Approximation

Skript zur Vorlesung

Hans Joachim Oberle

Inhalt

1. Beste Approximationen

Die Problemstellung.

In der Approximationstheorie geht es darum, eine vorgegebene Funktion, die beispielsweise nur aufwendig ausgewertet werden kann, oder gewisse Daten einer komplizierten Funktion durch eine einfache Funktion zu approximieren. Die Zutaten eines Approximationsproblems sind also

- Eine Funktion oder gewisse Daten einer Funktion f; f ist Element eines linearen Raumes R.
- Eine Menge $V \subset R$ von approximierenden Funktionen.
  V könnte z.B. bestehen aus alle Geraden, oder allen Polynomen von einem gewissen Höchstgrad, oder aus allen kubischen Splinefunktionen, oder aus allen rationalen Funktionen von gewissen Höchstgraden für Zähler und Nenner. Die ersten genannten Beispiele liefern *lineare* Räume für V; man spricht dann von einer *linearen Approximationsaufgabe*. Das letzte Beispiel (rationale Funktionen) ergibt jedoch eine *nichtlineare Approximationsaufgabe*.
- Ein Abstandsbegriff; dies ist im Allg. eine Norm $\|\cdot\|$ für R.

Das Problem der Bestapproximation (1.1)

Gegeben sei ein reeller normierter Raum $(R, \|\cdot\|)$, eine nichtleere Teilmenge $V \subset R$ und ein Element $f \in R$. Gesucht ist eine Approximation $p^* \in V$ mit der Eigenschaft

$$\forall\, p \in V: \quad \| f - p^* \| \;\leq\; \| f - p \|. \tag{1.2}$$

Eine Approximation mit dieser Eigenschaft heißt eine *Bestapproximation* von f aus V. Die Differenz $e := f - p^*$ bezeichnet den *Fehler* der Approximation.

Es sei angemerkt, dass es in praktischen Anwendungen häufig nicht notwendig ist, die (oder eine) Bestapproximation in obigem Sinn zu bestimmen. Vielmehr genügt es, zu einer vorgegebenen Fehlerschranke $\varepsilon > 0$, eine approximierende Funktion p zu berechnen, für die $\| f - p \| \leq \varepsilon$ gilt. Natürlich ist i. Allg. nicht klar, wie groß ε gewählt werden kann und natürlich möchte man auch wissen, wie weit der Fehler einer berechnete Approximation von dem einer Bestapproximation abweicht.

Fragen, die im Zusammenhang mit einer allgemeinen Approximationsaufgabe zu beantworten sind:

1.) Gibt es zu vorgegebenen $(R, \|\cdot\|, f, V, \varepsilon)$ ein $p \in V$ mit $\| f - p \| \leq \varepsilon$?

2.) Gibt es ein minimales $p^* \in V$, d.h. eine Bestapproximation, die (1.2) erfüllt?

3.) Die Konvergenzfrage: Sei beispielsweise $V = V_N := \Pi_N$ der lineare Raum der reellen Polynome vom Grad kleiner oder gleich N. Sei ferner p_N^* die (als existent vorausgesetzte) Bestapproximation bezüglich einer geeigneten Norm. Gilt dann $\lim_{N\to\infty} p_N^* = f$?

4.) Welche numerischen Verfahren stehen zur Berechnung von p bzw. p^* zur Verfügung?

Wir beginnen mit zwei Beispielen.

Beispiel 1.3. Gegeben seien die Meßpunkte

t_k :	-1	0	1	2
y_k :	-2	0	2.5	3

Gesucht ist eine Gerade $y = x_1\, t \,+\, x_0$, die möglichst gut durch diese Meßpunkte geht. x_0 und x_1 sind die Unbekannten.

a) Der Abstand einer Geraden von den vorgegebenen Punkte wird im Sinn der *kleinsten Quadrate* gemessen $(m = 3)$:

$$F_2(x_0, x_1) \;=\; \left[\sum_{k=0}^{m} (x_0 + x_1\, t_k - y_k)^2\right]^{1/2} \;=\; \|A\, x \;-\; b\|_2 \tag{1.4}$$

mit

$$A \;:=\; \begin{pmatrix} 1 & t_0 \\ 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \end{pmatrix}, \quad b \;:=\; \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix}, \quad x \;:=\; \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}.$$

Die Approximationsaufgabe hat hier also die Gestalt eines *linearen Ausgleichsproblems*.

Aus der Numerik ist bekannt, dass dieses Problem – wegen $\mathrm{Rang}(A) = 2$ – eine eindeutig bestimmte Lösung x^* besitzt und sich diese mit Hilfe der Normalgleichungen

$$A^{\mathrm{T}} A\, x^* \;=\; A^{\mathrm{T}}\, b$$

(numerisch schlecht, da instabil!) oder (besser und stabil) mittels einer *Orthogonalzerlegung* (QR-Zerlegung, Householder oder Givens Transformation) berechnen lässt.

Für die obigen Daten liefert eine einfache Handrechnung über die Normalgleichungen die Lösung $x_0^* = 0,\ x_1^* = 1.75$. Die Fehlernormen sind

$$\|\, A\, x^* \;-\; b\,\|_2 \;=\; 0.93541\ldots \qquad \|\, A\, x^* \;-\; b\,\|_\infty \;=\; 0.75$$

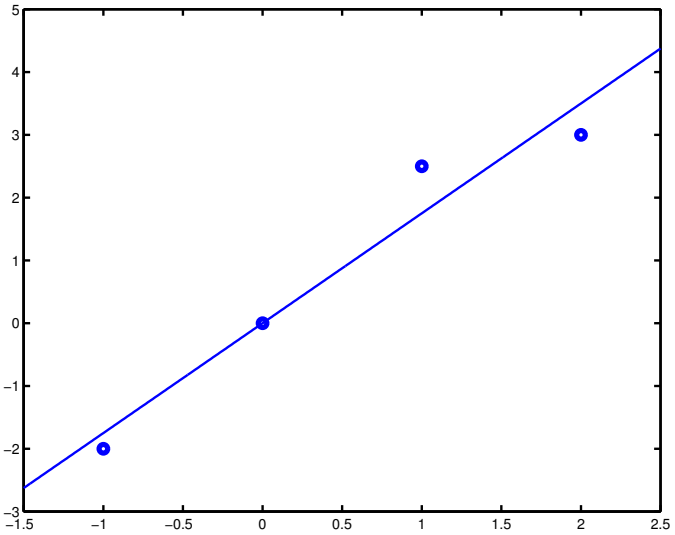


Abb. 1.1 Approximation im Sinn der L_2-Norm.

b) Wir messen den Abstand einer Geraden von den vorgegebenen Punkten in der Maximumsnorm, d.h. wir versuchen, den (betragsmäßig) größten Abstand von den Meßpunkten zu minimieren:

$$F_\infty(x_0, x_1) \;=\; \max\{\, |x_0 + x_1\, t_k - y_k| :\;\; k = 0, \ldots, m \,\} \;=\; \|A\, x \;-\; b\|_\infty. \qquad (1.5)$$

Man spricht hierbei von einem *Minimax Problem* bzw. von einer *Tschebyscheffschen Ausgleichsaufgabe*, benannt nach Pafnuti Lwowitsch Tschebyscheff (1821–1894).

Geometrisch macht man sich anhand der Abb.1.1 klar: Man muss die approximierende Gerade nun so legen, dass der maximale Fehler an *drei* der vorgegebenen Abszissen t_0, t_1, t_2, t_3 mit *alternierendem Vorzeichen* angenommen wird. Dies ist der Kern des später bewiesenen *Alternantensatzes*.

In unserem Beispiel sind dies die Abszissen t_0, t_2 und t_3. Bezeichnet δ den maximalen Fehler, so ergibt sich hiermit das folgende lineare Gleichungssystem in den Unbekannten x_0, x_1 und δ

$$\begin{aligned} y_0 \;-\; (x_0 \;+\; x_1\, t_0) \;&=\; \;\;\delta \\ y_2 \;-\; (x_0 \;+\; x_1\, t_2) \;&=\; -\delta \\ y_3 \;-\; (x_0 \;+\; x_1\, t_3) \;&=\; \;\;\delta. \end{aligned}$$

Mit den obigen Daten lautet die eindeutig bestimmte Lösung $\hat{x}_0 = 1/4$, $\hat{x}_1 = 5/3$, $\delta = -7/12$. Die Fehlernormen sind

$$\|\, A\,\hat{x} \;-\; b\,\|_2 \;=\; 1.040833\ldots \qquad \|\, A\,\hat{x} \;-\; b\,\|_\infty \;=\; 0.58333\ldots$$

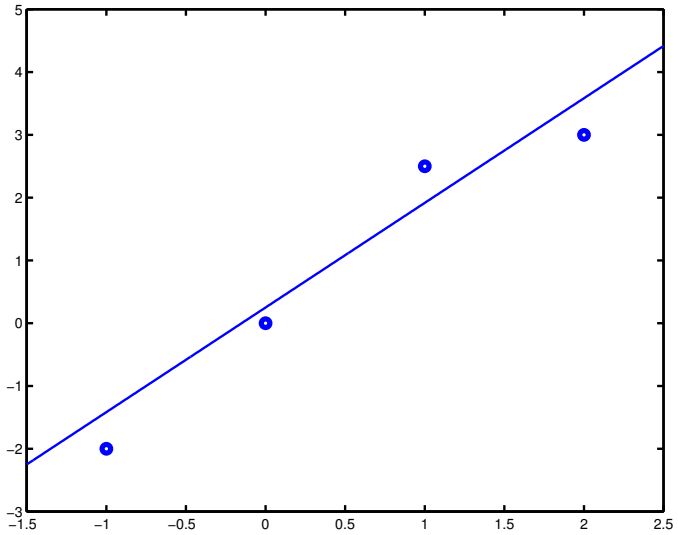


Abb. 1.2 Approximation im Sinn der Maximumsnorm.

Beispiel 1.6. Die Funktion $f(t) := \dfrac{\sin t}{t}$ soll im Intervall $[0, \pi/2]$ durch eine Parabel der Form $p(t) = x_0 + x_1 t^2$ approximiert werden.

Man beachte, dass f eine analytische Funktion in t^2 auf $\mathbb{R}$ ist und die Taylor-Entwicklung $f(t) = \sum_{k=0}^{\infty} (-1)^k \, t^{2k}/(2\,k+1)!$ besitzt.

a) *Taylor-Approximation*: Wir verwenden als Approximation die ersten beiden Summanden der Taylor-Reihe von f, $p_T(t) = 1 - t^2/6$. Für den Fehler der Approximation ergibt sich die Abschätzung $(0 < \Theta < 1)$:

$$|f(t) - p_T(t)| \;=\; \frac{1}{5!}\,|\,\sin^{(5)}(\Theta\,t)\,|\,t^4 \;\le\; \frac{1}{5!}\,\left(\frac{\pi}{2}\right)^4 \;\approx\; 0.050733.$$

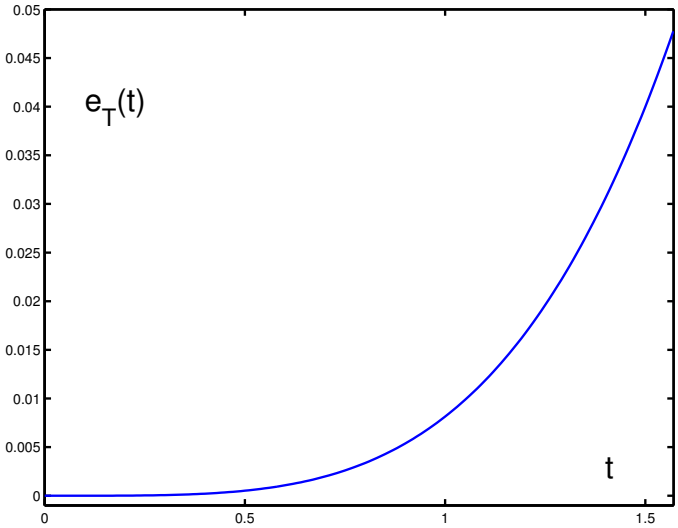


Abb. 1.3 Fehlerfunktion e_T für die Taylor-Approximation.

b) *L_2-Approximation*: Wir bestimmen $p_2^* \in V := \{p(t) = x_0 + x_1\, t^2 : \; x_i \in \mathbb{R}\}$ so, dass die Fehlerfunktion bzgl. der L_2-Norm

$$F_2(x_0, x_1) \;=\; \| f - p \|_2 \;=\; \left(\int_0^{\pi/2} (f(t) - p(t))^2 \, dt \right)^{0.5}$$

minimal wird.

Setzt man f und p hierin ein und multipliziert den Integranden aus, so ergibt sich

$$F_2(x_0, x_1)^2 \;=\; \int_0^{\pi/2} (\frac{\sin t}{t})^2 \, dt \;-\; 2\, x_0 \int_0^{\pi/2} \frac{\sin t}{t} \, dt \;-\; 2\, x_1 \int_0^{\pi/2} t \, \sin t \, dt$$

$$+ \;\; \frac{\pi}{2}\, x_0^2 \;+\; 2\, x_0\, x_1 \int_0^{\pi/2} t^2 \, dt \;+\; x_1^2 \int_0^{\pi/2} t^4 \, dt.$$

Die rechte Seite bildet eine quadratische Form in (x_0, x_1) mit einer symmetrischen und positiv definiten Koeffizientenmatrix der quadratischen Terme. Daher folgt, dass F_2 ein striktes globales Minimum $x^* \in \mathbb{R}^2$ besitzt, und, dass sich dieses aus der notwendigen Bedingung $\nabla F_2^2(x^*) = 0$ berechnen lässt.

Es ergibt sich das folgende lineare Gleichungssystem

$$\begin{pmatrix} \pi/2 & \int_0^{\pi/2} t^2 \, dt \\ \int_0^{\pi/2} t^2 \, dt & \int_0^{\pi/2} t^4 \, dt \end{pmatrix} \begin{pmatrix} x_0^* \\ x_1^* \end{pmatrix} \;=\; \begin{pmatrix} \int_0^{\pi/2} (\sin t)/t \, dt \\ \int_0^{\pi/2} t \, \sin t \, dt \end{pmatrix}$$

und hieraus die Lösung[1]

$$p_2^*(t) \;=\; x_1^*\, t^2 \;+\; x_0^*, \quad x_0^* \;\approx\; 0.9959262, \quad x_1^* \;\approx\; -0.1498806$$

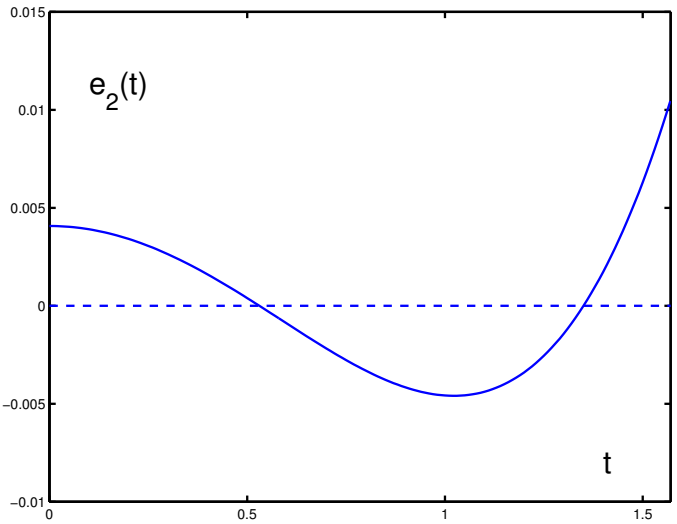


Abb. 1.4 Fehlerfunktion e_2 für die L_2-Approximation.

[1]Eine brauchbare Näherung für den Integralsinus findet man z.B. in Abramowitz, Stegun

c) *Tschebyscheff-Approximation*:

Wir bestimmen $p^*_\infty \in V := \{p(t) = x_0 + x_1\, t^2 : \; x_i \in \mathbb{R}\}$ so, dass die Fehlerfunktion bzgl. der Maximumsnorm

$$F_\infty(x_0, x_1) \;=\; \| f - p \|_\infty \;=\; \max\{\, | f(t) - p(t) | : \;\; 0 \le t \le \pi/2\}$$

minimal wird. Man beachte, dass F_∞ i. Allg. keine differenzierbare Funktion von $x_0,\; x_1$ ist.

Wir gehen analog zum diskreten Fall vor. Zunächst substituieren wir $\tau := t^2$. Damit ist die Funktion

$$\tilde{f}(\tau) \;:=\; f(\sqrt{\tau}) \;=\; \frac{\sin\sqrt{\tau}}{\sqrt{\tau}}, \quad 0 \;<\; \tau \;\le\; (\pi/2)^2$$

durch eine *Gerade* $\tilde{p}(\tau) := x_1\,\tau \;+\; x_0$ im Sinn kleinster Maximalabweichung zu approximieren. Wie im diskreten Fall gibt es dann drei Stellen, $\tau_0 = 0,\;\; \tau_1 \in]0, (\pi/2)^2[$ und $\tau_2 \;=\; (\pi/2)^2$, an denen die maximale Abweichung mit alternierendem Vorzeichen angenommen wird. Bezeichnet wieder δ die Maximalabweichung, so erhält man nach Rücktransformation auf die Variable t das folgende nunmehr nichtlineare Gleichungssystem in den Unbekannten $x_0,\;\; x_1,\;\; t_1$ und δ

$$\begin{aligned} f(0) - p(0) &= \;\;\delta \\ f(t_1) - p(t_1) &= -\delta \\ f(\pi/2) - p(\pi/2) &= \;\;\delta \\ f'(t_1) - p'(t_1) &= \;\;0 \end{aligned}$$

Die eindeutig bestimmte Lösung (numerisch mit dem Newton Verfahren bestimmt) lautet

$$\begin{aligned} x_0 \;&\approx\; 0.99419399, \qquad & x_1 \;&\approx\; -0.14727246, \\ \delta \;&\approx\; 0.58060118\text{E-}2, \qquad & t_1 \;&\approx\; 1.1023718. \end{aligned}$$

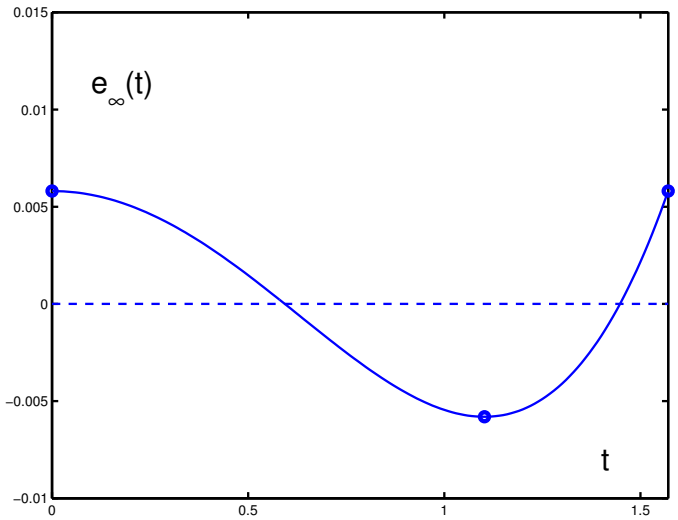


Abb. 1.5 Fehlerfunktion e_∞ für die Tschebyscheff-Approximation.

Wir kommen zur allgemeinen Problemstellung zurück und fassen nochmals zusammen.

Allgemeine Approximationsaufgabe (1.7)

Gegeben sei ein normierter linearer Raum $(R, \|\cdot\|)$ über $\mathbb{R}$ (manchmal auch über $\mathbb{C}$), eine nichtleere Teilmenge $V \subset R$ und ein zu approximierendes Element $f \in R$.

a) Die Zahl $d_V(f) := \inf\{\|f-p\| : p \in V\}$ heißt *Minimalabweichung* oder *Minimalabstand* von f zu V. Manchmal wird auch die Bezeichnung $\operatorname{dist}(f, V)$ verwendet.

b) Jede Abbildung $P : R \to V$ heißt ein *Approximationsoperator*. Ein solcher Operator heißt ein *Projektor*, falls $P \circ P = P$ gilt, er heißt ein *Projektor auf* V, falls $\forall\, p \in V : P(p) = p$ gilt.

c) Ein Element $p^* \in V$ mit $\|f - p^*\| = d_V(f)$ heißt *Minimallösung* oder *Bestapproximation* von f aus V.

Bemerkung (1.8)

Es gilt $d_V(f) = 0 \Leftrightarrow f \in \overline{V}$. Dabei bezeichnet $\overline{V}$ den topologischen Abschluss von V. In diesem Fall existiert nur dann eine Bestapproximation, falls $f \in V$.

Satz (1.9)

a) Die Funktion $d_V : R \to \mathbb{R}$ ist Lipschitz-stetig mit der Abschätzung

$$\forall\, f_1,\, f_2 \in R : \; |d_V(f_1) - d_V(f_2)| \;\le\; \|f_1 - f_2\|.$$

b) Ist V ein linearer Teilraum von R, so gelten die folgenden Halbnorm Eigenschaften $(f,\, f_1,\, f_2 \in R,\, \alpha \in \mathbb{R})$

$$\begin{aligned} &d_V(\alpha\, f) \;=\; |\alpha|\, d_V(f), \quad d_V(f_1 + f_2) \;\le\; d_V(f_1) \;+\; d_V(f_2), \\ &\forall\, p \in V : \; d_V(f + p) \;=\; d_V(f). \end{aligned}$$

Beweis: zu a): Zu $\varepsilon > 0$ existiert ein $p_\varepsilon \in V$ mit $\|f_2 - p_\varepsilon\| \le d_V(f_2) + \varepsilon$. Damit folgt

$$\begin{aligned} d_V(f_1) \;&\le\; \|f_1 - p_\varepsilon\| \;=\; \|f_1 - f_2 + f_2 - p_\varepsilon\| \\ &\le\; \|f_1 - f_2\| \;+\; \|f_2 - p_\varepsilon\| \\ &\le\; \|f_1 - f_2\| \;+\; d_V(f_2) + \varepsilon, \end{aligned}$$

und damit $d_V(f_1) - d_V(f_2) \le \|f_1 - f_2\| + \varepsilon$. Vertauschung der Rollen von f_1 und f_2 und Grenzwertbildung $\varepsilon \downarrow 0$ liefert die Behauptung.

zu b): Für $\alpha = 0$ ist die erste Ungleichung erfüllt. Für $\alpha \neq 0$ gilt

$$d_V(\alpha\, f) \;=\; \inf_{p \in V} \|\alpha\, f - p\| \;=\; |\alpha| \inf_{\tilde{p} \in V} \|f - \tilde{p}\| \;=\; |\alpha|\, d_V(f).$$

Zur zweiten Ungleichung:

$$\begin{aligned} d_V(f_1+f_2) &= \inf_{p\in V}\|f_1+f_2-p\| = \inf_{p_1,p_2\in V}\|f_1+f_2-p_1-p_2\| \\ &\le \inf_{p_1\in V}\|f_1-p_1\| + \inf_{p_2\in V}\|f_2-p_2\| = d_V(f_1)+d_V(f_2). \end{aligned}$$

Die letzte Relation schließlich ist unmittelbar klar. □

Geometrische Interpretation (1.10)

$d = d_V(f)$ ist der kleinste Radius einer Kugel $K_d(f) = \{g \in R : \|f-g\| < d\}$ mit der Eigenschaft $\forall\, r > d : K_r(f) \cap V \neq \emptyset$.

Zugleich ist d der größte Radius einer Kugel $K_d(f)$ mit der Eigenschaft $\forall\, r < d : K_r(f) \cap V = \emptyset$.

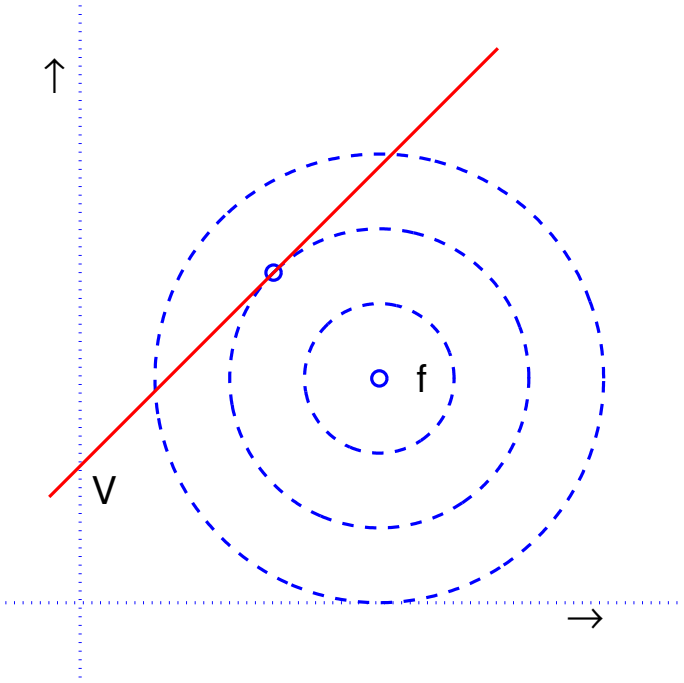


Abb. 1.6 Geometrische Interpretation.

Standardproblem (1.11)

In dieser Vorlesung wird häufig der folgende Standardfall betrachtet werden: $R := \mathrm{C}[a,b]$ mit einer der folgenden L_p–Normen

$$\|f\|_1 := \int_a^b |f(t)|\,dt, \quad \|f\|_2 := \left[\int_a^b f(t)^2\,dt\right]^{1/2}, \quad \|f\|_\infty := \max\{|f(t)| : a \le t \le b\}.$$

Als approximierende Funktionen werden beispielsweise die Polynomräume $V_N := \Pi_N$ (Polynome von Grad $\le N$) verwendet.

Die drei angegebenen Normen sind auf $\mathrm{C}[a,b]$ nicht äquivalent, es gelten aber die folgenden Abschätzungen.

Satz (1.12)

Für $f \in \mathrm{C}[a,b]$ gelten $\|f\|_1 \le \sqrt{b-a}\,\|f\|_2 \le (b-a)\,\|f\|_\infty$.

Beweis: Die Cauchy–Schwarzsche Ungleichung für das Standard–Skalarprodukt $\langle u,v\rangle := \int_a^b u(t)\,v(t)\,dt$ liefert die Abschätzung

$$\begin{aligned}\|f\|_1 &= \int_a^b |f(t)|\cdot 1\,dt \le \left(\int_a^b |f(t)|^2\,dt\right)^{1/2}\left(\int_a^b 1\,dt\right)^{1/2}\\ &= \sqrt{b-a}\,\|f\|_2 \le \sqrt{b-a}\left(\int_a^b \|f\|_\infty^2\,dt\right)^{1/2} = (b-a)\,\|f\|_\infty \qquad \square\end{aligned}$$

Der obige Satz zeigt, dass eine gute Approximation bzgl. $\|\cdot\|_\infty$ auch für die Normen $\|\cdot\|_1$ und $\|\cdot\|_2$ brauchbar ist.

Dass dies umgekehrt nicht der Fall zu sein braucht, zeigt etwa das folgende Beispiel:

$$R = \mathrm{C}[0,1], \quad f(t) = 1, \quad p_n(t) = t^{1/n}, \ n \in \mathbb{N}.$$

Man findet hierfür $\|f-p_n\|_1 = 1/(n+1)$, $\|f-p_n\|_2 = (2/(n^2+3n+2))^{1/2}$ und $\|f-p_n\|_\infty = 1$. Damit ist (p_n) also bzgl. $\|\cdot\|_1$ und $\|\cdot\|_2$ eine *Minimalfolge*, bzgl. der Maximumsnorm jedoch nicht.

Wir sehen uns im Folgenden noch zwei kritische Beispiele für Approximationen bzgl. $\|\cdot\|_1$ und $\|\cdot\|_\infty$ an.

Beispiele (1.13)

a) Sei $R := \mathrm{C}[-1,1]$ mit der Norm $\|\cdot\|_1$ und weiter $f(t) := 1$, sowie $V := \{p_\alpha : p_\alpha(t) = \alpha\,t, \quad \alpha \in \mathbb{R}\}$. Man findet

$$\|f-p_\alpha\|_1 = \int_{-1}^1 |1-\alpha\,t|\,dt = \begin{cases} 2, & -1 \le \alpha \le 1,\\ |\alpha| + 1/|\alpha|, & |\alpha| > 1.\end{cases}$$

Damit sind alle Geraden p_α, $-1 \le \alpha \le 1$ Bestapproximationen von f aus V. Es liegt keine Eindeutigkeit vor.

b) Das zweite Beispiel ist ähnlich aufgebaut: $R := \mathrm{C}[-1,1]$ mit der Norm $\|\cdot\|_\infty$, $f(t) := 1$, sowie $V := \{p_\alpha : p_\alpha(t) = \alpha\,(1+t), \quad \alpha \in \mathbb{R}\}$. Man findet

$$\|f-p_\alpha\|_\infty = \max_{-1\le t\le 1} |1-\alpha(1+t)| = \begin{cases} 1, & 0 \le \alpha \le 1,\\ |1-2\alpha|, & |\alpha-0.5| > 0.5.\end{cases}$$

Auch hier liegt demnach keine Eindeutigkeit der Bestapproximation vor.

2. Existenz, Eindeutigkeit und Stabilität

Vorgegeben sei eine Approximationsaufgabe für einen normierten $\mathbb{R}$–Vektorraum $(R, \|\cdot\|)$ und eine nichtleere Teilmenge $V \subset R$.

Existenz.

Wir geben zunächst zwei Sätze an, die die Existenz einer Bestapproximation, d.h. einer Lösung der obigen Approximationsaufgabe garantieren.

Satz (2.1) (Existenz I)

Ist V eine *kompakte* Teilmenge von R, so existiert eine Bestapproximation von V an $f \in R$.

Beweis: Die Abbildung $p \mapsto \|f - p\|$ ist stetig und nimmt daher auf V ein Minimum an. □

Satz (2.2) (Existenz II)

Ist V ein *endlich dimensionaler* linearer Teilraum von R, so existiert zu jeden $f \in R$ eine Bestapproximation.

Beweis: Die Menge $V_0 := \{p \in V : \ \|p\| \le 2\,\|f\|\}$ ist kompakt – als abgeschlossene und beschränkte Teilmenge eines endlich dimensionalen Teilraumes. Nach (2.1) existiert daher ein $p^* \in V$ mit $\|f - p^*\| \le \|f - p\|$, für alle $p \in V_0$.

Für die anderen $p \in V \setminus V_0$ gilt aber ebenfalls wegen $0 \in V$:

$$\|f - p\| \ \ge\ \|p\| - \|f\| \ >\ 2\,\|f\| - \|f\| \ =\ \|f - 0\| \ \ge\ \|f - p^*\|. \qquad \square$$

Unter weiteren Voraussetzungen an R lassen sich die Voraussetzungen an V abschwächen

Satz (2.3) (Existenz III)

Ist $(R, \langle\cdot,\cdot\rangle)$ ein reeller Hilbert-Raum und V ein *abgeschlossener* linearer Teilraum von R, so existiert zu jedem $f \in R$ eine Bestapproximation.

Beweis: Sei $(p_n) \in V^{\mathbb{N}}$ eine Minimalfolge, also $\|p_n - f\| \to d_V(f)$ $(n \to \infty)$. Mit der Parallelogrammgleichung $\|x + y\|^2 + \|x - y\|^2 \ =\ 2(\|x\|^2 + \|y\|^2)$ folgt dann

$$\begin{aligned} \|p_n \ -\ p_m\|^2 \ &=\ 2\,\|p_n - f\|^2 \ +\ 2\,\|p_m - f\|^2 \ -\ 4\|\frac{p_n + p_m}{2} - f\|^2 \\ &\le\ 2\big(\|p_n - f\|^2 \ +\ \|p_m - f\|^2\big) \ -\ 4\,d_V(f)^2. \end{aligned}$$

Damit ist (p_n) eine Cauchy-Folge und somit, da R vollständig ist, konvergent. Da V abgeschlossen ist, ist auch $p^* := \lim_{n\to\infty} p_n \in V$, und somit $\|p^* - f\| = d_V(f)$. □

Wir geben noch eine Kennzeichnung der Bestapproximation an für den Fall, dass R ein Hilbert-Raum ist:

Satz (2.4) (Orthogonalität)

Ist V ein *abgeschlossener* linearer Teilraum eines Hilbert-Raumes $(R, \langle\cdot,\cdot\rangle)$, so gilt für $f \in R \setminus V$ und $p^* \in V$: p^* Bestapproximation $\Leftrightarrow$ $e := f - p^* \perp V$.

Für $V \neq R$ ist somit aufgrund des Existenzsatzes (2.3) insbesondere $V^\perp \neq \{0\}$.

Beweis: Ist $p^* \in V$ Bestapproximation von f und $e := f - p^*$, so folgt für $p \in V \setminus \{0\}$ und $\alpha \in \mathbb{R}$:

$$\begin{aligned} \|f - p^*\|^2 &\leq \|f - p^* + \alpha p\|^2 \\ &= \|f - p^*\|^2 + 2\alpha\langle f - p^*, p\rangle + \alpha^2 \|p\|^2, \end{aligned}$$

also für alle α: $0 \leq 2\alpha\langle e, p\rangle + \alpha^2 \|p\|^2$.

Bildet man hier die Grenzwerte $\alpha \uparrow 0$ und $\alpha \downarrow 0$, so ergibt sich $\langle e, p\rangle = 0$.

Die Umkehrung folgt ebenfalls aus der obigen Relation mit $\alpha = 1$. □

Konvexität.

Um auf einfache Art zu Eindeutigkeitsaussagen zu kommen, werden häufig Konvexitätsannahmen verwendet.

Definition (2.5) Eine Teilmenge $S \subset R$ heißt *konvex*, falls

$$\forall\, x,\, y \in S : \forall\, \theta \in]0,1[: x + \theta\,(y - x) \in S.$$

Sie heißt *strikt konvex*, falls sogar

$$\forall\, x \neq y \in S : \forall\, \theta \in]0,1[: x + \theta\,(y - x) \in S^0.$$

Dabei bezeichnet S^0 das topologisch Innere von S.

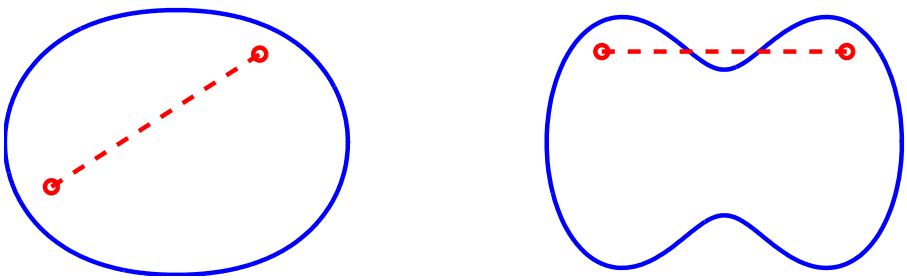


Abb.2.1. Konvexe und nicht konvexe Menge.

Satz (2.6) Offene oder abgeschlossene Normkugeln

$$K_r(f) := \{g \in R : \|g - f\| < r\}, \quad \overline{K}_r(f) := \{g \in R : \|g - f\| \leq r\}$$

sind stets konvex.

Beweis: (o.E.d.A. für offene Kugeln) Sind $f_0,\, f_1 \in K_r(f)$ und $\theta \in]0,1[$, so folgt

$$\begin{aligned} &\|(f_0 + \theta(f_1 - f_0)) - f\| = \|(1-\theta)(f_0 - f) + \theta(f_1 - f)\| \\ &\leq (1-\theta)\|f_0 - f\| + \theta\|f_1 - f\| < (1-\theta)r + \theta r = r. \quad \square \end{aligned}$$

Satz (2.7) Ist $V \subset R$ nichtleer und konvex, so ist auch die Menge $B(f, V)$ der Bestapproximationen von V an f konvex.

Beweis: Mit $r := d_V(f)$ ist $B(f, V) = V \cap \overline{K}_r(f)$ als Schnitt konvexer Mengen konvex. $\square$

Definition (2.8) Die Norm $\|\cdot\|$ des Raumes R heißt *strikt konvex*, falls die abgeschlossenen Einheitskugel $\overline{K}_1(0)$ strikt konvex ist, falls also gilt

$$\forall\, f, g : \; f \neq g \;\wedge\; \|f\|,\, \|g\| \leq 1 \;\wedge\; \theta \in]0,1[\;\Rightarrow\; \|f + \theta(g - f)\| < 1.$$

Man sagt dann auch, dass der Raum R *strikt normiert* sei. Es ist klar, dass in einem strikt normierten Raum dann auch *alle* Normkugeln (offen oder abgeschlossen) strikt konvexe Mengen sind.

Satz (2.9) Die folgenden Eigenschaften sind paarweise äquivalent

a) $\|\cdot\|$ ist strikt konvex,

b) $\|f + g\| = \|f\| + \|g\|,\; g \neq 0 \;\Rightarrow\; f = \alpha\, g,\; \alpha \geq 0,$

c) $\|f\| = \|g\| = 1,\; f \neq g \;\Rightarrow\; \|f + g\| < 2.$

Beweis:

a) $\Rightarrow$ c): Man setze $\theta := 0.5$.

c) $\Rightarrow$ b): Falls $f = 0$ ist, so gilt die Behauptung mit $\alpha = 0$. Es sei also $f \neq 0$. Ferner gelte: $\|f + g\| = \|f\| + \|g\|$ und es seien $\|g\|,\, \|f\| > 0$.

Dann folgt mit der Dreiecksungleichung und der Abschätzung $\|x - y\| \geq \|x\| - \|y\|$:

$$
\begin{aligned}
1 \;&\geq\; \frac{1}{2}\left\|\frac{f}{\|f\|}+\frac{g}{\|g\|}\right\| \;=\; \frac{1}{2}\left\|\left(\frac{f}{\|f\|}+\frac{g}{\|f\|}\right)-\left(\frac{g}{\|f\|}-\frac{g}{\|g\|}\right)\right\| \\
&\geq\; \frac{1}{2}\left(\left\|\frac{f}{\|f\|}+\frac{g}{\|f\|}\right\|-\left\|\frac{g}{\|f\|}-\frac{g}{\|g\|}\right\|\right) \\
&=\; \frac{1}{2}\left(\frac{\|f+g\|}{\|f\|}-\left(\frac{1}{\|f\|}-\frac{1}{\|g\|}\right)\|g\|\right) \\
&=\; \frac{1}{2}\left(\frac{\|f\|+\|g\|}{\|f\|}-\frac{\|g\|}{\|f\|}+\frac{\|g\|}{\|g\|}\right) \;=\; 1
\end{aligned}
$$

In dieser Ungleichungskette gilt damit durchgehend Gleichheit, insbesondere folgt aus der ersten Ungleichung

$$\left\|\frac{f}{\|f\|}+\frac{g}{\|g\|}\right\| \;=\; 2$$

Wegen c) muss daher $f/\|f\| \;=\; g/\|g\|$ sein, also $f=\alpha g$ mit $\alpha=\|f\|/\|g\|>0$.

b) $\Rightarrow$ a): Seien $\|f\|,\ \|g\|\leq 1,\ f\neq g$ und $\theta\in\,]0,1[$. Wegen

$$\|f+\theta(g-f)\| \;\leq\; (1-\theta)\|f\| \;+\; \theta\|g\|$$

gilt die Behauptung in a), falls $\|f\|<1$ oder $\|g\|<1$ ist.

Seien also $\|f\|=\|g\|=1$. Dann folgt

$$\|(1-\theta)f\| \;+\; \|\theta g\| \;=\; (1-\theta)\|f\| \;+\; \theta\|g\| \;=\; 1$$

und $\|(1-\theta)f \;+\; \theta g\| \;\leq\; 1$.

Wäre sogar $\|(1-\theta)f+\theta g\|=1$, so würde aus b) folgen, dass $(1-\theta)f=\alpha\theta g,\ \alpha\geq 0$.

Damit ist aber auch $f=\kappa g,\ \kappa\geq 0$ und aus $\|f\|=\|g\|=1$ folgt $\kappa=1$ und $f=g$, im Widerspruch zur Voraussetzung.

Daher folgt $\|(1-\theta)f \;+\; \theta g\| \;<\; 1$, was zu zeigen war. □

Bemerkung (2.10) Für den $\mathbb{R}^n$ sind die Normen $\|\cdot\|_1$ und $\|\cdot\|_\infty$ *nicht* strikt konvex, wohingegen $\|\cdot\|_2$ strikt konvex ist. Allgemein gilt

Satz (2.11) Euklidische bzw. unitäre Vektorräume $(R,\langle\cdot,\cdot\rangle)$ sind stets strikt konvex.

Beweis: Wir führen den Beweis für den reellen Fall und verwenden Satz (2.9).

Dazu seien $f,\, g \,\in\, R$ mit $f \neq g$ und $\|f\| \,=\, \|g\| \,=\, 1$. Für $\theta \in \mathbb{R}$ betrachten wir

$$\begin{aligned}\Phi(\theta) \;&:=\; \|f \,+\, \theta\,(g-f)\|^2 \;=\; \langle f \,+\, \theta\,(g-f), f \,+\, \theta\,(g-f)\rangle \\ &=\; \|f\|^2 \,+\, 2\,\theta\,\langle f, g-f\rangle \,+\, \theta^2\,\|f-g\|^2.\end{aligned}$$

Damit ist Φ bezüglich des Parameters θ eine nach oben geöffnete Parabel (höchster Koeffizient $\|f-g\|^2 > 0$) mit $\Phi(0) = \Phi(1) = 1$. Hiermit folgt

$$\forall\, \theta \,\in]0,1[:\;\; \Phi(\theta) \,<\, 1.$$

Speziell für $\theta = 1/2$ ergibt sich die Aussage aus Satz (2.9) c). □

Satz von Clarkson (2.12)

Die L_p-Räume $(L^p, \|\cdot\|_p)$ sind für $1 < p < \infty$ gleichmäßig konvex, d.h. zu $\varepsilon > 0$ existiert stets ein $\delta > 0$, so dass

$$\forall f, g \in L^p,\; \|f\| = \|g\| = 1 :\;\; \|0.5(f+g)\|_p \,>\, 1-\delta \Rightarrow\;\; \|f-g\|_p \,<\, \varepsilon.$$

Insbesondere sind damit nach Satz (2.9) c) die L^p-Normen $\|\cdot\|_p$, $1 < p < \infty$ auch strikt konvex.

Beweis: Siehe z.B. Hirzebruch, Scharlau.

Eindeutigkeit.

Satz (2.13) (Eindeutigkeit I)

Ist V strikt konvex, so existiert zu jedem $f \in R$ höchstens eine Bestapproximation von f aus V.

Beweis: O.E.d.A. sei der Minimalabstand $d := d_V(f) > 0$ positiv. Angenommen, $p_1 \neq p_2$ seien Bestapproximationen von f aus V, also $p_1,\, p_2 \in V$,

$$\|f \,-\, p_1\| \;=\; \|f \,-\, p_2\| \;=\; d.$$

Nach (2.7) ist dann auch $p^* \,=\, (p_1 + p_2)/2$ eine Bestapproximation von f aus V und wegen der strikten Konvexität zugleich ein innerer Punkt von V, also

$$\exists\, \varepsilon \in]0, d[:\;\; K_\varepsilon(p^*) \subset V, \quad \|f \,-\, p^*\| \;=\; d.$$

Für

$$q \;:=\; p^* + \frac{\varepsilon}{2\,d}\,(f - p^*) \;\in\; K_\varepsilon(p^*) \;\subset\; V$$

gilt dann

$$\|f \,-\, q\| \;=\; \big(1 \,-\, \frac{\varepsilon}{2\,d}\big)\;\|f \,-\, p^*\| \;<\; d.$$

Dies ist ein Widerspruch zur Minimalität von p^*. □

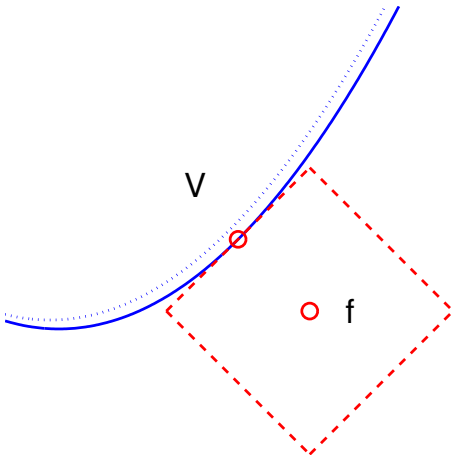


Abb. 2.2 Strikt konvexer Approximationsbereich.

Satz (2.14) (Eindeutigkeit II)

Ist V konvex, und die Norm $\|\cdot\|$ strikt konvex, so existiert zu jedem $f \in R$ höchstens eine Bestapproximation von f aus V.

Beweis: Sind $p_1 \neq p_2 \in V$ Bestapproximationen von f, so ist nach (2.7) auch $p := (p_1+p_2)/2 \in V$ eine Bestapproximation von f aus V. Nach Voraussetzung liegt diese jedoch im Innern der Normkugel $\overline{K}_d(f)$, $d = d_V(f)$, d.h. $\|f-p\| < d_V(f)$, im Widerspruch zum Minimalität von p_1 und p_2. □

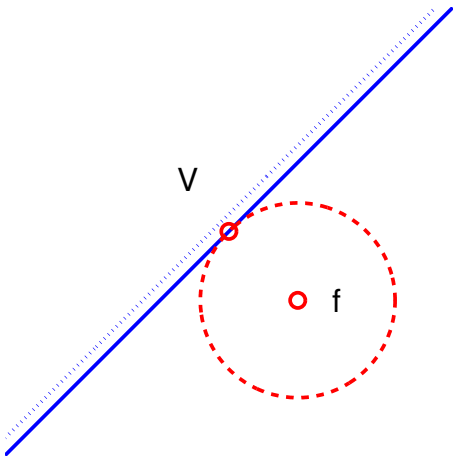


Abb. 2.3 Strikt konvexe Norm.

Folgerungen und Bemerkungen (2.15)

a) Ist R strikt normiert und V ein endlich dimensionaler linearere Teilraum von R, so existiert genau eine Bestapproximation von f aus V. Dies folgt aus (2.2) und (2.14). Das Gleiche gilt, falls R ein Hilbert-Raum und V ein abgeschlossener linearer Teilraum von R ist.

b) Aufgrund der Beispiele (1.13) aus Abschnitt 1 sind die Normen $\|\cdot\|_1$ und $\|\cdot\|_\infty$ auf $C[a,b]$ nicht strikt konvex.

c) Man kann zeigen: Ist R nicht strikt normiert, so gibt es ein $f \in R$ und einen endlich dimensionalen linearen Teilraum V von R, so dass f mehrere Bestapproximationen aus V besitzt; vgl. auch die Beispiele (1.13). Beweis im Skript von Geiger, Glashoff.

Stabilität.

Neben der Existenz und Eindeutigkeit einer Bestapproximation wird für Anwendungen benötigt, dass diese zumindest stetig von der zu approximierenden Funktion f abhängt. Wir bezeichnen diese Eigenschaft als *Stabilität* des Approximationsproblems.

Sei also wieder $(R, \|\cdot\|)$ ein normierter Raum, $V \subset R$ eine nichtleere Teilmenge. Wir nehmen an, dass zu jedem $f \in R$ eine eindeutig bestimmte Bestapproximation $p_V(f) \in V$ von f aus V existiert. Die Abbildung $p_V : R \to V$ heißt *metrische Projektion*.

Tatsächlich ist p_V ein *Projektor auf* V, d.h. es gilt

$$\forall p \in V : \quad p_V(p) = p.$$

Satz (2.16) (Stabilität)

Ist V kompakt, so ist die metrische Projektion p_V stetig.

Beweis: Für eine Folge $f_k \in R$ gelte $f_k \to f \in R\,(k \to \infty)$. Wir wollen hieraus folgern, dass auch $p_V(f_k) \to p_V(f)\,(k \to \infty)$ konvergiert. Wir führen den Beweis indirekt und nehmen an, dass die obige Konvergenz nicht gelte.

Dann gäbe es eine Teilfolge (f_{k_j}) und ein $\delta > 0$, so dass $\|p_V(f_{k_j}) - p_V(f)\| \geq \delta$ für alle $j \in \mathbb{N}$. Da V kompakt ist, besitzt die Folge $(p_V(f_{k_j}))$ eine weitere in V konvergente Teilfolge, die der Einfachheit halber ebenfalls mit $(p_V(f_{k_j}))$ bezeichnet werde und für die somit

$$p_V(f_{k_j}) \to p^* \in V\,(j \to \infty) \quad \text{und} \quad \|p_V(f_{k_j}) - p_V(f)\| \geq \delta > 0 \tag{2.17}$$

gelten.

Zu einem vorgegebenen $\varepsilon > 0$ sei nun $J(\varepsilon) \in \mathbb{N}$ so gewählt, dass für alle $j \geq J(\varepsilon)$ gelten

$$\|p_V(f_{k_j}) - p^*\| < \varepsilon/3 \quad \text{und} \quad \|f - f_{k_j}\| < \varepsilon/3. \tag{2.18}$$

Damit folgt nun für $j \geq J(\varepsilon)$

$$\begin{aligned}
\|f - p^*\| &\leq \|f - f_{k_j}\| + \|f_{k_j} - p_V(f_{k_j})\| + \|p_V(f_{k_j}) - p^*\| \\
&< \|f_{k_j} - p_V(f_{k_j})\| + 2\,\varepsilon/3 \\
&\leq \|f_{k_j} - p_V(f)\| + 2\,\varepsilon/3 \\
&\leq \|f_{k_j} - f\| + \|f - p_V(f)\| + 2\,\varepsilon/3 \\
&\leq \|f - p_V(f)\| + \varepsilon.
\end{aligned}$$

Erläuterung: Die erste Ungleichung gilt aufgrund der Dreiecksungleichung, die zweite wegen (2.18), die dritte Ungleichung folgt, da $p_V(f_{k_j})$ Bestapproximation von f_{k_j} aus V ist. Dann wird wieder die Dreiecksungleichung angewendet und schliesslich nochmals (2.18).

Aus der obigen Ungleichungskette folgt nun, da $\varepsilon > 0$ beliebig war, $\|f - p^*\| = \|f - p_V(f)\|$. Damit ist p^* Bestapproximation von f aus V und somit wegen der vorausgesetzten Eindeutigkeit $p^* = p_V(f)$.
Dies ist aber ein Widerspruch zu (2.17). □

Einschließung.

Wir beginnen mit einigen Vorbemerkungen über lineare Funktionale.

Es sei wie zuvor $(R, \|\cdot\|)$ ein reeller normierter Raum. Dann bildet die Menge R^* der stetigen, linearen Funktionale $\ell : R \to \mathbb{R}$ ebenfalls einen reellen Vektorraum, den so genannten *Dualraum* zu R

$$R^* := \{\, \ell \mid \quad \ell : R \to \mathbb{R} \quad \text{linear und stetig} \,\}. \tag{2.19}$$

Es sei daran erinnert, dass die Stetigkeit eines linearen Funktionals ℓ äquivalent ist zur *Beschränktheit* des Funktionals

$$\ell \text{ stetig} \;\Leftrightarrow\; \exists\, C > 0 :\; \forall f \in R :\; (\|f\| \leq 1 \;\Rightarrow\; |\ell(f)| \leq C). \tag{2.20}$$

Beweis:
$\Rightarrow$: Aus der Stetigkeit von ℓ folgt

$$\forall\, \varepsilon > 0 :\; \exists\, \delta > 0 :\; \|f\| < \delta \;\Rightarrow\; |\ell(f)| < \varepsilon.$$

Speziell für $\varepsilon = 1$ ergibt sich mit $\delta = \delta(1)$:

$$\forall\, f : \|f\| \leq 1 \;\Rightarrow\; |\ell(\frac{\delta}{2}\, f)| < 1 \;\Rightarrow\; |\ell(f)| \leq \frac{2}{\delta} =: C.$$

$\Leftarrow$: Aus $f_k \to f$ $(k \to \infty)$ folgt für $f_k \neq f$:

$$|\ell(f_k) - \ell(f)| = |\ell(f_k - f)| = \|f_k - f\|\, |\ell(\frac{f_k - f}{\|f_k - f\|})| \leq C\, \|f_k - f\| \to 0. \quad \square$$

Für stetige lineare Funktionale $\ell \in R^*$ ist daher der Ausdruck

$$\|\ell\| \;:=\; \sup_{f\neq 0} \frac{|\ell(f)|}{\|f\|} \;<\; \infty \tag{2.21}$$

endlich und hierdurch ist eine Norm $\|\cdot\|$ auf R^*, die so genannte *Operatornorm* definiert.

Der Dualraum eines reellen normierten Raumes $(R, \|\cdot\|)$ ist somit ebenfalls ein reeller normierter Raum. Er ist sogar vollständig, d.h. ein Banach-Raum, vgl. Hirzebruch, Scharlau.

Beispiele (2.22)

a) Auf dem Raum der stetigen Funktionen $(\mathrm{C}[a,b], \|\cdot\|_\infty)$ mit der Maximumsnorm $\|f\|_\infty := \max\{|f(t)| : a \le t \le b\}$ ist $\ell(f) := \int_a^b f(t)\,dt$ ein stetiges lineares Funktional mit

$$|\ell(f)| \;=\; |\int_a^b f(t)\,dt| \;\le\; (b-a)\,\|f\|_\infty.$$

Da hierin für $f = 1$ Gleichheit gilt, folgt $\|\ell\| = (b-a)$.

b) Für $(\mathrm{C}[a,b], \|\cdot\|_\infty)$ und einem festen Punkt $t_0 \in [a,b]$ ist das *Punktfunktional* $\ell(f) := f(t_0)$ ein stetiges lineares Funktional mit $\|\ell\| = 1$.

c) In Verallgemeinerung hiervon ist zu einer festen Zerlegung $a \le t_0 < \ldots < t_m \le b$ und gegebenen Koeffizienten $\lambda_i \in \mathbb{R}$ auch durch

$$\ell(f) \;:=\; \sum_{i=0}^{m} \lambda_i\, f(t_i), \quad f \in \mathrm{C}[a,b], \tag{2.23}$$

ein stetiges lineares Funktional $\ell \in \mathrm{C}[a,b]^*$ gegeben mit $\|\ell\|_\infty = \sum_{i=0}^m |\lambda_i|$. Auch hierbei spricht man von einem Punktfunktional.

d) Auf $(\mathrm{C}^1[a,b], \|\cdot\|_\infty)$ ist durch $\ell(f) := f'(t_0)$ ein lineares Funktional gegeben, $a < t_0 < b$. Dieses ist jedoch *nicht stetig*!

O.E.d.A. sei $t_0 = 0$. Für $\varepsilon > 0$ betrachten wir die C^1-Funktion

$$f(t,\varepsilon) \;:=\; \arctan(t/\varepsilon).$$

Hierfür ergibt sich $\|f(\cdot,\varepsilon)\|_\infty \le \pi/2$ und $\ell(f(\cdot,\varepsilon)) = f'(0,\varepsilon) = 1/\varepsilon$. Damit folgt

$$\frac{|\ell(f(\cdot,\varepsilon))|}{\|f(\cdot,\varepsilon)\|_\infty} \;\ge\; \frac{2}{\varepsilon\,\pi} \;\to\; \infty \quad (\varepsilon \downarrow 0).$$

Die Existenz stetiger linearer Funktionale lässt sich mit Hilfe des folgenden Satzes von Hahn-Banach zeigen:

Satz (2.24) (Hahn, Banach)
Jedes stetige lineare Funktional $\ell_0 \in U^*$ auf einem linearen Teilraum U von R besitzt eine stetige, lineare Fortsetzung auf R, $\ell \in R^*$, mit gleicher Norm: $\ell|_U = \ell_0$, und $\|\ell\| = \|\ell_0\|_U$.

Beweis: Siehe z.B. Hirzebruch, Scharlau.

Ist der zugrunde liegende Raum ein Hilbert-Raum $(R, \langle\cdot,\cdot\rangle)$, so ist der Dualraum $(R^*, \|\cdot\|)$ isometrisch (und bijektiv) zum Ausgangesraum. Dies ist Inhalt des folgenden Rieszschen Darstellungssatzes.

Satz (2.25) (Riesz)
Sei $(R, \langle\cdot,\cdot\rangle)$ ein Hilbert-Raum. Jedem $x \in R$ wird dann durch $\lambda_x(y) := \langle x, y\rangle$ ein stetiges lineares Funktional $\lambda_x \in R^*$ zugeordnet. Die hierdurch definierte Abbildung $\lambda : R \to R^*$ ist eine Isometrie (linear, bijektiv und normerhaltend). Insbesondere gibt es zu jedem stetigen linearen Funktional $\ell \in R^*$ genau ein $x \in R$ mit $\ell = \langle x, \cdot\rangle$.

Beweis:

Dass $\lambda_x : R \to \mathbb{R}$ eine lineare Abbildung ist, folgt unmittelbar aus den Skalarprodukteigenschaften. Mittels der Cauchy-Schwarzschen Ungleichung folgt weiterhin:

$$|\lambda_x(y)| \;=\; |\langle x, y\rangle| \;\leq \|x\|\,\|y\|.$$

Damit ist klar, dass λ_x stetig ist und dass $\|\lambda_x\| \leq \|x\|$ gilt. Speziell für $y := x$ ergibt sich $|\lambda_x(x)| = \|x\|^2$. Damit folgt schließlich $\|\lambda_x\| = \|x\|$.

Die Abbildung $\lambda : R \to R^*$ ist also wohldefiniert, normerhaltend und auch linear, wie man ebenfalls unmittelbar den Skalarprodukteigenschaften entnimmt.

Ist $x \in \mathrm{Kern}(\lambda)$, also $\lambda_x = 0$, so folgt $\langle x, y\rangle = 0$ für alle $y \in R$. Dann ist aber auch $x = 0$. Der Kern von λ ist also trivial und somit ist λ injektiv.

Es bleibt zu zeigen, dass λ auch surjektiv ist.

Sei dazu $\ell \in R^*$ und $V := \mathrm{Kern}(\ell)$. Da ℓ stetig ist, ist V ein *abgeschlossener* linearer Teilraum von R und für $\ell \neq 0$ ist auch $V \neq R$.

Sei nun gemäß Satz (2.4) $z \in R \setminus \{0\}$ mit $z \perp V$. Wir setzen $x := \alpha\, z$ mit $\alpha := \ell(z)/\|z\|^2$. Für $y \in R$ folgt dann:

$$\langle x, y\rangle \;=\; \langle \alpha\, z, y\rangle \;=\; \langle \alpha\, z, \left(y - \frac{\ell(y)}{\ell(z)}\, z\right) \;+\; \frac{\ell(y)}{\ell(z)}\, z\rangle$$

$$=\; 0 \;+\; \frac{\ell(y)}{\ell(z)}\, \alpha\, \langle z, z\rangle \;=\; \ell(y),$$

wobei der erste Summand verschwindet, da $\left(y - \frac{\ell(y)}{\ell(z)}\, z\right) \in V$ und $z \perp V$.

Damit ist gezeigt, dass $\ell = \lambda_x$ gilt, also λ surjektiv ist. $\square$

Definition (2.26)

Sei nun wieder $(R, \|\cdot\|)$ ein reeller normierter Raum und $V \subset R$ ein nichtleerer linearer Teilraum von R. Wir sagen, ein stetiges lineares Funktional $\ell \in R^*$ ist *orthogonal* zu V, falls $\forall\, y \in V : \ell(y) = 0$. Die Menge aller zu V orthogonalen stetigen linearen Funktionale bildet offensichtlich einen linearen Teilraum von R^*. Dieser wird mit $V^\perp$ bezeichnet. $V^\perp$ heißt der *Orthogonalraum* zu V.

Im Fall eines reellen Hilbert-Raumes sind nach dem Rieszschen Satz alle stetigen linearen Funktionale von der Form $\ell = \lambda_x,\ x \in R$. Die Orthogonalität zu einem linearen Teilraum V,

$$0 \;=\; \ell(y) \;=\; \lambda_x(y) \;=\; \langle x, y\rangle, \quad \forall\, y \in V,$$

stimmt daher mit dem üblichen Orthogonalitätsbegriff, $x \perp V$, überein.

Mit Hilfe des Orthogonalraumes gelangen wir nun zu einer *unteren Schranke* für die Minimalabweichung $d_V(f)$ einer Approximationsaufgabe. Eine obere Schranke erhalten wir dagegen sehr leicht durch irgendein Element $p \in V$, da ja immer $d_V(f) \le \|f - p\|$ gilt.

Satz (2.27) (Dualität)

Sei $(R, \|\cdot\|)$ ein reeller normierter Raum, $f \in R$ und V ein linearer Teilraum von R.

a) Ist $\ell \in V^\perp$ und $\|\ell\| \le 1$, so folgt $|\ell(f)| \le d_V(f)$.

b) Es gibt ein stetiges lineares Funktional $\ell_f \in V^\perp$ mit $\|\ell_f\| \le 1$ und $|\ell_f(f)| = d_V(f)$.

Anmerkungen (2.28)

a) ℓ_f heißt wegen der obigen Eigenschaften auch ein *maximales (stetiges, lineares) Funktional*. Ist $d_V(f) > 0$, also $f \notin \overline{V}$, so ist auch $\|\ell_f\| = 1$.

b) Gilt für ein stetiges, lineares Funktional $\ell \in V^\perp,\ \ell \neq 0$ und ein $p \in V$

$$|\ell(f)| \;=\; \|\ell\|\,\|f - p\,\|,$$

so ist p eine Bestapproximation von f aus V und $\ell/\|\ell\|$ ist ein maximales lineares Funktional.

c) Zur Bestimmung der Minimalabweichung $d_V(f)$ kann man aufgrund des obigen Dualitätssatzes anstelle der eigentlichen Approximationsaufgabe auch das folgende *duale Approximationsproblem* lösen:

$$\text{Maximiere die Funktion } |\ell(f)| \text{ über } \ell \in V^\perp,\ \|\ell\| \le 1. \tag{2.29}$$

d) Ist $(R, \langle\cdot,\cdot\rangle)$ ein Hilbert-Raum, so lässt sich mit Hilfe des Rieszschen Satzes das duale Approximationsproblem folgendermaßen formulieren: Maximiere $|\langle f, g\rangle|$ über $g \in R$ unter den Nebenbedingungen $\|g\| \le 1$ und $\forall\, p \in V:\ \langle g, p\rangle = 0$.

Beweis zu (2.27)

zu a) Ist $\ell \in V^\perp,\ \|\ell\| \le 1$ und $p \in V,$ so folgt

$$|\ell(f)| \;=\; |\ell(f) - \ell(p)| \;=\; |\ell(f-p)| \;\le\; \|\ell\|\,\|f-p\,\| \;\le\; \|f-p\,\|,$$

und somit auch $|\ell(f)| \le d_V(f)$.

zu b) Im Fall $d_V(f) = 0$ erfüllt $\ell_f := 0$ die Behauptung. Sei also im Folgenden $d_V(f) > 0$. Wir betrachten den linearen Teilraum $U := \mathrm{Spann}(V \cup \{f\})$. Jedes $g \in U$ besitzt dann eine *eindeutige* Darstellung der Form

$$g \;=\; \alpha_g\, f \,+\, p_g, \quad \alpha_g \in \mathbb{R}, \quad p_g \in V.$$

Hiermit zeigt man nun direkt (Übungsaufgabe): Die Abbildung $\ell_0 : U \to \mathbb{R}$, $\ell_0(g) := \alpha_g\, d_V(f)$ ist ein stetiges lineares Funktional auf U mit den Eigenschaften:

$$\ell_0\big|_V \;=\; 0, \quad \|\ell_0\|_{U^*} \;=\; 1, \quad \ell_0(f) \;=\; d_V(f).$$

Die Behauptung ergibt sich hieraus nun mit Hilfe des Satzes von Hahn und Banach (2.24). □

Beispiel (2.30)

Sei $R := \mathrm{C}[a,b],\ \|\cdot\| = \|\cdot\|_\infty,\ V = \mathrm{Spann}(p_0, \ldots, p_n)$, wobei die $p_j \in R$ linear unabhängig seien. Schließlich sei $f \in \mathrm{C}[a,b] \setminus V$.

Das Approximationsproblem lautet somit: Bestimme $x = (x_0, \ldots, x_n) \in \mathbb{R}^{n+1}$ mit

$$\|f \;-\; \sum_{j=0}^{n} x_j\, p_j\|_\infty \quad \text{minimal.}$$

Für das duale Approximationsproblem schränken wir uns auf Punktfunktionale

$$\ell(q) \;=\; \sum_{j=0}^{N} \lambda_j\, q(t_j), \quad q \in \mathrm{C}[a,b]$$

ein, wobei $a \le t_0 < t_1 < \ldots < t_N \le b$ ein geeignetes Gitter sei. A priori ist natürlich nicht gesichert, dass die Einschränkung auf Punktfunktionale zulässig ist. Nach (2.22)c) gilt für ein solches Funktional

$$\|\ell\| \;=\; \sum_{j=0}^{N} |\lambda_j|, \quad \ell \in V^\perp \;\Leftrightarrow\; \forall\, k = 0, 1, \ldots, n:\ \sum_{j=0}^{N} \lambda_j\, p_k(t_j) \;=\; 0.$$

Damit lautet das duale Approximationsproblem:

Man bestimme eine Unterteilung $a \leq t_0 < t_1 < \ldots < t_N \leq b$ und Koeffizienten $\lambda_0, \ldots, \lambda_N \in \mathbb{R}$, so dass

$$\Big| \sum_{j=0}^{N} \lambda_j \, f(t_j) \Big|$$

maximiert wird unter den Nebenbedingungen

$$\sum_{j=0}^{N} |\lambda_j| \;=\; 1, \quad \forall\, k = 0, 1, \ldots, n : \; \sum_{j=0}^{N} \lambda_j \, p_k(t_j) \;=\; 0.$$

3. Approximationsoperatoren

Allgemeines.

Wir betrachten wieder eine Approximationsaufgabe für einen normierten $\mathbb{R}$–Vektorraum $(R, \|\cdot\|)$ und nehmen nun an, dass die zur Approximation verwendete Menge einen linearen Teilraum $V \subset R$ bildet.

Wie schon in Abschnitt 1 erwähnt wurde, heißt jede Abbildung $P : R \to V$ ein *Approximationsoperator*. Erfüllt P die Projektoreigenschaft

$$\forall\, p \in V : \quad P(p) \;=\; p, \tag{3.1}$$

so heißt P *ein Projektor von R auf V*.

Ist P ein *stetiger, linearer* Operator, so wird durch

$$\|P\| \;:=\; \sup_{f \neq 0} \frac{\|P(f)\|}{\|f\|} \tag{3.2}$$

die *Operatornorm* von P definiert. Hierdurch ist eine Norm auf dem Raum $L(R, V)$ der stetigen, linearen Approximationsoperatoren definiert.

Beispiel (3.3) Sei $R := \mathrm{C}[0,1]$ und $V := \Pi_1[0,1]$, das ist der lineare Teilraum der Polynomfunktionen auf $[0,1]$ vom Höchstgrad eins. Der Operator P sei ein Interpolationsoperator, definiert durch

$$P(f)(t) \;:=\; f(0) \,+\, t\,(f(1) - f(0)).$$

Bzgl. der L_2-Norm ist dieser Operator unbeschränkt. Man untersuche dazu z.B. die Funktionenfolge $f_n(t) = t^n$ für $n \to \infty$. Man findet $\|P(f_n)\|_2 = 1/\sqrt{3}$ und $\|f_n\| = 1/\sqrt{2\,n+1}$.

Für die Maximumsnorm dagegen ergibt sich

$$\|P(f)\|_\infty \;=\; \max\left(|f(0)|, |f(1)|\right) \;\leq\; \|f\|_\infty$$

Da hierbei aber auch Gleichheit vorkommen kann (z.B. für affin-lineares f), erhalten wir $\|P\|_\infty = 1$.

Die folgende *Abschätzung für den Approximationsfehler* ist benannt nach Henry Léon Lebesgue (1875–1941).

Satz (3.4) (Lemma von Lebesgue)

Für einen stetigen linearen Projektor $P: R \to V$ auf V gilt

$$\forall\, f \in R: \ \|f - P(f)\| \ \leq \ (1 + \|P\|)\, d_V(f).$$

Beweis: Für $p \in V$ gilt

$$\begin{aligned} \|f - P(f)\| &= \|(f-p) + P(p-f)\| \\ &\leq \|f-p\| \ + \ \|P(p-f)\| \\ &\leq \big(1 + \|P\|\big)\, \|f-p\|. \end{aligned}$$

Die Infimumsbildung bzgl. $p \in V$ liefert die Behauptung. □

Bemerkungen (3.5)

a) Wegen des obigen Zusammenhangs heißt $\|P\|$ auch die *Lebesgue-Konstante* des Approximationsoperators P.

b) Ein kleiner Wert von $\|P\|$ bedeutet eine gute Approximationseigenschaft von $P(f)$ an f.

c) Für das Beispiel (3.3) mit der Maximumsnorm $\|\cdot\|_\infty$ erhalten wir aus dem Lemma von Lebesgue die Abschätzung

$$\|f - P(f)\|_\infty \ \leq \ 2 \min_{p \in \Pi_1} \|f - p\|_\infty$$

Diese Abschätzung ist auch scharf, denn für $f(t) := t^2$ gilt

$$\begin{aligned} P(f)(t) &= t, & \|f - P(f)\|_\infty &= 1/4, \\ p^*(t) &= t - 1/8, & \|f - p^*\|_\infty &= d_{\Pi_1}(f) = 1/8. \end{aligned}$$

Anwendung (3.6)

Zu einer vorgegebenen stetigen Funktion $f \in \mathrm{C}[a,b]$ und einer Fehlerschranke $\varepsilon > 0$ sei ein Polynom $p \in \Pi_n[a,b]$ gesucht mit $\|f - p\| \leq \varepsilon$. Dabei sei n a priori nicht bekannt und soll möglichst klein gewählt werden (vgl. auch den Weierstraßschen Approximationssatz). Seien nun weiter $P_n: \mathrm{C}[a,b] \to \Pi_n[a,b]$ lineare Projektoren mit bekannten Lebesgue-Konstanten $\|P_n\|_\infty$.

Gilt dann für ein vorgebenes $n \in \mathbb{N}$:

$$\|f - P_n(f)\| \ > \ \big(1 + \|P_n\|\big)\, \varepsilon,$$

so gibt es *kein* Polynom $p \in \Pi_n$ mit $\|f - p\|_\infty \leq \varepsilon$. Es bleibt also nichts anderes übrig, als n zu vergrößern.

Polynominterpolation.

Wir wiederholen einige Grundtatsachen der Interpolation durch Polynome. Es sei eine Funktion $f \in \mathrm{C}[a,b]$ vorgegeben und es sei $n \in \mathbb{N}_0$.

Satz (3.7) (Lagrange) Zu $(n+1)$ verschiedenen Punkten $t_i \in [a,b]$, $i = 0,\ldots,n$, existiert genau ein Polynom $p \in \Pi_n[a,b]$ mit $p(t_i) = f(t_i), i = 0,\ldots,n$. Dieses ist gegeben durch $p = \sum_{k=0}^{n} f(t_k)\,\ell_k$, wobei die Lagrange-Polynome ℓ_k definiert sind durch

$$\ell_k(t) \;:=\; \prod_{j=0, j\neq k}^{n} (t-t_j)/(t_k-t_j), \qquad k = 0,\ldots,n.$$

Beweis: Es gilt $\ell_k \in \Pi_n[a,b]$ mit $\ell_k(t_i) = \delta_{i\,k}$. Daher ist auch $p = \sum_{k=0}^{n} f(t_k)\,\ell_k \in \Pi_n$ und es gilt $p(t_i) = f(t_i),\ i = 0,\ldots,n$.

Sind $p,\ \tilde{p} \in \Pi_n[a,b]$ Polynome mit $p(t_i) = \tilde{p}(t_i) = f(t_i)$, so ist $(p-\tilde{p})$ ein Polynom höchstens n-ten Grades mit (wenigstens) $(n+1)$ Nullstellen; damit folgt $p = \tilde{p}$. □

Bei festen Interpolationsknoten $t_0,\ldots,t_n \in [a,b]$ ist durch $f \mapsto p$ eine lineare und stetige (bzgl. $\|\cdot\|_\infty$) Abbildung $P_n : \mathrm{C}[a,b] \to \Pi_n[a,b]$ erklärt.

P_n heißt der *Interpolationsoperator* (bzgl. Polynominterpolation) zu den Knoten $t_0,\ldots,t_n$.

P_n ist offenbar auch ein Projektor auf $\Pi_n[a,b]$, so dass das Lemma von Lebesgue angewendet werden kann. Hiernach gilt für jede stetige Funktion $f \in \mathrm{C}[a,b]$

$$\|f - P_n(f)\|_\infty \;\leq\; \big(1 + \|P_n\|_\infty\big)\, d_{\Pi_n}(f). \tag{3.8}$$

Satz (3.9) (Interpolationsfehler) Für $f \in \mathrm{C}^{n+1}[a,b]$ lässt sich der Interpolationsfehler $e_n(t) := f(t) - P_n(f)(t)$ folgendermaßen darstellen

$$e_n(t) \;=\; \frac{f^{(n+1)}(\tau)}{(n+1)!} \;\prod_{j=0}^{n} (t-t_j).$$

Dabei ist $\tau = \tau(t) \in \,]\min(t,t_0,\ldots,t_n), \max(t,t_0,\ldots,t_n)[$ eine (unbekannte) Zwischenstelle.

Beweis: Für $t \in [a,b] \setminus \{t_0,\ldots,t_n\}$ betrachte man die Hilfsfunktion

$$g(x) \;:=\; f(x) \;-\; p(x) \;-\; \big(f(t)-p(t)\big) \prod_{j=0}^{n} \frac{x-t_j}{t-t_j}.$$

g hat damit die $(n+2)$ Nullstellen $t, t_0,\ldots,t_n$. $(n+1)$ malige Differentiation und

die Anwendung des Satzes von Rolle ergibt

$$0 \;=\; g^{(n+1)}(\tau) \;=\; f^{(n+1)}(\tau) \;-\; \big(f(t)-p(t)\big)\,\frac{(n+1)!}{\prod_{j=0}^{n}(t-t_j)}. \qquad \square$$

Beispiel (3.10) (Carl Runge; 1856–1927)

Die Funktion $f(t) := 1/(1+t^2)$ soll im Intervall $[-5,5]$ interpoliert werden. Wir wählen (zunächst) äquidistante Stützstellen $t_j = -5 + 10\,j/n$, $j = 0,\ldots,n$. In Abb. 3.1 ist das Interpolationspolynom (gestrichelt) und die Ausgangsfunktion für $n = 6$ aufgetragen. Man erkennt, dass der Fehler insbesondere in der Nähe der Intervallenden groß ist.

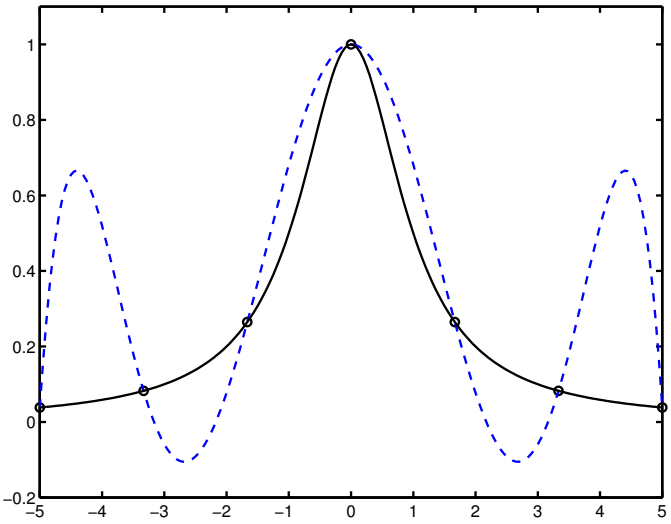


Abb. 3.1 Beispiel von Runge, $n = 6$.

Dass dieses Fehlerverhalten sich für größere Werte von n noch verschärft, kann man der folgenden Tabelle aus Powell entnehmen. Hier ist der Interpolationsfehler im Punkte $t_{n-1/2} = 0.5\,(t_{n-1} + t_n)$ für verschiedene Werte von n aufgetragen.

n	$f(t_{n-1/2})$	$p(t_{n-1/2})$	$e_n(t_{n-1/2})$
2	0.137931	0.759615	−0.621684
4	0.066390	−0.356826	0.423216
6	0.054463	0.607879	−0.553416
8	0.049651	−0.831017	0.880668
10	0.047059	1.578721	−1.531662
12	0.045440	−2.755000	2.800440
14	0.044334	5.332743	−5.288409
16	0.043530	−10.173867	10.217397
18	0.042920	20.123671	−20.080751
20	0.042440	−39.952449	39.994889

Der Grund für das Verhalten liegt im starken Anwachsen des Knotenpolynoms $\omega_n(t) := \prod_{j=0}^{n}(t - t_j)$ in der Fehlerdarstellung (3.9).

Eine Idee zur Verbesserung des Fehlerverhaltens ist daher, die Knoten t_j so zu wählen, dass die Maximumsnorm $\|\omega_n\|_\infty$ des Knotenpolynoms möglichst klein wird. Es ist dazu offenbar sinnvoll, zu verlangen, dass die Maxima/Minima von ω_n auf $[a, b]$ sämtlich gleichen Betrag haben. Dies führt auf die Tschebyscheff-Polynome:

Definition (3.11) Die Funktionen $T_n(x) := \cos(n \arccos x)$, $-1 \le x \le 1$, $n \in \mathbb{N}_0$, genügen der Dreiterm-Rekursion

$$\begin{aligned} T_{k+1}(x) &= 2\,x\,T_k(x) - T_{k-1}(x), \quad k \in \mathbb{N}, \\ T_0(x) &= 1, \quad T_1(x) = x. \end{aligned} \tag{3.12}$$

Sie sind daher Polynome, $T_n \in \Pi_n$, $\operatorname{grad} T_n = n$, und heißen Tschebyscheff-Polynome erster Art[1] .

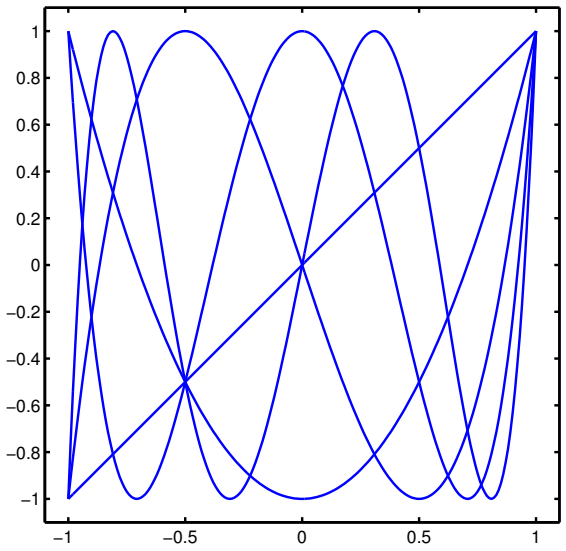


Abb. 3.2 Tschebyscheff Polynome T_n, $n = 1 : 1 : 5$.

Satz (3.13) (Tschebyscheff-Polynome)

Die Tschebyscheff-Polynome T_n besitzen folgende Eigenschaften

a) $T_{n+1}(x) = 2\,x\,T_n(x) - T_{n-1}(x)$, $n \in \mathbb{N}_0$, $x \in \mathbb{R}$; $T_0(x) = 1$; $T_1(x) = x$

b) $T_{n+1} \in \Pi_{n+1}$, $\quad T_{n+1}(x) = 2^n\,x^{n+1} + \ldots$

c) $|\,T_{n+1}(x)\,| \le 1$, $\quad -1 \le x \le 1$

d) Nullstellen von T_{n+1}: $\quad x_k = \cos\left(\frac{1+2(n-k)}{2(n+1)}\,\pi\right)$, $k = 0, \ldots, n$

e) Extremalstellen von T_{n+1} auf $[-1, 1]$: $\quad x_k^E = \cos\left(\frac{n+1-k}{n+1}\,\pi\right)$, $k = 0, \ldots, n+1$,
mit $\quad T_{n+1}(x_k^E) = (-1)^{n+1-k}$, $k = 0, \ldots, n+1$, $\quad$ und
$-1 = x_0^E < x_1^E < \ldots < x_{n+1}^E = 1$.

[1] Pafnuti Lwowitsch Tschebyscheff, 1821 – 1894

Wir haben also mit T_{n+1} ein Polynom gefunden, das auf $[-1,1]$ lauter Maxima und Minima mit gleichen Betrag und alternierendem Vorzeichen besitzt. Wir skalieren noch auf ein beliebiges Intervall $[a,b]$ und höchsten Koeffizienten 1:

$$\begin{aligned} x \in [-1,1]; \quad t &= a + \frac{x+1}{2}(b-a), \\ \omega_n(t) &= \frac{1}{2^n} T_{n+1}(x) = \frac{1}{2^n} T_{n+1}\big(2\,\frac{t-a}{b-a} - 1\big). \end{aligned} \tag{3.14}$$

Satz (3.15) (Minimax Eigenschaft)

Das skalierte Tschebyscheff-Polynom $\omega_n \in \Pi_{n+1}[a,b]$ aus (3.14) minimiert die Maximumsnorm $\|p\|_\infty := \max_{t\in[a,b]} |p(t)|$ über alle normierten Polynome $p(t) = t^{n+1} + \ldots \in \Pi_{n+1}[a,b]$.

Beweis: Gäbe es ein normiertes Polynom $p(t) = t^{n+1} + \ldots \in \Pi_{n+1}[a,b]$ mit $\|p\|_\infty < 1/2^n$, so wäre $q := \omega_n - p \in \Pi_n[a,b]$ ein Polynom vom Höchstgrad n mit $q(t_k^E) > 0$ für alle $k \in \{0,\ldots,n+1\}$ mit $n-k$ ungerade, und $q(t_k^E) < 0$ für alle $k \in \{0,\ldots,n+1\}$ mit $n-k$ gerade. Dabei sind t_k^E die gemäß (3.14) skalierten Extremalstellen von T_{n+1}. Nach dem Zwischenwertsatz hat q daher wenigstens $(n+1)$ Nullstellen, muss also identisch verschwinden. □

Die gesuchten *Tschebyscheff–Knoten* sind also gerade die transformierten Nullstellen von T_{n+1}, also

$$t_k = a + \frac{x_k+1}{2}(b-a), \quad x_k = \cos\big(\frac{1+2(n-k)}{2(n+1)}\,\pi\big), \quad k = 0,\ldots,n. \tag{3.16}$$

Beispiel (3.17) (Runge) Wir gehen nochmal auf das Rungesche Beispiel ein und wählen nunmehr Tschebyscheff-Knoten. Dabei legen wir das Intervall $[a,b]$ so fest, dass $t_0 = -5$ und $t_n = 5$ gelten. In Abb. 3.3 ist das Resultat für $n = 6$ wiedergegeben, in der Abb. 3.4 ist die Fehlerfunktion $e_n = f - p_n$ für $n = 20$ aufgezeichnet.

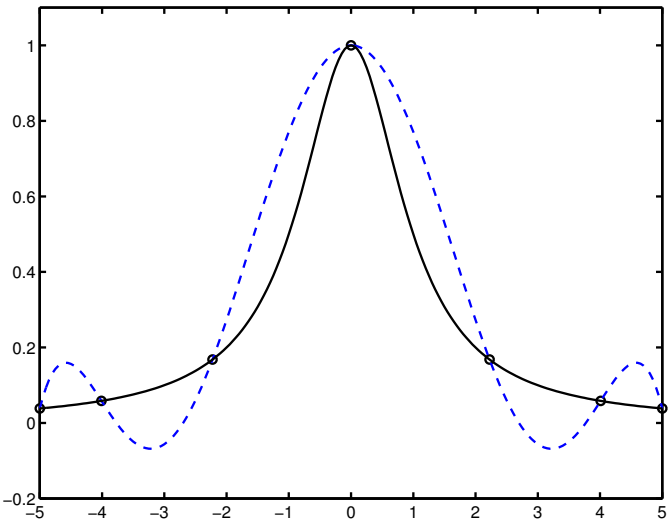


Abb. 3.3 Beispiel von Runge, Tschebyscheff-Knoten, $n = 6$.

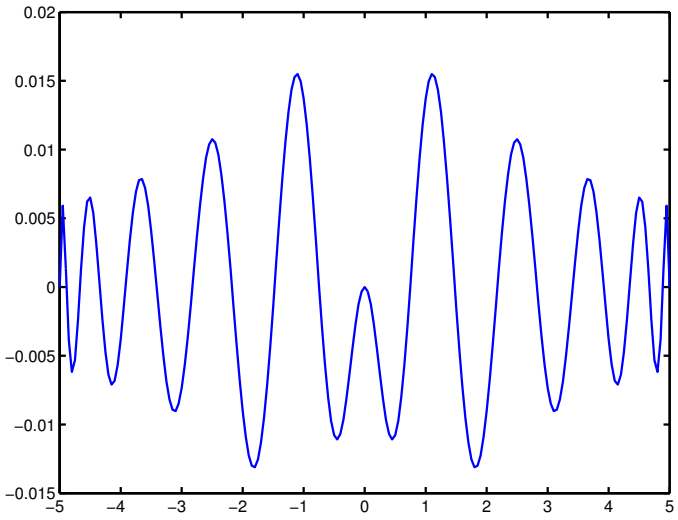


Abb. 3.4 Beispiel von Runge, Tschebyscheff-Knoten, $n = 20$.

Wir erkennen in diesem Beispiel, dass die Approximationsgüte für Tschebyscheff-Knoten deutlich besser ist, als für den Fall äquidistanter Knoten, ja dass sogar möglicherweise Konvergenz $\|e_n\|_\infty \to 0$ für $n \to \infty$ vorliegt.

Um allgemein zu Konvergenzaussagen zu gelangen, könnte man von der Fehlerdarstellung (3.9) ausgehen und zeigen, dass $\|f^{n+1}\|_\infty/(n+1)!$ weniger stark wächst als 2^n (im Fall der Tschebyscheff-Knoten). Alternativ könnte man auch das Lebesguesche Lemma (3.4) bzw. (3.8) heranziehen. Nach dem Weierstraßschen Approximationssatz gilt $\lim\limits_{n\to\infty} d_{\Pi_n}(f) = 0$. Würde demnach die Operatornorm $\|P_n\|_\infty$ beschränkt bleiben für $n \to \infty$, so würde auch der Interpolationsfehler $\|f - P_n(f)\|_\infty$ nach (3.4) gegen Null konvergieren.

Satz (3.18) Für den Interpolationsoperator $P_n : \mathrm{C}[a,b] \to \Pi_n[a,b]$ zu festen Knoten $a \le t_0 < \ldots < t_n \le b$ gilt

$$\|P_n\|_\infty \;=\; \max\Big\{ \sum_{k=0}^{n} |\ell_k(t)| \;:\; a \le t \le b \Big\},$$

dabei bezeichnet ℓ_k das k-te Lagrange-Polynom.

Beweis:
$$\begin{aligned}
\|P_n\|_\infty &= \sup\big\{\|P_n(f)\|_\infty : \; f \in \mathrm{C}[a,b],\; \|f\|_\infty \le 1\big\} \\
&= \sup\Big\{ \max_{a\le t\le b} \big| \sum_{k=0}^{n} f(t_k)\,\ell_k(t)\big| : \; f \in \mathrm{C}[a,b],\; \|f\|_\infty \le 1\Big\} \\
&= \max\Big\{ \sum_{k=0}^{n} |\ell_k(t)| : \; a \le t \le b\Big\}.
\end{aligned}$$

Bei der letzten Umformung ist zunächst $\le$ klar, es lässt sich aber auch ein f konstruieren, für das Gleichheit gilt! □

Bemerkungen (3.19)

a) Die Operatornorm $\|P_n\|$ gibt an, um wieviel der Interpolationsfehler vom kleinstmöglichen Approximationsfehler $d_{\Pi_n}(f)$ abweicht, vgl. (3.8). Dabei hängt, wie die folgende Tabelle zeigt, $\|P_n\|$ wesentlich von der Knotenwahl ab. $\|P_n^{(1)}\|$ bezieht sich dabei auf äquidistante Knoten im Intervall $[-5, 5]$, $\|P_n^{(2)}\|$ auf Tschebyscheff-Knoten mit $t_0 = -5$ und $t_n = 5$.

n	$\|P_n^{(1)}\|$	$\|P_n^{(2)}\|$
2	$0.12500e+01$	$0.12500e+01$
4	$0.22078e+01$	$0.15702e+01$
6	$0.45493e+01$	$0.17825e+01$
8	$0.10946e+02$	$0.19416e+01$
10	$0.29900e+02$	$0.20687e+01$
12	$0.89324e+02$	$0.21747e+01$
14	$0.28321e+03$	$0.22655e+01$
16	$0.93451e+03$	$0.23450e+01$
18	$0.31713e+04$	$0.24154e+01$
20	$0.10987e+05$	$0.24789e+01$

b) Trotz der guten Ergebnisse für die Tschebyscheff-Knoten, kann man zeigen, dass $\|P_n^{(2)}\|$ wie $\ln(n+1)$ gegen ∞ divergiert. Dass man damit nicht notwendiger Weise zu einer Konvergenzaussage gelangt, zeigt der folgende

Satz (3.20) (Faber) Zu jeder Folge von Intervallunterteilungen Δ_n von $[a, b]$ gibt es eine stetige Funktion $f \in C[a, b]$, so dass die zugehörigen Interpolationspolynome P_n von f zu Δ_n nicht gleichmäßig gegen f konvergieren.

Dividierte Differenzen.

Für die numerischen Berechnung eines Interpolationspolynoms werden zumeist die Newtonschen dividierten Differenzen verwendet. Wie wiederholen dieses Verfahren, das ja aus der Numerik bekannt ist, in gebotener Kürze.

Zu $(n+1)$ verschiedenen Knoten $(t_j), j = 0, \ldots n$ im Intervall $[a, b]$ und $f \in \mathrm{C}[a, b]$ wird definiert:

$f[t_0, \ldots, t_n]$: Koeffizient von t^n des Interpol.poly. p_n zu $(t_j, f(t_j)), j = 0, \ldots, n$.

Aus der Lagrange-Darstellung erhält man die folgende explizite Darstellung der dividierten Differenzen

$$f[t_0, \ldots, t_n] \;=\; \sum_{k=0}^{n} \frac{f(t_k)}{\prod_{j \neq k}(t_k - t_j)} \tag{3.21}$$

Die wichtigsten Eigenschaften der dividierten Differenzen werden im folgenden Satz

zusammengefasst

Satz (3.22) (Dividierte Differenzen)

a) Zu $f \in C^n[a,b]$ existiert eine Zwischenstelle $\tau \in]\min t_j, \max t_j[$ mit

$$f[t_0,\ldots,t_n] \;=\; \frac{f^{(n)}(\tau)}{n!}\,.$$

b) Bezeichnet p_k das Interpolationspolynom zu den Stützstellen $(t_j, f(t_j))$, $j = 0,\ldots,k$, so gilt

$$p_0(t) = f_0, \quad p_{k+1}(t) = p_k(t) + f[t_0,\ldots,t_{k+1}]\,(t-t_0)\ldots(t-t_k).$$

Damit folgt insbesondere die *Newton-Darstellung* des Interpolationspolynoms

$$p_n(t) \;=\; \sum_{k=0}^{n} f[t_0,\ldots,t_k] \prod_{j=0}^{k-1}(t-t_j). \tag{3.23}$$

c) Die dividierten Differnzen lassen sich rekursiv berechnen gemäß

$$\begin{aligned} f[t_j] \;&=\; f(t_j), \quad j = 0,\ldots,n \\ f[t_j,\ldots,t_{j+k}] \;&=\; \frac{f[t_{j+1},\ldots,t_{j+k}] - f[t_j,\ldots,t_{j+k-1}]}{t_{j+k}-t_j}, \\ &\qquad\qquad k = 1,\ldots,n, \quad j = 0,\ldots,n-k \end{aligned} \tag{3.24}$$

Beweis:

zu a) Die Fehlerfunktion $e_n := f - p_n$ hat wenigstens die $(n+1)$ Nullstellen $t_0,\ldots,t_n$. Nach dem Satz von Rolle hat daher $e_n^{(n)} = f^{(n)} - n!\,f[t_0,\ldots,t_n]$ wenigstens eine Nullstelle in dem betrachteten Intervall.

zu b) Das Polynom $q(t) := p_{k+1}(t) - p_k(t) - f[t_0,\ldots,t_{k+1}]\,(t-t_0)\ldots(t-t_k)$ hat nach Definition den Höchstgrad k und zugleich wenigstens $k+1$ Nullstellen $t_0,\ldots,t_k$. Daher muss q verschwinden.

zu c) Ist $p_{j,k} \in \Pi_k$ das Interpolationspolynom zu den Stützstellen $(t_i, f(t_i))$, $i = j,\ldots,j+k$, so findet man durch Einsetzen der Interpolationsknoten die Rekursion

$$p_{j,k}(t) \;=\; \frac{(t-t_j)\,p_{j+1,k-1}(t) \;+\; (t_{j+k}-t)\,p_{j,k-1}(t)}{t_{j+k} \;-\; t_j}\,.$$

Vergleich der Koeffizienten von t^k in der obigen Relation liefert die Behauptung. □

Bemerkungen (3.25)

a) Die dividierten Differenzen werden mit der Rekursion (3.24) berechnet (Dreiecksstableau). Dazu genügt es, ein eindimensionales Array der Länge $n+1$ zu verwenden. Die Auswertung des Interpolationspolynoms erfolgt dann über (3.23) mit einem angepassten Horner-Schema.

b) Das obige Rechenschema der dividierten Differenzen lässt sich mühelos auf den Fall der so genannten *Hermite-Interpolation* (benannt nach Charles Hermite, 1822-1901) übertragen.

Hier ist zu einer vorgegebenen C^1-Funktion $f \in C^1[a,b]$ und einem Gitter $a \leq t_0 < \ldots < t_n \leq b$ ein Polynom $p \in \Pi_{2n+1}[a,b]$ gesucht, so dass $p^{(k)}(t_j) = f^{(k)}(t_j)$ für alle j und $k = 0, 1$ gelten.

Auch diese Hermitesche Interpolationsaufgabe besitzt eine eindeutig bestimmte Lösung, die sich mittels (3.23) berechnen lässt. Dazu hat man alle Knoten in dem Tableau der dividierten Differenzen doppelt zu nehmen, wobei definiert wird: $f[t_j, t_j] := f'(t_j)$. Ansonsten ist das Tableau wie in (3.24) zu berechnen.

4. Der Weierstraßsche Approximationssatz [2]

Wir geben in diesem Abschnitt einen konstruktiven Beweis des Weierstraßschen Approximationssatzes, der mit den so genannten Bernstein-Polynomen (Felix Bernstein, 1878-1956) arbeitet.

Definition (4.1) Ein Operator $L : \mathrm{C}[a,b] \to \mathrm{C}[a,b]$ heißt *monoton*, falls

$$\forall\, f,g \in \mathrm{C}[a,b] : \quad f \le g \;\Rightarrow\; L(f) \le L(g).$$

Ist $L : \mathrm{C}[a,b] \to \mathrm{C}[a,b]$ ein *linearer* Operator, so ist L genau dann monoton, falls er *positiv* ist, d.h.

$$\forall\, f \in \mathrm{C}[a,b] : \quad f \ge 0 \;\Rightarrow\; L(f) \ge 0.$$

Bemerkung (4.2)

Lineare, monotone Operatoren sind stetig bzgl. $\|\cdot\|_\infty$ mit Operatornorm $\|L\|_\infty = \|L(1)\|_\infty$.

Beweis: Aus $f \le \|f\|_\infty 1$ folgt durch Anwendung des Operators L:

$$L(f) \le \|f\|_\infty L(1), \quad \text{also} \quad \|L(f)\|_\infty \le \|f\|_\infty \|L(1)\|_\infty.$$

Da speziell für $f = 1$ hier Gleichheit gilt, folgt die Behauptung. □

Satz (4.3) (Korovkin I)

Sei $L_n : \mathrm{C}[a,b] \to \mathrm{C}[a,b]$ eine Folge linearer, monotoner Operatoren. Gilt dann für die drei Funktionen $f_k(t) := t^k$, $k = 0,1,2$, die gleichmäßige Konvergenz $L_n(f_k) \to f_k$, $n \to \infty$, bzgl. $\|\cdot\|_\infty$, so folgt die gleichmäßige Konvergenz $L_n(f) \to f$, $n \to \infty$, für *alle* stetigen Funktionen $f \in \mathrm{C}[a,b]$.

Beweis: Wir führen den Beweis in drei Teilschritten.

Schritt (A): Ist $\Phi_n : \mathrm{C}[a,b] \to \mathbb{R}$ eine Folge linearer, positiver Funktionale und gelten mit $\psi_\alpha(x) := (x-\alpha)^2$ und $\alpha \in [a,b]$ die Bedingungen

$$\Phi_n(1) \to 1, \quad \Phi_n(\psi_\alpha) \to 0, \quad (n \to \infty),$$

so folgt $\Phi_n(f) \to f(\alpha)$ für *jedes* $f \in \mathrm{C}[a,b]$.

Beweis zu (A): Zunächst ist f auf $[a,b]$ beschränkt, es gibt also ein $M > 0$ mit $-M \le f(x) \le M$ für alle $x \in [a,b]$. Damit gilt auch

$$\forall x \in [a,b] : \; -2\,M \le f(x) - f(\alpha) \le 2\,M. \tag{1}$$

[2] Karl Weierstrass, 1815 – 1897

Aus der Stetigkeit von f folgt weiter für beliebiges $\varepsilon > 0$ die Existenz eines $\delta > 0$ mit

$$\forall\, x \in [a,b]:\ |x-\alpha| < \delta \ \Rightarrow\ -\varepsilon \,\le\, f(x)-f(\alpha) \,\le\, \varepsilon. \tag{2}$$

(1) und (2) zusammen ergeben die Abschätzung

$$\forall\, x \in [a,b]:\ -\varepsilon - \frac{2\,M}{\delta^2}\,\psi_\alpha(x) \ \le\ f(x)-f(\alpha) \ \le\ \varepsilon + \frac{2\,M}{\delta^2}\,\psi_\alpha(x). \tag{3}$$

Für $|x-\alpha| < \delta$ folgt dies direkt aus (2), da ja δ und M positiv und $\psi_\alpha(x)$ nichtnegativ sind. Ist dagegen $|x-\alpha| \ge \delta$, so ist $\psi_\alpha(x)/\delta^2 \ge 1$, so dass sich (3) aus (1) ergibt.

Auf (3) wenden wir nun die positiven Funktionale Φ_n an und erhalten

$$-\varepsilon\,\Phi_n(1) \,-\, \frac{2\,M}{\delta^2}\,\Phi_n(\psi_\alpha) \ \le\ \Phi_n(f) \,-\, f(\alpha)\,\Phi_n(1) \ \le\ \varepsilon\,\Phi_n(1) \,+\, \frac{2\,M}{\delta^2}\,\Phi_n(\psi_\alpha).$$

Für $n \to \infty$ folgt mit den Voraussetzungen: Jeder Häufungspunkt der Folge $\Phi_n(f)$ liegt im Intervall $[f(\alpha)-\varepsilon, f(\alpha)+\varepsilon]$. Da dies nun für *jedes* $\varepsilon > 0$ gilt, folgt die Behauptung.

Schritt (B): Ist $\Phi_n : \mathrm{C}[a,b] \to \mathbb{R}$ eine Folge linearer, positiver Funktionale und gelten mit $\alpha \in [a,b]$ die Bedingungen

$$\Phi_n(1) \to 1, \quad \Phi_n(x) \to \alpha, \quad \Phi_n(x^2) \to \alpha^2, \quad (n \to \infty),$$

so folgt $\Phi_n(f) \to f(\alpha)$ für *jedes* $f \in \mathrm{C}[a,b]$.

Beweis zu (B): Mit den obigen Voraussetzungen folgt durch Ausmultplizieren:

$$\Phi_n(\psi_\alpha) \ =\ \Phi_n(x^2) - 2\,\alpha\,\Phi_n(x) + \alpha^2\,\Phi_n(1) \ \to\ 0, \quad (n \to \infty),$$

und damit die Behauptung nach (A).

Schritt (C): Wir übertragen (A), (B) auf eine Folge linearer, monotoner (positiver) Operatoren $L_n : \mathrm{C}[a,b] \to \mathrm{C}[a,b]$. Dazu sehen wir den Parameter $\alpha \in [a,b]$ nun als variabel an. Wie in (A) findet man

$$\forall\, x,\alpha \in [a,b]:\ -2\,M \,\le\, f(x)-f(\alpha) \,\le\, 2\,M, \tag{1'}$$

$$\forall\, x,\alpha \in [a,b]:\quad |x-\alpha| < \delta \ \Rightarrow\ -\varepsilon \,\le\, f(x)-f(\alpha) \,\le\, \varepsilon, \tag{2'}$$

$$\forall\, x,\alpha \in [a,b]:\ -\varepsilon - \frac{2\,M}{\delta^2}\,\psi_\alpha(x) \ \le\ f(x)-f(\alpha) \ \le\ \varepsilon + \frac{2\,M}{\delta^2}\,\psi_\alpha(x). \tag{3'}$$

Bei (2′) hat man die gleichmäßige Stetigkeit von f zu beachten, so dass δ tatsächlich nur von ε abhängt.

Die Anwendung von L_n auf (3′) (bei festem α) ergibt

$$-\varepsilon\,L_n(1) \,-\, \frac{2\,M}{\delta^2}\,L_n(\psi_\alpha) \ \le\ L_n(f) \,-\, f(\alpha)\,L_n(1) \ \le\ \varepsilon\,L_n(1) \,+\, \frac{2\,M}{\delta^2}\,L_n(\psi_\alpha).$$

oder

$$-(\varepsilon - f(\alpha))\, L_n(1) \;-\; \frac{2\,M}{\delta^2}\, L_n(\psi_\alpha) \;\;\le\;\; L_n(f) \;\;\le\;\; (\varepsilon + f(\alpha))\, L_n(1) \;+\; \frac{2\,M}{\delta^2}\, L_n(\psi_\alpha).$$

Nun konvergieren $L_n(1)(t) \to 1$ und $L_n(\psi_\alpha)(t) \to (t^2 - 2\alpha t + \alpha^2)$ gleichmäßig auf $[a,b]$ für $n \to \infty$. Letzteres sieht man wieder durch Ausmultiplizieren und Anwendung der Voraussetzungen des Satzes. Zum vorgegebenen $\varepsilon > 0$ gibt es also ein $N = N(\varepsilon)$, so dass der Abstand zu den Grenzwerten für alle $n \ge N$ und $t \in [a,b]$ höchstens ε (im ersten Fall) bzw. $\varepsilon\,\delta^2$ (für den zweiten Grenzwert) beträgt. Speziell für $t = \alpha$ ergibt sich damit die Abschätzung $(n \ge N(\varepsilon))$

$$-(\varepsilon - f(\alpha))\,(1+\varepsilon) \;-\; 2\,M\,\varepsilon \;\;\le\;\; L_n(f)(\alpha) \;\;\le\;\; (\varepsilon + f(\alpha))\,(1+\varepsilon) \;+\; 2\,M\,\varepsilon,$$

und damit die *gleichmäßige* Konvergenz $L_n(f)(\alpha) \to f(\alpha)$ für $n \to \infty$. □

Für den Raum der reellen, stetigen, $2\,\pi$ - periodischen Funktionen

$$\mathrm{C}_{2\,\pi} \;:=\; \{f \in \mathrm{C}(\mathbb{R}) :\; \forall\, t \in \mathbb{R} :\; f(t + 2\,\pi) = f(t)\} \tag{4.4}$$

lässt sich die folgende Variante des Korovkinschen Satzes zeigen:

Satz (4.5) (Korovkin II)

Sei $L_n : \mathrm{C}_{2\,\pi} \to \mathrm{C}_{2\,\pi}$ eine Folge linearer, monotoner Operatoren. Gilt dann für die drei Funktionen $f_0(t) := 1$, $f_1(t) := \cos t$ und $f_2(t) := \sin t$ die gleichmäßige Konvergenz $L_n(f_k) \to f_k$, $n \to \infty$, so folgt hieraus die gleichmäßige Konvergenz $L_n(f) \to f$, $n \to \infty$, für *alle* Funktionen $f \in \mathrm{C}_{2\,\pi}$.

Definition (4.6) (Bernstein, 1912)

Die Operatoren $B_n : \mathrm{C}[0,1] \to \Pi_n$, $n \in \mathbb{N}_0$, definiert durch

$$B_n(f)(t) \;:=\; \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) t^k\,(1-t)^{n-k} f(\frac{k}{n}), \qquad 0 \le t \le 1,$$

heißen *Bernstein - Operatoren.*

Bemerkungen (4.7)

a) $B_n(f)(t)$ ist stets eine Konvexkombination der teilnehmenden Funktionswerte $f(k/n)$. Man beachte insbesondere die Ähnlichkeit zur binomischen Formel

$$B_n(1)(t) \;=\; \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) t^k\,(1-t)^{n-k} \;=\; (t + (1-t))^n \;=\; 1.$$

Man beachte aber auch die Ähnlichkeit zur Lagrange-Darstellung des Interpolationspolynoms in den Knoten $t_k := k/n$.

b) B_n ist offensichtlich ein linearer Operator. Er ist auch positiv und damit monoton,

$$f \geq 0 \;\Rightarrow\; B_n(f) \geq 0,$$

er ist jedoch *kein* Projektor. Für das Lagrange-Polynom ℓ_j zu den Knoten t_k ergibt sich nämlich

$$B_n(\ell_j)(t) \;=\; \left(\begin{array}{c} n \\ j \end{array} \right) t^j \, (1-t)^{n-j} \;\neq\; \ell_j(t).$$

Satz (4.8)

Für alle stetigen Funktionen $f \in \mathrm{C}[0,1]$ konvergieren die Bernstein-Approximationen auf $[0,1]$ gleichmäßig gegen f: $B_n(f) \to f$, $(n \to \infty)$.

Beweis: Nach dem Satz von Korovkin genügt es zu zeigen, dass $B_n(f_k) \to f_k$, gleichmäßig auf $[0,1]$, für $f_k(t) := t^k$ und $k = 0, 1, 2$ gilt.

$\mathbf{k = 0}$: $\quad B_n(1)(t) \;=\; 1 \;\to\; 1, \quad (n \to \infty)$

$\mathbf{k = 1}$:

$$\begin{aligned} B_n(t)(t) \;&=\; \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) t^k \, (1-t)^{n-k} \, \frac{k}{n} \\ &=\; \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)! \, ((n-1)-(k-1))!} \, t^k \, (1-t)^{(n-1)-(k-1)} \\ &=\; t \, (t + (1-t))^{n-1} \;=\; t \;\to\; t, \quad (n \to \infty) \end{aligned}$$

$\mathbf{k = 2}$: Durch elementare Umformung rechnet man nach

$$B_n(t^2)(t) \;=\; \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) t^k \, (1-t)^{n-k} (k/n)^2 \;=\; \frac{n-1}{n} \, t^2 \;+\; \frac{1}{n} \, t.$$

Damit wird

$$\begin{aligned} \| B_n(t^2) \;-\; t^2 \|_\infty \;&=\; \max_{t \in [0,1]} \left| \frac{n-1}{n} \, t^2 \;+\; \frac{1}{n} \, t \;-\; t^2 \right| \\ &=\; \frac{1}{n} \max_{t \in [0,1]} \, | -t^2 + t \, | \;=\; \frac{1}{4n} \;\to\; 0. \end{aligned}$$

□

Folgerung (4.9) (Approximationssatz I; Weierstraß 1885)

Zu einer stetigen Funktion $f \in \mathrm{C}[a,b]$ und $\varepsilon > 0$ existiert ein Polynom $p \in \Pi[a,b]$ mit $\|f - p\|_\infty \leq \varepsilon$. Anders ausgedrückt: Der Polynomraum $\Pi[a,b]$ liegt dicht in $\mathrm{C}[a,b]$ bezüglich der $\| \cdot \|_\infty$ - Norm.

Bemerkungen (4.10)

a) Für praktische Approximationen sind die Bernstein-Operatoren nur bedingt geeignet, da die Konvergenz sehr langsam ist. Dies wird durch die im Beweis zu Satz (4.8) aufgezeigte Beziehung $\|B_n(t^2) - t^2\|_\infty = 1/(4n)$ verdeutlicht. Andererseits hat die Approximation $B_n(f)$ theoretisch interessante Eigenschaften, vgl. z.B. den Satz (4.13). In Abbildung 4.1 ist die Bernstein Approximation zu $f(t) := \sin(3\,\pi t)$ für $n = 10$ und $n = 100$ dargestellt.

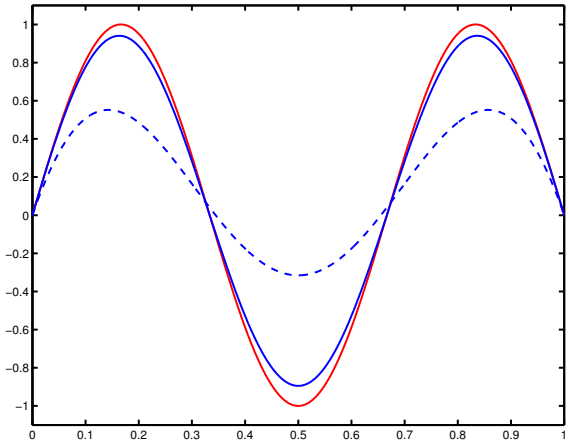


Abb. 4.1 Bernstein Approximation, $n = 10,\ 100$.

b) Zur numerischen Auswertung der Bernstein-Operatoren lassen sich die *Bernstein-Polynome*

$$B_k^n(t) \;:=\; \left(\begin{array}{c} n \\ k \end{array} \right) t^k\,(1-t)^{n-k}, \qquad k = 0, \ldots, n, \tag{4.11}$$

verwenden. Diese lassen sich wie folgt rekursiv berechnen

$$B_k^n(t) \;=\; t\,B_{k-1}^{n-1}(t) \,+\, (1-t)\,B_k^{n-1}(t),\;\; B_0^0(t) := 1,\; B_{-1}^{n-1}(t) := B_n^{n-1}(t) := 0. \tag{4.12}$$

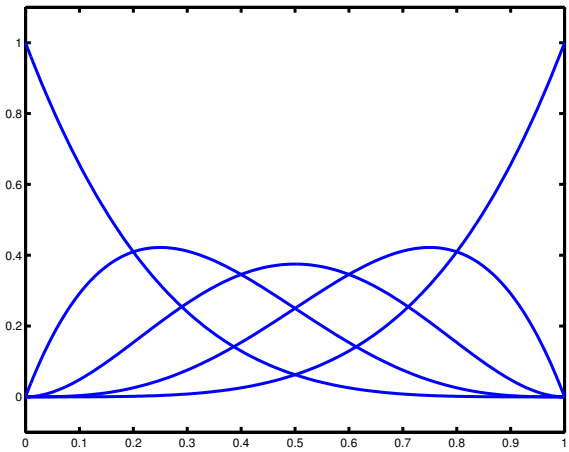


Abb. 4.2 Bernstein Polynome B_k^4.

Satz (4.13)

Für alle stetig differenzierbaren Funktionen $f \in \mathrm{C}^1[0,1]$ konvergieren auch die ersten Ableitungen der Bernstein Approximationen $B_n(f)$ gleichmäßig auf $[0,1]$ gegen f': $B_n(f)' \to f',\ n \to \infty$.

Beweis: Nach dem Satz (4.8) gilt $B_n(f') \to f'$ gleichmäßig auf $[0,1]$. Es genügt daher zu zeigen, dass

$$\|B_n(f') \ - \ B_{n+1}(f)'\|_\infty \ \to \ 0, \quad (n \to \infty).$$

Hierzu formen wir um

$$\begin{aligned}
B_{n+1}(f)'(t) &= \frac{d}{d\,t}\Big\{\sum_{k=0}^{n+1}\left(\begin{array}{c} n+1 \\ k \end{array}\right) t^k\,(1-t)^{n+1-k} f(\frac{k}{n+1})\Big\} \\
&= \sum_{k=1}^{n+1} \frac{(n+1)!}{(k-1)!\,(n+1-k)!}\,t^{k-1}\,(1-t)^{n+1-k} f(\frac{k}{n+1}) \\
&- \sum_{k=0}^{n} \frac{(n+1)!}{k!\,(n-k)!}\,t^k\,(1-t)^{n-k} f(\frac{k}{n+1}) \\
&= \sum_{k=0}^{n} \frac{(n+1)!}{k!\,(n-k)!}\,t^k\,(1-t)^{n-k}\big[f(\frac{k+1}{n+1}) - f(\frac{k}{n+1})\big] \\
&= \sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right) t^k\,(1-t)^{n-k} f\big[\frac{k}{n+1}, \frac{k+1}{n+1}\big] \\
&= \sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right) t^k\,(1-t)^{n-k} f'(\tau_k), \quad \frac{k}{n+1} < \tau_k < \frac{k+1}{n+1}.
\end{aligned}$$

Damit folgt

$$\begin{aligned}
|B_n(f')(t) - B_{n+1}(f)'(t)| &= \Big|\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right) t^k\,(1-t)^{n-k}\big(f'(\frac{k}{n}) - f'(\tau_k)\big)\Big| \\
&\le \max_{k=0,\ldots,n}\big|f'(\frac{k}{n}) - f'(\tau_k)\big| \ \to \ 0.
\end{aligned}$$

□

Wir sehen uns im Folgenden noch kurz eine Variante des Weierstraßschen Approximationssatzes für periodische Funktionen an. Dazu sei mit

$$\mathrm{T}_n \ := \ \{f \in \mathrm{C}_{2\,\pi} : \ f(t) = \frac{a_0}{2} \ + \ \sum_{k=1}^{n}\,[a_k \cos(kt) + b_k \sin(kt)],\ a_k, b_k \in \mathbb{R}\} \quad (4.14)$$

der $(2n+1)$-dimensionale Vektorraum der (reellen) trigonometrischen Polynome vom Maximalgrad n bezeichnet. Bekanntlich bilden die Funktionen $\cos(kt)$, $\sin(kt)$ eine

orthogonale Basis von T_n bezüglich des Standard - Skalarproduktes. Jedem $f \in \mathrm{C}_{2\pi}$ wird die Fourier - Summe

$$\begin{aligned} S_n(f)(t) &= \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kt) + b_k \sin(kt)] \\ a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t)\ \cos(kt)\, dt, \qquad k \geq 0, \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(t)\ \sin(kt)\, dt, \qquad k > 0, \end{aligned} \tag{4.15}$$

zugeordnet. $S_n(f)$ lässt sich als Projektion von $\mathrm{C}_{2\pi}$ auf T_n deuten. Man beachte, dass die Fourier-Koeffizienten a_k und b_k vom Approximationsgrad n unabhängig sind. Wir werden uns später mit der Approximationsgüte $S_n(f) - f$ und der Konvergenzeigenschaft $S_n(f) \to f$ ausführlicher beschäftigen. Wir geben zunächst eine Integraldarstellung der Fourier-Summe an, die sich aus einfachen trigonometrischen Umformungen ergibt.

Satz (4.16) (Integraldarstellung)

Für $f \in \mathrm{C}_{2n}$ gilt

$$S_n(f)(t) = \frac{1}{\pi} \int_0^{2\pi} f(x)\, \frac{\sin[(n+1/2)(x-t)]}{2\, \sin[(x-t)/2]}\, dx$$

Der Kern in dieser Integraldarstellung $D_n(\theta) := \dfrac{\sin[(n+1/2)\theta]}{2\, \sin[\theta/2]}$ heißt auch *Dirichlet - Kern.*

Beweis: Nach Einsetzen der Koeffizienten in (4.15) ergibt sich

$$\begin{aligned} S_n(f)(t) &= \frac{1}{\pi} \int_0^{2\pi} f(\theta) \left\{ \frac{1}{2} + \sum_{k=1}^{n} \cos(kt)\cos(k\theta) + \sin(kt)\sin(k\theta) \right\} d\theta \\ &= \frac{1}{\pi} \int_0^{2\pi} f(\theta) \left\{ \frac{1}{2} + \sum_{k=1}^{n} \cos(k(t-\theta)) \right\} d\theta \\ &= \frac{1}{\pi} \int_0^{2\pi} f(t+\theta) \left\{ \frac{1}{2} + \sum_{k=1}^{n} \cos(k\theta) \right\} d\theta \\ &= \frac{1}{\pi} \int_0^{2\pi} f(t+\theta)\, D_n(\theta)\, d\theta. \end{aligned}$$

Die vorletzte Gleichung ergibt sich aufgrund der $(2\,\pi)$ - Periodizität des Integranden, die letzte Gleichung durch Ausmultiplizieren von $\sin(\theta/2)\{1/2 + \sum_{k=1}^{n} \cos(k\theta)\}$ und Anwendung der Relation $\cos\alpha \sin\beta = 0.5\,(\sin(\alpha+\beta) - \sin(\alpha-\beta))$. □

Die Fourier-Summe selbst ist noch nicht zur Anwendung des zweiten Korovkinschen Satzes geeignet, da S_n nicht positiv ist. Statt dessen definieren wir den *Fejér - Operator* durch

$$F_n(f)(t) \;:=\; \frac{1}{n}\sum_{k=0}^{n-1} S_k(f)(t), \qquad f \in \mathrm{C}_{2\pi}. \tag{4.17}$$

Setzt man die Integraldarstellungen (4.16) hierin ein, so ergibt sich mittels trigonometrischer Umformung die folgende Darstellung für den Fejér-Operator

$$F_n(f)(t) \;:=\; \frac{1}{n\,\pi}\int_0^{2\pi} f(x)\,\frac{\sin^2[(n/2)(x-t)]}{2\,\sin^2[(x-t)/2]}\,dx \tag{4.18}$$

Der Kern dieses Integralausdrucks $\sigma_n(\theta) \;:=\; \dfrac{\sin^2[(n/2)\theta]}{2\,n\,\sin^2[\theta/2]}$ heißt entsprechend der *Fejér - Kern* von f.

Anhand der Darstellung (4.18) sieht man, dass F_n ein linearer und monotoner (positiver) Operator auf $\mathrm{C}_{2\pi}$ ist. Ferner ergibt sich direkt aus der Definition der Fourier-Summe

$$\begin{aligned} F_n(1)(t) &= 1 \;\to\; 1 \\ F_n(\cos)(t) &= \frac{n-1}{n}\cos t \;\to\; \cos t \\ F_n(\sin)(t) &= \frac{n-1}{n}\sin t \;\to\; \sin t, \end{aligned} \tag{4.19}$$

wobei die Konvergenz jeweils gleichmäßig ist. Damit sind die Voraussetzungen des Korovkinschen Satzes (4.5) erfüllt und wir erhalten den zweiten Weierstraßschen Approximationssatz.

Satz (4.20) (Approximationssatz II; Weierstraß)

Zu einer stetigen, $2\,\pi$-periodischen Funktion $f \in \mathrm{C}_{2\pi}$ und $\varepsilon > 0$ existiert ein trigonometrisches Polynom $p \in \mathrm{T} := \bigcup_{n\in\mathbb{N}} \mathrm{T}_n$ mit $\|f - p\,\|_\infty \le \varepsilon$. Anders ausgedrückt: Die trigonometrischen Polynome T liegen dicht in $\mathrm{C}_{2\pi}$ bezüglich der $\|\cdot\|_\infty$ - Norm.

5. Splinefunktionen

Interpolation mit Splines.

Die im dritten Kapitel geschilderten Schwierigkeiten bei der Interpolation mit Polynomen höheren Grades bzgl. der Approximationsgüte lassen sich dadurch überwinden, dass man auf *stückweise definierte Polynome* als approximierende Funktionen ausweicht.

Schränken wir uns zunächst auf den häufig auftretenden Fall der stückweise *kubischen* Polynome ein, so ergibt sich die folgende Problemstellung:

Gegeben seien Interpolationsknoten in einem Intervall $[a, b]$:

$$\Delta : \quad a = t_0 < t_1 < \ldots < t_n = b, \tag{5.1}$$

sowie Daten $f_j = f(t_j),\ j = 0, \ldots, n$ einer vorgegebenen Funktion $f \in \mathrm{C}[a, b]$. Gesucht ist eine (stetige) Funktion $s : [a, b] \to \mathbb{R}$ mit den Eigenschaften

$$\text{(a)} \quad s|_{[t_j, t_{j+1}]} \in \Pi_3, \qquad \text{(b)} \quad s(t_j) = f_j, j = 0, \ldots, n. \tag{5.2}$$

Setzt man das kubische Polynom im j-ten Teilintervall $[t_j, t_{j+1}]$ mit

$$s(t) \ = \ p_j(t) \ = \ a_j \ + \ b_j(t - t_j) \ + \ c_j(t - t_j)^2 \ + \ d_j(t - t_j)^3 \tag{5.3}$$

an, so erkennt man, dass zur Festlegung der Koeffizienten pro Teilintervall zwei Informationen fehlen.

Zur Festlegung könnte man zusätzlich Ableitungen f'_j vorschreiben, also neben (5.2) fordern

$$p'_j(t_j) = f'_j, \qquad p'_j(t_{j+1}) = f'_{j+1}, \qquad j = 0, \ldots, n - 1. \tag{5.4}$$

Man beachte, dass die Daten f'_j willkürlich vorgegeben werden, zumal nur $f \in \mathrm{C}[a, b]$ vorausgesetzt war. Zugleich schränkt man die approximierenden Funktionen damit auf C^1-Funktionen ein.

Mittels einfacher Rechnung ergeben sich hieraus die eindeutig bestimmten Koeffizienten

$$\begin{aligned} a_j &= f_j, \\ b_j &= f'_j, \\ c_j &= \frac{3\, f[t_j, t_{j+1}] \ - \ 2\, f'_j \ - \ f'_{j+1}}{t_{j+1} \ - \ t_j}, \\ d_j &= \frac{f'_j \ + \ f'_{j+1} \ - \ 2\, f[t_j, t_{j+1}]}{(t_{j+1} \ - \ t_j)^2}. \end{aligned} \tag{5.5}$$

Die Splinefunktion lässt sich nun leicht mittels (5.5) und (5.3) (per Intervallabfrage und Anwendung des Horner-Schemas) auswerten.

Es gibt verschiedene Wege, geeignete Ableitungswerte f'_j vorzuschreiben.

A. Kubische Hermite-Interpolation. Hierbei sind die Daten f'_j vom Problem her vorgeschrieben, f ist also eine C^1-Funktion.

B. Kubische Bessel-Interpolation. Man bestimmt f'_j als Ableitung einer interpolierenden Parabel zu drei benachbarten Knoten:

$$\begin{aligned}
&\text{für } j = 1, \ldots, n-1: \\
&\qquad p_j \in \Pi_2 \text{ interpoliere } (t_i, f_i),\ i = j-1, j, j+1; \\
&\qquad f'_j := p'_j(t_j); \\
&\text{für } j = 0: \quad f'_0 := p'_1(t_0); \\
&\text{für } j = n: \quad f'_n := p'_{n-1}(t_n).
\end{aligned}$$

C. Kubische Spline-Interpolation. Man bestimme die f'_j so, dass s sogar eine C^2-Funktion wird, also $p''_{j-1}(t_j) = p''_j(t_j)$, $j = 1, 2, \ldots, n-1$, gilt. Offenbar fehlen dann aber noch zwei zusätzliche Bedingungen.

Aus (5.3) und (5.5) erhält man damit das folgende lineare Gleichungssystem ($j = 1, 2, \ldots, n-1$):

$$\begin{aligned}
&\frac{1}{h_{j-1}} f'_{j-1} + 2\left(\frac{1}{h_{j-1}} + \frac{1}{h_j}\right) f'_j + \frac{1}{h_j} f'_{j+1} = r_j, \\
&r_j := 3\left(\frac{f[t_{j-1}, t_j]}{h_{j-1}} + \frac{f[t_j, t_{j+1}]}{h_j}\right),
\end{aligned} \tag{5.6}$$

wobei $h_j := t_{j+1} - t_j$, $j = 0, \ldots, n-1$. In Matrix-Schreibweise lautet das Gleichungssystem (5.6)

$$\begin{pmatrix}
\frac{1}{h_0} & 2(\frac{1}{h_0} + \frac{1}{h_1}) & \frac{1}{h_1} & & \\
& \ddots & \ddots & \ddots & \\
& & \frac{1}{h_{n-2}} & 2(\frac{1}{h_{n-2}} + \frac{1}{h_{n-1}}) & \frac{1}{h_{n-1}}
\end{pmatrix}
\begin{pmatrix} f'_0 \\ \vdots \\ \vdots \\ f'_n \end{pmatrix}
=
\begin{pmatrix} r_1 \\ \vdots \\ r_{n-1} \end{pmatrix}. \tag{5.7}$$

Schreibt man nun f'_0 und f'_n geeignet vor, so kann man diese Ausdrücke auf die rechte Seite bringen und man erhält aus (5.7) ein lineares Gleichungssystem mit einer quadratischen $(n-1, n-1)$ Koeffizientenmatrix, die *tridiagonal*, *symmetrisch* und *strikt diagonaldominant* ist, somit auch insbesondere regulär. Das lineare Gleichungssystem besitzt daher eine eindeutig bestimmte Lösung und diese lässt sich numerisch stabil und effizient mittels *Cholesky–Zerlegung* der Koeffizientenmatrix lösen.

Anmerkung zur Cholesky-Zerlegung: Bei vorgegebenen Ableitungen $f'_0,\ f'_n$ erhält man aus (5.7) das lineare Gleichungssystem

$$\begin{bmatrix} 2(\frac{1}{h_0}+\frac{1}{h_1}) & \frac{1}{h_1} & & & \\ \frac{1}{h_1} & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & \frac{1}{h_{n-2}} \\ & & & \frac{1}{h_{n-2}} & 2(\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}}) \end{bmatrix} \begin{pmatrix} f'_1 \\ \vdots \\ \vdots \\ \vdots \\ f'_{n-1} \end{pmatrix} = \begin{pmatrix} r_1 - f'_0/h_0 \\ r_2 \\ \vdots \\ r_{n-2} \\ r_{n-1} - f'_n/h_{n-1} \end{pmatrix}.$$

Für die Dreieckszerlegung der Koeffizientenmatrix verwendet man nun den Ansatz

$$\begin{pmatrix} a_1 & c_1 & & 0 \\ c_1 & a_2 & \ddots & \\ & \ddots & \ddots & c_{n-2} \\ 0 & & c_{n-2} & a_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & & & 0 \\ u_2 & 1 & & \\ & \ddots & \ddots & \\ 0 & & u_{n-1} & 1 \end{pmatrix} \begin{pmatrix} v_1 & c_1 & & 0 \\ & v_2 & \ddots & \\ & & \ddots & c_{n-2} \\ 0 & & & v_{n-1} \end{pmatrix},$$

so dass sich die $u_i,\ v_i$ mit dem folgenden **Algorithmus** berechnen lassen

$$\begin{aligned} & v_1 \ := \ 2\,\big(1/h_0 \,+\, 1/h_1\big); \\ & \text{für } \ j \,=\, 2,\, 3,\, \ldots\,, n-1 \\ & \qquad u_j \ := \ 1/(h_{j-1}\, v_{j-1}), \\ & \qquad v_j \ := \ 2\,\big(1/h_{j-1} \,+\, 1/h_j\big) \ - \ u_j/h_{j-1}; \end{aligned}$$

Im Anschluss ist die **Vorwärts- / Rückwärtssubstitution** durchzuführen

$$\begin{aligned} & z_1 \ := \ r_1 \,-\, f'_0/h_0; \\ & \text{für } \ j \,=\, 2,\, 3,\, \ldots\,, n-2 \\ & \qquad z_j \ := \ r_j \,-\, u_j\, z_{j-1}; \\ & z_{n-1} \ := \ \big(r_{n-1} \,-\, f'_n/h_{n-1}\big) \,-\, u_{n-1}\, z_{n-2}; \\ & f'_{n-1} \ := \ z_{n-1}/v_{n-1}; \\ & \text{für } \ j \,=\, n-2,\, n-3,\, \ldots\,, 1 \\ & \qquad f'_j \ := \ \big(z_j \,-\, f'_{j+1}/h_j\big)/v_j; \end{aligned}$$

Bei vorgegebenen Randableitungen $f'_0 = f'(t_0),\ f'_n = f'(t_n)$ ist also die *interpolierende kubische Splinefunktion* eindeutig bestimmt und werde mit s_f bezeichnet.

Andere Randvorgaben (5.8)

(i) *Natürliche Randbedingungen*: $s''(t_0) \,=\, s''(t_n) \,=\, 0.$

(ii) *Periodische Randbedingungen*: Ist f eine periodische Funktion, gilt also $f(t_0) = f(t_n),$ so fordert man zusätzlich $s'(t_0) = s'(t_n)$ und $s''(t_0) = s''(t_n).$

Definition (5.9)

Der lineare Raum der *kubischen Splinefunktionen* zum Gitter $\Delta = \{t_0 < \ldots < t_n\}$ werde definiert durch

$$S_3(\Delta) := \{s : [t_0, t_n] \to \mathbb{R} : \ s \in \mathrm{C}^2[t_0, t_n], \ s|_{[t_j, t_{j+1}]} \in \Pi_3(\forall j)\}$$

Satz (5.10) (Extremal- und Approximationseigenschaften)

Sei $f \in \mathrm{C}^2[a, b]$ und bezeichne

$$\mathrm{Int}[f] := \{g \in \mathrm{C}^2[a, b] : \ g(t_j) = f_j, \ j = 0, \ldots, n, \ g'(t_i) = f'(t_i), \ i = 0, n\}$$

die Menge der (die Daten der Funktion f) interpolierenden C^2-Funktionen. Dann gelten für den interpolierenden kubischen Spline $s_f \in \mathrm{Int}[f] \cap S_3(\Delta)$ die folgenden Eigenschaften

a) $$\forall\, g \in \mathrm{Int}[f] : \quad \int_a^b (s_f''(t))^2\, dt \ \leq \ \int_a^b (g''(t))^2\, dt,$$

b) $$\forall\, s \in S_3(\Delta) : \quad \int_a^b (f''(t) - s_f''(t))^2\, dt \ \leq \ \int_a^b (f''(t) - s''(t))^2\, dt.$$

Interpretation:

Unter allen interpolierenden C^2-Funktionen minimiert der kubische Spline das so genannte **Holladay-Funktional** $I(g) := \int_a^b (g''(t))^2\, dt$. Dieses kann gedeutet werden als eine Approximation der Krümmung von s_f.

Unter allen Splinefunktionen auf dem Gitter Δ ist s_f diejenige, für die s_f'' die zweite Ableitung f'' am Besten approximiert (im L_2-Sinn).

Beweis: Zunächst zeigen wir die folgende Orthogonalitätsrelation

$$\forall\, s \in S_3(\Delta),\ g \in \mathrm{Int}[f] : \ \int_a^b s''(t)\, (g''(t) - s_f''(t))\, dt \ = \ 0. \tag{5.11}$$

Mittels zweimaliger partieller Integration lässt sich das Integral nämlich umformen zu

$$\int_a^b s''(g'' - s_f'') \ = \ \sum_{j=0}^{n-1} \left\{ s''(g' - s_f') \big|_{t_j}^{t_{j+1}} \ - \ s'''(g - s_f) \big|_{t_j}^{t_{j+1}} \ + \ \int_{t_j}^{t_{j+1}} s^{(4)}(g - s_f) \right\}$$

Der erste Summand verschwindet wegen der Stetigkeit von $s''(g' - s_f')$ und wegen $g' = s_f' = f'$ in $t = a, b$.

Der zweite Summand verschwindet wegen $g = s_f = f$ in $t = t_j, \ j = 0, \ldots, n$.

Der dritte Summand verschwindet schließlich wegen $s^{(4)} = 0$.

Damit ist die Orthogonalitätsrelation (5.11) gezeigt. Wir kommen zum eigentlichen Beweis:

zu a):
$$\begin{aligned}\int_a^b (g''-s_f'')^2 &= \int_a^b (g'')^2 - \int_a^b (s_f'')^2 - 2\int_a^b s_f''\,(g''-s_f'') \\ &=_{(5.12)} \int_a^b (g'')^2 - \int_a^b (s_f'')^2 \geq 0.\end{aligned}$$

zu b):
$$\begin{aligned}\int_a^b (f''-s'')^2 &= \int_a^b (f''-s_f''+s_f''-s'')^2 \\ &= \int_a^b (f''-s_f'')^2 + \int_a^b (s_f''-s'')^2 \\ &+ 2\int_a^b (f''-s_f'')\,s_f'' - 2\int_a^b (f''-s_f'')\,s'' \\ &=_{(5.11)} \int_a^b (f''-s_f'')^2 + \int_a^b (s_f''-s'')^2 \\ &\geq \int_a^b (f''-s_f'')^2.\end{aligned}$$

□

Beispiel (5.12)

Für das Rungesche Beispiel, vgl. (3.10) und (3.17), ist in Abbildung 5.1 die kubische Spline-Interpolierende für $n = 6$, äquidistante Knoten und natürliche Randbedingungen aufgezeichnet.

Abbildung 5.2 zeigt die Fehlerfunktion für das gleiche Beispiel mit $n = 20$. Man vergleiche auch die Abbildungen 3.3 und 3.4.

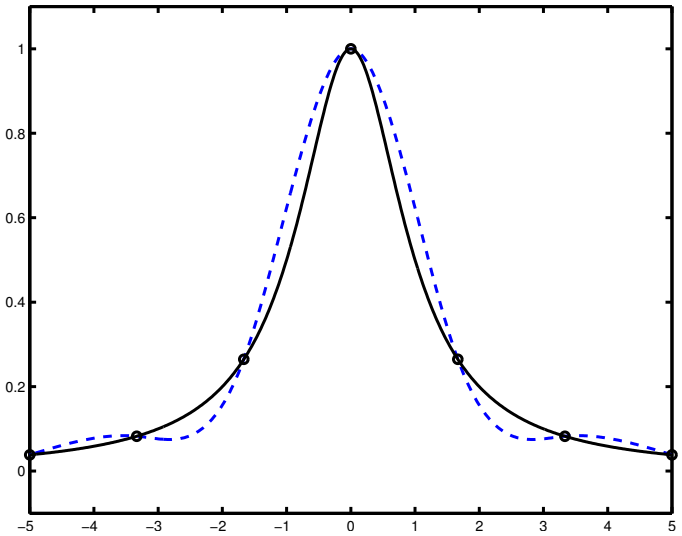


Abb. 5.1 Beispiel von Runge, natürlicher kubischer Spline, $n = 6$.

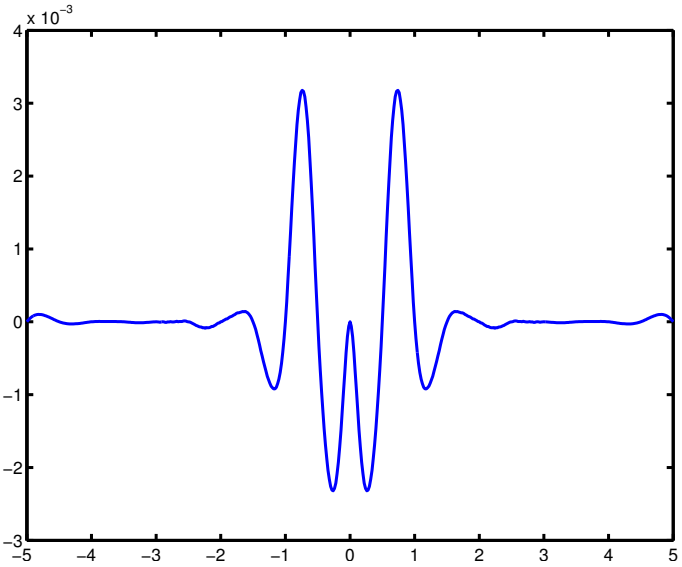


Abb. 5.2 Beispiel von Runge, natürlicher kubischer Spline, $n = 20$.

Definition (5.13)

In Verallgemeinerung von (5.9) definieren wir für $m \in \mathbb{N}_0$ und Knoten $a = t_0 < t_1 < \ldots < t_n = b$ den linearen Raum der *Splinefunktionen vom Grad* m und zum Gitter $\Delta = \{t_0 < \ldots < t_n\}$ durch

$$S_m(\Delta) := \{s : [t_0, t_n] \to \mathbb{R} : \ s \in \mathrm{C}^{m-1}[t_0, t_n], \ \ s|_{[t_j, t_{j+1}[} \in \Pi_m (\forall j), \ \ s|_{[t_{n-1}, t_n]} \in \Pi_m\}$$

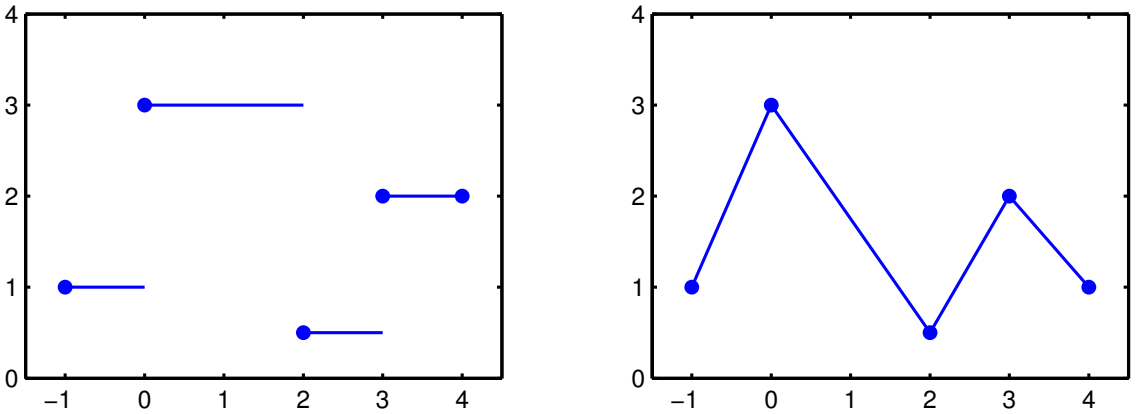


Abb. 5.3 Spline vom Grad $m = 0$ (Treppenfunktion) und $m = 1$ (Polygonzug).

Satz (5.14)

$S_m(\Delta)$ ist ein linearer Teilraum des Vektorraums $\mathbb{R}^{[t_0, t_n]}$ der Dimension $\dim S_m(t_0, \ldots, t_n) = m + n$.

Beweis: Dass $S_m(\Delta)$ ein reeller Vektorraum ist, ist unmittelbar klar. Zur Dimensionsbestimmung verwenden wir eine Basisdarstellung der Splines $s \in S_m(\Delta)$. Nach

Definition hat man eine eindeutige Darstellung $s(t) = p_j(t),\ p_j \in \Pi_m,$ auf dem Intervall $[t_j, t_{j+1}[$. Daher gilt

$$\textbf{(i)} \qquad p_0 \ \in \ \text{Spann}\{1,\ (t-t_0),\ \ldots\ ,\ (t-t_0)^m\}$$

Ferner ist $s \in S_m$ nach Definition eine C^{m-1}-Funktion. Daher muss $p_{j+1} - p_j$ in t_{j+1} eine m-fache Nullstelle besitzen. Es gilt somit

$$\textbf{(ii)} \qquad p_{j+1}(t) \ - \ p_j(t) \ = \ \beta_{j+1}\,(t-t_{j+1})^m, \quad t_{j+1} \le t \le t_n,$$

wobei β_{j+1} eindeutig bestimmt ist. Setzt man also

$$(t-t_j)_+ \ := \ \begin{cases} 0, & \text{für} \ \ t \ < \ t_j, \\ (t-t_j), & \text{für} \ \ t \ \ge \ t_j, \end{cases} \tag{5.15}$$

so erhält man die Spline-Darstellung

$$s(t) \ = \ \sum_{i=0}^{m} \alpha_i\,(t-t_0)^i \ + \ \sum_{j=1}^{n-1} \beta_j\,(t-t_j)_+^m \tag{5.16}$$

mit eindeutig bestimmten Koeffizienten $\alpha_j,\ \beta_j$. Insbesondere ist (5.16) eine Basisdarstellung für $S_m(\Delta)$. □

B-Splines.

Die obige Basis-Darstellung (5.16) ist für die numerische Auswertung wenig geeignet. Zum Einen ist sie anfällig gegenüber Auslöschung (also Verlust an Genauigkeit) zum Anderen ist sie nicht *lokal*, d.h. lokale Störungen in s (z.B. Messungenauigkeit der zu interpolierenden Funktionswerte an einer bestimmten Stelle) wirken sich im Allg. auf *alle* Koeffizienten in der Darstellung (5.16) aus.

Gesucht ist dagegen eine Basisdarstellung, deren Basis-Splines (*B-Splines*) einen möglichst kleinen (kompakten) Träger besitzen. Eine solche *B-Spline-Darstellung* hat die Form

$$s(t) \ = \ \sum_{j=-m}^{n-1} \alpha_j\, B_{m\,j}(t), \tag{5.17}$$

wobei man mit einem beidseitig *erweiterten Gitter*

$$t_{-m} \ < \ t_{-m+1} \ < \ \ldots \ < \ t_0 \ < \ \ldots \ < t_n \ < \ \ldots \ < \ t_{n+m} \tag{5.18}$$

arbeitet (m beliebige zusätzliche Knoten jeweils links von t_0 und rechts von t_n) und $B_{m\,j} \in S_m(t_{-m}, \ldots, t_{n+m})$ ein Spline vom Grad m mit Träger $[t_j, t_{j+m+1}]$ bezeichnet.

Für $m = 3$ hat man beispielsweise die B-Splines $B_{3,\,-3},\ B_{3,\,-2},\ \ldots,\ B_{3,\,n-1}$, wobei $B_{3,\,j}$ den Träger $[t_j, t_{j+4}]$ besitzt.

Die Vorteile der B-Spline-Darstellung (5.17) gegenüber der Darstellung (5.16) liegen auf der Hand:

- Es sind stets nur $m+1$ Summanden in (5.17) auszuwerten, nämlich für $t \in [t_j, t_{j+1}[$ lediglich die Summanden mit den B-Splines $B_{m,j-m}, \ldots, B_{m,j}$.

- Die Darstellung ist *lokal*, d.h. eine Störung in f_j beeinflusst höchstens die Koeffizienten $\alpha_{j-m}, \ldots, \alpha_j$.

- In vielen Anwendungen (z.B. in finiten Element-Programmen) sind Integrale der Form $\int_{t_0}^{t_n} s(t)\, f(t)\, dt$ auszuwerten. Setzt man die Darstellung (5.17) hierin ein, so verbleibt die Berechnung der Integrale $\int B_{m\,j} f(t)\, dt$ über einem jeweils *kleinen* Träger.

Konstruktion der B-Splines.

Wir suchen einen Spline $B_{m,j} \in S_m(t_{-m}, \ldots, t_{n+m})$ mit den Eigenschaften $B_{m,j} \neq 0$ und $B_{m,j}(t) = 0$ für alle $t \notin [t_j, t_{j+p}]$, wobei p möglichst klein sein soll.

Nach (5.16) hat man damit eine Darstellung

$$B_{m,j}(t) \;=\; \sum_{i=j}^{j+p} d_i\, (t - t_i)_+^m, \tag{5.19}$$

wobei wir fordern $\forall\, t > t_{j+p}: \quad \sum_{i=j}^{j+p} d_i\, (t - t_i)^m \;=\; 0.$ Multipliziert man dies mittels binomischer Formel aus, so ergibt sich

$$\forall\, t > t_{j+p}: \quad \sum_{r=0}^{m} \binom{m}{r} (-1)^r \left(\sum_{i=j}^{j+p} d_i t_i^r \right) t^{m-r} \;=\; 0$$

und damit das folgende homogene lineare Gleichungssystem zur Bestimmung der $d_j, \ldots, d_{j+p}$:

$$\forall\, r = 0, 1, \ldots, m: \quad \sum_{i=j}^{j+p} t_i^r\, d_i \;=\; 0. \tag{5.20}$$

Die Koeffizientenmatrix dieses linearen Gleichungssytems hat maximalen Rang ($= m+1$; die ersten m Spalten bilden gerade die Vandermonde Matrix zu $t_j, \ldots, t_{j+m}$) und ist daher nur für $p \geq m+1$ singulär.

Wir wählen p minimal, also $p = m+1$. Dann hat (5.20) einen eindimesionalen Lösungsraum und wir können noch einen Normalierungsparameter frei wählen.

Für eine spätere Anwendung halten wir fest:

Bemerkung (5.21)

Gilt für ein $p \leq m$ und Koeffizienten $d_j, \ldots, d_{j+p}$

$$\forall\, t_{j+p} < t < t_{j+p+1}: \quad \sum_{i=j}^{j+p} d_i\, (t - t_i)_+^m \;=\; 0,$$

so verschwinden notwendigerweise alle Koeffizienten $d_j = \ldots = d_{j+p} \;=\; 0$.

Um eine explizite Lösungsdarstellung zu erhalten, verwenden wir die Lagrange-Darstellung des Interpolationspolynoms und nutzen aus, dass die Monome t^r, $r = 0, \ldots, m+1$, bei den Interpolationsknoten $t_j, \ldots, t_{j+m+1}$ exakt interpoliert werden, also

$$\forall\, r = 0, \ldots, m+1: \quad t^r \;=\; \sum_{i=j}^{j+m+1} t_i^r\, \ell_i(t),$$

wobei ℓ_i die entsprechenden Lagrange-Polynome bezeichnen. Vergleicht man hierin die Koeffizienten von t^{m+1}, so ergibt sich mit (3.7)

$$\forall\, r = 0, \ldots, m: \quad 0 \;=\; \sum_{i=j}^{j+m+1} t_i^r \left(\prod_{\nu=j, \nu\neq i}^{j+m+1} \frac{1}{t_i - t_\nu} \right).$$

Durch Vergleich mit (5.20) und geeigneter Normierung der d_i (die Normierung ist so, dass $\sum_{j=-m}^{n-1} B_{m,j} = 1$ ist; vgl. (5.26)) ergibt sich die folgende explizite Darstellung der B-Splines $(j = -m, \ldots, n-1)$

$$B_{m\,j}(t) \;=\; (t_{j+m+1} - t_j) \sum_{i=j}^{j+m+1} \left(\prod_{\nu=j, \nu\neq i}^{j+m+1} \frac{1}{t_\nu - t_i} \right) (t - t_i)_+^m. \tag{5.22}$$

Beispiele: $\mathbf{m = 0}$:

$$\begin{aligned} B_{0,j}(t) &= (t_{j+1} - t_j) \left\{ \frac{(t-t_j)_+^0}{t_{j+1} - t_j} + \frac{(t - t_{j+1})_+^0}{t_j - t_{j+1}} \right\} \\ &= (t-t_j)_+^0 - (t-t_{j+1})_+^0 \\ &= \begin{cases} 1, & \text{für} \quad t_j \le t < t_{j+1} \\ 0, & \text{sonst} \end{cases} \end{aligned}$$

$\mathbf{m = 1}$:

$$\begin{aligned} B_{1,j}(t) &= (t_{j+2} - t_j) \left\{ \frac{(t-t_j)_+}{(t_{j+1} - t_j)\,(t_{j+2} - t_j)} + \frac{(t - t_{j+1})_+}{(t_j - t_{j+1})\,(t_{j+2} - t_{j+1})} \right. \\ &\quad \left. + \frac{(t - t_{j+2})_+}{(t_j - t_{j+2})\,(t_{j+1} - t_{j+2})} \right\} \\ &= \begin{cases} \dfrac{t - t_j}{t_{j+1} - t_j}, & \text{für} \quad t_j \le t < t_{j+1} \\ \dfrac{t_{j+2} - t}{t_{j+2} - t_{j+1}}, & \text{für} \quad t_{j+1} \le t < t_{j+2} \\ 0, & \text{sonst} \end{cases} \end{aligned}$$

Zur numerischen Auswertung ist (5.22) natürlich nur schlecht geeignet. Hier kann man sich zunutze machen, dass die B-Splines $B_{m,j}$ analog zu den Newtonschen dividierten Differenzen und den Bernstein–Polynomen einer Dreiterm-Rekursion genügen:

Satz (5.23)

Die $B_{m,j}$ genügen der Dreiterm-Rekursion $(\ell = 1, \ldots, m,\ \ j = -m, \ldots, n+m-\ell-1)$

$$B_{\ell,j}(t) \ = \ \frac{t-t_j}{t_{j+\ell}-t_j}\, B_{\ell-1,j}(t) \ + \ \frac{t_{j+\ell+1}-t}{t_{j+\ell+1}-t_{j+1}}\, B_{\ell-1,j+1}(t)$$

Start der Rekursion:

$$B_{0,j} \ := \ \begin{cases} 1, & \text{falls} \quad t_j \le t < t_{j+1} \\ 0, & \text{sonst} \end{cases} \ , \ \ j = -m, \ldots, n+m-1.$$

Beweis: (mittels vollständiger Induktion über ℓ)

Für $\ell = 1$ rechnet man aus der Rekursion (5.23) unmittelbar die obige explizite Darstellung für $m = 1$ nach.

Für $B_{\ell-1,j}(t)$ und $B_{\ell-1,j+1}(t)$ gelten nach Induktionsvoraussetzung die entsprechenden Darstellungen (5.22). Wir haben zu zeigen, dass diese dann auch für $B_{\ell,j}$ – berechnet nach (5.23) – gilt. Dazu setzen wir die Darstellungen in die rechte Seite von (5.23) ein und formen um

$$\begin{aligned}
B_{\ell,j}(t) \ &= \ \frac{t-t_j}{t_{j+\ell}-t_j}\, B_{\ell-1,j}(t) \ + \ \frac{t_{j+\ell+1}-t}{t_{j+\ell+1}-t_{j+1}}\, B_{\ell-1,j+1}(t) \\
&= \ (t-t_j) \sum_{i=j}^{j+\ell} \left(\prod_{\nu=j,\nu\neq i}^{j+\ell} \frac{1}{t_\nu - t_i} \right) (t-t_i)_+^{\ell-1} \\
&\quad + \ (t_{j+\ell+1}-t) \sum_{i=j+1}^{j+\ell+1} \left(\prod_{\nu=j+1,\nu\neq i}^{j+\ell+1} \frac{1}{t_\nu - t_i} \right) (t-t_i)_+^{\ell-1} \\
&= \ \sum_{i=j}^{j+\ell} (t-t_j)\,(t_{j+\ell+1}-t_i) \left(\prod_{\nu=j,\nu\neq i}^{j+\ell+1} \frac{1}{t_\nu - t_i} \right) (t-t_i)_+^{\ell-1} \\
&\quad + \ \sum_{i=j+1}^{j+\ell+1} (t_{j+\ell+1}-t)\,(t_j-t_i) \left(\prod_{\nu=j,\nu\neq i}^{j+\ell+1} \frac{1}{t_\nu - t_i} \right) (t-t_i)_+^{\ell-1} \\
&= \ \sum_{i=j}^{j+\ell+1} \big\{ (t-t_j)\,(t_{j+\ell+1}-t_i) + (t_{j+\ell+1}-t)\,(t_j-t_i) \big\} \left(\prod_{\nu=j,\nu\neq i}^{j+\ell+1} \frac{1}{t_\nu - t_i} \right) (t-t_i)_+^{\ell-1}
\end{aligned}$$

Durch Ausmultiplizieren der geschweiften Klammer findet man

$$\{ \ldots \} \ = \ (t_{j+\ell+1}-t_j)\,(t-t_i)$$

und somit genau die Darstellung (5.22) für $B_{\ell,j}$. $\square$

Tableau zur Berechnung von $\mathbf{B}_{m,j}$: $(j = -m, \ldots, n-1)$

$$
\begin{array}{llll}
B_{0,j} & & & \\
B_{0,j+1} & B_{1,j} & & \\
B_{0,j+2} & B_{1,j+1} & \ddots & \\
\vdots & \vdots & & \\
B_{0,j+m} & B_{1,j+m-1} & \ldots & B_{m,j}
\end{array}
$$

Anmerkung : Zur numerischen Auswertung von (5.23) genügt es, ein *eindimensionales* Array zu verwenden, etwa $B_0, \ldots, B_m$, und das Tableau wie folgt zu speichern:

$$
\begin{array}{llll}
B_0 & & & \\
B_1 & B_0 & & \\
B_2 & B_1 & \ddots & \\
\vdots & \vdots & & \\
B_m & B_{m-1} & \ldots & B_0
\end{array}
$$

Es ergibt sich dann der folgende Algorithmus zur Berechnung von $B_{m,j}(t)$.

Algorithmus (5.24):

Für $k = 0, 1, \ldots, m$:

$$B_k := \begin{cases} 1, & \text{falls} \quad t_{j+k} \le t < t_{j+k+1}, \quad (\le \text{ bei } k = m) \\ 0, & \text{sonst;} \end{cases}$$

Für $\ell = k-1, k-2, \ldots, 0$:

$$B_\ell := \frac{t - t_{j+\ell}}{t_{j+k} - t_{j+\ell}} B_\ell + \frac{t_{j+k+1} - t}{t_{j+k+1} - t_{j+\ell+1}} B_{\ell+1};$$

$$B_{m,j}(t) = B_0.$$

Satz (5.25) : Es gilt:

$$B_{m,j}(t) \begin{cases} > 0 & \text{für} \quad t_j < t < t_{j+m+1}, \\ = 0 & \text{für} \quad t \le t_j, \quad t \ge t_{j+1+m}. \end{cases}$$

Beweis: Man sieht dies unmittelbar mittels der Rekursion (5.23) per vollständiger Induktion. □

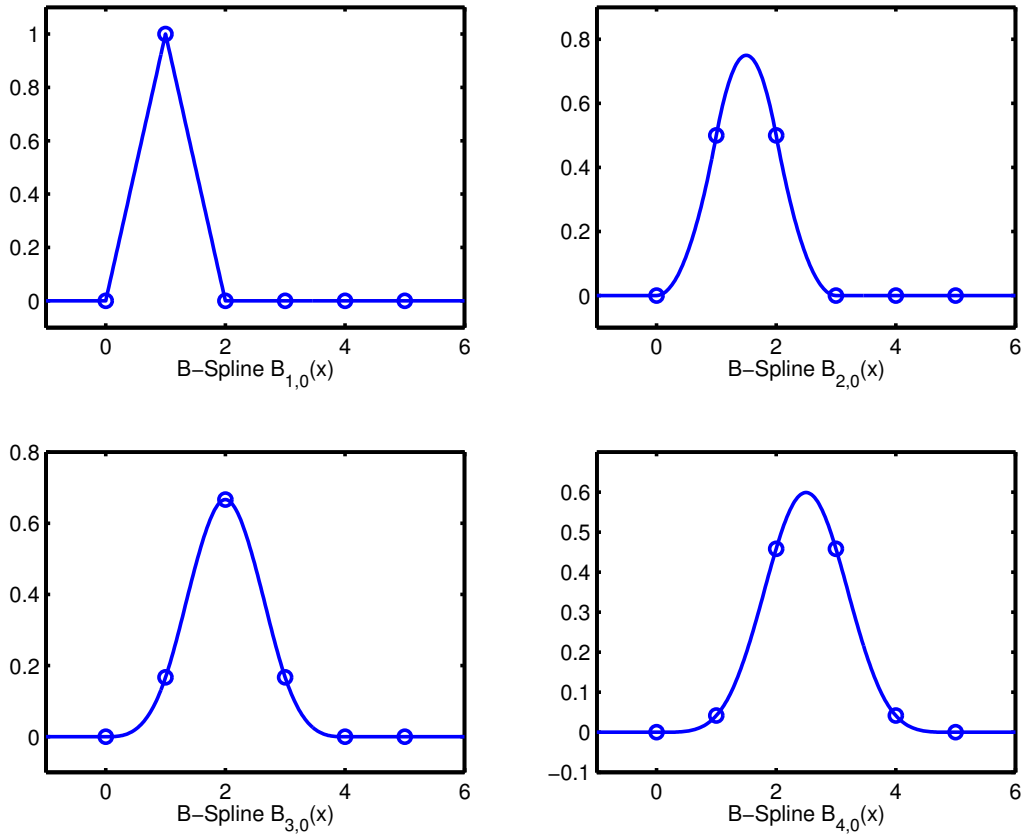


Abb. 5.4 B-Splines $B_{m,0},\ m = 1, 2, 3, 4.$

Satz (5.26) :

Auf dem Intervall $[t_0, t_n]$ bilden die B-Splines $B_{m,j},\ j = -m, \ldots, n-1$ eine *Zerlegung der Eins*:

$$\forall\, t \in [t_0, t_n] : \quad \sum_{j=-m}^{n-1} B_{m,j}(t) \;=\; 1.$$

Beweis: (per vollst. Induktion über m)

$\mathbf{m = 0}$: Klar nach Definition der $B_{0,j}$.

$\mathbf{m - 1 \Rightarrow m}$: Summation mittels (5.23) liefert

$$\begin{aligned}
\sum_{j=-m}^{n-1} B_{m,j}(t) &= \sum_{j=-m}^{n-1} \Big\{ \frac{t - t_j}{t_{j+m} - t_j}\, B_{m-1,j}(t) \;+\; \frac{t_{j+m+1} - t}{t_{j+m+1} - t_{j+1}}\, B_{m-1,j+1}(t) \Big\} \\
&= \sum_{j=-m}^{n-1} \frac{t - t_j}{t_{j+m} - t_j}\, B_{m-1,j}(t) \;+ \sum_{j=-m+1}^{n} \frac{t_{j+m} - t}{t_{j+m} - t_j}\, B_{m-1,j}(t) \\
&= \frac{t - t_{-m}}{t_0 - t_{-m}}\, B_{m-1,-m}(t) \;+ \sum_{j=-m+1}^{n-1} B_{m-1,j}(t) \;+ \frac{t_{n+m} - t}{t_{n+m} - t_n}\, B_{m-1,n}(t) \\
&= \quad 1,
\end{aligned}$$

da die Splines $B_{m-1,-m}$ und $B_{m-1,n}$ auf $[t_0, t_n]$ verschwinden und aufgrund der Induktionsannahme. $\square$

Satz (5.27) :

Die B-Splines $B_{m,j},\ j = -m, \ldots, n-1$ bilden eine Basis von $S_m(t_0, \ldots, t_n)$.

Beweis: Wegen $\dim S_m(t_0, \ldots, t_n) = m+n$ genügt es zu zeigen, dass die B-Splines $B_{m,j},\ j = -m, \ldots, n-1$, linear unabhängig sind. Gelte also mit $\alpha_j \in \mathbb{R}$:

$$\forall t \in [t_0, t_n]: \quad s(t) := \sum_{j=-m}^{n-1} \alpha_j B_{m,j}(t) = 0. \tag{1}$$

Wir haben zu zeigen, dass alle α_j verschwinden. Setzt man in die obige Gleichung die $B_{m,j}$ gemäß (5.22) ein, so erhält man nach Umsortierung der Summe eine Darstellung der Form

$$\begin{aligned} s(t) &= \sum_{j=-m}^{n-1} \alpha_j \sum_{i=j}^{j+m+1} c_{i,j}\,(t-t_i)_+^m = \sum_{j=-m}^{n-1} \alpha_j \sum_{i=-m}^{n+m} c_{i,j}\,(t-t_i)_+^m \\ &= \sum_{i=-m}^{n+m} \Big(\sum_{j=-m}^{n-1} \alpha_j\, c_{i,j} \Big)\,(t-t_i)_+^m =: \sum_{i=-m}^{n+m} d_i\,(t-t_i)_+^m. \end{aligned}$$

Hierbei bezeichnen $c_{i,j}$ die (nicht verschwindenden) Koeffizienten aus (5.22). Genauer gilt $c_{i,j} \neq 0$ für $j \leq i \leq j+m+1$ und es wird gesetzt $c_{i,j} := 0$ für alle $i < j$ und für alle $i > j+m+1$.

Betrachtet man nun die Teilsumme $\tilde{s}(t) := \sum_{i=-m}^{0} d_i\,(t-t_i)_+^m$, so gilt nach Voraussetzung $\forall\, t \in [t_0, t_1]:\ \tilde{s}(t) = 0$.

Aufgrund der Bemerkung (5.21) über den minimalen Träger eines Splines folgt hieraus aber $d_j = 0,\ j = -m, \ldots, 0$. Damit ergibt sich rekursiv

$$\begin{aligned} j = -m: \qquad & d_{-m} = \sum_{j=-m}^{n-1} \alpha_j\, c_{-m,j} = \alpha_{-m}\, c_{-m,-m} = 0 \\ & \Rightarrow \quad \alpha_{-m} = 0, \end{aligned}$$

$$\begin{aligned} j = -m+1: \quad & d_{-m+1} = \sum_{j=-m+1}^{n-1} \alpha_j\, c_{-m+1,j} = \alpha_{-m+1}\, c_{-m+1,-m+1} = 0 \\ & \Rightarrow \quad \alpha_{-m+1} = 0, \end{aligned}$$

u.s.w. Insgesamt folgt daraus $\forall\ -m \leq j \leq 0:\ \alpha_j = 0$. Es bleibt somit nach (1):

$$\forall\, t \in [t_0, t_n]: \quad s(t) := \sum_{j=1}^{n-1} \alpha_j\, B_{m,j}(t) = 0.$$

Mit (5.25) sieht man aber unmittelbar, dass diese Relation nur für $\alpha_j = 0,\ 1 \le j \le n-1$ gelten kann. □

Anmerkungen (5.28)

a) Die B-Splines $B_{m,j}$ sind C^{m-1}-Funktionen auf $\mathbb{R}$. Man kann die zusätzlichen Knoten $t_{-m}, \ldots, t_{-1}$ und $t_{n+1}, \ldots, t_{n+m}$ im Sinn eines Grenzwertes als Mehrfachknoten in t_0 bzw. t_n wählen. Hierdurch reduziert sich die Differenzierbarkeitsordnung in t_0 und t_n. Die $B_{m,j}$ lassen sich auch in diesem Fall, also bei der Knotenwahl

$$t_{-m} \;=\; \ldots \;=\; t_0 \;<\; \ldots \;<\; t_n \;=\; t_{n+1} \;=\; \ldots \;=\; t_{n+m}, \tag{5.29}$$

mit Hilfe der Dreiterm-Rekursion (5.23)

$$B_{\ell,j}(t) \;=\; \frac{t - t_j}{t_{j+\ell} - t_j}\, B_{\ell-1,j}(t) \;+\; \frac{t_{j+\ell+1} - t}{t_{j+\ell+1} - t_{j+1}}\, B_{\ell-1,j+1}(t)$$

berechnen. Dabei ist die folgende *Zusatzregel* zu berücksichtigen: Tritt in einem der beiden Summanden von (5.23) der Nenner Null auf, so ist der entsprechende Summand wegzulassen.

b) Wählt man die Knoten t_j nach (5.29) und bildet jeweils die Mittelwerte $x_j := \big(\sum_{i=1}^{m} t_{i+j}\big)/m,\ \ j = -m, \ldots, n-1$, so wird zu $f \in C[t_0, t_n]$ durch

$$s_f(t) \;:=\; \sum_{j=-m}^{n-1} f(x_j)\, B_{m,j}(t) \tag{5.30}$$

ein Spline $s_f \in S_m(t_0, \ldots, t_n)$ definiert.

s_f ist *kein* interpolierender Spline, jedoch wird durch $s_m(f) := s_f$, analog zu (4.6) und (4.18), ein positiver, linearer Approximationsoperator $s_m : \mathrm{C}[t_0, t_n] \to S_m(t_0, \ldots, t_n)$ definiert.

s_f heißt nach Schoenberg *variationsvermindernder Spline*. Anwendungen finden variationsvermindernde Splines im CAD (computer aided design).

Der Satz von Schoenberg, Whitney.

Wir betrachten die folgende **Interpolationsaufgabe (5.31)**:

Zu vorgegebenen *Spline-Knoten* $a = t_0 < \ldots < t_n = b$ und (evtl. von diesen verschiedenen) *Interpolationsknoten* $a \le x_1 < \ldots < x_{n+m} \le b$ wird zu $f \in \mathrm{C}[a, b]$ eine Splinefunktion $s_f \in S_m(t_0, \ldots, t_n)$ gesucht mit

$$\forall\, i = 1, \ldots, n+m: \quad s_f(x_i) \;=\; f(x_i).$$

Man beachte, dass Spline-Knoten und Interpolationsknoten in der obigen Formulierung der Interpolationsaufgabe durchaus verschieden sein können. Andererseits

treten keine zusätzlichen Randbedingungen wie in (5.8) auf, da die Anzahl der Interpolationsknoten mit der Dimension des Spline-Raumes übereinstimmt.

Wir fragen nun, für welche Knotenwahl (t_j), (x_i) die obige Interpolationsaufgabe eindeutig lösbar ist. Eine Antwort gibt der folgende

Satz (5.32) (Schoenberg, Whitney, 1953)

Die Interpolationsaufgabe (5.31) besitzt genau dann für jedes $f \in \mathrm{C}[a,b]$ eine eindeutige Lösung, wenn die folgende *Knotenbedingung* erfüllt ist:

$$\forall\, j = 1, 2, \ldots, n-1: \quad x_j \;<\; t_j \;<\; x_{j+m+1}. \tag{5.33}$$

Bemerkung: Man mache sich die Aussage (5.33) für den Fall $m = 1$ klar. Hier ist der Spline ein Polygonzug. Sie bedeutet:

$$x_1 \;\in\; [t_0, t_1[, \quad x_i \;\in]t_{i-2}, t_i[\;\; (i = 2, \ldots, n), \quad x_{n+1} \;\in]t_{n-1}, t_n].$$

Beweisidee zu (5.32): Wir gehen wieder von einem beidseitig erweiterten Spline-Gitter (5.18) aus. Man überlegt sich zunächst, dass die Aussage (5.33) dann äquivalent ist zu den Vorzeichenbedingungen

$$\forall\, i = 1, 2, \ldots, n+m: \quad B_{m,i-m-1}(x_i) \;>\; 0. \tag{5.34}$$

Ferner bedeutet die eindeutige Lösbarkeit der Interpolationsaufgabe, dass das folgende lineare Gleichungssystem bei beliebiger rechten Seite eine eindeutig bestimmte Lösung $\alpha_j,\ j = -m, \ldots, n-1,$ besitzt

$$\sum_{j=-m}^{n-1} \alpha_j \, B_{m,j}(x_i) \;=\; f(x_i), \quad i = 1, 2, \ldots, n+m. \tag{5.35}$$

Wir zeigen (5.35) $\Rightarrow$ (5.34): Wäre $B_{m,i_0-m-1}(x_{i_0}) = 0$ für ein $i_0 \in \{1, 2, \ldots, n+m\}$, so würde folgen $x_{i_0} \le t_{i_0-m-1}$ oder $x_{i_0} \ge t_{i_0}$.

Im ersten Fall $(x_{i_0} \le t_{i_0-m-1})$ ist $B_{m,j}(x) = 0$ für $j \ge i_0 - m - 1$ und $x \le x_{i_0}$. Die ersten i_0 Gleichungen von (5.35) lauten damit

$$\sum_{j=-m}^{i_0-m-2} \alpha_j \, B_{m,j}(x_i) \;=\; f(x_i), \quad i = 1, 2, \ldots, i_0.$$

Dies sind aber i_0 Gleichungen für $i_0 - 1$ Unbekannte. Es gibt also rechte Seiten $f(x_i)$, für die das Gleichungssystem keine Lösung besitzt.

Analog haben im zweiten Fall $(x_{i_0} \ge t_{i_0})$ die letzten $n + m + 1 - i_0$ Gleichungen von (5.35) die Form

$$\sum_{j=i_0-m}^{n-1} \alpha_j \, B_{m,j}(x_i) \;=\; f(x_i), \quad i = i_0, i_0+1, \ldots, n+m.$$

Auch dieses Gleichungssystem hat weniger Unbekannte als Gleichungen, ist also nicht für alle rechten Seiten lösbar.

Zur Umkehrung zeigt man, dass das homogene lineare Gleichungssystem (also $f(x_i) = 0$ in (5.35)) unter der Voraussetzung (5.34) nur die triviale Lösung $\alpha_j = 0$ besitzt.

Auf diesen technischeren Teil des Beweises verzichten wir hier und verweisen dazu auf das Lehrbuch von Powell. □

Bemerkungen (5.36)

a) Unter der Voraussetzung des Schoenberg, Whitneyschen Satzes ist das für die numerische Berechnung zu lösende lineare Gleichungssystem (5.35) eindeutig lösbar, die Koeffizientenmatrix ist also regulär, nicht negativ und sie besitzt zudem Bandstruktur. Genauer: $B_{m,j-m-1}(x_i) = 0$, für $|i - j| \geq m + 1$. Das lineare Gleichungssystem lässt sich mit dem Gaußschen Eliminationsverfahren (ohne Pivotsuche!) effizient und numerisch stabil lösen.

b) Haben Spline- und Interpolationsknoten die im Schoenberg, Whitneyschen Satz geforderte Anordnung, so ist durch die Zuordnung

$$s_m : \mathrm{C}[a,b] \to S_m(t_0,\ldots,t_n), \quad s_m(f) = \sum_{j=-m}^{n-1} \alpha_j B_{m,j}, \tag{5.37}$$

mit α_j nach (5.35), ein *stetiger linearer Projektor* gegeben. Insbesondere lässt sich die Güte der Spline- Approximation mit dem Lemma von Lebesgue abschätzen

$$\|f - s_m(f)\|_\infty \leq \big(1 + \|s_m\|_\infty\big)\, d_{S_m(t_0,\ldots,t_n)} f \tag{5.38}$$

ähnlich wie bei der Interpolation durch Polynome ist die Wahl der Interpolationsknoten für die Größe der Operatornorm entscheidend.

c) Wählt man die Spline-Knoten gemäß (5.29) (also mehrfache Knoten in den beiden Intervallenden) und die Interpolationsknoten gemäß

$$x_i := (t_{i-m} + \ldots + t_{i-1})/m, \quad i = 1,\ldots,n+m, \tag{5.39}$$

so sind die Voraussetzungen des Schoenberg, Whitneyschen Satzes erfüllt.

Für $m = 2$ gilt dann die folgende Abschätzung für die Operatornorm $\|s_2\|_\infty \leq 2$.

Für $m = 3$ gilt unter gleichen Voraussetzungen $\|s_3\|_\infty \leq 27$ (de Boor, 1975).

d) Wir beschliessen den Abschnitt mit zwei Abschätzungen für die Minimalabweichung $d_{S_m(\Delta)} f$ einer stetigen Funktion $f \in \mathrm{C}[a,b]$ zum Splineraum $S_m(\Delta)$, wobei wie bisher $\Delta := \{a = t_0 < \ldots < t_n = b\}$ das Spline-Gitter und

$$h := \|\Delta\| := \max\{|t_{j+1} - t_j| : \ j = 0,\ldots,n-1\} \tag{5.40}$$

den maximale Knotenabstand (die *Feinheit* des Gitters) bezeichnet.

Allgemein lässt sich dann die folgende Abschätzung zeigen

$$f \in \mathrm{C}[a,b] \;\Rightarrow\; d_{S_m(\Delta)} f \;\leq\; \omega_f((m+1)\,h/2). \tag{5.41}$$

Dabei ist $\omega_f(\delta) := \sup\{|f(x)-f(y)| : |x-y| \leq \delta\}$ der *Stetigkeitsmodul* der Funktion f.

Insbesondere konvergiert die Minimalabweichung gegen Null mit $h \downarrow 0$. Der Raum der Splinefunktionen vom Grad m liegt also dicht in $\mathrm{C}[a,b]$.

Ist $f \in \mathrm{C}^1[a,b]$ sogar eine C^1-Funktion, so gilt bekanntlich (Mittelwertsatz) $\omega_f(\delta) \leq \|f^{(1)}\|_\infty\,\delta$ und damit folgt aus (5.41)

$$f \in \mathrm{C}^1[a,b] \;\Rightarrow\; d_{S_m(\Delta)} f \;\leq\; \frac{m+1}{2}\,\|f^{(1)}\|_\infty\,\|\Delta\|. \tag{5.42}$$

In Verallgemeinerung von (5.42) gilt für eine C^k-Funktion mit $k \leq m+1$

$$f \in \mathrm{C}^k[a,b],\;\; k \leq (m+1) \;\Rightarrow\; d_{S_m(\Delta)} f \;\leq\; \frac{(m+1)!}{2^k\,(m+1-k)!}\,\|f^{(k)}\|_\infty\,\|\Delta\|^k. \tag{5.43}$$

Aus S. Karlin: To I. J. Schoenberg and his mathematics, J. Approx. Theory, Vol.8, 1973:

... Schoenberg is noted worldwide for his realisation of the importance of spline functions for general mathematical analysis and in approximation theory, their key relevance in numerical procedures for solving differential equations with initial and/or boundary conditions, and their role in the solution of a whole host of variational problems. The fundamental papers by Schoenberg [two papers in 1946] form a monument in the history of the subject as well as its inauguration.

Aus R. Askey and C. de Boor: Im Memoriam: I. J. Schoenberg (1903–1990), J. Approx. Theory, Vol. 63, 1990:

For the next 15 years, Schoenberg had splines all to himself. This changed around 1960, when computers became more widespread and splines first assumed their role as the premier tool for data fitting and computer-aided geometric design. Schoenberg's more than 40 papers on splines after 1960 gave much impetus to the rapid development of the field.

6. L_2 - Approximation

Euklidische Räume.

Im Folgenden sei $(R, \langle\cdot,\cdot\rangle)$ ein Euklidischer Vektorraum, also ein reeller, linearer Raum mit einem Skalarprodukt. Bekanntlich lauten die Skalarprodukt-Axiome

$$\begin{aligned} \langle f,f\rangle &\geq 0, \quad \langle f,f\rangle = 0 \Leftrightarrow f = 0, \\ \langle f,g\rangle &= \langle g,f\rangle, \\ \langle \alpha f_1 + \beta f_2, g\rangle &= \alpha\langle f_1,g\rangle + \beta\langle f_2,g\rangle. \end{aligned} \tag{6.1}$$

Durch $\|f\| := \sqrt{\langle f,f\rangle}$ wird *die zugehörige Norm* auf R definiert.

Cauchy-Schwarzsche Ungleichung:

$$|\langle f,g\rangle| \leq \|f\|\,\|g\|; \tag{6.2}$$

dabei gilt die Gleichheit genau dann, wenn f, g linear abhängig sind.

Parallelogrammgleichung:

$$\|f+g\|^2 + \|f-g\|^2 = 2\left(\|f\|^2 + \|g\|^2\right). \tag{6.3}$$

Die Norm eines Euklidischen Raumes ist stets strikt konvex, vgl. (2.11), damit ist die Eindeutigkeit der linearen Approximationsaufgabe gesichert, vgl. (2.14).

Charakterisierungssatz (6.4)

Ist $f \in R$ und ist V ein linearer Teilraum von R, so gilt: $p^* \in V$ ist genau dann eine Bestapproximation von f aus V, wenn der Fehler $e := f - p^*$ auf V senkrecht steht, also für alle $p \in V$ gilt: $\langle e,p\rangle = 0$.

Beweis: Siehe Satz (2.4). Die Abgeschlossenheit von V und die Vollständigkeit von R wurde dabei lediglich für die Existenz einer Bestapproximation benötigt, vgl. auch (2.3). □

Aus dem Charakterisierungssatz folgt unmittelbar die folgende Relation (*Satz des Pythagoras*) für eine Bestapproximation p^* von f

$$\|f\|^2 = \|f-p^*\|^2 + \|p^*\|^2. \tag{6.5}$$

Im folgenden Satz beschreiben wir, wie sich die Bestapproximation im Fall eines *endlich* dimensionalen linearen Raumes V berechnen lässt.

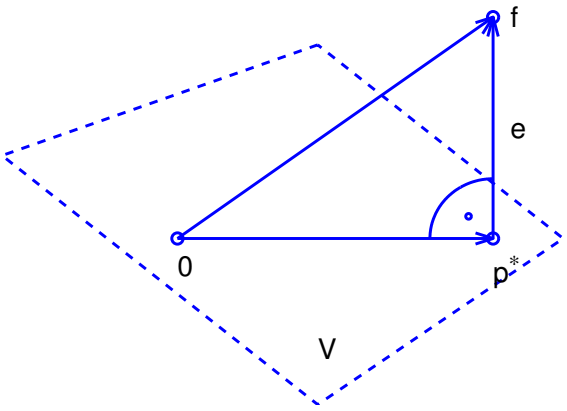


Abb. 6.1 Charakterisierungssatz.

Satz (6.6)

Sei V ein endlich dimensionaler linearer Teilraum von R.

a) Ist $(h_0, \ldots, h_n)$ eine Basis von V, also $\dim(V) = n+1$, so ist die Bestapproximation $p^* = \sum_0^n \alpha_j\, h_j$ von f aus V gegeben durch das lineare Gleichungssystem

$$\begin{bmatrix} \langle h_0, h_0\rangle & \ldots & \langle h_n, h_0\rangle \\ \vdots & & \vdots \\ \langle h_0, h_n\rangle & \ldots & \langle h_n, h_n\rangle \end{bmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \langle f, h_0\rangle \\ \vdots \\ \langle f, h_n\rangle \end{pmatrix}. \tag{6.7}$$

Die Koeffizientenmatrix $G(h_0, \ldots, h_n) \in \mathbb{R}^{(n+1,n+1)}$ ist symmetrisch und regulär (beachte Existenzsatz II), sie heißt *Gramsche Matrix*.

b) Ist $(q_0, \ldots, q_n)$ eine Orthonormalbasis (ONB) von V, so ist die Bestapproximation gegeben durch

$$p^* \;=\; \sum_{j=0}^{n} \langle f, q_j\rangle\, q_j. \tag{6.8}$$

(6.8) heißt auch *Fourier-Entwicklung* von f bzgl. der ONB $(h_0, \ldots, h_n)$, die Koeffizienten $\langle f, q_j\rangle$ heißen auch *Fourier-Koeffizienten*.

Beweis: zu a): Nach (6.4) ist $p^* = \sum \alpha_j\, h_j$ genau dann Bestapproximation, wenn gilt

$$\forall\, k = 0, 1, \ldots, n: \qquad \langle f - \textstyle\sum_j \alpha_j\, h_j, h_k\rangle \;=\; 0$$

$$\Leftrightarrow \quad \forall\, k = 0, 1, \ldots, n: \qquad \sum_j \langle h_j, h_k\rangle\, \alpha_j \;=\; \langle f, h_k\rangle.$$

Dies ist ein eindeutig lösbares lineares Gleichungssystem zur Bestimmung der α_j.

zu b) Im Fall einer ONB wird die Koeffizientenmatrix zur Einheitsmatrix $G(q_0, \ldots, q_n) = I_{n+1}$, und damit folgt $\alpha_k = \langle f, q_k\rangle$ für alle $k = 0, 1, \ldots, n$.

□

Bemerkung (6.9)

Sei $(q_0, \ldots, q_n)$ ONB von V. Wegen (6.5) folgt für die Bestapproximation p^* von f aus V

$$d_V(f)^2 \;=\; \|f - p^*\|^2 \;=\; \|f\|^2 \;-\; \|p^*\|^2 \;=\; \|f\|^2 \;-\; \sum_{k=0}^{n} |\langle f, q_k\rangle|^2$$

und damit insbesondere die *Besselsche Ungleichung*

$$\sum_{k=0}^{n} |\langle f, q_k\rangle|^2 \;\leq\; \|f\|^2. \tag{6.10}$$

Gleichheit liegt genau dann vor, wenn $f \in V$.

Folgerung (6.11)

Ist $(q_k)_{k\in\mathbb{N}}$ eine Folge orthonormaler Vektoren aus R und $V_n := \text{Spann}(q_0, \ldots, q_n)$, so folgt für $f \in R$ aus der Besselschen Ungleichung, dass die Reihe $\sum_0^\infty |\langle f, q_k\rangle|^2$ (absolut) konvergiert, insbesondere gilt damit

$$\lim_{n\to\infty} \langle f, q_n\rangle \;=\; 0, \tag{6.12}$$

die Fourier-Koeffizienten bilden also eine Nullfolge!

Gilt darüber hinaus, dass der lineare Teilraum $V := \bigcup V_n$ in R dicht liegt (vgl. die Weierstraßschen Approximationssätze bzw. die Aussagen über die Splineräume), so gilt $\lim_{n\to\infty} d_{V_n}(f) = 0$. Damit folgt aus (6.9) die so genannte *Parseval-Gleichung*

$$\sum_{k=0}^{\infty} |\langle f, q_k\rangle|^2 \;=\; \|f\|^2. \tag{6.13}$$

Beispiel (6.14)

Die Funktionen

$$\frac{1}{\sqrt{2}},\;\; \cos t,\;\; \sin t,\;\; \sin(2\,t),\;\; \cos(2\,t), \ldots$$

bilden ein Orthonormalsystem bzgl. des inneren Produktes

$$\langle f, g\rangle \;=\; \frac{1}{\pi} \int_0^{2\,\pi} f(t)\, g(t)\, dt, \qquad f,\, g \in \mathrm{C}_{2\,\pi}.$$

Die Bestapproximation einer Funktion $f \in \mathrm{C}_{2\,\pi}$ bezüglich des Raumes T_n (vgl. (4.14)) lautet damit

$$\begin{aligned} p^*(t) \;&=\; \frac{a_0}{2} \;+\; \sum_{k=1}^{n} \left[a_k \cos(k\,t) \;+\; b_k \sin(k\,t)\right], \\ a_k \;&=\; \langle f, \cos(kt)\rangle, \quad b_k \;=\; \langle f, \sin(kt)\rangle. \end{aligned}$$

Beispiel (6.15)

Gegeben seien die Meßdaten

t_k :	1	2	3
y_k :	2.0	2.8	4.2

Gesucht ist eine bestapproximierende Gerade $p(t) = c_0 + c_1 t$ bezüglich der gewichteten Norm ($m > 0$)

$$\|g\|^2 \;:=\; m\,g(t_0)^2 \;+\; 10\,g(t_1)^2 \;+\; 10\,g(t_2)^2; \quad g \in R := C(\{t_0, t_1, t_2\}).$$

Offensichtlich ist die Norm zugehörig zu dem entsprechenden Skalarprodukt, so dass wir die L_2-Theorie anwenden können.

Wir wählen zunächst die Basis $h_0(t) = 1,\ h_1(t) = t$. Das lineare Gleichungssystem (6.7) lautet dann

$$\begin{bmatrix} m+20 & m+50 \\ m+50 & m+130 \end{bmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} \;=\; \begin{pmatrix} 2\,m+70 \\ 2\,m+182 \end{pmatrix}.$$

Dieses Gleichungssystem entspricht der Normalgleichung des linearen Ausgleichsproblems. Man sieht, dass die Koeffizientenmatrix für große Parameter m *schlecht konditioniert* ist. Die Berechnung der Bestapproximation über diesen Weg ist daher *numerisch instabil*.

Die Lösung des obigen Gleichungssystems lautet

$$c_0 \;=\; \frac{0.96\,m}{m+2}, \quad c_1 \;=\; \frac{1.04\,m+2.8}{m+2}.$$

Alternativ könnte man eine orthogonale Basis wählen, die sich mit dem Gram-Schmidtschen Verfahren aus (h_0, h_1) gewinnen lässt (auf die Normierung wird hier verzichtet). Man erhält

$$p_0(t) \;=\; 1, \qquad p_1(t) \;=\; t-\alpha, \quad \alpha \;=\; \frac{m+50}{m+20}.$$

Das zugehörige lineare Gleichungssystem (6.7) erhält dann die Form

$$\begin{bmatrix} m+20 & 0 \\ 0 & \dfrac{50\,m+100}{m+20} \end{bmatrix} \begin{pmatrix} \tilde{c}_0 \\ \tilde{c}_1 \end{pmatrix} \;=\; \begin{pmatrix} 2\,m+70 \\ \dfrac{52\,m+140}{m+20} \end{pmatrix}.$$

Hieraus lassen sich die $\tilde{c}_j$ numerisch stabil berechnen. Durch Umskalierung der ersten Gleichung lässt sich zudem erreichen, dass die Konditionszahl der Koeffizientenmatrix für $m \to \infty$ beschränkt bleibt.

Beispiel (6.16)

Sei $R := \mathrm{C}[0,1]$, $\langle f,g\rangle := \int_0^1 f(t)g(t)\,dt$ und $V_n := \Pi_n[0,1]$.

Wählt man die Monome $(1,\, t, \ldots,\, t^n)$ als Basis von V_n, so erhält man für die Gramsche Matrix

$$G(1,t,\ldots,t^n) \;=\; \begin{bmatrix} 1 & 1/2 & \ldots & 1/(n+1) \\ 1/2 & 1/3 & \ldots & 1/(n+2) \\ \vdots & & & \vdots \\ 1/(n+1) & & \ldots & 1/(2n+1) \end{bmatrix}.$$

Dies ist bekanntlich die *Hilbert-Matrix*, die für größere Werte von n schlecht konditioniert ist.

Als Abhilfe ergibt sich auch hier die Möglichkeit, eine orthogonale Basis mit Hilfe des Gram-Schmidtschen Verfahrens zu konstruieren. Man erhält hierfür umskalierte Legendre-Polynome.

Orthogonale Polynome.

Sei $[a,b] \subset \mathbb{R}$ ein kompaktes Intervall mit $a < b$. Sei $R := \mathrm{C}[a,b]$ und $\omega :]a,b[\to \mathbb{R}$ eine *stetige, positive* Gewichtsfunktion, für die das Integral $\int_a^b \omega(t) f(t)\,dt$ für alle Funktionen $f \in \mathrm{C}[a,b]$ (im uneigentlichen Sinn) existiert.

Durch

$$\langle f,g\rangle \;:=\; \int_a^b \omega(t)\, f(t)\, g(t)\, dt \tag{6.17}$$

wird dann ein Skalarprodukt auf R erklärt.

Als Teilräume für eine L_2-Approximation untersuchen wir die Polynomräume $V_n := \Pi_n[a,b]$ und bestimmen hierzu eine Orthogonal- bzw. Orthonormalbasis von V_n bzgl. $\langle\cdot,\cdot\rangle$. Da die Normierung in der üblichen Form, nämlich $\|q_k\| = 1$, noch Freiheiten lässt und zudem kompliziertere Wurzel auftreten, wählen wir statt dessen die *Normierungsbedingung*

$$p_k \;=\; t^k + a_{k-1}t^{k-1} + \ldots + a_0, \tag{6.18}$$

also höchster Koeffizient $=$ 1, die die Eindeutigkeit erzwingt und zudem $p_k \in \Pi_k \setminus \Pi_{k-1}$ ergibt.

Satz (6.19) (Über Orthogonalpolynome)

a) Es gibt eine eindeutig bestimmte Folge von Orthogonalpolynomem $(p_k)_{k\in\mathbb{N}_0}$ mit $p_k \in \Pi_k \setminus \Pi_{k-1}$, $\langle p_j, p_k\rangle = 0$ für $j \neq k$, und der Normierungsbedingung (6.18).

b) Die (p_k) genügen der *Dreitermrekursion*

$$\begin{aligned} p_k(t) &= (t-a_k)\, p_{k-1}(t) - b_k\, p_{k-2}(t), \quad k = 2,3,\ldots, \\ a_k &= \frac{\langle t\, p_{k-1}, p_{k-1}\rangle}{\langle p_{k-1}, p_{k-1}\rangle}, \quad b_k = \frac{\langle t\, p_{k-1}, p_{k-2}\rangle}{\langle p_{k-2}, p_{k-2}\rangle}, \end{aligned} \tag{6.20}$$

Startwerte sind $p_0(t) := 1$, $p_1(t) := t - a_1$.

c) Das Orthogonalpolynom p_k besitzt genau k einfache Nullstellen, die sämtlich im offenen Intervall $]a,b[$ liegen.

Beweis: zu a), b): Mit $k \geq 1$ und $p_0 := 1$ seien bereits normierte, orthogonale Polynome $p_0, \ldots, p_{k-1}$ konstruiert worden. Für das gesuchte Polynom p_k gilt dann aufgrund der Normierungsbedingung $p_k - t\, p_{k-1} \in \Pi_{k-1}$ und somit nach Satz (6.6)b) (Fourier-Entwicklung)

$$p_k - t\, p_{k-1} = \sum_{j=0}^{k-1} \frac{\langle p_k - t p_{k-1}, p_j\rangle}{\langle p_j, p_j\rangle}\, p_j = -\sum_{j=0}^{k-1} \frac{\langle t p_{k-1}, p_j\rangle}{\langle p_j, p_j\rangle}\, p_j\,.$$

Die letzte Gleichheit besteht, da p_k senkrecht auf Π_{k-1} stehen soll.

Für $j < k-2$ gilt nun $\langle t\, p_{k-1}, p_j\rangle = \langle p_{k-1}, t\, p_j\rangle = 0$, da $t\, p_j \in \Pi_{k-2}$. Damit bleiben von der obigen Summe nur zwei Summanden bestehen und wir haben

$$\begin{aligned} p_k &= \Big(t - \frac{\langle t p_{k-1}, p_{k-1}\rangle}{\langle p_{k-1}, p_{k-2}\rangle}\Big)\, p_{k-1} - \frac{\langle t p_{k-1}, p_{k-2}\rangle}{\langle p_{k-2}, p_{k-2}\rangle}\, p_{k-2} \\ &= (t-a_k)\, p_{k-1} - b_k\, p_{k-2}. \end{aligned}$$

Umgekehrt zeigt man auch unmittelbar, dass durch obige Relation ein normiertes Polynom p_k definiert wird, das auf Π_{k-1} senkrecht steht.

zu c): Seien $t_1 < t_2 < \ldots < t_m \in]a,b[$ die $m \in \{0,1,\ldots,k\}$ Punkte, an denen p_k das Vorzeichen wechselt. Man beachte: Alle Nullstellen eines Polynoms $p \neq 0$ sind isolierte Nullstellen! Nullstellen, an denen kein Vorzeichenwechsel stattfindet werden nicht mitgezählt. Somit wechselt $q(t) := (t-t_1)\ldots(t-t_m)$ an den gleichen Stellen das Vorzeichen. Damit folgt, dass $\omega \cdot q \cdot p_k$ überhaupt keinen Vorzeichenwechsel in $]a,b[$ besitzt, also

$$\langle q, p_k\rangle = \int_a^b \omega(t)\, q(t)\, p_k(t)\, dt \neq 0.$$

gilt.

Nun steht p_k nach Konstruktion auf Π_{k-1} senkrecht, damit muss $m = \mathrm{grad}(q) \geq k$, also $m = k$ sein. Da p_k als Polynom k-ten Grades aber nicht mehr als k Nullstellen

besitzen kann, sind die $t_1, \ldots, t_m$ sämtliche und lauter einfache Nullstellen von p_k, die zudem alle in $]a, b[$ liegen. □

Beispiel (6.21) (Legendre-Polynome)

Hier ist $[a, b] = [-1, 1]$, $\omega(t) = 1$. Die Orthogonalpolynome sind gegeben durch die *Formel von Rodrigues*

$$P_n(t) = \frac{1}{2^n\, n!}\, \frac{d^n}{d\, t^n}\, [(t^2-1)^n] \tag{6.22}$$

Normierung:

$$P_n(t) = \frac{1 \cdot 3 \cdot 5 \ldots (2n-1)}{n!}\, t^n + \ldots$$

$$\int_{-1}^{1} P_k(t)\, P_\ell(t)\, dt = \begin{cases} 0, & \text{für } k \neq \ell, \\ 2/(2n+1), & \text{für } k = \ell. \end{cases}$$

Dreitermrekursion:

$$\begin{aligned} &P_{n+1}(t) = \frac{2n+1}{n+1}\, t\, P_n(t) - \frac{n}{n+1}\, P_{n-1}(t), \quad n \geq 1, \\ &P_0(t) = 1, \quad P_1(t) = t. \end{aligned} \tag{6.23}$$

Differentialgleichung: (Legendresche Differentialgleichung)

$$(t^2-1)\, y'' + 2\, t\, y' - n(n+1)\, y = 0, \quad n \geq 0. \tag{6.24}$$

Zweite Lösung von (6.24): $Q_n(t) = P_n(t) \int \frac{d\, t}{(t^2-1)\, P_n^2(t)}$.

Beispiel (6.25) (Tschebyscheff-Polynome)

Hier ist $[a, b] = [-1, 1]$, $\omega(t) = 1/\sqrt{1-t^2}$. Die Orthogonalpolynome sind gegeben durch

$$T_n(t) = \cos[n \arccos t], \quad t \in [-1, 1], \tag{6.26}$$

Normierung:

$$T_n(t) = 2^{n-1}\, t^n + \ldots$$

$$\int_{-1}^{1} \frac{T_k(t)\, T_\ell(t)}{\sqrt{1-t^2}}\, dt = \begin{cases} 0, & \text{für } k \neq \ell, \\ \pi/2, & \text{für } k = \ell \neq 0, \\ \pi & \text{für } k = \ell = 0. \end{cases}$$

Dreitermrekursion:

$$\begin{aligned} &T_{n+1}(t) = 2\, t\, T_n(t) - T_{n-1}(t), \quad n \geq 1, \\ &T_0(t) = 1, \quad T_1(t) = t. \end{aligned} \tag{6.27}$$

Differentialgleichung: (Tschebyscheffsche Differentialgleichung)

$$(1-t^2)\,y'' \;-\; t\,y' \;-n^2\,y \;=\; 0, \quad n \geq 0. \tag{6.28}$$

Viele weitere Beispiele (auch für unbeschränkte Intervalle) findet man im Handbuch mathematischer Funktionen von Abramowitz und Stegun.

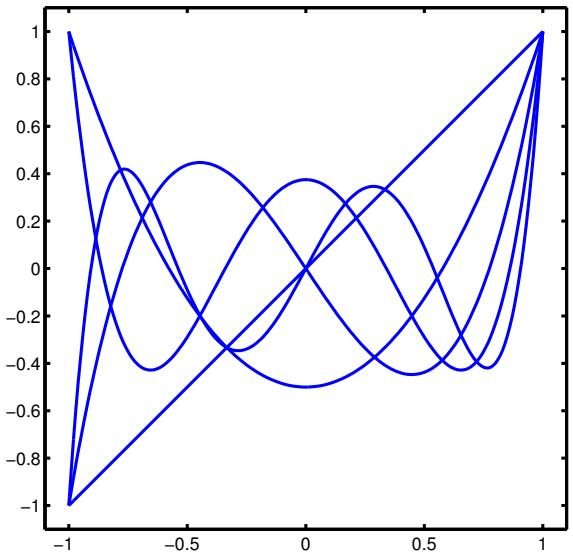


Abb. 6.2 Legendre Polynome P_n, $n = 1, \ldots, 5$.

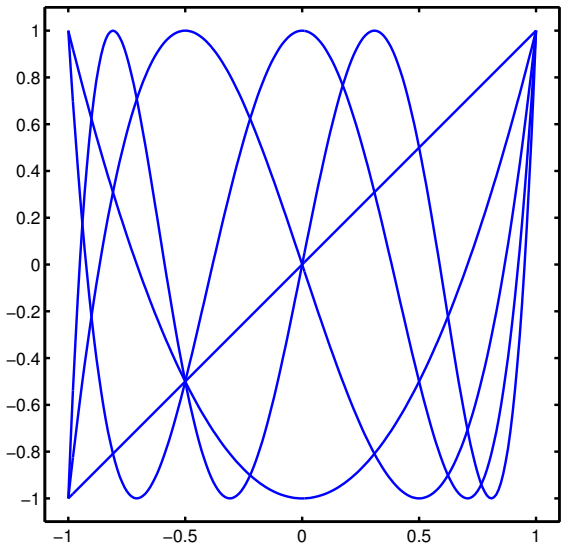


Abb. 6.3 Tschebyscheff Polynome T_n, $n = 1, \ldots, 5$.

Satz (6.29)

Für die normierten Orthogonalpolynome aus (6.19) gilt: p_n minimiert die Norm $\|p\|$ über allen Polynomen $p \in \Pi_n$, die die Normierungsbedingung (6.18) erfüllen.

Beweis: Mit Π_n^0 sei der affin-lineare Teilraum der Polynome aus Π_n bezeichnet, die die Normalisierungsbedingung (6.18) erfüllen. Anstelle der Minimierung von $\|p\|$ über Π_n^0 lässt sich äquivalent $\|p_n - q\|$ über $q \in \Pi_{n-1}$ minimieren. Wir suchen also eine Bestapproximation von p_n aus dem linearen Teilraum Π_{n-1}.

Nach Satz (6.6) hat man für die Bestapproximation q^* die Fourier-Entwicklung

$$q^* \;=\; \sum_{k=0}^{n-1} \frac{\langle p_n, p_k\rangle}{\langle p_k, p_k\rangle}\, p_k.$$

Aufgrund der Orthogonalitätseigenschaft ist damit $q^* = 0$. □

Satz (6.30) (Algorithmus von Clenshaw)

Eine Entwicklung nach Orthogonalpolynomen $f_n(t) = \sum_0^n c_k p_k(t)$, wobei p_k die normierten Orthogonalpolynome gemäß (6.19) sind, lässt sich mit dem folgenden Algorithmus von Clenshaw auswerten

$$\begin{aligned}
&z_n \;:=\; c_n;\\
&z_{n-1} \;:=\; c_{n-1} \,+\, (t - a_n)\, z_n;\\
&\text{für } k \;=\; n-2 : -1 : 0\\
&\qquad z_k \;:=\; c_k \,+\, (t - a_{k+1})\, z_{k+1} \,-\, b_{k+2}\, z_{k+2};\\
&\text{end } k;\\
&f_n \;:=\; z_0;
\end{aligned}$$

Beweis: Nach (6.19) gilt die Dreitermrekursion

$$\begin{aligned}
p_0 \;&=\; 1, \quad p_1 \;=\; (t - a_1)\\
p_k \;&=\; (t - a_k)\, p_{k-1} \,-\, b_k\, p_{k-2}, \quad k \geq 2
\end{aligned}$$

In Matrixschreibweise lautet diese $A\,p \;=\; r$ mit

$$\begin{bmatrix}
1 & & & & \\
-(t-a_1) & 1 & & & \\
b_2 & -(t-a_2) & 1 & & \\
& \ddots & \ddots & \ddots & \\
& & b_n & -(t-a_n) & 1
\end{bmatrix}
\begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}
\;=\;
\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Mit $c := (c_0, \ldots, c_n)^{\mathrm{T}}$ gilt nun

$$f_n \;=\; p^{\mathrm{T}} c \;=\; (A^{-1} r)^{\mathrm{T}} c \;=\; r^{\mathrm{T}} A^{-\mathrm{T}} c \;=\; r^{\mathrm{T}} z \;=\; z_0,$$

wobei $z := A^{-\mathrm{T}} c$ oder $A^{\mathrm{T}} z = c$. Dieses lineare Gleichungssystem lautet explizit

$$\begin{bmatrix} 1 & -(t-a_1) & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -(t-a_{n-1}) & b_n & \\ & & & 1 & -(t-a_n) & \\ & & & & 1 & \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Die Rückwärtsubstitution für dieses lineare Gleichungssystem ergibt die Rekursion aus der Behauptung des Satzes. □

Tschebyscheff – Entwicklung.

Wie im letzten Teilabschnitt sei $R := \mathrm{C}[a,b]$ und $\omega :]a,b[\to \mathbb{R}$ eine stetige und positive Gewichtsfunktion, für die das Integral $\int_a^b \omega(t) f(t)\, dt$ für alle $f \in \mathrm{C}[a,b]$ existiert. Wieder sei mit $\langle \cdot, \cdot \rangle$ das Skalarprodukt (6.17) bezeichnet und $\| \cdot \|$ sei die zugehörige Norm. Sind (p_n) die gemäß Satz (6.19) (eindeutig bestimmten) Orthogonalpolynome, so werde mit $q_n := p_n / \|p_n\|$ das zugehörige *Orthonormalsystem* bezeichnet. Wir haben dann $q_n \in \Pi_n \setminus \Pi_{n-1}$ und $\|q_n\| = 1$.

Zu $f \in R$ ist damit gemäß Satz (6.6) die Bestapproximation aus Π_n (bzgl. $\| \cdot \|$) gegeben durch die Fourier-Entwicklung

$$R_n(f) \;:=\; \sum_{k=0}^{n} c_k\, q_k, \quad c_k \;:=\; \langle f, q_k \rangle. \tag{6.31}$$

Man beachte, dass die Fourier-Koeffizienten c_k unabhängig von n sind.

Zu $f \in R$ lässt sich damit die folgende *formale Orthogonalentwicklung* definieren

$$f \;\sim\; \sum_{k=0}^{\infty} \langle f, q_k \rangle\, q_k. \tag{6.32}$$

Wir fragen nun nach der Konvergenz der rechten Seite in (6.32) und nach hinreichenden Bedingungen, unter denen in (6.32) Gleichheit gilt.

Satz (6.33)

Zu $f \in \mathrm{C}[a,b]$ bezeichne $T_n(f)$ die (bzw. eine) Bestapproximation von f aus Π_n bezüglich $\|\cdot\|_\infty$ und $R_n(f)$ die Bestapproximation von f bezüglich $\|\cdot\|$. Dann gelten

$$\begin{aligned}
&\textbf{a)} && \|f - T_n(f)\|_\infty \;\to\; 0 && (n \to \infty),\\
&\textbf{b)} && \|f - T_n(f)\| \;\to\; 0 && (n \to \infty),\\
&\textbf{c)} && \|f - R_n(f)\| \;\to\; 0 && (n \to \infty).
\end{aligned}$$

Man beachte, dass i. Allg. *nicht* $\|f - R_n(f)\|_\infty \to 0$ folgt.

Beweis:

zu a): Weierstraßscher Approximationssatz.

zu b):

$$\begin{aligned}
\|f - T_n(f)\|^2 &= \int_a^b \big(f(t) - T_n(f)(t)\big)^2\, \omega(t)\, dt\\
&\le \|f - T_n(f)\|_\infty^2 \int_a^b \omega(t)\, dt\\
\Rightarrow \qquad \|f - T_n(f)\| &\le \sqrt{\int_a^b \omega(t)\, dt}\;\; \|f - T_n(f)\|_\infty.
\end{aligned}$$

zu c): $\|f - R_n(f)\| \le \|f - T_n(f)\|.$ □

Satz (6.34)

Die *Tschebyscheff-Entwicklung* einer Funktion $f \in \mathrm{C}[-1,1]$ ist gegeben durch

$$R_n(f)(x) := \frac{a_0}{2} + \sum_{k=1}^{n} a_k\, T_k(x), \qquad a_k := \frac{2}{\pi} \int_{-1}^{1} \frac{f(x) T_k(x)}{\sqrt{1-x^2}}\, dx$$

Ist $f \in \mathrm{C}^2[-1,1]$, so konvergiert $R_n(f)$ gleichmäßig und absolut auf $[-1,1]$ gegen f.

Beweis: Mit $x = \cos\theta$, $\tilde f(\theta) := f(\cos\theta)$ gilt $a_k = \dfrac{2}{\pi} \displaystyle\int_0^\pi \tilde f(\theta)\, \cos(k\,\theta)\, d\theta.$

Für $k > 0$ liefert die zweimalige partielle Integration

$$a_k = -\frac{2}{\pi\, k^2} \int_0^\pi \tilde f''(\theta)\, \cos(k\,\theta)\, d\theta,$$

wobei die ausintegrierten Bestandteile verschwinden. Damit gilt die Abschätzung

$|a_k| \leq M/k^2$ und somit nach dem Majorantenkriterium die gleichmäßige und absolute Konvergenz von $R_n(f)$.

Die Grenzfunktion $F := \lim_{n\to\infty} R_n(f)$ ist somit (als gleichmäßiger Limes stetiger Funktionen) stetig und es folgt

$$\|f - F\| \;\leq\; \|f - R_n(f)\| \;+\; \|R_n(f) - F\|.$$

Der erste Summand konvergiert nach (6.33) gegen Null, der zweite aufgrund der Definition von F. Somit ist also $F = f$. □

Bemerkung (6.35)

An obigem Beweis erkennt man, dass die T-Entwicklung einer Funktion $f \in \mathrm{C}[-1,1]$ direkt mit der Fourier-Entwicklung einer $2\,\pi$-periodischen, geraden Funktion zusammenhängt, genauer

$$\tilde{f}(\theta) \;=\; f(\cos\theta) \;\sim\; \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} a_k\, \cos(k\,\theta),$$

$$a_k \;=\; \frac{1}{\pi}\int_0^{2\,\pi} \tilde{f}(\theta)\, \cos(k\theta)\, d\theta \;=\; \frac{2}{\pi}\int_0^{\pi} \tilde{f}(\theta)\, \cos(k\theta)\, d\theta$$

Die Rücktransformation auf x liefert gerade die T-Entwicklung der Funktion f.

Aussagen über Konvergenz bzw. Konvergenzgeschwindigkeit lassen sich also aus den entsprechenden Aussagen über Fourier-Reihen ableiten.

Im Vorgriff auf spätere Untersuchungen sei der Satz von Dini und Lipschitz zitiert. Er ist benannt nach Ulisse Dini (1845 - 1918) und Rudolf Lipschitz (1832 - 1903).

Satz (6.36) (Dini, Lipschitz)

Erfüllt $f \in \mathrm{C}[-1,1]$ die *Dini-Lipschitz-Bedingung*

$$\lim_{\delta\downarrow 0} \big[\omega_f(\delta)\, \ln\delta\big] \;=\; 0,$$

so gilt $\|f - R_n(f)\|_\infty \to 0 \;\; (n \to \infty)$.

Dabei bezeichnet $\omega_f(\delta) := \sup\{|f(x) - f(y)| : \; |x - y| \leq \delta\}$ den *Stetigkeitsmodul* vom f.

Bemerkung (6.37)

In Verallgemeinerung der Aussage von Satz (6.36) lässt sich in Bezug auf die Konvergenzgeschwindigkeit der Tschebyscheff-Entwicklung zeigen:

Für $f \in \mathrm{C}^m[-1,1], \quad m \geq 2$, gilt die folgende Abschätzung der Tschebyscheff-Koeffizienten

$$|a_k| \;\leq M_m/k^m, \quad k \in \mathbb{N}.$$

Insbesondere konvergieren damit die T-Entwicklungen analytischer Funktionen sehr schnell!

Satz (6.38)

Für die Operatornorm $\|R_n\|_\infty$ der Tschebyscheff Approximation

$$R_n(f)(x) \;:=\; {\sum_{k=0}^{n}}' a_k\, T_k(x), \quad a_k \;:=\; \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx, \quad f \in \mathrm{C}[-1,1],$$

gelten die folgenden Darstellungen

$$\|R_n\|_\infty \;=\; \frac{1}{\pi}\int_0^\pi |\,\frac{\sin[(n+1/2)\,\theta\,]}{\sin(\theta/2)}\,|\,d\,\theta \;=\; \frac{1}{2\,n+1} \;+\; \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{k}\,\tan\big(\frac{k\,\pi}{2\,n+1}\big).$$

(Das Summensymbol $\sum'$ bedeutet, dass des erste Summand mit Faktor 1/2 genommen wird.)

Beweisskizze:

Mit der Substitution $x = \cos t$ wird

$$\begin{aligned} R_n(f)(\cos t) \;&=\; \frac{2}{\pi}\,{\sum_{k=0}^{n}}'\big[\int_0^\pi f(\cos(\theta))\,\cos(k\,\theta)\,d\,\theta\big]\,\cos(k\,t) \\ &=\; \frac{2}{\pi}\int_0^\pi f(\cos(\theta))\,{\sum_{k=0}^{n}}'\cos(k\,t)\,\cos(k\,\theta)\,d\,\theta \end{aligned}$$

und damit (analog zu früheren Überlegungen)

$$\begin{aligned} \|R_n\|_\infty \;&=\; \max_{t\in[0,\pi]}\frac{2}{\pi}\int_0^\pi |\,{\sum_{k=0}^{n}}'\cos(k\,\theta)\,\cos(k\,t)|\,d\,\theta \\ &=\; \max_{t\in[0,\pi]}\frac{1}{\pi}\int_{-\pi}^\pi |\,{\sum_{k=0}^{n}}'\frac{1}{2}\big(\cos(k\,(t+\theta)) \;+\; \cos(k\,(t-\theta))\big)\,|\,d\,\theta \\ &\le\; \max_{t\in[0,\pi]}\frac{1}{2\,\pi}\int_{-\pi}^\pi |\,{\sum_{k=0}^{n}}'\cos(k\,(t+\theta))| \;+\; |\,{\sum_{k=0}^{n}}'\cos(k\,(t-\theta))|\,d\,\theta \\ &=\; \frac{1}{\pi}\int_{-\pi}^\pi |\,{\sum_{k=0}^{n}}'\cos(k\,\theta)|\,d\,\theta \\ &=\; \frac{2}{\pi}\int_0^\pi |\,{\sum_{k=0}^{n}}'\cos(k\,\theta)|\,d\,\theta. \end{aligned}$$

Da der letzte Term gerade mit der ersten Summe für $t = 0$ übereinstimmt, gilt in der obigen Relation durchgehend Gleichheit.

Die erste Darstellung in der Behauptung ergibt sich dann aus

$$\begin{aligned} 2 \sum_{k=0}^{n}{}' \cos(k\,\theta) &= 1 \,+\, 2\,\cos(\theta) \,+\, \ldots \,+\, 2\,\cos(n\,\theta) \\ &= \sum_{k=-n}^{n} \mathrm{e}^{i\,k\,\theta} \\ &= \begin{cases} 2n+1 & : \quad \theta \in 2\,\pi\,\mathbb{Z}, \\ \dfrac{\sin[(n+1/2)\,\theta\,]}{\sin(\theta/2)} & : \qquad \text{sonst.} \end{cases} \end{aligned}$$

Die zweite Darstellung in der Behauptung erhält man aus der expliziten Aufspaltung des Integrals zwischen den Nullstellen des Integranden $\theta_k = k\,\pi/(n+1/2),\ \ k = 0, \ldots, n$. Die technischen Details hierzu findet man im Buch von Powell. □

Die zweite Darstellung der Operatornorm $\|R_n\|_\infty$ ermöglicht nun deren explizite Berechnung. In der folgenden Tabelle sind einige Werte angegeben.

n	$\|R_n\|_\infty$	n	$\|R_n\|_\infty$
2	$0.16422e+01$	12	$0.22940e+01$
4	$0.18801e+01$	14	$0.23542e+01$
6	$0.20290e+01$	16	$0.24065e+01$
8	$0.21377e+01$	18	$0.24529e+01$
10	$0.22234e+01$	20	$0.24945e+01$

Die Werte der Operatornorm $\|R_n\|_\infty$ liegen somit in der gleichen Größenordnung wie die des Operators für die Polynom-Interpolation zu Tschebyscheff-Knoten; vgl. Abschnitt 3.

Es ist allerdings zu beachten, dass die Operatornorm die Approximationsgüte für (nur) stetige Funktionen $f \in \mathrm{C}[-1,1]$ widerspiegelt. Für glattere Funktionen f liefert die T-Approximation dagegen i.Allg. erheblich bessere Approximationen!

7. Approximation periodischer Funktionen

Fourier–Reihen.

Wir betrachten wieder den Raum aller stetigen, $2\,\pi$-periodischen Funktionen

$$\mathrm{C}_{2\,\pi} \;:=\; \{f \in \mathrm{C}(\mathbb{R}) :\; \forall\, t \in \mathbb{R} :\; f(t + 2\,\pi) = f(t)\} \tag{7.1}$$

mit dem Standard-Skalarprodukt

$$\langle f, g\rangle \;:=\; \frac{1}{\pi} \int_0^{2\,\pi} f(t)\, g(t)\, dt. \tag{7.2}$$

Die zugehörige Norm auf $\mathrm{C}_{2\,\pi}$ werde wieder mit $\|\cdot\|$ oder besser $\|\cdot\|_2$ bezeichnet.

Der folgende Satz fasst nochmals das Beipiel (6.14) zusammen

Satz (7.3)

Die Funktionen $\dfrac{1}{\sqrt{2}},\ \cos t,\ \sin t, \ldots,\ \cos(n\,t),\ \sin(n\,t)$ bilden ein Orthonormalsystem bzgl. $\langle\cdot,\cdot\rangle$. Sie sind somit zugleich eine ONB des von ihnen aufgespannten linearen Teilraumes

$$\mathrm{T}_n \;:=\; \{f \in \mathrm{C}_{2\,\pi} :\; f(t) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kt) + b_k \sin(kt)],\ a_k, b_k \in \mathbb{R}\}. \tag{7.4}$$

Die L_2–Bestapproximation einer Funktion $f \in \mathrm{C}_{2\,\pi}$ aus T_n lautet daher

$$\begin{aligned} S_n(f)(t) \;&=\; f_n(t) \;=\; \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kt) + b_k \sin(kt)] \\ a_k \;&=\; \frac{1}{\pi} \int_0^{2\,\pi} f(t)\, \cos(kt)\, dt, \qquad b_k \;=\; \frac{1}{\pi} \int_0^{2\,\pi} f(t)\, \sin(kt)\, dt. \end{aligned} \tag{7.5}$$

Für $n \to \infty$ erhält man hieraus formal die *Fourier-Entwicklung* einer Funktion $f \in \mathrm{C}_{2\,\pi}$ (Jean-Baptiste-Joseph Fourier; 1768–1830)

$$f(t) \;\sim\; \frac{a_0}{2} + \sum_{k=1}^{\infty} [a_k \cos(kt) + b_k \sin(kt)]. \tag{7.6}$$

Wir fragen nach der Konvergenz dieser *Fourier-Reihe*, nach der Konvergenzgeschwindigkeit und untersuchen, unter welchen Voraussetzungen in (7.6) punktweise bzw. gleichmäßige Konvergenz vorliegt.

Zunächst können wir analog zu Satz (6.33) die L_2-Konvergenz der Fourier-Reihe feststellen.

Satz (7.7)

Zu $f \in C_{2\pi}$ bezeichne $T_n(f)$ die (bzw. eine) Bestapproximation von f aus T_n bezüglich $\|\cdot\|_\infty$. Dann gelten

$$\begin{aligned} &\textbf{a)} && \|f - T_n(f)\|_\infty \;\to\; 0 && (n \to \infty), \\ &\textbf{b)} && \|f - T_n(f)\|_2 \;\to\; 0 && (n \to \infty), \\ &\textbf{c)} && \|f - S_n(f)\|_2 \;\to\; 0 && (n \to \infty). \end{aligned}$$

Beweis: Der Beweis erfolgt analog zu dem von Satz (6.33). Teil a) ergibt sich aus dem zweiten Weierstraßschen Approximationssatz (4.20).
Teil b) folgt aus der Abschätzung

$$\|f - T_n(f)\|_2 \;\le\; \sqrt{2}\;\|f - T_n(f)\|_\infty.$$

Teil c) folgt schließlich aus der Minimaleigenschaft von $S_n(f)$, nämlich
$\|f - S_n(f)\|_2 \le \|f - T_n(f)\|_2$. □

Erinnert sei auch an die *Parseval Gleichung* (6.13), die sich aus der Dichtheit der trigonometrischen Polynome in $C_{2\pi}$ ergibt,

$$\frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2\right) \;=\; \|f\|_2^2. \tag{7.8}$$

Aus (7.8) folgt insbesondere, dass die Fourier-Koeffizienten $a_k,\ b_k$ Nullfolgen bilden.

Nun zur Untersuchung der punktweisen bzw. gleichmäßigen Konvergenz der Fourier-Reihe. Zunächst ist klar, dass Satz (7.7) lediglich die *Konvergenz im quadratischen Mittel* zeigt, woraus nicht auf die punktweise Konvergenz geschlossen werden kann. Tatsächlich hat Paul du Bois-Reymond (1831–1889) eine stetige Funktion $f \in C_{2\pi}$ angegeben, deren Fourier-Reihe in mindestens einem Punkt divergiert.

Die Jackson–Sätze.

Ausgangspunkt für die folgenden Untersuchungen ist das Lemma von Lebesgue (3.4). Hierbei ist zu beachten, dass $S_n : C_{2\pi} \to T_n$ ein stetiger linearer Projektor ist. Die Stetigkeit ergibt sich beispielsweise aus der Integraldarstellung (vgl. (4.16))

$$S_n(f)(t) \;=\; \frac{1}{\pi} \int_0^{2\pi} f(t+\theta)\, D_n(\theta)\, d\theta, \qquad D_n(\theta) \;:=\; \frac{\sin[(n+1/2)\theta]}{2\,\sin[\theta/2]}. \tag{7.9}$$

Das Lemma von Lebesgue ist also anwendbar und lautet hier konkret

$$\|f - S_n(f)\|_\infty \;\le\; \big(1 + \|S_n\|_\infty\big)\, E_n(f), \qquad E_n(f) \;:=\; \inf_{p \in \mathrm{T}_n} \|f - p\,\|_\infty. \tag{7.10}$$

Dabei bezeichnet $E_n(f)$ den Minimalabstand von f zu T_n bezüglich der Maximumsnorm.

Wir beginnen mit einer Abschätzung für $\|S_n\|_\infty$.

Satz (7.11)

a) $$\|S_n\|_\infty \;=\; \frac{1}{\pi} \int\limits_0^\pi |\, \frac{\sin[(n+1/2)\,\theta\,]}{\sin(\theta/2)}\,|\, d\,\theta$$

b) $$\frac{4}{\pi^2}\, \ln(1+n) \;\le\; \|S_n\|_\infty \;\le\; 1 \;+\; \ln(2\,n+1).$$

Beweis:

zu a) Dies folgt direkt aus der Integraldarstellung (7.9). Man beachte, dass $D_n(\theta)$ eine gerade, $2\,\pi$-periodische Funktion ist.

zu b) Seien $\theta_k := (k\,\pi)/(n+1/2)$ die Nullstellen von $D_n(\theta)$. Damit gilt

$$\begin{aligned}
\|S_n\|_\infty \;&\ge\; \frac{1}{\pi} \sum_{k=0}^{n-1} \int\limits_{\theta_k}^{\theta_{k+1}} |\, \frac{\sin[(n+1/2)\,\theta\,]}{\theta/2}\,|\, d\,\theta \\
&\ge\; \frac{2}{\pi} \sum_{k=0}^{n-1} \frac{1}{\theta_{k+1}} \int\limits_{\theta_k}^{\theta_{k+1}} |\, \sin[(n+1/2)\,\theta\,]\,|\, d\,\theta \\
&=\; \frac{4}{\pi^2} \sum_{k=0}^{n-1} \frac{1}{k+1} \;\ge\; \frac{4}{\pi^2}\; \ln(n+1).
\end{aligned}$$

Für die rechte Seite verwenden wir die folgenden beiden Abschätzungen

$$|\, \frac{\sin[(n+1/2)\,\theta\,]}{\sin(\theta/2)}\,| \;=\; 2\,|\, {\sum_{k=0}^{n}}' \cos(k\,\theta)\,| \;\le\; 2\,n+1,$$

sowie
$$|\, \frac{\sin[(n+1/2)\,\theta\,]}{\sin(\theta/2)}\,| \;\le\; \frac{1}{\theta/\pi} \;=\; \frac{\pi}{\theta}.$$

Für irgendein $\mu \in]0,1[$ folgt somit

$$\|S_n\|_\infty \;\le\; \frac{1}{\pi}\,\Big(\int\limits_0^\mu (2\,n+1)\; d\theta \;+\; \int\limits_\mu^\pi \frac{\pi}{\theta}\; d\theta\Big) \;=\; \frac{(2\,n+1)\,\mu}{\pi} \;+\; \ln\frac{\pi}{\mu}.$$

Speziell für $\mu = \pi/(2\,n+1)$ ergibt sich die angegebene obere Schranke. □

Um nun mittels (7.10) Konvergenz zu zeigen, benötigen wir die hinreichend schnelle Konvergenz von $E_n(f) \to 0 \quad (n \to \infty)$. Hierzu dienen die verschiedenen Sätze von Jackson, benannt nach Dunham Jackson (1888–1946), einem Schüler von Edmund Landau.

Satz (7.12) (Jackson I)

Für $f \in \mathrm{C}^{(1)}_{2\pi} := \mathrm{C}_{2\pi} \cap \mathrm{C}^1(\mathbb{R})$ und $n \in \mathbb{N}_0$ gilt

$$E_n(f) \;\leq\; \frac{\pi}{2(n+1)}\; \|f'\|_\infty.$$

Beweis: Zunächst zeigt man mittels partieller Integration die Darstellung

$$f(t) \;=\; \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(\theta)\, d\theta \;+\; \frac{1}{2\,\pi} \int_{-\pi}^{\pi} \theta\; f'(\theta + t + \pi)\, d\theta. \tag{7.13}$$

Der erste Summand ist konstant. Da T_n die konstanten Funktionen enthält, genügt es, den zweiten Summanden durch trigonometrische Polynome aus T_n zu approximieren.

$$E_n(f) \;=\; \inf\{\, \|\frac{1}{2\,\pi} \int_{-\pi}^{\pi} \theta\; f'(\theta + t + \pi)\, d\theta \;-\; q(t)\|_\infty :\; q \in \mathrm{T}_n\}. \tag{7.14}$$

Nun ist für $g \in \mathrm{C}_{2\pi}$ und $p \in \mathrm{T}_n$ die folgende Funktion

$$q(t) \;:=\; \int_{-\pi}^{\pi} p(\theta)\, g(\theta + t)\, d\theta$$

auch stets wieder ein trigonometrisches Polynom in T_n. Aufgrund der Periodizität von p und g hat man nämlich

$$\begin{aligned} q(t) \qquad &= \int_{-\pi}^{\pi} p(\theta - t)\, g(\theta)\, d\theta \\ p(\theta - t) \;&=\; \frac{a_0(\theta)}{2} \;+\; \sum_{k=1}^{n} a_k(\theta)\, \cos(kt) \;+\; b_k(\theta)\, \sin(kt). \end{aligned}$$

Wir reduzieren also die Infimumsbildung in (7.14) auf die trigonometrischen Polynome q der obigen Form, wobei wir $g(t) := f'(t + \pi)$ wählen. Damit erhalten wir aus (7.14)

$$\begin{aligned} E_n(f) \;&\leq\; \inf\{\, \|\frac{1}{2\,\pi} \int_{-\pi}^{\pi} (\theta \;-\; p(\theta))\; f'(\theta + t + \pi)\, d\theta\, \|_\infty :\; p \in \mathrm{T}_n\} \\ &\leq\; \frac{1}{2\,\pi}\; \|f'\|_\infty\; \inf\{ \int_{-\pi}^{\pi} |\theta \;-\; p(\theta)|\, d\theta \;:\; p \in \mathrm{T}_n\}. \end{aligned} \tag{7.15}$$

Damit ist nun unser Problem, eine obere Schranke für $E_n(f)$ zu finden, zurückgeführt worden auf eine L_1-Approximationsaufgabe, nämlich auf das Problem, die Funktion $g(t) = t$ in L_1-Sinn auf dem Intervall $[-\pi, \pi]$ durch ein trigonometrisches Polynom $p \in T_n$ zu approximieren.

Wir verzichten darauf, die L_1-Optimalität der folgenden Konstruktion zu beweisen, und geben statt dessen nur die Lösung dieser Approximationsaufgabe an. Dazu bedarf es noch einiger Vorbemerkungen über trigonometrischer Interpolation.

(a) Zu $(2n+1)$ Knoten $t_0 < t_1 < \ldots < t_{2n} \in [-\pi, \pi[$ und Funktionswerten f_j gibt es genau ein trigonometrisches Polynom $p \in \mathrm{T}_n$ mit $p(t_j) = f_j,\ j = 0, \ldots, 2n$.
(Beweis: Transformation in ein komplexes Polynom.)

(b) Ein trigonometrisches Polynom $p \in \mathrm{T}_n,\ p \neq 0,$ kann in $[-\pi, \pi[$ höchstens $2n$ Nullstellen haben.
(Beweis: Gäbe es mehr als $2n$ Nullstellen, hätte man einen Widerspruch zu (a).)

(c) Wir setzen $t_k := \dfrac{k\pi}{n+1}$, $k = 1, \ldots, n$. Dann gibt es genau ein trigonometrisches Polynom $p \in \mathrm{T}_n$ der Form $p(t) = \sum_1^n b_k \sin(kt)$ mit $p(t_k) = t_k,\ k = 1, \ldots, n$.
(Beweis: Wende (a) an auf die Stützstellen $(\pm t_k, \pm t_k)$ und $(0, 0)$ und zeige ebenfalls mittels (a), dass alle a_k verschwinden müssen.)

(d) Die Fehlerfunktion $e(t) := t - p(t)$ hat in $]0, \pi[$ genau die Nullstellen $t_k,\ k = 1, \ldots, n$.
(Beweis: e hat nach Konstruktion die Nullstellen t_k und 0. Hätte e noch eine weitere Nullstelle in $]0, \pi[$, so wäre $e'(t) = 1 - p'(t)$ ein *gerades* trigonometrisches Polynom in T_n und hätte nach dem Satz von Rolle wenigstens $(n+1)$ Nullstellen in $]0, \pi[$, also wenigstens $(2n+2)$ Nullstellen in $[-\pi, \pi[$; Widerspruch!)

Für die obige Konstruktion werten wir nun die L_1–Norm aus

$$\|e\|_1 \;=\; \int_{-\pi}^{\pi} |\theta \,-\, p(\theta)|\, d\theta \;=\; 2 \int_{0}^{\pi} |\theta \,-\, p(\theta)|\, d\theta.$$

Dazu setzen wir $\sigma_{n+1}(t) := \mathrm{sign}[\sin((n+1)t)]$ und finden

$$\begin{aligned}
\int_0^{\pi} |t \,-\, p(t)|\, dt \;&=\; \Big| \int_0^{\pi} \sigma_{n+1}(t)\, (t \,-\, p(t))\, dt \,\Big| \\
&=\; \Big| \int_0^{\pi} \sigma_{n+1}(t)\, t\, dt \;-\; \sum_{k=1}^{n} b_k \int_0^{\pi} \sigma_{n+1}(t)\, \sin(k\,t)\, dt \,\Big| \\
&=\; \Big| \sum_{k=0}^{n} (-1)^k \int_{t_k}^{t_{k+1}} t\, dt \,\Big| \\
&=\; \frac{\pi^2}{2(n+1)}\,.
\end{aligned}$$

Zur vorletzten Gleichung zeigt man explizit, dass sämtliche Integrale $\int \sigma_{n+1}(t)\,\sin(k\,t)\,dt$ verschwinden. Die letzte Gleichheit ist ebenfalls durch explizite Berechnung der Integrale $\int t\,dt$ zu zeigen.

Insgesamt ist damit eine obere Schranke, nämlich $\|e\|_1 \le \pi^2/(n+1)$ bewiesen, wobei diese Schranke (ohne Beweis) bestmöglich ist. Setzt man diese nun in (7.15) ein, so ergibt sich schließlich die Behauptung des Satzes. □

Bemerkung (7.16)

Der in der Abschätzung (7.12) auftretende Faktor $\pi/(2n+2)$ lässt sich nicht verbessern.

Beweis: Zu $\varepsilon > 0$ lässt sich eine Funktion $f_\varepsilon \in \mathrm{C}^{(1)}_{2\pi}$ konstruieren mit den Eigenschaften

$$f_\varepsilon(k\pi/(n+1)) \;=\; (-1)^k,\; k = 0, \pm 1, \ldots, \pm(n+1), \quad \|f_\varepsilon'\|_\infty \;\le\; \frac{2(n+1)}{\pi}\,(1+\varepsilon).$$

Sei p_ε nun eine Bestapproximation von f_ε aus T_n bzgl. $\|\cdot\|_\infty$. Wäre $\|f_\varepsilon - p_\varepsilon\|_\infty < 1$, so hätte p_ε in den t_k das gleiche Vorzeichen wie f_ε und somit nach Zwischenwertsatz wenigstens $2n+2$ Nullstellen in $]-\pi, \pi[$. Widerspruch!
Damit folgt aber

$$E_n(f_\varepsilon) \;\ge\; 1 \;\ge\; \frac{\pi}{2\,(n+1)\,(1+\varepsilon)}\,\|f_\varepsilon'\|_\infty.$$

Für $\varepsilon \downarrow 0$ geht die Abschätzung gegen die des ersten Jackson-Satzes. □

Satz (7.17) (Jackson II)

Ist $f \in \mathrm{C}_{2\pi}$ Lipschitz-stetig mit einer Lipschitz-Konstanten L, so gilt für $n \in \mathbb{N}_0$

$$E_n(f) \;\le\; \frac{\pi}{2(n+1)}\;L.$$

Beweis: Für $\delta > 0$ setze man $F_\delta(t) \;:=\; \dfrac{1}{2\,\delta}\displaystyle\int_{t-\delta}^{t+\delta} f(\theta)\,d\theta.$ Dann ist $F_\delta \in \mathrm{C}^{(1)}_{2\pi}$ und es gilt $F_\delta'(t) \;=\; \dfrac{1}{2\,\delta}\,\big(f(t+\delta) - f(t-\delta)\big)$, also $\|F_\delta'\|_\infty \le L$.

Ist $p \in \mathrm{T}_n$ nun Bestapproximation von F_δ aus T_n bzgl. $\|\cdot\|_\infty$, so folgt

$$E_n(f) \;\le\; \|f - p\,\|_\infty \;\le\; \|f - F_\delta\|_\infty + \|F_\delta - p\,\|_\infty \;\le\; \|f - F_\delta\|_\infty + \frac{\pi}{2(n+1)}\,L.$$

Ferner lässt sich abschätzen

$$\|f - F_\delta\|_\infty \;=\; \max_t\,|\frac{1}{2\delta}\int_{t-\delta}^{t+\delta}(f(t) - f(\theta))d\theta| \;\le\; \max_t\,\frac{L}{2\delta}\int_{t-\delta}^{t+\delta}|t-\theta|d\theta \;=\; \frac{L\delta}{2}$$

und somit $\|f - F_\delta\|_\infty \to 0 \quad (\delta \downarrow 0)$. □

Satz (7.18) (Jackson III)

Ist $f \in \mathrm{C}_{2\pi}$, so gilt für $n \in \mathbb{N}_0$

$$E_n(f) \;\le\; \frac{3}{2}\,\omega(\frac{\pi}{n+1}).$$

Dabei ist $\omega(\delta) := \sup\{|f(x)-f(y)| : |x-y| \le \delta\}$ der Stetigkeitsmodul von f.

Beweis: Für $\delta > 0$ sei F_δ wie im Beweis zu (7.17) erklärt. Dann folgt mit $F'_\delta(t) = \big(f(t+\delta) - f(t-\delta)\big)/(2\,\delta)$ die Abschätzung

$$\|F'_\delta\|_\infty \;\le\; \frac{\omega(2\,\delta)}{2\,\delta} \;\le\; \frac{\omega(\delta)}{\delta}.$$

Die letzte Ungleichung ist eine direkte Folge aus der Definition des Stetigkeitsmoduls. Nach dem Jackson-Satz I gilt damit $E_n(F_\delta) \le \frac{\pi}{2(n+1)\delta}\,\omega(\delta)$. Weiterhin ist

$$\begin{aligned}\|f - F_\delta\|_\infty &= \max_t \frac{1}{2\,\delta}\,\Big|\int_{t-\delta}^{t+\delta} \big(f(t) - f(\theta)\big)\,d\theta\Big| \\ &\le \max_t \frac{1}{2\,\delta}\int_{t-\delta}^{t+\delta} \omega(\delta)\,d\theta \;=\; \omega(\delta).\end{aligned}$$

Insgesamt ergibt sich somit wieder

$$E_n(f) \;\le\; \|f - F_\delta\|_\infty + E_n(F_\delta) \;\le\; \big(1 + \frac{\pi}{2(n+1)\,\delta}\big)\,\omega(\delta).$$

Speziell für $\delta = \pi/(n+1)$ ergibt sich die Behauptung des Satzes. □

Bemerkung (7.19)

Da für eine stetige Funktion $f \in \mathrm{C}_{2\pi}$ der Stetigkeitsmodul gegen Null konvergiert $\omega(\pi/(n+1)) \to 0$ $(n \to \infty)$ folgt aus dem dritten Jackson-Satz auch der zweite Weierstraßsche Approximationssatz zurück: Jede stetige, 2π-periodische Funktion lässt sich bzgl. der Maximumsnorm beliebig genau durch trigonometrische Polynome approximieren.

Kombiniert man nun diese Abschätzung für die Minimalabweichung $E_n(f)$ mit der Abschätzung (7.11b) für die Operatornorm $\|S_n\|_\infty$, so ergibt sich der folgende Satz von Dini und Lipschitz.

Satz (7.20) (Dini, Lipschitz)

Erfüllt $f \in \mathrm{C}_{2\pi}$ die *Dini-Lipschitz-Bedingung*,

$$\lim_{\delta \downarrow 0} \big[\,\omega(\delta)\,\ln\delta\,\big] \;=\; 0,$$

so konvergiert die Fourier-Reihe $S_n(f)$ gleichmäßig gegen f für $n \to \infty$.

Beweis: Aufgrund des Lemmas von Lebesgue (7.10) sowie der Abschätzungen (7.11) und (7.18) ergibt sich

$$\begin{aligned} \|f - S_n(f)\|_\infty &\leq \big(1 + \|S_n\|_\infty\big)\, E_n(f) \\ &\leq \big(2 \,+\, \ln(2\,n+1)\big)\; \frac{3}{2}\, \omega(\frac{\pi}{n+1}). \end{aligned}$$

Für $n \geq 3$ stellt man fest, dass $\ln(2\,n+1) \leq \ln(2\,\pi) - \ln(\frac{\pi}{n+1})$ und $\ln(2\,\pi) < 2$ gelten.

Somit folgt

$$\|f - S_n(f)\|_\infty \;\leq\; 6\,\omega(\frac{\pi}{n+1}) \;-\; 1.5\, \ln(\frac{\pi}{n+1})\, \omega(\frac{\pi}{n+1}) \;\to\; 0 \quad (n \to \infty).$$ □

Bemerkung (7.21)

Über die Konvergenzgeschwindigkeit der Fourier-Entwicklung, wenn weitere Glattheitsvoraussetzungen an die Funkton f gestellt werden, gibt schließlich der vierte Jackson-Satz Auskunft, den wir nur kurz zitieren wollen

$$f \,\in\, \mathrm{C}^{(k)}_{2\,\pi} \,:=\, \mathrm{C}_{2\,\pi} \cap \mathrm{C}^k(\mathbb{R}) \;\Rightarrow\; E_n(f) \,\leq\, \big(\frac{\pi}{2n+2}\big)^k \;\|f^{(k)}\|_\infty.$$

Berechnung von Fourier–Koeffizienten.

Wir beginnen mit der **komplexen Darstellung** der Partialsumme

$$S_n(f) \;=\; \frac{a_0}{2} \,+\, \sum_{k=1}^{n} \left[a_k \cos(kt) + b_k \sin(kt)\right] \tag{7.22}$$

der Fourier-Entwicklung einer Funktion $f \in \mathrm{C}_{2\,\pi}$. Stellt man $\cos t$ und $\sin t$ als Real- und Imaginärteil von e^{it} dar, so ergibt sich

$$\begin{aligned} S_n(f) &= \frac{a_0}{2} \,+\, \sum_{k=1}^{n} \big[\, \frac{a_k}{2}\, (\mathrm{e}^{ikt} + \mathrm{e}^{-ikt}) \,+\, \frac{b_k}{2i}\, (\mathrm{e}^{ikt} - \mathrm{e}^{-ikt}) \,\big] \\ &= \frac{a_0}{2} \,+\, \sum_{k=1}^{n} \big[\, \frac{a_k - i\,b_k}{2}\, \mathrm{e}^{ikt} \;+\; \frac{a_k + i\,b_k}{2}\, \mathrm{e}^{-ikt} \,\big] \qquad (7.23) \\ &= \sum_{k=-n}^{n} c_k \mathrm{e}^{ikt}\,. \end{aligned}$$

Umrechnung der Koeffizienten: $(k = 1, 2, \ldots, n)$

$$\begin{aligned} c_0 &= \frac{a_0}{2}, \qquad c_k = \frac{1}{2}(a_k - i\,b_k), \qquad c_{-k} = \frac{1}{2}(a_k + i\,b_k), \\ a_0 &= 2\,c_0, \qquad a_k = c_k + c_{-k}, \qquad b_k = i\,(c_k - c_{-k}). \end{aligned} \tag{7.24}$$

Fourier–Koeffizienten: $(k \in \mathbb{N}_0,\ j \in \mathbb{Z})$

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t)\,\cos(kt)\,dt, \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(t)\,\sin(kt)\,dt, \\ c_j &= \frac{1}{2\pi} \int_0^{2\pi} f(t)\,\mathrm{e}^{-i\,j\,t}\,dt. \end{aligned} \tag{7.25}$$

Die Grundidee zur numerischen Berechnung der Fourier-Koeffizienten besteht darin, die Integrale in (7.25) durch Trapezsummen zu approximieren. Wegen der Periodizität der Integranden sind diese zur numerischen Quadratur besonders gut geeignet.

Für hinreichend große $N \in \mathbb{N}$ setzen wir

$$h := \frac{2\pi}{N}, \quad t_k := k\,h, \quad f_k := f(t_k), \quad k = 0, 1, \ldots, N. \tag{7.26}$$

Mit $f_0 = f_N$ (Periodizität) ergeben sich dann die folgenden Näherungen durch die Trapezsumme

$$\begin{aligned} a_k &\approx \frac{h}{\pi}\left\{\frac{f_0}{2} + \sum_{j=1}^{N-1} f_j\,\cos(k\,t_j) + \frac{f_N}{2}\right\} \\ &= \frac{2}{N}\sum_{j=0}^{N-1} f_j\,\cos(j\,k\,h) =: A_k, \quad k = 0, \ldots, n, \\ b_k &\approx \frac{2}{N}\sum_{j=0}^{N-1} f_j\,\sin(j\,k\,h) =: B_k, \quad k = 1, \ldots, n, \end{aligned} \tag{7.27}$$

$$S_n(f) \approx \frac{A_0}{2} + \sum_{k=1}^{n} [\,A_k\,\cos(k\,t) + B_k\,\sin(k\,t)\,] =: \widetilde{S_n}(f)(t).$$

Man beachte, dass die Approximationen $A_k \approx a_k$ und $B_k \approx b_k$ i. Allg. nur für kleine k brauchbar sind. So sind die A_k, B_k gemäß Definition periodisch im Index, $A_{k+N} = A_k$, $B_{k+N} = B_k$, während die tatsächlichen Fourier-Koeffizienten Nullfolgen bilden. Eine *Faustregel* besagt, dass man $N \geq 2\,n$ wählen sollte.

Andere Interpretationen:

a) Man kann die diskreten Fourier-Koeffizienten $A_k,\ B_k$ (eigentlich $A_{k,N},\ B_{k,N}$) auch als exakte Fourier-Koeffizienten einer *diskretisierten Approximationsaufgabe* interpretieren. Zu $N \geq 2\,n+1$ werde definiert

$$\begin{aligned} R &:= \mathrm{C}_{2\,\pi}(B), \qquad B := \left\{ \frac{2\,\pi}{N}\,k : \ k = 0, 1, \ldots, N-1 \right\}, \\ \langle f, g \rangle &:= \frac{2}{N} \sum_{k=0}^{N-1} f(k\,h)\, g(k\,h). \end{aligned} \tag{7.28}$$

Die Funktionen $\dfrac{1}{\sqrt{2}}, \cos t, \sin t, \ldots, \cos(n\,t), \sin(n\,t)$ bilden eine ONB des von ihnen aufgespannten (diskreten) Raumes $\widetilde{\mathrm{T}}_n$ bzgl. $\langle \cdot, \cdot \rangle$. Die $A_k,\ B_k$ liefern dann die Lösung der L_2–Approximationsaufgabe, ein vorgebenes $f \in R$ durch ein Element aus $\widetilde{\mathrm{T}}_n$ zu approximieren. Genauer ist die Lösung der Approximationsaufgabe gegeben durch

$$\widetilde{S_n}(f)(t) \ = \ \frac{A_0}{2} \ + \ \sum_{k=1}^{n} [\, A_k \, \cos(k\,t) \ + \ B_k \, \sin(k\,t)\,],$$

wobei die $A_k,\ B_k$ durch (7.27) gegeben sind.

b) Die $A_k,\ B_k$ lassen sich auch als Lösung einer *trigonometrischen Interpolationaufgabe* interpretieren.

Zu Interpolationsdaten $(t_k, f_k),\ k = 0, 1, \ldots, N-1,$ mit t_k aus (7.26), ist eine interpolierende trigonometrische Summe der Form

$$q(t) \ = \ \begin{cases} \dfrac{A_0}{2} \ + \ \displaystyle\sum_{k=1}^{n} [\, A_k \, \cos(k\,t) \ + \ B_k \, \sin(k\,t)\,], & \text{für } N = 2\,n+1, \\ \dfrac{A_0}{2} \ + \ \displaystyle\sum_{k=1}^{n-1} [\, A_k \, \cos(k\,t) \ + \ B_k \, \sin(k\,t)\,] \ + \ A_n \, \cos(n\,t), & \text{für } N = 2\,n. \end{cases} \tag{7.29}$$

Es lässt sich dann zeigen, dass diese Interpolationsaufgabe eine eindeutig bestimmte Lösung besitzt, welche gerade durch die $A_k,\ B_k$ gemäß (7.27) gegeben ist.

Komplexe Variante: Man sucht ein (komplexes) trigonometrisches Polynom der Form

$$p(t) \ = \ \sum_{k=0}^{N-1} C_k \, \mathrm{e}^{ikt}, \tag{7.30}$$

das die gegebenen Daten $(t_k, f_k),\ k = 0, \ldots, N-1,$ interpoliert.

Man erhält wieder eine eindeutig bestimmte Lösung, nämlich

$$C_k \ = \ \frac{1}{N} \sum_{j=0}^{N-1} f_j \, \mathrm{e}^{-i\,j\,k\,h}, \quad k = 0, 1, \ldots, N-1. \tag{7.31}$$

Auch hier lässt sich die reelle Lösung (7.27) und die komplexe Lösung (7.31) ineinander umrechnen. Man erhält analog zum kontinuierlichen Approximationsproblem, vgl. (7.22), $(k = 1, \ldots, n)$

$$\begin{aligned} C_0 &= \frac{A_0}{2}, \qquad C_k = \frac{1}{2}(A_k - i\,B_k), \qquad C_{N-k} = \frac{1}{2}(A_k + i\,A_k), \\ A_0 &= 2\,C_0, \qquad A_k = C_k + C_{N-k}, \qquad B_k = i\,(C_k - C_{N-k}). \end{aligned} \tag{7.32}$$

Fazit: Sowohl für die numerische Berechnung der Fourier-Koeffizienten wie auch für die Auswertung der trigonometrischen Partialsummen sind jeweils trigonometrische Summen der Form (7.27) bzw. (7.31) auf einem festen Gitter - oder kontinuierlich (7.22) bzw. (7.23) auszuwerten. Dabei ist es gleichgültig, ob man die reelle oder die komplexe Darstellung verwendet, da beide inneinander umgerechnet werden können.

Der Algorithmus von Goertzel und Reinsch.

Der Algorithmus berechnet zu gegebenen Daten $f_0, \ldots, f_{N-1} \in \mathbb{R}$ und $t \in \mathbb{R}$ die trigonometrischen Summen

$$A(t) := \sum_{k=0}^{N-1} f_k \cos(k\,t), \quad B(t) := \sum_{k=0}^{N-1} f_k \sin(k\,t). \tag{7.33}$$

Die Grundidee des Algorithmus ist – ganz analog zum Clenshaw Verfahren – die Verwendung einer Dreiterm-Rekursion für die Terme $c_k := \cos(k\,t)$ und $s_k := \sin(k\,t)$. Mittels trigonometrischer Umformung zeigt man

$$\begin{aligned} c_0 &= 1, \qquad c_1 = \cos t, \qquad c_{k+1} = 2\,c_1\,c_k - c_{k-1}, \ k \geq 1, \\ s_0 &= 0, \qquad s_1 = \sin t, \qquad s_{k+1} = 2\,c_1\,s_k - s_{k-1}, \ k \geq 1. \end{aligned} \tag{7.34}$$

Mit der gleichen Technik, die wir für den Algorithmus von Clenshaw angewendet haben, vgl. (6.30), erhalten wir den folgenden Algorithmus von Goertzel (1958):

Algorithmus von Goertzel (7.35)

$$U_N := 0, \quad U_{N-1} := f_{n-1},$$

$$c := \cos t, \quad F := 2\,c,$$

für $k = N-2,\ N-3,\ \ldots,\ 1$

$$U_k := f_k + F\,U_{k+1} - U_{k+2}$$

$$A(t) := f_0 + c\,U_1 - U_2, \quad B(t) := U_1 \sin t.$$

Anmerkungen: Zu Berechnung aller $A_k,\ B_k,\ t = t_k = 2\,\pi\,k/N,$ benötigt der Algorithmus $O(N^2)$ wesentliche Operationen.

Für kleine $|t|$ ist der Algorithmus numerisch instabil. Der Grund liegt darin, dass der Algorithmus mit $c = \cos t$ arbeitet. Kleine Änderungen in $c \approx 1$ entsprechen dabei großen Änderungen in t.

Stabilisierung nach Reinsch (1968)

Man ersetzt die Rekursion (7.34) durch eine solche, die mit $\sin^2(t/2)$ bzw. $\cos^2(t/2)$ anstelle von $\cos t$ arbeitet. Hintergrund sind die trigonometrischen Relationen

$$\sin^2(\frac{t}{2}) \;=\; \frac{1}{2}\,(1-\cos t), \qquad \cos^2(\frac{t}{2}) \;=\; \frac{1}{2}\,(1+\cos t).$$

Fall 1: $\cos t \;>\; 0.$

$$\begin{aligned} U_k &= f_k \;+\; 2\,c_1\,U_{k+1} \;-\; U_{k+2} \\ &= f_k \;+\; 2\,(c_1-1)\,U_{k+1} \;+\; 2\,U_{k+1} \;-\; U_{k+2} \\ \Rightarrow\ (U_k - U_{k+1}) &= f_k \;+\; 2\,(c_1-1)\,U_{k+1} \;+\; (U_{k+1} - U_{k+2}). \end{aligned}$$

Mit $D_k := U_k \;-\; U_{k+1}$ gilt somit

$$D_k \;=\; f_k \;-\; 4\,\sin^2(t/2)\,U_{k+1} \;+\; D_{k+1}$$

und $$A \;=\; f_0 \;+\; c_1\,U_1 \;-\; U_2 \;=\; U_0 \;-\; c_1\,U_1 \;=\; D_0 \;+\; 2\,\sin^2(t/2)\,U_1.$$

Fall 2: $\cos t \;\leq\; 0.$

$$\begin{aligned} U_k &= f_k \;+\; 2\,c_1\,U_{k+1} \;-\; U_{k+2} \\ &= f_k \;+\; 2\,(c_1+1)\,U_{k+1} \;-\; 2\,U_{k+1} \;-\; U_{k+2} \\ \Rightarrow\ (U_k + U_{k+1}) &= f_k \;+\; 2\,(c_1+1)\,U_{k+1} \;-\; (U_{k+1} + U_{k+2}). \end{aligned}$$

Mit $D_k := U_k \;+\; U_{k+1}$ gilt somit

$$D_k \;=\; f_k \;+\; 4\,\cos^2(t/2)\,U_{k+1} \;-\; D_{k+1}$$

und $$A \;=\; f_0 \;+\; C_1\,U_1 \;-\; U_2 \;=\; U_0 \;-\; C_1\,U_1 \;=\; D_0 \;-\; 2\,\cos^2(t/2)\,U_1.$$

Algorithmus von Goertzel und Reinsch (7.36)

$$\begin{aligned} &\text{Falls } \cos t \;>\; 0: \quad \sigma \;:=\; 1, \qquad F \;:=\; -4\,\sin^2(t/2), \\ &\text{sonst}: \qquad\qquad\quad \sigma \;:=\; -1, \quad F \;:=\; 4\,\cos^2(t/2), \\ &U_N \;:=\; 0, \quad D_{N-1} \;:=\; f_{N-1}, \end{aligned}$$

$$\text{für } k = N-2,\ N-3,\ \ldots,\ 0$$

$$U_{k+1} := D_{k+1} + \sigma\, U_{k+2},$$

$$D_k := f_k + F\, U_{k+1} + \sigma\, D_{k+1},$$

$$A(t) := D_0 - 0.5\, F\, U_1, \quad B(t) := U_1 \sin t.$$

In dieser Form ist der Algorithmus numerisch stabil und nur wenig aufwendiger als die ursprüngliche Version von Goertzel.

Die schnelle Fourier-Transformation (FFT).

Der Algorithmus von Goertzel und Reinsch wertet die trigonometrischen Summen an beliebigen Stellen $t \in \mathbb{R}$ aus. Er ist jedoch mit $O(N^2)$ Operationen immer noch recht aufwendig.

Ist man jedoch nur an den diskreten Fourier-Koeffizienten A_k, B_k interessiert, oder möchte man das trigonometrische Polynom nur auf dem Gitter $k\,h$, $h = 2\,\pi/N$ auswerten, so lassen sich effizientere Algorithmen angeben, die mit $O(N\ \log_2 N)$ Operationen auskommen. Solche Verfahren sind unter dem Namen FFT (fast Fourier transform) bekannt. Sie sind durch eine Arbeit von J.W. Cooley und J.W. Tuckey (Math. Comput. 19, 1965) populär geworden und bilden auch heute ein aktuelles Forschungsgebiet.

Wir gehen im Folgenden von der *komplexen Darstellung* aus und nehmen an, dass N eine Zweierpotenz ist. Der reelle Fall wird mittels der Rücktransformation (7.32) erledigt.

Gegeben: $N = 2^{r+1}, \quad r \in \mathbb{N}_0$

$$h = 2\,\pi/N, \quad t_k = k\,h, \quad f_k = f(t_k),\ k = 0, \ldots, N-1,$$

Gesucht: $$C_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j\, \mathrm{e}^{-i\,j\,k\,h}, \quad k = 0, \ldots, N-1.$$

Die Grundidee besteht darin, die Summe in zwei Summanden aufzuteilen, die selber als Fourier-Koeffizienten zu einem gröberen Gitter mit doppelter Schrittweite interpretiert werden können.

Setzen wir $m := N/2$ und sortieren die Summe C_k nach geraden und ungeraden Indizes, so ergibt sich für $k = 0, \ldots, N-1$

$$
\begin{aligned}
C_k &= \frac{1}{N}\left[\sum_{j=0}^{m-1} f_{2j}\,\mathrm{e}^{-i\,(2j)\,k\,h} \;+\; \sum_{j=0}^{m-1} f_{2j+1}\,\mathrm{e}^{-i\,(2j+1)\,k\,h}\right] \\
&= \left[\frac{1}{N}\sum_{j=0}^{m-1} f_{2j}\,\mathrm{e}^{-i\,j\,k\,(2h)}\right] \;+\; \mathrm{e}^{-i\,k\,\pi/m}\;\left[\frac{1}{N}\sum_{j=0}^{m-1} f_{2j+1}\,\mathrm{e}^{-i\,j\,k\,(2h)}\right] \\
&= G_k \;+\; \mathrm{e}^{-i\,k\,\pi/m}\;U_k, \qquad \text{wobei definiert wird} \\
G_k &:= \frac{1}{N}\sum_{j=0}^{m-1} f_{2j}\,\mathrm{e}^{-i\,j\,k\,(2\pi/m)}, \qquad U_k := \frac{1}{N}\sum_{j=0}^{m-1} f_{2j+1}\,\mathrm{e}^{-i\,j\,k\,(2\pi/m)}.
\end{aligned}
$$

Nun sind die G_k und U_k periodisch im Index:

$$G_{k+m} = G_k, \quad U_{k+m} = U_k, \quad k = 0,\,1,\,\ldots\,,\,m-1,$$

so dass sich der Rechenaufwand zur Berechnung der G_k und U_k gegenüber der Berechnung der C_k im Wesentlichen halbiert. Wir fassen den Reduktionsschritt noch einmal zusammen.

Reduktionsschritt (7.37) Mit $m := N/2$ berechne man

$$
\begin{aligned}
&\text{für} \quad k = 0,\,1,\ldots,\,m-1 \\
&\qquad G_k := \frac{1}{N}\sum_{j=0}^{m-1} f_{2j}\,\mathrm{e}^{-i\,j\,k\,(2\pi/m)}, \\
&\qquad U_k := \frac{1}{N}\sum_{j=0}^{m-1} f_{2j+1}\,\mathrm{e}^{-i\,j\,k\,(2\pi/m)}, \\
&\qquad C_k := G_k \;+\; \mathrm{e}^{-i\,k\,\pi/m}\;U_k, \\
&\qquad C_{k+m} := G_k \;-\; \mathrm{e}^{-i\,k\,\pi/m}\;U_k. \\
&\text{end } k.
\end{aligned}
$$

Dieser Reduktionsschritt wird nun iteriert bis nur noch triviale Fourier-Transformationen (also $m = 1$) auszuführen sind. Die einzelnen Fourier-Transformationen für $m = 1,\,2,\,4,\ldots,2^r$ werden aus diesem gemäß (7.37) zusammengesetzt.

Der Algorithmus besteht aus zwei Phasen: Zunächst werden die Daten so umsortiert, wie sie nach Durchführung aller Reduktionsschritte auftreten. In der zweiten Phase werden diese Daten dann gemäß (7.37) wieder zusammengesetzt.

Phase 1: Sortieren der Daten:

Wir betrachten zunächst den Sortierschritt für $N = 8$:

Ausgangsdaten:	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Schritt 1:	f_0	f_2	f_4	$f_6\|$	f_1	f_3	f_5	f_7
Schritt 2:	f_0	$f_4\|$	f_2	$f_6\|$	f_1	$f_5\|$	f_3	f_7
Schritt 3:	$f_0\|$	$f_4\|$	$f_2\|$	$f_6\|$	$f_1\|$	$f_5\|$	$f_3\|$	f_7

Betrachtet man die Dualdarstellung der Indizes bei den obigen Sortierschritten, so sieht man für den ersten Schritt

$$\begin{array}{cccc} (2j) & |*\ \ldots\ *|0| & (2j+1) & |*\ \ldots\ *|1| \\ \downarrow & \searrow\ \searrow & \downarrow & \searrow\ \searrow \\ (j) & |0|*\ \ldots\ *| & (m+j) & |1|*\ \ldots\ *| \end{array}$$

Für sämtliche r Sortierschritte erhält man die Index-Zuordnung

$$j \;=\; (j_r\, j_{r-1}\, \ldots\, j_1\, j_0)_2 \quad\mapsto\quad \overline{j} \;=\; (j_0\, j_1\, \ldots\, j_{r-1}\, j_r)_2,$$

d.h. der Index des f-Wertes in Anfangsposition j ergibt sich durch *Bitumkehr* der Ziffernfolge in der Dualdarstellung des Index j.

Für das Beispiel $N = 8$ ergibt sich

j	$(j)_2$	$(\overline{j})_2$	$\overline{j}$
0	(0 0 0)	(0 0 0)	0
1	(0 0 1)	(1 0 0)	4
2	(0 1 0)	(0 1 0)	2
3	(0 1 1)	(1 1 0)	6
4	(1 0 0)	(0 0 1)	1
5	(1 0 1)	(1 0 1)	5
6	(1 1 0)	(0 1 1)	3
7	(1 1 1)	(1 1 1)	7

Algorithmus FFT – 1.Teil (7.38)

$C_0 \;=\; f_0/N; \quad \overline{j} \;=\; 0;$
for $\quad j \;=\; 1, 2, \ldots, N-1$
  $\ell \;=\; N/2;$
  while $\quad \ell + \overline{j} \geq N,$
    $\ell \;=\; \ell/2;$
  end
  $\overline{j} \;=\; \overline{j} \;+\; 3\,\ell \;-\; N;$
  $C_{\overline{j}} \;:=\; f_j/N;$
end j

Erläuterung: In Schritt j wird zum Index j der neues Index $\overline{j}$ berechnet. Dabei wird in der while–Schleife der alte Index $\overline{j}$ auf führende Einsen getestet

$$\begin{aligned} \overline{j} &= (1,\ldots,1,0,j_{\ell+2},\ldots,j_r)_2 \\ \ell &= (0,\ldots,0,1,0,\ \ldots\ ,0)_2 \\ \overline{j}_{\text{neu}} &= (0,\ldots,0,1,j_{\ell+2},\ldots,j_r)_2 \\ \overline{j}+3\,\ell-n &= (0,\ldots,0,1,j_{\ell+2},\ldots,j_r)_2 \end{aligned}$$

Phase 2: Nachdem die f_j–Werte nun in die richtige Reihenfolge gebracht worden sind, erfolgen die Reduktionsschritte (7.37)

Algorithmus FFT – 2.Teil (7.39)

for $\ell = 0,1,\ldots,r$
  $m = 2^\ell$; $\tilde{m} = 2\cdot m$;
  for $k = 0,\ 1,\ldots,m-1$
    $c = \exp(-i\,k\,\pi/m)$;
    for $j = 0 : \tilde{m} : N-\tilde{m}$
      $G = C_{j+k}$;
      $U = c\cdot C_{j+k+m}$;
      $C_{j+k} = G + U$;
      $C_{j+k+m} = G - U$;
    end j
  end k
end ℓ

8. Tschebyscheff-Approximation: Theorie

Im Folgenden untersuchen wir Bestapproximationen bezüglich der Maximumsnorm. Die Wurzeln dieser Theorie gehen auf Pafnuti Lwowitsch Tschebyscheff (1821–1894) zurück. Tschebyscheff untersuchte polynomiale und rationale Bestapproximationen für stetige Funktionen $f \in \mathrm{C}[a,b]$ bezüglich $\|\cdot\|_\infty$. Die zentrale Aussagen dieser Theorie, der so genannte Alternantensatz (auch Satz von Tschebyscheff genannt) wurde jedoch von Blichfeld (1901) und Kirchfelder (1902) bewiesen. Tschebyscheff zeigte die folgende schwächere Aussage:

Ist $p \in \Pi_n$ eine Bestapproximation zu $f \in \mathrm{C}^1[a,b]$, so gibt es wenigstens $(n+2)$ kritische Punkte, das sind Randpunkte des Intervalls oder stationäre Punkte der Fehlerfunktion $e := f - p$.

Haarsche Räume.

Es sei $B \subset \mathbb{R}$ eine nichtleere und kompakte Menge und $R := \mathrm{C}(B)$ ausgestattet mit der Maximumsnorm $\|f\|_\infty := \max_{x\in B} |f(x)|$.

Natürlich lässt sich auch eine komplexe Variante der obigen Voraussetzungen formulieren, also $B \subset \mathbb{C}$, und die meisten der folgenden Aussagen bleiben unverändert oder mit geringen Modifikationen gültig. Anders sieht es jedoch beim mehrdimensionalen Fall, also $B \subset \mathbb{R}^m$, bzw. $B \subset \mathbb{C}^m$ mit $m > 1$ aus, wie wir sogleich sehen werden.

Gegeben sei wieder ein endlich dimensionaler, linearer Teilraum, $V \subset R$. Typische Beispiele sind mit $B = [a,b]$ die Polynomräume, $V := \Pi_n[a,b]$ mit $\dim V = n+1$, die trigonometrischen Polynome $V = T_n[a,b]$ mit $\dim V = 2n+1$, oder die Spline-Räume $V = S_m(t_0,\ldots,t_n)$ mit $\dim V = m+n$.

Definition und Satz (8.1)

Sei $V = V_n$ ein Teilraum von R der Dimension $\dim V = n+1$. Dann sind die folgenden drei Eigenschaften von V äquivalent:

(H1) Jedes Element $p \in V$, $p \neq 0$, hat höchstens n Nullstellen.

(H2) Zu $(n+1)$ Stützstellen $(t_j, f_j) \in B \times \mathbb{R}$, $j = 0,1,\ldots,n$, mit paarweise verschiedenen t_j gibt es genau eine interpolierende Funktion $p \in V$.

(H3) Ist $(h_0,\ldots,h_n)$ irgendeine Basis von V und sind $t_0,\ldots,t_n \in B$ paarweise verschieden, so ist die Matrix

$$D(t_0,\ldots,t_n) \;:=\; \begin{pmatrix} h_0(t_0) & \ldots & h_n(t_0) \\ \vdots & & \vdots \\ h_0(t_n) & \ldots & h_n(t_n) \end{pmatrix}$$

regulär in $\mathbb{R}^{(n+1,n+1)}$.

Erfüllt V diese Eigenschaften, so heißt V ein *Haarscher Raum*, benannt nach Alfred Haar (1885 – 1933). Eine beliebige Basis eines Haarschen Raumes heißt auch ein *Haarsches System* oder ein *Tschebyscheff-System* in $\mathrm{C}(B)$.

Beweis: Die Äquivalenz von (H2) und (H3) ist aus der Numerik wohlbekannt: Die eindeutige Existenz einer interpolierenden Funktion $p = \sum \alpha_j h_j \in V$ ist äquivalent zur eindeutigen Lösbarkeit des linearen Gleichungssystems

$$\begin{pmatrix} h_0(t_0) & \ldots & h_n(t_0) \\ \vdots & & \vdots \\ h_0(t_n) & \ldots & h_n(t_n) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

und damit zur Regularität der Koeffizientenmatrix.

Ist (H3) nicht erfüllt, so gibt es eine Basis $(h_0,\ldots,h_n)$ von V und paarweise verschiedene Punkte $t_0,\ldots,t_n \in B$, so dass das obige lineare Gleichungssystem für $f_j = 0$ eine Lösung $\alpha \neq 0$ besitzt. $p := \sum \alpha_j h_j$ ist damit eine Element aus $V \setminus \{0\}$ mit wenigstens $(n+1)$ Nullstellen. Damit ist auch (H1) nicht erfüllt. Umgekehrt: Gilt (H3), so hat das homogene lineare Gleichungssystem nur die triviale Lösung $\alpha \neq 0$. Jedes $p \in V$, $p \neq 0$, hat daher höchstens n Nullstellen. □

Beispiel (8.2)

Die Monome $(1, t, \ldots, t^n)$ bilden ein Haarsches System in $\mathrm{C}[a,b]$, $a < b$.

Die Aussage (H1) hängt dabei mit dem Fundamentalsatz der Algebra zusammen, (H2) entspricht der Existenz und Eindeutigkeit des Problems der Interpolation durch Polynome, (H3) entspricht schließlich der Regularität der Vandermonde-Matrix.

Das Beispiel lässt sich unmittelbar auf $\mathbb{C}$ übertragen: $(1, z, \ldots, z^n)$ ist ein Haarsches System in $\mathrm{C}(B;\mathbb{C})$, wobei $B \subset \mathbb{C}$ wenigstens $(n+1)$ Punkte enthalten muss.

Beispiel (8.3)

Die trigonometrischen Funktionen $(1, \cos t, \sin t, \ldots, \cos(nt), \sin(nt))$ bilden ein Haarsches System in $\mathrm{C}[0, 2\,\pi[$.

Beweis: Man sieht dies anhand der komplexen Darstellung

$$p(t) \;=\; \frac{a_0}{2} \;+\; \sum_{k=1}^{n} [a_k \cos(kt) + b_k \sin(kt)] \;=\; \sum_{k=-n}^{n} \gamma_k \mathrm{e}^{ikt} \;=\; \mathrm{e}^{-int} \sum_{k=0}^{2n} \gamma_{k-n} z^k$$

mit $\gamma_0 = a_0/2$, $\gamma_k = (a_k - i b_k)/2$, $\gamma_{-k} = (a_k + i b_k)/2$ $(k > 0)$ und $z = \mathrm{e}^{it}$.

Zu $(2n+1)$ verschiedenen Punkten in $[0, 2\,\pi[$ sind auch die $z_j := \mathrm{e}^{it_j}$ paarweise verschieden. Daher gibt es zu den Knoten $(z_j, \mathrm{e}^{int_j}\, f_j)$, $j = 0, \ldots, 2n$, ein eindeutig bestimmtes Interpolationspolynom $q(z) = \sum_{k=0}^{2n} \gamma_{k-n} z^k$. Wegen

$$f_j \;=\; \mathrm{e}^{-int_j} \sum_{k=0}^{2n} \gamma_{k-n}\, z_j^k \;=\; \sum_{k=0}^{2n} \gamma_{k-n}\, \mathrm{e}^{i(k-n)t_j} \;=\; \sum_{k=0}^{2n} \overline{\gamma_{k-n}}\, \mathrm{e}^{i(n-k)t_j} \;=\; \mathrm{e}^{-int_j} \sum_{\ell=0}^{2n} \overline{\gamma_{n-\ell}}\, z_j^\ell$$

gilt $\gamma_{k-n} = \overline{\gamma_{n-k}}$, $k = 0, \ldots, 2n$. Damit ist das zugehörige interpolierende trigonometrische Polynom $p(t)$ – mit *reellen* Koeffizienten a_k, b_k – ebenfalls eindeutig bestimmt. □

Durch gerade bzw. ungerade Fortsetzung einer Funktion aus $\mathrm{C}([0, \pi[)$ bzw. $\mathrm{C}(]0, \pi[)$ ergibt sich:

$(1, \cos t, \ldots, \cos(nt))$ ist ein Haarsches System in $\mathrm{C}([0, \pi[)$, $(\sin t, \ldots, \sin(nt))$ ist ein Haarsches System in $\mathrm{C}(]0, \pi[)$.

Beispiel (8.4)

Für $\lambda_0 < \lambda_1 < \ldots < \lambda_n$ bilden die Funktionen $(\mathrm{e}^{\lambda_0 t}, \ldots, \mathrm{e}^{\lambda_n t})$ ein Haarsches System in $\mathrm{C}[a, b]$.

Beweis: (per Induktion über n)
Für $n = 0$ ist die Gültigkeit von (H1) klar. Zum Induktionsschritt: Hat $p(t) = \sum_{k=0}^{n} a_k \mathrm{e}^{\lambda_k t}$ $(n+1)$ Nullstellen in $[a, b]$, so hat $q(t) := (\mathrm{e}^{-\lambda_0 t}\, p(t))'$ nach dem Satz von Rolle dort wenigstens n Nullstellen. Da aber $q \in \mathrm{Spann}(\mathrm{e}^{(\lambda_1 - \lambda_0)t}, \ldots, \mathrm{e}^{(\lambda_n - \lambda_0)t})$ folgt aus der Induktionsvoraussetzung, dass $q = 0$ ist. Damit ergibt sich aber auch $p = 0$. □

Eine Variante des obigen Beweises zeigt weiter: Für $\lambda_0 < \lambda_1 < \ldots < \lambda_n$ und $m_i \in \mathbb{N}_0$, $i = 0, \ldots, n$, ist $(\mathrm{e}^{\lambda_0 t}, t\,\mathrm{e}^{\lambda_0 t}, \ldots, t^{m_0}\,\mathrm{e}^{\lambda_0 t}, \ldots, t^{m_n}\,\mathrm{e}^{\lambda_n t})$ ein Haarsches System in $\mathrm{C}[a, b]$, $a < b$.

Ferner folgt mittels Substitution $\mathrm{e}^t \to x$:

Für $\lambda_0 < \lambda_1 < \ldots < \lambda_n$ ist $(x^{\lambda_0}, \ldots, x^{\lambda_n})$ ein Haarsches System in $\mathrm{C}[a, b]$, $0 < a < b$.

Beispiel (8.5) (Mairhuber, Proc. AMS 7, 1956)

Die folgende Konstruktion von Mairhuber zeigt, dass im (reellen) mehrdimensionalen Fall, $\mathrm{C}(B)$, $B \subset \mathbb{R}^m$, $m \geq 2$, i. Allg. kein Haarsches System existiert.

Es sei $m = 2$ und die Menge $B \subset \mathbb{R}^2$ enthalte eine y-förmige Teilmenge. Legt man die Punkte $z_0, \ldots, z_n$ wie in Abb. 8.1 und nimmt an, dass $h_0, \ldots, h_n$ ein Haarsches System ist, so darf $d(z_0, \ldots, z_n) := \det D(z_0, \ldots, z_n)$ nicht verschwinden. d hängt dabei stetig von den $z_0, \ldots, z_n$ ab.

Wir führen nun folgende Punktverschiebung durch: z_0 wandere über Position a zur Position b, anschließend wandere z_1 über die Position a zur alten Position von z_0 und sodann z_0 von Position b über a zur alten Position von z_1.

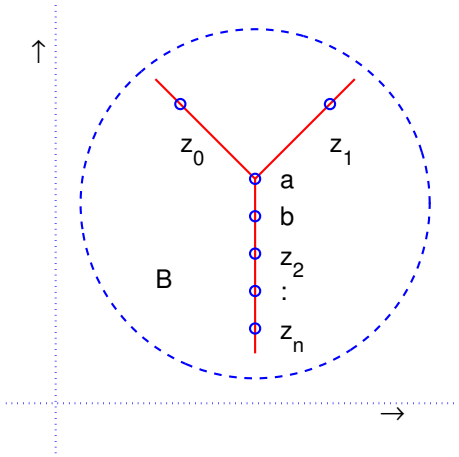


Abb. 8.1: Beispiel von Mairhuber

Bei diesem ganzen Prozeß bleiben die Punkte jeweils paarweise verschieden, so dass $d(z_0, \ldots, z_n)$ sein Vorzeichen nicht wechselt. Andererseits entsteht die Endposition durch Vertauschung der ersten beiden Zeilen in der Matrix $D(z_0, \ldots, z_n)$. Damit folgt aber $d(z_0, z_1, \ldots, z_n) = -d(z_1, z_0, \ldots, z_n)$ und wir erhalten einen Widerspruch! □

Für den eindimensionalen, reellen Fall mit $B = [a, b]$ sind die folgenden weiteren Eigenschaften Haarscher Räume hilfreich.

Satz (8.6)

Sei V ein $(n+1)$ dimensionaler Haarscher Teilraum von $\mathrm{C}[a, b]$. Dann gelten

(H4) Hat $p \in V \setminus \{0\}$ im Intervall $[a, b]$ m Nullstellen, von denen k Nullstellen ohne Vorzeichenwechsel sind (diese liegen dann im offenen Intervall $]a, b[$), so gilt $m + k \leq n$.

(H5) Zu $k \leq m \in \mathbb{N}_0$, $m + k = n$, und paarweise verschiedenen Punkten $t_1, \ldots, t_k \in]a, b[$, $t_{k+1}, \ldots, t_m \in [a, b]$ existiert eine Funktion $p \in V \setminus \{0\}$, die genau die Nullstellen $t_1, \ldots, t_m$ besitzt und für die $t_1, \ldots, t_k$ Nullstellen ohne Vorzeichenwechsel sind.

(H6) Zu $m \leq n$ vorgegebenen paarweise verschiedenen Punkten $t_1, \ldots, t_m \in]a, b[$, existiert eine Funktion $p \in V \setminus \{0\}$, die genau die Nullstellen $t_1, \ldots, t_m$ besitzt und diese sämtlich Nullstellen mit Vorzeichenwechsel sind.

Beweis: Für den Beweis dieser Aussagen sei auf das Lehrbuch von Powell (Anhang) verwiesen. □

Das Kolmogoroff-Kriterium.

Es sei wieder $B \subset \mathbb{R}$ eine kompakte Menge, $R = \mathrm{C}(B)$ ausgestattet mit der Maximumsnorm $\|\cdot\|_\infty$. Weiter sei $f \in \mathrm{C}(B)$ und V ein $(n+1)$-dimensionaler linearer Teilraum von $\mathrm{C}(B)$.

Definition (8.7)

Zu $p \in V$ heißt $e := f - p$ die *Fehlerfunktion* der Approximation p von f. Die Menge

$$A \;=\; A(f,p) \;:=\; \{t \in B :\; |e(t)| \;=\; \|e\|_\infty\}$$

heißt *Menge der Extremalpunkte* von p bezüglich f.

Es gibt nun die folgende Charakterisierung einer Bestapproximation durch ihre Extremalpunkte (nach Andrey Nikolaevich Kolmogoroff; 1903 – 1987)

Satz (8.8) (Kolmogoroff, 1948)

$p \in V$ ist genau dann eine Bestapproximation von $f \in \mathrm{C}(B)$ bezüglich V, $\|\cdot\|_\infty$, wenn das folgende Kriterium gilt

$$\forall\, q \in V :\quad \min\{(f(t) - p(t))\, q(t) :\; t \in A(f,p)\} \;\leq\; 0.$$

Beweis:

⇒: Angenommen, das Kolmogoroff-Kriterium gilt nicht, d.h. es gibt ein $q \in V$ und ein $\varepsilon > 0$, so dass

$$\forall\, t \in A(f,p) :\quad (f(t) - p(t))\, q(t) \;>\; 2\,\varepsilon.$$

Da $A(f,p)$ kompakt ist, gibt es eine in B offene Menge $U \supset A(f,p)$ mit

$$\forall\, t \in U :\quad (f(t) - p(t))\, q(t) \;>\; \varepsilon.$$

Mit $M := \|q\|_\infty$, $p_1 := p + \lambda q$, $\lambda > 0$ folgt dann für $t \in U$:

$$\begin{aligned}
(f(t) - p_1(t))^2 \;&=\; ((f(t) - p(t)) - \lambda q(t))^2 \\
&=\; (f(t) - p(t))^2 \;-\; 2\,\lambda\,(f(t) - p(t))\, q(t) \;+\; \lambda^2 (q(t))^2 \\
&<\; \|e\|_\infty^2 \;-\; 2\,\lambda\,\varepsilon \;+\; \lambda^2\, M^2 \\
&<\; \|e\|_\infty^2 \;-\; \lambda\,\varepsilon, \qquad \text{für} \quad 0 < \lambda < \varepsilon / M^2.
\end{aligned}$$

Da $B \setminus U$ kompakt ist und $A(f,p) \subset U$ existiert ein $\delta > 0$ mit $\forall\, t \in B \setminus U :$ $|f(t) - p(t)| < \|e\|_\infty - \delta$.

Für $\lambda < \delta/(2M)$ und $t \in B \setminus U$ folgt:

$$\begin{aligned}
|f(t) - p_1(t)| \;&\leq\; |f(t) - p(t)| \;+\; \lambda\,|q(t)| \\
&\leq\; \|e\|_\infty \;-\; \delta \;+\; \delta/(2M)\, M \;=\; \|e\|_\infty \;-\; \delta/2.
\end{aligned}$$

Insgesamt ist damit gezeigt: $\|f - p_1\|_\infty \;<\; \|f - p\|_\infty$, im Widerspruch zur Minimaleigenschaft von p.

$\Leftarrow$: Gelte das Kolmogoroff-Kriterium und sei $p_1 \in V$, $q := p_1 - p$.
Dann existiert ein $t_0 \in A(f,p)$ mit $(f(t_0) - p(t_0))\, q(t_0) \leq 0$. Damit folgt

$$\begin{aligned}(f(t_0) - p_1(t_0))^2 &= ((f(t_0) - p(t_0)) - q(t_0))^2 \\ &= (f(t_0) - p(t_0))^2 \; - \; 2\,(f(t_0) - p(t_0))\, q(t_0) \; + \; q(t_0)^2 \\ &\geq (f(t_0) - p(t_0))^2 \;=\; \|f - p\|^2.\end{aligned}$$

Es folgt $\|f - p_1\|_\infty \geq \|f - p\|_\infty, \; \forall\, p_1 \in V$. Damit ist p Bestapproximation von f aus V. $\square$

Bemerkung (8.9)

Für den komplexen Fall $B \subset \mathbb{C}$ (kompakt) lautet das Kolmogoroff-Kriterium

$$\forall\, q \in V: \quad \min\{\, \mathrm{Re}\,[(f(t) - p(t))\, \overline{q(t)}\,] : \; t \in A(f,p)\,\} \;\leq\; 0.$$

Der Beweis erfolgt analog zu dem von Satz (8.8).

Beispiel (8.10)

Sei $R := \mathrm{C}[1,2]$, $V := \Pi_1[1,2]$ und $f(t) = t^2$. Zunächst bestimmen wir analog zu den Beispielen 1.3b) und 1.6c) einen Kandidaten $p(t) = p^*(t) = a_0 + a_1\, t$ aus dem nichtlinearen Gleichungssystem

$$\begin{aligned} f(1) \; - \; p(1) &= \;\; \delta, \\ f(\tau) \; - \; p(\tau) &= -\delta, \\ f(2) \; - \; p(2) &= \;\; \delta, \\ f'(\tau) \; - \; p'(\tau) &= \;\; 0. \end{aligned}$$

Eine einfache Rechnung liefert die (eindeutige) Lösung $a_0 = -17/8$, $a_1 = 3$ und $\tau = 1.5$. Damit ist

$$\|f - p^*\|_\infty \;=\; \delta \;=\; 1/8, \quad A(f,p^*) \;=\; \{1, 1.5, 2\}.$$

Ist p^* nun tatsächlich eine Bestapproximation? Das Kolmogoroff-Kriterium besagt, dass für alle $q \in \Pi_1$ gelten muss:

$$\min\{\delta\, q(1),\; -\delta\, q(3/2),\; \delta\, q(2)\} \;\leq\; 0.$$

Wäre dies nicht der Fall, so müsste $q(1) > 0$, $q(3/2) < 0$ und $q(2) > 0$ sein, im Widerspruch zu $q \in \Pi_1$. Damit ist gezeigt, dass das Kolmogoroff-Kriterium erfüllt ist; p^* ist also tatsächlich Bestapproximation von f aus V!

Eine Folgerung aus dem Kolmogoroff-Kriterium ist der folgende Eindeutigkeitssatz für $\|\cdot\|_\infty$–Bestapproximationen.

Satz (8.11) (Haarscher Eindeutigkeitssatz)

Sei wieder $B \subset \mathbb{R}$ kompakt, $R := \mathrm{C}(B)$ ausgestattet mit $\|\cdot\|_\infty$. Ferner sei $V \subset R$ ein $(n+1)$–dimensionaler Haarscher Teilraum von R. Dann gelten

a) Ist $p \in V$ Bestapproximation von $f \in R \setminus V$ bezüglich V, so enthält $A(f,p)$ wenigstens $(n+2)$ Punkte.

b) Zu jedem $f \in R$ gibt es genau eine Bestappoximation aus V.

Beweis: zu a) Nehmen wir an, es gäbe $(n+1)$ paarweis verschiedene Punkte $t_0, \ldots, t_n \in B$ mit $A(f,p) \subset \{t_0, \ldots, t_n\}$. Nach (H2) existiert dann $q \in V$ mit $q(t_j) = e(t_j) := f(t_j) - p(t_j)$. Für alle $t_j \in A(f,p)$ gilt somit

$$(f(t_j) - p(t_j))\, q(t_j) \;=\; e(t_j)^2 \;=\; \|f - p\,\|_\infty^2 \;>\; 0.$$

Das Kolmogoroff-Kriterium ist damit *nicht* erfüllt und somit p keine Bestapproximation. Widerspruch!

zu b) Seien p_1, p_2 Bestapproximationen an f aus V und gelte o.B.d.A. $f \in \mathrm{C}(B) \setminus V$. Dann ist auch $p = (p_1 + p_2)/2$ eine Bestapproximation.

Nach Teil a) existieren wenigstens $(n+2)$ Extremalpunkte $t_0, \ldots, t_{n+1} \in A(f,p)$. Es gilt also

$$f(t_j) - p(t_j) \;=\; \sigma_j\, d_V(f), \quad j = 0, \ldots, n+1; \quad |\sigma_j| \;=\; 1,$$

$$\Rightarrow\; |\frac{1}{2}\,(f(t_j) - p_1(t_j)) \;+\; \frac{1}{2}\,(f(t_j) - p_2(t_j))| \;=\; d_V(f).$$

Da aber zugleich $|f(t_j) - p_k(t_j)| \le d_V(f)$ gilt, folgt hiermit

$$f(t_j) - p_1(t_j) \;=\; f(t_j) - p_2(t_j) \;=\; \sigma_j\, d_V(f), \qquad j = 0, \ldots, n+1.$$

Damit ist aber $(p_2 - p_1)(t_j) = 0$, $j = 0, \ldots, n+1$, und damit nach (H1): $p_2 = p_1$. □

Bemerkungen (8.12)

a) Der Beweis des Haarschen Eindeutigkeitssatzes gilt analog im komplexen Fall $(B \subset \mathbb{C})$. Zum Beweisteil b) beachte man, dass $(\mathbb{C}, |\cdot|)$ strikt normiert ist.

b) Es gilt in gewissem Sinn die Umkehrung des Haarschen Eindeutigkeitssatzes: Ist V ein endlich dimensionaler Teilraum und erfüllt V nicht die Haarsche Bedingung, so existiert ein $f \in \mathrm{C}(B)$ zu dem es mehrere Bestapproximationen gilt; vgl. Schönhage, Satz 6.4.

Die Aussagen (8.12)b) und (8.11) lassen sich wie folgt zusammenfassen

Folgerung (8.13)

Für einen linearen Teilraum $V \subset \mathrm{C}(B)$ mit $\dim V = n+1$ sind äquivalent

a) Jedes $p \in V \setminus \{0\}$ hat höchstens n Nullstellen,

b) Zu jedem $f \in \mathrm{C}(B)$ gibt es genau eine Bestapproximation von f aus V.

Neben der Frage der Eindeutigkeit einer Bestapproximation ist auch die Frage nach

der *strikten* Eindeutigkeit von Interesse, gemeint ist damit ein mindestens lineares Anwachsen des Fehlers mit dem Abstand von der Bestapproximation.

Definition (8.14)

$p \in V$ heißt eine *strikt eindeutige* Bestapproximation von f bzgl. V, falls es ein $\gamma > 0$ gibt mit

$$\forall\, q \in V: \ \|f - q\|_\infty \ \geq \ \|f - p\|_\infty \ + \ \gamma \, \|p - q\|_\infty .$$

Im reellen Fall lässt sich tatsächlich aus der Haarschen Bedingung die Existenz einer strikt eindeutigen Bestapproximation folgern, vgl. Nürnberger; Theorem 3.18.

Satz (8.15)

Für einen endlich dimensionalen linearen Teilraum $V \subset \mathrm{C}[a,b]$ sind äquivalent

a) V ist Haarscher Teilraum,

b) Zu jedem $f \in \mathrm{C}[a,b]$ gibt es eine strikt eindeutige Bestapproximation von f aus V.

Alternanten.

Im Folgenden sei $B \subset [a,b]$ kompakt, $R = \mathrm{C}(B)$, V sei ein $(n+1)$–dimensionaler Teilraum von $\mathrm{C}[a,b]$. Ferner enthalte B wenigstens $(n+2)$ Punkte.

Satz (8.16) (de la Vallee-Pouissin)

Erfüllt V die Haarsche Bedingung bzgl. $\mathrm{C}[a,b]$ und gibt es zu $f \in \mathrm{C}(B)$ und $p \in V$ Punkte $t_0 < t_1 < \ldots < t_{n+1} \in B$, so dass mit einem $\sigma \in \{-1, 1\}$ gilt

$$\mathrm{sign}((f-p)(t_j)) \ = \ \sigma\,(-1)^j, \quad j = 0, \ldots, n+1,$$

so folgt $$\min_j |(f-p)(t_j)| \ \leq \ d_V(f) \ = \ \inf_{q \in V} \max_{t \in B} |f(t) - q(t)|.$$

Gleichheit kann hierbei nur für $|(f-p)(t_j)| = |(f-p)(t_k)|, \ \forall\, j,k$ auftreten.

Beweis:

Sei $(h_0, \ldots, h_n)$ eine Basis von V. Dann hat die folgende Matrix aufgrund der Haarschen Bedingung maximalen Rang $(= n+1)$

$$D(t_0, \ldots, t_{n+1}) \ = \ \begin{pmatrix} h_0(t_0) & \ldots & h_n(t_0) \\ \vdots & & \vdots \\ h_0(t_{n+1}) & \ldots & h_n(t_{n+1}) \end{pmatrix} \ \in \ \mathbb{R}^{(n+2,n+1)} .$$

Es gibt somit einen nichtverschwindenden Vektor $(\lambda_0, \ldots, \lambda_{n+1}) \in \mathbb{R}^{n+2} \setminus \{0\}$ mit den Eigenschaften

$$\begin{aligned} &\text{(a)} \qquad \sum_{j=0}^{n+1} \lambda_j \, h_k(t_j) \;=\; 0, \qquad k \;=\; 0,1,\ldots,n \\ &\text{(b)} \qquad \sum_{j=0}^{n+1} |\,\lambda_j\,| \;=\; 1. \end{aligned} \tag{8.17}$$

Da ferner *jede* quadratische Teilmatrix von $D(t_0,\ldots,t_{n+1})$ aus $n+1$ Zeilen nach der Haarschen Bedingung regulär ist, verschwindet auch keins der λ_j.

Für das lineare Funktional $\ell \in R^*$ mit $\ell(q) := \sum_0^{n+1} \lambda_j \, q(t_j)$ gilt somit $\ell \in V^\perp$ und $\| \ell \| = 1$.

Aufgrund des Dualitätsprinzips, Satz (2.27), folgt $|\,\ell(f)\,| \;\le\; d_V(f)$.

Seien nun $q_k \in V$, $k = 0,1,\ldots,n$, bestimmt durch die Interpolationsbedingungen

$$q_k(t_j) \;=\; 0, \;\; j \in \{0,\ldots,n+1\} \setminus \{k,k+1\}, \qquad q_k(t_k) \;=\; 1.$$

Nach (H2) sind die q_k hiermit eindeutig bestimmt, $q_k \neq 0$ und, da q_k nach (H1) höchstens n Nullstellen besitzen kann, ist auch $q_k(t_{k+1}) \;>\; 0$.

Damit folgt

$$\begin{aligned} 0 \;&=\; \ell(q_k) \;=\; \lambda_k \cdot 1 \;+\; \lambda_{k+1} \; q_k(t_{k+1}) \\ \Rightarrow \;\; \operatorname{sign} \lambda_{k+1} \;&=\; -\operatorname{sign} \lambda_k, \qquad \lambda_j \;\neq\; 0. \end{aligned}$$

und somit

$$\begin{aligned} d_V(f) \;\ge\; |\,\ell(f)\,| \;&=\; |\,\ell(f-p)\,| \;=\; |\sum_{j=0}^{n+1} \lambda_j \,(f-p)(t_j)\,| \\ &=\; \sum_{j=0}^{n+1} |\,\lambda_j\,|\,|\,(f-p)(t_j)\,| \;\ge\; \big(\sum_{j=0}^{n+1} |\,\lambda_j\,|\big) \;\min_j |\,(f-p)(t_j)\,| \\ &=\; \min_j |\,(f-p)(t_j)\,|\,. \end{aligned}$$

Gleichheit kann höchstens dann vorliegen, falls alle $|\,(f-p)(t_j)\,|$ gleich sind. □

Folgerung (8.18)

Gilt für $(n+2)$ Extremalpunkte $t_0 < t_1 < \ldots < t_{n+1}$ von p bezüglich f mit einem festen $\sigma \in \{-1,1\}$

$$(f-p)(t_j) \;=\; \sigma\,(-1)^j \,\| f-p \|_\infty, \quad j = 0,\ldots,n+1,$$

so ist p Bestapproximation von f aus V.

Hierzu gilt nun auch die Umkehrung:

Satz (8.19) (Alternantensatz)

Sei wieder $f \in R = \mathrm{C}(B)$, $B \subset [a,b]$ eine kompakte Menge, die wenigstens $(n+2)$ Punkte enthält. Ferner sei V ein $(n+1)$–dimensionaler Haarscher Teilraum von $\mathrm{C}[a,b]$.
Ein Element $p \in V$ ist genau dann Bestapproximation von f aus V, wenn es $(n+2)$ Punkte $t_0 < t_1 < \ldots < t_{n+1}$ aus B gibt mit

$$(f-p)(t_j) \;=\; \sigma\,(-1)^j \,\|\, f-p \,\|_\infty, \quad j = 0,\ldots,n+1, \qquad \sigma \in \{-1,1\}. \quad (8.20)$$

Das Tupel $(t_0,\ldots,t_{n+1})$ heißt dann eine *Alternante* der Fehlerfunktion $e = f-p$.

Beweis:

Nach (8.18) genügt es zu zeigen, dass es zu jeder Bestapproximation $p \in V$ eine Alternante der Fehlerfunktion gibt.

Sei also $p \in V$ Bestapproximation von $f \in \mathrm{C}(B)$ und nehmen wir an, dass es keine Alternante zu p gäbe. Dann existiert eine Unterteilung

$$a \;=\; \tau_0 \;<\; \tau_1 \;<\; \ldots \;<\; \tau_m \;<\; \tau_{m+1} \;=\; b,$$

wobei $m \le n$ ist, sowie ein $\delta > 0$ und ein $\sigma \in \{-1,1\}$, so dass für die Fehlerfunktion $e := f-p$ und $t \in B \cap [\tau_j, \tau_{j+1}]$, $j = 0,1,\ldots,m$ gilt

$$\sigma\,(-1)^j \;=\; +1 \quad \Rightarrow \quad e(t) \;\in\;]\; -\,\|\,e\,\| + \delta\,,\; \|\,e\,\|\,[$$

$$\sigma\,(-1)^j \;=\; -1 \quad \Rightarrow \quad e(t) \;\in\; [\; -\|\,e\,\|\,,\; \|\,e\,\| - \delta\,[\,.$$

Wir wenden nun (H6) an, vgl. (8.6). Demnach gibt es eine Funktion $q \in V$, die genau die Nullstellen $\tau_1,\ldots,\tau_m$ in $[a,b]$ besitzt und dazwischen die folgende Vorzeichenverteilung besitzt

$$\operatorname{sign} q(t) \;=\; \sigma\,(-1)^j, \qquad t \in]\tau_j, \tau_{j+1}[\,.$$

Für hinreichend kleines $\varepsilon > 0$ erfüllt $\tilde{e} := e - \varepsilon\, q$ damit die Bedingung $\|\,\tilde{e}\,\|_\infty < \|\,e\,\|_\infty$. Das bedeutet aber, dass $p + \varepsilon q \in V$ eine bessere Approximation von f liefert als p. Widerspruch! □

Beispiel (8.21)

Die Funktion $f := \sin$ sei auf einem Intervall $[a,b]$ durch eine Konstante $p \in \Pi_0$ zu approximieren. Enthält das Intervall nun wenigstens zwei Punkte der Form $t_k = (2k+1)\,\pi/2$, $k \in \mathbb{Z}$, so ist $p^* = 0$ Bestapproximation. Mit zwei benachbarten Punken t_k und t_{k+1} ist ja bereits eine Alternante der Länge zwei gegeben.

Enthält das Intervall $[a,b]$ sogar drei Punkte der obigen Form, so ist $p^* = 0$ sogar Bestapproximation bezüglich Π_1.

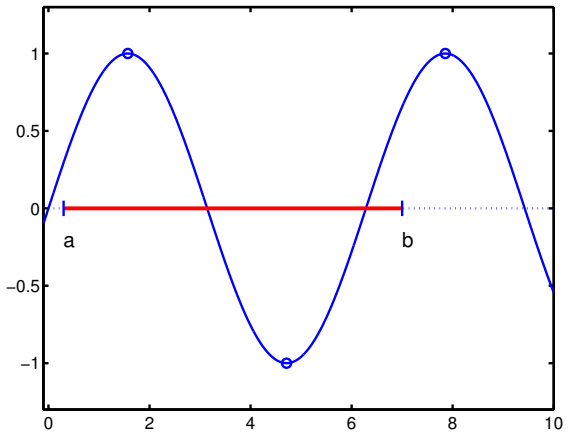


Abb. 8.2 Alternante zu Beispiel (8.21)

Beispiel (8.22)

Die Funktion $f := \sin$ soll im Intervall $[0, \pi/2]$ durch eine Parabel der Form $p(t) = a_0\, t \,+\, a_1\, t^2$ approximiert werden.

Es ist zu beachten, dass der lineare Raum V der Polynome dieser Form *keinen* Haarschen Raum über $[0, \pi/2]$ bildet. Man kann sich damit behelfen, dass man 0 nicht zur Alternante hinzunimmt und das Problem auf einem Intervall $[a, \pi/2]$ mit kleinem $a > 0$ betrachtet (dort ist V ein Haarscher Raum).

Die numerische Berechnung ergibt die Alternante

$$t_0 \doteq 0.28373316, \quad t_1 \doteq 1.10612446, \quad t_2 \doteq 1.57079633,$$

die Bestapproximation: $p^*(t) \doteq 1.13662336\, t \,-\, 0.31121899\, t^2$

und die Minimalabweichung: $\|f - p^*\|_\infty \doteq 0.017501718.$

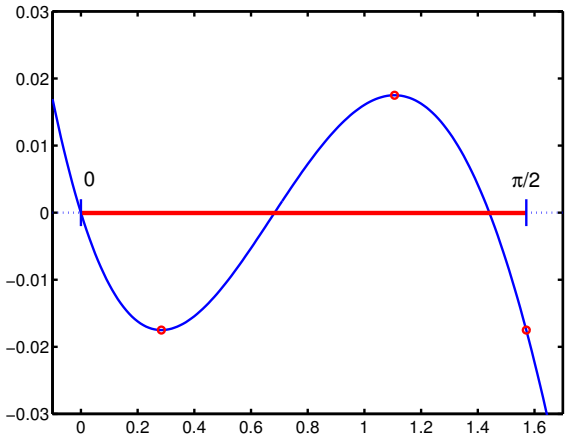


Abb. 8.3 Fehlerfunktion zu Beispiel (8.22)

Beispiel (8.23)

Die Funktion $f(t) := 1/(1+t)$, $0 \le t \le 1$, ist durch ein Polynom $p \in \Pi_2[0,1]$ zu approximieren. Eine numerische Berechnung ergibt die Alternante

$$t_0 \;=\; 0, \quad t_1 \doteq 0.20710678, \quad t_2 \doteq 0.70710678, \quad t_3 \;=\; 1,$$

die Bestapproximation $p^*(t) \;=\; a_0 \,+\, a_1\, t \,+\, a_2\, t^2$ mit

$$a_0 \doteq 0.99264069, \;\; a_1 \doteq -0.82842712, \;\; a_2 \doteq 0.34314575,$$

die Minimalabweichung lautet $\|f - p^*\|_\infty \doteq 0.735931288 \times 10^{-2}$.

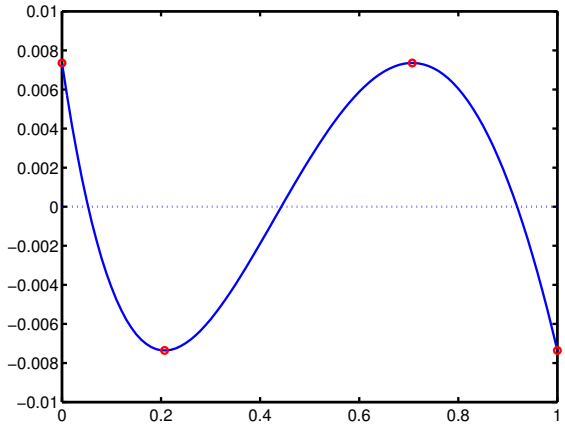


Abb. 8.4 Fehlerfunktion zu Beispiel (8.23)

Satz (8.24)

Sei $R = \mathrm{C}[a,b]$ und V ein $(n+1)$-dimensionaler Teilraum von R, der die konstanten Funktionen enthält. Schließlich sei $f \in R \setminus V$, so dass sowohl V, wie auch $\mathrm{Spann}(V \cup \{f\})$ Haarsche Teilräume sind.
Für die Bestapproximation p von f aus V gilt dann: Die Fehlerfunktion $e := f - p$ besitzt genau $(n+2)$ Extremalpunkte $a = t_0 < \ldots < t_{n+1} = b$. Zwischen benachbarten Extremalpunkten ist e streng monoton. Insbesondere ist die Alternante gemäß (8.19) eindeutig bestimmt und enthält die Randpunkte des Intervalls.

Beweis:

Für ein beliebiges $c \in \mathbb{R}$ ist $f - p - c \in \mathrm{Spann}(V \cup \{f\})$. Daher hat $f - p - c$ höchstens $n+1$ Nullstellen in $[a,b]$. Die Behauptung ergibt sich hieraus mit dem Alternantensatz und dem Zwischenwertsatz. □

Für das Beispiel (8.23) sind die Voraussetzungen des Satzes (8.24) erfüllt. Die Alternante ist also eindeutig bestimmt und enthält die Randpunkte $t_0 = 0$ und $t_4 = 1$.

Beispiel (8.25)

Sei $B := \{t_0, \ldots, t_{n+1}\} \subset [a,b]$, V_n ein $(n+1)$–dimensionaler Haarscher Teilraum von $\mathrm{C}[a,b]$. $h_{n+1} \in \mathrm{C}[a,b] \setminus V$ bezeichne ein neues Basiselement, so dass $V_{n+1} := \mathrm{Spann}(V_n \cup \{h_{n+1}\})$ ebenfalls ein Haarscher Teilraum von $\mathrm{C}[a,b]$ ist.

Zu $f \in \mathrm{C}(B)$ konstruieren wir $g_1, g_2 \in V_{n+1}$ durch die Interpolationsbedingungen

$$g_1(t_j) \;=\; f(t_j), \quad g_2(t_j) \;=\; (-1)^j, \quad j = 0,\, 1,\, \ldots\,,\, n+1.$$

Ferner bestimmen wir $\mu \in \mathbb{R}$ durch die Forderung $p := g_1 - \mu\, g_2 \;\in\; V_n$. Damit stellen wir fest

$$(f-p)(t_j) \;=\; \mu\,(-1)^j, \quad j = 0, 1, \ldots, n+1.$$

Da es in B keine weiteren Punkte gibt, ist $(t_0, \ldots, t_{n+1})$ eine Alternante zur Fehlerfunktion $e := f - p$. Damit ist p Bestapproximation von $f \in \mathrm{C}(B)$ bezüglich V_n.

Beispiel (8.26)

Die Bestapproximation der Funktion $f(t) := t^{n+1} \in \mathrm{C}[-1,1]$ bezüglich $\Pi_n[-1,1]$ ist gegeben durch $p(t) := t^{n+1} \;-\; 2^{-n}\, T_{n+1}(t)$ mit $d_{\Pi_n}(f) = 2^{-n}$.

Die Begründung ergibt sich unmittelbar aus den Eigenschaften der Tschebyscheff–Polynome, vgl. Satz (3.13). Zunächst ist p tatsächlich ein Polynom n-ten Grades, ferner hat die Fehlerfunktion $e \;=\; f - p \;=\; 2^{-n} T_{n+1}$ die Alternante

$$t_k^E \;=\; \cos\big(\frac{n+1-k}{n+1}\,\pi\big),\; k = 0,\, \ldots\,,\, n+1. \qquad \square$$

Wir formulieren noch die Folgerung aus dem Alternantensatz, die sich für die Tschebyscheff-Approximation von $2\,\pi$–periodischen Funktionen durch trigonometrische Polynome ergibt.

Satz (8.27)

Sei $f \in \mathrm{C}_{2\,\pi}$ und $n \in \mathbb{N}$. Dann gibt es genau eine Bestapproximation von f aus T_n bezüglich $\|\cdot\|_\infty$. Diese ist charakterisiert durch eine Alternante der Länge $2\,n+2$ im halboffenen Intervall $[0, 2\,\pi[$.

Beweis: Anwendung des Alternantensatzes für das Intervall $[0,b]$ und Betrachtung des Grenzübergangs $b \uparrow 2\,\pi$. $\square$

Satz (8.28) (Fehlerdarstellung)

Sei $f \in \mathrm{C}^{n+1}[-1,1]$ und $E_n(f) := d_{\Pi_n[-1,1]}(f)$ (bzgl. $\|\cdot\|_\infty$).

a) Gilt für eine Vergleichsfunktion $f_0 \in \mathrm{C}^{n+1}[-1,1]$ die Abschätzung $|\, f^{(n+1)}(t)\,| \;\le\; f_0^{(n+1)}(t),\; \forall\, t \in [-1,1]$, so folgt $E_n(f) \;\le\; E_n(f_0)$.

b) Es gibt ein $\tau \in [-1,1]$ mit

$$E_n(f) \;\leq\; \frac{|\, f^{(n+1)}(\tau)\,|}{2^n\,(n+1)!}. \tag{8.29}$$

Beweis:

zu a): Ist p Bestapproximation von f aus $\Pi_n[-1,1]$, so existiert eine Alternante $(t_0,\ldots,t_{n+1})$ mit

$$(f-p)(t_j) \;=\; (-1)^j\,\mu, \quad |\mu| \;=\; E_n(f).$$

Es sei nun $p_0 \in \Pi_n$ die Bestapproximation von f_0 auf $B := \{t_0,\ldots,t_{n+1}\}$. Dann gilt analog

$$(f_0-p_0)(t_j) \;=\; (-1)^j\,\mu_0, \quad |\mu_0| \;\leq\; E_n(f_0).$$

Für $F := \mu_0\,(f-p) \,-\, \mu\,(f_0-p_0)$ gilt dann $F(t_j)=0,\ j=0,\ldots,n+1$ und somit nach dem Satz von Rolle

$$\exists\,\tau \,\in\, [-1,1]:\;\; F^{(n+1)}(\tau) \;=\; \mu_0\,f^{(n+1)}(\tau) \,-\, \mu\,f_0^{(n+1)}(\tau) \;=\; 0.$$

Gilt nun $|\, f^{(n+1)}(t)\,| \,<\, f_0^{(n+1)}(t)$, für alle $t \in [-1,1]$, so folgt

$$E_n(f) \;=\; |\mu| \;=\; |\mu_0|\,\frac{|\, f^{(n+1)}(\tau)\,|}{|\, f_0^{(n+1)}(\tau)\,|} \;\leq\; |\mu_0| \;\leq\; E_n(f_0).$$

Gilt dagegen nur $|\, f^{(n+1)}(t)\,| \,\leq\, f_0^{(n+1)}(t)$, für alle $t \in [-1,1]$, so wende man die obige Überlegung auf $\widetilde{f}_0(t) := f_0(t) \,+\, \varepsilon\,t^{n+1},\ \varepsilon>0$, an.

Es ist dann $\widetilde{f}_0^{(n+1)}(t) \;=\; f_0^{(n+1)}(x) \,+\, (n+1)!\,\varepsilon \;>\; f_0^{(n+1)}(t)$ und somit

$$E_n(f) \;\leq\; E_n(\widetilde{f}_0) \;\leq\; E_n(f_0) \,+\, 2^{-n}\,\varepsilon.$$

Für $\varepsilon \downarrow 0$ folgt die Behauptung.

zu b): Man setze $f_0(t) := c\,t^{n+1}/(n+1)!$.

Für die Minimalabweichung gilt nach (8.26): $E_n(f_0) \;=\; c\;2^{-n}/(n+1)!$.

Sei nun $\tau \in [-1,1]$ mit

$$c \;:=\; |\, f^{(n+1)}(\tau)\,| \;=\; \max\{\,|\, f^{(n+1)}(t)\,|:\;\; -1 \leq t \leq 1\,\}$$

Dann folgt $|\, f^{(n+1)}(t)\,| \,\leq\, c \,=\, f_0^{(n+1)}(t)$ und damit nach a)

$$E_n(f) \;\leq\; \frac{|\, f^{(n+1)}(\tau)\,|}{2^n\,(n+1)!}$$

□

9. Tschebyscheff-Approximation: Numerik

Der Remez–Algorithmus.

Der meist verwendete Algorithmus zur Lösung von Approximationsaufgaben bezüglich der Tschebyscheff-Norm ist nach dem russischem Mathematiker Evgeny Yakovlevich Remez (1896–1975) benannt.

Wir gehen wieder von der folgenden Standard-Situation aus: $[a,b] \subset \mathbb{R}$ sei ein kompaktes Intervall, $R := \mathrm{C}[a,b]$ sei ausgestattet mit der Maximumsnorm, V sei ein $(n+1)$-dimensionaler Haarscher Teilraum von R.

Ferner sei $B \subset [a,b]$ eine kompakte Teilmenge, die wenigstens $(n+2)$ Punkte enthält. Schließlich sei $f \in \mathrm{C}(B)$.

Wir beginnen mit einer Vorüberlegung, die sich aus dem Beweis des Satzes von de la Vallee-Pouissin ergibt. Es sei $M = \{t_0 < \ldots < t_{n+1}\}$ eine $(n+2)$ elementige Teilmenge von B, die als Schätzung für eine Alternante, vgl. (8.19), zu interpretieren ist. Die Bestapproximation von f bezüglich V **auf M** lässt sich dann folgendermaßen ermitteln.

a) Man bestimme $\lambda_0, \ldots, \lambda_{n+1}$ aus dem homogenen linearen Gleichungssystem

$$\begin{pmatrix} h_0(t_0) & \ldots & h_0(t_{n+1}) \\ \vdots & & \vdots \\ h_n(t_0) & \ldots & h_n(t_{n+1}) \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} . \tag{9.1}$$

Skalierung: $\displaystyle\sum_{k=0}^{n+1} |\lambda_k| = 1, \quad \lambda_0 > 0.$

b) Man bestimme die Koeffizienten a_k in der Darstellung $p = \sum a_k\, h_k$ sowie die Minimalabweichung μ aus den folgenden Bedingungen, vgl. (8.19) und (8.25),

$$f(t_j) - \sum_{k=0}^{n} a_k\, h_k(t_j) = \mu\ \mathrm{sign}\,\lambda_j, \quad j = 0, \ldots, n+1.$$

Dies als lineares Gleichungssystem geschrieben ergibt

$$\begin{pmatrix} h_0(t_0) & \ldots & h_n(t_0) & \mathrm{sign}\,\lambda_0 \\ h_0(t_1) & \ldots & h_n(t_1) & \mathrm{sign}\,\lambda_1 \\ \vdots & & \vdots & \vdots \\ h_0(t_{n+1}) & \ldots & h_n(t_{n+1}) & \mathrm{sign}\,\lambda_{n+1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \\ \mu \end{pmatrix} = \begin{pmatrix} f(t_0) \\ \vdots \\ f(t_n) \\ f(t_{n+1}) \end{pmatrix} . \tag{9.2}$$

Das lineare Gleichungssystem (9.2) ist aufgrund der Voraussetzungen stets eindeutig lösbar (Übungsaufgabe). Ferner folgt aus dem Beweis zum Satz von de la Vallee-Pouissin, dass die λ_k sämtlich von Null verschieden sind und im Vorzeichen alternieren - hierzu wird die Haarsche Bedingung für V als Teilmenge von $C[a,b]$ benötigt.

Zusammen mit der Skalierung $\lambda_0 > 0$ ergibt sich somit

$$\operatorname{sign} \lambda_j \;=\; (-1)^j, \qquad j = 0, \ldots, n+1. \tag{9.3}$$

Damit lässt sich das lineare Gleichungssystem (9.2) ohne explizite Kenntnis der λ_j lösen, nämlich vermöge

$$\begin{pmatrix} h_0(t_0) & \ldots & h_n(t_0) & 1 \\ h_0(t_1) & \ldots & h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ h_0(t_{n+1}) & \ldots & h_n(t_{n+1}) & (-1)^{n+1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \\ \mu \end{pmatrix} = \begin{pmatrix} f(t_0) \\ \vdots \\ f(t_n) \\ f(t_{n+1}) \end{pmatrix}. \tag{9.4}$$

Ablauf des Remez–Algorithmus.

I. Man arbeitet mit einer Schätzung $M_\nu = \{t_0^{(\nu)} < \ldots < t_{n+1}^{(\nu)}\} \subset B$ für die Alternante. Dabei bezeichnet ν den Iterationsindex, die Menge M_ν heißt auch ν-te Referenz. Man löst nun das lineare Gleichungssysten (9.4) und bestimmt damit

a) die Bestapproximation $p^{(\nu)} = \sum_{j=0}^{n} a_j^{(\nu)} h_j \in V$ von f auf M_ν und

b) die zugehörige Minimalabweichung $\mu = \mu_\nu$.

Bemerkung (9.5)

Man beachte auch (8.25) für die Konstruktion der Bestapproximation $p^{(\nu)}$ mittels Interpolationstechniken.

II. Im nächsten Schritt bestimmen wir den tatsächlichen Approximationsfehler in Bezug auf die Menge B

$$\delta_\nu \;:=\; \max\{|(f - p^{(\nu)})(t)| : \; t \in B\} \tag{9.6}$$

Gilt $|\mu_\nu| = \delta_\nu$, so ist $p^{(\nu)}$ nach (8.18) die gesuchte Bestapproximation. Für die numerische Realisierung ersetzen wir diese Relation durch das *Abbruchkriterium*

$$\delta_\nu - |\mu_\nu| \;\leq\; \mathrm{TOL} \cdot |\mu_\nu|. \tag{9.7}$$

Dabei bezeichne TOL eine vorgegebene (relative) Toleranzschranke.

III. Iterationsschritt:

Ist das obige Abbruchkriterium nicht erfüllt, so ist eine neue (bessere) Referenz $M_{\nu+1}$ zu bestimmen. Wir verlangen dabei, dass die folgenden Konvergenz erzeugenden Bedingungen erfüllt sind

$$\begin{aligned}
&\textbf{(a)} \qquad |(f-p^{(\nu)})(t_j^{(\nu+1)})| \;\geq\; |\mu_\nu|, \quad \forall\; j, \\
&\textbf{(b)} \qquad \exists\, j_0: \quad |(f-p^{(\nu)})(t_{j_0}^{(\nu+1)})| \;\geq\; (|\mu_\nu|+\delta_\nu)/2, \\
&\textbf{(c)} \qquad \operatorname{sign}[(f-p^{(\nu)})(t_j^{(\nu+1)})] \;=\; \sigma(-1)^j, \quad \sigma\in\{\pm 1\}.
\end{aligned} \tag{9.8}$$

Im Folgende zeigen wir, dass sich aus diesen drei Eigenschaften tatsächlich die Konvergenz des Verfahrens folgt.

Satz (9.9)

Unter der Annahme $\delta_\nu > |\mu_\nu|$ folgt mit (9.8): $|\mu_{\nu+1}| > |\mu_\nu|$.

Beweis:

Für das maximale lineare Funktional $\ell_{\nu+1}\in V^\perp$ auf $M_{\nu+1}$ gilt

$$\begin{aligned}
|\mu_{\nu+1}| \;&=\; |\ell_{\nu+1}(f)| \;=\; |\sum_{j=0}^{n+1}\lambda_j^{(\nu+1)}\, f(t_j^{(\nu+1)})| \\
&=\; |\sum_{j=0}^{n+1}\lambda_j^{(\nu+1)}\,(f(t_j^{(\nu+1)})-p^{(\nu)}(t_j^{(\nu+1)}))|
\end{aligned}$$

Da die $\lambda_j^{(\nu+1)}$ nicht verschwinden und alternierendes Vorzeichen haben und nach (9.8) (c) auch die Fehlerfunktion $e^{(\nu)}$ auf der neuen Referenz $M_{\nu+1}$ alterniert, ergibt sich

$$|\mu_{\nu+1}| \;=\; \sum_{j=0}^{n+1}|\lambda_j^{(\nu+1)}|\,|f(t_j^{(\nu+1)})-p^{(\nu)}(t_j^{(\nu+1)})|.$$

Wegen $\sum_j|\lambda_j^{(\nu+1)}|=1$ folgt nun mit (9.8) (a) und (b)

$$\begin{aligned}
|\mu_{\nu+1}| \;-\; |\mu_\nu| \;&=\; \sum_{j=0}^{n+1}|\lambda_j^{(\nu+1)}|\,\big(|f(t_j^{(\nu+1)})-p^{(\nu)}(t_j^{(\nu+1)})| \;-\; |\mu_\nu|\big). \\
&\geq\; |\lambda_{j_0}^{(\nu+1)}|\,\big((|\mu_\nu|+\delta_\nu)/2 \;-\; |\mu_\nu|\big) \\
&=\; (1/2)\,|\lambda_{j_0}^{(\nu+1)}|\,\big(\delta_\nu \;-\; |\mu_\nu|\big) \;>\; 0 \qquad\qquad \square
\end{aligned}$$

Satz (9.10)

Die zu einer beliebigen Referenz $M = \{t_0 < t_1 < \ldots < t_{n+1}\} \subset B$ gemäß (9.1) bestimmten λ_j sind nicht nur sämtlich von Null verschieden und alternierend, sondern sogar gleichmäßig von Null weg beschränkt, d.h. zu jedem $c > 0$ existiert $d = d(f) > 0$, so dass für alle Referenzen M gilt

$$|\sum_{j=0}^{n+1} \lambda_j(M) f(t_j)| \ \geq \ c \ \Rightarrow \ \forall\, j: \ |\lambda_j(M)| \ \geq \ d$$

Beweis:

Würde die Behauptung nicht gelten, so gäbe es ein $c > 0$ und eine Folge von Referenzen $M_\nu = \{t_0^{(\nu)} < \ldots < t_{n+1}^{(\nu)}\} \subset B$ mit $|\sum \lambda_j^{(\nu)} f(t_j^{(\nu)})| \geq c, \ \forall \nu,$ so dass für ein $j_0 \in \{0, \ldots, n+1\}$ gilt $\lambda_{j_0}^{(\nu)} \to 0 \ (\nu \to \infty)$. Da B kompakt ist und $\sum |\lambda_j^{(\nu)}| = 1$ lassen sich konvergente Teilfolgen von $(\lambda_j^{(\nu)})$ und $(t_j^{(\nu)})$ finden mit $t_j^{(\nu)} \to t_j, \nu \to \infty$ und $\lambda_j^{(\nu)} \to \lambda_j, \ \nu \to \infty$. Dabei ist natürlich $\lambda_{j_0} = 0$.

Sei nun $q \in V$ durch die Interpolationsbedingungen $q(t_j) = f(t_j)$ für $j \neq j_0$ bestimmt. Dann folgt

$$\begin{aligned} \sum_{j=0}^{n+1} \lambda_j^{(\nu)} f(t_j^{(\nu)}) \ &= \ \sum_{j=0}^{n+1} \lambda_j^{(\nu)} (f-q)(t_j^{(\nu)}) \\ &\to \ \sum_{j=0}^{n+1} \lambda_j (f-q)(t_j) \ = \ 0 \quad (\nu \to \infty) \end{aligned}$$

Widerspruch zur Annahme $|\sum \lambda_j^{(\nu)} f(t_j^{(\nu)})| \geq c > 0$. □

Satz (9.11) (Konvergenz des Remez-Algorithmus)

Erfüllt eine Folge von Referenzen $M_\nu = \{t_0^{(\nu)} < \ldots < t_{n+1}^{(\nu)}\} \subset B, \ \nu \in \mathbb{N}_0$, die Konvergenz erzeugenden Bedingungen (9.8), so konvergieren die zugehörigen $p^{(\nu)} \in V$ auf B gleichmäßig gegen die Bestapproximation p^* von f bezüglich V. Diese Konvergenzaussage gilt unabhängig von der Startreferenz M_0.

Beweis:

(i) Nach (9.9) wächst die Folge der $|\mu_\nu|$ streng monoton. Daher folgt insbesondere

$$\sum_{j=0}^{n+1} \lambda_j^{(\nu)} f(t_j^{(\nu)}) \ = \ |\mu_\nu| \ \geq \ |\mu_1| \ > \ 0.$$

Nach (9.10) und dem Beweis zu (9.9) folgt hiermit

$$|\mu_{\nu+1}| \ - \ |\mu_\nu| \ \geq \ \frac{1}{2} \, |\lambda_{j_0}^{(\nu+1)}| \, (\delta_\nu - |\mu_\nu|) \ \geq \ \frac{d}{2} \, (\delta_\nu - |\mu_\nu|), \tag{9.12}$$

wobei o.E.d.A. $d \in]0,2[$ gewählt werden kann. Damit ergibt sich

$$\begin{aligned} d_V(f) \,-\, |\mu_{\nu+1}| \;&\leq\; d_V(f) \,-\, |\mu_\nu| \,-\, \frac{d}{2}\,(\delta_\nu - |\mu_\nu|) \\ &\leq\; (1-d/2)\,[\,d_V(f) - |\mu_\nu|\,] \end{aligned}$$

Mit $q := (1-d/2) \in]0,1[$ liefert die obige Abschätzung

$$0 \;\leq\; (d_V(f) - |\mu_{\nu+1}|) \;\leq\; q^\nu\,(d_V(f) - |\mu_0|) \;\to\; 0 \;\; (\nu \to \infty)$$

und damit

$$\lim_{k\to\infty} |\mu_\nu| \;=\; d_V(f). \tag{9.13}$$

(ii) Aus (9.12) folgt analog zur obigen Abschätzung

$$\begin{aligned} 0 \;&\leq\; \delta_\nu - d_V(f) \;\leq\; \frac{2}{d}\,(|\mu_{\nu+1}| - |\mu_\nu|) \,+\, |\mu_\nu| \,-\, d_V(f) \\ &=\; \frac{2}{d}\,(|\mu_{\nu+1}| - d_V(f)) \,-\, (\frac{2}{d} - 1)\,(|\mu_\nu| - d_V(f)) \\ &\leq\; (\frac{2}{d} - 1)\,[\,d_V(f) - |\mu_\nu|\,]. \end{aligned}$$

Hieraus folgt: Es gibt eine Konstante $C > 0$ mit

$$0 \;\leq\; \|f - p^{(\nu)}\| - d_V(f) \;\leq\; C\,q^\nu. \tag{9.14}$$

(iii) Aufgrund der strikten Eindeutigkeit der Bestapproximation, vgl. (8.15), gilt mit einem $\gamma = \gamma(f) > 0$ für alle $q \in V$:

$$\|f - q\| \;\geq\; \|f - p^*\| \,+\, \gamma\,\|q - p^*\|,$$

wobei p^* die Bestapproximation von f aus V bezeichnet. Setzt man $q = p^{(\nu)}$, so folgt mit (9.14)

$$\gamma\,\|p^{(\nu)} - p^*\| \;\leq\; \|f - p^{(\nu)}\| \,-\, \|f - p^*\| \;\leq\; C\,q^\nu.$$

Damit ist gezeigt

$$\|p^{(\nu)} - p^*\| \;\leq\; \frac{C}{\gamma}\,q^\nu, \quad \text{mit} \quad 0 < q < 1. \qquad \square$$

Bemerkung (9.15)

I. Allg. müssen die Referenzen M_ν nicht notwendigerweise konvergieren; sie sind ja i. Allg. auch nicht eindeutig bestimmt. Aus Kompaktheitsgründen gibt es jedoch stets eine Teilfolge der (M_ν), die gegen eine Alternante konvergiert.

Der kritische Punkt des Remez-Verfahrens ist es nunmehr, bei vorliegender Referenz M_ν, ein Verfahren zur Konstruktion von $M_{\nu+1}$ anzugeben, so dass die Konvergenz

erzeugenden Eigenschaften (9.8) erfüllt sind. Prinzipiell sind dabei zwei Verfahren in Gebrauch, die als *Einzelaustausch* bzw. *Simultanaustausch* bezeichnet werden.

IV. Einzelaustausch: Man bestimme einen Punkt $\tau \in [a,b]$ mit

$$|\,(f-p^{(\nu)})(\tau)\,| \;\geq\; \frac{1}{2}\,(\delta_\nu + |\mu_\nu|).$$

Hierzu könnte man beispielsweise einen Punkt $\tau \in [a,b]$ bestimmen, an dem $|(f-p^{(\nu)})|$ näherungsweise maximal wird. Dies könnte durch numerische Bestimmung einer Nullstelle der Ableitung $(f-p^{(\nu)})'$ mittels Bisektion oder mittels Newton-Verfahren geschehen, oder (einfacher) durch Absuchen auf einem festen, feinen Gitter des Intervalls $[a,b]$.

Austauschregeln (9.16) $(e_\nu := f - p^{(\nu)})$

(a) Falls $t_{j_0}^{(\nu)} < \tau < t_{j_0+1}^{(\nu)}$:

$$t_{j_0}^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_j^{(\nu)}\,(j \neq j_0), \quad \text{für } \operatorname{sign} e_\nu(t_{j_0}^{(\nu)}) = \operatorname{sign} e_\nu(\tau),$$

$$t_{j_0+1}^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_j^{(\nu)}\,(j \neq j_0+1), \quad \text{für } \operatorname{sign} e_\nu(t_{j_0+1}^{(\nu)}) = \operatorname{sign} e_\nu(\tau),$$

(b) Falls $\tau < t_0^{(\nu)}$:

$$t_0^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_j^{(\nu)}\,(j \neq 0), \quad \text{für } \operatorname{sign} e_\nu(t_0^{(\nu)}) = \operatorname{sign} e_\nu(\tau),$$

$$t_0^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_{j-1}^{(\nu)}\,(j \neq 0), \quad \text{für } \operatorname{sign} e_\nu(t_0^{(\nu)}) \neq \operatorname{sign} e_\nu(\tau),$$

(c) Falls $\tau > t_{n+1}^{(\nu)}$:

$$t_{n+1}^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_j^{(\nu)}\,(j \leq n), \quad \text{für } \operatorname{sign} e_\nu(t_{n+1}^{(\nu)}) = \operatorname{sign} e_\nu(\tau),$$

$$t_{n+1}^{(\nu+1)} := \tau,\; t_j^{(\nu+1)} := t_{j+1}^{(\nu)}\,(j \leq n), \quad \text{für } \operatorname{sign} e_\nu(t_{n+1}^{(\nu)}) \neq \operatorname{sign} e_\nu(\tau).$$

Beispiel (9.17)

Zu approximieren sei

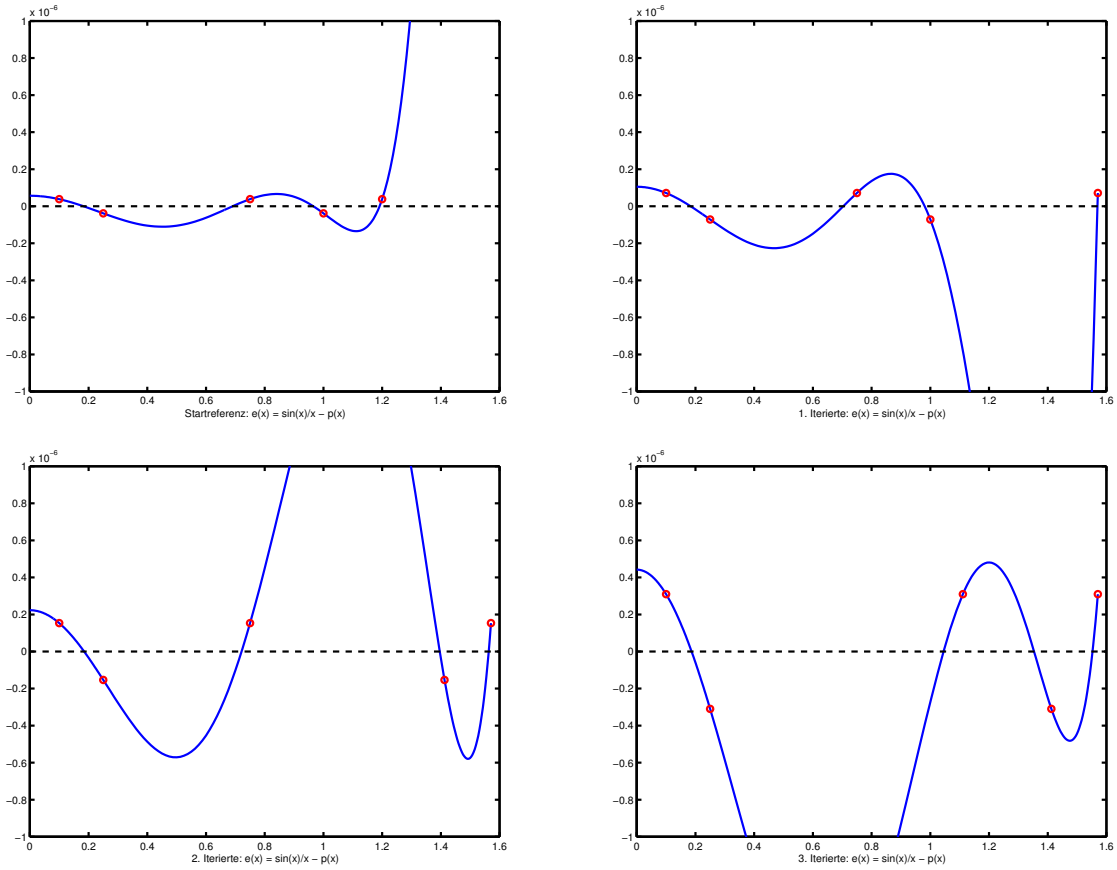
$$f(t) \;=\; \frac{\sin t}{t}, \qquad 0 \;\leq\; t \;\leq\; \pi/2$$

durch ein gerades Polynom $p(t) = a_0 + a_1\,t^2 + a_2\,t^4 + a_3\,t^6$.

In der folgenden Tabelle sind die Referenzen der ersten Iterationsschritte des Algorithmus angegeben

iter	t_0	t_1	t_2	t_3	t_4
0	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.10000e+01$	$0.12000e+01$
1	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.10000e+01$	$0.15708e+01$
2	$0.10000e+00$	$0.25000e+00$	$0.75000e+00$	$0.14125e+01$	$0.15708e+01$
3	$0.10000e+00$	$0.25000e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
4	$0.10000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
5	$0.00000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14125e+01$	$0.15708e+01$
6	$0.00000e+00$	$0.63943e+00$	$0.11109e+01$	$0.14499e+01$	$0.15708e+01$
10	$0.00000e+00$	$0.59967e+00$	$0.11091e+01$	$0.14506e+01$	$0.15708e+01$

Für die Koeffizienten der Bestapproximation erhält die folgenden Näherungen $a_0 = 0.99999e+00$, $a_1 = -0.16666e+00$, $a_2 = 0.83132e-02$ und $a_3 = -0.18524e-03$. Für die Minimalabweichung ergibt sich $\delta = 0.75439e-06$.



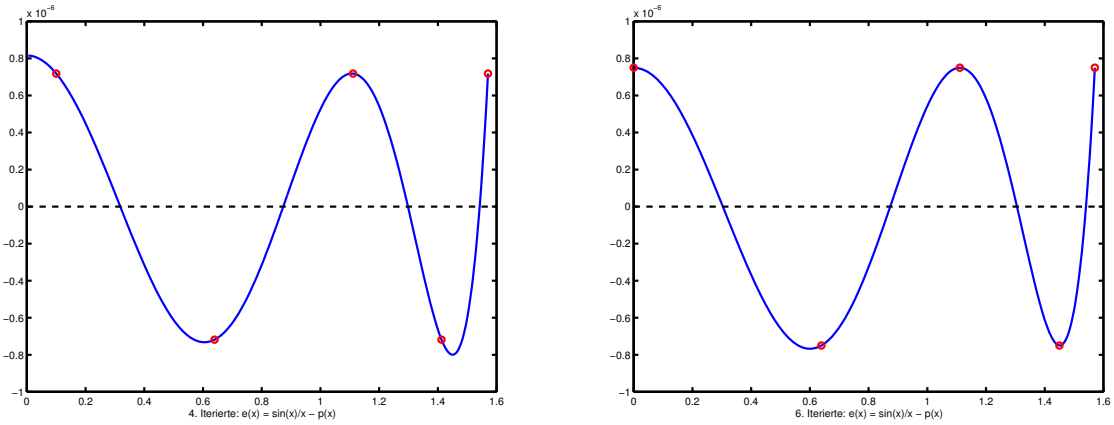


Abb. 9.1: Die ersten Iterationen des Remez-Algorithmus für Beispiel (9.17)

V. Simultanaustausch:

Hierbei zerlegt man das Intervall $[a,b]$ in mindestens $(n+2)$ Teilintervalle, in denen die aktuelle Fehlerfunktion $e_\nu = f - p^{(\nu)}$ abwechselnd nur nichtnegativ bzw. nichtpositiv ist. Man bestimmt dann näherungsweise die Maxima bzw. Minima der Fehlerfunktion in diesen Teilintervallen und wählt hieraus $(n+2)$ Punkte als neue Referenz aus.

Eine Variante dieses Verfahren ergibt sich durch die Bestimmung der lokalen Maxima/Minima der Fehlerfunktion mittels des Newton-Verfahrens für die Ableitung e'_ν. Als Startwerte wählt man die Punkte aus der Referenz M_ν und iteriert wie folgt

$$\begin{aligned} t_0^{(\nu+1)} &= a, \quad t_{n+1}^{(\nu+1)} = b, \quad \text{falls dieses bekannt ist,} \\ t_j^{(\nu+1)} &= t_j^{(\nu)} - \frac{e'_\nu(t_j^{(\nu)})}{e''_\nu(t_j^{(\nu)})}, \quad j = 1,\ldots,n \end{aligned} \tag{9.18}$$

Bemerkungen (9.19)

- Der Remez-Algorithmus mit Simultanaustausch ist i. Allg. schneller als der mit Einzelaustausch. Allerdings ist die Technik zur Sicherung der globalen Konvergenz mühsamer.

- Beim Einzelaustausch lassen sich zur Lösung des linearen Gleichungssystems (9.4) so genannte *update–Techniken* verwenden.

- Für den Remez-Algorithmus mit Simultanaustausch lässt sich unter gewissen Zusatzvoraussetzungen (u.a. die C^2–Eigenschaft von f, h_j) die *quadratische Konvergenz* des Verfahrens zeigen, d.h.

$$\exists\, C > 0: \quad [d_V(f) - |\mu_{\nu+1}|\,] \;\le\; C\,[d_V(f) - |\mu_\nu|\,]^2$$

- Die Wahl der Startreferenz M_0 ist wegen der globalen Konvergenz i. Allg. nicht sehr kritisch; man kann M_0 beispielsweise über eine L_2–Approximation – z.B. eine Tschebyscheff–Entwicklung – erhalten.

Das Newton–Verfahren.

Als Alternative zum Remez–Algorithmus bietet sich an, das nichtlineare Gleichungssystem (8.20) des Alternantensatzes mit einer geeigneten Variante des Newton-Verfahrens zu lösen. Gehören etwa beide Randpunkte des Intervalls $[a, b]$ zur Alternante (ein hinreichendes Kriterium gibt Satz (8.24) an), so liefert der Alternantensatz das folgende nichtlineare Gleichungssystem

$$\begin{aligned} \sum_{j=0}^{n} a_j\, h_j(t_k) \;-\; f(t_k) \;+\; (-1)^k\, \mu \;&=\; 0, \quad k = 0, \ldots, n+1, \\ \sum_{j=0}^{n} a_j\, h_j'(t_k) \;-\; f'(t_k) \;&=\; 0, \quad k = 1, \ldots, n. \end{aligned} \tag{9.20}$$

Dies sind $(2\,n\,+\,2)$ Gleichungen in den $(2\,n\,+\,2)$ Unbekannten $(a_0, \ldots, a_n, t_1, \ldots, t_n, \mu)$. Bezeichnet man mit $e(t) := f(t) - \sum a_j h_j(t)$ wieder die Fehlerfunktion, so lautet die zugehörige Newton-Gleichung, d.h. das lineare Gleichungssystem zur Berechnung der Newton–Korrekturen Δa_j, Δt_k und $\Delta\mu$:

$$\begin{aligned} \sum_{j=0}^{n} h_j(t_k)\,\Delta a_j \;-\; e'(t_k)\,\Delta t_k \;+\; (-1)^k \Delta\mu \;&=\; e(t_k) - (-1)^k \mu, \\ & \qquad k = 0, 1, \ldots, n+1 \\ \sum_{j=0}^{n} h_j'(t_k)\,\Delta a_j \;-\; e''(t_k)\,\Delta t_k \;&=\; e'(t_k), \qquad k = 1, \ldots, n. \end{aligned} \tag{9.21}$$

Dabei ist $\Delta t_0 := \Delta t_{n+1} := 0$, $a_j^{(\nu+1)} := a_j^{(\nu)} + \Delta a_j$, $t_k^{(\nu+1)} := t_k^{(\nu)} + \Delta t_k$ und $\mu^{(\nu+1)} := \mu^{(\nu)} + \Delta\mu$.

Schreibt man dieses lineare Gleichungssystem auf die Unbekannten $a_j^{(\nu+1)}$, Δt_k und $\mu^{(\nu+1)}$ um (dies sollte man nicht machen, wenn man Dämpfungsstrategien verwenden möchte), so ergibt sich mit $t_k = t_k^{(\nu)}$

$$\begin{aligned} \sum_{j=0}^{n} h_j(t_k)\, a_j^{(\nu+1)} \;-\; e'(t_k)\,\Delta t_k \;+\; (-1)^k \mu^{(\nu+1)} \;&=\; f(t_k), \\ & \qquad k = 0, 1, \ldots, n+1 \\ \sum_{j=0}^{n} h_j'(t_k)\, a_j^{(\nu+1)} \;-\; e''(t_k)\,\Delta t_k \;&=\; f'(t_k), \qquad k = 1, \ldots, n. \end{aligned} \tag{9.22}$$

Man vergleiche diese Relationen mit den entsprechenden Gleichungen (9.4) und (9.18) des Remez-Verfahrens.

Zusammenhang zur Linearen Optimierung.

Wir betrachten eine diskrete Approximationsaufgabe im Tschebyscheffschen Sinn. Dazu sei $B \subset [a, b]$ eine endliche Menge, $B = \{t_1, \ldots, t_m\}$ mit $\#B = m > n + 2$.

Wir können ferner davon ausgehen, dass i. Allg. $m \gg n$ gelten wird. Wieder sei $V = V_n$ ein $(n+1)$ dimensionaler linearer Teilraum von $\mathrm{C}(B)$ mit Basis $(h_0, \ldots, h_n)$.

Die *Approximationsaufgabe* lautet: Man bestimme $(a_0, \ldots, a_n) \in \mathbb{R}^{n+1}$, so dass

$$I(a_0, \ldots, a_n) \;:=\; \max\{\, |f(t_k) - \sum_{j=0}^{n} a_j h_j(t_k)| :\;\; k = 1, \ldots, m \,\} \tag{9.23}$$

minimal wird.

Diese Approximationsaufgabe lässt sich nun unmittelbar in eine lineare Optimierungsaufgabe transformieren. Dazu definieren wir

$$\delta \;:=\; \max\{\, |f(t_k) - \sum_{j=0}^{n} a_j h_j(t_k)| :\;\; k = 1, \ldots, m \,\} \tag{9.24}$$

(9.23) ist dann äquivalent zur *linearen Optimierungsaufgabe*:

Bestimme $(a_0, \ldots, a_n, \delta) \in \mathbb{R}^{n+2}$, so dass $J := \delta$ minimal wird unter den Nebenbedingungen

$$\begin{aligned} \sum_{j=0}^{n} a_j\, h_j(t_k) \;-\; \delta \;&\leq\; f(t_k) \\ -\sum_{j=0}^{n} a_j\, h_j(t_k) \;-\; \delta \;&\leq\; -f(t_k), \qquad k = 1, \ldots, m \end{aligned} \tag{9.25}$$

Um die Standardformulierung einer linearen Optimierungsaufgabe zu erhalten führen wir die folgenden Definitionen ein

$$\begin{aligned} \mathbf{A}^{\mathrm{T}} \;&:=\; \begin{pmatrix} h_0(t_1) & \ldots & h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ h_0(t_m) & \ldots & h_n(t_m) & -1 \\ -h_0(t_1) & \ldots & -h_n(t_1) & -1 \\ \vdots & & \vdots & \vdots \\ -h_0(t_m) & \ldots & -h_n(t_m) & -1 \end{pmatrix} \;\in\; \mathbb{R}^{(2m, n+2)} \\ \mathbf{b}^{\mathrm{T}} \;&:=\; [f(t_1) \ldots f(t_m),\;\; -f(t_1) \ldots - f(t_m)] \;\in \mathbb{R}^{2\,m} \\ \mathbf{z}^{\mathrm{T}} \;&:=\; [a_0 \ldots a_n,\;\; \delta] \;\in \mathbb{R}^{n+2} \\ \mathbf{c}^{\mathrm{T}} \;&:=\; [0 \ldots 0,\; -1] \;\in \mathbb{R}^{n+2}. \end{aligned} \tag{9.26}$$

Damit lautet die lineare Optimierungsaufgabe schließlich

$$\text{Maximiere } J_D(\mathbf{z}) \;=\; \mathbf{c}^{\mathrm{T}}\mathbf{z}; \quad \text{Nebenbedingungen:} \quad \mathbf{A}^{\mathrm{T}}\,\mathbf{z} \;\leq\; \mathbf{b}. \tag{9.27}$$

Bemerkungen (9.28)

Die obige Darstellung (9.27) heißt auch **Dualform** einer linearen Optimierungsaufgabe (vgl. Numerik-Vorlesung bzw. Opfer: Numerische Mathematik). Die Zielfunktion $J_D(\mathbf{z})$ ist auf der zulässigen Menge $Z := \{\mathbf{z} : \mathbf{A}^{\mathrm{T}}\mathbf{z} \leq \mathbf{b}\}$ nach oben beschränkt (durch $-\delta = 0$). Ferner ist Z nichtleer, da jeder Punkt $\mathbf{z} = (\mathbf{0}, \delta)^{\mathrm{T}}$ mit $\delta \geq \|f\|_\infty$ zulässig ist. Nach der Theorie der linearen Optimierungsaufgaben (vgl. J.Werner: Numerische Mathematik 2; Satz 2.4) folgt hieraus, dass (9.27) wenigstens eine Lösung $\mathbf{z}^*$ besitzt.

Da i.Allg. $m \gg n$ gelten wird, empfiehlt es sich nicht, die Ungleichungen $\mathbf{A}^{\mathrm{T}}\mathbf{z} \leq \mathbf{b}$ durch Einführung von Schlupfvariablen in Gleichungen zu transformieren. Vielmehr ist es i. Allg. vorteilhaft, anstelle des dualen Problems (9.27) das zugehörige *primale Problem* zu lösen (etwa mit dem Simplexverfahren). Das primale lineare Optimierungsproblem in Normalform lautet

$$\text{Minimiere } J_P(\mathbf{y}) \;=\; \mathbf{b}^{\mathrm{T}}\mathbf{y}; \quad \text{Nebenbedingungen: } \quad \mathbf{A}\,\mathbf{y} \;=\; \mathbf{c}, \quad \mathbf{y} \;\geq\; \mathbf{0}. \tag{9.29}$$

Dabei ist $\mathbf{y} \in \mathbb{R}^{2\,m}$.

Schreibt man diese lineare Optimierungsaufgabe mittels der Definitionen (9.26) wieder explizit auf, so ergibt sich das **primale Optimierungsproblem**

$$\text{Minimiere} \qquad J_P(\mathbf{y}) \;=\; \sum_{k=1}^{m} f(t_k)\,(y_k - y_{m+k}) \tag{9.30}$$

unter den Nebenbedingungen

$$\begin{aligned} &\sum_{k=1}^{m} h_j(t_k)\,(y_k - y_{m+k}) \;=\; 0, \qquad j = 0, \ldots, n \\ &\sum_{k=1}^{2\,m} y_k \;=\; 1, \qquad y_k \;\geq\; 0, \qquad k = 1, \ldots, 2\,m. \end{aligned} \tag{9.31}$$

Bemerkungen (9.32)

a) Die Normierung $\sum y_k = 1$ lässt sich abschwächen zu $\sum y_k \leq 1$. Gilt nämlich für einen zulässigen Punkt $\mathbf{y}$ des relaxierten Problems $\sum y_k < 1$, so erhöhe man irgendein y_{k_0} und das zugehörige y_{m+k_0} um den gleichen Wert $(1 - \sum y_k)/2 > 0$. Der neue Vektor $\widetilde{\mathbf{y}}$ ist dann zulässig für (9.30) bei gleichem Wert der Zielfunktion.

b) Weiterhin lässt sich für das relaxierte lineare Optimierungsproblem durch den Übergang

$$\begin{pmatrix} y_k \\ y_{m+k} \end{pmatrix} \;\to\; \begin{pmatrix} y_k \;-\; \min\{y_k, y_{m+k}\} \\ y_{m+k} \;-\; \min\{y_k, y_{m+k}\} \end{pmatrix} \;\geq\; \mathbf{0}$$

erreichen, dass die *Komplementaritätsbedingung* $y_k\, y_{m+k} \;=\; 0$ erfüllt ist.

Relaxiertes Primales Optimierungsproblem

$$\text{Minimiere} \qquad J_P(\mathbf{y}) \;=\; \sum_{k=1}^{m} f(t_k)\,(y_k - y_{m+k}) \tag{9.33}$$

unter den Nebenbedingungen

$$\begin{aligned} &\sum_{k=1}^{m} h_j(t_k)\,(y_k - y_{m+k}) \;=\; 0, \qquad j = 0,\ldots,n \\ &\sum_{k=1}^{2\,m} y_k \;\le\; 1, \qquad y_k\,y_{m+k} \;=\; 0, \qquad k = 1,\ldots,m \\ &y_k \;\ge\; 0, \qquad k = 1,\ldots,2\,m. \end{aligned} \tag{9.34}$$

Ist $\mathbf{y}$ zulässiger Basisvektor des primalen Optimierungsproblems, so gibt es $I = \{k_0,\ldots,k_{n+1}\} \subset \{1,\ldots,m\}$ mit $y_k = y_{m+k} = 0$ für alle $k \notin I$. Unter Beachtung der Komplementaritätsbedingung sei nun

$$\begin{aligned} \lambda_i \;&:=\; y_{k_i} \;-\; y_{m+k_i}, \\ |\lambda_i| \;&=\; y_{k_i} \;+\; y_{m+k_i}, \qquad i = 0,\ldots,n+1. \end{aligned} \tag{9.35}$$

Damit findet man

$$\begin{aligned} J_P \;&=\; \sum_{i=0}^{n+1} f(t_{k_i})\,\lambda_i \\ \sum_{i=0}^{n+1} h_j(t_{k_i})\,\lambda_i \;&=\; 0, \quad j = 0,\ldots,n \\ \sum_{i=0}^{n+1} |\lambda_i| \;&=\; 1. \end{aligned} \tag{9.36}$$

Sei $\mathbf{y}$ nun *optimale Basislösung*, $\mathbf{z} = (a_0,\ldots,a_n,\delta)^{\mathrm{T}}$ sei Lösung des dualen Problems (9.27). Wegen des Dualitätssatzes , vgl. Vorlesung über Numerik, gilt dann

$$J_D \;=\; J_P \;=\; \mathbf{b}^{\mathrm{T}}\mathbf{y} \;=\; \mathbf{c}^{\mathrm{T}}\mathbf{z}$$

Damit folgt

$$\begin{aligned} 0 \;&=\; \mathbf{y}^{\mathrm{T}}(\mathbf{b} - \mathbf{A}^{\mathrm{T}}\mathbf{z}) \\ &=\; \sum_{k=1}^{m} y_k\,[\,f(t_k) \;-\; \sum_{j=0}^{n} a_j\,h_j(t_k) \;+\; \delta\,] \\ &+\; \sum_{k=1}^{m} y_{m+k}\,[\,-f(t_k) \;+\; \sum_{j=0}^{n} a_j\,h_j(t_k) \;+\; \delta\,] \end{aligned}$$

Hier sind alle Summanden nichtnegativ, so dass sich die folgenden Komplementaritätsbedingungen ergeben

$$\begin{aligned} y_k\,[\,f(t_k)\;-\;\sum_{j=0}^{n} a_j\,h_j(t_k)\;+\;\delta\,]\;&=\;0,\\ y_{m+k}\,[\,-f(t_k)\;+\;\sum_{j=0}^{n} a_j\,h_j(t_k)\;+\;\delta\,]\;&=\;0. \end{aligned} \tag{9.37}$$

Zusammen mit (9.35) und (9.36) folgt nun

$$\begin{aligned} k\;\notin\;I\qquad\qquad &\Rightarrow\;\; y_k\;=\;y_{m+k}\;=\;0,\\ k_i\;\in\;I,\quad \lambda_i>0\;\;&\Rightarrow\;\; f(t_{k_i})\;-\;\sum_{j=0}^{n} a_j\,h_j(t_{k_i})\;=\;-\,\delta,\\ k_i\;\in\;I,\quad \lambda_i<0\;\;&\Rightarrow\;\; f(t_{k_i})\;-\;\sum_{j=0}^{n} a_j\,h_j(t_{k_i})\;=\;\;\delta\,. \end{aligned}$$

Damit lässt sich (9.37) als ein lineares Gleichungssystem zur Berchnung von $\mathbf{z} = (a_0,\ldots,a_n,\delta)^{\mathrm{T}}$ ansehen. Nach Herleitung ist dieses Gleichungssystem stets lösbar; die Eindeutigkeit der Lösung ist jedoch nur für $\lambda_k \neq 0$, für alle $k \in I$, gewährleistet.

Wir fassen das Ergebnis zusammen

Satz (9.38)

a) $p = \sum a_j h_j$ ist genau dann Bestapproximation von f aus V auf $B = \{t_1,\ldots,t_m\}$, wenn es Punkte $t_{k_0} < \ldots < t_{k_{r+1}} \in B$ gibt, $0 \le r \le n$ und $\lambda_i \neq 0$ mit

$$\begin{aligned} &f(t_{k_i})\;-\;\sum_{j=0}^{n} a_j\,h_j(t_{k_i})\;=\;\mathrm{sign}\lambda_i\,\|f-p\|_\infty,\quad i=0,\ldots,r+1,\\ &\sum_{i=0}^{r+1}\lambda_i\,h_j(t_{k_i})\;=\;0,\quad j=0,\ldots,n,\\ &\sum_{i=0}^{r+1}|\lambda_i|\;=\;1. \end{aligned}$$

b) Ist $\mathbf{y}$ optimale Basislösung zu (9.29), so ist das lineare Gleichungssystem (9.37) für $\mathbf{z} = (a_0,\ldots,a_n,\delta)^{\mathrm{T}}$ lösbar. Jede Lösung liefert eine Bestapproximation von f aus V auf B.

Bemerkung (9.39)

Genügt V der Haarschen Bedingung, so ist jede (zulässige) Basislösung des primalen Problems nichtentartet, d.h. $\forall i = 0,\ldots,n+1 :\; y_{k_i} \neq 0$. Ferner lässt sich analog zum Beweis des Satzes von de la Vallee, Pouissin folgern, dass die λ_i nicht verschwinden und alternieren.

Beispiel (9.40) (aus J. Werner: Numerische Mathematik 2)

Zu minimieren sei $\max\{|e^{t_j} - p(t_j)| : j\}$ über $p \in \Pi_4$ auf dem Gitter $t_j = (j-1)/10$, $j = 1, \ldots, 11$.

Das zugehörige (primale) lineare Optimierungsproblem hat die folgende Form

$$\text{Minimiere} \quad J_P \;=\; \mathbf{b}^{\mathrm{T}}\,\mathbf{y}, \qquad \mathbf{y} \in \mathbb{R}^{22}$$

unter den Nebenbedingungen

$$\mathbf{A}\mathbf{y} \;=\; \mathbf{c}, \qquad \mathbf{y} \;\geq\; \mathbf{0}.$$

Dabei ist

$$\mathbf{A}^{\mathrm{T}} \;:=\; \begin{pmatrix} 1 & t_1 & \ldots & t_1^4 & -1 \\ \vdots & & & \vdots & \vdots \\ 1 & t_{11} & \ldots & t_{11}^4 & -1 \\ -1 & -t_1 & \ldots & -t_1^4 & -1 \\ \vdots & & & \vdots & \vdots \\ -1 & -t_{11} & \ldots & -t_{11}^4 & -1 \end{pmatrix} \;\in\; \mathbb{R}^{(22,6)}$$

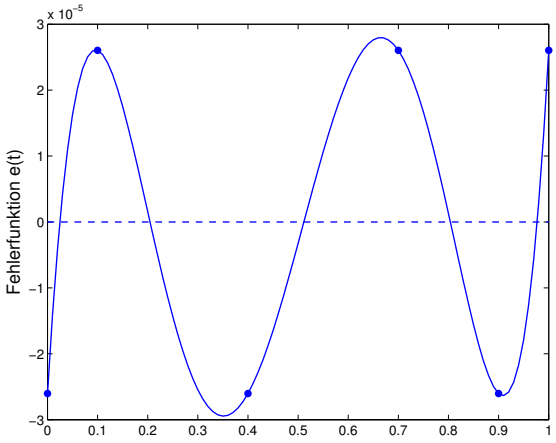
$$\mathbf{b}^{\mathrm{T}} \;:=\; [e^{t_1} \ldots e^{t_{11}},\; -e^{t_1} \ldots - e^{t_{11}}] \;\in \mathbb{R}^{2\,m}$$

$$\mathbf{c}^{\mathrm{T}} \;:=\; [0 \ldots 0,\; -1] \;\in \mathbb{R}^6.$$

Die numerische Lösung des primalen Problems mit Hilfe der MATLAB Routine **linprog** ergibt eine optimale nichtentartete Basislösung mit den Basisindizes $J_B = (1, 5, 10, 13, 19, 22)$.

Das zugehörige lineare Gleichungssystem (9.37) mit diesen Indizes liefert schließlich die Lösung

$$\mathbf{a} \;\approx\; \begin{pmatrix} 1.000026 \\ 0.998714 \\ 0.510077 \\ 0.139716 \\ 0.069722 \end{pmatrix}, \qquad \delta \;\approx\; 2.602631e-05.$$



10. L_1–Approximation

Problemstellung.

Wir betrachten den reellen Vektorraum $R = C[a,b]$, $a < b$, sowie die zugehörige L_1–Norm

$$\| f \|_1 \;:=\; \int_a^b |f(t)|\, dt \tag{10.1}$$

Weiter sei $V \subset R$ ein $(n+1)$-dimensionaler Teilraum und $f \in R \setminus V$. Wir suchen eine L_1–Bestapproximation $p^* \in V$, also

$$\forall\, p \in V: \quad \|f - p^*\|_1 \;\le\; \|f - p\|_1.$$

Wie bisher wird zu einer Approximation $p \in V$ mit $e := f - p$ die *Fehlerfunktion* bezeichnet.
Zur Berechnung von $\|e\|_1$ definieren wir die *Vorzeichenfunktion* $s: [a,b] \to \mathbb{R}$ gemäß

$$s(t) \;:=\; \operatorname{sign}\, e(t) \;:=\; \begin{cases} 1, & \text{falls } e(t) > 0, \\ 0, & \text{falls } e(t) = 0, \\ -1, & \text{falls } e(t) < 0, \end{cases} \tag{10.2}$$

Damit folgt $\|f - p\|_1 \;=\; \|e\|_1 \;=\; \int_a^b s(t)\, e(t)\, dt.$

Beispiel (10.3)

Sei $f \in C^1[a,b]$ streng monoton wachsend, $f' > 0$ und $V = \Pi_0[a,b]$.

Wir haben nun ein $\tau \in]a,b[$ zu bestimmen, so dass für $p_\tau(t) := f(\tau)$ gilt $\|f - p_\tau\|_1$ minimal!

Für die Funktion

$$\Phi(\tau) \;:=\; \|f - p_\tau\|_1 \;=\; \int_a^\tau (f(\tau) - f(t))\, dt \;+\; \int_\tau^b (f(t) - f(\tau))\, dt$$

findet man

$$\begin{aligned} \Phi'(\tau) \;&=\; (f(\tau) - f(\tau)) \;+\; \int_a^\tau f'(\tau)\, dt \;-\; (f(\tau) - f(\tau)) \;-\; \int_\tau^b f'(\tau)\, dt \\ &=\; f'(\tau)\,(2\,\tau - a - b) \end{aligned}$$

Φ' hat also die einzige Nullstelle $\tau^* = (a+b)/2$. Wegen $f' > 0$ ist ferner $\Phi'(t) < 0$ für $t \in [a, \tau^*[$ und $\Phi'(t) > 0$ für $t \in]\tau^*, b]$.

Damit ist τ^* also ein striktes globales Minimum von Φ.

Man beachte die bemerkenswerte Eigenschaft, dass τ^* unabhängig von f ist.

Charakterisierung der Bestapproximation.

Wir nehmen an, dass die Nullstellenmenge

$$Z \;:=\; \{t \in [a,b] :\, e(t) = 0\} \tag{10.4}$$

aus endlich vielen kompakten Teilintervall von $[a,b]$ besteht. Diese können auch einpunktig sein.

Satz (10.5) (Kripke, Rivlin, 1965)

p ist genau dann L_1–Bestapproximation von f aus V, wenn gilt

$$\forall\, q \in V: \quad \Big|\int_a^b s(t)\, q(t)\, dt\,\Big| \;\le\; \int_Z |q(t)|\, dt\,.$$

Bemerkungen (10.6)

a) Für das Beispiel (10.3) ist $Z = \{\tau\}$ und $s(t) = \operatorname{sign}(t-\tau)$. Nach (10.5) ist τ so zu bestimmen, dass

$$\forall\, q \in \Pi_0: \quad \int_a^b s(t)\, q(t)\, dt \;=\; 0.$$

Hieraus folgt sofort $\tau = (a+b)/2$.

b) Besteht Z nur aus endlich vielen Punkten, so besagt der Charakterisierungssatz

$$p \quad L_1 - \text{Bestapproximation} \quad \Leftrightarrow \quad \forall\, q \in V:\; \langle \operatorname{sign}(f-p), q\rangle \;=\; 0$$

Man vergleiche dies auch mit den Charakterisierungen

$$p \quad L_2 - \text{Bestapproximation} \quad \Leftrightarrow \quad \forall\, q \in V:\; \langle f-p, q\rangle \;=\; 0$$

$$p \quad L_\infty - \text{Bestapproximation} \quad \Leftrightarrow \quad \forall\, q \in V:\; \min_{A(f,p)}(f-p)\, q \;\le\; 0$$

Beweis zu (10.5)

⇒: Sei $p \in V$ Bestapproximation, so dass die Charakterisierung

$$\forall\, q \in V: \quad \Big|\int_a^b s(t)\, q(t)\, dt\,\Big| \;\le\; \int_Z |q(t)|\, dt\,.$$

nicht erfüllt ist. Dann existiert ein $q \in V$ mit

$$\eta \;:=\; |\int_a^b s\,q\,| \;-\; \int_Z |\,q\,| \;>\; 0. \qquad (*)$$

O.B.d.A. können wir annehmen, dass $\int sq > 0$ und $\|q\|_\infty = 1$ gelten.

Wir erweitern Z durch die Nachbarmenge $(\Theta > 0)$

$$Z_\Theta \;:=\; \{t \in [a,b] :\; 0 < |f(t) - p(t)| \le \Theta\}.$$

Wegen der Voraussetzung an Z und der Stetigkeit von $|e|$ ist Z_Θ meßbar und für hinreichend kleines $\Theta > 0$ gilt

$$\int_{Z_\Theta} dt \;<\; \eta/2.$$

Schließlich setzen wir $Z_R := [a,b] \setminus (Z \cup Z_\Theta)$.

Zur Abschätzung von $\|f - (p + \Theta q)\|_1$ zerlegen wir das Integral über $[a,b]$ in die drei Teilintegrale über Z, Z_Θ und Z_R.

(i) Für $t \in Z$ hat man $|\,f - p - \Theta\,q\,| \;=\; \Theta\,|\,q\,|$. Damit gilt

$$\int_Z |f - p - \Theta\,q| \;=\; \Theta \int_Z |q|.$$

(ii) Für $t \in Z_\Theta$ gilt

$$|f - p - \Theta\,q| \;\le\; |f - p\,| \;+\; \Theta\,|q| \;\le\; |f - p\,| \;+\; \Theta\,(2 - sq)$$

Die letzte Abschätzung gilt, da $\|q\|_\infty = 1$, also $2 - sq \ge 1$. Somit

$$\int_{Z_\Theta} |f - p - \Theta\,q| \;\le\; \int_{Z_\Theta} |f - p| \;+\; \Theta \int_{Z_\Theta} (2 - sq).$$

(iii) Für $t \in Z_R$ hat man $|f-p| > \Theta$ und damit $\operatorname{sign}(f-p-\Theta q) = \operatorname{sign}(f-p) = s$. Hiermit ergibt sich schließlich

$$|\,f - p - \Theta\,q\,| \;=\; |\,f - p\,| \;-\; \Theta\,s\,q,$$

also

$$\int_{Z_R} |f - p - \Theta\,q| \;=\; \int_{Z_R} |f - p| \;-\; \Theta \int_{Z_R} s\,q.$$

Insgesamt ergibt sich die Abschätzung:

$$\begin{aligned} \|f - (p + \Theta q)\|_1 \;&\le\; \|f - p\|_1 \;+\; \Theta \int_Z |\,q\,| \;+\; \Theta \int_{Z_\Theta} (2 - sq) \;-\; \Theta \int_{Z_R} s\,q \\ &\le\; \|f - p\|_1 \;+\; \Theta \int_Z |\,q\,| \;+\; 2\,\Theta \int_{Z_\Theta} dt \;-\; \Theta \int_a^b s\,q \end{aligned}$$

$$\leq \quad \|f-p\|_1 \;+\; 2\,\Theta \int\limits_{Z_\Theta} dt \;-\; \Theta \left(\int\limits_a^b s\,q \;-\; \int\limits_Z |\,q\,| \right)$$

$$<_{(*)} \quad \|f-p\|_1 \;+\; \Theta\,\eta \;-\; \Theta\,\eta \;=\; \|f-p\|_1.$$

Damit ist aber p keine L_1–Bestapproximation von f.

$\Leftarrow$: Wir haben zu zeigen, dass aus der Charakterisierung

$$\forall\, q \in V: \quad \Big|\int\limits_a^b s(t)\,q(t)\,dt\,\Big| \;\leq\; \int\limits_Z |q(t)|\,dt\,.$$

folgt, dass p Bestapproximation ist. Dazu schätzen wir ab:

$$\begin{aligned}
\|\,f-p+q\,\|_1 &= \int\limits_a^b |\,f(t)-p(t)+q(t)\,|\,dt \\
&= \int\limits_a^b |\,s\,(f-p+q)\,| \;+\; \int\limits_Z |\,f-p+q\,| \\
&\geq \int\limits_a^b s\,(f-p+q) \;+\; \int\limits_Z |\,f-p+q\,| \\
&= \int\limits_a^b s\,(f-p) \;+\; \int\limits_a^b s\,q \;+\; \int\limits_Z |\,q\,| \\
&\geq \int\limits_a^b s\,(f-p) \;=\; \|\,f-p\,\|_1 \qquad \square
\end{aligned}$$

L_1–Approximation für Haarsche Räume.

Satz (10.7)

Ist V ein Haarscher Teilraum von $C[a,b]$, $p \in V$ L_1–Bestapproximation und hat $e := f-p$ nur endlich viele Nullstellen, so wechselt e wenigstens $(n+1)$ mal das Vorzeichen.

Beweis:

Hat e genau m Nullstellen mit Vorzeichenwechsel und ist $m \leq n$, so existiert nach (H6) eine Funktion $q \in V$, die genau diese m Nullstellen besitzt und dort das Vorzeichen wechselt. O.B.d.A. gilt somit

$$\forall\, t \in [a,b] \setminus Z: \quad s(t)\,q(t) \;>\; 0.$$

Damit folgt aber

$$\int_a^b s(t)\, q(t)\, dt \;>\; 0, \qquad \int_Z |\, q(t)\,|\, dt \;=\; 0,$$

im Widerspruch zur Charakterisierung (10.5). □

Satz (10.8) (Eindeutigkeit)

Ist V ein Haarscher Teilraum, so gibt es zu jedem $f \in \mathrm{C}[a,b]$ höchstens eine L_1–Bestapproximation (und damit auch genau eine nach Satz (2.2)).

Beweis: Sind p_1, p_2 Bestapproximationen von f, so folgt für $p := (p_1 + p_2)/2$

$$|\, f(t) - p(t)\,| \;\le\; \frac{1}{2}\,|\, f(t) - p_1(t)\,| \;+\; \frac{1}{2}\,|\, f(t) - p_2(t)\,|\,.$$

Somit ist auch p eine L_1–Bestapproximation und die Integrale über die obige Ungleichung müssen jeweils gleich sein, damit muss aus Stetigkeitsgründen auch punktweise Gleichheit gelten

$$\forall\, t \in [a,b]: \quad |\, f(t) - p(t)\,| \;=\; \frac{1}{2}\,|\, f(t) - p_1(t)\,| \;+\; \frac{1}{2}\,|\, f(t) - p_2(t)\,|\,.$$

Nach (10.7) hat $(f-p)$ aber $(n+1)$ Nullstellen in $[a,b]$. Diese müssen nach Obigem zuleich Nullstellen von $(f-p_1)$ und von $(f-p_2)$ sein. Damit hat aber auch $(p_2-p_1) \in V$ diese $(n+1)$ Nullstellen. Da V ein Haarscher Raum ist, folgt $p_1 = p_2$. □

Satz (10.9) (L_1–Knoten)

Ist V ein Haarscher Teilraum, $p \in V$ L_1–Bestapproximation von $f \in R$ und hat die Fehlerfunktion $e := f - p$ genau $(n+1)$ Nullstellen, so hängen diese nicht von f ab.

Beweis: Die Fehlerfunktion $e := f - p$ habe genau die Nullstellen $\tau_0 < \ldots < \tau_n$ in $[a,b]$.

Zu $g \in R$ sei weiter $q \in V$ L_1-Bestapproximation, so dass die Fehlerfunktion $\widetilde{e} := g - q$ genau die Nullstellen $\sigma_0 < \ldots < \sigma_n$ in $[a,b]$ besitzt.

Nach (10.7) haben alle Nullstellen Vorzeichenwechsel, liegen also insbesondere im offenen Intervall $]a,b[$. Die Anfangwerte $e(a)$ und $\widetilde{e}(a)$ verschwinden nicht und o.E.d.A. nehmen wir an, dass $e(a)\,\widetilde{e}(a) > 0$ und $\tau_0 < \sigma_0$ gelten.

Da V ein Haarscher Raum ist, gibt es ein $h \in V$, das genau die Nullstellen $\tau_1, \ldots, \tau_n$ besitzt und dort jeweils das Vorzeichen wechselt. Wieder gelte o.E.d.A. $e(a)h(a) < 0$.

Nach dem Charakterisierungssatz (10.5) gilt nun mit $s(t) := \operatorname{sign} e(t)$, $\widetilde{s}(t) := \operatorname{sign} \widetilde{e}(t)$:

$$\int_a^b s(t)\, h(t)\, dt \;=\; \int_a^b \widetilde{s}(t)\, h(t)\, dt \;=\; 0$$

und somit auch

$$\int_a^b (\widetilde{s}(t) - s(t))\, h(t)\, dt = 0. \qquad (*)$$

Nach Konstruktion ist $s(t) = \widetilde{s}(t)$ für $t \in [a, \tau_0[$ und $s(t)\, h(t) > 0, \ \forall\, t \in]\tau_0, b] \setminus \{\tau_1, \ldots, \tau_n\}$. Hieraus lässt sich nun unmittelbar ableiten

$$\forall\, t \in [a, b]: \ [s(t) - \widetilde{s}(t)]\, h(t) \geq 0.$$

Zusammen mit $(*)$ folgt somit

$$\forall\, t \in [a, b]: \ [s(t) - \widetilde{s}(t)]\, h(t) = 0.$$

Dies kann aber nur gelten, wenn alle Nullstellen übereinstimmen, also $\forall\, j: \ \sigma_j = \tau_j$ gilt. □

Die Nullstellen der Fehlerfunktion $e = f - p$ sind also (sofern es nur $(n+1)$ Nullstellen gibt) unabhängig von f. Sie heißen die *zu V gehörigen L_1–Knoten* $\tau_0 < \ldots < \tau_n$. Sind diese Knoten bekannt, so lässt sich zu vorgegebenem $f \in R$ die L_1–Bestapproximation p als Lösung der folgenden Interpolationsaufgabe ermitteln

$$p \in V, \text{ mit } p(\tau_j) = f(\tau_j), \ j = 0, \ldots, n.$$

Satz (10.10) (Polynomräume)

Die L_1–Knoten für $V = \Pi_n[-1, 1]$ sind die inneren Extremalstellen des Tschebyscheff-Polynoms T_{n+2}:

$$\tau_k = \cos\left[\frac{n+1-k}{n+2}\,\pi\right], \quad k = 0, 1, \ldots, n.$$

Beweis:

Die τ_k seien wie oben gegeben, ferner sei $\tau_{-1} := -1$ und $\tau_{n+1} := 1$. Die Vorzeichenfunktion lautet also

$$s(t) = \begin{cases} (-1)^j & : \ \tau_{j-1} < t < \tau_j \\ 0 & : \ t = \tau_j \end{cases}$$

und es ist zu zeigen, dass für alle $k = 0, 1, \ldots, n$ gilt

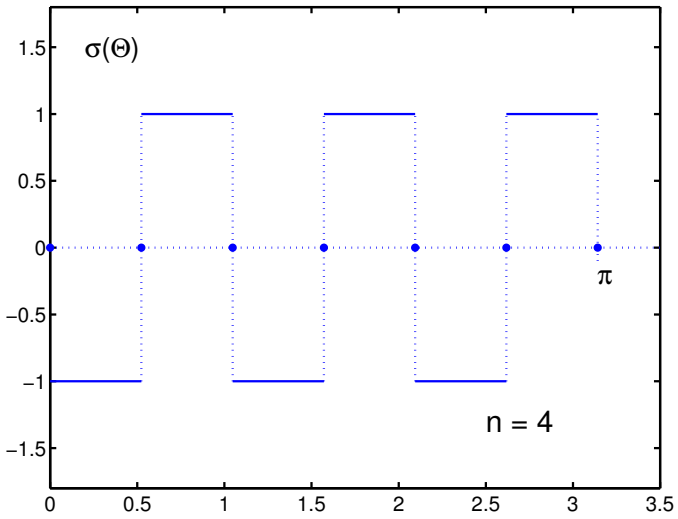
$$\int_{-1}^{1} s(t)\, T_k(t)\, dt = 0. \qquad (10.11)$$

Die Substitution $t = \cos\Theta, \ dt = -\sin\Theta\, d\Theta$ ergibt die zu (10.11) äquivalente Bedingung

$$\int_0^{\pi} \sigma(\Theta)\, \cos(k\Theta)\, \sin(\Theta)\, d\Theta = 0, \qquad k = 0, \ldots, n, \qquad (10.12)$$

wobei

$$\sigma(\Theta) \;:=\; s(\cos\Theta) \;=\; \begin{cases} (-1)^{n-j} & : \;\; \dfrac{j-1}{n+2}\,\pi < \Theta < \dfrac{j}{n+2}\,\pi \\ 0 & : \;\; \Theta \;=\; \dfrac{j}{n+2}\,\pi. \end{cases} \tag{10.13}$$



Wir setzen σ zu einer ungeraden und $2\,\pi$–periodischen Funktion auf $\mathbb{R}$ fort. Damit gilt

$$\forall\,\Theta \in \mathbb{R}: \quad \sigma(\Theta \,+\, \frac{\pi}{n+2}) \;=\; -\sigma(\Theta) \tag{10.14}$$

und es folgt

$$\begin{aligned} \int_0^\pi \sigma(\Theta)\,\cos(k\Theta)\,\sin\Theta\,d\Theta \;&=\; \frac{1}{2}\int_0^\pi \sigma(\Theta)\,[\,\sin((k+1)\Theta) - \sin((k-1)\Theta)\,]\,d\Theta \\ &=\; \frac{1}{4}\int_{-\pi}^\pi \sigma(\Theta)\,[\,\sin((k+1)\Theta) - \sin((k-1)\Theta)\,]\,d\Theta. \end{aligned}$$

Wir setzen nun

$$I_m \;:=\; \int_{-\pi}^\pi \sigma(\Theta)\,\sin(m\,\Theta)\,d\Theta \tag{10.15}$$

Verschiebung des $2\,\pi$–periodischen Integranden mit $\Theta \,=\, \varphi + \dfrac{\pi}{n+2}$ ergibt unter Verwendung von (10.14)

$$
\begin{aligned}
I_m &= \int_{-\pi}^{\pi} \sigma(\varphi + \frac{\pi}{n+2})\ \sin(m\,[\varphi + \frac{\pi}{n+2}])\,d\varphi \\
&= -\int_{-\pi}^{\pi} \sigma(\varphi)\ \left\{\sin(m\,\varphi)\ \cos(\frac{m\pi}{n+2})\ +\ \cos(m\varphi)\ \sin(\frac{m\pi}{n+2})\right\}\,d\varphi \\
&= -\ \cos(\frac{m\pi}{n+2})\ I_m,
\end{aligned}
$$

wobei sich die letzte Gleichheit dadurch ergibt, dass $\sigma(\varphi)$ eine ungerade, $\cos(m\,\varphi)$ jedoch eine gerade Funktion ist.

Hieraus folgt insbesondere $I_m = 0$ für alle $m = 0, 1, \ldots, n+1$ und somit auch

$$
\int_0^{\pi} \sigma(\Theta)\ \cos(k\Theta)\ \sin\Theta\,d\Theta\ =\ 0, \qquad k = 1, \ldots, n.
$$

Für $k = 0$ folgt dies direkt aus $I_1 = 0$. Damit ist (10.12) gezeigt. □

Folgerung (10.16)

Aus dem obigen Beweis halten wir fest: Setzt man für ein vorgegebenes $n \in \mathbb{N}$

$$
\tau_{j,n}\ :=\ \frac{j\,\pi}{n+1}, \qquad j = 0, \ldots, n+1,
$$

sowie

$$
\sigma_n(t)\ :=\ \begin{cases} (-1)^j & :\ \tau_{j,n} < t < \tau_{j+1,n} \\ 0 & :\ t\ =\ \tau_{j,n}, \end{cases}
$$

so gilt $I_{m,n}\ :=\ \int_0^{\pi} \sigma_n(t)\ \sin(m\,t)\,dt\ =\ 0, \quad$ für $m = 0, \ldots, n.$

Beispiel (10.17)

Gesucht sei die L_1–Bestapproximation von $f(t) := t$ auf $[0, \pi]$ bezüglich des folgenden linearen Teilraumes von $C[0, \pi]$:

$$
V_n\ =\ \text{Spann}\{\sin(kt) :\ k = 1, 2, \ldots, n\}.
$$

Da V_n auf dem offenen Intervall $]0, \pi[$ ein Haarscher Raum der Dimension n ist, suchen wir Punkte $0 < \tau_1 < \ldots < \tau_n < \pi$, so dass mit der zugehörigen Vorzeichenfunktion σ_n gilt

$$
\forall\, q \in V_n :\quad \int_0^{\pi} \sigma_n(t)\,q(t)\,dt\ =\ 0.
$$

Nach Folgerung (10.16) sind diese Punkte durch $\tau_j = j\pi/(n+1)$ gegeben. Die Bestapproximation erhält man also durch Lösung der *trigonometrischen Interpolationsaufgabe*

$$p(t) \;=\; \sum_{k=1}^{n} a_k \,\sin(kt); \quad \text{mit} \quad p(\tau_j) = \tau_j, \;\; j = 1, \ldots, n.$$

Wir berechnen die *Minimalabweichung*

$$\begin{aligned}
\int_0^\pi |t - p(t)|\, dt \;&=\; |\int_0^\pi \sigma_n(t)\,(t - p(t))\, dt\,| \;=\; |\int_0^\pi \sigma_n(t)\, t\, dt\,| \\
&=\; |\sum_{j=0}^{n} (-1)^j \int_{\tau_j}^{\tau_{j+1}} t\, dt\,| \;=\; |\sum_{j=0}^{n} (-1)^j\,(\tau_{j+1}^2 \,-\, \tau_j^2)/2\,| \\
&=\; \frac{\pi^2}{2\,(n+1)^2}\,|\sum_{j=0}^{n} (-1)^j\,(2j+1)\,| \;=\; \frac{\pi^2}{2\,(n+1)}\,.
\end{aligned}$$

Man vergleiche hierzu auch den Beweis des ersten Jackson-Satzes (7.12).

11. Darstellung von Kurven und Flächen

Bézier–Kurven.

Unser Ziel ist es, polynomiale Kurven auf dem Rechner möglichst effizient darzustellen. Hierzu nutzen wir die Basisdarstellung mit Hilfe der Bernstein-Polynome aus; man vergleiche hierzu auch Abschnitt 4.

Definition (11.1)

Für $n \in \mathbb{N}_0$ sind die *Bernstein–Polynome* vom Grad n definiert durch

$$B_k^n(t) \;:=\; \left(\begin{array}{c} n \\ k \end{array} \right) t^k \, (1-t)^{n-k}, \quad k \;=\; 0, 1, \ldots, n.$$

Satz (11.2)

Die Bernstein-Polynome erfüllen die folgenden elementaren Eigenschaften

a) $B_k^n(t)$ besitzt eine k-fache Nullstelle in $t = 0$ und eine $(n-k)$-fache Nullstelle in $t = 1$. Ferner hat man die Symmetrie

$$B_k^n(t) \;=\; B_{n-k}^n(1-t), \quad t \in \mathbb{R}, \quad k = 0, 1, \ldots, n.$$

b) Die $B_k^n(t)$ sind auf dem Intervall $[0,1]$ nichtnegativ und haben in $t_k^n := k/n$ ein striktes globales Maximum (bezogen auf $[0,1]$).

c) Es gelten für $t \in \mathbb{R}$

$$1 \;=\; \sum_{k=0}^{n} B_k^n(t), \quad t \;=\; \sum_{k=0}^{n} \frac{k}{n} \, B_k^n(t), \quad t^2 \;=\; \sum_{k=0}^{n} \frac{k\,(k-1)}{n\,(n-1)} \, B_k^n(t). \tag{11.3}$$

Insbesondere bilden die B_k^n, $k = 0, \ldots, n$, eine *Zerlegung der Eins*.

d) Die Bernstein-Polynome lassen sich rekursiv über den folgenden Neville-artigen Algorithmus auswerten

$$\begin{array}{l} B_0^0(t) \;:=\; 1, \\ \text{für} \quad m \;=\; 1, \ldots, n \\ \qquad B_{-1}^{m-1}(t) \;:=\; B_m^{m-1}(t) \;:=\; 0, \\ \qquad B_k^m(t) \;:=\; t\, B_{k-1}^{m-1}(t) \;+\; (1-t)\, B_k^{m-1}(t), \quad k = 0, \ldots, m, \\ \text{end } m. \end{array} \tag{11.4}$$

e) *Weierstraßscher Approximationmssatz:*

Für $f \in \mathrm{C}[0,1]$ konvergieren die Bernstein-Approximationen

$$B_n(f)(t) \;:=\; \sum_{k=0}^{n} f(\frac{k}{n})\; B_k^n(t), \qquad 0 \le t \le 1,$$

für $n \to \infty$ gleichmäßig auf $[0,1]$ gegen die Funktion f.

f) Die Bernstein-Polynome $(B_0^n, \ldots, B_n^n)$ bilden eine Basis des Polynomraums Π_n.

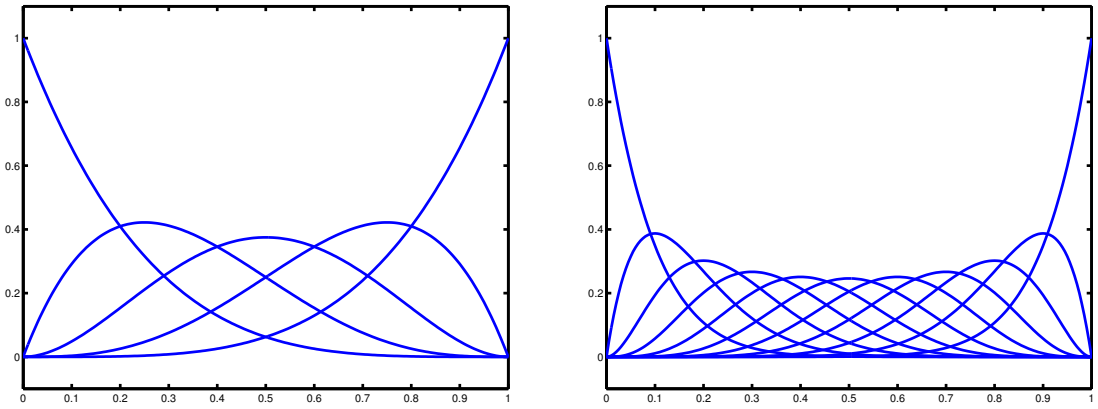


Abb. 11.1 Bernstein-Polynome B_k^n für $n = 4$ und $n = 10$

Definition (11.5)

Aufgrund des Satzes (11.2) f) hat jedes Polynom $p \in \Pi_n$ eine eindeutig bestimmte Darstellung als Linearkombination der Bernstein Polynome

$$p(t) \;=\; \sum_{k=0}^{n} a_k\, B_k^n(t) \tag{11.6}$$

Die Darstellung (11.6) heißt die *Bézier-Darstellung* [1] des Polynoms p.
Die Koeffizienten $a_0, \ldots, a_n$ heißen die *Bèzier-Punkte* von p, das Polygon mit den Ecken $(k/n, a_k)$, $k = 0, \ldots, n$, heißt das *Bézier-Polygon* zu p.

Bemerkung (11.7)

Für die Ableitung der Bernstein-Polynome ergibt sich aus (11.1)

$$\frac{d}{d\,t} B_k^n(t) \;:=\; n\,(B_{k-1}^{n-1}(t) \;-\; B_k^{n-1}(t)), \quad k = 0, \ldots, n, \tag{11.8}$$

wobei wie in (11.4) $B_{-1}^{n-1}(t) \;:=\; B_n^{n-1}(t) \;:=\; 0$ gesetzt wird.

[1] Pierre Etienne Bézier (1910-1999); Paris

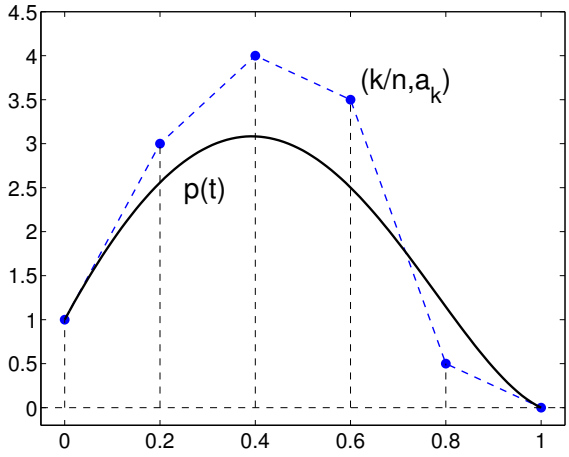


Abb. 11.2 Bézier-Polygon und zugehöriges Polynom

Verwendet man die Beziehung (11.8) zur Differentiation des Polynoms $p(t)$, so ergibt sich

$$p'(t) \;:=\; n \sum_{k=0}^{n-1} (a_{k+1} - a_k)\, B_k^{n-1}(t). \tag{11.9}$$

An dieser Beziehung liest man ab, dass das Bézier-Polynom in den Randpunkten $t = 0$ und $t = 1$ tangential am Bézier-Polygon verläuft, vgl. Abbildung 11.2.

Weiterhin folgt aus der Bézier-Darstellung (11.6) und den Eigenschaften $B_k^n(t) \geq 0$ sowie $\sum_{k=0}^{n} B_k^n(t) = 1$, dass die Werte $p(t)$ ganz in der konvexen Hülle der Bézier-Punkte $(a_0, \ldots, a_n)$ verlaufen müssen:

$$p(t) \;\in\; \mathrm{conv}(a_0, \ldots, a_n). \tag{11.10}$$

Bei Kenntnis der Bézier-Punkte $(a_0, \ldots, a_n)$ ist damit der ungefähre Verlauf des Bézier-Polynoms einzuschätzen. Ferner lässt sich durch Veränderung einzelner Bézier-Punkte gezielt Einfluß auf den Verlauf des Bézier-Polynoms nehmen. Man nennt die Bézier-Punkte daher auch *Kontrollpunkte*.

Die genannten Eigenschaften von Bézier-Polynomen lassen sich unmittelbar auf den vektorwertigen Fall einer polynomialen Kurve $\mathbf{p}(t) \in \mathbb{R}^m$ übertragen. Die Bézier-Darstellung lautet dann

$$\mathbf{p}(t) \;=\; \sum_{k=0}^{n} \mathbf{a}_k\, B_k^n(t), \tag{11.11}$$

wobei die Bézier-Punkte nun Vektoren $\mathbf{a}_k \in \mathbb{R}^m$ sind. Das Bézier-Polygon lässt sich dann als ein Polygonzug im $\mathbb{R}^m$ interpretieren mit den Ecken $(\mathbf{a}_0, \ldots, \mathbf{a}_n)$ und es gelten analog zum skalaren Fall:

$$\begin{aligned} &\mathbf{p}(t) \;\in\; \mathrm{conv}(\mathbf{a}_0, \ldots, \mathbf{a}_n). \\ &\mathbf{p}'(0) \,||\, (\mathbf{a}_1 - \mathbf{a}_0), \qquad \mathbf{p}'(1) \,||\, (\mathbf{a}_n - \mathbf{a}_{n-1}). \end{aligned} \tag{11.12}$$

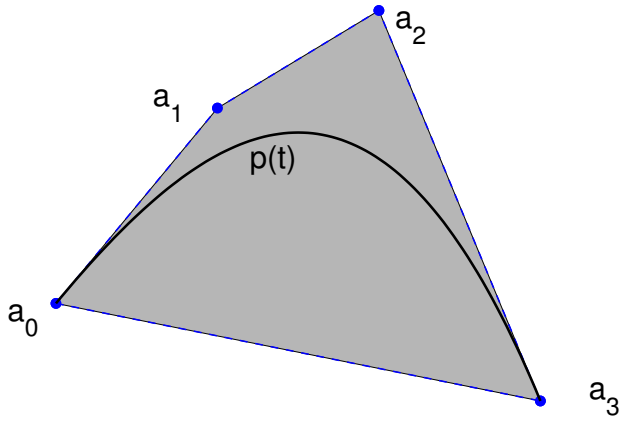


Abb. 11.3 Bézier-Polygon und zugehöriges Polynom im $\mathbb{R}^2$

Der Algorithmus von de Casteljau[2].

Der Wert $p(t)$ eines Bézier-Polynoms (11.6) lässt sich durch fortgesetzte lineare Interpolationen mit Hilfe eines Neville-artigen Algorithmus berechnen. Der Einfachheit halber betrachten wir hier wieder den skalaren Fall und definieren in Verallgemeinerung von (11.6) die folgenden Bézier-Polynome

$$a_i^m(t) \;:=\; \sum_{k=0}^{m} a_{i+k}\, B_k^m(t), \qquad m = 0, \ldots, n, \quad i = 0, \ldots, n-m. \tag{11.13}$$

Offensichtlich sind die $a_i^m(t)$ Polynome aus Π_m. Es gilt $a_i^0(t) = a_i,\ \ i = 0, \ldots, n$, sowie $a_0^n(t) = p(t)$.

Satz (11.14)

Die Bézier-Polynome $a_i^m(t)$, $0 \le t \le 1$, lassen sich wie folgt rekursiv berechnen:

$$\begin{aligned} a_i^0(t) \;&:=\; a_i, \quad i = 0, \ldots, n, \\ a_i^m(t) \;&:=\; (1-t) \cdot a_i^{m-1}(t) \;+\; t \cdot a_{i+1}^{m-1}(t), \quad m = 1, \ldots, n, \quad i = 0, \ldots, n-m. \end{aligned}$$

Damit ist $p(t) = a_0^n(t)$.

Ferner gilt für die Ableitung des Bézier-Polynoms $p'(t) = n \cdot (a_1^{n-1}(t) - a_0^{n-1}(t))$.

Beweis: Wir verwenden die Rekursion (11.4) für die Bernstein-Polynome:

[2]Paul de Faget de Casteljau (geb. 1930 in Besançon); Paris

$$
\begin{aligned}
a_i^m(t) &= \sum_{k=0}^{m} a_{i+k}\, B_k^m(t) \\
&= \sum_{k=0}^{m} a_{i+k} \left(t \cdot B_{k-1}^{m-1}(t) \;+\; (1-t) \cdot B_k^{m-1}(t)\right) \\
&= t \cdot \sum_{k=0}^{m} a_{i+k}\, B_{k-1}^{m-1}(t) \;+\; (1-t) \cdot \sum_{k=0}^{m} a_{i+k}\, B_k^{m-1}(t)) \\
&= t \cdot \sum_{k=0}^{m-1} a_{i+1+k}\, B_k^{m-1}(t) \;+\; (1-t) \cdot \sum_{k=0}^{m-1} a_{i+k}\, B_k^{m-1}(t)) \\
&= t \cdot a_{i+1}^{m-1}(t) \;+\; (1-t) \cdot a_i^{m-1}(t)).
\end{aligned}
$$

Für die Ableitung des Bézier-Polynoms ergibt sich mit (11.9)

$$
p'(t) \;=\; n \sum_{k=0}^{n-1} (a_{k+1} - a_k)\, B_k^{n-1}(t) \;=\; n \left(a_1^{n-1}(t) \;-\; a_0^{n-1}(t)\right). \qquad \square
$$

Tableau von de Casteljau (11.15)

$$
\begin{array}{llll}
a_0 & & & \\
a_1 & a_0^1(t) & & \\
a_2 & a_1^1(t) & \ddots & \\
\vdots & \vdots & & a_0^{n-1}(t) \\
a_n & a_{n-1}^1(t) \quad \ldots & & a_1^{n-1}(t) \qquad a_0^n(t)
\end{array}
$$

Konkretes Zahlenbeispiel: $(n = 3)$

$$
\begin{aligned}
p(t) &= 1 \cdot B_0^3(t) \;+\; 4 \cdot B_1^3(t) \;+\; 3 \cdot B_3^3(t) \;+\; 0 \cdot B_3^3(t) \\
&= (1-t)^3 \;+\; 12 \cdot (1-t)^2\, t \;+\; 9 \cdot (1-t)\, t^2
\end{aligned}
$$

Für $t = 0.4$ (also $1 - t = 0.6$) erhält man das Tableau:

$$
\begin{array}{llll}
1 & & & \\
4 & 2.2 & & \\
3 & 3.6 & 2.76 & \\
0 & 1.8 & 2.88 & 2.808
\end{array}
$$

Somit ist $p(0.4) = 2.808$ und $p'(0,4) = 3 \cdot (2.88 - 2.76) = 0.36$.

Der de Casteljau–Algorithmus in (11.14) beschreibt in der Tat eine iterierte lineare Interpolation, wobei jeweils zwei bereits konstruierte benachbarte Punkte im gleichen Verhältnis t unterteilt werden.

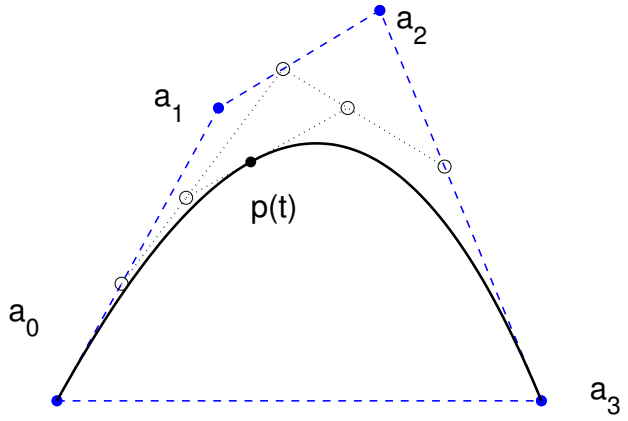


Abb. 11.4 Konstruktion von de Casteljau für $t = 0.4$

Iteriert man diese Konstruktion, wie in Abb. 11.4 angedeutet, so erhält man einen Punkt des Bézier-Polynoms und die Tangente in diesem Punkt. Damit hat man nun einen sehr einfaches geometrisches Werkzeug, um geeignete Kurven zu konstruieren. Diese Hilfmittel wird vielfach im 'Computer Aided Geometric Design' (CAGD) eingesetzt.

Bézier–Flächen.

Unter einer *Parameterdarstellung einer Fläche* im $\mathbb{R}^3$ versteht man eine stetig differenzierbare Abbildung

$$\mathbf{\Phi}: \ [0,1]^2 \ \to \ \mathbb{R}^3, \qquad (u,v)^{\mathrm{T}} \ \mapsto \ \mathbf{x} \ = \ \mathbf{\Phi}(u,v).$$

Zur Approximation von Flächen verwendet man häufig komponentenweise und lokal Polynomräume in zwei Variablen

$$\Pi_{n,m} \ := \ \{\, p(u,v) \ = \ \sum_{i=0}^{n} \sum_{j=0}^{m} a_{ij}\, u^i\, v^j \ : \ u,\, v \in \mathbb{R} \,\}. \tag{11.16}$$

Offensichtlich ist $\Pi_{n,m}$ ein $(n+1)\cdot(m+1)$–dimensionaler $\mathbb{R}$–Vektorraum und man sieht unmittelbar, dass die Produktpolynome

$$B_k^n(u) \cdot B_\ell^m(v), \qquad 0 \le k \le n, \quad 0 \le \ell \le m,$$

eine Basis von $\Pi_{n,m}$ bilden.

Jedes Vektorpolynom $\mathbf{p}(u,v) \in \mathbb{R}^3$ besitzt daher eine eindeutig bestimmte *Bézier-Darstellung*

$$\mathbf{p}(u,v) \;=\; \sum_{i=0}^{n} \sum_{j=0}^{m} \mathbf{a}_{ij} \cdot B_i^n(u) \cdot B_j^m(v). \tag{11.17}$$

Wiederum heißen die Koeffizienten $\mathbf{a}_{ij} \in \mathbb{R}^3$ die *Bézier-Punkte* von (11.17) und die durch die Parameterdarstellung $\mathbf{x} = \mathbf{p}(u,v)$ definierte Fläche heißt *Bézier-Fläche*.

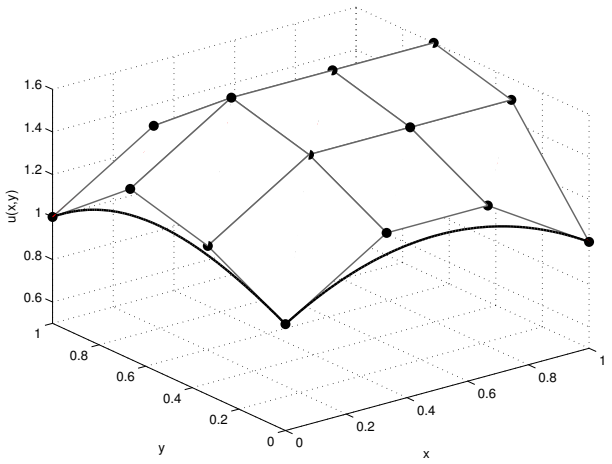


Abb. 11.5 Bikubische Bézier-Fläche im $\mathbb{R}^3$

Die Berechnung der Flächenpunkte $\mathbf{p}(u,v)$ bei vorgegebenen Parametern (u,v) erfolgt durch iterative Verwendung des Algorithmus von de Casteljau für den eindimensionalen Fall. Dazu schreibt man (11.17) wie folgt um

$$\begin{aligned} \mathbf{p}_i(v) \;&=\; \sum_{j=0}^{m} \mathbf{a}_{ij} \cdot B_j^m(v), \quad i \;=\; 0,\ldots,n, \\ \mathbf{p}(u,v) \;&=\; \sum_{i=0}^{n} \mathbf{p}_i(v) \cdot B_i^n(u). \end{aligned} \tag{11.18}$$

Jeder dieser Ausdrücke ist ein eindimensionales Bézier-Polynom zu den Bézier-Punkten $\mathbf{a}_{i0},\ldots\mathbf{a}_{im}$, bzw. $\mathbf{p}_0(v),\ldots\mathbf{p}_n(v)$. Es sind also zur Auswertung von (11.17) jeweils $(n+2)$ eindimensionale Bézier-Polynome im $\mathbb{R}^3$ mittels (11.14) zu berechnen.

Im Folgenden gehen wir noch auf kurz auf zwei wichtige Eigenschaften von Bézier-Flächen ein.

Konvexe Hüllen Eigenschaft:

Es gelten $B_i^n(u) \cdot B_j^m(v) \geq 0$ sowie $\sum_{i=0}^{n} \sum_{j=0}^{m} B_i^n(u) \cdot B_j^m(v) \;=\; 1.$

Damit ist $\mathbf{p}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{m} \mathbf{a}_{ij} \cdot B_i^n(u) \cdot B_j^m(v)$ eine Konvexkombination der Bézier-Punkte $\mathbf{a}_{ij}$. Die Bézier-Fläche $\mathbf{x} = \mathbf{p}(u,v)$ verläuft also ganz in der konvexen Hülle der Bézier-Punkte.

Partielle Ableitungen:

Analog zum eindimensionalen Fall (vgl. Satz (11.14)) lassen sich die partiellen Ableitungen der Bézier-Fläche sehr leicht mit Hilfe des Algorithmus von de Casteljau berechnen. Es gilt nämlich

$$\begin{aligned} \frac{\partial}{\partial u}\mathbf{p}(u,v) &= n \cdot \sum_{i=0}^{n-1}\sum_{j=0}^{m} (\mathbf{a}_{i+1,j} - \mathbf{a}_{ij})\, B_i^n(u)\, B_j^m(v) \\ \frac{\partial}{\partial v}\mathbf{p}(u,v) &= m \cdot \sum_{i=0}^{n}\sum_{j=0}^{m-1} (\mathbf{a}_{i,j+1} - \mathbf{a}_{ij})\, B_i^n(u)\, B_j^m(v). \end{aligned} \tag{11.19}$$

Insbesondere lässt sich hiermit ein Einheits-Normalenvektor an die Bézier-Fläche wie folgt berechnen

$$\mathbf{n}(u,v) = \frac{\mathbf{p}_u(u,v) \times \mathbf{p}_v(u,v)}{\|\mathbf{p}_u(u,v) \times \mathbf{p}_v(u,v)\|}. \tag{11.20}$$

B–Spline Kurven und Flächen.

Ganz ähnliche Darstellungen und Berechnungsmethoden erhält man, wenn man in den Gleichungen (11.11) und (11.17) die Bernstein-Polynome durch B-Splines zu einem festen Gitter

$$t_{-m} < t_{-m+1} < \ldots < t_0 < \ldots < t_n < \ldots < t_{n+m}$$

ersetzt. Die zu (11.11) analoge Darstellung lautet dann

$$\mathbf{p}(t) = \sum_{j=-m}^{n-1} \mathbf{a}_{j+m}\, B_{mj}(t). \tag{11.21}$$

Wieder heißen die $\mathbf{a}_0, \ldots \mathbf{a}_{m+n-1}$ *Kontrolpunkte* (manchmal auch *de Boor-Punkte*) der B-Spline Kurve (11.21).

Wegen $B_{m,j}(t) \geq 0$ und $\sum_{j=-m}^{n-1} B_{m,j}(t) = 1$, vgl. (5.25) und (5.26), erfüllen B-Spline Kurven (und analog B-Spline Flächen) ebenfalls die Konvexe Hüllen Eigenschaft $\mathbf{p}(t) \in \mathrm{conv}(\mathbf{a}_0, \ldots, \mathbf{a}_{m+n-1})$.

Die einfache Berechnung mit Hilfe des de Casteljau–Algorithmus lässt sich ebenfalls auf B-Spline Kurven und Flächen übertragen. Anstelle der Dreiterm-Rekursion (11.4) für die Bernstein-Polynome verwendet man hierzu die Dreiterm-Rekursion (5.23) für die B-Splines.

Gegenüber den Bézier-Kurven und Fläche haben die B-Spline Kurven und Flächen einen wichtigen Vorteil, der sich aufgrund der kompakten Träger der B-Splines ergibt, vgl. (5.25): Fehler in einem Kontrolpunkt $\mathbf{a}_{j+m}$ beeinflussen die B-Spline Kurve $\mathbf{p}(t)$ nur lokal, nämlich im Intervall $t_j \leq t \leq t_{j+m+1}$, dem Träger von $B_{m,j}$.