.16) definierten Tschebyscheff–Polynome $T_k(x)$ die alternative Darstellung*

$$T_k(x) = \cos(k \arccos x) . \tag{1.17}$$

Beweis: Der Beweis erfolgt durch vollständige Induktion nach k. Mit $T_0(x) = 1$ und $T_1(x) = x$ gilt offenbar die Induktionsverankerung für $k = 0$ und für $k = 1$.
Aus dem Additionstheorem

$$\cos\alpha + \cos\beta \;=\; 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}$$

folgt

$$\cos\alpha \;=\; 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2} - \cos\beta$$

und mit

$$\alpha \;:=\; (k+1)\arccos x, \qquad \beta \;:=\; (k-1)\arccos x$$

ergibt sich dann die Behauptung

$$\begin{aligned}\cos[(k+1)\arccos x] &= 2\cos[k\arccos x]\cos\arccos x - \cos[(k-1)\arccos x] \\ &= 2xT_k(x) - T_{k-1}(x) \;=\; T_{k+1}(x)\end{aligned}$$

für alle $k = 1, 2, \ldots$. ∎

Aus der Darstellung (1.17) lassen sich nun einige wichtige Eigenschaften der Tschebyscheff–Polynome $T_k(x)$ ablesen. Zunächst ist

$$\max_{x\in[-1,+1]} |T_k(x)| \;=\; \max_{x\in[-1,+1]} |\cos(k\arccos x)| \;=\; 1\,. \tag{1.18}$$

Für

$$\widetilde{x}_i^{(k)} = \cos\frac{i\pi}{k} \quad \text{für } i = 0, \ldots, k \tag{1.19}$$

gilt dabei

$$T_k(\widetilde{x}_i^{(k)}) = \cos(k\arccos\widetilde{x}_i^{(k)}) = \cos\, i\pi = (-1)^i \quad \text{für } i = 0, \ldots, k.$$

Im Intervall $[-1, 1]$ hat das Tschebyscheff–Polynom $T_k(x)$ also k Vorzeichenwechsel, d.h. $T_k(x)$ besitzt in $[-1, 1]$ k Nullstellen. Diese ergeben sich aus der Forderung

$$T_k(x) = \cos(k\arccos x) \;=\; 0$$

bzw.

$$k\arccos x \;=\; \frac{\pi}{2} + i\pi \quad \text{für } i \in \mathbb{N}.$$

Daraus folgt

$$x_i^{(k)} = \cos\frac{(1+2i)\pi}{2k} \quad \text{für } i = 0, \ldots, k-1. \tag{1.20}$$

Neben der hier durch die Rekursionsvorschrift (1.16) erfolgten Definition der Tschebyscheff–Polynome und der dazu äquivalenten Darstellung (1.17) durch trigonometrische Polynome ermöglichen die Tschebyscheff–Polynome eine dritte Darstellung, die insbesondere für die Funktionsauswertung von $T_k(x)$ für $x > 1$ wesentlich sein wird.

Lemma 1.3. *Für $k = 0, 1, 2, \ldots$ gelten für die durch* (1.16) *definierten Tschebyscheff–Polynome $T_k(x)$ die Darstellungen*

$$T_k(x) = \frac{1}{2}\Big[(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k\Big] \tag{1.21}$$

$$= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^k + (x+\sqrt{x^2-1})^{-k}\Big]. \tag{1.22}$$

Beweis: Der Beweis erfolgt durch vollständige Induktion nach k. Für $k = 0$ und für $k = 1$ gilt (1.21) offensichtlich. Als Induktionsvoraussetzung gelte also

$$\begin{aligned} T_{k-1}(x) &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^{k-1} + (x-\sqrt{x^2-1})^{k-1}\Big], \\ T_k(x) &= \frac{1}{2}\Big[(x+\sqrt{x^2-1})^{k} + (x-\sqrt{x^2-1})^{k}\Big]. \end{aligned}$$

Mit der rekursiven Definition (1.16) der Tschebyscheff–Polynome $T_k(x)$ folgt

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \\ &= x\left[(x+\sqrt{x^2-1})^{k} + (x-\sqrt{x^2-1})^{k}\right] \\ &\qquad -\frac{1}{2}\Big[(x+\sqrt{x^2-1})^{k-1} + (x-\sqrt{x^2-1})^{k-1}\Big] \\ &= \frac{1}{2}(x+\sqrt{x^2-1})^{k-1}\Big[2x\,(x+\sqrt{x^2-1}) - 1\Big] \\ &\qquad +\frac{1}{2}(x-\sqrt{x^2-1})^{k-1}\Big[2x\,(x-\sqrt{x^2-1}) - 1\Big]. \end{aligned}$$

Die erste Behauptung (1.21) ergibt sich nun aus

$$2x(x \pm \sqrt{x^2-1}) - 1 \;=\; x^2 + 2x\sqrt{x^2-1} + x^2 - 1 \;=\; (x \pm \sqrt{x^2-1})^2.$$

Die zweite Behauptung (1.22) folgt unmittelbar aus

$$x - \sqrt{x^2-1} \;=\; \frac{[x-\sqrt{x^2-1}][x+\sqrt{x^2-1}]}{x+\sqrt{x^2-1}} \;=\; \frac{1}{x+\sqrt{x^2-1}}.$$

■

Für ein beliebiges Intervall $[a, b]$ mit $0 < a < b$ definiert

$$x \;=\; \frac{b+a-2t}{b-a} \tag{1.23}$$

eine Transformation von $[-1, +1]$ auf $[a, b]$. Diese Transformation ermöglicht die Definition der skalierten Tschebyscheff–Polynome

$$\widetilde{T}_k(t) \;:=\; \frac{T_k\big(\frac{b+a-2t}{b-a}\big)}{T_k\big(\frac{b+a}{b-a}\big)} \quad \text{für } t \in [a, b]. \tag{1.24}$$

Nach Konstruktion ist $\widetilde{T}_k(0) = 1$, d.h. $\widetilde{T}_k \in \Pi_k^1$ mit

$$\Pi_k^1 = \Big\{p_k \in \Pi_k \,:\, p_k(0) = 1\Big\}.$$

Die modizierten Tschebyscheff–Polynome $\widetilde{T}_n \in \Pi_n^1$ sind die Polynome vom Polynomgrad n mit dem kleinsten Maximum im Intervall $[a, b]$:

Satz 1.3. *Für $0 < a < b$ sind die modifizierten Tschebyscheff–Polynome $\widetilde{T}_n(t)$ Lösung der Minimierungsaufgabe*

$$\min_{p_n \in \Pi_n^1} \max_{t\in[a,b]} \left|p_n(t)\right| \;=\; \max_{t\in[a,b]} \left|\widetilde{T}_n(t)\right| \;=\; \frac{2q^n}{1+q^{2n}} \quad \textit{mit} \quad q \;=\; \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}}\,.$$

Beweis: Der Beweis erfolgt indirekt durch die Annahme, es existiere ein Polynom $q_n \in \Pi_n^1$ mit

$$\max_{t\in[a,b]} \left|q_n(t)\right| \;<\; \max_{t\in[a,b]} \left|\widetilde{T}_n(t)\right|.$$

Für die durch (1.19) gegebenen Argumente

$$\widetilde{x}_i^{(n)} \;=\; \cos\frac{i\pi}{n} \in [-1,1]$$

ist

$$\widetilde{t}_i^{(n)} \;:=\; \frac{1}{2}\left[(b+a)-(b-a)\widetilde{x}_i^{(n)}\right] \in [a,b] \quad \text{für } i=0,\ldots,n.$$

Dann wird durch

$$\widetilde{T}_n(\widetilde{t}_i^{(n)}) \;=\; \frac{T_n(\widetilde{x}_i^{(n)})}{T_n(\frac{b+a}{b-a})} \;=\; \frac{(-1)^i}{T_n(\frac{b+a}{b-a})}$$

das Maximum bzw. das Minimum des modifizierten Tschebyscheff–Polynoms $\widetilde{T}_n(t)$ in $[a,b]$ angenommen. In diesen Punkten gilt

$$\left|q_n(\widetilde{t}_i^{(n)})\right| \;\le\; \max_{t\in[a,b]} \left|q_n(t)\right| \;<\; \frac{1}{T_n(\frac{b+a}{b-a})} \quad \text{für } i=0,\ldots,n.$$

Insbesondere für $i=2j$ ist $\widetilde{T}_n(\widetilde{t}_{2j}^{(n)}) > 0$ und somit gilt

$$-\widetilde{T}_n(\widetilde{t}_{2j}^{(n)}) < q_n(\widetilde{t}_{2j}^{(n)}) < \widetilde{T}_n(\widetilde{t}_{2j}^{(n)}).$$

Entsprechend ergibt sich für $i=2j+1$ $\widetilde{T}_n(\widetilde{t}_{2j+1}^{(n)}) < 0$ und somit gilt

$$\widetilde{T}_n(\widetilde{t}_{2j+1}^{(n)}) < q_n(\widetilde{t}_{2j+1}^{(n)}) < -\widetilde{T}_n(\widetilde{t}_{2j+1}^{(n)}).$$

Für das Polynom $r_n(t) := \widetilde{T}_n(t) - q_n(t) \in \Pi_n$ folgt dann

$$r_n(\widetilde{t}_{2j}^{(n)}) \;=\; \widetilde{T}_n(\widetilde{t}_{2j}^{(n)}) - q_n(\widetilde{t}_{2j}^{(n)}) \;>\; 0$$

und

$$r_n(\widetilde{t}_{2j+1}^{(n)}) \;=\; \widetilde{T}_n(\widetilde{t}_{2j+1}^{(n)}) - q_n(\widetilde{t}_{2j+1}^{(n)}) \;<\; 0\,.$$

Zwischen den $n+1$ paarweise verschiedenen Stellen $\widetilde{t}_i^{(n)}$ finden also n Vorzeichenwechsel statt, d.h. das Polynom $r_n(t)$ besitzt im Intervall $[a,b]$ mindestens n Nullstellen. Wegen $\widetilde{T}_n \in \Pi_n^1$ und $q_n \in \Pi_n^1$ ist weiterhin

$$r_n(0) \;=\; \widetilde{T}_n(0) - q_n(0) \;=\; 1-1 \;=\; 0$$

und somit ist Null eine weitere Nullstelle von $r_n(t)$. Damit besitzt das Polynom $r_n(t)$ vom Polynomgrad n auf der reellen Achse mindestens $n+1$ Nullstellen. Daraus folgt $r_n(t) \equiv 0$ für alle $t \in \mathbb{R}$ und somit $q_n = \widetilde{T}_n$ im Widerspruch zur Annahme.
Zu bestimmen bleibt der maximale Wert von

$$\max_{t\in[a,b]} \left|\widetilde{T}_n(t)\right| \;=\; \frac{1}{T_n(\frac{b+a}{b-a})}\,.$$

Mit der Darstellung (1.22) und

$$\begin{aligned} q &= \frac{b+a}{b-a} + \sqrt{\left(\frac{b+a}{b-a}\right)^2 - 1} \\ &= \frac{1}{b-a}\left[(b+a) + \sqrt{(b+a)^2 - (b-a)^2}\right] \\ &= \frac{1}{b-a}\left[b+a+2\sqrt{ab}\right] \\ &= \frac{(\sqrt{b}+\sqrt{a})^2}{(\sqrt{b}+\sqrt{a})(\sqrt{b}-\sqrt{a})} = \frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}} \end{aligned}$$

ist schließlich

$$T_n\Big(\frac{b+a}{b-a}\Big) \;=\; \frac{1}{2}\left[q^n + q^{-n}\right] \;=\; \frac{q^{2n}+1}{2q^n}\,.$$

■

Für das Tschebyscheff–Polynom $T_{n+1} \in \Pi_{n+1}$ mit den Nullstellen $x_i^{(n+1)}$ gilt die Darstellung

$$T_{n+1}(x) = \alpha \prod_{i=0}^{n} \left(x - x_i^{(n+1)}\right) = \alpha\, x^{n+1} + p_n(x)$$

mit einem Polynom $p_n(x)$ vom Polynomgrad n. Andererseits ergibt sich aus der rekursiven Definition (1.16) der Tschebyscheff–Polynome die Darstellung

$$T_{n+1}(x) = 2^n\, x^{n+1} + p_n(x)$$

wieder mit einem Polynom $p_n \in \Pi_n$ vom Polynomgrad n. Durch Vergleich der führenden Koeffizienten folgt $\alpha = 2^n$ und somit

$$T_{n+1}(x) = 2^n \prod_{i=0}^{n} \left(x - x_i^{(n+1)}\right).$$

Für $x \in [-1,+1]$ gilt dann

$$\left|\prod_{i=0}^{n} \left(x - x_i^{(n+1)}\right)\right| \;=\; \left|2^{-n}\, T_{n+1}(x)\right| \;\le\; 2^{-n}\,.$$

Werden also im Intervall $[-1,+1]$ die Nullstellen $x_i^{(n+1)}$ des Tschebyscheff–Polynoms $T_{n+1}(x)$ als Interpolationsknoten x_i gewählt, d.h. ist

$$f_n(x_i^{(n+1)}) \;=\; f(x_i^{(n+1)}) \quad \text{für} \quad i = 0, \ldots, n,$$

dann ergibt sich aus (1.15) die Fehlerabschätzung

$$\max_{x\in[-1,+1]} |f(x) - f_n(x)| \;\leq\; \frac{2^{-n}}{(n+1)!} \max_{x\in[-1,+1]} |f^{(n+1)}(x)|\,. \tag{1.25}$$

Für ein beliebig gegebenes Intervall $[a,b]$ können die Stützstellen aus $[-1,+1]$ durch eine geeignete Transformation entsprechend übertragen werden: Für die Nullstellen $x_i^{(n+1)}$ von $T_{n+1}(x)$ ergeben sich die transformierten Stützstellen

$$t_i^{n+1} = \frac{1}{2}\Big[(b+a) - (b-a)x_i^{(n+1)}\Big] \in [a,b] \quad \text{für } i = 0, \ldots, n.$$

Beschreibt $f_n(t)$ das Interpolationspolynom der im Intervall $[a,b]$ gegebenen Funktion $f(t)$ mit

$$f_n(t_i^{(n+1)}) = f(t_i^{(n+1)}) \quad \text{für } i = 0, \ldots, n,$$

dann lautet die Darstellung (1.14) des Fehlers für $t \in [a,b]$ und Rückführung auf $x \in [-1,1]$

$$\begin{aligned}
f(t) - f_n(t) &= \frac{1}{(n+1)!}\, f^{(n+1)}(\xi(t)) \prod_{j=0}^{n} (t - t_j^{(n+1)}) \\
&= \frac{1}{(n+1)!}\, f^{(n+1)}(\xi(t)) \prod_{j=0}^{n} \left[\frac{1}{2}(b-a)\Big(x_i^{(n+1)} - x\Big)\right] \\
&= \frac{1}{(n+1)!}\, f^{(n+1)}(\xi(t))\; (-1)^{n+1} \left(\frac{b-a}{2}\right)^{n+1} \prod_{j=0}^{n} \Big(x - x_i^{(n+1)}\Big) \\
&= \frac{1}{(n+1)!}\, f^{(n+1)}(\xi(t))\; (-1)^{n+1} \left(\frac{b-a}{2}\right)^{n+1} 2^{-n}\, T_{n+1}(x)\,.
\end{aligned}$$

Dann ergibt sich die Fehlerabschätzung

$$\max_{t\in[a,b]} \Big|f(t) - f_n(t)\Big| \;\leq\; \frac{2^{-n}}{(n+1)!} \left(\frac{b-a}{2}\right)^{n+1} \max_{t\in[a,b]} \Big|f^{(n+1)}(t)\Big|, \tag{1.26}$$

welche für $[a,b] = [-1,1]$ mit (1.25) übereinstimmt.

Beispiel 1.6. *Die Nullstellen $x_i^{(n+1)}$ des Tschebyscheff–Polynoms $T_{n+1}(x)$ im Intervall $[-1,1]$ sind gegeben durch*

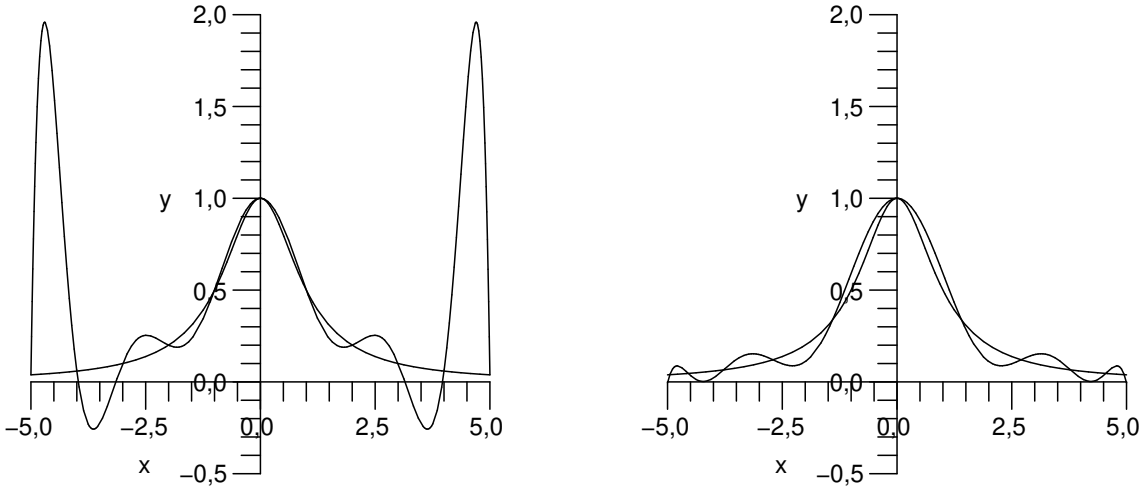
$$x_i^{(n+1)} \;=\; \cos\frac{(1+2i)\pi}{2(n+1)} \quad \textit{für } i = 0, 1, 2, \ldots, n.$$

Für die Interpolation der Funktion $f(t) = \dfrac{1}{1+t^2}$ *im Intervall* $[-5,5]$ *ergeben sich dann die transformierten Stützstellen*

$$t_i^{(n+1)} \;=\; -5\, x_i^{(n+1)} \;=\; -5\cos\frac{(1+2i)\pi}{2(n+1)}, \quad \textit{für } i = 0,1,2,\ldots,n$$

und es gilt die Fehlerabschätzung

$$\max_{t\in[-5,5]} \Big| f(t) - f_n(t) \Big| \;\le\; \frac{2^{-n}}{(n+1)!}\, 5^{n+1} \max_{t\in[-5,5]} \Big| f^{(n+1)}(t) \Big|.$$



$$t_i = -5 + \frac{i}{10}, i = 0,\ldots,10 \qquad\qquad t_i = -5\cos\frac{(1+2i)\pi}{22}, i = 0,\ldots,10$$

Abbildung 1.6: Interpolationspolynome $f_{10}(t)$ *von* $f(t) = \dfrac{1}{1+t^2}$.

1.4 Hermite–Interpolation

Setzt man in den $n+1$ paarweise verschiedenen Stützstellen $x_0 < x_1 < \ldots x_n$ neben den Funktionswerten $f(x_i)$ auch die ersten Ableitungen $f'(x_i)$ der zu interpolierenden Funktion $f(x)$ voraus, so kann aus der Kenntnis dieser Daten ein Interpolationspolynom

$$f_{2n+1}(x) \;=\; \sum_{k=0}^{2n+1} a_k x^k \tag{1.27}$$

gewonnen werden. Die Interpolationsgleichungen zur Bestimmung der $2n+2$ unbekannten Zerlegungskoeffizienten $a_0, \ldots, a_{2n+1}$ lauten dann

$$f_{2n+1}(x_i) = f(x_i), \quad f'_{2n+1}(x_i) = f'(x_i) \quad \text{für } i = 0, \ldots, n. \tag{1.28}$$

Das resultierende Interpolationspolynom $f_{2n+1}(x)$ wird als Hermitesches Interpolationspolynom bezeichnet. Die eindeutige Lösbarkeit des linearen Gleichungssystems (1.28) ergibt sich aus der Eindeutigkeit der Interpolationsaufgabe.

Satz 1.4. *Für paarweise verschiedene Stützstellen $x_0 < x_1 < \ldots < x_n$ besitzt die Hermitesche Interpolationsaufgabe* (1.28) *eine eindeutig bestimmte Lösung $f_{2n+1}(x)$.*

Beweis: Seien $f_{2n+1}(x)$ und $g_{2n+1}(x)$ zwei Lösungen der Hermiteschen Interpolationsaufgabe (1.28), d.h. es gelten

$$f_{2n+1}(x_i) = g_{2n+1}(x_i) = f(x_i), \quad f'_{2n+1}(x_i) = g'_{2n+1}(x_i) = f'(x_i) \quad \text{für } i = 0, \ldots, n.$$

Dann hat das Polynom $r_{2n+1}(x) = f_{2n+1}(x) - g_{2n+1}(x)$ die $n+1$ doppelten Nullstellen x_i, insgesamt also $2n+2$ Nullstellen. Andererseits ist nach Konstruktion der maximale Polynomgrad von $r_{2n+1}(x)$ $2n+1$. Daraus folgt $r_{2n+1}(x) \equiv 0$ und somit die Gleichheit $f_{2n+1}(x) = g_{2n+1}(x)$, d.h. die Lösung der Hermiteschen Interpolationsaufgabe (1.28) ist eindeutig. Da (1.28) einem quadratischen linearen Gleichungssystem der Dimension $2n+2$ entspricht, folgt daraus auch die Lösbarkeit von (1.28). ∎

Beispiel 1.7. *Für die Hermite–Interpolation der Funktion $f(x) = \sin x$ im Intervall $[0, \frac{\pi}{2}]$ sei $n = 1$. In den Stützstellen $x_0 = 0$ und $x_1 = \frac{\pi}{2}$ ist dann*

$$f(0) = 0, \quad f'(0) = 1, \quad f(\frac{\pi}{2}) = 1, \quad f'(\frac{\pi}{2}) = 0.$$

Für das Interpolationspolynom

$$f_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

ist

$$f'_3(x) = a_1 + 2a_2 x + 3a_3 x^2 \,.$$

Die Interpolationsgleichungen (1.28) *lauten also*

$$a_0 = 0, \quad a_1 = 1, \quad a_0 + a_1\frac{\pi}{2} + a_2\frac{\pi^2}{4} + a_3\frac{\pi^3}{8} = 1, \quad a_1 + 2a_2\frac{\pi}{2} + 3a_3\frac{\pi^2}{4} = 0\,.$$

Als Lösung ergibt sich

$$a_0 = 0, \quad a_1 = 1, \quad a_2 = \frac{4(3-\pi)}{\pi^2}, \quad a_3 = \frac{4(\pi-4)}{\pi^3}$$

und somit

$$f_3(x) = x + \frac{4(3-\pi)}{\pi^2}x^2 + \frac{4(\pi-4)}{\pi^3}x^3\,.$$

Analog zu Satz 1.1 kann eine Fehlerabschätzung für den Interpolationsfehler des Hermite–Interpolationspolynoms (1.27) hergeleitet werden.

Satz 1.5. *Sei $f(x)$ im Intervall $[a,b]$ $2n+2$–mal stetig differenzierbar, und sei $f_{2n+1}(x)$ das Hermitesche Interpolationspolynom vom Grad $2n+1$ mit*

$$f_{2n+1}(x_i) \;=\; f(x_i), \quad f'_{2n+1}(x_i) \;=\; f'(x_i) \quad \textit{für } i = 0,\ldots,n$$

in $n+1$ paarweise verschiedenen Stützstellen $x_i \in [a,b]$. Für den Interpolationsfehler gilt dann die Darstellung

$$f(x) - f_{2n+1}(x) \;=\; \frac{1}{(2n+2)!} f^{(2n+2)}(\xi(x)) \prod_{j=0}^{n} (x-x_j)^2 \quad \textit{für } x \in [a,b] \tag{1.29}$$

mit einer geeigneten Zwischenwertstelle $\xi(x) \in [a,b]$.

Beweis: Für die von einem reellen Parameter $\alpha \in \mathbb{R}$ abhängige Funktion

$$g_\alpha(x) \;:=\; f(x) - f_{2n+1}(x) - \alpha \prod_{j=0}^{n} (x-x_j)^2$$

gilt nach Konstruktion für alle $\alpha \in \mathbb{R}$

$$g_\alpha(x_i) \;=\; g'_\alpha(x_i) \;=\; 0 \quad \text{für } i = 0,\ldots,n.$$

Für ein beliebiges $\bar{x} \in [a,b]$ mit $\bar{x} \neq x_i$ für $i = 0,\ldots,n$ ist

$$\bar{\alpha} \;=\; \frac{f(\bar{x}) - f_{2n+1}(\bar{x})}{\prod\limits_{j=0}^{n} (x-x_j)^2}$$

wohldefiniert und es folgt

$$g_{\bar{\alpha}}(\bar{x}) \;=\; 0\,.$$

Es ist $\bar{x} \in [x_{i^*-1}, x_{i^*}]$ für genau einen Index $1 \le i^* \le n$. Somit gilt

$$g_{\bar{\alpha}}(x_{i^*-1}) = g_{\bar{\alpha}}(\bar{x}) = g_{\bar{\alpha}}(x_{i^*}) = 0.$$

Dann existieren Zwischenwertstellen $\xi_{i^*,1} \in (x_{i^*-1}, \bar{x})$ und $\xi_{i^*,2} \in (\bar{x}, x_{i^*})$ mit

$$g'_{\bar{\alpha}}(\xi_{i^*,1}) = g'_{\bar{\alpha}}(\xi_{i^*,2}) = 0\,.$$

Für die paarweise verschiedenen Argumente

$$x_{i^*-1} < \xi_{i^*,1} < \xi_{i^*,2} < x_{i^*}$$

gilt also

$$g'_{\bar{\alpha}}(x_{i^*-1}) \;=\; g'_{\bar{\alpha}}(\xi_{i^*,1}) \;=\; g'_{\bar{\alpha}}(\xi_{i^*,2}) \;=\; g'_{\bar{\alpha}}(x_{i^*}) \;=\; 0\,.$$

Damit besitzt $g''_{\bar{\alpha}}(x)$ im Intervall (x_{i^*-1}, x_{i^*}) drei voneinander verschiedene Nullstellen. Für $i = 1, \ldots, n$ sei nun $[x_{i-1}, x_i]$ ein Intervall mit $\bar{x} \notin [x_{i-1}, x_i]$. Aus

$$g_{\bar{\alpha}}(x_{i-1}) = g_{\bar{\alpha}}(x_i) = 0$$

folgt die Existenz einer Zwischenwertstelle $\xi_i \in (x_{i-1}, x_i)$ mit $g'_{\bar{\alpha}}(\xi_i) = 0$. Für

$$x_{i-1} < \xi_i < x_i$$

gilt also

$$g'_{\bar{\alpha}}(x_{i-1}) \;=\; g'_{\bar{\alpha}}(\xi_i) \;=\; g'_{\bar{\alpha}}(x_i) \;=\; 0,$$

woraus die Existenz von zwei paarweise verschiedenen Nullstellen von $g''_{\bar{\alpha}}(x)$ im Intervall (x_{i-1}, x_i) folgt.
Die Funktion $g_{\bar{\alpha}}(x)$ hat also im Intervall $[a, b]$ insgesamt

$$2(n-1) + 3 \;=\; 2n + 1$$

voneinander verschiedene Nullstellen. Die rekursive Anwendung des Satzes von Rolle ergibt nun die Existenz einer Nullstelle $\xi(\bar{x}) \in (a, b)$ von $g_{\bar{\alpha}}^{(2n+2)}(x)$ mit

$$0 \;=\; g_{\bar{\alpha}}^{(2n+2)}(\xi(\bar{x})) \;=\; f^{(2n+2)}(\xi(\bar{x})) - \bar{\alpha}(2n+2)!\,,$$

woraus

$$\bar{\alpha} \;=\; \frac{1}{(2n+2)!} f^{(2n+2)}(\xi(\bar{x}))$$

und somit die behauptete Darstellung folgt. ∎

1.5 Stückweise polynomiale Interpolation

Die bisher verwendeten Ansatzfunktionen zur Bestimmung des Interpolationspolynoms sind global, d.h. sie sind stets im gesamten Intervall $[a, b]$ auszuwerten. Die Anwendung der Fehlerabschätzung (1.15) für ein Interpolationspolynom n–ten Grades erfordert darüberhinaus die Stetigkeit der $(n+1)$–ten Ableitung der zu interpolierenden Funktion. Für viele Anwendungen ist dies aber eine zu starke Restriktion. Deshalb sollen im folgenden Approximationsmethoden betrachtet werden, die neben lokalen Ansatzfunktionen auch Fehlerabschätzungen für Funktionen mit geringerer Regularität ermöglichen.
Gegeben seien im Intervall $[a, b]$ $n+1$ voneinander verschiedene Stützstellen x_i mit

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Zum Beispiel gilt für gleichmässig verteilte Stützstellen

$$x_i = a + i\frac{b-a}{n} = a + ih \quad \text{für } i = 0, \ldots, n$$

mit der Schrittweite

$$h = \frac{b-a}{n} \to 0 \quad \text{für } n \to \infty.$$

In den Intervallen $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, wird nun die Interpolation einer gegebenen Funktion $f(x)$ durch ein lokales Interpolationspolynom $f_{i,p_i}(x)$ vom Polynomgrad p_i betrachtet. Für die Stützstellen im Intervall $[x_{i-1}, x_i]$ gelte dabei

$$x_{i-1} = x_{i,0} < x_{i,1} < \ldots < x_{i,p_i} = x_i \,.$$

Die lokalen Interpolationsgleichungen lauten also

$$f_{i,p_i}(x_{i,k}) = f(x_{i,k}) \quad \text{für } k = 0, \ldots, p_i.$$

Wegen

$$f_{i-1,p_{i-1}}(x_{i-1,p_{i-1}}) = f(x_{i-1,p_{i-1}}) = f(x_{i-1}) = f(x_{i,0}) = f_{i,p_i}(x_{i,0})$$

folgt dann die globale Stetigkeit des lokal definierten Interpolationspolynoms.
Aus der Fehlerabschätzung (1.15) ergibt sich für den lokalen Interpolationsfehler

$$\max_{x\in[x_{i-1},x_i]} |f(x) - f_{i,p_i}(x)| \le \frac{1}{(p_i+1)!} \max_{x\in[x_{i-1},x_i]} |f^{(p_i+1)}(x)| \max_{x\in[x_{i-1},x_i]} \left| \prod_{j=0}^{p_i} (x - x_{i,j} \right| . \quad (1.30)$$

Insbesondere für eine lokal lineare Interpolation mit $p_i = 1$ für alle $i = 1, \ldots, n$ sind $x_{i,0} = x_{i-1}$ und $x_{i,1} = x_i$ und es folgt die Fehlerabschätzung

$$\begin{aligned} \max_{x\in[x_{i-1},x_i]} |f(x) - f_{i,1}(x)| &\le \frac{1}{2} \max_{x\in[x_{i-1},x_i]} |f''(x)| \max_{x\in[x_{i-1},x_i]} |(x - x_{i-1})(x - x_i)| \\ &= \frac{1}{8}(x_i - x_{i-1})^2 \max_{x\in[x_{i-1},x_i]} |f''(x)| \,. \end{aligned} \quad (1.31)$$

Sind in den Stützstellen x_i die Funktionswerte der zu interpolierenden Funktion $f(x)$ durch $f_i = f(x_i)$ gegeben, so folgt für die lokal lineare Interpolierende $f_n(x) := f_{i,1}(x)$ die Darstellung

$$f_n(x) = f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}}[f_i - f_{i-1}] \quad \text{für } x \in [x_{i-1}, x_i],\ i = 1, \ldots, n. \quad (1.32)$$

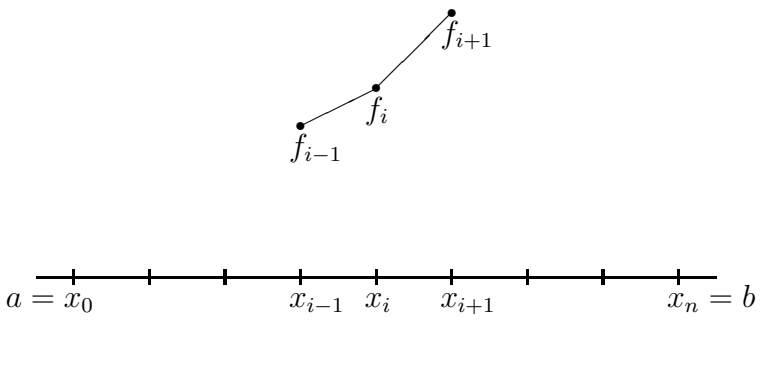


Abbildung 1.7: Stückweise lineare Interpolation.

Die punktweise Fehlerabschätzung (1.31) setzt die Beschränktheit der zweiten Ableitung $f''(x)$ der zu interpolierenden Funktion $f(x)$ voraus. Im folgenden soll deshalb eine Abschätzung des Interpolationsfehlers in der L_2–Norm

$$\int_{x_{i-1}}^{x_i} \Big[f(x) - f_n(x) \Big]^2 \, dx$$

gewonnen werden. Dabei werden wir immer wieder auf die Cauchy–Schwarz Ungleichung

$$\int_a^b \Big| f(x) g(x) \Big| \, dx \le \left(\int_a^b [f(x)]^2 \, dx \right)^{1/2} \left(\int_a^b [g(x)]^2 \, dx \right)^{1/2} \tag{1.33}$$

für quadrat–integrierbare Funktionen $f(x)$ und $g(x)$ zurückgreifen.

Lemma 1.4. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer lokal stetig differenzierbaren Funktion $f(x)$. Dann gilt*

$$\int_{x_{i-1}}^{x_i} [f(x) - f_n(x)]^2 dx \;\le\; \frac{1}{8} (x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 \, dx \, . \tag{1.34}$$

Beweis: Sei $\bar{x}_i = \frac{1}{2}(x_{i-1} + x_i)$ der Mittelpunkt des Intervalls $[x_{i-1}, x_i]$. Für $x \in (x_{i-1}, \bar{x}_i)$ folgt aus den Interpolationsgleichungen $f(x_{i-1}) = f_n(x_{i-1})$ zunächst

$$f(x) - f_n(x) \;=\; f(x) - f(x_{i-1}) + f_n(x_{i-1}) - f_n(x) \;=\; \int_{x_{i-1}}^{x} [f'(\xi) - f_n'(\xi)] \, d\xi .$$

Dann gilt, unter Verwendung der Cauchy–Schwarz Ungleichung (1.33),

$$\begin{aligned}
|f(x) - f_n(x)|^2 &= \left| \int_{x_{i-1}}^{x} [f'(\xi) - f_n'(\xi)] \, d\xi \right|^2 \\
&\le \left[\int_{x_{i-1}}^{x} 1 \cdot |f'(\xi) - f_n'(\xi)| \, d\xi \right]^2 \\
&\le \int_{x_{i-1}}^{x} 1^2 \, d\xi \int_{x_{i-1}}^{x} \Big[f'(\xi) - f_n'(\xi) \Big]^2 d\xi \\
&\le (x - x_{i-1}) \int_{x_{i-1}}^{\bar{x}_i} \Big[f'(\xi) - f_n'(\xi) \Big]^2 d\xi \quad \text{für } x \in (x_{i-1}, \bar{x}_i).
\end{aligned}$$

Integration nach $x \in (x_{i-1}, \bar{x}_i)$ liefert dann

$$\begin{aligned}
\int_{x_{i-1}}^{\bar{x}_i} \Big[f(x) - f_n(x) \Big]^2 dx &\le \int_{x_{i-1}}^{\bar{x}_i} (x - x_{i-1}) \, dx \int_{x_{i-1}}^{\bar{x}_i} \Big[f'(\xi) - f_n'(\xi) \Big]^2 d\xi \\
&= \frac{1}{2} (\bar{x}_i - x_{i-1})^2 \int_{x_{i-1}}^{\bar{x}_i} \Big[f'(\xi) - f_n'(\xi) \Big]^2 d\xi \\
&= \frac{1}{8} (x_i - x_{i-1})^2 \int_{x_{i-1}}^{\bar{x}_i} \Big[f'(\xi) - f_n'(\xi) \Big]^2 d\xi .
\end{aligned}$$

Entsprechend gilt

$$\int_{\bar{x}_i}^{x_i} \Big[f(x) - f_n(x)\Big]^2 dx \;\le\; \frac{1}{8}(x_i - x_{i-1})^2 \int_{\bar{x}_i}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi$$

und durch Addition beider Anteile folgt die Behauptung. ∎

Die Abschätzung (1.47) beschreibt eine Abschätzung des Interpolationsfehlers durch den Fehler in den Ableitungen. Dieser kann im folgenden weiter abgeschätzt werden.

Lemma 1.5. *Sei $f_n(x)$ das durch* (1.32) *definierte stückweise lineare Interpolationspolynom einer lokal zweimal stetig differenzierbaren Funktion $f(x)$. Dann gilt*

$$\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2\, dx \;\le\; \frac{1}{3}(x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} [f''(x)]^2\, dx\,. \tag{1.35}$$

Beweis: Wegen

$$\int_{x_{i-1}}^{x_i} \Big[f'(s) - f_n'(s)\Big]\, ds \;=\; f(x_i) - f(x_{i-1}) - f_n(x_i) + f_n(x_{i-1}) \;=\; 0$$

ergibt sich

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi \;&=\; \int_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi) - \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} \Big[f'(s) - f_n'(s)\Big] ds\Big]^2 d\xi \\
&=\; \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \Big[\int_{x_{i-1}}^{x_i} \Big[(f'(\xi) - f_n'(\xi)) - (f'(s) - f_n'(s))\Big] ds\Big]^2 d\xi \\
&=\; \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} \int_{s}^{\xi} [f''(t) - f_n''(t)]\, dt\, ds \right]^2 d\xi\,.
\end{aligned}$$

Für die lokal linear Interpolierende $f_n(t)$ ist $f_n''(t) = 0$ und daher ist

$$\int_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi \;=\; \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} \int_{s}^{\xi} f''(t)\, dt\, ds \right]^2 d\xi\,.$$

Durch wiederholte Anwendung der Cauchy–Schwarz Ungleichung (1.33) folgt weiterhin

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} \Big[f'(\xi) - f_n'(\xi)\Big]^2 d\xi \;&=\; \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} 1 \cdot \int_{s}^{\xi} f''(t)\, dt\, ds \right]^2 d\xi \\
&\le\; \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} 1^2 ds \int_{x_{i-1}}^{x_i} \left[\int_{s}^{\xi} f''(t) dt \right]^2 ds\, d\xi \\
&=\; \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \left[\int_{s}^{\xi} 1 \cdot f''(t) dt \right]^2 ds\, d\xi \\
&\le\; \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \left| \int_{s}^{\xi} 1^2 dt \right| \cdot \left| \int_{s}^{\xi} [f''(t)]^2 dt \right| ds\, d\xi \\
&\le\; \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} |\xi - s|\, ds\, d\xi \int_{x_{i-1}}^{x_i} [f''(t)]^2 dt\,.
\end{aligned}$$

Mit

$$\begin{aligned}\int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i}|\xi-s|\,ds\,d\xi &= 2\int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{\xi}(\xi-s)\,ds\,d\xi = \int_{x_{i-1}}^{x_i}\left[-(\xi-s)^2\right]_{x_{i-1}}^{\xi}d\xi\\ &= \int_{x_{i-1}}^{x_i}(\xi-x_{i-1})^2 d\xi \;=\; \frac{1}{3}(x_i-x_{i-1})^3\end{aligned}$$

folgt schließlich die Behautpung,

$$\int_{x_{i-1}}^{x_i}[f'(\xi)-f_n'(\xi)]^2 d\xi \;\le\; \frac{1}{3}(x_i-x_{i-1})^2\int_{x_{i-1}}^{x_i}[f''(t)]^2 dt.$$

■

Durch Verknüpfung der Fehlerabschätzungen (1.47) und (1.35) ergibt sich:

Folgerung 1.2. *Sei $f_n(x)$ das durch* (1.32) *definierte stückweise lineare Interpolationspolynom einer lokal zweimal stetig differenzierbaren Funktion $f(x)$. Dann gilt*

$$\int_{x_{i-1}}^{x_i}[f(x)-f_n(x)]^2\,dx \;\le\; \frac{1}{24}(x_i-x_{i-1})^4\int_{x_{i-1}}^{x_i}[f''(x)]^2\,dx\,. \tag{1.36}$$

Durch Summation der lokalen Fehlerabschätzungen ergibt sich nun:

Folgerung 1.3. *Sei $f_n(x)$ das durch* (1.32) *definierte stückweise lineare Interpolationspolynom einer lokal zweimal stetig differenzierbaren Funktion $f(x)$. Dann gelten die Fehlerabschätzungen*

$$\int_a^b[f(x)-f_n(x)]^2\,dx \le \frac{1}{24}\sum_{i=1}^{n}(x_i-x_{i-1})^4\int_{x_{i-1}}^{x_i}[f''(x)]^2\,dx \le \frac{1}{24}\,h^4\int_a^b[f''(x)]^2\,dx \tag{1.37}$$

und

$$\int_a^b[f'(x)-f_n'(x)]^2\,dx \le \frac{1}{3}\sum_{i=1}^{n}(x_i-x_{i-1})^2\int_{x_{i-1}}^{x_i}[f''(x)]^2\,dx \le \frac{1}{3}\,h^2\int_a^b[f''(x)]^2\,dx \tag{1.38}$$

mit der globalen Maschenweite

$$h := \max_{i=1,\ldots,n} h_i, \qquad h_i := x_i - x_{i-1}\,.$$

Die Fehlerabschätzungen (1.35) und (1.36) setzen voraus, dass die zu interpolierende Funktion $f(x)$ wenigstens zweimal stetig differenzierbar ist. Ist diese Voraussetzung nicht erfüllt, können die Fehlerabschätzungen (1.35) und (1.36) nicht angewendet werden. Für eine einmal stetig differenzierbare Funktion gilt die folgende Abschätzung:

Lemma 1.6. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer stetig differenzierbaren Funktion $f(x)$. Dann gilt*

$$\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 \, dx \;\le\; 4 \int_{x_{i-1}}^{x_i} [f'(x)]^2 \, dx \,. \tag{1.39}$$

Beweis: Mit der Dreiecksungleichung gilt zunächst

$$\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 dx \;\le\; 2 \int_{x_{i-1}}^{x_i} [f'(x)]^2 dx + 2 \int_{x_{i-1}}^{x_i} [f_n'(x)]^2 dx \,.$$

Für die linear Interpolierende $f_n(x)$ folgt, wieder unter Verwendung der Cauchy–Schwarz Ungleichung (1.33),

$$\begin{aligned} \int_{x_{i-1}}^{x_i} [f_n'(x)]^2 dx &= \int_{x_{i-1}}^{x_i} \left[\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \right]^2 dx \\ &= \frac{1}{x_i - x_{i-1}} \left[\int_{x_{i-1}}^{x_i} f'(s) \, ds \right]^2 \le \int_{x_{i-1}}^{x_i} [f'(s)]^2 \, ds. \end{aligned}$$

Daraus folgt unmittelbar die Behauptung. ∎

Für die Abschätzung des Interpolationsfehlers in L_2 ergibt sich dann:

Folgerung 1.4. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer lokal stetig differenzierbaren Funktion $f(x)$. Dann gilt*

$$\int_{x_{i-1}}^{x_i} [f(x) - f_n(x)]^2 \, dx \;\le\; \frac{1}{2}(x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} [f'(x)]^2 \, dx \,. \tag{1.40}$$

Folgerung 1.5. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer lokal zweimal stetig differenzierbaren Funktion $f(x)$. Dann gelten die Fehlerabschätzungen*

$$\int_a^b [f(x) - f_n(x)]^2 \, dx \le \frac{1}{2} \sum_{i=1}^{n} (x_i - x_{i-1})^2 \int_{x_{i-1}}^{x_i} [f'(x)]^2 \, dx \le \frac{1}{2} \, h^2 \int_a^b [f'(x)]^2 \, dx \tag{1.41}$$

und

$$\int_a^b [f'(x) - f_n'(x)]^2 \, dx \le 4 \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} [f'(x)]^2 \, dx \;=\; 4 \int_a^b [f'(x)]^2 \, dx. \tag{1.42}$$

Die Fehlerabschätzungen (1.37), (1.38), (1.41) und (1.42) können für $s = 1, 2$ und $\sigma = 0, 1$ kompakt in der folgenden Form geschrieben werden,

$$\int_a^b [f^{(\sigma)}(x) - f_n^{(\sigma)}]^2 \, dx \;\le\; c(s, \sigma) \, h^{2(s-\sigma)} \int_a^b [f^{(s)}(x)]^2 \, dx \,, \tag{1.43}$$

mit

$$c(2,0) = \frac{1}{24}, \quad c(2,1) = \frac{1}{3}, \quad c(1,0) = \frac{1}{2}, \quad c(1,1) = 4\,.$$

Mit

$$\|f\|_0 := \sqrt{\int_a^b [f(x)]^2 dx}, \quad \|f\|_1 := \sqrt{\int_a^b [f'(x)]^2 dx}, \quad \|f\|_2 := \sqrt{\int_a^b [f''(x)]^2 dx}$$

können wir die Fehlerabschätzung (1.43) auch in der Form

$$\|f - f_n\|_\sigma \le c\, h^{s-\sigma}\, \|f\|_s\,. \tag{1.44}$$

schreiben. Im folgenden zeigen wir, dass diese für die Fälle $\sigma \in (0,1)$ und $s \in (1,2)$ verallgemeinert werden kann.
Für eine gegebene Funktion $u(x)$, $x \in (a,b)$, betrachten wir die Kosinus–Reihe

$$u(x) = \sum_{k=0}^{\infty} u_k \cos k\pi \, \frac{x-a}{b-a}$$

mit den Koeffizienten

$$u_0 = \frac{1}{b-a} \int_a^b u(x)\, dx, \quad u_k = \frac{2}{b-a} \int_a^b u(x) \cos k\pi \, \frac{x-a}{b-a}\, dx, \quad k \in \mathbb{N}.$$

Dabei haben wir die Orthogonalität

$$\int_a^b \cos k\pi \, \frac{x-a}{b-a} \cos \ell\pi \, \frac{x-a}{b-a}\, dx = \begin{cases} 0 & \text{für } k \neq \ell, \\ b-a & \text{für } k = \ell = 0, \\ \dfrac{b-a}{2} & \text{für } k = \ell \neq 0, \end{cases}$$

ausgenutzt. Dann ergibt sich

$$\begin{aligned} \int_a^b [u(x)]^2\, dx &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} u_k u_\ell \int_a^b \cos k\pi \, \frac{x-a}{b-a} \cos \ell\pi \, \frac{x-a}{b-a}\, dx \\ &= (b-a)\, u_0^2 + \frac{b-a}{2} \sum_{k=1}^{\infty} u_k^2, \end{aligned}$$

und für

$$u'(x) = -\frac{1}{b-a} \sum_{k=1}^{\infty} u_k\, k\pi \, \sin k\pi \, \frac{x-a}{b-a}$$

folgt analog

$$\begin{aligned}\int_a^b [u'(x)]^2\,dx &= \frac{1}{(b-a)^2}\sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} u_k u_\ell\, k\pi\,\ell\pi \int_a^b \sin k\pi\,\frac{x-a}{b-a}\,\sin \ell\pi\,\frac{x-a}{b-a}\,dx\\ &= \frac{1}{2}\,\frac{1}{b-a}\sum_{k=1}^{\infty} u_k^2 (k\pi)^2.\end{aligned}$$

Ausgehend von

$$\|u\|_0^2 := \int_a^b [u(x)]^2\,dx = (b-a)\left[u_0^2 + \frac{1}{2}\sum_{k=1}^{\infty} u_k^2\right]$$

und

$$\|u\|_1^2 := \int_a^b [u'(x)]^2\,dx = \frac{1}{2}(b-a)\sum_{k=1}^{\infty} u_k^2\left(\frac{k\pi}{b-a}\right)^2$$

definieren wir für $\sigma \in (0,1)$

$$\|u\|_\sigma^2 := \frac{1}{2}(b-a)\sum_{k=1}^{\infty} u_k^2\left(\frac{k\pi}{b-a}\right)^{2\sigma}. \tag{1.45}$$

Unter Verwendung der Hölder–Ungleichung

$$\sum_{k=0}^{\infty} a_k b_k \le \left(\sum_{k=1}^{\infty} a_k^p\right)^{1/p}\left(\sum_{k=1}^{\infty} b_k^q\right)^{1/q}, \quad \frac{1}{p}+\frac{1}{q}=1, \quad 0 \le a_k, b_k,$$

erhalten wir

$$\begin{aligned}\|u\|_\sigma^2 &= \frac{1}{2}(b-a)\sum_{k=1}^{\infty} u_k^2\left(\frac{k\pi}{b-a}\right)^{2\sigma}\\ &= \frac{1}{2}(b-a)\sum_{k=1}^{\infty} u_k^{2-2\sigma}\left(u_k\,\frac{k\pi}{b-a}\right)^{2\sigma}\\ &\le \frac{1}{2}(b-a)\left(\sum_{k=1}^{\infty} u_k^{(2-2\sigma)p}\right)^{1/p}\left(\sum_{k=1}^{\infty}\left(u_k\,\frac{k\pi}{b-a}\right)^{2\sigma q}\right)^{1/q}.\end{aligned}$$

Insbesondere für

$$p = \frac{1}{1-\sigma}, \quad q = \frac{1}{\sigma}, \quad \frac{1}{p}+\frac{1}{q} = 1-\sigma+\sigma = 1$$

folgt

$$\|u\|_\sigma^2 \le \frac{1}{2}(b-a)\left(\sum_{k=1}^{\infty} u_k^2\right)^{1-\sigma}\left(\sum_{k=1}^{\infty} u_k^2\left(\frac{k\pi}{b-a}\right)^2\right)^{\sigma},$$

d.h.

$$\|u\|_\sigma^2 \le \|u\|_0^{2(1-\sigma)} \, \|u\|_1^{2\sigma} .$$

Aus den Fehlerabschätzungen (1.37) und (1.38) folgt somit

$$\begin{aligned}
\|f - f_n\|_\sigma^2 &\le \left(\int_a^b [f(x) - f_n(x)]^2 \, dx \right)^{1-\sigma} \left(\int_a^b [f'(x) - f_n'(x)]^2 \, dx \right)^{\sigma} \\
&\le \left(\frac{1}{24} h^4 \int_a^b [f''(x)]^2 \, dx \right)^{1-\sigma} \left(\frac{1}{3} h^2 \int_a^b [f''(x)]^2 \, dx \right)^{\sigma} \\
&= c(s) \, h^{4-2\sigma} \int_a^b [f''(x)]^2 \, dx,
\end{aligned}$$

d.h. (1.44) für $s = 2$ und $\sigma \in (0,1)$.
Die Fehlerabschätzungen (1.35), (1.36), (1.39) und (1.47) setzen die zwei– bzw. einmalige Differenzierbarkeit der zu interpolierenden Funktion $f(x)$ voraus. Im folgenden betrachten wir eine Fehlerabschätzung für eine Funktion $f(x)$ mit dazwischenliegender Regularität.

Lemma 1.7. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer stetig differenzierbaren Funktion $f(x)$ mit*

$$\int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]^2}{|x-y|^{1+2s}} \, dy \, dx < \infty , \quad s \in (0,1).$$

Dann gilt

$$\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 dx \le (x_i - x_{i-1})^{2s} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]^2}{|x-y|^{1+2s}} \, dy \, dx . \tag{1.46}$$

Beweis: Es ist

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 dx &= \int_{x_{i-1}}^{x_i} \left[f'(x) - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \right]^2 dx \\
&= \int_{x_{i-1}}^{x_i} \left[f'(x) - \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} f'(y) \, dy \right]^2 dx \\
&= \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} [f'(x) - f'(y)] \, dy \right]^2 dx .
\end{aligned}$$

Für $s \in (0,1)$ folgt mit der Cauchy–Schwarz Ungleichung (1.33)

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} [f'(x) - f_n'(x)]^2 dx &= \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]}{|x-y|^{1/2+s}} \, |x-y|^{1/2+s} \, dy \right]^2 dx \\
&\le \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]^2}{|x-y|^{1+2s}} \, dy \int_{x_{i-1}}^{x_i} |x-y|^{1+2s} \, dy \, dx \\
&\le \frac{1}{(x_i - x_{i-1})^2} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]^2}{|x-y|^{1+2s}} \, dy \int_{x_{i-1}}^{x_i} (x_i - x_{i-1})^{1+2s} \, dy \, dx \\
&= (x_i - x_{i-1})^{2s} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \frac{[f'(x) - f'(y)]^2}{|x-y|^{1+2s}} \, dy \, dx
\end{aligned}$$

die Behauptung. ∎

Für die Abschätzung des Interpolationsfehlers ergibt sich:

Folgerung 1.6. *Sei $f_n(x)$ das durch (1.32) definierte stückweise lineare Interpolationspolynom einer stetig differenzierbaren Funktion $f(x)$ mit*

$$\int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i} \frac{[f'(x)-f'(y)]^2}{|x-y|^{1+2s}}\,dy\,dx < \infty\,, \quad s \in (0,1).$$

Dann gilt

$$\int_{x_{i-1}}^{x_i} [f(x)-f_n(x)]^2 dx \;\le\; \frac{1}{8}(x_i-x_{i-1})^{2+2s} \int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i} \frac{[f'(x)-f'(y)]^2}{|x-y|^{1+2s}}\,dy\,dx\,, \tag{1.47}$$

und durch Summation folgt

$$\int_a^b [f(x)-f_n(x)]^2 dx \;\le\; \frac{1}{8}h^{2+2s} \int_a^b\int_a^b \frac{[f'(x)-f'(y)]^2}{|x-y|^{1+2s}}\,dy\,dx\,, \tag{1.48}$$

d.h. (1.44) für $\widetilde{s} := 1+s \in (1,2)$ und $\sigma = 0$. Insbesondere ist

$$\|f\|_{\widetilde{s}}^2 := \int_a^b\int_a^b \frac{[f'(x)-f'(y)]^2}{|x-y|^{1+2s}}\,dy\,dx\,. \tag{1.49}$$

Bemerkung 1.1. *Die Beschränkung von $s \in [1,2]$ in der Fehlerabschätzung (1.44) wird einerseits durch die Differenzierbarbeit von f bestimmt, andererseits durch den Polynomgrad $p = 1$ der stückweise linearen Interpolation, d.h. es gilt $s \le p+1$. Entsprechend können diese Fehlerabschätzungen auf lokale Interpolationspolynome höheren Polynomgrades p übertragen werden.*
Die Fehlerabschätzung (1.44) bleibt richtig für $\sigma \in (\frac{1}{2},1)$, wobei für die zu interpolierende Funktion die Stetigkeit vorauszusetzen ist. Dies soll hier jedoch nicht weiter betrachtet werden. Später werden vergleichbare Fehlerabschätzungen für Projektionsverfahren mit stückweise konstanten Basisfunktionen hergeleitet.

Beispiel 1.8. *Sei $f(x) = \sin x$ für $x \in [a,b] = [0, \frac{\pi}{2}]$. Dann ist*

$$\|f\|_2 = \|f''\|_0 = \left[\int_0^{\frac{\pi}{2}} (-\sin x)^2 dx\right]^{1/2} \;=\; \left[\frac{\pi}{4}\right]^{1/2} = \frac{1}{2}\sqrt{\pi}\,.$$

Für eine stückweise lineare Interpolation bezüglich der gleichmässig verteilten Stützstellen

$$x_i = ih \quad \text{für } i = 0,\ldots,n, \quad h = \frac{\pi}{2n}$$

folgt die Fehlerabschätzung

$$\|f-f_n\|_0 \;\le\; \frac{\sqrt{6}}{12}\left(\frac{\pi}{2n}\right)^2 \frac{1}{2}\sqrt{\pi} \;=\; \frac{\sqrt{6}}{96}\,\pi^{5/2}\,\frac{1}{n^2}\,.$$

In Tabelle 1.1 werden zusätzlich zur Fehlerabschätzung auch die tatsächlichen Fehler angegeben. Diese bestätigen insbesondere die quadratische Konvergenzordnung der stückweise linearen Interpolation.

	Fehlerabschätzung	Fehler $\|f - f_n\|_0$
2	0.111588	0.049236
4	0.027897	0.012434
8	0.006974	0.003116
16	0.001744	0.000787

Tabelle 1.1: Vergleich Interpolationsfehler mit Fehlerabschätzung.

Für eine globale Darstellung des durch (1.32) definierten stückweise linearen Interpolationspolynoms

$$I_n f(x) = \sum_{k=0}^{n} f(x_k)\varphi_k(x) \tag{1.50}$$

können Basisfunktionen

$$\varphi_k(x) = \begin{cases} 1 & \text{für } x = x_k, \\ 0 & \text{für } x = x_\ell \neq x_k, \\ \text{stückweise linear} & \text{sonst} \end{cases}$$

definiert werden, siehe hierzu auch Abbildung 1.8.

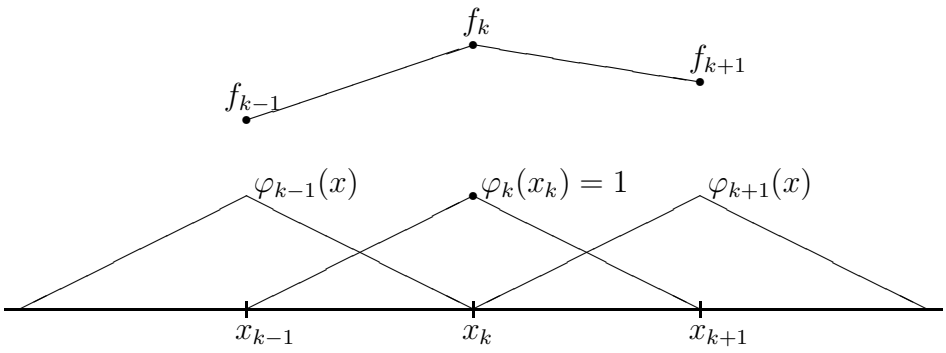


Abbildung 1.8: Ansatzfunktionen $\varphi_k(x)$ sowie $\varphi_{k\pm1}(x)$.

Für die Basisfunktionen $\varphi_k(x)$ ergibt sich daraus die funktionale Darstellung

$$\varphi_k(x) = \begin{cases} \dfrac{x - x_{k-1}}{x_k - x_{k-1}} & \text{für } x \in [x_{k-1}, x_k], \\ \dfrac{x_{k+1} - x}{x_{k+1} - x_k} & \text{für } x \in [x_k, x_{k+1}], \\ 0 & \text{sonst.} \end{cases} \tag{1.51}$$

Die Verwendung lokaler Basisfunktionen, z.B. stückweise linearer Ansatzfunktionen, ermöglicht eine einfache Auswertung des Interpolationspolynoms. Jedoch verlangt die Interpolationsaufgabe die Stetigkeit der zu approximierenden Funktion. Diese Voraussetzung kann durch die Verwendung von Projektionsmethoden vermieden werden.

1.6 Projektionsmethoden

Die Approximation einer gegebenen Funktion $f(x)$ durch ein Interpolationspolynom $f_n(x)$ erfordert zumindest die Stetigkeit der zu approximierenden Funktion $f(x)$. Diese Voraussetzung kann durch geeignete Projektionsmethoden abgeschwächt werden. Gesucht ist eine Approximation

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

mit zunächst beliebigen Ansatzfunktionen $\varphi_k(x)$, die den zugehörigen Fehler

$$\begin{aligned}
\int_a^b [f(x) - f_n(x)]^2 \, dx &= \int_a^b \left[f(x) - \sum_{k=0}^{n} a_k \varphi_k(x) \right]^2 dx \\
&= \int_a^b [f(x)]^2 \, dx - 2 \sum_{k=0}^{n} a_k \int_a^b f(x) \varphi_k(x) \, dx + \sum_{k=0}^{n} \sum_{\ell=0}^{n} a_k a_\ell \int_a^b \varphi_k(x) \varphi_\ell(x) \, dx \\
&= \int_a^b [f(x)]^2 \, dx - 2 \sum_{k=0}^{n} b_k f_k + \sum_{k=0}^{n} \sum_{\ell=0}^{n} b_k b_\ell m_{k\ell}
\end{aligned}$$

minimiert. Dabei sind

$$f_k = \int_a^b f(x) \varphi_k(x) \, dx, \quad m_{k\ell} = \int_a^b \varphi_k(x) \varphi_\ell(x) \, dx \quad \text{für } k, \ell = 0, \ldots, n.$$

Aus der notwendigen Minimierungsbedingung

$$\frac{\partial}{\partial a_j} \int_a^b \left[f(x) - \sum_{k=0}^{n} a_k \varphi_k(x) \right]^2 dx = 0 \quad \text{für alle } j = 0, \ldots, n$$

folgt, unter Ausnutzung der Symmetrie $m_{k\ell} = m_{\ell k}$,

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_j}\left[\int_a^b [f(x)]^2\,dx - 2\sum_{k=0}^n a_k f_k + \sum_{k=0}^n\sum_{\ell=0}^n a_k a_\ell m_{k\ell}\right] \\ &= -2f_j + \frac{\partial}{\partial a_j}\left[a_j^2 m_{jj} + \sum_{\ell=0,\ell\neq j}^n a_j a_\ell m_{j\ell} + \sum_{k=0,k\neq j}^n\sum_{\ell=0}^n a_k a_\ell m_{k\ell}\right] \\ &= -2f_j + 2a_j m_{jj} + \sum_{\ell=0,\ell\neq j}^n a_\ell m_{j\ell} + \sum_{k=0,k\neq j}^n a_k m_{kj} \\ &= -2f_j + 2\sum_{k=0}^n a_k m_{kj}\,. \end{aligned}$$

Dies ist gleichbedeutend mit

$$\sum_{k=0}^n a_k \int_a^b \varphi_k(x)\varphi_j(x)\,dx \;=\; \int_a^b f(x)\varphi_j(x)\,dx \quad \text{für } j = 0,\ldots,n, \tag{1.52}$$

bzw. mit dem linearen Gleichungssystem

$$M_h \underline{a} = \underline{f} \tag{1.53}$$

mit der Massematrix

$$M_h[j,k] = \int_a^b \varphi_k(x)\varphi_j(x)\,dx \tag{1.54}$$

und dem Vektor der rechten Seite,

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

für $j,k = 0,\ldots,n$. Wegen

$$M_h[j,k] \;=\; \int_a^b \varphi_k(x)\varphi_j(x)\,dx \;=\; \int_a^b \varphi_j(x)\varphi_k(x)\,dx \;=\; M_h[k,j]$$

für alle $k,j = 0,\ldots,n$ ist die Massematrix $M_h = M_h^\top$ symmetrisch, und wegen

$$\begin{aligned} (M_h\underline{a},\underline{a}) &= \sum_{j=0}^n\sum_{k=0}^n M_h[j,k]a_k a_j \;=\; \sum_{j=0}^n\sum_{k=0}^n a_k a_j \int_a^b \varphi_k(x)\varphi_j(x)\,dx \\ &= \int_a^b \sum_{k=0}^n a_k\varphi_k(x)\sum_{j=0}^n a_j\varphi_j(x)\,dx \;=\; \int_a^b \left[\sum_{k=0}^n a_k\varphi_k(x)\right]^2 dx \;>\; 0 \end{aligned}$$

für alle $\underline{a} \in \mathbb{R}^{n+1}$ mit $\|\underline{a}\|_2 > 0$ positiv definit, wenn die Basisfunktionen $\varphi_k(x)$ linear unabhängig sind. Insbesondere folgt daraus die Invertierbarkeit der Massematrix M_h, und somit die eindeutige Lösbarkeit des linearen Gleichungssystems (1.53).

Bei Verwendung der stückweise linearen Basisfunktionen (1.51) ergibt sich für die Einträge (1.54) der Massematrix

$$M_h[j,k] \;=\; \int_a^b \varphi_k(x)\varphi_j(x)\,dx \;=\; 0 \quad \text{für } j \neq k, k \pm 1.$$

Für die Nebendiagonaleinträge und $j = k \pm 1$ ist

$$\begin{aligned} M_h[k \pm 1, k] &= \int_{x_k}^{x_{k+1}} \frac{x_{k+1} - x}{x_{k+1} - x_k} \frac{x - x_k}{x_{k+1} - x_k}\,dx \\ &= \frac{1}{(x_{k+1} - x_k)^2} \int_0^{x_{k+1}-x_k} (x_{k+1} - x_k - t)\,t\,dt \;=\; \frac{1}{6}(x_{k+1} - x_k). \end{aligned}$$

Für die Hauptdiagonaleinträge und $j = k$ ist schließlich

$$\begin{aligned} M_h[k,k] &= \int_{x_{k-1}}^{x_k} \left[\frac{x - x_{k-1}}{x_k - x_{k-1}}\right]^2 dx + \int_{x_k}^{x_{k+1}} \left[\frac{x_{k+1} - x}{x_{k+1} - x_k}\right]^2 dx \\ &= \frac{1}{3}(x_k - x_{k-1}) + \frac{1}{3}(x_{k+1} - x_k)\,. \end{aligned}$$

Für eine gleichmäßige Unterteilung mit $h = x_{k+1} - x_k$ für alle $k = 1, \ldots, n$ folgt somit

$$M_h \;=\; \frac{h}{6} \begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & \ddots & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

Die Massematrix M_h ist symmetrisch und positiv definit sowie schwach besetzt, d.h. M_h besitzt $2 + 3(n-1) + 2$ Nichtnulleinträge. Damit können für die Lösung des linearen Gleichungssystems $M_h \underline{a} = \underline{f}$ effiziente Lösungsverfahren verwendet werden. Darauf soll an dieser Stelle jedoch nicht weiter eingegangen werden.
Die voneinander linear unabhängigen Basisfunktionen $\varphi_k(x)$ bilden einen linearen Raum

$$S_n \;=\; \operatorname{span}\{\varphi_k\}_{k=0}^n.$$

Die durch die Lösung des linearen Gleichungssystems (1.53) eindeutig bestimmte Approximation $f_n \in S_n$ ist also Lösung des Variationsproblems

$$\int_a^b f_n(x)\varphi_j(x)\,dx \;=\; \int_a^b f(x)\varphi_j(x)\,dx \quad \text{für } j = 0, \ldots, n,$$

bzw. von

$$\int_a^b f_n(x)g_n(x)\,dx \;=\; \int_a^b f(x)g_n(x)\,dx \quad \text{für alle } g_n \in S_n. \tag{1.55}$$

Offenbar gilt dann die Galerkin–Orthogonalität

$$\int_a^b [f(x) - f_n(x)] g_n(x)\, dx \;=\; 0 \quad \text{für alle } g_n \in S_n. \tag{1.56}$$

Die Approximation $f_n = Q_n f \in S_n$ wird als L_2–Projektion von f bezeichnet.

Lemma 1.8. *Sei $Q_n f \in S_n$ die durch die Lösung des Variationsproblems* (1.55) *eindeutig bestimmte L_2–Projektion einer gegebenen Funktion f. Dann gilt die Fehlerabschätzung*

$$\|f - Q_n f\|_0 \;\le\; \|f - g_n\|_0 \quad \textit{für alle } g_n \in S_n, \tag{1.57}$$

d.h. es gilt

$$\|f - f_n\|_0 \;=\; \min_{g_n \in S_n} \|f - g_n\|_0 \,. \tag{1.58}$$

Beweis: Für den Fehler $f - Q_n f$ folgt mit der Galerkin–Orthogonalität (1.56) und der Cauchy–Schwarz Ungleichung (1.33)

$$\begin{aligned}
\|f - Q_n f\|_0^2 \;&=\; \int_a^b [f(x) - Q_n f(x)][f(x) - Q_n f(x)]\, dx \\
&= \int_a^b [f(x) - Q_n f(x)][f(x) - g_n(x)]\, dx + \int_a^b [f(x) - Q_n f(x)][g_n(x) - f_n(x)]\, dx \\
&= \int_a^b [f(x) - Q_n f(x)][f(x) - g_n(x)]\, dx \\
&\le \|f - Q_n f\|_0 \|f - g_n\|_0
\end{aligned}$$

für alle $g_n \in S_n$. Daraus folgt unmittelbar die Fehlerabschätzung (1.57). ∎

Für das Beispiel der L_2–Projektion in den Raum der stückweise linearen Funktionen kann die Fehlerabschätzung (1.57) mit der Fehlerabschätzung (1.44) kombiniert werden.

Folgerung 1.7. *Sei $Q_n f \in S_n$ die durch die Lösung des Variationsproblems* (1.55) *eindeutig bestimmte stückweise lineare L_2–Projektion einer gegebenen Funktion f mit $\|f\|_s < \infty$ für $s \in [1, 2]$. Sei weiterhin $I_n f$ das stückweise lineare Interpolationspolynom von f. Dann gilt die Fehlerabschätzung*

$$\|f - Q_n f\|_0 \;\le\; \|f - I_n f\|_0 \;\le\; c\, h^s \, \|f\|_s \,. \tag{1.59}$$

Beispiel 1.9. *Wie in Beispiel* 1.8 *betrachten wir die stückweise lineare Approximation der Funktion $f(x) = \sin x$ für $x \in [0, \frac{\pi}{2}]$ bezüglich gleichmässig verteilten Stützstellen. Die Fehler der Interpolation $I_n f$ sowie der L_2–Projektion $Q_n f$ sind in Tabelle* 1.2 *angegeben. In beiden Fällen beobachtet man eine quadratische Konvergenz, wobei im Fall der Projektion ein um den Faktor* 2 *reduzierter Fehler auftritt.*

N	$\|f - I_n f\|_0$	$\|f - Q_n f\|_0$
2	4.92 –2	2.14 –2
4	1.24 –2	5.17 –3
8	3.12 –3	1.28 –3
16	7.87 –4	3.19 –4
32	1.95 –4	7.96 –5
64	4.87 –5	1.99 –5

Tabelle 1.2: Vergleich der L_2–Fehler von Interpolation und L_2–Projektion.

Für die Abschätzung der Ableitung des Fehlers der L_2–Projektion muss anders wie bei der Interpolation vorgegangen werden. Für eine stückweise lineare Funktion $g_n \in S_n$ zeigen wir zunächst eine lokale inverse Ungleichung.

Lemma 1.9. *Für eine stückweise lineare Funktion $g_n \in S_n$ mit*

$$g_n(x) = g_{k-1} + \frac{x - x_{k-1}}{x_k - x_{k-1}}[g_k - g_{k-1}] \quad \textit{für } x \in (x_{k-1}, x_k), k = 1, \ldots, n$$

gilt die inverse Ungleichung

$$\int_{x_{k-1}}^{x_k} [g_n'(x)]^2 \, dx \leq 12 \, (x_k - x_{k-1})^{-2} \int_{x_{k-1}}^{x_k} [g_n(x)]^2 \, dx \, . \tag{1.60}$$

Beweis: Die Behauptung folgt unmittelbar aus

$$\int_{x_{k-1}}^{x_k} [g_n'(x)]^2 \, dx = \frac{1}{x_k - x_{k-1}} (g_k - g_{k-1})^2 \leq \frac{2}{x_k - x_{k-1}} \Big[g_k^2 + g_{k-1}^2 \Big]$$

und

$$\begin{aligned} \int_{x_{k-1}}^{x_k} [g_n(x)]^2 \, dx &= \frac{1}{6} \, (x_k - x_{k-1}) \left(\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} g_k \\ g_{k-1} \end{pmatrix}, \begin{pmatrix} g_k \\ g_{k-1} \end{pmatrix} \right) \\ &\geq \frac{1}{6} \, (x_k - x_{k-1}) \Big[g_k^2 + g_{k-1}^2 \Big]. \end{aligned}$$

■

Für eine global gleichmässige Unterteilung mit $h = x_k - x_{k-1}$ für alle $k = 1, \ldots, n$ folgt aus der lokalen inversen Ungleichung unmittelbar die globale inverse Ungleichung

$$\int_a^b [g_n'(x)]^2 dx \leq 12 \, h^{-2} \int_a^b [g_n(x)]^2 \, dx \quad \text{für alle } g_n \in S_n. \tag{1.61}$$

Lemma 1.10. *Sei $Q_n f \in S_n$ die durch die Lösung des Variationsproblems* (1.55) *eindeutig bestimmte lineare L_2–Projektion einer gegebenen Funktion f mit $\|f\|_s < \infty$ für $s \in [1,2]$. Für eine global gleichmässige Unterteilung folgt dann die Fehlerabschätzung*

$$\|f - Q_n f\|_1 \leq c\, h^{s-1}\, \|f\|_s\,. \tag{1.62}$$

Beweis: Mit der Dreiecksungleichung ist zunächst

$$\|f - Q_n f\|_1 \leq \|f - I_n f\|_1 + \|I_n f - Q_n f\|_1\,,$$

wobei $I_n f$ die stückweise linear Interpolierende von f bezeichnet. Mit der globalen inversen Ungleichung (1.61) für $I_n f - Q_n f \in S_n$ folgt dann

$$\|f - Q_n f\|_1 \leq \|f - I_n f\|_1 + \sqrt{12}\, h^{-1} \|I_n f - Q_n f\|_0\,.$$

Nochmalige Verwendung der Dreiecksungleichung ergibt

$$\|f - Q_n f\|_1 \leq \|f - I_n f\|_1 + \sqrt{12}\, h^{-1} \|f - I_n f\|_0 + \sqrt{12}\, h^{-1} \|f - Q_n f\|_0,$$

und die Behauptung folgt aus der Fehlerabschätzung (1.44) für $\sigma = 0$ und $\sigma = 1$, bzw. aus der Fehlerabschätzung (1.59). ∎

Die Anwendung der Fehlerabschätzung (1.59) setzt die Beschränktheit von $\|f\|_\sigma < \infty$ für $\sigma \in [1,2]$ voraus. Am Beispiel der L_2–Projektion mit stückweise konstanten Ansatzfunktionen

$$\psi_k(x) \;=\; \begin{cases} 1 & \text{für } x \in (x_{k-1}, x_k), \\ 0 & \text{sonst} \end{cases} \tag{1.63}$$

für $k = 1, \ldots, n$ sollen nun Approximationen

$$Q_n f(x) = \sum_{k=1}^{n} a_k \psi_k(x)$$

untersucht werden, wenn die zu approximierende Funktion f eine geringere Differenzierbarkeit aufweist. Die L_2–Projektion $Q_n f$ ergibt sich, vergleiche (1.52), als eindeutige Lösung des Variationsproblems

$$\sum_{k=1}^{n} a_k \int_a^b \psi_k(x)\psi_j(x)\,dx \;=\; \int_a^b f(x)\psi_j(x)\,dx \quad \text{für } j = 1, \ldots, n. \tag{1.64}$$

Wegen

$$\int_a^b \psi_k(x)\psi_j(x)\,dx \;=\; \begin{cases} x_k - x_{k-1} & \text{für } j = k, \\ 0 & \text{sonst} \end{cases}$$

folgt

$$a_k \;=\; \frac{1}{x_k - x_{k-1}} \int_a^b f(x)\psi_k(x)\,dx \;=\; \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} f(x)\,dx.$$

Lemma 1.11. *Sei $Q_n f$ die durch die Lösung des Variationsproblems* (1.64) *eindeutig bestimmte stückweise konstante L_2–Projektion einer gegebenen, lokal stetig differenzierbaren Funktion $f(x)$. Dann gilt die Fehlerabschätzung*

$$\int_{x_{k-1}}^{x_k} [f(x) - Q_n f(x)]^2 \, dx \;\le\; \frac{1}{3}(x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} [f'(x)]^2 \, dx \,. \tag{1.65}$$

Beweis: Für $x \in (x_{k-1}, x_k)$ ist $Q_n f(x) = a_k \psi_k(x) = a_k$ und somit ist

$$f(x) - Q_n f(x) \;=\; \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} [f(x) - f(y)] \, dy \;=\; \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \int_y^x f'(s) \, ds \, dy.$$

Dann gilt, unter Verwendung der Cauchy–Schwarz Ungleichung (1.33),

$$\begin{aligned}
[f(x) - Q_n f(x)]^2 &= \frac{1}{(x_k - x_{k-1})^2} \left[\int_{x_{k-1}}^{x_k} 1 \cdot \int_y^x f'(s) \, ds \, dy \right]^2 \\
&\le \frac{1}{(x_k - x_{k-1})^2} \int_{x_{k-1}}^{x_k} 1^2 \, ds \int_{x_{k-1}}^{x_k} \left[\int_y^x f'(s) \, ds \right]^2 dy \\
&= \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \left[\int_y^x f'(s) \, ds \right]^2 dy.
\end{aligned}$$

Durch erneute Anwendung der Cauchy–Schwarz Ungleichung (1.33) folgt

$$\begin{aligned}
[f(x) - Q_n f(x)]^2 &= \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \left[\int_y^x 1 \cdot f'(s) \, ds \right]^2 dy \\
&\le \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \left| \int_y^x 1^2 \, ds \right| \cdot \left| \int_y^x [f'(s)]^2 \, ds \right| dy \\
&\le \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} |x - y| \, dy \int_{x_{k-1}}^{x_k} [f'(s)]^2 \, ds \,.
\end{aligned}$$

Integration bezüglich $x \in (x_{k-1}, x_k)$ ergibt

$$\int_{x_{k-1}}^{x_k} [f(x) - Q_n f(x)]^2 \, dx \;\le\; \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \int_{x_{k-1}}^{x_k} |x - y| \, dy \, dx \int_{x_{k-1}}^{x_k} [f'(s)]^2 \, ds \,.$$

Mit

$$\int_{x_{k-1}}^{x_k} \int_{x_{k-1}}^{x_k} |x - y| \, dy \, dx \;=\; \frac{1}{3}(x_k - x_{k-1})^3$$

folgt schließlich die Behauptung. ∎

Summation der lokalen Fehlerabschätzungen (1.65) ergibt die globale Fehlerabschätzung

$$\int_a^b [f(x) - Q_n f(x)]^2 \, dx \;\le\; \frac{1}{3} \sum_{k=1}^{n} (x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} [f'(x)]^2 \, dx \,. \tag{1.66}$$

Die Voraussetzung der Quadratintegrierbarkeit der ersten Ableitung $f'(x)$ der zu approximierenden Funktion $f(x)$ gewährleistet also ein lineares Konvergenzverhalten der stückweise konstanten L_2–Projektion $Q_n f$ für den Fehler $f - Q_n f$ in der L_2–Norm. Ausgangspunkt für eine noch schwächere Fehlerabschätzung ist für $s \in (0,1)$

$$\begin{aligned} f(x) - Q_n f(x) &= \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} [f(x) - f(y)]\, dy \\ &= \frac{1}{x_k - x_{k-1}} \int_{x_{k-1}}^{x_k} \frac{f(x) - f(y)}{|x-y|^{\frac{1}{2}+s}} \, |x-y|^{\frac{1}{2}+s}\, dy. \end{aligned}$$

Mit der Cauchy–Schwarz Ungleichung (1.33) folgt daraus

$$\begin{aligned} [f(x) - Q_n f(x)]^2 &= \frac{1}{(x_k - x_{k-1})^2} \left[\int_{x_{k-1}}^{x_k} \frac{f(x) - f(y)}{|x-y|^{\frac{1}{2}+s}} \, |x-y|^{\frac{1}{2}+s}\, dy \right]^2 \\ &\le \frac{1}{(x_k - x_{k-1})^2} \int_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}}\, dy \int_{x_{k-1}}^{x_k} |x-y|^{1+2s}\, dy \\ &\le (x_k - x_{k-1})^{2s} \int_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}}\, dy\,. \end{aligned}$$

Integration bezüglich $x \in (x_{k-1}, x_k)$ liefert jetzt die lokale Fehlerabschätzung

$$\int_{x_{k-1}}^{x_k} [f(x) - Q_n f(x)]^2\, dx \;\le\; (x_k - x_{k-1})^{2s} \int_{x_{k-1}}^{x_k} \int_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}}\, dy\, dx \tag{1.67}$$

und durch Summation folgt die globale Fehlerabschätzung

$$\int_a^b [f(x) - Q_n f(x)]^2\, dx \;\le\; \sum_{k=1}^{n} (x_k - x_{k-1})^{2s} \int_{x_{k-1}}^{x_k} \int_{x_{k-1}}^{x_k} \frac{[f(x) - f(y)]^2}{|x-y|^{1+2s}}\, dy\, dx\,. \tag{1.68}$$

Aus der Galerkin–Orthogonalität (1.56) folgt für $g_n = f_n = Q_n f$ unter Verwendung der Cauchy–Schwarz Ungleichung (1.33)

$$\begin{aligned} \|f - Q_n f\|_0^2 &= \int_a^b [f(x) - Q_n f(x)]\, [f(x) - Q_n f(x)]\, dx \\ &= \int_a^b [f(x) - Q_n f(x)]\, f(x)\, dx \\ &\le \left(\int_a^b [f(x) - Q_n f(x)]^2\, dx \right)^{1/2} \left(\int_a^b [f(x)]^2\, dx \right)^{1/2} \end{aligned}$$

die triviale Fehlerabschätzung

$$\|f - Q_n f\|_0 \;\le\; \|f\|_0\,. \tag{1.69}$$

Definieren wir

$$\|f\|_s := \left(\int_a^b \int_a^b \frac{[f(x)-f(y)]^2}{|x-y|^{1+2s}}\, dy\, dx \right)^{1/2} \text{ für } s \in (0,1), \tag{1.70}$$

so können die obigen Fehlerabschätzungen wie folgt zusammengefaßt werden. Man beachte, daß durch (1.45) und (1.70) zueinander äquivalente Normen definiert werden.

Satz 1.6. *Sei $Q_n f$ die stückweise konstante L_2–Projektion einer gegebenen Funktion f mit $\|f\|_s < \infty$ für $s \in [0,1]$. Dann gilt die Fehlerabschätzung*

$$\|f - Q_n f\|_0 \le c\, h^s\, \|f\|_s\,. \tag{1.71}$$

Bemerkung 1.2. *Die Fehlerabschätzung* (1.71) *für stückweise konstante Ansatzfunktionen ($p=0$) entspricht formal der Fehlerabschätzung* (1.59) *für stückweise lineare Ansatzfunktionen ($p=1$) und für $s \in [1,2]$. Tatsächlich bleibt* (1.59) *richtig für $s \in [0,2]$, wobei der Fall $s=0$ wie in* (1.69) *folgt. Die Fehlerabschätzungen für $s \in (0,1)$ folgen dann durch Anwendung des sogenannten Interpolationssatzes aus den Fehlerabschätzungen für $s=0$ und $s=1$. Für L_2–Projektionen $Q_n f$ mit Basisfunktionen mit lokalem Polynomgrad p gilt die Fehlerabschätzung* (1.71) *für alle $s \in [0, p+1]$ unter der Voraussetzung $\|f\|_s < \infty$, wobei $\|f\|_s$ in Erweiterung zu* (1.49) *und* (1.70) *entsprechend definiert ist.*

Am Beispiel der stückweise konstanten Basisfunktionen $\psi_k(x)$ soll abschliessend auf einen Vergleich von Interpolation $I_n f$ und L_2–Projektion $Q_n f$ eingegangen werden. Bezeichnet $\widehat{x}_k := \frac{1}{2}(x_{k-1} + x_k)$ den Mittelpunkt von (x_{k-1}, x_k), ist ist die stückweise konstante Interpolation einer gegebenen Funktion $f(x)$ gegeben durch

$$I_n f(x) = \sum_{k=1}^{n} f(\widehat{x}_k)\psi_k(x)\,. \tag{1.72}$$

Lemma 1.12. *Sei $I_n f$ die durch* (1.72) *erklärte stückweise konstante Interpolierende einer gegebenen, lokal stetig differenzierbaren Funktion $f(x)$. Dann gilt die Fehlerabschätzung*

$$\int_{x_{k-1}}^{x_k} [f(x) - I_n f(x)]^2\, dx \;\le\; \frac{1}{4}(x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} [f'(x)]^2\, dx\,. \tag{1.73}$$

Beweis: Für $x \in (x_{k-1}, x_k)$ ist $I_n f(x) = f(\widehat{x}_k)$ und somit

$$f(x) - I_n f(x) = f(x) - f(\widehat{x}_k) = \int_{\widehat{x}_k}^{x} f'(s)\, ds\,.$$

Mit der Cauchy–Schwarz Ungleichung (1.33) folgt dann

$$\begin{aligned}
[f(x) - I_n f(x)]^2 &= \left[\int_{\widehat{x}_k}^{x} 1 \cdot f'(s)\, ds \right]^2 \\
&\le |x - \widehat{x}_k| \left| \int_{\widehat{x}_k}^{x} [f'(s)]^2\, ds \right| \\
&\le |x - \widehat{x}_k| \int_{x_{k-1}}^{x_k} [f'(s)]^2\, ds\,.
\end{aligned}$$

Integration nach x liefert dann

$$\begin{aligned}\int_{x_{k-1}}^{x_k} [f(x) - I_n f(x)]^2\, dx &\leq \int_{x_{k-1}}^{x_k} |x - \widehat{x}_k|\, dx \int_{x_{k-1}}^{x_k} [f'(s)]^2\, ds \\ &= \frac{1}{4}(x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} [f'(s)]^2\, ds.\end{aligned}$$

■

Summation der lokalen Fehlerabschätzungen (1.73) liefert die globale Abschätzung

$$\int_a^b [f(x) - I_n f(x)]^2\, dx \leq \frac{1}{4}\sum_{k=1}^{n}(x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} [f'(x)]^2\, dx\,. \tag{1.74}$$

Für einen Vergleich von Interpolation und L_2–Projektion beschränken wir uns auf das Intervall $[a, b] = [0, 1]$ mit einer gleichmässigen Unterteilung in Stützstellen $x_k = kh$, $k = 0, \ldots, n$, mit einer Schrittweite $h = 1/n$.

Satz 1.7. *Sei $f(x)$ eine lokal zweimal stetig differenzierbare Funktion. Dann stimmen die L_2–Fehler der stückweise konstanten Interpolation $I_n f$ und der stückweise konstanten L_2–Projektion $Q_n f$ bis auf Terme höherer Ordnung überein, d.h. es gilt*

$$\|f - I_n f\|_0 - \frac{1}{\sqrt{96}}\, h^2\, \|f\|_2 \leq \|f - Q_n f\|_0 \leq \|f - I_n f\|_0\,. \tag{1.75}$$

Der Beweis von Satz 1.7 beruht auf einer Anwendung der Taylorschen Formel in der folgenden Form.

Satz 1.8 (Taylorsche Formel)**.** *Sei $f(x)$ in einer Umgebung von x_0 $(n+1)$–mal stetig differenzierbar. Dann gilt*

$$f(x) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{1}{n!}\int_{x_0}^{x} (x - s)^n f^{(n+1)}(s)\, ds\,. \tag{1.76}$$

Beweis: Für eine stetig differenzierbare Funktion ist zunächst

$$f(x) - f(x_0) = \int_{x_0}^{x} f'(s)\, ds\,.$$

Wird hinreichende Differenzierbarkeit vorausgesetzt, so folgt mit partieller Integration

$$\begin{aligned}f(x) - f(x_0) &= \int_{x_0}^{x} f'(s)\, ds \\ &= (s - x)f'(s)\Big|_{x_0}^{x} - \int_{x_0}^{x} (s - x) f''(s)\, ds \\ &= (x - x_0)f'(x_0) + \int_{x_0}^{x} (x - s) f''(s)\, ds,\end{aligned}$$

d.h.

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \int_{x_0}^{x} (x - s)f''(s)\, ds.$$

Für $n = 1$ gilt also die Induktionsvoraussetzung

$$f(x) = \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{1}{n!} \int_{x_0}^{x} (x - s)^n f^{(n+1)}(s)\, ds\,.$$

Nochmalige partielle Integration ergibt

$$\begin{aligned} f(x) &= \sum_{k=0}^{n} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) - \frac{1}{(n+1)!}(x - s)^{n+1} f^{(n+1)}(s)\Big|_{x_0}^{x} \\ &\qquad + \frac{1}{(n+1)!} \int_{x_0}^{x} (x - s)^{n+1} f^{(n+2)}(s)\, ds \\ &= \sum_{k=0}^{n+1} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{1}{(n+1)!} \int_{x_0}^{x} (x - s)^{n+1} f^{(n+2)}(s)\, ds, \end{aligned}$$

d.h. die Induktionsbehauptung für $n + 1$. ∎

Beweis von Satz 1.7: Mit den Stützstellen $x_k = kh$ für $k = 0, \ldots, n$ sind die stückweise konstante L_2–Projektion $Q_n f$ und die stückweise konstante Interpolierende $I_n f$ gegeben durch, für $\widehat{x}_k = \frac{1}{2}(x_{k-1} + x_k)$,

$$Q_n f(x) = \sum_{k=1}^{n} a_k \psi_k(x), \quad a_k = \frac{1}{h} \int_{x_{k-1}}^{x_k} f(x)\, dx, \quad I_n f(x) = \sum_{k=1}^{n} b_k \psi_k(x), \quad b_k = f(\widehat{x}_k).$$

Die L_2–Projektion $Q_h f$ ist definiert als Lösung eines Minimierungsproblems, mit (1.57) für $g_n = I_n f$ gilt daher

$$\|f - Q_n f\|_0 \le \|f - I_n f\|_0.$$

und somit die obere Abschätzung. Andererseits gilt mit der Dreiecksungleichung

$$\|f - I_n f\|_0 \le \|f - Q_n f\|_0 + \|Q_n f - I_n f\|_0$$

und für den zweiten Summanden gilt

$$\|Q_n f - I_n f\|_0^2 = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} [Q_n f(x) - I_n f(x)]^2\, dx = h \sum_{k=1}^{n} (a_k - b_k)^2\,.$$

Durch Anwendung der Taylorschen Formel für eine lokal zweimal stetig differenzierbare Funktion folgt für die Differenz der Zerlegungskoeffizienten

$$\begin{aligned} a_k - b_k &= \frac{1}{h} \int_{x_{k-1}}^{x_k} f(x)\, dx - f(\widehat{x}_k) \;=\; \frac{1}{h} \int_{x_{k-1}}^{x_k} [f(x) - f(\widehat{x}_k)]\, dx \\ &= \frac{1}{h} \int_{x_{k-1}}^{x_k} \left[(x - \widehat{x}_k) f'(\widehat{x}_k) + \int_{\widehat{x}_k}^{x} (x - s) f''(s)\, ds \right] dx \\ &= \frac{1}{h} \int_{x_{k-1}}^{x_k} \int_{\widehat{x}_k}^{x} (x - s) f''(s)\, ds\, dx\,. \end{aligned}$$

Somit ergibt sich, unter Verwendung der Cauchy–Schwarz Ungleichung (1.33),

$$\begin{aligned}
(a_k - b_k)^2 &= \frac{1}{h^2}\left(\int_{x_{k-1}}^{x_k}\int_{\widehat{x}_k}^{x}(x-s)f''(s)\,ds\,dx\right)^2 \\
&\le \frac{1}{h^2}\int_{x_{k-1}}^{x_k}1^2\,dx\int_{x_{k-1}}^{x_k}\left(\int_{\widehat{x}_k}^{x}(x-s)f''(s)\,ds\right)^2 dx \\
&= \frac{1}{h}\int_{x_{k-1}}^{x_k}\left(\int_{\widehat{x}_k}^{x}(x-s)f''(s)\,ds\right)^2 dx \\
&\le \frac{1}{h}\int_{x_{k-1}}^{x_k}\left|\int_{\widehat{x}_k}^{x}(x-s)^2\,ds\right|\,\left|\int_{\widehat{x}_k}^{x}[f''(s)]^2\,ds\right| dx \\
&\le \frac{1}{h}\int_{x_{k-1}}^{x_k}\left|\int_{\widehat{x}_k}^{x}(x-s)^2\,ds\right| dx\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds.
\end{aligned}$$

Durch Berechnung des verbeibenden Integrals folgt

$$\begin{aligned}
(a_k - b_k)^2 &\le \frac{1}{h}\int_{x_{k-1}}^{x_k}\left|\int_{\widehat{x}_k}^{x}(x-s)^2\,ds\right| dx\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds \\
&= \frac{1}{h}\int_{x_{k-1}}^{x_k}\left|-\frac{1}{3}(x-s)^3\Big|_{\widehat{x}_k}^{x}\right| dx\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds \\
&= \frac{1}{3}\frac{1}{h}\int_{x_{k-1}}^{x_k}|x-\widehat{x}_k|^3\,dx\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds \\
&= \frac{2}{3}\frac{1}{h}\int_{\widehat{x}_k}^{x_k}(x-\widehat{x}_k)^3\,dx\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds \\
&= \frac{1}{96}\,h^3\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds\,.
\end{aligned}$$

Damit gilt

$$\|Q_nf - I_nf\|_0^2 \le \frac{1}{96}h^4\sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}[f''(s)]^2\,ds = \frac{1}{96}\,h^4\int_0^1[f''(x)]^2\,dx$$

und somit

$$\|f - I_nf\|_0 \le \|f - Q_nf\|_0 + \frac{1}{\sqrt{96}}\,h^2\,\|f''\|_0\,,$$

woraus unmittelbar die Behauptung folgt. ∎

Beispiel 1.10. *Gegeben sei die Funktion $f(x) = x^2$ für $x \in [0,1]$. Für die Schrittweite $h = 1/n$ ergeben sich die Stützstellen $x_k = kh$, $k = 0,\ldots,n$, und die Elementmittelpunkte*

$$\widehat{x}_k = \frac{1}{2}(x_{k-1} + x_k) = \frac{1}{2}(2k-1)h \quad \textit{für } k = 1,\ldots,n.$$

Für die Zerlegungskoeffizienten b_k der stückweise konstanten Interpolierenden $I_n f$ ergibt sich

$$b_k = f(\hat{x}_k) = \hat{x}_k^2 = \frac{1}{4}h^2(2k-1)^2$$

während für die Zerlegungskoeffizienten a_k der L_2–Projektion $Q_n f$

$$a_k = \frac{1}{h}\int_{x_{k-1}}^{x_k} x^2\,dx = \frac{1}{3}\frac{1}{h}[x_k^3 - x_{k-1}^3] = \frac{h^2}{3}[k^3-(k-1)^3] = h^2\Big[k^2-k+\frac{1}{3}\Big]$$

folgt.
Für den Fehler der stückweise konstanten Interpolierenden $I_n f$ ergibt sich

$$\begin{aligned}
\int_0^1 [f(x)-I_h f(x)]^2\,dx &= \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}\left[x^2-\frac{1}{4}h^2(2k-1)^2\right]^2 dx \\
&= \frac{1}{3}h^5\sum_{k=1}^{n}\Big[k^2-k+\frac{23}{80}\Big] \\
&= \frac{1}{3}h^5\left[\frac{1}{6}n(n+1)(2n+1)-\frac{1}{2}n(n+1)+\frac{23}{80}n\right] \\
&= \frac{1}{9}h^2-\frac{11}{720}h^4
\end{aligned}$$

und im Fall der L_2–Projektion $Q_n f$ ist

$$\begin{aligned}
\int_0^1 [f(x)-Q_h f(x)]^2\,dx &= \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k}\left[x^2-h^2\Big(k^2-k+\frac{1}{3}\Big)\right]^2 dx \\
&= \frac{1}{45}h^5\sum_{k=1}^{n}\big[15k^2-15k+4\big] \\
&= \frac{1}{45}h^5\left[\frac{15}{6}n(n+1)(2n+1)-\frac{15}{2}n(n+1)+4n\right] \\
&= \frac{1}{9}h^2-\frac{1}{45}h^4\,.
\end{aligned}$$

Bis auf Terme vierter Ordnung stimmen also Interpolations– und Projektionsfehler überein.

Kapitel 2

Numerische Integration

Für eine im Intervall $[a, b]$ gegebene Funktion $f(x)$ ist das bestimmte Integral

$$I \;=\; \int_a^b f(x)\,dx \tag{2.1}$$

durch eine geeignete Näherungsformel

$$I_n \;=\; \sum_{k=0}^{n} f(x_k)\,\omega_k \tag{2.2}$$

mit paarweise verschiedenen Stützstellen x_k und Integrationsgewichten ω_k zu berechnen. Ein numerisches Integrationsverfahren (2.2) heißt von der Ordnung p, falls p die größte ganze Zahl ist, für die das Verfahren alle Polynome kleineren Grades als p exakt integriert,

$$\int_a^b q_m(x)\,dx \;=\; \sum_{k=0}^{n} q_m(x_k)\,\omega_k \quad \text{für alle } q_m(x) \;=\; \sum_{j=0}^{m} a_j x^j, \quad m < p.$$

Eine erste Idee für die Herleitung numerischer Integrationsformeln besteht im Ersetzen der Funktion $f(x)$ durch das Interpolationspolynom $f_n \in \Pi_n$ mit

$$f_n(x_i) = f(x_i) \quad \text{für } i = 0, \ldots, n.$$

Bei Verwendung von Lagrange–Polynomen (1.11) lautet das Interpolationspolynom

$$f_n(x) \;=\; \sum_{k=0}^{n} f(x_k) L_k^n(x), \quad L_k^n(x) \;=\; \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j},$$

und für die Integrationsformel (2.2) ergibt sich

$$I_n \;=\; \int_a^b f_n(x)\,dx \;=\; \sum_{k=0}^{n} f(x_k) \int_a^b L_k^n(x)\,dx \;=\; \sum_{k=0}^{n} f(x_k)\,\omega_k \tag{2.3}$$

mit den Integrationsgewichten

$$\omega_k = \int_a^b L_k^n(x)\,dx = \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j}\,dx \quad \text{für } k = 0, \ldots, n. \tag{2.4}$$

Aus der Darstellung (1.14) des Interpolationsfehlers

$$f(x) - f_n(x) \;=\; \frac{1}{(n+1)!}\, f^{(n+1)}(\xi(x)) \prod_{j=0}^{n} (x - x_j)$$

mit einer geeigneten Zwischenwertstelle $\xi(x) \in (a, b)$ folgt für den Fehler der numerischen Integrationsformel (2.3)

$$I - I_n \;=\; \int_a^b [f(x) - f_n(x)]\,dx \;=\; \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x)) \prod_{j=0}^{n} (x - x_j)\,dx. \tag{2.5}$$

Die Integrationsformel (2.3) ist von der Ordnung $n + 1$, d.h. Polynome $f(x) = f_m(x)$ mit dem Polynomgrad $m \leq n$ werden exakt integriert. Ist die Integrationsformel insbesondere exakt für konstante Funktionen, dann folgt für $f(x) = 1$

$$I \;=\; \int_a^b dx \;=\; b - a \;=\; I_n \;=\; \sum_{k=0}^{n} \omega_k$$

und somit

$$\frac{1}{b-a} \sum_{k=0}^{n} \omega_k \;=\; 1.$$

Für eine stabile numerische Auswertung der numerischen Integrationsformel (2.3) ist weiterhin die Positivität der Integrationsgewichte, $\omega_k > 0$, zu fordern.
Die Integrationsformel (2.3) und die Fehlerabschätzung (2.5) gelten für eine beliebige Wahl der paarweise verschiedenen Stützstellen x_k. Im folgenden betrachten wir zunächst im Intervall $[a, b]$ gleichmässig verteilte Stützstellen.

2.1 Newton–Cotes Integrationsformeln

Für eine äquidistante Verteilung der Stützstellen,

$$x_k \;=\; a + k\,\frac{b-a}{n} \;=\; a + kh \quad \text{für } k = 0, \ldots, n, \quad h = \frac{b-a}{n},$$

ergibt sich für die Berechnung der Integrationsgewichte (2.4) für $k = 0, \ldots, n$

$$\omega_k \;=\; \int_a^b L_k^n(x)\,dx \;=\; \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x - x_j}{x_k - x_j}\,dx \;=\; \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x - (a + jh)}{(k - j)h}\,dx\,.$$

Mit der Substitution

$$x = a + th \quad \text{für } t \in [0, n], \quad dx = h\,dt$$

folgt

$$\omega_k \;=\; h \int_0^n \prod_{j=0, j\neq k}^{n} \frac{t-j}{k-j}\,dt \;=\; \frac{b-a}{n}\,\widetilde{\omega}_k$$

mit

$$\widetilde{\omega}_k \;=\; \int_0^n \prod_{j=0, j\neq k}^{n} \frac{t-j}{k-j}\,dt \quad \text{für } k = 0, \ldots, n.$$

Die resultierenden numerischen Integrationsformeln sind die Newton–Cotes–Formeln

$$I_n \;=\; \frac{b-a}{n} \sum_{k=0}^{n} f(x_k)\,\widetilde{\omega}_k. \tag{2.6}$$

Beispiel 2.1. *Für $n = 1$ sind die Stützstellen durch*

$$x_0 = a, \quad x_1 = b$$

gegeben und für die Integrationsgewichte ergibt sich

$$\begin{aligned} \widetilde{\omega}_0 \;&=\; \int_0^1 \frac{t-1}{0-1}\,dt \;=\; \int_0^1 (1-t)\,dt \;=\; \frac{1}{2}, \\ \widetilde{\omega}_1 \;&=\; \int_0^1 \frac{t-0}{1-0}\,dt \;=\; \int_0^1 t\,dt \;=\; \frac{1}{2}. \end{aligned}$$

Damit ist

$$I_1 \;=\; (b-a)\left[\frac{1}{2}f(a) + \frac{1}{2}f(b)\right] \;=\; \frac{b-a}{2}\,[f(a) + f(b)] \tag{2.7}$$

die Trapezregel. Für den Fehler (2.5) *ergibt sich*

$$I - I_1 \;=\; \frac{1}{2}\int_a^b f''(\xi(x))(x-a)(x-b)\,dx.$$

Die Substitution

$$s(x) \;=\; \int (x-a)(x-b)\,dx \;=\; \frac{1}{3}x^3 - \frac{1}{2}(a+b)x^2 + abx$$

ergibt eine für $x \in (a, b)$ streng monoton fallende Funktion, für die die Umkehrfunktion $x = x(s)$ existiert. Durch Anwendung des Mittelwertsatzes der Integralrechnung folgt dann

$$I - I_1 \;=\; \frac{1}{2}\int_{s(a)}^{s(b)} f''(\xi(x(s)))\,ds \;=\; \frac{1}{2}[s(b) - s(a)]\,f''(\xi(x(\bar{s}))) \;=\; -\frac{1}{12}\,f''(\eta)\,(b-a)^3$$

mit einer Zwischenwertstelle

$$\eta = f''(\xi(x(s))) \in (a,b).$$

Insbesondere gilt

$$\int_a^b f(x)\,dx \;=\; \frac{b-a}{2}\,[f(a)+f(b)] - \frac{1}{12}f''(\eta)(b-a)^3\,. \tag{2.8}$$

Damit ist die Trapezregel ein Verfahren zweiter Ordnung, d.h. lineare Funktionen werden exakt integriert.

Beispiel 2.2. *Wird als Stützstelle nur der Mittelpunkt*

$$x_0 \;=\; \frac{a+b}{2}$$

betrachtet, so ergibt eine Taylor–Entwicklung (1.76) *für* $x \in (a,b)$

$$f(x) \;=\; f(x_0) + f'(x_0)\,(x-x_0) + \frac{1}{2}f''(\xi(x))\,(x-x_0)^2$$

mit einer geeigneten Zwischenwertstelle $\xi(x) \in (a,b)$*. Dann folgt*

$$\begin{aligned} I \;=\; \int_a^b f(x)\,dx \;&=\; \int_a^b \left[f(x_0) + f'(x_0)\,(x-x_0) + \frac{1}{2}f''(\xi(x))(x-x_0)^2\right] dx \\ &=\; (b-a)\,f(x_0) + \frac{1}{2}\int_a^b f''(\xi)(x-x_0)^2 dx\,. \end{aligned}$$

Die Substitution

$$s(x) = \int (x-x_0)^2 dx \;=\; \frac{1}{3}(x-x_0)^3$$

ergibt eine für $x \in (a,b)$ *streng monoton steigende Funktion, für die die Umkehrfunktion* $x = x(s)$ *existiert. Durch Anwendung des Mittelwertsatzes der Integralrechnung folgt dann*

$$\begin{aligned} \frac{1}{2}\int_a^b f''(\xi)(x-x_0)^2\,dx \;&=\; \frac{1}{2}\int_{s(a)}^{s(b)} f''(\xi(x(s)))\,ds \\ &=\; \frac{1}{2}\,[s(b)-s(a)]\,f''(\xi(x(\bar{s}))) \\ &=\; \frac{1}{6}\left[(b-x_0)^3 - (a-x_0)^3\right] f''(\eta) \;=\; \frac{1}{24}(b-a)^3 f''(\eta) \end{aligned}$$

mit einer Zwischenwertstelle $\eta = \xi(x(\bar{s})) \in (a,b)$*. Für die Mittelpunktformel*

$$I_0 \;=\; (b-a)\,f\Big(\frac{a+b}{2}\Big) \tag{2.9}$$

folgt dann die Darstellung

$$\int_a^b f(x)\,dx \;=\; (b-a)f\Big(\frac{a+b}{2}\Big) + \frac{1}{24}(b-a)^3 f''(\eta) \tag{2.10}$$

mit einer Zwischenwertstelle $\eta \in (a,b)$. Die Mittelpunktformel ist wieder ein Verfahren zweiter Ordnung, d.h. lineare Funktionen werden exakt integriert. Im Vergleich zur Trapezregel ist aber nur eine Funktionsauswertung von $f(x)$ erforderlich.
Betrachten wir für $f(x)$ das lineare Hermitesche Interpolationspolynom (1.27) *für $n = 0$,*

$$f_1(x) \;=\; f(x_0) + f'(x_0)(x - x_0)\,,$$

dann gilt die Darstellung (1.29)

$$f(x) \;=\; f_1(x) + \frac{1}{2}f''(\xi(x))(x-x_0)^2$$

und die Abschätzung des Integrationsfehlers folgt wie oben.

Beispiel 2.3. *Für $n = 2$ sind die Stützstellen durch*

$$x_0 = a, \quad x_1 = \frac{1}{2}(a+b), \quad x_2 = b$$

gegeben und für die Integrationsgewichte ergibt sich

$$\widetilde{\omega}_0 = \frac{1}{3}, \quad \widetilde{\omega}_1 = \frac{4}{3}, \quad \widetilde{\omega}_2 = \frac{1}{3}.$$

Die resultierende Integrationsformel

$$I_2 \;=\; \frac{1}{6}(b-a)\left[f(a) + 4\,f\Big(\frac{a+b}{2}\Big) + f(b)\right] \tag{2.11}$$

ist die Simpson–Regel. Wie bei der Mittelpunktformel betrachten wir jetzt für die zu integrierende Funktion $f(x)$ das Hermitesche Interpolationspolynom $f_3(x)$ mit

$$f_3(x_0) = f(x_0), \quad f_3(x_1) = f(x_1), \quad f_3(x_2) = f(x_2), \quad f_3'(x_1) = f'(x_1).$$

Für dieses gilt die Darstellung

$$\begin{aligned} f_3(x) \;=\;& f(x_0)\frac{(x-x_1)^2}{(x_0-x_1)^2}\frac{x-x_2}{x_0-x_2} + f(x_2)\frac{(x-x_1)^2}{(x_2-x_1)^2}\frac{x-x_0}{x_2-x_0} \\ & + f(x_1)\frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}[\alpha x+\beta] + f'(x_1)\frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}(x-x_1) \end{aligned}$$

mit

$$\alpha = \frac{x_0+x_2-2x_1}{(x_1-x_0)(x_1-x_2)}, \quad \beta = 1-\alpha x_1\,.$$

Weiters gilt (1.29), *d.h.*

$$f(x) = f_3(x) + \frac{1}{4!} f^{(4)}(\xi(x))(x-x_0)(x-x_2)(x-x_1)^2 .$$

Mit

$$\begin{aligned}
\int_a^b \frac{(x-x_1)^2}{(x_0-x_1)^2}\frac{x-x_2}{x_0-x_2}dx &= \frac{1}{6}(b-a), \\
\int_a^b \frac{(x-x_1)^2}{(x_2-x_1)^2}\frac{x-x_0}{x_2-x_0}dx &= \frac{1}{6}(b-a), \\
\int_a^b \frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}(\alpha x+\beta)dx &= \frac{2}{3}(b-a), \\
\int_a^b \frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}(x-x_1)dx &= 0
\end{aligned}$$

erhalten wir

$$I = I_2 + \frac{1}{24}\int_a^b f^{(4)}(\xi(x))(x-x_0)(x-x_2)(x-x_1)^2\,dx .$$

Daraus folgt

$$\int_a^b f(x)\,dx \;=\; \frac{b-a}{6}\left[f(a) + 4\,f\Big(\frac{a+b}{2}\Big) + f(b)\right] - \frac{1}{2880}\,(b-a)^5\,f^{(4)}(\eta) \tag{2.12}$$

mit einer Zwischenwertstelle $\eta \in (a,b)$. Die Simpson–Regel ist ein Verfahren vierter Ordnung, d.h. kubische Polynome werden exakt integriert.

Beispiel 2.4. *Betrachtet wird das bestimmte Integral*

$$I = \int_0^a \sin x\,dx \;=\; 1 - \cos a\,.$$

Für $f(x) = \sin x$ ist $f''(x) = -\sin x$ und somit folgt

$$|f''(\xi)| \;\le\; f(a) \quad \textit{für } \xi \in (0,a)\,.$$

Damit ergeben sich für die Mittelpunktformel (2.9), *die Trapezregel* (2.7) *und für die Simpsonregel* (2.11) *die folgenden Fehlerabschätzungen*

$$|I - I_0| \le \frac{1}{24}a^3 \sin a, \quad |I - I_1| \le \frac{1}{12}a^3 \sin a, \quad |I - I_2| \le \frac{1}{2880}a^5 \sin a.$$

Die in Tabelle 2.4 *aufgeführten numerischen Ergebnisse spiegeln die theoretischen Fehlerabschätzungen wieder, wobei die einfache Mittelpunktformel der Trapezregel überlegen scheint.*

	Mittelpunkt		Trapez		Simpson	
a	Theorie	Fehler	Theorie	Fehler	Theorie	Fehler
$\frac{\pi}{2}$	1.61 –1	1.11 –1	3.23 –1	2.15 –1	3.32 –3	2.28 –3
$\frac{\pi}{4}$	1.43 –2	7.67 –3	2.85 –2	1.52 –2	7.34 –5	3.94 –5
$\frac{\pi}{8}$	9.66 –4	4.91 –4	1.93 –3	9.81 –4	1.24 –6	6.31 –7

Tabelle 2.1: Fehler der Newton–Cotes Integrationsformeln.

Als notwendiges Kriterium für die Konvergenz der bisherigen numerischen Integrationsformeln ist

$$|b-a| < 1$$

vorauszusetzen. Der allgemeine Fall kann durch zusammengesetzte Integrationsformeln

$$I \;=\; \int_a^b f(x)\,dx \;=\; \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,dx$$

mit Stützstellen

$$x_k = a + k\frac{b-a}{n} \qquad \text{für } k = 0,\ldots,n$$

und numerischer Integration der verbleibenden Integrale behandelt werden. Mit der Simpson–Regel folgt zum Beispiel

$$\begin{aligned} I_n &= \sum_{k=1}^{n} \frac{1}{6}(x_k - x_{k-1}) \Big[f(x_{k-1}) + 4\, f\Big(\frac{x_{k-1}+x_k}{2}\Big) + f(x_k) \Big] \\ &= \frac{b-a}{6n} \sum_{k=0}^{n} \Big[f(x_{k-1}) + 4\, f\Big(\frac{x_{k-1}+x_k}{2}\Big) + f(x_k) \Big]. \end{aligned}$$

2.2 Gauß–Legendre Integrationsformeln

Bei den bisherigen Betrachtungen wurden die paarweise verschiedenen Integrationspunkte x_k als gegeben vorausgesetzt. Allgemein enthält die Integrationsformel

$$I_n \;=\; \sum_{k=0}^{n} f(x_k)\,\omega_k$$

$2(n+1)$ frei wählbare Parameter (x_k, ω_k), $k = 0,\ldots,n$. Diese können aus der Forderung der exakten Integration von Polynomen $f(x) = x^\alpha$ für $\alpha = 0,\ldots,2n+1$ gewonnen werden,

$$\int_0^1 x^\alpha\,dx \;=\; \sum_{k=0}^{n} x_k^\alpha\,\omega_k\,.$$

Beispiel 2.5. *Für $n = 2$ und das Integrationsintervall $[a, b] = [0, 1]$ ergibt sich zur Bestimmung der Parameter (x_0, ω_0), (x_1, ω_1) und (x_2, ω_2) das nichtlineare Gleichungssystem*

$$\int_a^b x^\alpha \, dx \;=\; \frac{1}{\alpha + 1} \;=\; \sum_{k=0}^{2} x_k^\alpha \, \omega_k \quad \textit{für } \alpha = 0, \ldots, 5.$$

Aus Symmetriegründen ist

$$x_0 = t, \quad x_1 = \frac{1}{2}, \quad x_2 = 1 - t \quad \textit{für } t \in [0, 1]$$

und

$$\omega_0 = \omega_2 = \omega$$

zu wählen. Aus der Gleichung für $\alpha = 0$,

$$\omega_0 + \omega_1 + \omega_2 = 1,$$

folgt dann

$$\omega_1 = 1 - 2\omega.$$

Man prüft leicht nach, daß dann auch die Gleichung für $\alpha = 1$ erfüllt ist,

$$\frac{1}{2} \;=\; \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 \;=\; \omega t + (1 - 2\omega)\frac{1}{2} + \omega(1 - t) \;=\; \frac{1}{2}.$$

Für $\alpha = 2$ bzw. für $\alpha = 3$ ergibt sich

$$\frac{1}{12} \;=\; \omega \left[2t^2 - 2t + \frac{1}{2} \right],$$

während für $\alpha = 4$ bzw. für $\alpha = 5$

$$\frac{11}{80} \;=\; \omega \left[2t^4 - 4t^3 + 6t^2 - 4t + \frac{7}{8} \right]$$

folgt. Gleichsetzen liefert

$$40t^4 - 80t^3 + 54t^2 - 14t + 1 \;=\; 0$$

mit den Lösungen

$$t_{1/2} = \frac{1}{2} \pm \frac{\sqrt{15}}{10}, \quad t_{3/4} = \frac{1}{2}.$$

Für

$$t = \frac{1}{2} - \frac{\sqrt{15}}{10}$$

folgt

$$\omega = \frac{5}{18}.$$

Somit lauten die Stützstellen

$$x_0 = \frac{1}{2} - \frac{\sqrt{15}}{10}, \quad x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2} + \frac{\sqrt{15}}{10}$$

und die zugehörigen Integrationsgewichte sind

$$\omega_0 = \frac{5}{18}, \quad \omega_1 = \frac{8}{18}, \quad \omega_2 = \frac{5}{18}.$$

Die resultierende Integrationsformel wird als Gauß–Legendre Integration bezeichnet.

Beispiel 2.6. *Wie in Beispiel 2.4 betrachten wir wieder das bestimmte Integral*

$$I = \int_0^a \sin x \, dx \; = \; 1 - \cos a.$$

Verglichen wird die Simpson–Regel (2.11) *mit der in Beispiel 2.5 hergeleiteten Gauß–Legendre Integrationsformel. Bei gleicher Anzahl von Stützstellen zeigt sich eine deutlich schnellere Konvergenz.*

a	Simpson–Regel	Gauß–Legendre
$\frac{\pi}{2}$	2.28 –3	8.12 –6
$\frac{\pi}{4}$	3.94 –5	3.48 –8
$\frac{\pi}{8}$	6.31 –7	1.39 –10

Tabelle 2.2: Fehler der Simpson–Regel und der Gauß–Legendre Integrationsformel

.

Es stellt sich die Frage, wie der in Beispiel 2.5 betrachtete Zugang und insbesondere die Lösung des nichtlinearen Gleichungssystems verallgemeinert werden kann. Zur Berechnung des Integrals (2.1),

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

wird eine numerische Integrationsfomel

$$I_n = \sum_{k=0}^{n} f(x_k) \, \omega_k$$

betrachtet, welche Polynome $f_m(x)$ von möglichst maximalen Polynomgrad $m > n$ exakt integriert, d.h.

$$\int_a^b f_m(x) \, dx \; = \; \sum_{k=0}^{n} f_m(x_k) \, \omega_k \, .$$

Bezeichnet

$$f_n(x) = \sum_{k=0}^{n} f_m(x_k) L_k^n(x)$$

das Interpolationspolynom vom Grad n, so ist

$$r_m(x) := f_m(x) - f_n(x)$$

ein Polynom vom Grad m mit den $n+1$ Nullstellen x_k, $k = 0, \ldots, n$. Für $r_m(x)$ gilt daher die Darstellung

$$r_m(x) = f_m(x) - f_n(x) = g_{m-(n+1)}(x) \prod_{j=0}^{n} (x - x_j)$$

mit einem beliebigen, aber durch $f_m(x)$ eindeutig bestimmten Polynom $g_{m-(n+1)}(x)$ vom Grad $m-(n+1)$. Insbesondere gilt also

$$f_m(x) = \sum_{k=0}^{n} f_m(x_k) L_k^n(x) + g_{m-(n+1)}(x)\, p_{n+1}(x) \tag{2.13}$$

mit

$$p_{n+1}(x) = \prod_{j=0}^{n} (x - x_j)\,.$$

Einsetzen in die Integrationsformel (2.1) ergibt

$$\int_a^b f_m(x)dx \;=\; \sum_{k=0}^{n} f_m(x_k) \int_a^b L_k^n(x)\, dx + \int_a^b g_{m-(n+1)}(x)\, p_{n+1}(x)\, dx \;=\; \sum_{k=0}^{n} f_m(x_k)\omega_k$$

mit den Integrationsgewichten

$$\omega_k = \int_a^b L_k^n(x)\, dx,$$

falls

$$\int_a^b g_{m-(n+1)}(x)\, p_{n+1}(x)\, dx \;=\; 0$$

erfüllt ist. Für

$$g_{m-(n+1)}(x) \;=\; \sum_{j=0}^{m-(n+1)} \gamma_j p_j(x)$$

mit noch zu bestimmenden linear unabhängigen Polynomen $p_j(x)$ vom Grad j und zugehörigen Koeffizienten γ_j folgt dies aus der Orthogonalität

$$\int_a^b p_j(x)\, p_{n+1}(x)\, dx \;=\; 0 \quad \text{für } j = 0, \ldots, m-(n+1).$$

Offenbar ist

$$m - (n+1) \leq n$$

zu fordern, d.h.

$$m \leq 2n+1,$$

Insbesondere werden Polynome maximalen Grades $2n+1$ exakt integriert. Benötigt wird also ein System $\{p_j\}_{j=0}^{n+1}$ von zueinander orthogonalen Polynomen $p_j(x)$ vom Grad j mit

$$\int_a^b p_j(x)\, p_\ell(x)\, dx \;=\; 0 \quad \text{für } \ell \neq j. \tag{2.14}$$

Für die Nullstellen orthogonaler Polynome gilt das folgende Resultat:

Lemma 2.1. *Gegeben sei ein System $\{p^k\}_{k=0}^{n+1}$ orthogonaler Polynome, d.h. es gilt* (2.14). *Das Polynom $p_{n+1}(x)$ besitzt in $[a,b]$ $n+1$ einfache reelle Nullstellen $x_i^{(n+1)}$.*

Beweis: Sei $x_0 = \alpha + i\beta$, $\beta \neq 0$, eine komplexe Nullstelle von $p_{n+1}(x)$. Da die Koeffizienten von $p_{n+1}(x)$ reell sind, ist auch $\overline{x}_0 = \alpha - i\beta$ Nullstelle von $p_{n+1}(x)$. Für das Polynom

$$q_{n-1}(x) := \frac{p_{n+1}(x)}{(x-x_0)(x-\overline{x}_0)} = \frac{p_{n+1}(x)}{(x-\alpha)^2+\beta^2} \quad \text{für } x \in [a,b]$$

mit dem Polynomgrad $n-1$ ergibt sich mit, siehe (2.14)

$$0 = \int_a^b p_{n+1}(x) q_{n-1}(x)\, dx = \int_a^b \frac{[p_{n+1}(x)]^2}{(x-\alpha)^2+\beta^2}\, dx > 0$$

ein Widerspruch, d.h. $p_{n+1}(x)$ kann keine komplexen Nullstellen besitzen.
Sei $x_0 \in \mathbb{R}$, $x_0 > b$ eine Nullstelle von $p_{n+1}(x)$. Für das Polynom

$$q_n(x) = \frac{p_{n+1}(x)}{x_0 - x} \quad \text{für } x \in (a,b)$$

folgt wieder mit (2.14)

$$0 = \int_a^b p_{n+1}(x) q_n(x)\, dx = \int_a^b \frac{[p_{n+1}(x)]^2}{(x_0-x)}\, dx > 0$$

ein Widerspruch, wodurch eine reelle Nullstelle $x_0 > b$ ausgeschlossen wird.
Durch Betrachtung von

$$q_n(x) = \frac{p_{n+1}(x)}{x - x_0} \quad \text{für } x \in (a,b), \quad x_0 < a,$$

wird analog eine reelle Nullstelle $x_0 < a$ ausgeschlossen.

Somit besitzt $p_{n+1}(x)$ in $[a,b]$ $n+1$ reelle Nullstellen. Für eine mehrfache Nullstelle x_0 ist

$$q_{n-1}(x) := \frac{p_{n+1}(x)}{(x-x_0)^2}$$

und mit (2.14) ergibt sich wegen

$$0 = \int_a^b p_{n+1}(x) q_{n-1}(x)\,dx = \int_a^b \frac{[p_{n+1}(x)]^2}{(x-x_0)^2}\,dx > 0$$

wieder ein Widerspruch. Damit gibt es im Intervall $[a,b]$ $n+1$ einfache reelle Nullstellen $x_i^{(n+1)}$ des Polynoms $p_{n+1}(x)$. ∎

Die Stützstellen der numerischen Integrationsformel ergeben sich dann aus den Nullstellen von

$$p_{n+1}(x) = \prod_{j=0}^{n}(x - x_j), \tag{2.15}$$

und die Integrationsgewichte ergeben sich entsprechend aus

$$\omega_k = \int_a^b L_k^n(x)\,dx = \int_a^b \prod_{j=0, j\neq k}^{n} \frac{x-x_j}{x_k-x_j}\,dx \quad \text{für } k = 0,\ldots,n. \tag{2.16}$$

Mit den durch (2.15) bestimmten Stützstellen x_k und den durch (2.16) gegebenen Integrationsgewichten ω_k ergibt sich die Integrationsformel

$$I_n \;=\; \sum_{k=0}^{n} f(x_k)\,\omega_k\,. \tag{2.17}$$

Ausgehend von der Basis $\{x^j\}_{j=0}^{n+1}$ der Monome x^j kann durch Anwendung des Orthogonalisierungsverfahrens von Gram–Schmidt ein System orthogonaler Polynome konstruiert werden, vergleiche Algorithmus 2.1.

Setze
$p_0(x) := 1\,.$
Für $k = 0,\ldots,n$ berechne
$$p_{k+1}(x) := x^{k+1} - \sum_{\ell=0}^{k} \beta_{k\ell} p_\ell(x), \quad \beta_{k\ell} := \frac{1}{\alpha_\ell}\int_a^b x^{k+1} p_\ell(x)\,dx\,,$$
$$\alpha_{k+1} := \int_a^b [p_{k+1}(x)]^2\,dx\,.$$

Algorithmus 2.1: Konstruktion orthogonaler Polynome.

Beispiel 2.7. *Für $n = 2$ und $[a, b] = [0, 1]$ ist zunächst*

$$p_0(x) = 1, \quad \alpha_0 = \int_0^1 [p_0(x)]^2\, dx = 1\,.$$

Für $k = 0$ ist

$$\beta_{00} = \frac{1}{\alpha_0}\int_0^1 x\, p_0(x)\, dx = \int_0^1 x\, dx \;=\; \frac{1}{2}$$

und somit

$$p_1(x) = x - \beta_{00}p_0(x) = x - \frac{1}{2}, \quad \alpha_1 = \int_0^1 [p_1(x)]^2\, dx = \frac{1}{12}\,.$$

Für $k = 1$ ist

$$\begin{aligned}
\beta_{10} &= \frac{1}{\alpha_0}\int_0^1 x^2\, p_0(x)\, dx = \int_0^1 x^2\, dx = \frac{1}{3},\\
\beta_{11} &= \frac{1}{\alpha_1}\int_0^1 x^2\, p_1(x)\, dx = 12\int_0^1 x^2\left(x - \frac{1}{2}\right) dx = 1
\end{aligned}$$

und somit

$$p_2(x) = x^2 - \beta_{11}p_1(x) - \beta_{10}p_0(x) = x^2 - \left(x - \frac{1}{2}\right) - \frac{1}{3} = x^2 - x + \frac{1}{6},$$

d.h.

$$\alpha_2 = \int_0^1 [p_2(x)]^2\, dx = \frac{1}{180}\,.$$

Für $k = 2$ ist

$$\begin{aligned}
\beta_{20} &= \frac{1}{\alpha_0}\int_0^1 x^3\, p_0(x)\, dx = \int_0^1 x^3\, dx = \frac{1}{4},\\
\beta_{21} &= \frac{1}{\alpha_1}\int_0^1 x^3\, p_1(x)\, dx = 12\int_0^1 x^3\left(x - \frac{1}{2}\right) dx = \frac{9}{10},\\
\beta_{22} &= \frac{1}{\alpha_2}\int_0^1 x^3\, p_2(x)\, dx = 180\int_0^1 x^3\left(x^2 - x + \frac{1}{6}\right) dx = \frac{3}{2}
\end{aligned}$$

und somit

$$\begin{aligned}
p_3(x) &= x^3 - \beta_{22}p_2(x) - \beta_{21}p_1(x) - \beta_{20}p_0(x)\\
&= x^3 - \frac{3}{2}\left(x^2 - x + \frac{1}{6}\right) - \frac{9}{10}\left(x - \frac{1}{2}\right) - \frac{1}{4}\\
&= x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\,.
\end{aligned}$$

Zu bestimmen sind die Nullstellen von $p_3(x)$ durch Lösen der kubischen Gleichung

$$20x^3 - 30x^2 + 12x - 1 = 0$$

mit den Lösungen

$$x_0 = \frac{1}{2} - \frac{\sqrt{15}}{10}, \quad x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2} + \frac{\sqrt{15}}{10}.$$

Für die zugehörigen Integrationsgewichte ist zunächst

$$\omega_0 \;=\; \int_0^1 \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2}\,dx \;=\; \frac{5}{18}$$

und die Werte für $\omega_1 = \dfrac{4}{9}$ *und* $\omega_2 = \dfrac{5}{18}$ *ergeben sich analog.*

Der Fehler $I - I_n$ der Integrationsformel (2.17) für eine beliebige Funktion $f(x)$ kann auf die Fehlerdarstellung (1.29) des Hermiteschen Interpolationspolynoms zurückgeführt werden.

Satz 2.1. *Sei* $\{p_j\}_{j=0}^{n+1}$ *ein System von orthogonalen Polynomen* $p_j(x)$ *vom Grad* j *mit*

$$\int_a^b p_j(x)\, p_\ell(x)\, dx = 0 \quad \textit{für } \ell \neq j.$$

Für $k = 0, \ldots, n$ *seien* x_k *die Nullstellen von* $p_{n+1}(x)$*. Sei* f *in* $[a, b]$ $(2n+2)$*–mal stetig differenzierbar. Dann gilt*

$$\int_a^b f(x)\, dx \;=\; \sum_{k=0}^{n} f(x_k)\,\omega_k + \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b \prod_{j=0}^{n}(x - x_j)^2\, dx \tag{2.18}$$

mit einer geeigneten Zwischenwertstelle $\eta \in (a, b)$*.*

Beweis: Für die gegebene Funktion f sei f_{2n+1} das Hermitesche Interpolationspolynom vom Grad $2n+1$ mit

$$f_{2n+1}(x_i) = f(x_i), \quad f'_{2n+1}(x_i) = f'(x_i) \quad \text{für } i = 0, \ldots, n.$$

Mit (1.29) ist

$$f(x) = f_{2n+1}(x) + \frac{1}{(2n+2)!}\, f^{(2n+2)}(\xi(x)) \prod_{j=0}^{n}(x - x_j)^2$$

mit einer geeigneten Zwischenwertstelle $\xi(x) \in (a, b)$ und somit folgt, nach Konstruktion der numerischen Integrationsformel zur exakten Integration von Polynomen vom Grad $2n+1$,

$$\begin{aligned}
\int_a^b f(x)\, dx &= \int_a^b f_{2n+1}(x)\, dx + \frac{1}{(2n+2)!}\int_a^b f^{(2n+2)}(\xi(x)) \prod_{j=0}^{n}(x - x_j)^2\, dx \\
&= \sum_{k=0}^{n} f_{2n+1}(x_k)\,\omega_k + \frac{1}{(2n+2)!}\int_a^b f^{(2n+2)}(\xi(x)) \prod_{j=0}^{n}(x - x_j)^2\, dx \\
&= \sum_{k=0}^{n} f(x_k)\,\omega_k + \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b \prod_{j=0}^{n}(x - x_j)^2 dx
\end{aligned}$$

mit einer Zwischenwerstelle $\eta \in (a,b)$. Im letzten Schritt wurde die Interpolationsbedingung $f_{2n+1}(x_k) = f(x_k)$ benutzt, sowie ein verallgemeinerter Mittelwertsatz der Integralrechung verwendet. ∎

Ausgehend von $p_0(x) = 1$ können orthogonale Polynome $p_k(x)$ durch das Gram–Schmidtsche Orthogonalisierungsverfahren bestimmt werden, siehe Algorithmus 2.1. Als Ausgangspolynom kann aber auch $xp_k(x)$ gewählt werden, d.h. für $k = 0, \ldots, n$ ist

$$p_{k+1}(x) = x\, p_k(x) - \sum_{\ell=0}^{k} \beta_{k\ell} p_\ell(x)$$

mit den Koeffizienten

$$\beta_{k\ell} = \frac{\displaystyle\int_a^b x p_k(x) p_\ell(x)\, dx}{\displaystyle\int_a^b [p_\ell(x)]^2\, dx} \quad \text{für } \ell = 0, \ldots, k.$$

Für $\ell < k-1$ ist $xp_\ell(x)$ ein Polynom vom Grad $\ell+1$. Dieses kann als Linearkombination der orthogonalen Polynome $\{p_j(x)\}_{j=0}^{\ell+1}$, dargestellt werden,

$$x p_\ell(x) = \sum_{j=0}^{\ell+1} c_j p_j(x), \quad c_j = \frac{\displaystyle\int_a^b x p_\ell(x) p_j(x)\, dx}{\displaystyle\int_a^b [p_j(x)]^2\, dx}.$$

Für den Zähler von $\beta_{k\ell}$ folgt dann

$$\int_a^b x p_k(x) p_\ell(x)\, dx = \sum_{j=0}^{\ell+1} c_j \int_a^b p_k(x) p_j(x)\, dx = 0 \quad \text{für alle } \ell < k-1.$$

Damit ist

$$p_0(x) = 1, \quad p_{k+1}(x) = x\, p_k(x) - \beta_{kk} p_k(x) - \beta_{kk-1} p_{k-1}(x) \quad \text{für } k = 0, \ldots, n, \tag{2.19}$$

mit den Koeffizienten

$$\beta_{kk} = \frac{\displaystyle\int_a^b x p_k(x) p_k(x)\, dx}{\displaystyle\int_a^b [p_k(x)]^2\, dx}, \quad \beta_{kk-1} = \frac{\displaystyle\int_a^b x p_k(x) p_{k-1}(x)\, dx}{\displaystyle\int_a^b [p_{k-1}(x)]^2\, dx}, \tag{2.20}$$

wobei $p_{-1}(x) = 0$ gesetzt sei.

Lemma 2.2. *Gegeben sei die Rekursionsvorschrift* (2.19) *und* (2.20). *Für* $[a,b] = [-1,1]$ *gilt*

$$p_{2j}(-x) = p_{2j}(x), \quad p_{2j+1}(-x) = -p_{2j+1}(x) \quad \textit{für } j = 0,1,2,\ldots$$

sowie

$$\beta_{kk} = 0\,.$$

Beweis: Offensichtlich ist $p_0(x) = 1$ gerade und es folgt

$$p_1(x) = x - \beta_{00}, \quad \beta_{00} = \frac{\displaystyle\int_{-1}^{1} x\,dx}{\displaystyle\int_{-1}^{1} dx} = 0,$$

d.h.

$$p_1(x) = x, \quad p_1(-x) = -p_1(x),$$

und für den Zähler von β_{11} folgt

$$\int_{-1}^{1} x\,[p_1(x)]^2\,dx = 0\,.$$

Nach Induktionsvoraussetzung für $k = 1$ gilt also

$$p_k(-x) = -p_k(x), \quad p_{k-1}(-x) = p_{k-1}(x), \quad \beta_{kk}(x) = 0.$$

Dann folgt

$$p_{k+1}(x) = x\,p_k(x) - \beta_{kk-1}p_{k-1}(x),$$

d.h.

$$\begin{aligned} p_{k+1}(-x) &= -x\,p_k(-x) - \beta_{kk-1}p_{k-1}(-x) \\ &= x\,p_k(x) - \beta_{kk-1}p_{k-1}(x) \;=\; p_{k+1}(x) \end{aligned}$$

ist gerade und für den Zähler von β_{k+1k+1} folgt

$$\int_{-1}^{1} x\,[p_{k+1}(x)]^2\,dx = 0, \quad \text{d.h.} \quad \beta_{k+1k+1} = 0.$$

Ist $p_k(x)$ gerade und $p_{k-1}(x)$ ungerade, so folgt entsprechend, daß $p_{k+1}(x)$ ungerade ist und $\beta_{k+1k+1} = 0$ gilt. ∎

Für das Intervall $[a,b] = [-1,1]$ folgt also die Rekursionsvorschrift

$$p_0(x) = 1, \quad p_{k+1}(x) = x\,p_k(x) - \beta_k p_{k-1}(x) \quad \text{für } k = 0,\ldots,n, \tag{2.21}$$

mit den Koeffizienten

$$\beta_k = \frac{\displaystyle\int_{-1}^{1} x p_k(x) p_{k-1}(x)\,dx}{\displaystyle\int_{-1}^{1} [p_{k-1}(x)]^2\,dx}. \tag{2.22}$$

Lemma 2.3. *Für die durch* (2.21) *und* (2.22) *erzeugte Folge orthogonaler Polynome gilt*

$$\int_{-1}^{1} [p_k(x)]^2 \, dx = \frac{2[p_k(1)]^2}{1+2k} \, . \tag{2.23}$$

Beweis: Für $p_0(x) = 1$ ist

$$\int_{-1}^{1} [p_0(x)]^2 \, dx = \int_{-1}^{1} dx = 2, \quad \frac{2[p_0(1)]^2}{1+2\cdot 0} = 2$$

und für $p_1(x) = x$ ist

$$\int_{-1}^{1} [p_1(x)]^2 \, dx = \int_{-1}^{1} x^2 \, dx = \frac{2}{3}, \quad \frac{2[p_1(1)]^2}{1+2\cdot 1} = \frac{2}{3}$$

ist die Behauptung offensichtlich richtig. Für $k > 1$ folgt durch partielle Integration

$$\int_{-1}^{1} [p_k(x)]^2 \, dx = x \, [p_k(x)]^2 \Big|_{-1}^{1} - 2 \int_{-1}^{1} x p_k(x) p_k'(x) \, dx.$$

Nach Konstruktion ist

$$p_k(x) = x^k + q_{k-1}(x),$$

d.h.

$$p_k'(x) = k \, x^{k-1} + q_{k-1}'(x)$$

bzw.

$$\begin{aligned} x \, p_k'(x) &= k \, x^k + x q_{k-1}'(x) \\ &= k \Big[p_k(x) - q_{k-1}(x) \Big] + x q_{k-1}'(x) \\ &= k \, p_k(x) + x q_{k-1}'(x) - k q_{k-1}(x) \\ &= k \, p_k(x) + r_{k-1}(x). \end{aligned}$$

Damit folgt

$$\int_{-1}^{1} x p_k(x) p_k'(x) \, dx = \int_{-1}^{1} p_k(x) [k p_k(x) + r_{k-1}(x)] \, dx = k \int_{-1}^{1} [p_k(x)]^2 \, dx,$$

woraus sich unmittelbar die Behauptung ergibt. ∎

Lemma 2.4. *Für den in* (2.22) *erklärten Koeffizienten gilt die Darstellung*

$$\beta_k = \frac{2k-1}{2k+1} \, \frac{[p_k(1)]^2}{[p_{k-1}(1)]^2} \, .$$

Beweis: Für den Zähler von β_k ist

$$\begin{aligned}\int_{-1}^{1} x p_k(x) p_{k-1}(x)\,dx &= \int_{-1}^{1} p_k(x) x [x^{k-1} + q_{k-2}(x)] dx \\ &= \int_{-1}^{1} p_k(x) [x^k + x q_{k-2}(x)] dx \\ &= \int_{-1}^{1} p_k(x) [p_k(x) - q_{k-1}(x) + x q_{k-2}(x)] dx \\ &= \int_{-1}^{1} [p_k(x)]^2\,dx\end{aligned}$$

Damit folgt

$$\beta_k = \frac{\displaystyle\int_{-1}^{1} [p_k(x)]^2\,dx}{\displaystyle\int_{-1}^{1} [p_{k-1}(x)]^2\,dx} = \frac{\dfrac{2[p_k(1)]^2}{1+2k}}{\dfrac{2[p_{k-1}(1)]^2}{1+2(k-1)}} = \frac{2k-1}{2k+1}\frac{[p_k(1)]^2}{[p_{k-1}(1)]^2}.$$

■

Damit folgt

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_{k+1}(x) = x p_k(x) - \frac{2k-1}{2k+1}\frac{[p_k(1)]^2}{[p_{k-1}(1)]^2} p_{k-1}(x) \tag{2.24}$$

und

$$p_0(1) = 1, \quad p_1(1) = 1, \quad p_{k+1}(1) = p_k(1) - \frac{2k-1}{2k+1}\frac{[p_k(1)]^2}{p_{k-1}(1)}\,. \tag{2.25}$$

Lemma 2.5. *Für die in* (2.25) *erklärte Rekursion gilt*

$$p_k(1) = \frac{k}{2k-1} p_{k-1}(1)\,. \tag{2.26}$$

Beweis: Für $k = 1$ ist (2.26) offensichtlich richtig. Mit der Induktionsvoraussetzung

$$p_k(1) = \frac{k}{2k-1} p_{k-1}(1)$$

ergibt sich mit

$$p_{k+1}(1) = p_k(1) - \frac{2k-1}{2k+1}\frac{[p_k(1)]^2}{p_{k-1}(1)} = \left[1 - \frac{k}{2k+1}\right] p_k(1) = \frac{k+1}{2(k+1)-1}\, p_k(1)$$

die Behauptung für $k+1$. ■

Für die durch (2.24) erklärten Polynome $p_k(x)$ definieren wir die skalierten Polynome

$$P_k(x) = \frac{p_k(x)}{p_k(1)} \quad \text{mit} \quad P_k(1) = 1\,.$$

Aus (2.24) folgt dann, unter Verwendung von (2.26),

$$\begin{aligned} p_{k+1}(x) &= x p_k(x) - \frac{2k-1}{2k+1}\frac{[p_k(1)]^2}{[p_{k-1}(1)]^2} p_{k-1}(x) \\ &= p_k(1)\left[x\,\frac{p_k(x)}{p_k(1)} - \frac{2k-1}{2k+1}\frac{p_k(1)}{p_{k-1}(1)}\frac{p_{k-1}(x)}{p_{k-1}(1)}\right] \\ &= p_k(1)\left[x\,P_k(x) - \frac{2k-1}{2k+1}\frac{p_k(1)}{p_{k-1}(1)} P_{k-1}(x)\right] \\ &= p_k(1)\left[x\,P_k(x) - \frac{k}{2k+1} P_{k-1}(x)\right]. \end{aligned}$$

Mit

$$p_{k+1}(1) = \frac{k+1}{2k+1}\,p_k(1)$$

ergibt sich dann

$$P_{k+1}(x) = \frac{p_{k+1}(x)}{p_{k+1}(1)} = \frac{2k+1}{k+1}\left[x\,P_k(x) - \frac{k}{2k+1} P_{k-1}(x)\right],$$

d.h.

$$(k+1)\,P_{k+1}(x) = (2k+1)\,x\,P_k(x) - k\,P_{k-1}(x) \quad \text{für } k = 1, 2, 3, \ldots. \tag{2.27}$$

Mit $P_0(x) = 1$ und $P_1(x)$ ist dies die Rekursionsvorschift der Legendre–Polynome. Aus (2.23) folgt sofort

$$\int_{-1}^{1} [P_n(x)]^2\,dx = \frac{2}{1+2n}\,.$$

Die Legendre–Polynome $P_n(x)$ können auch als Lösung der Differentialgleichung

$$(x^2-1)P_n''(x) + 2xP_n'(x) - n(n+1)P_n(x) = 0 \quad \text{für } x \in (-1,1)$$

erklärt werden. Ein Potenzreihenansatz und die Skalierungsbedingung $P_n(1) = 1$ führt dann zu der als Rodrigues–Formel bekannten Darstellung

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n\,. \tag{2.28}$$

Die Orthogonalität und die Rekursion der Legendre–Polynome folgt dann unter Verwendung der Differentialgleichung.

Für die ersten sechs Legendre–Polynome erhalten wir

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{1}{2}(3x^2 - 1), \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x), \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \\ P_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x)\,. \end{aligned}$$

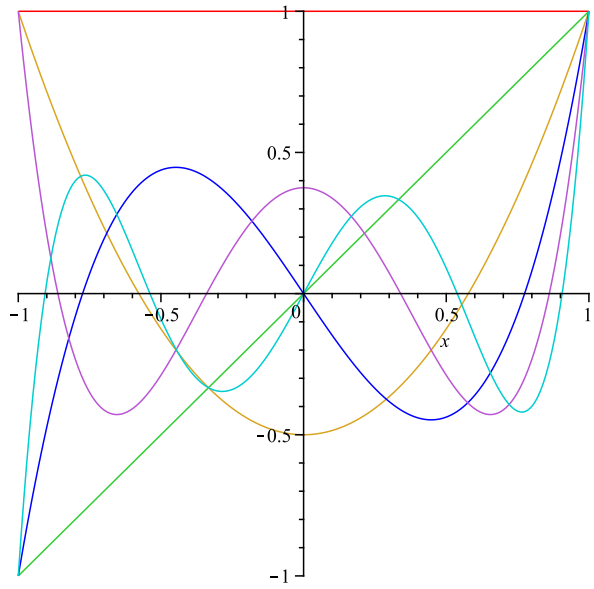


Abbildung 2.1: Legendre–Polynome $P_0(x), \ldots, P_5(x)$.

2.3 Gauß–Tschebyscheff Integrationsformeln

Abschließend sollen geeignete numerische Integrationsformeln für gewichtete Integrale der Form

$$I = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx, \quad I_n = \sum_{k=0}^{n} f(x_k)\omega_k \tag{2.29}$$

betrachtet werden. Wie bei der Herleitung der Gauß–Legendre Integrationsformeln sollen zunächst Polynome $f_m(x)$ von möglichst maximalen Polynomgrad $m = 2n + 1$ exakt

integriert werden. Für Polynome f_m vom Grad $m = 2n + 1$ gilt die Darstellung (2.13),

$$f_m(x) = \sum_{k=0}^{n} f_m(x_k) L_k^n(x) + g_{m-(n+1)}(x) \prod_{j=0}^{n} (x - x_j).$$

Einsetzen in die Integrationsformel (2.29) ergibt

$$\int_{-1}^{1} \frac{f_m(x)}{\sqrt{1-x^2}} dx = \sum_{k=0}^{n} f_m(x_k) \int_{-1}^{1} \frac{L_k^n(x)}{\sqrt{1-x^2}} dx + \int_{-1}^{1} \frac{g_{m-(n+1)}(x) p_{n+1}(x)}{\sqrt{1-x^2}} dx = \sum_{k=0}^{n} f_m(x_k) \omega_k$$

mit den Integrationsgewichten

$$\omega_k = \int_{-1}^{1} \frac{L_k^n(x)}{\sqrt{1-x^2}} dx, \quad p_{n+1}(x) = \prod_{j=0}^{n} (x - x_j),$$

falls

$$\int_{-1}^{1} \frac{g_{m-(n+1)}(x) p_{n+1}(x)}{\sqrt{1-x^2}} dx \; = \; 0$$

erfüllt ist. Gesucht sind also orthogonale Polynome $p_j(x)$ vom Grad j mit

$$\int_{-1}^{1} \frac{p_j(x) p_\ell(x)}{\sqrt{1-x^2}} \, dx \; = \; 0 \quad \text{für } \ell \neq j.$$

Lemma 2.6. *Für die durch* (1.16) *definierten Tschebyscheff–Polynome* $T_k(x)$ *gilt die Orthogonalität*

$$\int_{-1}^{1} \frac{T_j(x) T_\ell(x)}{\sqrt{1-x^2}} dx \; = \; \begin{cases} 0 & \text{für } \ell \neq j, \\ \dfrac{\pi}{2} & \text{für } \ell = j \neq 0, \\ \pi & \text{für } \ell = j = 0. \end{cases} \tag{2.30}$$

Beweis: Mit der Substitution

$$x \; = \; \cos\varphi, \quad \frac{dx}{d\varphi} \; = \; -\sin\varphi, \quad \varphi \in [\pi, 0]$$

und der Darstellung (1.17) ist zunächst

$$\int_{-1}^{1} \frac{T_j(x) T_\ell(x)}{\sqrt{1-x^2}} dx = \int_{-1}^{1} \frac{\cos(j \arccos x) \, \cos(\ell \arccos x)}{\sqrt{1-x^2}} dx = \int_{0}^{\pi} \cos j\varphi \cos \ell\varphi \, d\varphi \, .$$

Aus dem Additionstheorem

$$\cos\alpha + \cos\beta \;=\; 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}$$

folgt mit

$$\alpha \;=\; (j+\ell)\varphi, \quad \beta \;=\; (j-\ell)\varphi$$

für den Integranden

$$\cos j\varphi\,\cos\ell\varphi \;=\; \frac{1}{2}\left[\cos(j+\ell)\varphi + \cos(j-\ell)\varphi\right].$$

Damit ist

$$\int_{-1}^{1} \frac{T_j(x)T_\ell(x)}{\sqrt{1-x^2}}dx \;=\; \frac{1}{2}\int_0^\pi \left[\cos(j+\ell)\varphi + \cos(j-\ell)\varphi\right]d\varphi\,.$$

Für $j \neq \ell$ und $m = j \pm \ell \neq 0$ folgt die Behauptung aus

$$\int_0^\pi \cos m\varphi\, d\varphi \;=\; 0\,.$$

Für $j = \ell = 0$ ist $T_0(x) = 1$ und somit

$$\int_{-1}^{1} \frac{T_0(x)T_0(x)}{\sqrt{1-x^2}}dx \;=\; \int_0^\pi d\varphi \;=\; \pi,$$

und für $j = \ell \neq 0$ ergibt sich

$$\int_{-1}^{1} \frac{T_j(x)T_j(x)}{\sqrt{1-x^2}}dx \;=\; \frac{1}{2}\int_0^\pi d\varphi \;=\; \frac{\pi}{2}\,.$$

■

Als Stützstellen x_k der numerischen Integrationsformel I_n sind also die Nullstellen $x_k^{(n+1)}$ des Tschebyscheff–Polynoms $T_{n+1}(x)$ zu wählen,

$$x_k^{(n+1)} \;=\; \cos\frac{(1+2k)\pi}{2(n+1)} \quad \text{für } k = 0,\ldots,n.$$

Zu berechnen bleiben die Integrationsgewichte

$$\omega_k = \int_{-1}^{1} \frac{L_k^n(x)}{\sqrt{1-x^2}} \quad \text{für } k = 0,\ldots,n.$$

Stellen wir das Lagrange–Polynom $L_k^n(x)$ in der Basis der Tschebyscheff–Polynome dar,

$$L_k^n(x) = \sum_{i=0}^{n} \alpha_i T_i(x),$$

so erhalten wir für die Zerlegungskoeffizienten

$$a_0 = \frac{1}{\pi} \int_{-1}^{1} \frac{L_k^n(x)}{\sqrt{1-x^2}}\, dx$$

bzw.

$$a_i = \frac{2}{\pi} \int_{-1}^{1} \frac{L_k^n(x) T_i(x)}{\sqrt{1-x^2}}\, dx \quad \text{für } i = 1, \ldots, n.$$

Da die Integrationsformel exakt für Polynome vom maximalen Grad $2n+1$ ist, folgt

$$a_0 = \frac{1}{\pi} \int_{-1}^{1} \frac{L_k^n(x)}{\sqrt{1-x^2}}\, dx = \frac{1}{\pi} \sum_{\ell=0}^{n} L_k^n(x_\ell)\omega_\ell = \frac{1}{\pi}\omega_k$$

bzw.

$$a_i = \frac{2}{\pi} \int_{-1}^{1} \frac{L_k^n(x) T_i(x)}{\sqrt{1-x^2}}\, dx = \frac{2}{\pi} \sum_{\ell=0}^{n} L_k^n(x_\ell) T_i(x_\ell)\omega_\ell = \frac{2}{\pi} T_i(x_k)\omega_k\,. \quad \text{für } i = 1, \ldots, n.$$

Dann ergibt sich

$$\begin{aligned} \int_{-1}^{1} \frac{[L_k^n(x)]^2}{\sqrt{1-x^2}}\, dx &= \sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i\alpha_j \int_{-1}^{1} \frac{T_i(x)T_j(x)}{\sqrt{1-x^2}}\, dx \\ &= \alpha_0^2\,\pi + \frac{\pi}{2}\sum_{i=1}^{n} \alpha_i^2 \\ &= \omega_k^2 \left[\frac{1}{\pi} + \frac{2}{\pi}\sum_{i=1}^{n} [T_i(x_k)]^2\right]. \end{aligned}$$

Andererseits ist

$$\int_{-1}^{1} \frac{[L_k^n(x)]^2}{\sqrt{1-x^2}}\, dx = \sum_{\ell=0}^{n} [L_k^n(x_\ell)]\omega_\ell = \omega_k\,,$$

und somit folgt

$$\begin{aligned} \omega_k &= \left[\frac{1}{\pi} + \frac{2}{\pi}\sum_{i=1}^{n} [T_i(x_k)]^2\right]^{-1} \\ &= \left[\frac{1}{\pi} + \frac{2}{\pi}\sum_{i=1}^{n} \left[\cos i\frac{(2k+1)\pi}{2(n+1)}\right]^2\right]^{-1}. \end{aligned}$$

Mit

$$\Big[\cos\frac{\alpha}{2}\Big]^2 = \frac{1}{2}\Big[1+\cos\alpha\Big]$$

folgt

$$\begin{aligned}\omega_k &= \pi\left[1+\sum_{i=1}^{n}\Big[1+\cos\frac{(2k+1)i\pi}{n+1}\Big]\right]^{-1}\\ &= \pi\left[n+1+\sum_{i=1}^{n}\cos\frac{(2k+1)i\pi}{n+1}\right]^{-1}.\end{aligned}$$

Für $n = 2m$ ist

$$\begin{aligned}\sum_{i=1}^{n}\cos\frac{(2k+1)i\pi}{n+1} &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2m+1}+\cos\frac{(2k+1)(2m+1-i)\pi}{2m+1}\right]\\ &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2m+1}+\cos\left((2k+1)\pi-\frac{(2k+1)i\pi}{2m+1}\right)\right]\\ &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2m+1}+\cos(2k+1)\pi\cos\frac{(2k+1)i\pi}{2m+1}\right]\\ &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2m+1}-\cos\frac{(2k+1)i\pi}{2m+1}\right] = 0.\end{aligned}$$

Entsprechend ist für $n = 2m+1$

$$\begin{aligned}\sum_{i=1}^{n}\cos\frac{(2k+1)i\pi}{n+1} &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2(m+1)}+\cos\frac{(2k+1)(2(m+1)-i)\pi}{2(m+1)}\right]+\cos(2k+1)\frac{\pi}{2}\\ &= \sum_{i=1}^{m}\left[\cos\frac{(2k+1)i\pi}{2(m+1)}+\cos\left((2k+1)\pi-\frac{(2k+1)i\pi}{2(m+1)}\right)\right]\\ &= 0\,.\end{aligned}$$

Für die Integrationsgewichte ω_k ergibt sich somit

$$\omega_k = \frac{\pi}{n+1} \quad \text{für } k = 0,\ldots,n.$$

Damit folgt die numerische Integrationsformel

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}dx = \frac{\pi}{n+1}\sum_{k=0}^{n}f(x_k^{(n+1)}) + \frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_{-1}^{1}\prod_{j=0}^{n}(x-x_j^{(n+1)})^2\frac{dx}{\sqrt{1-x^2}} \qquad (2.31)$$

mit einer geeigneten Zwischenwertstelle $\eta \in (-1,+1)$. Die Integrationsformel (2.31) ist nach Konstruktion exakt für alle Polynome vom Grad $2n+1$. Insbesondere für $f(x) = T_k(x)T_\ell(x)$ mit $k,\ell = 0,\ldots,n$ folgt also

$$\int_{-1}^{1} \frac{T_k(x)T_\ell(x)}{\sqrt{1-x^2}}dx = \frac{\pi}{n+1}\sum_{i=0}^{n} T_k(x_i^{(n+1)})T_\ell(x_i^{(n+1)}).$$

Mit der Orthogonalität (2.30) folgt daraus

$$\sum_{i=0}^{n} T_k(x_i^{(n+1)})T_\ell(x_i^{(n+1)}) = \begin{cases} 0 & \text{für } k \neq \ell, \\ \frac{1}{2}(n+1) & \text{für } k = \ell \neq 0, \\ n+1 & \text{für } k = \ell = 0. \end{cases} \tag{2.32}$$

Wird für das globale Interpolationspolynom $f_n(x)$ vom Grad n der Ansatz

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

mit den Tschebyscheff–Polynomen $T_k(x)$ gewählt, so lauten die Interpolationsgleichungen in den Nullstellen $x_i^{(n+1)}$ des Tschebyscheff–Polynoms $T_{n+1}(x)$

$$f_n(x_i^{(n+1)} = \sum_{k=0}^{n} a_k T_k(x_i^{(n+1)}) = f(x_i^{(n+1)}) \quad \text{für } i = 0,\ldots,n.$$

Für die Bestimmung der Zelregungskoeffizienten ist also ein lineares Gleichungssystem mit einer vollbesetzten Matrix, d.h. mit einer Matrix mit $(n+1)^2$ Nichtnulleinträgen, zu lösen. Die Anwendung eines direkten Lösungsverfahrens, zum Beispiel des Gaußschen Eliminationsverfahrens, erfordert $\mathcal{O}(n^3)$ wesentliche Operationen, d.h. eine Verdoppelung von n verachtfacht die erforderliche Rechenzeit.
Zu bestimmen bleiben die Zerlegungskoeffizienten von f_n als Lösung des zugehörigen linearen Gleichungssystems. Der Ansatz

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

mit Tschebyscheff–Polynomen $T_k(x)$ führt dann auf die Interpolationsgleichungen

$$f_n\Big(\bar{x}_i^{(n+1)}\Big) = \sum_{k=0}^{n} a_k T_k\Big(\bar{x}_i^{(n+1)}\Big) = f_i \quad \text{für } i = 0,\ldots,n.$$

Zur Bestimmung der Zerlegungskoeffizienten a_k werden die Interpolationsgleichungen mit $T_\ell(\bar{x}_i^{(n+1)})$ multipliziert und über $i = 0,\ldots,n$ summiert,

$$\sum_{i=0}^{n}\sum_{k=0}^{n} a_k T_k\Big(\bar{x}_i^{(n+1)}\Big)T_\ell\Big(\bar{x}_i^{(n+1)}\Big) = \sum_{i=0}^{n} f_i T_\ell\Big(\bar{x}_i^{(n+1)}\Big)\ .$$

Es gilt

$$\sum_{i=0}^{n} T_k\Big(\bar{x}_i^{(n+1)}\Big) T_\ell\Big(\bar{x}_i^{(n+1)}\Big) = \begin{cases} 0 & \text{für } k \neq \ell, \\ \frac{1}{2}(n+1) & \text{für } k = \ell \neq 0, \\ n+1 & \text{für } k = \ell = 0 \end{cases}$$

und somit

$$a_0 = \frac{1}{(n+1)} \sum_{i=0}^{n} f_i \quad \text{für } k = 0$$

bzw.

$$a_k = \frac{2}{n+1} \sum_{i=0}^{n} f_i T_k\Big(\bar{x}_i^{(n+1)}\Big) \quad \text{für } k = 1, \ldots, n.$$

Dies ist gleichbedeutend mit

$$a_k = \frac{2}{n+1} \sum_{i=0}^{n} f_i \cos k \frac{(2i+1)\pi}{2(n+1)} \quad \text{für } k = 1, \ldots, n.$$

Eine effiziente Berechnung der Zerlegungskoeffizienten a_k kann schließlich durch eine schnelle Fouriertransformation realisiert werden, siehe zum Beispiel [14].

2.4 Schnelle Fouriertransformation

In diesem Abschnitt sollen effiziente Verfahren zur Realisierung der diskreten Kosinus–Transformation

$$a_k = \sum_{j=0}^{n-1} f_j \cos \frac{2\pi kj}{n} \quad \text{für } k = 0, \ldots, n-1 \tag{2.33}$$

beziehungsweise der diskreten Sinus–Transformation

$$b_k = \sum_{j=1}^{n-1} f_j \sin \frac{2\pi kj}{n} \quad \text{für } k = 1, \ldots, n-1 \tag{2.34}$$

betrachtet werden. Durch Übergang ins Komplexe sind diese gleichbedeutend mit der komplexen Fouriertransformation

$$c_k = \sum_{j=0}^{n-1} f_j e^{-i2\pi kj/n} \quad \text{für } k = 0, \ldots, n-1. \tag{2.35}$$

Anwendungen dieser Art ergeben sich beispielsweise bei der Interpolation mit Tschebyscheff–Polynomen oder bei der Beschreibung zirkulanter Matrizen.
Eine direkte Auswertung der Koeffizienten c_k in (2.35) erfordert n^2 komplexe Multiplikationen. Ziel ist deshalb die Herleitung eines schnelleren Berechnungsverfahrens. Die Idee dafür

besteht in der Rückführung der ursprünglichen Aufgabe auf eine Folge ähnlicher Probleme kleinerer Dimension.
Sei $n = 2m$. Dann ergibt sich für die Koeffizienten c_k mit geradzahligem Index $k = 2\ell$ für $\ell = 0, \ldots, m-1$ durch Aufspalten der Summe

$$\begin{aligned} c_{2\ell} &= \sum_{j=0}^{n-1} f_j e^{-i2\pi 2\ell j/n} \\ &= \sum_{j=0}^{m-1} \left[f_j e^{-i2\pi 2\ell j/n} + f_{m+j} e^{-i2\pi 2\ell (m+j)/n} \right] \\ &= \sum_{j=0}^{m-1} \left[f_j + f_{m+j} e^{-i2\pi 2\ell m/n} \right] e^{-i2\pi 2\ell j/n} \\ &= \sum_{j=0}^{m-1} \left[f_j + f_{m+j} \right] e^{-i2\pi \ell j/m} . \end{aligned}$$

Für die Koeffizienten c_k mit ungeradem Index $k = 2\ell + 1$ und $\ell = 0, \ldots, m-1$ ergibt sich analog

$$\begin{aligned} c_{2\ell+1} &= \sum_{j=0}^{n-1} f_j e^{-i2\pi (2\ell+1) j/n} \\ &= \sum_{j=0}^{m-1} \left[f_j e^{-i2\pi (2\ell+1) j/n} + f_{m+j} e^{-i2\pi (2\ell+1)(m+j)/n} \right] \\ &= \sum_{j=0}^{m-1} \left[f_j + f_{m+j} e^{-i2\pi (2\ell+1) m/n} \right] e^{-i2\pi (2\ell+1) j/n} \\ &= \sum_{j=0}^{m-1} \left[f_j - f_{m+j} \right] e^{-i2\pi j/n} \, e^{-i2\pi \ell j/m}. \end{aligned}$$

Damit kann die Fourier–Transformation (2.35) für n Koeffizienten realisiert werden durch zwei Fourier–Transformation für $m = n/2$ Koeffizienten, wobei die modifizierten Koeffizienten

$$\hat{f}_j = f_j + f_{m+j}, \quad \hat{f}_{m+j} = \left[f_j - f_{m+j} \right] e^{-i2\pi j/n} \tag{2.36}$$

für $j = 0, \ldots, m-1$ zu berechnen sind. Für $n = 2^p$ ist dieses Vorgehen rekursiv anwendbar, und nach p Reduktionsschritten sind n Fourier–Transformationen für jeweils einen Koeffizienten durchzuführen. Bei der Berechnung der Koeffizienten mit ungeradem Index sind dabei jeweils $m = n/2$ komplexe Multiplikationen zu realisieren. Damit ergeben sich insgesamt

$$p \frac{n}{2} = \frac{1}{2} n \log n$$

komplexe Multiplikationen.

Beispiel 2.8. *Betrachtet wird die komplexe Fourier–Transformation* (2.35) *für die Berechnung von* $n = 2^3 = 8$ *Koeffizienten,*

$$c_k \;=\; \sum_{j=0}^{7} f_j e^{-i2\pi kj/8} \quad \textit{für } k = 0, \ldots, n-1.$$

Die rekursive Anwendung der Vorschrift (2.36) *für die Berechnung der Koeffizienten* c_k *ergibt das folgende Schema:*

c_0	$f_0^1 = f_0 + f_4$	c_0	$f_0^2 = f_0^1 + f_2^1$	c_0	$f_0^3 = f_0^2 + f_1^2$
c_2	$f_1^1 = f_1 + f_5$	c_4	$f_1^2 = f_1^1 + f_3^1$	c_4	$f_1^3 = f_0^2 - f_1^2$
c_4	$f_2^1 = f_2 + f_6$	c_2	$f_2^2 = (f_0^1 - f_2^1)\omega_4^0$	c_2	$f_2^3 = f_2^2 + f_3^2$
c_6	$f_3^1 = f_3 + f_7$	c_6	$f_3^2 = (f_1^1 - f_3^1)\omega_4^1$	c_6	$f_3^3 = f_2^2 - f_3^2$
c_1	$f_4^1 = (f_0 - f_4)\omega_8^0$	c_1	$f_4^2 = f_4^1 + f_6^1$	c_1	$f_4^3 = f_4^2 + f_5^2$
c_3	$f_5^1 = (f_1 - f_5)\omega_8^1$	c_5	$f_5^2 = f_5^1 + f_7^1$	c_5	$f_5^3 = f_4^2 - f_5^2$
c_5	$f_6^1 = (f_2 - f_6)\omega_8^2$	c_3	$f_6^2 = (f_4^1 - f_6^1)\omega_4^0$	c_3	$f_6^3 = f_6^2 + f_7^2$
c_7	$f_7^1 = (f_3 - f_7)\omega_8^3$	c_7	$f_7^2 = (f_5^1 - f_7^1)\omega_4^1$	c_7	$f_7^3 = f_6^2 - f_7^2$

Dabei sind

$$\omega_m^k \;=\; e^{i2\pi k/m} \quad \textit{für } k = 0, \ldots, m-1$$

die komplexen Einheitswurzeln.
Nach der dreifachen Anwendung der Vorschrift (2.36) *sind schließlich die erhaltenen Werte* f_j^3 *den Koeffizienten* c_k *zuzuordnen. Dabei sind die durch die Umnummerierungen erfolgten Permuationen der Indizes der Koeffizienten* c_k *zu berücksichtigen. Dies kann durch eine Spiegelung der Binärdarstellung der Indizes erfolgen:*

Zuordnung	*Index von* f_j^3	*binär*	*Index von* c_k	*binär*
$c_0 = f_0^3$	0	0 0 0	0	0 0 0
$c_4 = f_1^3$	1	0 0 1	4	1 0 0
$c_2 = f_2^3$	2	0 1 0	2	0 1 0
$c_6 = f_3^3$	3	0 1 1	6	1 1 0
$c_1 = f_4^3$	4	1 0 0	1	0 0 1
$c_5 = f_5^3$	5	1 0 1	5	1 0 1
$c_3 = f_6^3$	6	1 1 0	3	0 1 1
$c_7 = f_7^3$	7	1 1 1	7	1 1 1

Für eine weitergehende Diskussion der schnellen Fouriertransformation sei hier auf [3] verwiesen, siehe auch [17] für eine Implementierung für allgemeines $n \in \mathbb{N}$.

Kapitel 3

Vektoren und Matrizen

In diesem Kapitel sollen die Grundlagen aus der linearen Algebra bereitgestellt werden, die später bei der Konstruktion effizienter Algorithmen für die Lösung linearer Gleichungssysteme benötigt werden. Neben den grundlegenden Begriffen wie zum Beispiel Normen von Vektoren und Matrizen wird die Singulärwertzerlegung beliebiger Matrizen hergeleitet. Das Orthogonalisierungsverfahren nach Gram–Schmidt bildet den Ausgangspunkt für die Herleitung von modernen Iterationsverfahren für lineare Gleichungssysteme.

3.1 Normen von Vektoren und Matrizen

Für $n \in \mathbb{N}$ ist $\mathbb{R}^n$ der Raum der n–dimensionalen Vektoren $\underline{u} \in \mathbb{R}^n$ mit Komponenten $u_i \in \mathbb{R}$ für $i = 1, \ldots, n$. Mit

$$\langle \cdot, \cdot \rangle \; : \; \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

wird ein beliebiges Skalarprodukt im Vektorraum $\mathbb{R}^n$ bezeichnet, das heißt es gilt die Distributivität

$$\langle \underline{u} + \underline{v}, \underline{w} \rangle \; = \; \langle \underline{u}, \underline{w} \rangle + \langle \underline{v}, \underline{w} \rangle,$$

die Kommutativität

$$\langle \underline{u}, \underline{v} \rangle \; = \; \langle \underline{v}, \underline{u} \rangle,$$

die Homogenität

$$\langle \alpha \underline{u}, \underline{v} \rangle \; = \; \alpha \, \langle \underline{u}, \underline{v} \rangle$$

für alle Vektoren $\underline{u}, \underline{v}, \underline{w} \in \mathbb{R}^n$ und $\alpha \in \mathbb{R}$ sowie die positive Definitheit

$$\langle \underline{u}, \underline{u} \rangle \; > \; 0$$

für alle $\underline{u} \in \mathbb{R}^n$ mit $\underline{u} \neq \underline{0}$. Insbesondere definiert

$$\langle \underline{u}, \underline{v} \rangle_2 \; := \; (\underline{u}, \underline{v}) \; = \; \sum_{i=1}^{n} u_i v_i$$

das Euklidische Skalarprodukt. Für einen Vektor $\underline{u} \in \mathbb{R}^n$ bezeichnet

$$\|\cdot\|_V \,:\, \mathbb{R}^n \to \mathbb{R}$$

eine beliebige Vektornorm, für welche die Normaxiome gelten, das heißt die positive Definitheit

$$\|\underline{u}\|_V \,\geq\, 0 \quad \text{für alle } \underline{u} \in \mathbb{R}^n, \quad \|\underline{u}\|_V \,= 0 \quad \text{genau dann, wenn } \underline{u} = \underline{0},$$

die Homogenität

$$\|\alpha\,\underline{u}\|_V \,=\, |\alpha|\,\|\underline{u}\|_V \quad \text{für alle } \underline{u} \in \mathbb{R}^n \text{ und } \alpha \in \mathbb{R},$$

sowie die Dreiecksungleichung

$$\|\underline{u} + \underline{v}\|_V \,\leq\, \|\underline{u}\|_V + \|\underline{v}\|_V \quad \text{für alle } \underline{u}, \underline{v} \in \mathbb{R}^n.$$

Beispiele für Vektornormen sind die Euklidische Norm

$$\|\underline{u}\|_2 \,:=\, \left(\sum_{i=1}^{n} u_i^2 \right)^{1/2},$$

die Maximumnorm

$$\|\underline{u}\|_\infty \,:=\, \max_{i=1,\ldots,n} |u_i| \,,$$

sowie die Summennorm

$$\|\underline{u}\|_1 \,:=\, \sum_{i=1}^{n} |u_i| \,.$$

Nach Definition ist

$$\|\underline{u}\|_2^2 \,=\, \sum_{i=1}^{n} u_i^2 \,=\, (\underline{u}, \underline{u}) \quad \text{für alle } \underline{u} \in \mathbb{R}^n$$

und es gilt die Cauchy–Schwarz–Ungleichung

$$(\underline{u}, \underline{v}) \,=\, \sum_{i=1}^{n} u_i v_i \,\leq\, \left(\sum_{i=1}^{n} u_i^2 \right)^{1/2} \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2} \,=\, \|\underline{u}\|_2 \|\underline{v}\|_2 \tag{3.1}$$

für alle $\underline{u}, \underline{v} \in \mathbb{R}^n$.

Zwei Vektornormen $\|\cdot\|_{V_1}$ und $\|\cdot\|_{V_2}$ heißen zueinander äquivalent, wenn unabhängig von $\underline{u} \in \mathbb{R}^n$ zwei positive Konstanten c_1 und c_2 existieren, so daß die Äquivalenzungleichungen

$$c_1\,\|\underline{u}\|_{V_1} \,\leq\, \|\underline{u}\|_{V_2} \,\leq\, c_2\,\|\underline{u}\|_{V_1} \quad \text{für alle } \underline{u} \in \mathbb{R}^n \tag{3.2}$$

erfüllt sind. Die Äquivalenzungleichungen sind scharf, wenn im allgemeinen unterschiedliche Vektoren $\underline{u} \in \mathbb{R}^n$ existieren, für die in (3.2) jeweils die Gleichheit gilt.

Lemma 3.1. *Für beliebiges $\underline{u} \in \mathbb{R}^n$ gelten die Äquivalenzungleichungen*

$$\begin{aligned} \|\underline{u}\|_\infty &\leq \|\underline{u}\|_1 \leq n \|\underline{u}\|_\infty, \\ \|\underline{u}\|_\infty &\leq \|\underline{u}\|_2 \leq \sqrt{n} \|\underline{u}\|_\infty, \\ \|\underline{u}\|_2 &\leq \|\underline{u}\|_1 \leq \sqrt{n} \|\underline{u}\|_2. \end{aligned}$$

Alle Abschätzungen sind scharf.

Beweis: Zunächst ist

$$\|\underline{u}\|_\infty = \max_{i=1,\ldots,n} |u_i| \leq \sum_{i=1}^{n} |u_i| = \|\underline{u}\|_1,$$

wobei die Gleichheit zum Beispiel für $\underline{u} = (1, 0, \ldots, 0)^\top$ angenommen wird. Für die Abschätzung in umgekehrter Richtung folgt

$$\|\underline{u}\|_1 = \sum_{i=1}^{n} |u_i| \leq n \max_{i=1,\ldots,n} |u_i| = n \|\underline{u}\|_\infty.$$

Diese ist scharf zum Beispiel für $\underline{u} = (1, \ldots, 1)^\top$.
Die Äquivalenz zwischen Maximumnorm $\|\cdot\|_\infty$ und Euklidischer Norm $\|\cdot\|_2$ folgt analog, das heißt

$$\|\underline{u}\|_\infty^2 = \left(\max_{i=1,\ldots,n} |u_i| \right)^2 = \max_{i=1,\ldots,n} |u_i|^2 \leq \sum_{i=1}^{n} |u_i|^2 = \|\underline{u}\|_2^2$$

sowie

$$\|\underline{u}\|_2^2 = \sum_{i=1}^{n} |u_i|^2 \leq n \max_{i=1,\ldots,n} |u_i|^2 = n \, ||\underline{u}||_\infty^2 \, .$$

Gleichheit gilt beispielsweise für $\underline{u} = (1, 0, \ldots, 0)^\top$ sowie für $\underline{u} = (1, \ldots, 1)^\top$.
Die Kombination der bereits gezeigten Ungleichungen ergibt für die Äquivalenz der Euklidischen Norm $\|\cdot\|_2$ zur Summennorm $\|\cdot\|_1$ die Ungleichungen

$$\frac{1}{n} \|\underline{u}\|_1 \leq \|\underline{u}\|_2 \leq \sqrt{n} \|\underline{u}\|_1. \tag{3.3}$$

Da aber in den einzelnen Äquivalenzungleichungen die Gleichheit jeweils für unterschiedliche Vektoren $\underline{u} \in \mathbb{R}^n$ angenommen wird, sind die resultierenden Äquivalenzungleichungen (3.3) nicht scharf und daher nicht optimal.
Mit der Cauchy–Schwarz–Ungleichung (3.1) folgt

$$\|\underline{u}\|_1 = \sum_{i=1}^{n} |u_i| = \sum_{i=1}^{n} (1 \cdot |u_i|) \leq \left(\sum_{i=1}^{n} 1^2 \right)^{1/2} \left(\sum_{i=1}^{n} |u_i|^2 \right)^{1/2} = \sqrt{n} \, \|\underline{u}\|_2 \, .$$

Diese ist scharf für $\underline{u} = (1, \ldots, 1)^\top$. Andererseits ist

$$\|\underline{u}\|_2^2 = \sum_{i=1}^{n} |u_i|^2 \leq \left(\sum_{i=1}^{n} |u_i| \right)^2 = \|\underline{u}\|_1^2$$

mit Gleichheit für $\underline{u} = (1, 0, \ldots, 0)^\top$. ■

Sei $B \in \mathbb{R}^{m \times n}$ eine beliebig gegebene Matrix mit Einträgen $B[k, \ell] = b_{k\ell} \in \mathbb{R}$ für $k = 1, \ldots, m$ und $\ell = 1, \ldots, n$. Mit

$$\|\cdot\|_M \; : \; \mathbb{R}^{m \times n} \to \mathbb{R}$$

wird eine beliebige Matrixnorm bezeichnet. Beispiele für Matrixnormen sind die Zeilensummennorm

$$\|B\|_\infty := \max_{k=1,\ldots,m} \sum_{\ell=1}^{n} |b_{k\ell}| \,,$$

die Spaltensummennorm

$$\|B\|_1 := \max_{\ell=1,\ldots,n} \sum_{k=1}^{m} |b_{k\ell}|$$

sowie die Frobenius–Norm (Hilbert–Schmidt–Norm)

$$\|B\|_F := \left(\sum_{k=1}^{m} \sum_{\ell=1}^{n} b_{k\ell}^2 \right)^{1/2} .$$

Für eine sowohl in $\mathbb{R}^n$ als auch in $\mathbb{R}^m$ gegebene Vektornorm $\|\cdot\|_V$ kann durch

$$\|B\|_M := \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n, B\underline{x} \in \mathbb{R}^m} \frac{\|B\underline{x}\|_V}{\|\underline{x}\|_V}$$

stets eine induzierte Matrixnorm definiert werden. Insbesondere induziert die Euklidische Vektornorm die Euklidische Matrixnorm

$$\|B\|_2 := \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{||B\underline{x}||_2}{||\underline{x}||_2}.$$

Lemma 3.2. *Die Zeilensummennorm $\|B\|_\infty$ wird durch die Maximumnorm $\|\underline{x}\|_\infty$ induziert.*

Beweis: Für die Maximumnorm von $B\underline{x} \in \mathbb{R}^m$ für einen beliebigen Vektor $\underline{x} \in \mathbb{R}^n$ ergibt sich

$$\|B\underline{x}\|_\infty = \max_{k=1,\ldots,m} \left| \sum_{\ell=1}^{n} b_{k\ell} x_\ell \right| \leq \|\underline{x}\|_\infty \max_{k=1,\ldots,m} \sum_{\ell=1}^{n} |b_{k\ell}| \,.$$

Für alle $\underline{x} \in \mathbb{R}^n$ mit $\|\underline{x}\|_\infty \neq 0$ ist somit

$$\frac{\|B\underline{x}\|_\infty}{\|\underline{x}\|_\infty} \;\leq\; \max_{k=1,\ldots,m} \sum_{\ell=1}^{n} |b_{k\ell}| \;=\; \|B\|_\infty,$$

woraus

$$\sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|B\underline{x}\|_\infty}{\|\underline{x}\|_\infty} \;\leq\; \|B\|_\infty$$

folgt. Für den Nachweis der umgekehrten Ungleichung bezeichne k_0 den Index, für welchen die Zeilensummennorm angenommen wird, das heißt

$$\|B\|_\infty \;=\; \max_{k=1,\ldots,m} \sum_{\ell=1}^{n} |b_{k\ell}| \;=\; \sum_{\ell=1}^{n} |b_{k_0\ell}|\,.$$

Sei $\widetilde{\underline{x}} \in \mathbb{R}^n$ definiert durch

$$\widetilde{x}_\ell \;=\; \begin{cases} \dfrac{b_{k_0\ell}}{|b_{k_0\ell}|} & \text{für } b_{k_0\ell} \neq 0, \\ 1 & \text{für } b_{k_0\ell} = 0 \end{cases}$$

und $\ell = 1, \ldots, n$. Nach Konstruktion ist $\|\widetilde{\underline{x}}\|_\infty = 1$. Dann ergibt sich

$$\|B\widetilde{\underline{x}}\|_\infty = \max_{k=1,\ldots,m} |\sum_{\ell=1}^{n} b_{k\ell}\widetilde{x}_\ell| \;\geq\; |\sum_{\ell=1}^{n} b_{k_0\ell}\widetilde{x}_\ell| \;=\; \sum_{\ell=1}^{n} |b_{k_0\ell}| \;=\; \|B\|_\infty,$$

und wegen $\|\widetilde{\underline{x}}\|_\infty = 1$ folgt

$$\|B\|_\infty \;\leq\; \frac{\|B\widetilde{\underline{x}}\|_\infty}{\|\widetilde{\underline{x}}\|_\infty} \;\leq\; \sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|B\underline{x}\|_\infty}{\|\underline{x}\|_\infty} \;\leq\; \|B\|_\infty$$

und somit die Gleichheit. ∎

Lemma 3.3. *Die Spaltensummennorm $\|B\|_1$ wird durch die Summennorm $\|\underline{x}\|_1$ induziert.*

Beweis: Für die Summennorm von $B\underline{x} \in \mathbb{R}^m$ ergibt sich

$$\begin{aligned} \|B\underline{x}\|_1 \;&=\; \sum_{k=1}^{m} \left| \sum_{\ell=1}^{n} b_{k\ell} x_\ell \right| \;\leq\; \sum_{k=1}^{m} \sum_{\ell=1}^{n} |b_{k\ell}|\,|x_\ell| \\ &\leq\; \left(\max_{\ell=1,\ldots,n} \sum_{k=1}^{m} |b_{k\ell}| \right) \sum_{\ell=1}^{n} |x_\ell| \;=\; \|B\|_1\,\|\underline{x}\|_1 \end{aligned}$$

für alle $\underline{x} \in \mathbb{R}^n$, und für $\|\underline{x}\|_1 \neq 0$ folgt

$$\sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|B\underline{x}\|_1}{\|\underline{x}\|_1} \;\leq\; \|B\|_1\,.$$

Sei nun ℓ_0 der Index, für den die Spaltensummennorm angenommen wird,

$$\|B\|_1 \;=\; \max_{\ell=1,\dots,n} \sum_{k=1}^{m} |b_{k\ell}| \;=\; \sum_{k=1}^{m} |b_{k\ell_0}|\,,$$

und sei $\widetilde{\underline{x}} = (\delta_{1\ell_0}, \dots, \delta_{n\ell_0})^\top$ mit $\|\widetilde{\underline{x}}\|_1 = 1$. Hierbei bezeichnet

$$\delta_{k\ell} \;=\; \begin{cases} 1 & \text{für } k = \ell, \\ 0 & \text{für } k \neq \ell \end{cases}$$

das Kroneckersymbol. Dann folgt

$$\|B\|_1 \;=\; \sum_{k=1}^{m} |b_{k\ell_0}| \;=\; \sum_{k=1}^{m} \left| \sum_{\ell=1}^{n} b_{k\ell} \widetilde{x}_\ell \right| \;=\; \|B\widetilde{\underline{x}}\|_1 \;=\; \frac{\|B\widetilde{\underline{x}}\|_1}{\|\widetilde{\underline{x}}\|_1} \;\leq\; \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{\|B\underline{x}\|_1}{\|\underline{x}\|_1}$$

und somit insgesamt die Behauptung

$$\|B\|_1 \;=\; \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{\|B\underline{x}\|_1}{\|\underline{x}\|_1}.$$

■

Eine Matrixnorm $\|\cdot\|_M$ heißt kompatibel beziehungsweise verträglich zur Vektornorm $\|\cdot\|_V$, wenn für beliebige Matrizen $B \in \mathbb{R}^{m\times n}$ und beliebige Vektoren $\underline{x} \in \mathbb{R}^n$ die Ungleichung

$$\|B\underline{x}\|_V \;\leq\; \|B\|_M \|\underline{x}\|_V$$

gilt. Für eine durch eine Vektornorm $\|\cdot\|_V$ induzierte Matrixnorm $\|\cdot\|_M$ folgt

$$\|B\|_M \;=\; \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{\|B\underline{x}\|_V}{\|\underline{x}\|_V} \;\geq\; \frac{\|B\underline{x}\|_V}{\|\underline{x}\|_V} \quad \text{für alle } \underline{x} \in \mathbb{R}^n, \|x\|_V \neq 0,$$

das heißt eine induzierte Matrixnorm $\|\cdot\|_M$ ist stets verträglich zu der sie erzeugenden Vektornorm $\|\cdot\|_V$. Ist eine Matrixnorm $\|\cdot\|_M$ durch eine Vektornorm $\|\cdot\|_V$ induziert, so ergibt sich für die Norm der Einheitsmatrix $I \in \mathbb{R}^{n\times n}$

$$\|I\|_M \;=\; \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{\|I\underline{x}\|_V}{\|\underline{x}\|_V} \;=\; \sup_{\underline{0} \neq \underline{x} \in \mathbb{R}^n} \frac{\|\underline{x}\|_V}{\|\underline{x}\|_V} \;=\; 1.$$

Abschließend soll ein Beispiel einer zu einer Vektornorm $\|\cdot\|_V$ verträglichen Matrixnorm $\|\cdot\|_M$ betrachtet werden, die durch keine Vektornorm induziert wird.

Beispiel 3.1. *Sei zunächst $m = n$. Für die Einheitsmatrix $I \in \mathbb{R}^{n\times n}$ gilt dann in der Frobenius–Norm $\|I\|_F = \sqrt{n}$, dies steht aber für $n > 1$ im Widerspruch zu $\|I\|_M = 1$ für eine induzierte Matrix–Norm $\|\cdot\|_M$. Deshalb kann die Frobenius–Norm $\|A\|_F$ durch keine Vektornorm $\|\underline{x}\|_V$ induziert sein.*

Für $B \in \mathbb{R}^{m\times n}$ folgt andererseits mit der Cauchy–Schwarz–Ungleichung (3.1)

$$\|B\underline{x}\|_2^2 \;=\; \sum_{k=1}^{m}\left(\sum_{\ell=1}^{n} b_{k\ell}x_\ell\right)^2 \;\le\; \sum_{k=1}^{m}\left(\sum_{\ell=1}^{n} b_{k\ell}^2\right)\left(\sum_{\ell=1}^{n} x_\ell^2\right) \;=\; \|B\|_F^2\|\underline{x}\|_2^2$$

und somit die Verträglichkeit der Frobenius–Norm $\|B\|_F$ zur Euklidischen Vektornorm $\|\underline{x}\|_2$.

Eine invertierbare Matrix $V \in \mathbb{R}^{n\times n}$ (beziehungsweise $U \in \mathbb{R}^{m\times m}$) heißt orthogonal, wenn ihre inverse Matrix V^{-1} durch die transponierte Matrix $V^\top$ gegeben ist, das heißt

$$V^\top V \;=\; VV^\top \;=\; I_n \in \mathbb{R}^{n\times n}, \quad U^\top U \;=\; UU^\top \;=\; I_m \in \mathbb{R}^{m\times m}.$$

Wegen

$$\|\underline{x}\|_2^2 \;=\; (\underline{x},\underline{x})_2 \;=\; (\underbrace{V^\top V}_{=\,I}\underline{x},\underline{x})_2 \;=\; (V\underline{x},V\underline{x})_2 \;=\; \|V\underline{x}\|_2^2$$

für beliebige Vektoren $\underline{x} \in \mathbb{R}^n$ folgt mit der Substitution $\underline{x} = V\underline{z}$

$$\|B\|_2 \;=\; \sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|B\underline{x}\|_2}{\|\underline{x}\|_2} \;=\; \sup_{\underline{0}\neq\underline{x}=V\underline{z}\in\mathbb{R}^n} \frac{\|BV\underline{z}\|_2}{\|V\underline{z}\|_2} \;=\; \sup_{\underline{0}\neq\underline{z}\in\mathbb{R}^n} \frac{\|BV\underline{z}\|_2}{\|\underline{z}\|_2} \;=\; \|BV\|_2.$$

Analog ergibt sich

$$\|B\|_2 \;=\; \sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|B\underline{x}\|_2}{\|\underline{x}\|_2} \;=\; \sup_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{\|UB\underline{x}\|_2}{\|\underline{x}\|_2} \;=\; \|UB\|_2\,.$$

Insgesamt gilt also für eine beliebige Matrix $B \in \mathbb{R}^{m\times n}$ und orthogonale Matrizen $V \in \mathbb{R}^{n\times n}$ beziehungsweise $U \in \mathbb{R}^{m\times m}$ die Gleichheit

$$\|B\|_2 \;=\; \|UB\|_2 \;=\; \|BV\|_2 \;=\; \|UBV\|_2, \tag{3.4}$$

das heißt die Euklidische Matrixnorm $\|B\|_2$ ist invariant bezüglich orthogonaler Transformationen.
Für $\ell = 1,\ldots,n$ bezeichne $\underline{b}^\ell = (b_{k\ell})_{k=1}^m$ die Spaltenvektoren der Matrix $B \in \mathbb{R}^{m\times n}$ mit der Euklidischen Vektornorm

$$\|\underline{b}^\ell\|_2^2 \;=\; \sum_{k=1}^{m} b_{k\ell}^2\,.$$

Damit ergibt sich für die Frobenius–Norm der Matrix B die Darstellung

$$\|B\|_F^2 \;=\; \sum_{k=1}^{m}\sum_{\ell=1}^{n} b_{k\ell}^2 \;=\; \sum_{\ell=1}^{n} \|\underline{b}^\ell\|_2^2\,.$$

Andererseits gilt für das Matrixprodukt UB mit einer orthogonalen Matrix $U \in \mathbb{R}^{m\times m}$

$$UB \;=\; \left(U\underline{b}^1,\ldots,U\underline{b}^n\right).$$

Aus der Invarianz der Euklidischen Vektornorm ergibt sich in der Frobenius–Norm

$$\|UB\|_F^2 \;=\; \sum_{\ell=1}^{n} \|U\underline{b}^\ell\|_2^2 \;=\; \sum_{\ell=1}^{n} \|\underline{b}^\ell\|_2^2 \;=\; \|B\|_F^2$$

und somit

$$\|UB\|_F \;=\; \|B\|_F\,.$$

Damit folgt auch, jeweils durch Übergang zur transponierten Matrix, für eine orthogonale Matrix $V \in \mathbb{R}^{n\times n}$

$$\|B\|_F \;=\; \|B^\top\|_F \;=\; \|V^\top B^\top\|_F \;=\; \|(V^\top B^\top)^\top\|_F \;=\; \|BV\|_F.$$

Insgesamt gilt für eine beliebige Matrix $B \in \mathbb{R}^{m\times n}$ und orthogonale Matrizen $V \in \mathbb{R}^{n\times n}$ und $U \in \mathbb{R}^{m\times m}$ die Gleichheit

$$\|B\|_F \;=\; \|UB\|_F \;=\; \|BV\|_F \;=\; \|UBV\|_F, \tag{3.5}$$

das heißt die Invarianz der Frobenius–Norm bezüglich orthogonaler Transformationen.
Ist eine quadratische Matrix $A \in \mathbb{R}^{n\times n}$ invertierbar, so definiert

$$\kappa_M(A) \;:=\; \|A\|_M \|A^{-1}\|_M \tag{3.6}$$

die Konditionszahl bezüglich der Matrixnorm $\|\cdot\|_M$. Insbesondere bezeichnet

$$\kappa_2(A) \;=\; \|A\|_2 \|A^{-1}\|_2 \tag{3.7}$$

die spektrale Konditionszahl. Eine Matrix $A \in \mathbb{R}^{n\times n}$ (beziehungsweise die Familie von Matrizen $A \in \mathbb{R}^{n\times n}$ für verschiedene $n \in \mathbb{N}$) heißt schlecht konditioniert, wenn ihre spektrale Konditionszahl $\kappa_2(A)$ proportional zur Dimension n anwächst.

3.2 Eigenwerte und Singulärwerte

Eine komplexe Zahl $\lambda(A) \in \mathbf{C}$ heißt Eigenwert der quadratischen Matrix $A \in \mathbb{R}^{n\times n}$, wenn das lineare Gleichungssystem

$$A\underline{x} \;=\; \lambda(A)\,\underline{x} \tag{3.8}$$

eine nicht triviale Lösung $\underline{x} \in \mathbb{R}^n$ mit $\|\underline{x}\|_V > 0$ besitzt. Diese heißt Eigenvektor zum Eigenwert $\lambda(A)$. Als notwendige Bedingung für die Existenz nichttrivialer Lösungen von (3.8) ergeben sich die μ voneinander verschiedenen Eigenwerte $\lambda_k(A)$ für $k = 1, \ldots, \mu \le n$ als Nullstellen des charakteristischen Polynoms

$$p(\lambda) \;:=\; \det(A - \lambda I) \;=\; (\lambda_1(A) - \lambda)^{\alpha_1} \ldots (\lambda_\mu(A) - \lambda)^{\alpha_\mu} \;=\; \prod_{k=1}^{\mu} (\lambda_k(A) - \lambda)^{\alpha_k}.$$

Die Potenzen $\alpha_k \in \mathbb{N}$ beschreiben die algebraische Vielfachheit des Eigenwertes $\lambda_k(A)$, und es gilt

$$\sum_{k=1}^{\mu} \alpha_k \;=\; n\,.$$

Durch Koeffizientenvergleich des charakteristischen Polynoms folgen

$$\mathrm{spur}(A) \;=\; \sum_{i=1}^{n} a_{ii} \;=\; \sum_{k=1}^{\mu} \alpha_k \lambda_k(A), \quad \det(A) \;=\; \prod_{k=1}^{\mu} [\lambda_k(A)]^{\alpha_k}.$$

Da ein Eigenwert $\lambda_k(A)$ Nullstelle des charakteristischen Polynoms $\det(A - \lambda I)$ ist, so ist auch sein konjugiert komplexer Wert $\overline{\lambda_k(A)}$ Nullstelle und somit Eigenwert von A. Wegen $\det(A - \lambda I) = \det(A^\top - \lambda I)$ sind diese auch Eigenwerte der transponierten Matrix $A^\top$. Die zum Eigenwert $\lambda_k(A)$ gehörenden Eigenvektoren bilden einen linearen Teilraum,

$$\mathcal{L}(\lambda_k(A)) \;:=\; \{\underline{x} \in \mathbb{R}^n \;:\; A\underline{x} = \lambda_k(A)\underline{x}\}\,, \quad \beta_k \;:=\; \dim \mathcal{L}(\lambda_k(A)),$$

dessen Dimension β_k die Anzahl der linear unabhängigen Eigenvektoren zum Eigenwert $\lambda_k(A)$ angibt. Diese heißt geometrische Vielfachheit des Eigenwerts $\lambda_k(A)$.
Durch

$$\varrho(A) \;:=\; \max_{k=1,\ldots,\mu \le n} |\lambda_k(A)|$$

wird schließlich der Spektralradius der Matrix A definiert.

Für symmetrische Matrizen $A = A^\top \in \mathbb{R}^{n\times n}$ sind die Eigenwerte $\lambda_k(A)$ für $k = 1, \ldots, n$ reell und die zugehörigen Eigenvektoren $\{\underline{v}^k\}_{k=1}^n$ bilden eine Orthonormalbasis mit

$$(\underline{v}^k, \underline{v}^\ell) \;=\; \delta_{k\ell} \quad \text{für alle } k, \ell = 1, \ldots, n.$$

Ein beliebiges Element $\underline{x} \in \mathbb{R}^n$ kann deshalb durch

$$\underline{x} \;=\; \sum_{k=1}^{n} \xi_k \underline{v}^k \quad \text{mit } \xi_k \;=\; (\underline{x}, \underline{v}^k) \tag{3.9}$$

dargestellt werden, und es gilt

$$\|\underline{x}\|_2^2 \;=\; (\underline{x}, \underline{x}) \;=\; (\sum_{k=1}^{n} \xi_k \underline{v}^k, \sum_{\ell=1}^{n} \xi_\ell \underline{v}^\ell) \;=\; \sum_{k=1}^{n}\sum_{\ell=1}^{n} \xi_k \xi_\ell (\underline{v}^k, \underline{v}^\ell) \;=\; \sum_{k=1}^{n} \xi_k^2$$

sowie

$$(A\underline{x}, \underline{x}) \;=\; \sum_{k=1}^{n}\sum_{\ell=1}^{n} \xi_k \xi_\ell (A\underline{v}^k, \underline{v}^\ell) \;=\; \sum_{k=1}^{n}\sum_{\ell=1}^{n} \xi_k \xi_\ell \lambda_k(A) (\underline{v}^k, \underline{v}^\ell) \;=\; \sum_{k=1}^{n} \lambda_k(A) \xi_k^2.$$

Eine symmetrische Matrix $A = A^\top \in \mathbb{R}^{n\times n}$ heißt positiv definit, falls alle Eigenwerte $\lambda_k(A)$ positiv sind. In diesem Fall folgt

$$(A\underline{x}, \underline{x}) \;=\; \sum_{k=1}^{n} \lambda_k(A)\xi_k^2 \;\geq\; \min_{k=1,\ldots,n} \lambda_k(A) \sum_{k=1}^{n} \xi_k^2 \;=\; \min_{k=1,\ldots,n} \lambda_k(A)\, \|\underline{x}\|_2^2$$

für alle $\underline{x} \in \mathbb{R}^n$. Weiterhin kann der Rayleigh–Quotient durch die extremalen Eigenwerte von A abgeschätzt werden, das heißt für alle $\underline{x} \in \mathbb{R}^n$ mit $\|\underline{x}\|_V > 0$ gilt

$$\min_{k=1,\ldots,n} \lambda_k(A) \;\leq\; \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})} \;\leq\; \max_{k=1,\ldots,n} \lambda_k(A).$$

Damit folgt

$$\lambda_{\min}(A) \;=\; \min_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})}, \quad \lambda_{\max}(A) \;=\; \max_{\underline{0}\neq\underline{x}\in\mathbb{R}^n} \frac{(A\underline{x}, \underline{x})}{(\underline{x}, \underline{x})}.$$

Gelten die Spektraläquivalenzungleichungen

$$c_1^A\,(\underline{x}, \underline{x}) \;\leq\; (A\underline{x}, \underline{x}) \;\leq\; c_2^A\,(\underline{x}, \underline{x}) \tag{3.10}$$

für alle $\underline{x} \in \mathbb{R}^n$ mit positiven Konstanten c_1^A und c_2^A, so folgt

$$c_1^A \;\leq\; \lambda_{\min}(A) \;\leq\; \lambda_{\max}(A) \;\leq\; c_2^A\,,$$

das heißt, die Konstanten c_1^A und c_2^A sind untere beziehungsweise obere Schranken der extremalen Eigenwerte der positiv definiten Matrix A.
Für eine symmetrische und positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ kann durch

$$\langle \underline{u}, \underline{v} \rangle_A \;:=\; (A\underline{u}, \underline{v}) \;=\; (\underline{u}, A\underline{v}) \;:\; \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{3.11}$$

das A–energetische Skalarprodukt erklärt werden. Die durch dieses Skalarprodukt induzierte Vektornorm

$$||\underline{x}||_A \;:=\; [\langle \underline{x}, \underline{x} \rangle_A]^{1/2} \;=\; (A\underline{x}, \underline{x})^{1/2} \tag{3.12}$$

wird als A–energetische Vektornorm bezeichnet.
Die durch die Eigenvektoren von $A = A^\top \in \mathbb{R}^{n\times n}$ gebildete Matrix

$$V \;=\; (\underline{v}^1, \ldots, \underline{v}^n) \in \mathbb{R}^{n\times n}$$

ist orthogonal, und es gilt

$$AV \;=\; \left(A\underline{v}^1, \ldots, A\underline{v}^n\right) \;=\; \left(\lambda_1(A)\underline{v}^1, \ldots, \lambda_n(A)\underline{v}^n\right) \;=\; VD$$

mit der durch die Eigenwerte von A definierten Diagonalmatrix

$$D \;=\; \operatorname{diag}\left(\lambda_k(A)\right)_{k=1}^n.$$

Multiplikation mit $V^\top$ von links ergibt wegen der Orthogonalität der Eigenvektoren

$$V^\top A V \;=\; D \tag{3.13}$$

beziehungsweise durch die Multiplikation mit $V^\top$ von rechts folgt die bekannte Faktorisierung der Matrix A,

$$A \;=\; V\,D\,V^\top \;=\; \sum_{k=1}^{n} \lambda_k(A)\,\underline{v}^k \underline{v}^{k,\top}. \tag{3.14}$$

Die Darstellung (3.14) ist einerseits Grundlage für die Definition einer Niedrig–Rang Approximation von A, andererseits ermöglicht sie die symmetrische Vorkonditionierung eines linearen Gleichungssystems $A\underline{x} = \underline{f}$ zur Verbesserung der spektralen Konditionszahl der vorkonditionierten Systemmatrix. Hierzu wird die Wurzel einer symmetrischen und positiv definiten Matrix A benötigt: Für positive Eigenwerte $\lambda_k(A) > 0$, $k = 1, \ldots, n$, kann die Diagonalmatrix

$$D^{1/2} \;=\; \mathrm{diag}\left(\sqrt{\lambda_k(A)}\right)_{k=1}^{n}$$

und somit die symmetrische und positiv definite Matrix

$$A^{1/2} \;=\; V D^{1/2} V^\top \tag{3.15}$$

erklärt werden. Nach Konstruktion gilt

$$A^{1/2}A^{1/2} \;=\; V D^{1/2} \underbrace{V^\top V}_{=\,I} D^{1/2} V^\top \;=\; V D V^\top \;=\; A\,.$$

Entsprechend kann

$$A^{-1/2} \;=\; (A^{1/2})^{-1} \;=\; V D^{-1/2} V^\top, \quad D^{-1/2} \;=\; \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_k(A)}}\right)_{k=1}^{n}$$

definiert werden. Mit der Transformation $\underline{x} = A^{-1/2}\underline{z}$ folgt aus den Spektraläquivalenzungleichungen (3.10) auch die Gültigkeit der Spektraläquivalenzungleichungen

$$\frac{1}{c_2^A}\,(\underline{z},\underline{z}) \;\le\; (A^{-1}\underline{z},\underline{z}) \;\le\; \frac{1}{c_1^A}\,(\underline{z},\underline{z}) \tag{3.16}$$

für alle $\underline{z} \in \mathbb{R}^n$.
Der Rang einer Matrix A beschreibt die Anzahl der linear unabhängigen Zeilen beziehungsweise Spalten von A. Die Darstellung (3.14) zeigt, daß der Rang einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ mit der Anzahl der nicht verschwindenden Eigenwerte zusammenfällt, das heißt es gilt

$$A \;=\; \sum_{k=1}^{\mathrm{rang}A} \lambda_k(A)\underline{v}^k \underline{v}^{k,\top},$$

falls eine entsprechende Nummerierung der Eigenwerte mit $\lambda_k(A) = 0$ für $k > \text{rang} A$ vorausgesetzt wird.
Aus der Norminvarianz (3.4) folgt schließlich

$$\|A\|_2 \;=\; \|VDV^\top\|_2 \;=\; \|D\|_2 \;=\; \max_{k=1,\ldots,n} |\lambda_k(A)| \;=\; \varrho(A),$$

beziehungsweise gilt mit der Invarianz (3.5) der Frobenius–Norm

$$\|A\|_F \;=\; \|VDV^\top\|_F \;=\; \|D\|_F \;=\; \sqrt{\sum_{k=1}^{n} [\lambda_k(A)]^2}.$$

Ist die Matrix A invertierbar, so sind die Eigenwerte der Inversen A^{-1} durch

$$\lambda_k(A^{-1}) = [\lambda_k(A)]^{-1}$$

gegeben. Ist A zusätzlich symmetrisch und positiv definit, so folgt für die spektrale Konditionszahl

$$\kappa_2(A) \;=\; \|A\|_2 \|A^{-1}\|_2 \;=\; \varrho(A)\varrho(A^{-1}) \;=\; \frac{\max\limits_{k=1,\ldots,n} |\lambda_k(A)|}{\min\limits_{k=1,\ldots,n} |\lambda_k(A)|} \;=\; \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Bei den obigen Überlegungen wurden Matrizen $A \in \mathbb{R}^{n\times n}$ betrachtet. Sei nun $B \in \mathbb{R}^{m\times n}$ eine beliebig gegebene Matrix mit $\text{rang}\, B \le \min\{m,n\}$. Dann definiert $A := B^\top B \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit $\text{rang}\, A \le \min\{m,n\}$ und n reellen Eigenwerten

$$\lambda_k(A) = \lambda_k(B^\top B)$$

sowie einem zugehörigen orthonormalen System $\{\underline{v}^k\}_{k=1}^n$ von Eigenvektoren. Dieses bildet eine Basis des $\mathbb{R}^n$, so daß jedes Element $\underline{x} \in \mathbb{R}^n$ wie in (3.9) dargestellt werden kann,

$$\underline{x} \;=\; \sum_{k=1}^{n} \xi_k \underline{v}^k \quad \text{mit } \xi_k \;=\; (\underline{x}, \underline{v}^k).$$

Wegen

$$\begin{aligned} 0 \;&\le\; \|B\underline{x}\|_2^2 \;=\; (B\underline{x}, B\underline{x}) \;=\; (B^\top B\underline{x}, \underline{x}) \;=\; (A\underline{x}, \underline{x}) \\ &=\; \sum_{k=1}^{n}\sum_{\ell=1}^{n} \xi_k\xi_\ell (A\underline{v}^k, \underline{v}^\ell) \;=\; \sum_{k=1}^{n}\sum_{\ell=1}^{n} \xi_k\xi_\ell \lambda_k(A)(\underline{v}^k, \underline{v}^\ell) \;=\; \sum_{k=1}^{n} \lambda_k(A)\,\xi_k^2 \end{aligned}$$

folgt $\lambda_k(A) \ge 0$ für alle $k = 1,\ldots,n$. Ohne Einschränkung der Allgemeinheit gelte $\lambda_k(A) > 0$ für alle $k = 1,\ldots,\mu = \text{rang}\, A \le \min\{m,n\}$ und $\lambda_k(A) = 0$ für $k = \mu+1,\ldots,n$. Nach (3.13) gilt die Faktorisierung

$$V^\top A V \;=\; V^\top B^\top B V \;=\; D \;=\; \text{diag}\,(\lambda_k(A))_{k=1}^n\,. \tag{3.17}$$

Wegen $\lambda_k(A) \geq 0$ für $k = 1, \ldots, \min\{m, n\}$ existieren die Singulärwerte

$$\sigma_k(B) \;=\; \sqrt{\lambda_k(A)} \;=\; \sqrt{\lambda_k(B^\top B)} \geq 0 \quad \text{für } k = 1, \ldots, \min\{m, n\}.$$

Insbesondere gilt $\sigma_k(B) > 0$ für $k = 1, \ldots, \mu \leq \min\{m, n\}$. Die Singulärwerte definieren eine Diagonalmatrix

$$\Sigma \;=\; \operatorname{diag}\left(\sigma_k(B)\right)_{k=1}^{\min\{m,n\}} \in \mathbb{R}^{m\times n} \tag{3.18}$$

und es gilt

$$D \;=\; \Sigma^\top \Sigma \in \mathbb{R}^{n\times n}\,.$$

Wird durch

$$\Sigma^+ \;=\; \begin{pmatrix} \frac{1}{\sigma_1(B)} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{\sigma_\mu(B)} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times m} \tag{3.19}$$

die Pseudoinverse zu Σ definiert, das heißt

$$\Sigma^+\Sigma \;=\; \begin{pmatrix} I_\mu & \\ & 0 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

dann folgt aus der Faktorisierung (3.17) durch Multiplikation mit der Pseudoinversen $\Sigma^{+,\top}$ von links

$$\Sigma^{+,\top} V^\top B^\top B V \;=\; \Sigma \in \mathbb{R}^{m\times n}$$

beziehungsweise

$$U^\top B V \;=\; \Sigma \tag{3.20}$$

mit

$$U \;=\; BV\Sigma^+ \in \mathbb{R}^{m\times m}.$$

Wegen

$$U^\top U \;=\; \Sigma^{+,\top} V^\top B^\top B V^\top \Sigma^+ \;=\; \Sigma^{+,\top} D \Sigma^+ \;=\; \begin{pmatrix} I_\mu & \\ & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}$$

ist $U^\top$ die Pseudoinverse zu U. Damit folgt aus (3.20) die Singulärwertzerlegung von $B \in \mathbb{R}^{m\times n}$,

$$B \;=\; U\Sigma V^\top \;=\; \sum_{k=1}^{\mu} \sigma_k(B)\, \underline{u}^k\, \underline{v}^{k,\top}, \tag{3.21}$$

das heißt $\mu = \operatorname{rang} B$ beschreibt die Anzahl der nicht verschwindenden Singulärwerte von B. Aus der Invarianz (3.4) der Euklidischen Matrixnorm folgt schließlich

$$\|B\|_2 = \|U\Sigma V^\top\|_2 = \|\Sigma\|_2 = \max_{k=1,\ldots,\mu} \sigma_k(B) = \max_{k=1,\ldots,\mu} \sqrt{\lambda_k(B^\top B)} = \sqrt{\varrho(B^\top B)}$$

beziehungsweise ist mit der Invarianz (3.5) der Frobenius–Norm

$$\|B\|_F \;=\; \|U\Sigma V^\top\|_F \;=\; \|\Sigma\|_F \;=\; \sqrt{\sum_{k=1}^{\mu}[\sigma_k(B)]^2}.$$

Multiplikation der Gleichung (3.20) von rechts mit $V^\top$ und Übergang zur Transponierten ergibt

$$B^\top U \;=\; V\Sigma$$

und somit folgt durch Vergleich der Spaltenvektoren

$$B^\top \underline{u}_k \;=\; \sigma_k(B)\,\underline{v}_k \quad \text{für } k = 1,\ldots,\min\{m,n\}.$$

Multiplikation der Gleichung (3.20) von links mit U liefert

$$BV \;=\; U\Sigma$$

und somit

$$B\underline{v}_k \;=\; \sigma_k(B)\,\underline{u}_k \quad \text{für } k = 1,\ldots,\min\{m,n\}.$$

3.3 Orthogonalisierung von Vektorsystemen

Für $m \in \mathbb{N}$ mit $m \le n$ heißt ein System[1] $\{\underline{w}^k\}_{k=0}^{m-1}$ von m nicht verschwindenden Vektoren $\underline{w}^k \in \mathbb{R}^n$, das heißt es gilt $\|\underline{w}^k\|_V > 0$, linear unabhängig, wenn die Gleichheit

$$\sum_{k=0}^{m-1} \alpha_k \underline{w}^k \;=\; \underline{0}$$

nur für die triviale Lösung

$$\alpha_0 = \cdots = \alpha_k = \cdots = \alpha_{m-1} = 0$$

erfüllt ist. Die Vektoren $\{\underline{w}^k\}_{k=0}^{m-1}$ heißen zueinander orthogonal bezüglich dem Skalarprodukt $\langle\cdot,\cdot\rangle$, falls

$$\langle \underline{w}^k, \underline{w}^\ell \rangle \;=\; 0 \quad \text{für alle } k,\ell = 0,\ldots,m-1 \text{ und } k \neq \ell$$

gilt, und orthonormal, wenn

$$\langle \underline{w}^k, \underline{w}^\ell \rangle \;=\; \delta_{k\ell} \quad \text{für alle } k,\ell = 0,\ldots,m-1$$

erfüllt ist. Für $m = n$ heißt das System $\{\underline{w}^k\}_{k=0}^{n-1}$ von n linear unabhängigen Vektoren Basis des $\mathbb{R}^n$, das heißt ein beliebiges Element $\underline{u} \in \mathbb{R}^n$ kann als Linearkombination der Basisvektoren $\{\underline{w}^k\}_{k=0}^{n-1}$ dargestellt werden.

[1] Im Hinblick auf die später beschriebenen Iterationsverfahren zur Lösung linearer Gleichungssysteme werden Vektorsysteme $\{\underline{w}^k\}_{k=0}^{n-1}$ stets von $k = 0,\ldots,n-1$ indiziert.

Beispiel 3.2. *Die Einheitsvektoren*

$$\underline{e}^k \;=\; \big(\delta_{(k+1)j}\big)_{j=1}^{n} \quad \textit{für } k = 0, \ldots, n-1$$

bilden eine Basis des $\mathbb{R}^n$*. Diese wird als kanonische Basis bezeichnet. Die Einheitsvektoren* $\underline{e}^k$ *sind orthonormal bezüglich dem Euklidischen Skalarprodukt, und für einen beliebigen Vektor* $\underline{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$ *gilt die Darstellung*

$$\underline{u} \;=\; \sum_{k=0}^{n-1} u_{k+1}\, \underline{e}^k \in \mathbb{R}^n.$$

Gegeben sei jetzt eine beliebige Basis $\{\underline{w}^k\}_{k=0}^{n-1}$ des $\mathbb{R}^n$, gesucht ist eine bezüglich dem Skalarprodukt $\langle\cdot,\cdot\rangle$ orthogonale Basis $\{\underline{p}^k\}_{k=0}^{n-1}$ mit

$$\langle \underline{p}^k, \underline{p}^\ell \rangle \;=\; 0 \quad \text{für } k, \ell = 0, \ldots, n-1 \text{ und } k \neq \ell.$$

Diese kann durch das Gram–Schmidtsche Orthogonalisierungsverfahren wie folgt konstruiert werden:

Setze
$\underline{p}^0 := \underline{w}^0$.
Für $k = 0, \ldots, n-2$ berechne
$$\underline{p}^{k+1} := \underline{w}^{k+1} - \sum_{\ell=0}^{k} \beta_{k\ell}\underline{p}^\ell, \quad \beta_{k\ell} = \frac{\langle \underline{w}^{k+1}, \underline{p}^\ell\rangle}{\langle \underline{p}^\ell, \underline{p}^\ell\rangle}.$$

Algorithmus 1.1: Orthogonalisierungsverfahren nach Gram–Schmidt.

Lemma 3.4. *Sei* $\{\underline{w}^k\}_{k=0}^{n-1}$ *ein System linear unabhängiger Vektoren. Dann ist das durch das Gram–Schmidtsche Orthogonalisierungsverfahren* (Algorithmus 1.1) *erzeugte Vektorsystem* $\{\underline{p}^k\}_{k=0}^{n-1}$ *orthogonal, das heißt es gilt*

$$\langle \underline{p}^k, \underline{p}^\ell \rangle \;=\; 0 \quad \textit{für } k, \ell = 0, \ldots, n-1,\ k \neq \ell$$

und

$$\langle \underline{p}^k, \underline{p}^k \rangle \;>\; 0 \quad \textit{für } k = 0, \ldots, n-1.$$

Beweis: Der Nachweis erfolgt durch vollständige Induktion nach k. Für $k = 0$ ist $\underline{p}^0 = \underline{w}^0$ und es gilt $\langle \underline{p}^0, \underline{p}^0\rangle > 0$. Dann ist $\underline{p}^1$ durch

$$\underline{p}^1 \;=\; \underline{w}^1 - \beta_{10}\underline{p}^0, \quad \beta_{10} \;=\; \frac{\langle \underline{w}^1, \underline{p}^0\rangle}{\langle \underline{p}^0, \underline{p}^0\rangle}$$

wohldefiniert, und die Orthogonalität folgt aus

$$\langle \underline{p}^1, \underline{p}^0\rangle \;=\; \langle \underline{w}^1 - \beta_{10}\underline{p}^0, \underline{p}^0\rangle \;=\; \langle \underline{w}^1, \underline{p}^0\rangle - \frac{\langle \underline{w}^1, \underline{p}^0\rangle}{\langle \underline{p}^0, \underline{p}^0\rangle}\langle \underline{p}^0, \underline{p}^0\rangle \;=\; 0.$$

Zu zeigen bleibt $\langle \underline{p}^1, \underline{p}^1 \rangle > 0$. Dieser Nachweis erfolgt **indirekt**, das heißt aus der Annahme $\langle \underline{p}^1, \underline{p}^1 \rangle = 0$ folgt

$$\underline{0} \;=\; \underline{p}^1 \;=\; \underline{w}^1 - \beta_{10}\underline{p}^0 \;=\; \underline{w}^1 - \beta_{10}\underline{w}^0$$

im Widerspruch zur linearen Unabhängigkeit der Vektoren $\underline{w}^0$ und $\underline{w}^1$.
Für $k = 1$ gelten somit die Induktionsvoraussetzungen

$$\langle \underline{p}^\ell, \underline{p}^j \rangle \;=\; 0 \quad \text{für alle } \ell, j = 0, \ldots, k \text{ mit } \ell \neq j$$

und

$$\langle \underline{p}^\ell, \underline{p}^\ell \rangle > 0 \quad \text{für alle } \ell = 0, \ldots, k.$$

Aus der Induktionsvoraussetzung für $k \in \mathbb{N}$ folgt durch Einsetzen der Koeffizienten β_{kj} für den Induktionsschritt $k+1$ die Orthogonalität

$$\langle \underline{p}^{k+1}, \underline{p}^j \rangle \;=\; \langle \underline{w}^{k+1}, \underline{p}^j \rangle - \sum_{\ell=0}^{k} \beta_{k\ell} \langle \underline{p}^\ell, \underline{p}^j \rangle \;=\; \langle \underline{w}^{k+1}, \underline{p}^j \rangle - \beta_{kj} \langle \underline{p}^j, \underline{p}^j \rangle \;=\; 0$$

für $j = 0, \ldots, k$. Zu zeigen bleibt $\langle \underline{p}^{k+1}, \underline{p}^{k+1} \rangle > 0$. Nach Konstruktion gilt

$$\underline{p}^\ell \in \operatorname{span}\left\{ \underline{w}^0, \ldots, \underline{w}^\ell \right\} \quad \text{für alle } \ell = 0, \ldots, k+1.$$

Die Annahme $\underline{p}^{k+1} = \underline{0}$ führt dann wegen

$$\underline{0} \;=\; \underline{p}^{k+1} \;=\; \underline{w}^{k+1} - \sum_{\ell=0}^{k} \beta_{k\ell} \underline{p}^\ell \;=\; \underline{w}^{k+1} - \sum_{\ell=0}^{k} \beta_{k\ell} \sum_{j=0}^{\ell} \alpha_{\ell j} \underline{w}^j$$

zum Widerspruch zur Voraussetzung der linearen Unabhängigkeit des Vektorsystems $\{\underline{w}^\ell\}_{\ell=0}^{k+1}$.
Damit ist Algorithmus 1.1 wohldefiniert. ∎

Sei $A \in \mathbb{R}^{n \times n}$ eine invertierbare Matrix mit $\operatorname{rang} A = n$. Dann bilden die Spaltenvektoren von A,

$$A \;=\; \left(\underline{a}^1, \ldots, \underline{a}^n \right) \in \mathbb{R}^{n \times n},$$

ein linear unabhängiges Vektorsystem $\{\underline{a}^k\}_{k=1}^n$. Die Anwendung des Orthogonalisierungsverfahrens nach Gram–Schmidt bezüglich dem Euklidischen Skalarprodukt und bei gleichzeitiger Normierung,

$$\underline{\hat{v}}^k = \underline{a}^k - \sum_{\ell=1}^{k-1} (\underline{a}^k, \underline{v}^\ell)\, \underline{v}^\ell, \quad \underline{v}^k = \frac{1}{\|\underline{\hat{v}}^k\|_2}\, \underline{\hat{v}}^k \quad \text{für } k = 1, \ldots, n,$$

liefert für die Spaltenvektoren von A die Darstellung

$$\underline{a}^k \;=\; \|\underline{\hat{v}}^k\|_2\, \underline{v}^k + \sum_{\ell=1}^{k-1} (\underline{a}^k, \underline{v}^\ell)\, \underline{v}^\ell \quad \text{für } k = 1, \ldots, n.$$

In Matrixschreibweise lautet diese

$$A = Q\,R \tag{3.22}$$

mit

$$Q = \left(\underline{v}^1, \ldots, \underline{v}^n\right) \in \mathbb{R}^{n\times n}, \quad Q^\top Q = I,$$

und der durch

$$R[\ell,k] = \begin{cases} (\underline{a}^k, \underline{v}^\ell) & \text{für } \ell = 1,\ldots,k-1, \\ \|\hat{\underline{v}}^k\|_2 & \text{für } \ell = k, \\ 0 & \text{für } \ell = k+1,\ldots,n. \end{cases}$$

für $k = 1,\ldots,n$ definierten oberen Dreiecksmatrix R. Durch das Orthogonalisierungsverfahren von Gram–Schmidt kann also die QR–Zerlegung (3.22) einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ berechnet werden.

Wird ein gegebenes linear unabhängiges Vektorsystem $\{\underline{w}^\ell\}_{\ell=0}^{n-1}$ bezüglich dem A–energetischen Skalarprodukt (3.11) orthogonalisiert, so nennt man das resultierende Vektorsystem $\{\underline{p}^\ell\}_{\ell=0}^{n-1}$ A–orthogonal beziehungsweise konjugiert, das heißt es gilt

$$\langle \underline{p}^k, \underline{p}^\ell \rangle_A = (A\underline{p}^k, \underline{p}^\ell) = (\underline{p}^k, A\underline{p}^\ell) = 0 \quad \text{für } k \neq \ell.$$