Approximation

Roland Pulch

Skript zur Vorlesung im Sommersemester 2024

Institut für Mathematik und Informatik
Universität Greifswald

Inhalt:

1. Approximation mit Polynomen und Splines

2. Approximation in normierten Räumen

3. Parameterbestimmung

Literatur:

H.R. Schwarz, N. Köckler: Numerische Mathematik. (8. Aufl.) Vieweg+Teubner 2011. (Kapitel 3)

R. Schaback, H. Wendland: Numerische Mathematik. (5. Aufl.) Springer 2005. (Kapitel 12)

R.A. DeVore, G.G. Lorentz: Constructive Approximation. Springer 1993.

O. Christensen, K.L. Christensen: Approximation Theory: From Taylor Polynomials to Wavelets. Birkhäuser 2005.

Inhaltsverzeichnis

1 Approximation mit Polynomen und Splines

Wir betrachten das folgende Problem. Dabei bezeichnet $C[a,b]$ die Menge aller stetigen Funktionen $f : [a,b] \to \mathbb{R}$, wobei stets $a < b$ vorausgesetzt wird.

Aufgabenstellung: Gegeben sei eine Funktion $f \in C[a,b]$. Aus einer Menge $\mathcal{G} \subset C[a,b]$ von Funktionen einfacher Gestalt soll ein $g \in \mathcal{G}$ gefunden werden, so dass $\|f - g\|_\infty$ klein wird.

Diese Aufgabenstellung ist allgemein formuliert, wodurch eine Approximation g noch nicht eindeutig definiert ist. Unter gewissen Voraussetzungen an die Menge $\mathcal{G}$ existiert eine Bestapproximation $\hat{g}$, d.h. es gilt dann

$$\|f - \hat{g}\|_\infty \leq \|f - g\|_\infty \qquad \text{für alle } \ g \in \mathcal{G}.$$

In diesem Kapitel wird die folgende Strategie angewendet: Es seien genau $n+1$ Knoten $a \leq x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n \leq b$ ausgewählt. An diesen Stellen erfolgen Auswertungen der Funktion f. Gesucht ist dann eine Funktion $g \in \mathcal{G}$ derart, dass die Abweichungen

$$|f(x_j) - g(x_j)| \qquad \text{für } \ j = 0, 1, \ldots, n \tag{1.1}$$

möglichst klein werden. Auch hier ist die Approximation g noch nicht eindeutig bestimmt.

In manchen Anwendungen ist die Funktion f sogar unbekannt und es liegt nur eine endliche Anzahl von Funktionsauswertungen vor, z.B. aus Messungen. In diesem Fall muss das obige Konzept verwendet werden.

Eine naheliegende Idee ist, die Funktion f mit einer Funktion $g \in \mathcal{G}$ an den Stützpunkten zu interpolieren, sofern dies möglich ist. Dann gilt entsprechend $f(x_j) = g(x_j)$ für alle $j = 0, 1, \ldots, n$ und die Differenzen (1.1) werden alle zu null. Die Hoffnung ist, dass dann g auch nahe bei f außerhalb der Knoten liegt und somit $\|f - g\|_\infty$ klein ausfällt.

1.1 Interpolation mit Polynomen

Die Interpolation kann mit Polynomen erfolgen. Wir bezeichnen die Menge aller Polynome vom Grad höchstens n mit $\mathcal{P}_n$. Ein Wertepaar (x_j, y_j) heißt Stützpunkt und x_j darin die Stützstelle. Es gilt der folgende Satz für das Interpolationsproblem.

Satz 1.1 *Zu Stützpunkten (x_j, y_j) für $j = 0, 1, \ldots, n$ mit $x_j \neq x_i$ für $j \neq i$ existiert ein eindeutiges Polynom $p \in \mathcal{P}_n$ mit $p(x_j) = y_j$ für $j = 0, 1, \ldots, n$.*

Beweis: siehe [20], Satz 2.1.1.1.

Das gesuchte Polynom $p \in \mathcal{P}_n$ besitzt die Gestalt

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{n-1} x^{n-1} + \alpha_n x^n$$

mit a priori unbekannten reellen Koeffizienten $\alpha_0, \ldots, \alpha_n$. Daraus ersieht man die Monom-Basis

$$\mathcal{P}_n = \operatorname{span}\left\{1, x^1, x^2, \ldots, x^{n-1}, x^n\right\}.$$

Für theoretische Untersuchungen bei der Polynominterpolation sind die Lagrangepolynome

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad \text{für } \; i = 0, 1, \ldots, n$$

als Basis geeignet, d.h. es gilt $\mathcal{P}_n = \operatorname{span}\{L_0, \ldots, L_n\}$. Die Lagrange-Polynome zeichnen sich durch die Eigenschaft

$$L_i(x_j) = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}$$

aus. Es gilt $\operatorname{grad}(L_i) = n$ für alle i. Die Konstruktion dieser Basis liefert die einfache Darstellung

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

des Interpolationspolynoms, d.h. die Koeffizienten sind aus den Stützpunkten bekannt. Ein entscheidender Nachteil ist jedoch, dass sich bei hinzufügen einer neuen Stützstelle (x_{n+1}, y_{n+1}) zu den bisherigen Stützpunkten dann alle Basispolynome ändern.

In der Praxis werden daher die Newton-Polynome eingesetzt. Sie sind definiert durch

$$N_i(x) = \prod_{j=0}^{i-1} x - x_j = (x - x_0)(x - x_1) \cdots (x - x_{i-1})$$

für $i = 1, \ldots, n$ und $N_0(x) = 1$. Es gilt $\mathcal{P}_n = \operatorname{span}\{N_0, \ldots, N_n\}$ sowie $\operatorname{grad}(N_i) = i$ für alle i. Das Interpolationspolynom besitzt somit die Darstellung

$$p(x) = \sum_{i=0}^{n} \beta_i N_i(x).$$

Die gesuchten Koeffizienten β_i können mit der Methode der Dividierten Differenzen berechnet werden. Der Rechenaufwand ist dabei proportional zu n^2. Die Auswertungen des Polynoms erfolgen dann mit einer Modifikation des Horner-Schemas für die Newton-Polynome (Aufwand proportional zu n). Alternativ kann

das Interpolationspolynom direkt ausgewertet werden mit dem Algorithmus von Aitken-Neville, wobei der Rechenaufwand wieder proportional zu n^2 ausfällt. Näheres hierzu findet man in [18, 20].

Wird zu den gegebenen Stützpunkten (x_j, y_j) für $j = 0, \ldots, n$ eine weiterer Stützpunkt (x_{n+1}, y_{n+1}) hinzugefügt, dann ändern sich die Koeffizienten $\beta_0, \ldots, \beta_n$ nicht und es muss nur der neue Koeffizient β_{n+1} berechnet werden.

Gegeben sei eine Folge $(\Delta_m)_{m\in\mathbb{N}}$ von Mengen aus Stützstellen

$$\Delta_m = \left\{ x_0^{(m)}, x_1^{(m)}, \ldots, x_{n_m}^{(m)} \right\}$$

mit $a \le x_0^{(m)} < x_1^{(m)} < \cdots < x_{n_m}^{(m)} \le b$ für alle m. Wir definieren

$$\rho(\Delta_m) = \max \left\{ \left| x_1^{(m)} - a \right|, \left| x_2^{(m)} - x_1^{(m)} \right|, \ldots, \left| x_{n_m}^{(m)} - x_{n_m-1}^{(m)} \right|, \left| b - x_{n_m}^{(m)} \right| \right\}$$

als Kennzahl für die Feinheit der Zerlegung. Gilt $\rho(\Delta_m) \to 0$, dann geht die Anzahl n_m der Stützpunkte notwendigerweise gegen unendlich.

Beispiele für Folgen von Stützstellen sind $(n_m = m)$:

i) *äquidistante Knoten*:
$h = \frac{b-a}{m}$, $x_j^{(m)} = a + jh$ für $j = 0, 1, \ldots, m$.

ii) *Tschebycheff-Knoten*:
in $[-1, 1]$: $\xi_j^{(m)} = \cos\left(\frac{2j+1}{2(m+1)}\pi\right)$ für $j = 0, 1, \ldots, m$,
in $[a, b]$: $x_j^{(m)} = a + \frac{b-a}{2}(\xi_j^{(m)} + 1)$ für $j = 0, 1, \ldots, m$.

Beide Folgen erfüllen $\rho(\Delta_m) \to 0$.

Die Frage ist nun, für welche Folgen $(\Delta_m)_{m\in\mathbb{N}}$ die zugehörige Folge $(p_m)_{m\in\mathbb{N}}$ der Interpolationspolynome gleichmäßig gegen f konvergiert.

Satz 1.2 (Marcinkiewicz 1939) *Zu jeder stetigen Funktion $f : [a, b] \to \mathbb{R}$ existiert eine Folge $(\Delta_m)_{m\in\mathbb{N}}$, so dass die zugehörigen Interpolationspolynome gleichmäßig gegen f konvergieren.*

Beweis: siehe [15].

Leider liefert der Beweis dieses Satzes kein Konstruktionsverfahren für die Stützstellen. Zudem gilt das folgende negative Resultat bezüglich der Approximation mit einer fest vorgegebenen Verfeinerung.

Satz 1.3 (Faber 1914) *Zu jeder festen Folge $(\Delta_m)_{m\in\mathbb{N}}$ existiert eine stetige Funktion $f : [a, b] \to \mathbb{R}$, so dass die zugehörigen Interpolationspolynome nicht gleichmäßig gegen f konvergieren.*

Beweis: siehe [4].

Unter stärkeren Voraussetzungen an die zu approximierende Funktion erhalten wir jedoch das gewünschte Verhalten.

Satz 1.4 *Sei $f : \mathbb{C} \to \mathbb{C}$ analytisch und $f|_{[a,b]}$ reellwertig. Dann konvergieren die Interpolationspolynome zu jeder Folge $(\Delta_m)_{m\in\mathbb{N}}$ mit $n_m \to \infty$ gleichmäßig gegen f.*

Beweis: siehe [11], Kapitel 5, Abschnitt 4.3.

Im Fall der Tschebyscheff-Knoten läßt sich die Voraussetzung noch wesentlich abschwächen.

Satz 1.5 *Ist $f : [a, b] \to \mathbb{R}$ eine Lipschitz-stetige Funktion, dann konvergiert die Folge der Interpolationspolynome zu den Tschebycheff-Knoten gleichmäßig gegen f.*

Je glatter die Funktion ist, desto schneller erfolgt die Konvergenz.

Satz 1.6 *Für $f \in C^{k+1}[a, b]$ mit $k > 1$ ist die Konvergenzgeschwindigkeit bei den Tschebycheff-Knoten gekennzeichnet durch $\|f - p_m\|_\infty = o(\frac{\log m}{m^k})$ für $m \to \infty$.*

Hinreichend für die Lipschitz-Stetigkeit als Voraussetzung in Satz 1.5 ist bereits $f \in C^1[a, b]$. In Satz 1.6 zeigt die Abschätzung $\log_{10} m \leq m$ für alle m die obere Schranke $o(m^{-(k-1)})$. Somit ist das Ziel für eine wichtige Teilmenge der stetigen Funktionen bereits erreicht. Ein Nachteil dieser Konstruktion ist, dass die Tschebycheff-Knoten nicht ineinander geschachtelt sind. Eine Erhöhung der Knotenanzahl erfordert somit eine komplette Neuberechnung des Interpolationspolynoms.

Beispiel von Runge: Die Funktion

$$f : [-a, a] \to \mathbb{R}, \quad f(x) = \frac{1}{1 + x^2}$$

wird betrachtet für $a \geq 4$, z.B. $a = 5$. Diese Funktion ist nicht analytisch in $\mathbb{C}$, jedoch Lipschitz-stetig in $[-a, a]$. Für äquidistante Knoten konvergieren die Interpolationspolynome nicht gleichmäßig gegen f, sondern das Interpolationspolynom

geht sogar gegen unendlich am Rand. Für die Tschebycheff-Knoten konvergieren die Interpolationspolynome gleichmäßig gegen f.

Es verbleibt die Frage nach einer Konstruktion der Stützstellen im Fall einer beliebigen stetigen Funktion. Als Ausblick betrachten wir das folgende rekursive Verfahren zur Konstruktion des Interpolationspolynoms.

„Gieriger Algorithmus“:

Sei $x_0 = \arg\max_{x\in[a,b]} |f(x)|$.
Setze $p_0 :\equiv f(x_0)$ und $\Delta_0 = \{x_0\}$.
für $n = 1, 2, 3, \ldots$
  bestimme $x_n = \arg\max_{x\in[a,b]} |f(x) - p_{n-1}(x)|$
  setze $\Delta_n = \Delta_{n-1} \cup \{x_n\}$
  bilde $p_n \in \mathcal{P}_n$ aus den Stützstellen Δ_n

Dieser Algorithmus ist „gierig“, da er die Stelle mit dem größten Fehler als neue Stützstelle verwendet und damit der Fehler an dieser Stelle im nächsten Interpolationspolynom zu null wird. Ein Vorteil ist, dass die Stützstellenmengen ineinander geschachtelt sind, d.h. p_{n+1} kann mit wenig Aufwand aus p_n bestimmt werden bei Verwendung der Newton-Basis. Das Verfahren vermeidet die Einschränkung aus dem Satz von Faber, da die Stützstellen in Abhängigkeit von f gewählt werden. Eine Konvergenzanalyse dieses Verfahrens ist noch offen.

1.2 Approximation mit Polynomen

In diesem Abschnitt betrachten wir sowohl die Approximation von kontinuierlichen Funktionen als auch diskreten Daten durch Polynome.

Approximation kontinuierlicher Funktionen

Bezüglich der Approximation mit Polynomen ohne Interpolationsbedingungen gilt der allgemeine Satz.

Satz 1.7 (Weierstraß) *Zu jeder Funktion $f \in C[a, b]$ existiert eine Folge $(p_n)_{n\in\mathbb{N}}$ von Polynomen, welche gleichmäßig gegen f konvergiert.*

Satz 1.2 impliziert bereits diese Aussage. Der Beweis des Satzes 1.7 kann jedoch auch konstruktiv erfolgen über die Bernstein-Polynome.

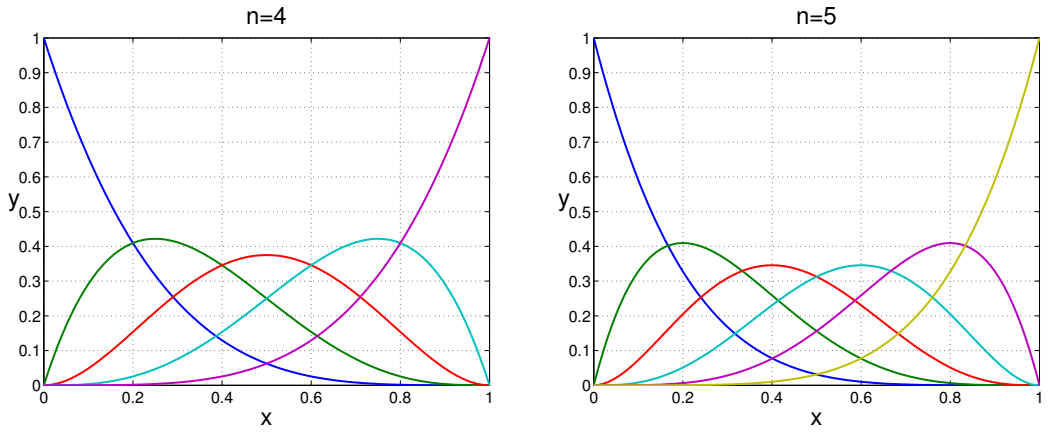


Abbildung 1: Beispiele für Bernstein-Polynome im Intervall $[0, 1]$.

Def. 1.1 *Für ein Intervall* $[a, b]$ *lauten die* Bernstein-Polynome *vom Grad* n

$$B_{i,n}^{[a,b]}(x) = \frac{1}{(b-a)^n} \binom{n}{i} (x-a)^i (b-x)^{n-i}$$

für $i = 0, 1, \ldots, n$.

Es gilt $\mathcal{P}_n = \text{span}\{B_{0,n}^{[a,b]}, \ldots, B_{n,n}^{[a,b]}\}$ und $\text{grad}(B_{i,n}^{[a,b]}) = n$ für alle n.

Weitere Eigenschaften der Bernstein-Polynome:

(i) Positivität: $B_{i,n}^{[a,b]}(x) > 0$ für alle $i = 0, 1, \ldots, n$ und $x \in (a, b)$,

(ii) Zerlegung der Eins:

$$\sum_{i=0}^{n} B_{i,n}(x) = 1 \quad \text{für alle } x \in [a, b]. \tag{1.2}$$

Die Eigenschaft (i) ist offensichtlich, während die Eigenschaft (ii) aus dem Binomischen Lehrsatz folgt. Eine Folgerung aus (i) und (ii) ist $B_{i,n}^{[a,b]}(x) \leq 1$ für alle $i = 0, 1, \ldots, n$ und $x \in [a, b]$. O.E.d.A. verwenden wir im folgenden das Intervall $[0, 1]$, wodurch sich die Gestalt der Bernstein-Polynome vereinfacht zu

$$B_{i,n}(x) = \binom{n}{i} x^i (1-x)^{n-i} \quad \text{für } i = 0, 1, \ldots, n\,.$$

Diese Polynome lassen sich wie folgt zur Approximation einsetzen.

Def. 1.2 *Zu einer stetigen Funktion $f : [0,1] \to \mathbb{R}$ heißt*

$$B_n(x) = \sum_{i=0}^{n} f(\tfrac{i}{n}) B_{i,n}(x) \tag{1.3}$$

das n-te Bernstein-Polynom.

Wir benötigen noch folgende Hilfsaussage.

Lemma 1.1 *Es gilt die Abschätzung*

$$\sum_{i=0}^{n} \left(x - \tfrac{i}{n}\right)^2 \binom{n}{i} x^i (1-x)^{n-i} \leq \frac{1}{4n}$$

für $0 \leq x \leq 1$.

Beweis:

Man benötigt hierfür (1.2) und die beiden Eigenschaften

$$\sum_{i=0}^{n} \frac{i}{n} \binom{n}{i} x^i (1-x)^{n-i} = x,$$
$$\sum_{i=0}^{n} \frac{i(n-i)}{n^2} \binom{n}{i} x^i (1-x)^{n-i} = \tfrac{n-1}{n} x(1-x).$$

Wir rechnen nach

$$\begin{aligned}\sum_{i=0}^{n} \frac{i}{n} \binom{n}{i} x^i (1-x)^{n-i} &= \sum_{i=1}^{n} \frac{i\,n!}{n\,i!(n-i)!} x^i (1-x)^{n-i} = \sum_{i=1}^{n} \binom{n-1}{i-1} x^i (1-x)^{n-i} \\ &= x \sum_{i=0}^{n-1} \binom{n-1}{i} x^i (1-x)^{n-1-i} = x(x+(1-x))^{n-1} \;=\; x\end{aligned}$$

und

$$\begin{aligned}&\sum_{i=0}^{n} \frac{i(n-i)}{n^2} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{i=1}^{n-1} \frac{i(n-i)n!}{n^2 i!(n-i)!} x^i (1-x)^{n-i} \\ &= \frac{n-1}{n} x(1-x) \sum_{i=1}^{n-1} \frac{(n-2)!}{(i-1)!(n-1-i)!} x^{i-1} (1-x)^{n-i-1} \\ &= \frac{n-1}{n} x(1-x) \sum_{i=0}^{n-2} \frac{(n-2)!}{i!(n-2-i)!} x^i (1-x)^{n-i-2} \\ &= \frac{n-1}{n} x(1-x)(x+(1-x))^{n-2} \;=\; \frac{n-1}{n} x(1-x).\end{aligned}$$

Sei als Abkürzung $R_i = \binom{n}{i} x^i (1-x)^{n-i}$. Dadurch erhalten wir

$$\begin{aligned}
\sum_{i=0}^{n} \left(x - \frac{i}{n} \right)^2 R_i &= \sum_{i=0}^{n} \left(x^2 - \frac{2i}{n} x + \frac{i^2}{n^2} \right) R_i \\
&= x^2 \sum_{i=0}^{n} R_i - 2x \sum_{i=0}^{n} \frac{i}{n} R_i + \sum_{i=0}^{n} \frac{i^2}{n^2} R_i \\
&= x^2 - 2x \cdot x - \sum_{i=0}^{n} \frac{i(n-i)}{n^2} R_i + \sum_{i=0}^{n} \frac{i}{n} R_i \\
&= -x^2 - \left(1 - \tfrac{1}{n}\right) x(1-x) + x \;=\; \tfrac{1}{n} x(1-x).
\end{aligned}$$

Man kann schließlich abschätzen $x(1-x) \leq \frac{1}{4}$ für $x \in [0,1]$. □

Es ergibt sich dann das gewünschte Verhalten.

Satz 1.8 *Für eine stetige Funktion $f : [0,1] \to \mathbb{R}$ konvergiert die Folge der Bernstein-Polynome gleichmäßig gegen f.*

Beweis:

Sei $\varepsilon > 0$. Da f auf $[0,1]$ gleichmäßig stetig ist gibt es ein $\delta > 0$, so dass für alle $x, y \in [0,1]$ gilt

$$|x - y| < \delta \quad \Rightarrow \quad |f(x) - f(y)| < \varepsilon.$$

Wegen (1.2) und (1.3) gilt

$$|f(x) - B_n(x)| = |f(x) \cdot 1 - B_n(x)| \leq \sum_{i=0}^{n} |f(x) - f(\tfrac{i}{n})| \binom{n}{i} x^i (1-x)^{n-i}.$$

Wir spalten die Summe in zwei Teile auf gemäß

$$A_n = \left\{ i \;:\; 0 \leq i \leq n,\; \left| x - \tfrac{i}{n} \right| < \delta \right\}, \qquad A_n' = \left\{ i \;:\; 0 \leq i \leq n,\; \left| x - \tfrac{i}{n} \right| \geq \delta \right\},$$

d.h. $A_n \cap A_n' = \emptyset$ und $A_n \cup A_n' = \{0, 1, \ldots, n\}$. Mit der gleichmäßigen Stetigkeit gilt

$$|f(x) - f(\tfrac{i}{n})| < \varepsilon \qquad \text{für } \; i \in A_n$$

und desweiteren die grobe Abschätzung mit Dreiecksungleichung

$$|f(x) - f(\tfrac{i}{n})| \leq |f(x)| + |f(\tfrac{i}{n})| \leq 2\|f\|_\infty \qquad \text{für } \; i \in A_n'.$$

Sei $R_i = \binom{n}{i} x^i (1-x)^{n-i}$. Damit können wir abschätzen unter Verwendung von Lem-

ma 1.1

$$\begin{aligned}\sum_{i=0}^{n} |f(x) - f(\tfrac{i}{n})|\, R_i &= \sum_{i\in A_n} |f(x) - f(\tfrac{i}{n})|\, R_i + \sum_{i\in A_n'} |f(x) - f(\tfrac{i}{n})|\, R_i \\ &< \varepsilon \sum_{i\in A_n} R_i + 2\|f\|_\infty \sum_{i\in A_n'} R_i \\ &\leq \varepsilon \sum_{i=0}^{n} R_i + \frac{2\|f\|_\infty}{\delta^2} \sum_{i=0}^{n} \left(x - \tfrac{i}{n}\right)^2 R_i \\ &\leq \varepsilon + \frac{\|f\|_\infty}{2n\delta^2}.\end{aligned}$$

Obwohl A_n, A_n' von x abhängen ist die obere Schranke nun gleichmäßig für alle x. Da n beliebig war, wählen wir n_0 derart, dass $\|f\|_\infty/(2n_0\delta^2) < \varepsilon$ erfüllt ist. Für alle $x \in [0,1]$ und alle $n \geq n_0$ gilt dann $|f(x) - B_n(x)| < 2\varepsilon$ und die gleichmäßige Konvergenz ist gezeigt. □

Als Folgerung aus Satz 1.8 ergibt sich Satz 1.7. Zur Konvergenzgeschwindigkeit der Approximation gilt die asymptotische Formel.

Satz 1.9 (Voronovskaya 1932) *Sei die Funktion $f : [0,1] \to \mathbb{R}$ beschränkt und $x_0 \in [0,1]$, so dass f in einer Umgebung von x_0 differenzierbar ist und $f''(x_0)$ existiert. Dann gilt*

$$\lim_{n\to\infty} n\,[f(x_0) - B_n(x_0)] = \tfrac{1}{2} x_0(1-x_0) f''(x_0).$$

Beweis: siehe Kapitel 10, Theorem 3.1 in [3].

Für ein stetiges f mit $f''(x_0) \neq 0$ in einem Punkt x_0 liegt damit Konvergenz in x_0 mit der Konvergenzgeschwindigkeit $\mathcal{O}(\frac{1}{n})$ vor. Die Geschwindigkeit der gleichmäßigen Konvergenz ist damit höchstens so schnell wie $\mathcal{O}(\frac{1}{n})$, d.h. relativ langsam im Vergleich zu beispielsweise der Splineinterpolation. Zudem bewirkt eine erhöhte Glattheit von f keine Konvergenzbeschleunigung.

Für nur Lipschitz-stetige Funktionen ergibt sich eine langsamere Konvergenz.

Satz 1.10 *Für eine Lipschitz-stetige Funktion $f : [0,1] \to \mathbb{R}$ mit Lipschitz-Konstante $L > 0$ gilt*

$$\|B_n - f\|_\infty \leq \frac{L}{2} \cdot \frac{1}{\sqrt{n}}.$$

Beweis: siehe [11], Kapitel 4, Abschnitt 2.5.

Für die Ableitungen ergibt sich jedoch ein erstaunlich starkes Resultat.

Satz 1.11 *Falls $f \in C^k[0,1]$, dann konvergieren die i-ten Ableitungen der Bernstein-Approximationen B_n gleichmäßig gegen die Ableitungen $f^{(i)}$ für jedes $i \leq k$.*

Beweis: siehe [3], Kapitel 10, Theorem 2.1.

Approximation diskreter Daten

Statt einer kontinuierlichen Funktion werden nun nur endlich viele Stützpunkte angenähert.

Aufgabenstellung: Gegeben Stützpunkte (x_i, y_i) für $i = 0, 1, \ldots, m$ mit $x_j \neq x_i$ für $j \neq i$. Gesucht ist ein Polynom $p \in \mathcal{P}_n$ mit $n < m$, so dass die Zielfunktion

$$J = \sum_{i=0}^{m} (y_i - p(x_i))^2 \tag{1.4}$$

minimal wird.

Wir setzen das gesuchte Polynom in der Monom-Basis an, d.h.

$$p(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1} x^{n-1} + \alpha_n x^n.$$

Gesucht ist daher

$$\min_{\alpha_0,\ldots,\alpha_n} \sum_{i=0}^{m} \left(y_i - (\alpha_0 + \alpha_1 x_i + \cdots + \alpha_{n-1} x_i^{n-1} + \alpha_n x_i^n) \right)^2 .$$

Im Spezialfall $n = m$ können für die Koeffizienten $\alpha_0, \ldots, \alpha_n$ das Interpolationspolynom eingesetzt werden. Dadurch wird die Zielfunktion (1.4) null und somit liegt bereits das eindeutige Minimum vor.

Bei dieser Approximationsaufgabe ist eine Polynom von niedrigem Grad n erwünscht, welches dennoch eine Menge von Stützpunkten mit hohem m gut approximiert. Daher wird die Frage der Konvergenz für $n \to \infty$ nicht gestellt. Für $n \geq m$ liegt das Interpolationspolynom vor und somit gelten die Aussagen aus Abschnitt 1.1.

Die obige Aufgabenstellung ist bei der Approximation von kontinuierlichen Funktionen f sinnvoll, wenn die Funktionsauswertungen $y_i = f(x_i) + \varepsilon_i$ mit Fehlern ε_i (z.B. Messfehlern) behaftet sind. Eine Interpolation der Werte würde damit unnötigerweise den Fehler mit erfassen.

Wir definieren die Vektoren

$$\theta = (\alpha_0, \ldots, \alpha_n)^\top \in \mathbb{R}^{n+1}, \quad y = (y_0, \ldots, y_m)^\top \in \mathbb{R}^{m+1}$$

und die Matrix

$$\Phi = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{pmatrix} \in \mathbb{R}^{(m+1)\times(n+1)}. \tag{1.5}$$

Die Aufgabenstellung läßt sich daher als lineares Ausgleichsproblem schreiben, d.h.

$$\min_{\theta\in\mathbb{R}^{n+1}} \|y - \Phi\theta\|_2^2$$

mit der Euklidischen Norm $\|\cdot\|_2$. Die Matrix Φ besitzt vollen Spaltenrang, da die Stützstellen paarweise verschieden sind. Somit existiert eine eindeutige Lösung des Ausgleichsproblems. Die Lösung θ kann über eine QR-Zerlegung von Φ berechnet werden, siehe [20]. Die QR-Zerlegung erfolgt mit entweder Householder-Transformationen oder Givens-Rotationen.

Ein Hinzufügen von weiteren Stützpunkte $(x_i, f(x_i))$ bedeutet hier die Vergrößerung der Matrix Φ um weitere Zeilen. Wurde eine Matrix Φ bereits transformiert, dann brauchen die Transformationen nur auf den neu hinzugekommenen Teil angewendet zu werden.

1.3 Interpolation mit Splines

Eine naheliegende Idee ist es, statt mit Polynomen die Interpolation mit stückweise polynomialen Funktionen durchzuführen. Sei eine Zerlegung Δ der Form $a = x_0 < x_1 < \cdots < x_n = b$ gegeben. Es bezeichnet

$$\rho(\Delta) = \max_{j=1,\ldots,n} |x_j - x_{j-1}|$$

die Feinheit der Zerlegung. Man beachte, dass die beiden Randpunkte hier Stützstellen sind.

Def. 1.3 *Die Menge der Splines vom Grad k zu einer Zerlegung Δ von $[a, b]$ ist definiert durch*

$$\mathcal{S}_k(\Delta) = \left\{ s \in C^{k-1}[x_0, x_n] \ : \ s|_{[x_{j-1},x_j]} \in \mathcal{P}_k \text{ für } j = 1, \ldots, n \right\}.$$

In der Praxis werden meist nur Splines vom Grad $k = 1, 2, 3, 4, 5$ eingesetzt, wobei der Fall $k = 3$ am häufigsten auftritt.

Lineare Splines

Im Fall $k = 1$ entstehen als Splines stetige Streckenzüge. Der lineare interpolierende Spline interpoliert dann eine Funktion f an den Stützpunkten $(x_j, f(x_j))$ für $j = 0, 1, \ldots, n$. Es folgt die Formel

$$s(x) = \frac{x_j - x}{x_j - x_{j-1}} f(x_{j-1}) + \frac{x - x_{j-1}}{x_j - x_{j-1}} f(x_j) \quad \text{für } x \in [x_{j-1}, x_j].$$

Die Stetigkeit von s ist mit dieser Konstruktion sichergestellt. Folgende Approximationsgüte kann gezeigt werden.

Satz 1.12 *Sei $x_j = a + jh$ für $j = 0, 1, \ldots, n$ mit $h = \frac{b-a}{n}$. Ist $f \in C^2[a, b]$, dann gilt für den linearen interpolierenden Spline s*

$$\|f - s\|_\infty \leq \tfrac{1}{8}\|f''\|_\infty h^2 \qquad \textit{und} \qquad \|f' - s'\|_\infty \leq \tfrac{1}{2}\|f''\|_\infty h,$$

wobei die Ableitungen definiert sind über $s'(x_0) = s'(x_0+)$, $s'(x_n) = s'(x_n-)$, $s'(x_j) = \frac{1}{2}(s'(x_j-) + s'(x_j+))$ für $j = 1, \ldots, n-1$.

Beweis: siehe [23].

Falls f nur stetig auf $[a, b]$ ist, dann kann man die gleichmäßige Konvergenz des linearen interpolierenden Splines noch durch die gleichmäßige Stetigkeit von f nachweisen.

Satz 1.13 *Für eine Folge $(\Delta_n)_{n\in\mathbb{N}}$ von Zerlegungen mit $\rho(\Delta_n) \to 0$ konvergieren die linearen interpolierenden Splines s gleichmäßig gegen $f \in C[a, b]$.*

Beweis:

Sei $\varepsilon > 0$. Da f gleichmäßig stetig ist gibt es ein $\delta > 0$, so dass

$$|x - x'| < \delta \quad \Rightarrow \quad |f(x) - f(x')| < \varepsilon$$

für alle $x, x' \in [a, b]$. Es gibt ein n_0, so dass $\rho(\Delta_n) < \delta$ für alle $n \geq n_0$. Für festes $x \in [a, b]$ existiert jeweils ein von n abhängiges $j \in \{1, \ldots, n\}$ mit $x \in [x_{j-1}, x_j]$. Dadurch gilt

$$s(x) - f(x) = \underbrace{\tfrac{x_j - x}{x_j - x_{j-1}}}_{=:\eta_j} f(x_{j-1}) + \underbrace{\tfrac{x - x_{j-1}}{x_j - x_{j-1}}}_{=:\xi_j} f(x_j) - f(x).$$

Es gilt stets $\eta_j + \xi_j = 1$ und $\eta_j, \xi_j \geq 0$. Daher können wir abschätzen

$$\begin{aligned} |s(x) - f(x)| &= |\eta_j f(x_{j-1}) + \xi_j f(x_j) - (\eta_j + \xi_j) f(x)| \\ &\leq \eta_j \, |f(x_{j-1}) - f(x)| + \xi_j \, |f(x_j) - f(x)| \\ &\leq \eta_j \varepsilon + \xi_j \varepsilon \;=\; \varepsilon \end{aligned}$$

wegen $|x - x_j| < \delta$ und $|x - x_{j-1}| < \delta$. Somit ist die gleichmäßige Konvergenz gezeigt. □

Die Approximation mit linearen Splines ist in der Praxis jedoch meistens ungeeignet, das diese Funktionen nicht stetig differenzierbar sind.

Konstruktion kubischer Splines

Wir verwenden im folgenden stets kubische Splines, d.h. $k = 3$. Dadurch folgt die stückweise Darstellung

$$s(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad \text{für } x_i \leq x \leq x_{i+1}. \tag{1.6}$$

Damit gilt für $i = 0, 1, \ldots, n-1$

$$s(x_i) = a_i, \quad s'(x_i) = b_i, \quad s''(x_i) = 2c_i, \quad s'''(x_i) = 6d_i.$$

Die $4n$ Koeffizienten a_i, b_i, c_i, d_i stellen Freiheitsgrade dar. Die Glattheitsbedingung $s \in C^2[x_0, x_n]$ ergibt jedoch bereits $3(n-1)$ Bedingungen.

Seien Stützpunkte (x_j, y_j) für $j = 0, 1, \ldots, n$ gegeben. Ein kubischer interpolierender Spline s erfüllt die Eigenschaft

$$s(x_j) = y_j \quad \text{für } j = 0, 1, \ldots, n. \tag{1.7}$$

Somit werden $n + 1$ Bedingungen gestellt. Mit dieser Forderung ist s noch nicht eindeutig bestimmt. Wir stellen die natürlichen Randbedingungen

$$s''(x_0) = s''(x_n) = 0, \tag{1.8}$$

wodurch insgesamt $4n$ Bedingungen vorliegen. Der kubische interpolierende Spline ist damit eindeutig festgelegt wie wir nachher zeigen.

Die Krümmung einer Funktion $f \in C^2$ wird definiert als

$$\kappa(x) = \frac{f''(x)}{\sqrt{1 + f'(x)^2}^3} \approx f''(x)$$

für kleine $|f'(x)|$. Wir definieren die Gesamtkrümmung

$$J(f) = \|f''\|_{L^2}^2 = \int_{x_0}^{x_n} (f''(x))^2 \, \mathrm{d}x. \tag{1.9}$$

Für den kubischen Spline $s \in C^2[x_0, x_n]$ mit der Eigenschaft (1.7) und den Randbedingungen (1.8) folgt $J(s) \leq J(f)$ für alle $f \in C^2[x_0, x_n]$ mit gleicher Interpolationseigenschaft, siehe [20]. Somit minimiert der kubische interpolierende Spline mit natürlichen Randbedingungen die Gesamtkrümmung (1.9).

Nun berechnen wir die Koeffizienten in (1.6). Wir definieren die Schrittweiten $h_i = x_i - x_{i-1}$ und die Werte

$$M_i = s''(x_i) \qquad \text{für } i = 0, 1, \ldots, n-1, n.$$

Die Randbedingungen (1.8) ergeben $M_0 = M_n = 0$. Die Eigenschaft $s^{(4)} \equiv 0$ und die Stetigkeit von s'' führt auf

$$s''(x) = M_i \frac{x_{i+1} - x}{h_{i+1}} + M_{i+1} \frac{x - x_i}{h_{i+1}} \qquad \text{für } x_i \le x \le x_{i+1}. \tag{1.10}$$

Integration liefert

$$\begin{aligned} s'(x) &= -M_i \tfrac{(x_{i+1}-x)^2}{2h_{i+1}} + M_{i+1} \tfrac{(x-x_i)^2}{2h_{i+1}} + B_i \\ s(x) &= M_i \tfrac{(x_{i+1}-x)^3}{6h_{i+1}} + M_{i+1} \tfrac{(x-x_i)^3}{6h_{i+1}} + B_i(x - x_i) + A_i \end{aligned}$$

mit Konstanten A_i, B_i. Die Interpolationsbedingungen (1.7) ergeben

$$M_i \tfrac{h_{i+1}^2}{6} + A_i = y_i, \qquad M_{i+1} \tfrac{h_{i+1}^2}{6} + B_i h_{i+1} + A_i = y_{i+1}$$

für $i = 0, 1, \ldots, n-1$ und stellen gleichzeitig die Stetigkeit von s sicher. Die Integrationskonstanten können nun bestimmt werden

$$A_i = y_i - M_i \tfrac{h_{i+1}^2}{6}, \qquad B_i = \tfrac{y_{i+1}-y_i}{h_{i+1}} - \tfrac{h_{i+1}}{6}(M_{i+1} - M_i).$$

Damit folgen die Konstanten in der Darstellung (1.6) als

$$\begin{aligned} a_i &= y_i \\ b_i &= \tfrac{y_{i+1}-y_i}{h_{i+1}} - \tfrac{h_{i+1}}{6}(2M_i + M_{i+1}) \\ c_i &= \tfrac{1}{2} M_i \\ d_i &= \tfrac{1}{6h_{i+1}}(M_{i+1} - M_i). \end{aligned}$$

Es verbleibt also nur das Problem die Koeffizienten M_i zu bestimmen. Die erste Ableitung des Splines kann dargestellt werden als

$$s'(x) = -M_i \frac{(x_{i+1} - x)^2}{2h_{i+1}} + M_{i+1} \frac{(x - x_i)^2}{2h_{i+1}} + \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{h_{i+1}}{6}(M_{i+1} - M_i)$$

für $x_i \le x \le x_{i+1}$. Als links- und rechtsseitige Grenzwerte folgen

$$\begin{aligned} s'(x_i-) &= \tfrac{y_i - y_{i-1}}{h_i} + \tfrac{h_i}{3} M_i + \tfrac{h_i}{6} M_{i-1}, \\ s'(x_i+) &= \tfrac{y_{i+1}-y_i}{h_{i+1}} - \tfrac{h_{i+1}}{3} M_i - \tfrac{h_{i+1}}{6} M_{i+1}. \end{aligned}$$

Die Stetigkeit von s' erzeugt die Bedingungen

$$h_i M_{i-1} + 2(h_i + h_{i+1}) M_i + h_{i+1} M_{i+1} = 6 \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right)$$

für $i = 1, 2, \ldots, n-1$. Da $M_0 = M_n = 0$ gilt, erhalten wir ein lineares Gleichungssystem für $M_1, \ldots, M_{n-1}$. Die zugehörige Tridiagonalmatrix

$$T = \begin{pmatrix} 2(h_1+h_2) & h_2 & & & & \\ h_2 & 2(h_2+h_3) & h_3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ & & & h_{n-1} & 2(h_{n-1}+h_n) \end{pmatrix} \tag{1.11}$$

ist symmetrische und positiv definit. Zudem ist T strikt diagonaldominant. Daher kann der Gauß-Algorithmus ohne Pivoting eingesetzt werden zur Lösung des Gleichungssystems. Der Rechenaufwand ist dann proportional zu n.

Desweiteren sei erwähnt, dass periodische Splines konstruierbar sind. Dabei werden periodische Stützstellen vorausgesetzt, was lediglich $y_0 = y_n$ bedeutet. Als Randbedingungen hat man dann

$$s'(x_0) = s'(x_n) \qquad \text{und} \qquad s''(x_0) = s''(x_n).$$

Zur Konstruktion bzw. Berechnung dieses Splines siehe [20].

Konvergenzaussagen bei kubischen Splines

Für hinreichend glatten Approximanden gilt das folgende Resultat.

Satz 1.14 *Sei $f \in C^4[a,b]$ und $s \in \mathcal{S}_3(\Delta)$ ein interpolierender kubischer Spline. Zudem bezeichne*

$$h_{\max} = \max\{h_1, \ldots, h_n\} \quad \textit{und} \quad h_{\min} = \min\{h_1, \ldots, h_n\}.$$

Falls

$$\max_{j=0,\ldots,n} |s''(x_j) - f''(x_j)| \le C \|f^{(4)}\|_\infty h_{\max}^2$$

mit einer Konstanten $C > 0$ erfüllt ist, dann folgt mit $c = \frac{h_{\max}}{h_{\min}}(C + \frac{1}{4})$ die Abschätzungen

$$\begin{aligned} \|s - f\|_\infty &\le c\|f^{(4)}\|_\infty h_{\max}^4, \\ \|s' - f'\|_\infty &\le 2c\|f^{(4)}\|_\infty h_{\max}^3, \\ \|s'' - f''\|_\infty &\le 2c\|f^{(4)}\|_\infty h_{\max}^2, \\ \|s''' - f'''\|_\infty &\le 2c\|f^{(4)}\|_\infty h_{\max}, \end{aligned}$$

wobei für das Resultat bezüglich s''' die Auswertungen in den Knoten nicht betrachtet werden.

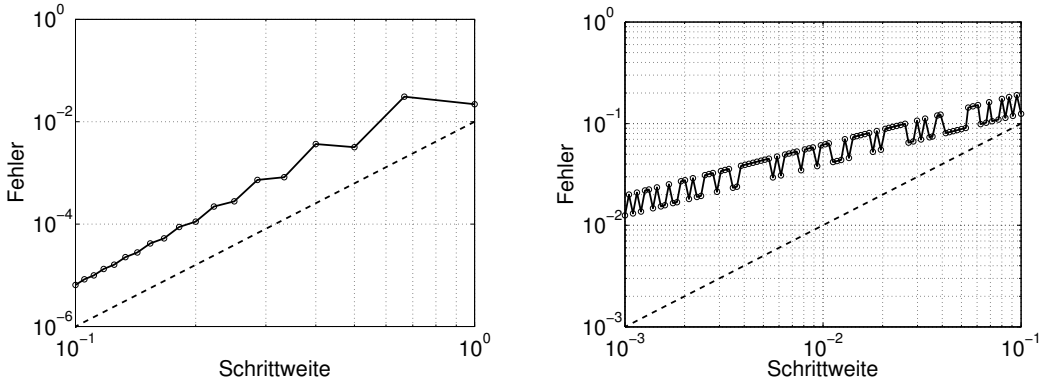


Abbildung 2: Approximationsfehler in Maximumnorm für kubischen interpolierenden Spline bei $f(x) = \frac{1}{1+x^2}$ in $[-5, 5]$ zusammen mit Gerade für H^4 (links) und bei $f(x) = \sqrt{|x|}$ in $[-1, 1]$ zusammen mit Gerade für H (rechts) in doppeltlogarithmischer Skala.

Beweis: siehe [16].

Für den kubischen interpolierenden Spline mit den natürlichen Randbedingungen (1.8) kann eine Konstante $C = \frac{3}{4}$ nachgewiesen werden falls $f''(a) = f''(b) = 0$ gilt. Diese Randwerte können erzeugt werden, indem zur ursprünglichen Funktion f ein Polynom (höchstens) dritten Grades addiert wird.

Im Fall von äquidistanten Stützstellen gilt $H = h_{\max} = h_{\min}$ und es folgt $c = 1$. Wir erhalten

$$\|s - f\|_\infty \leq \|f^{(4)}\|_\infty H^4,$$

d.h. eine Konvergenzgeschwindigkeit $\mathcal{O}(\frac{1}{n^4})$. Im Gegensatz zur Polynominterpolation ist die Wahl äquidistanter Knoten bei der Splineinterpolation günstig.

Abbildung 2 zeigt den Approximationsfehler in der Maximumnorm für äquidistante Stützstellen beim Beispiel von Runge, wo die Funktion beliebig oft differenzierbar ist, und bei einer Wurzelfunktion, die an einer Stelle nicht differenzierbar ist. Entsprechend erkennen wir bei der hinreichend glatten Funktion eine Konvergenz von vierter Ordnung, obwohl die Voraussetzung $f''(a) = f''(b) = 0$ nicht erfüllt ist. Bei der nicht-glatten Funktion liegt die Konvergenz auch vor, jedoch mit einer Geschwindigkeit langsamer als lineare Konvergenz.

Desweiteren wird in Abbildung 3 die Konvergenzgeschwindigkeit des kubischen interpolierenden Splines bei äquidistanten Knoten verglichen mit der Polynominterpolation bei Tschebycheff-Knoten. Beim Beispiel von Runge ist die Funktion beliebig oft differenzierbar, wodurch die totale Variation aller Ableitungen

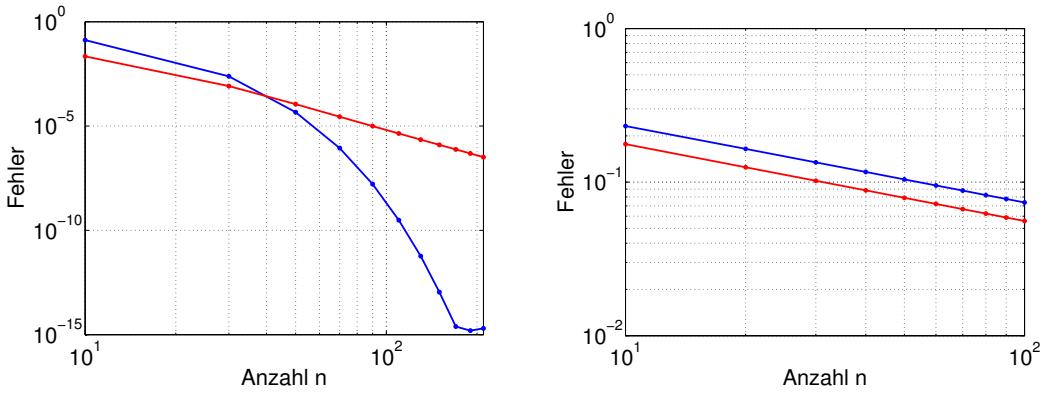


Abbildung 3: Approximationsfehler bei $f(x) = \frac{1}{1+x^2}$ in $[-5, 5]$ (links) und bei $f(x) = \sqrt{|x|}$ in $[-1, 1]$ (rechts) für kubischen interpolierenden Spline (rot) und für Polynominterpolation mit Tschebycheff-Knoten (blau) in doppelt-logarithmischer Skala.

existiert. Nach Satz 1.6 wird die Konvergenzgeschwindigkeit bei Tschebycheff-Knoten beliebig schnell. Dies ist auch erkennbar bis zu einem Fehler in der Größenordnung der Maschinengenauigkeit ($\varepsilon_0 \approx 10^{-16}$). Bei der Wurzelfunktion ist für die Tschebycheff-Knoten die Voraussetzung für die Konvergenz aus Satz 1.5 nicht erfüllt. Trotzdem beobachten wir eine Konvergenz mit der gleichen Geschwindigkeit wie für den Spline.

Es verbleibt die Frage, ob bei nur stetiger Funktion f auf $[a, b]$ der kubisch interpolierende Spline gleichmäßig gegen f konvergiert falls $\rho(\Delta_n) \to 0$. Hierzu gibt es nur Resultate für den periodischen Spline.

Satz 1.15 *Sei $f \in C[0, 1]$ und $f(0) = f(1)$. Zu einer Folge von Zerlegungen $(\Delta_n)_{n\in\mathbb{N}}$ des Intervalls $[0, 1]$ mit $\rho(\Delta_n) \to 0$ definiert man $K_n = \frac{h_{\max}}{h_{\min}}$. Falls die Menge aller K_n beschränkt bleibt, dann konvergiert die Folge der kubischen interpolierenden periodischen Splines gleichmäßig gegen f.*

Beweis: siehe [19].

Offensichtlich ist die Voraussetzung für den Satz 1.15 für äquidistante Stützstellen gegeben, da dann $K_n = 1$ für alle n gilt. Die Beschränktheit der K_n bedeutet, dass die Gitterfolge nicht entartet, d.h. $h_{\min}$ kann nicht beliebig klein im Vergleich zu $h_{\max}$ werden.

Konvergenzaussagen bei Splines allgemeinen Grades

Wir betrachten jetzt Splines von einem beliebigen Grad $2m-1$ für ein $m \geq 2$. Eine zu approximierende Funktion $f \in C^m[a,b]$ sei gegeben. Es gilt $\dim(\mathcal{S}_{2m-1}(\Delta)) = n + 2m - 1$. Dann gibt es bei interpolierenden Splines $2m-2$ Freiheitsgrade, die mit einem der folgenden Randbedingungstypen festgelegt werden.

(i) *natürliche Randbedingungen*

$$s^{(j)}(a) = s^{(j)}(b) = 0 \qquad \text{für } j = m, \ldots, 2m-2.$$

(ii) *vollständige Randbedingungen*

$$s^{(j)}(a) = f^{(j)}(a) \quad \text{und} \quad s^{(j)}(b) = f^{(j)}(b) \qquad \text{für } j = 1, \ldots, m-1.$$

(iii) *periodische Randbedingungen*

$$s^{(j)}(a) = s^{(j)}(b) \qquad \text{für } j = 1, \ldots, 2m-2$$

unter der Voraussetzung $f^{(j)}(a) = f^{(j)}(b)$ für $j = 0, \ldots, m-1$.

Man kann zeigen, dass diese Bedingungen jeweils einen interpolierenden Spline eindeutig festlegen. Für die Konvergenz der Approximation gilt dann das folgende Resultat.

Satz 1.16 *Sei $f \in C^m[a,b]$ und $s \in \mathcal{S}_{2m-1}(\Delta)$ der interpolierende Spline eines der Typen (i),(ii),(iii). Dann gilt die Abschätzuung*

$$\|f^{(j)} - s^{(j)}\|_\infty \leq \frac{m!}{\sqrt{m}} \, \frac{1}{j!} \, \|f^{(m)}\|_{L_2} \, \rho(\Delta)^{m-j-\frac{1}{2}}$$

für beliebiges $m \geq 2$ und $j = 0, 1, \ldots, m-1$.

Beweis: siehe [11], Kapitel 6, Abschnitt 5.4.

Man beachte, dass hier die L_2-Integralnorm von $f^{(m)}$ auftritt, welche sich aber durch die Maximumnorm abschätzen läßt über

$$\|g\|_{L_2} = \sqrt{\int_a^b g(x)^2 \, \mathrm{d}x} \leq \sqrt{b-a} \, \|g\|_\infty.$$

Im Fall $m = 2$ der kubischen Splines folgen die nächsten beiden Abschätzungen mit $h_{\max} = \rho(\Delta)$

$$\begin{aligned} \|f - s\|_\infty &\leq \sqrt{2} \, \|f''\|_{L_2} \, h_{\max}^{\frac{3}{2}}, \\ \|f' - s'\|_\infty &\leq \sqrt{2} \, \|f''\|_{L_2} \, h_{\max}^{\frac{1}{2}}, \end{aligned}$$

Die gleichmäßige Konvergenz ist somit für die Funktion und ihre erste Ableitung garantiert. Im Gegensatz zu den Resultaten aus dem vorhergehenden Abschnitt brauchen hier keine Forderungen an die minimale Schrittweite $h_{\min}$ gestellt zu werden.

Die Konvergenzraten aus Satz 1.16 sind nicht optimal als Preis dafür, dass die Aussagen für beliebigen Grad $2m-1$ gelten. Daher lassen sich für konkrete Gerade oft bessere Konvergenzgeschwindigkeiten ohne höhere Glattheitsforderungen finden. Optimale Schranken und kleinstmögliche Konstanten für kubische interpolierende Splines mit vollständigen Randbedingungen unter der Annahme $f \in C^4$ finden sich in [10].

1.4 Ausgleichsspline

Die folgende Vorgehensweise folgt im wesentlichen der Beschreibung in [17]. Seien die Stützpunkte (x_j, y_j) für $j = 0, 1, \ldots, n$ gegeben mit $x_0 < x_1 < \cdots < x_n$. Dabei kann n groß sein. Falls die Werte y_j (und/oder x_j) fehlerbehaftet sind beispielsweise durch Messfehler, dann ist eine Interpolation der Stützpunkte nicht sinnvoll. Daher approximieren wir die Stützpunkte mit einer Funktion $f \in C^2$. Abbildung 4 verdeutlicht diesen Ansatz.

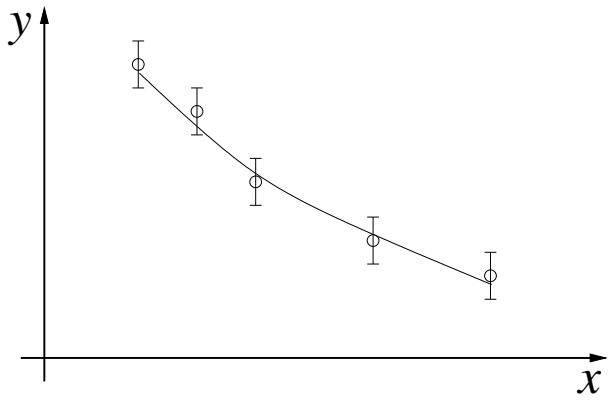


Abbildung 4: Interpolation von Daten mit Messfehlern.

Minimierung mit Nebenbedingung

Da wir Oszillationen in der Approximation f vermeiden möchten, stellen wir die Minimierungsaufgabe

$$\min_{f \in C^2[x_0, x_n]} J(f) \qquad \text{mit} \quad J(f) = \int_{x_0}^{x_n} (f''(x))^2 \, \mathrm{d}x \tag{1.12}$$

unter der Nebenbedingung

$$\sum_{i=0}^{n} \left(\frac{f(x_i) - y_i}{w_i} \right)^2 \leq S \tag{1.13}$$

mit Gewichten $w_i > 0$ und dem Glättungsparameter $S \geq 0$. Die Gewichte w_i werden in Abhängigkeit von der Größe der Messfehler gewählt. Ein sinnvoller Bereich für den Glättungsparameter ist $S \in [N - \sqrt{2N}, N + \sqrt{2N}]$ mit $N = n+1$.

Für $S = 0$ folgt $f(x_i) = y_i$ für alle i und die optimale Funktion f ist gerade der interpolierende kubische Spline mit natürlichen Randbedingungen. Ohne die Nebenbedingung (1.13) ergibt sich das Minimum der Aufgabe (1.12) aus $f'' \equiv 0$ ($J(f) = 0$), d.h. eine Gerade $f(x) = \alpha x + \beta$. Diesen Fall kann man als $S \to \infty$ interpretieren. Die bestapproximierende Gerade bezüglich der Stützpunkte und der Gewichte ist die Ausgleichsgerade. Die Ausgleichsgerade ist Lösung des linearen Ausgleichsproblems

$$\min_{z \in \mathbb{R}^2} \|D^{-1}(Az - y)\|_2^2 \tag{1.14}$$

mit

$$D = \begin{pmatrix} w_0 & & & \\ & w_1 & & \\ & & \ddots & \\ & & & w_n \end{pmatrix}, \quad A = \begin{pmatrix} x_0 & 1 \\ x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \quad z = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Sei $f^*(x) = \alpha^* x + \beta^*$ die eindeutige Lösung des Ausgleichsproblems (1.14). Wir erhalten für das Residuum

$$S_0 = \sum_{i=0}^{n} \left(\frac{\alpha^* x_i + \beta^* - y_i}{w_i} \right)^2.$$

Für $S \in [0, S_0]$ existiert eine eindeutige Lösung der Minimierungsaufgabe (1.12) mit Nebenbedingung (1.13). Für $S > S_0$ ist die Lösung nicht eindeutig und wir wählen die Ausgleichsgerade f^* in diesem Fall. Man beachte im folgenden, dass jede Gerade auch einen kubischen Spline mit natürlichen Randbedingungen darstellt.

Die Ungleichung (1.13) kann in eine Gleichung umgeformt werden durch Einführung einer Variablen $z \in \mathbb{R}$

$$\sum_{i=0}^{n} \left(\frac{f(x_i) - y_i}{w_i} \right)^2 + z^2 - S = 0.$$

Wir koppeln diese Bedingung an die Minimierung (1.12) über einen Lagrange-Parameter p. Es folgt das Funktional

$$\tilde{J}(f,p,z) = \int_{x_0}^{x_n} (f''(x))^2 \,\mathrm{d}x + 2p \left(\sum_{i=0}^{n} \left(\frac{f(x_i)-y_i}{w_i} \right)^2 + z^2 - S \right).$$

Eine Variationsrechnung kann nun durchgeführt werden basierend auf diesem Funktional, siehe [9].

Sei s die Lösung der Aufgabe (1.12),(1.13). Wir betrachten ein stückweise kubisches Polynom

$$s(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3 \quad \text{für } x \in [x_i, x_{i+1}] \tag{1.15}$$

mit $i = 0, 1, \ldots, n-1$. An den Stützstellen sind die links- und rechtsseitigen Grenzwerte dann

$$s^{(k)}(x_i+) = \lim_{h\to 0, h>0} s^{(k)}(x_i+h), \qquad s^{(k)}(x_i-) = \lim_{h\to 0, h>0} s^{(k)}(x_i-h).$$

Wegen $s \in C^2$ erhalten wir die Bedingungen

$$s(x_i+) = s(x_i-), \quad s'(x_i+) = s'(x_i-), \quad s''(x_i+) = s''(x_i-) \tag{1.16}$$

für $i = 1, \ldots, n-1$. Wir fordern die Randbedingungen

$$s''(x_0) = s''(x_n) = 0. \tag{1.17}$$

Desweiteren wird die Sprungbedingung

$$s'''(x_i-) - s'''(x_i+) = 2p\,\frac{s(x_i)-y_i}{w_i^2} \qquad \text{für } i = 0, 1, \ldots, n \tag{1.18}$$

mit $s'''(x_0-) = s'''(x_n+) = 0$ gestellt. Wenn der Lagrange-Parameter p gegeben ist, dann liefern (1.16),(1.17),(1.18) hier $4n$ Gleichungen für die $4n$ unbekannten Koeffizienten in (1.15).

Berechnung der Koeffizienten

Mit den Schrittweiten $h_i = x_{i+1} - x_i$ erhalten wir

$$\begin{aligned} s(x) &= a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3, \\ s'(x) &= b_i + 2c_i(x-x_i) + 3d_i(x-x_i)^2, \\ s''(x) &= 2c_i + 6d_i(x-x_i), \\ s'''(x) &= 6d_i \end{aligned}$$

für $x \in [x_i, x_{i+1}]$ und $i = 0, 1, \ldots, n-1$. Es folgt:

- Die dritte Bedingung aus (1.16) und die Randbedingungen (1.17) liefern

$$2c_i + 6d_i h_i = 2c_{i+1} \qquad \Rightarrow \qquad d_i = \frac{c_{i+1} - c_i}{3h_i}$$

für $i = 0, 1, \ldots, n-1$ mit $c_0 = c_n = 0$.

- Die erste Bedingung aus (1.16) führt auf

$$a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = a_{i+1} \qquad \Rightarrow \qquad b_i = \frac{a_{i+1} - a_i}{h_i} - c_i h_i - d_i h_i^2$$

für $i = 0, 1, \ldots, n-1$.

- Die zweite Bedingung aus (1.16) ergibt

$$b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1}.$$

Mit den Beziehungen zwischen b_i und d_i erreichen wir

$$\frac{h_{i-1}}{3} c_{i-1} + \frac{2}{3}(h_{i-1} + h_i) c_i + \frac{h_i}{3} c_{i+1} = \frac{1}{h_{i-1}} a_{i-1} - \left(\frac{1}{h_{i-1}} + \frac{1}{h_i} \right) a_i + \frac{1}{h_i} a_{i+1}$$

für $i = 1, \ldots, n-1$.

Wir definieren die Vektoren $c = (c_1, \ldots, c_{n-1})^\top$ und $a = (a_0, \ldots, a_n)^\top$ sowie die Matrizen $T \in \mathbb{R}^{(n-1)\times(n-1)}$ und $Q \in \mathbb{R}^{(n+1)\times(n-1)}$

$$T = \begin{pmatrix} t_{11} & t_{12} & & 0 \\ t_{21} & \ddots & \ddots & \\ & \ddots & \ddots & t_{n-2,n-1} \\ 0 & & t_{n-1,n-2} & t_{n-1,n-1} \end{pmatrix}, \qquad Q = \begin{pmatrix} q_{11} & & 0 \\ q_{21} & \ddots & \\ q_{31} & \ddots & q_{n-1,n-1} \\ & \ddots & q_{n,n-1} \\ 0 & & q_{n+1,n-1} \end{pmatrix}$$

mit $t_{ii} = \frac{2}{3}(h_{i-1} + h_i)$, $t_{i+1,i} = t_{i,i+1} = \frac{1}{3} h_i$, $q_{ii} = \frac{1}{h_{i-1}}$, $q_{i+1,i} = -\frac{1}{h_{i-1}} - \frac{1}{h_i}$, $q_{i+2,i} = \frac{1}{h_i}$. Die Tridiagonalmatrix T ist (bis auf ein Vielfaches) identisch mit der Matrix (1.11) aus der kubischen Splineinterpolation. Es folgt das Gleichungssystem

$$Tc = Q^\top a. \tag{1.19}$$

Daher koppelt T die Koeffizienten c_i aus der zweiten Ableitung mit einem Differenzenschema aus den Werten a_i.

Die Sprungbedingung (1.18) ergibt direkt

$$6(d_{i-1} - d_i) = 2p \frac{a_i - y_i}{w_i^2} \tag{1.20}$$

für $i = 0, \ldots, n$ mit $d_{-1} = d_n = 0$. Ersetzen der d_i durch die c_i liefert

$$Qc = pD^{-2}(y - a).$$

Durch die Gleichungen

$$Q^\top(D^2Qc) = Q^\top(p(y-a)) = pQ^\top y - pQ^\top a = pQ^\top y - pTc$$

kann die Berechnung von a und c aus (1.19),(1.20) entkoppelt werden zu

$$(Q^\top D^2 Q + pT)c = pQ^\top y, \qquad p \neq 0: \quad a = y - \frac{1}{p}D^2Qc. \tag{1.21}$$

Die Matrix $Q^\top D^2 Q + pT$ ist symmetrisch und positiv definit für $p \geq 0$ wegen

$$x^\top(Q^\top D^2 Q + pT)x = \underbrace{\|DQx\|_2^2}_{>0} + p\underbrace{x^\top T x}_{>0} > 0 \qquad \text{für jedes } x \neq 0.$$

Man beachte, dass T positiv definit ist und Q vollen Rang besitzt. Desweiteren ist $Q^\top D^2 Q + pT$ eine Bandmatrix mit Breite 5. Für gegebenen Parameter p können wir sukzessive c_i, a_i, d_i, b_i berechnen.

Bestimmung des Lagrange-Parameters

Die Aufgabe ist somit auf die Identifizierung des Lagrange-Parameters p zurückgeführt. Die Funktion s erfüllt die Bedingung

$$\sum_{i=0}^{n} \left(\frac{s(x_i) - y_i}{w_i}\right)^2 + z^2 - S = 0.$$

Die Summe besitzt die Darstellung

$$\sum_{i=0}^{n} \left(\frac{s(x_i) - y_i}{w_i}\right)^2 = \|D^{-1}(y-a)\|_2^2.$$

Mit

$$D^{-1}(y-a) = \tfrac{1}{p}DQc = DQ(Q^\top D^2 Q + pT)^{-1}Q^\top y,$$

was auch für $p = 0$ gilt wegen der Stetigkeit, folgt

$$F(p)^2 = S - z^2$$

mit

$$F(p) = \|DQ(Q^\top D^2 Q + pT)^{-1}Q^\top y\|_2. \tag{1.22}$$

Es kann gezeigt werden, dass $F(p)^2$ eine streng monoton fallende und konvexe Funktion für $p \geq 0$ ist. Also ist auch $F(p)$ injektiv.

Deshalb erhalten wir eine Verbindung zwischen p und z. Wir fordern

$$p \geq 0 \qquad \text{und} \qquad pz = 0. \tag{1.23}$$

Somit treten zwei Fälle auf:

- 1. Fall: $p = 0$
  Die Gleichung (1.21) liefert direkt $c = 0$. Es folgt $d_i = 0$ sowie $b_i = b_{i+1}$ und $a_{i+1} = a_i + h_i b_i$ für $i = 0, 1, \ldots, n-1$. Somit ist die Funktion $s \in C^2$ eine Gerade. Es gilt $J(s) = 0$ für (1.9). Im Unterfall $S = S_0$ ist dann die Ausgleichsgerade die eindeutige Lösung des Minimierungsproblems mit Nebenbedingung. Für jede andere Gerade ist das Residuum des Ausgleichsproblems größer als S_0. In (1.14) gilt $Az = a$. Mit $F(p)^2 = \|D^{-1}(y-a)\|_2^2$ folgt $F(0)^2 = S_0$. Im Unterfall $S > S_0$ ist die Lösung des Minimierungsproblems mit Nebenbedingung nicht eindeutig. Dann wähle die Ausgleichsgerade.

- 2. Fall: $p > 0$
  Bedingung (1.23) impliziert $z = 0$. Der Lagrange-Parameter ist Lösung der nichtlinearen Gleichung
  $$F(p)^2 - S = 0.$$
  Da F^2 eine konvexe Funktion ist, liegt im Newton-Verfahren sogar globale Konvergenz vor. Man kann als Startwert beispielsweise $p^{(0)} = 0$ setzen.

  Es folgt
  $$(Q^\top D^2 Q + pT)^{-1} \quad \overset{p \gg 1}{\approx} \quad \tfrac{1}{p} T^{-1} \quad \overset{p \to \infty}{\longrightarrow} \quad 0 (\in \mathbb{R}^{(n-1)\times(n-1)})$$
  und damit
  $$\lim_{p \to \infty} F(p) = 0 (\in \mathbb{R}).$$
  Dementsprechend besitzt die nichtlineare Gleichung eine Lösung für jeden Glättungsparameter $0 < S < S_0$. Die Funktion $F(p)$ ist in Abbildung 5 dargestellt.

  Die Newton-Iteration für die äquivalente Gleichung $F(p) - \sqrt{S} = 0$ lautet
  $$p^{(j+1)} = p^{(j)} - \frac{F(p^{(j)}) - \sqrt{S}}{F'(p^{(j)})} = p^{(j)} - \frac{F(p^{(j)})^2 - F(p^{(j)})\sqrt{S}}{F(p^{(j)})F'(p^{(j)})}. \qquad (1.24)$$
  Man kann zeigen, dass
  $$F(p)F'(p) = pu^\top T(Q^\top D^2 Q + pT)^{-1} Tu - u^\top Tu$$
  mit $u = p^{-1}c = (Q^\top D^2 Q + pT)^{-1} Q^\top y$. Es folgt $F(p) = \|DQu(p)\|$.

Wir erkennen, dass entweder s zur Ausgleichsgeraden aus (1.14) resultiert ($p = 0$) oder die Ungleichung (1.13) wird zur Gleichung ($p > 0$, $z = 0$).

Wir zeigen noch nachträglich die Monotonie der Funktion (1.22).

Lemma 1.2 *In der Funktion F aus (1.22) sei $Q^\top y \neq 0$. Dann ist F^2 streng monoton fallend und konvex für alle $p \geq 0$.*

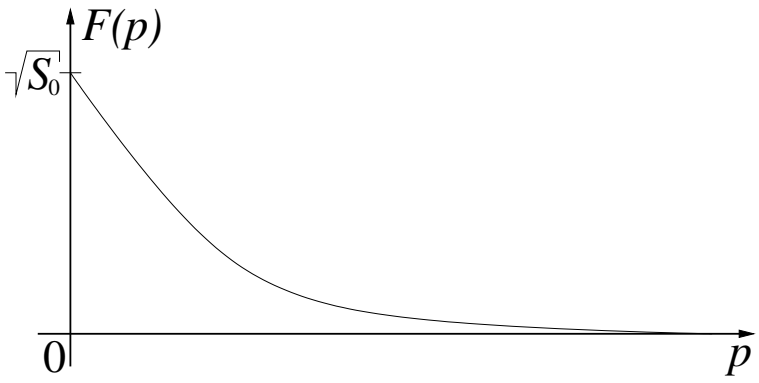


Abbildung 5: Funktion $F(p)$ in Abhängigkeit vom Lagrange-Parameter p.

Beweis:

Die Matrix T ist symmetrisch und positiv definit. Dadurch gilt $T = U^\top \hat{D} U$ mit Diagonalmatrix $\hat{D} = \mathrm{diag}(\hat{d}_1, \ldots, \hat{d}_{n-1})$ und orthogonaler Matrix U. Für

$$T^{\frac{1}{2}} = U^\top \hat{D}^{\frac{1}{2}} U \qquad \text{mit} \qquad \hat{D}^{\frac{1}{2}} = \mathrm{diag}\left(\sqrt{\hat{d}_1}, \ldots, \sqrt{\hat{d}_{n-1}}\right)$$

gilt dann $T^{\frac{1}{2}} T^{\frac{1}{2}} = T$. Es ist auch $T^{\frac{1}{2}}$ symmetrisch und positiv definit. Dadurch erhalten wir

$$Q^\top D^2 Q + pT = T^{\frac{1}{2}} (T^{-\frac{1}{2}} Q^\top D^2 Q T^{-\frac{1}{2}} + pI) T^{\frac{1}{2}}$$

sowie

$$(Q^\top D^2 Q + pT)^{-1} = T^{-\frac{1}{2}} (T^{-\frac{1}{2}} Q^\top D^2 Q T^{-\frac{1}{2}} + pI)^{-1} T^{-\frac{1}{2}}.$$

Es bezeichne $S = T^{-\frac{1}{2}} Q^\top D^2 Q T^{-\frac{1}{2}}$. Die Matrix S ist symmetrisch und positiv definit, da $Q^\top D^2 Q$ positiv definit ist und $T^{-\frac{1}{2}}$ symmetrisch sowie regulär ist. Es seien $\lambda_1, \ldots, \lambda_{n-1} > 0$ die Eigenwerte von S und $v_1, \ldots, v_{n-1}$ eine zugehörige Orthonormalbasis aus Eigenvektoren.

Wir berechnen

$$\begin{aligned} F(p)^2 &= (DQT^{-\frac{1}{2}}(S+pI)^{-1}T^{-\frac{1}{2}}Q^\top y)^\top (DQT^{-\frac{1}{2}}(S+pI)^{-1}T^{-\frac{1}{2}}Q^\top y) \\ &= y^\top Q T^{-\frac{1}{2}}(S+pI)^{-1}T^{-\frac{1}{2}}Q^\top D^2 Q T^{-\frac{1}{2}}(S+pI)^{-1}T^{-\frac{1}{2}}Q^\top y \\ &= z^\top (S+pI)^{-1} S (S+pI)^{-1} z \end{aligned}$$

mit $z = T^{-\frac{1}{2}} Q^\top y$. Da $T^{-\frac{1}{2}}$ regulär ist, gilt nach Voraussetzung $z \neq 0$.

In der Basisdarstellung der Eigenvektoren ist $z = \alpha_1 v_1 + \cdots + \alpha_{n-1} v_{n-1}$ mit $\alpha_j \neq 0$ für mindestens ein j. Die Matrix $(S+pI)^{-1} S (S+pI)^{-1}$ besitzt die gleichen Eigenvektoren wie S und die zugehörigen Eigenwerte lauten

$$\mu_i = \frac{\lambda_i}{(\lambda_i + p)^2} \quad \text{für } i = 1, \ldots, n-1.$$

Dadurch gilt

$$F(p)^2 = \sum_{i=1}^{n-1} \frac{\alpha_i^2 \lambda_i}{(\lambda_i + p)^2} = \sum_{\alpha_i \neq 0} \frac{\alpha_i^2 \lambda_i}{(\lambda_i + p)^2}.$$

Wir differenzieren

$$\frac{\mathrm{d}}{\mathrm{d}p} F(p)^2 = -2 \sum_{\alpha_i \neq 0} \frac{\alpha_i^2 \lambda_i}{(\lambda_i + p)^3}$$

und

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2} F(p)^2 = 6 \sum_{\alpha_i \neq 0} \frac{\alpha_i^2 \lambda_i}{(\lambda_i + p)^4}$$

Offensichtlich gilt $(F^2)' < 0$ und $(F^2)'' > 0$ für alle $p \geq 0$. □

Algorithmus

Zur Berechnung des Ausgleichssplines wird hier die einzelne nichtlineare Gleichung $F(p) - \sqrt{S} = 0$ mit der Newton-Iteration gelöst. Die Auswertung von (1.24) erfolgt in den Schritten:

1. Berechnung der Cholesky-Zerlegung
$$R^\top R = Q^\top D^2 Q + pT.$$
2. Löse $R^\top R u = Q^\top y$ mit Vorwärtssubstitution $R^\top r = Q^\top y$ und Rückwärtssubstitution $Ru = r$.
3. Berechne $F = DQu$, $\bar{F} = F^\top F$, $f = Tu$, $\bar{f} = u^\top f$.
4. Löse $R^\top v = f$ und bestimme $g = v^\top v$.

Nun wird die Newton-Iteration (1.24) zu

$$p^{(j+1)} = p^{(j)} - \frac{\bar{F} - \sqrt{S\bar{F}}}{p^{(j)} g - \bar{f}}.$$

Die Iteration kann durchgeführt werden bis die Lösung auf Maschinengenauigkeit vorliegt. Eine geeignete Abbruchbedingung lautet $F(p^{(j)})^2 \leq S$.

Dieser Algorithmus ist im Fall $S = 0$ nicht durchführbar, da kein entsprechender Lagrange-Parameter existiert ($p \to +\infty$ für $S \to 0$).

Verifikation der Optimalität

In den vorhergehenden Abschnitten haben wir eine kubische Splinefunktion identifiziert, welche die Bedingung (1.13) erfüllt. Jetzt ist noch zu zeigen, dass diese Funktion die Minimierungsaufgabe löst.

Satz 1.17 *Der Ausgleichsspline s definiert durch die Bedingungen (1.15), (1.16), (1.17), (1.18), (1.23) stellt ein Minimum laut (1.12) unter der Nebenbedingung (1.13) dar, d.h. für alle Funktionen $f \in C^2[x_0, x_n]$ mit*

$$\sum_{i=0}^{n} \left(\frac{f(x_i) - y_i}{w_i} \right)^2 \leq S, \tag{1.25}$$

folgt

$$\int_{x_0}^{x_n} f''(x)^2 \, \mathrm{d}x \geq \int_{x_0}^{x_n} s''(x)^2 \, \mathrm{d}x.$$

Beweis:

Wir erhalten

$$\begin{aligned} \int_{x_0}^{x_n} f''(x)^2 \, \mathrm{d}x &= \int_{x_0}^{x_n} (f''(x) - s''(x))^2 + 2(f''(x) - s''(x))s''(x) + s''(x)^2 \, \mathrm{d}x \\ &\geq 2\int_{x_0}^{x_n} (f''(x) - s''(x))s''(x) \, \mathrm{d}x + \int_{x_0}^{x_n} s''(x)^2 \, \mathrm{d}x. \end{aligned}$$

Nachzuweisen ist somit

$$\int_{x_0}^{x_n} (f''(x) - s''(x))s''(x) \, \mathrm{d}x \geq 0 \tag{1.26}$$

für alle $f \in C^2[x_0, x_n]$ welche (1.25) erfüllen. Partielle Integration liefert

$$\begin{aligned} & \int_{x_0}^{x_n} (f''(x) - s''(x))s''(x) \, \mathrm{d}x \\ = & \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (f''(x) - s''(x))s''(x) \, \mathrm{d}x \\ = & \sum_{i=0}^{n-1} [(f' - s')s'']_{x_i}^{x_{i+1}} - [(f - s)s''']_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} (f(x) - s(x))s^{(4)}(x) \, \mathrm{d}x. \end{aligned}$$

Wegen $f, s \in C^2[x_0, x_n]$ und den Randbedingungen (1.17) für s verschwindet der erste Term. Da s ein stückweise kubisches Polynom ist, siehe (1.15), gilt $s^{(4)} \equiv 0$ und der dritte Term verschwindet. Mit $s'''(x_0-) = s'''(x_n+) = 0$ und der

Sprungbedingung (1.18) wird der zweite Term weiter umgeformt

$$\begin{aligned}
&\int_{x_0}^{x_n} (f''(x) - s''(x))s''(x)\,\mathrm{d}x \\
= \; & -\sum_{i=0}^{n} (f(x_i) - s(x_i))(s'''(x_i-) - s'''(x_i+)) \\
= \; & -\sum_{i=0}^{n} (f(x_i) - s(x_i))2p\frac{s(x_i) - y_i}{w_i^2} \\
= \; & -2p\sum_{i=0}^{n} \frac{(f(x_i) - y_i)(s(x_i) - y_i)}{w_i^2} - \frac{(s(x_i) - y_i)^2}{w_i^2} \\
= \; & 2p\left(S - z^2 - \sum_{i=0}^{n} \frac{(f(x_i) - y_i)(s(x_i) - y_i)}{w_i^2}\right).
\end{aligned}$$

Für $p = 0$ ist die Ungleichung (1.26) erfüllt. Für $p > 0$ liefert die Bedingung (1.23) dann $z = 0$. Wir haben zu zeigen, dass

$$\sum_{i=0}^{n} \frac{(f(x_i) - y_i)(s(x_i) - y_i)}{w_i^2} \leq S.$$

Mit der Definition $u = D^{-1}(f - y)$ und $v = D^{-1}(s - y)$ entspricht diese Behauptung $u^\top v \leq S$. Die Nebenbedingung (1.13) führt direkt auf $\|u\|_2^2 \leq S$, $\|v\|_2^2 \leq S$. Mit der Cauchy-Schwarzschen Ungleichung schätzen wir ab

$$|u^\top v| \leq \|u\|_2\|v\|_2 \leq \sqrt{S} \cdot \sqrt{S} = S.$$

Somit ist die Eigenschaft (1.26) gezeigt. □

Beispiel 1:

Wir verwenden eine Menge aus fünf Stützpunkten. Die Gewichte werden auf $w_i = 1$ für alle i gesetzt. Abbildung 6 zeigt die resultierenden Ausgleichssplines für verschiedene Parameter S. Zum einen ergibt sich eine Funktion ähnlich zum kubischen interpolierenden Spline im Fall $S \approx 0$. Zum anderen entsteht die Ausgleichsgerade für hohe Werte S.

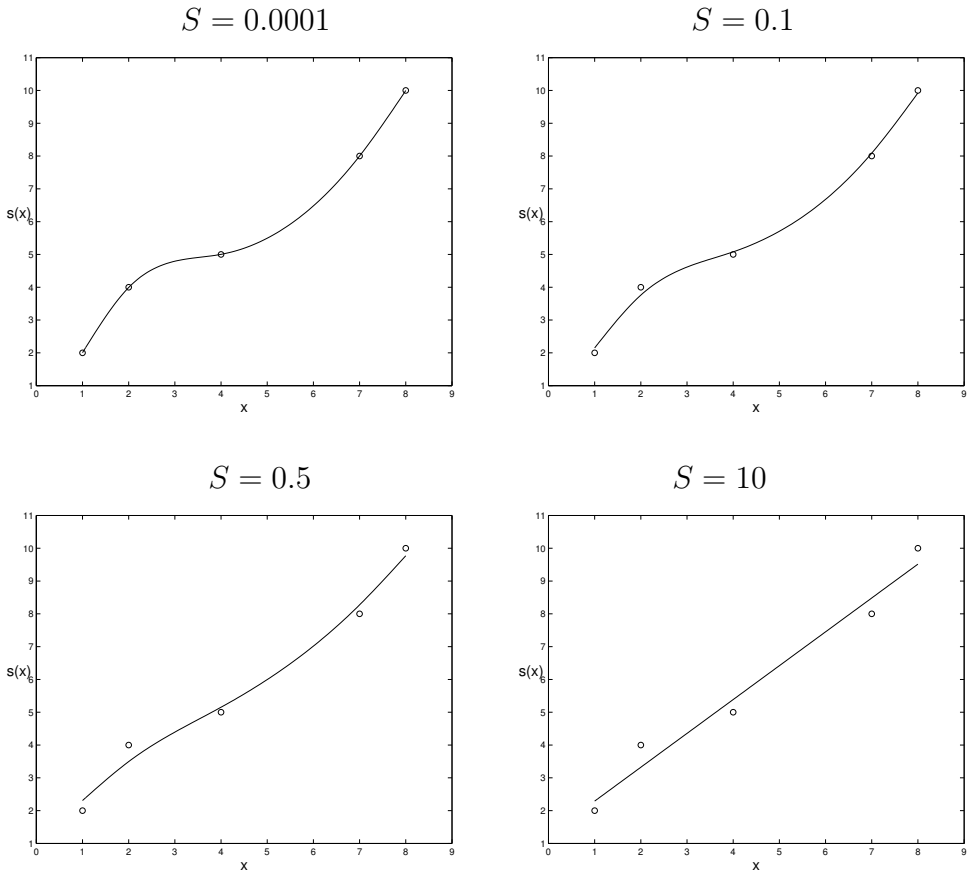


Abbildung 6: Ausgleichsspline für identische Stützpunkte und unterschiedliche Glättungsparameter S.

Beispiel 2:

Sei $x_i = a + ih$ mit $h = \frac{b-a}{n}$ für $i = 0, 1, \ldots, n$. Wir betrachten die Punkte

$$y_i = g(x_i) + r_i \qquad \text{für } \; i = 0, 1, \ldots, n$$

mit der vorgegebenen Funktion

$$g(x) = 1 + \sin(2\pi x).$$

Die Werte r_i stellen Zufallszahlen aus einer Gleichverteilung in $[-0.1, 0.1]$ dar. Der Erwartungswert und die Varianz ergeben sich daher zu

$$\mathbb{E}(r_i) = 0, \quad \text{Var}(r_i) = \sigma^2 = \tfrac{0.2^2}{12} = \tfrac{1}{300} \quad \text{für } \; i = 0, 1, \ldots, n.$$

Alternativ können in der Praxis auch paarweise verschiedene Varianzen auftreten.

Zu den Daten (x_i, y_i) kann man versuchen eine Funktion der Form

$$f(x) = \alpha + \beta \sin(\omega x + \varphi)$$

mit a priori unbekannten Parametern $\alpha, \beta, \omega, \varphi \in \mathbb{R}$ anzupassen. Falls nur α und β unbekannt sind, dann folgt ein lineares Ausgleichsproblem. Jedoch ergibt sich ein nichtlineares Gleichungssystem falls alle Parameter als unbekannt angenommen werden. Zudem erfordert die Auswahl des Ansatzes für f eine Untersuchung der Gestalt bzw. Struktur der Daten (x_i, y_i).

Im Gegensatz hierzu wenden wir den Ansatz des Ausgleichssplines an, welcher unabhängig von der Gestalt der Eingabedaten ist. Wir setzen $n = 100$, $a = 0$, $b = 1.5$ und wählen die Gewichte $w_i = \sigma$ für alle i. Abbildung 7 stellt die entstehenden Ausgleichssplines für verschiedene Parameter S dar. Die Wahl $S = n+1$ liefert eine gute Approximation der zugrunde liegenden Funktion g. Viel größere Werte S führen auf eine schlechte Approximation sowohl der Daten als auch der Funktion g, da zu stark geglättet wird. Viel kleinere Werte S erzeugen unerwünschte Oszillationen im Spline, wobei die Stützpunkte zwar besser approximiert werden, jedoch die Funktion g schlechter erfasst wird.

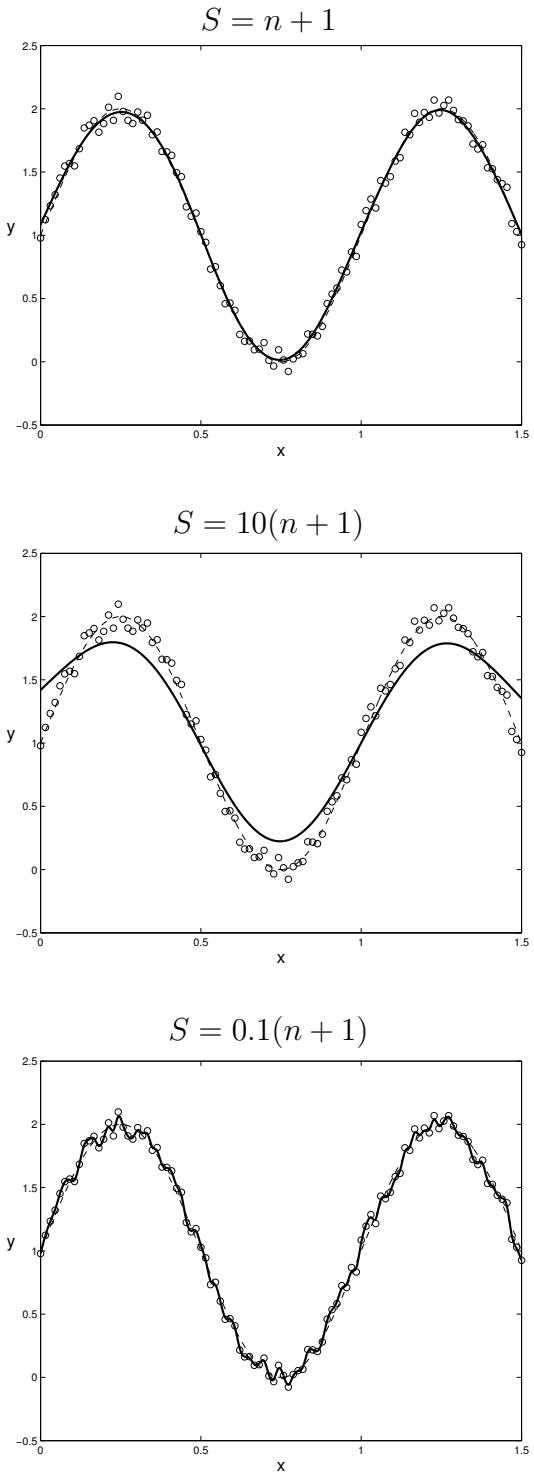


Abbildung 7: Ausgleichsspline (—) für identische Stützpunkte (∘) und verschiedene Parameter S zusammen mit der ursprünglichen Funktion g (- - -).

2 Approximation in normierten Räumen

In diesem Kapitel betrachten wir eine allgemeine Approximationstheorie in normierten Vektorräumen. Als günstisten Fall erweisen sich dabei Hilbert-Räume. Die Ansätze zur Approximation werden konkret bei Fourier-Reihen und Wavelets eingesetzt.

2.1 Allgemeine Approximationstheorie

Zunächst wiederholen wir einige Begriffe aus der Topologie.

Definitionen aus der Topologie

Wir betrachten einen reellen normierten Vektorraum $(V, \|\cdot\|)$, welcher eine beliebige Dimension haben kann. Die Norm induziert eine Metrik d über

$$d(v_1, v_2) := \|v_1 - v_2\| \quad \text{für } v_1, v_2 \in V.$$

Zu einem $u \in V$ und $\varepsilon > 0$ definieren wir die Kugel

$$B_\varepsilon(u) := \{v \in V \ : \ \|v - u\| < \varepsilon\}.$$

Für ein Folge $(v_i)_{i\in\mathbb{N}} \subset V$ schreiben wir

$$\hat{v} = \lim_{i\to\infty} v_i \qquad \text{falls} \qquad \lim_{i\to\infty} \|v_i - \hat{v}\| = 0.$$

Damit folgen die nachstehenden Begriffe.

Def. 2.1 *Sei $(V, \|\cdot\|)$ ein normierter Vektorraum.*

(i) Eine Menge $M \subseteq V$ heißt offen, wenn

$$\forall v \in M \ \exists \varepsilon > 0 \ : \ B_\varepsilon(v) \subset M.$$

(ii) Eine Menge $M \subseteq V$ heißt abgeschlossen, wenn $V \backslash M$ offen ist.

(iii) Eine Menge $M \subseteq V$ heißt beschränkt, wenn ein $\varepsilon > 0$ existiert mit $M \subset B_\varepsilon(0)$.

(iv) Eine Menge $M \subseteq V$ heißt vollständig, wenn jede Cauchy-Folge aus M einen Grenzwert in M besitzt.

Ist der ganze Raum V vollständig, dann nennt man $(V, \|\cdot\|)$ einen Banach-Raum. Von besonderer Bedeutung sind auch kompakte Mengen, welche dadurch gekennzeichnet sind, dass jede offene Überdeckung eine endliche Teilüberdeckung besitzt.

Def. 2.2 *Eine Menge $K \subset V$ heißt kompakt, wenn aus*

$$K \subset \bigcup_{i \in I} O_i$$

mit offenen Mengen O_i die Existenz endlich vieler Indizes $i_1, i_2, \ldots, i_m$ mit

$$K \subset O_{i_1} \cup O_{i_2} \cup \cdots \cup O_{i_m}$$

folgt.

Eine kompakte Menge ist stets beschränkt und abgeschlossen. Die Umkehrung gilt meistens nicht. Jedoch ist im $\mathbb{R}^n$ jede beschränkte und abgeschlossene Menge auch kompakt nach dem Satz von Heine-Borel. Eine wichtige Eigenschaft kompakter Mengen ist der Satz von Bolzano-Weierstraß.

Satz 2.1 *Sei $K \subset V$ eine kompakte Teilmenge. Dann besitzt jede Folge aus K eine konvergente Teilfolge mit Grenzwert in K.*

Beweis: siehe [7], Satz 9.

Man beachte, dass diese Aussage auch ohne die Vollständigkeit von V gilt.

Satz 2.2 *Ein beliebiger Vektorraum V mit endlicher Dimension n ist isomorph zu $\mathbb{R}^n$.*

Beweis: siehe [6], Abschnitt 2.2.4, Korollar 2.

Aus diesem Satz folgt sofort, dass ein endlichdimensionaler Vektorraum vollständig ist, da $\mathbb{R}^n$ für jedes n vollständig ist. Insbesondere gilt die Aussage für endlichdimensionale Untervektorräume $U \subset V$ eines möglicherweise unendlichdimensionalen Vektorraums V.

Satz 2.3 *Ein endlichdimensionaler Untervektorraum $U \subseteq V$ eines normierten Vektorraums $(V, \|\cdot\|)$ ist abgeschlossen.*

Beweis:

Wir zeigen, dass $V\backslash U$ offen ist. Sei $v \in V\backslash U$. Angenommen, es gelte für kein $\varepsilon > 0$ die Inklusion $B_\varepsilon(v) \subset V\backslash U$. Dann enthielte jede Kugel um v Elemente aus U und somit gäbe es eine Folge aus U, die gegen v konvergiert. Da nach Satz 2.2 U isomorph zu einem $\mathbb{R}^n$ ist, folgt wegen der Vollständigkeit von $\mathbb{R}^n$ dann für den Grenzwert $v \in U$. Durch diesen Widerspruch gibt es ein $\varepsilon > 0$ mit $B_\varepsilon(v) \subset V\backslash U$. □

Besonders günstig erweisen sich Vektorräume V mit einem Skalarprodukt $\langle \cdot, \cdot \rangle$, d.h. einer positiv definiten, symmetrischen Bilinearform. Das Skalarprodukt erzeugt eine Norm durch

$$\|v\| = \sqrt{\langle v, v \rangle} \quad \text{für } v \in V.$$

Somit liegt auch ein normierter Vektorraum vor. Ist $(V, \langle \cdot, \cdot \rangle)$ vollständig, dann liegt ein Hilbert-Raum vor. Ist $(V, \langle \cdot, \cdot \rangle)$ nicht vollständig, so spricht man von einem Prä-Hilbert-Raum.

Strikte Normen

Desweiteren benötigen wir den Konvexitätsbegriff.

Def. 2.3 *Sei $(V, \|\cdot\|)$ ein normierter Vektorraum.*

(i) Eine Menge $M \subseteq V$ heißt konvex, falls gilt

$$v_1, v_2 \in M \quad \Rightarrow \quad \lambda v_1 + (1-\lambda) v_2 \in M$$

für alle $\lambda \in [0, 1]$.

(ii) Eine Menge $M \subseteq V$ heißt streng konvex, falls mit $v_1, v_2 \in M$ und $v_1 \neq v_2$ folgt

$$\forall \lambda \in (0,1) \exists \varepsilon > 0 \ : \ B_\varepsilon(\lambda v_1 + (1-\lambda) v_2) \subset M.$$

Für normierte Räume betrachten wir damit die folgende Verschärfung.

Def. 2.4 *Ein normierter Vektorraum $(V, \|\cdot\|)$ heißt strikt normiert, falls die abgeschlossene Einheitskugel $\{v \in V \ : \ \|v\| \leq 1\}$ streng konvex ist.*

Äquivalent zur strikten Normiertheit ist die Bedingung

$$\|v_1\|, \|v_2\| \leq 1, \ \ v_1 \neq v_2 \quad \Rightarrow \quad \|\lambda v_1 + (1-\lambda) v_2\| < 1.$$

Gilt $\|v_1\| \leq 1$ und $\|v_2\| < 1$ sowie $v_1 \neq v_2$, dann folgt

$$\|\lambda v_1 + (1-\lambda)v_2\| \leq |\lambda| \cdot \|v_1\| + |1-\lambda| \cdot \|v_2\| < |\lambda| + |1-\lambda| = 1$$

für $0 < \lambda < 1$, wodurch die Bedingung aus der strengen Konvexität erfüllt ist. Somit braucht nur der Fall $\|v_1\| = \|v_2\| = 1$ betrachtet zu werden. Notwendig und hinreichend für die strikte Normiertheit ist daher das Kriterium

$$\|v_1\| = \|v_2\| = 1, \quad v_1 \neq v_2 \quad \Rightarrow \quad \|\lambda v_1 + (1-\lambda)v_2\| < 1.$$

Auf dem $\mathbb{R}^n$ haben wir als übliche Normen

$$\|x\|_p = \sqrt[p]{\sum_{j=1}^{n} |x_j|^p} \tag{2.1}$$

für $p \geq 1$ mit der Maximumnorm

$$\|x\|_\infty = \max_{j=1,\ldots,n} |x_j|$$

als Grenzfall. Die Normen (2.1) sind genau dann strikt, wenn $1 < p < \infty$ gilt.

Der folgende Satz liefert eine Aussage für (Prä-)Hilbert-Räume.

Satz 2.4 *Wird in einem normierten Vektorraum $(V, \|\cdot\|)$ die Norm von einem Skalarprodukt induziert, dann ist der Raum strikt normiert.*

Beweis:

Sei $\|v_1\| = \|v_2\| = 1$ und $v_1 \neq v_2$. Wir müssen

$$\|(1-\lambda)v_1 + \lambda v_2\| < 1 \quad \text{für} \ \ \lambda \in (0,1)$$

zeigen. O.E.d.A. sei $0 < \lambda \leq \frac{1}{2}$. Dieser Fall kann dargestellt werden als

$$(1-\lambda)v_1 + \lambda v_2 = \tfrac{1}{2}\big(\underbrace{v_1}_{=:\tilde{v}_1} + \underbrace{(v_1 + 2\lambda(v_2 - v_1))}_{=:\tilde{v}_2}\big).$$

Es gilt $\|\tilde{v}_1\| = 1$ sowie

$$\|\tilde{v}_2\| = \|(1-2\lambda)v_1 + 2\lambda v_2\| \leq |1-2\lambda| \cdot \|v_1\| + |2\lambda| \cdot \|v_2\| = |1-2\lambda| + |2\lambda| = 1.$$

Gilt $\|\tilde{v}_2\| < 1$ dann ist die obige Bedingung sofort erfüllt. Es genügt somit den Fall $\lambda = \frac{1}{2}$ zu betrachten.

Da die Norm von einem Skalarprodukt erzeugt wird, gilt die Parallelogrammgleichung

$$\|v_1 + v_2\|^2 + \|v_1 - v_2\|^2 = 2\left(\|v_1\|^2 + \|v_2\|^2\right) \qquad \text{für alle } v_1, v_2.$$

Sei $v_1 \neq v_2$ und $\|v_1\| = \|v_2\| = 1$. Die Parallelogrammgleichung liefert

$$\begin{aligned}\|\tfrac{1}{2}v_1 + \tfrac{1}{2}v_2\|^2 = \tfrac{1}{4}\|v_1 + v_2\|^2 &= \tfrac{1}{4}\left(2\|v_1\|^2 + 2\|v_2\|^2 - \|v_1 - v_2\|^2\right) \\ &< \tfrac{1}{2}\|v_1\|^2 + \tfrac{1}{2}\|v_2\|^2 = 1.\end{aligned}$$

Somit folgt $\|\frac{1}{2}v_1 + \frac{1}{2}v_2\| < 1$. □

Bestapproximationen

Die folgende Definition gilt für beliebige Teilmengen.

Def. 2.5 *Sei $(V, \|\cdot\|)$ ein normierter Vektorraum und $M \subset V$ mit $M \neq \emptyset$ sowie $v \in V$ fest gegeben.*

(i) Die reelle Zahl

$$E_M(v) = \inf_{u \in M} \|u - v\|$$

wird als Minimalabstand des Vektors v zur Teilmenge M bezeichnet.

(ii) Eine Folge $(u_i)_{i \in \mathbb{N}} \subset M$ heißt M-Minimalfolge an v, wenn

$$\lim_{i \to \infty} \|u_i - v\| = E_M(v).$$

(iii) Man nennt einen Vektor $\hat{u} \in M$ eine Bestapproximation in M oder auch M-Proximum falls

$$\|\hat{u} - v\| = \inf_{u \in M} \|u - v\|.$$

Der Minimalabstand existiert immer und ist eindeutig. Die Existenz einer Minimalfolge ergibt sich aus der Eigenschaft des Infimums. Eine naheliegende Fragestellung ist nun, für welche Teilmengen die Existenz und Eindeutigkeit einer Bestapproximation garantiert werden kann.

Beispiele:

1. Wir betrachten $(\mathbb{R}^2, \|\cdot\|_2)$ und $M = \{x \in \mathbb{R}^2 \,:\, \|x\|_2 \leq 1\}$. Zu $v \notin M$ gibt es ein eindeutiges M-Proximum, nämlich

$$\hat{u} = \frac{1}{\|v\|_2}\, v.$$

Es folgt $E_M(v) = \|v\|_2 - 1$. Dass $\hat{u}$ ein M-Proximum ist erkennt man wie folgt: Sei $\tilde{u} \in \mathbb{R}^2$ mit $\|\tilde{u} - v\|_2 < \|\hat{u} - v\|_2$. Es folgt

$$\|v\|_2 \leq \|\tilde{u} - v\|_2 + \|\tilde{u}\|_2 < \|\hat{u} - v\|_2 + \|\tilde{u}\|_2 = \|v\|_2 - 1 + \|\tilde{u}\|_2.$$

Daraus folgt sofort $\|\tilde{u}\|_2 > 1$ und somit $\tilde{u} \notin M$.

2. Wir verwenden $(C[0,1], \|\cdot\|_\infty)$ und $M = \{e^{\beta x} \ : \ \beta > 0\}$ sowie $v \equiv \frac{1}{2}$. Es gilt

$$\|e^{\beta x} - \tfrac{1}{2}\|_\infty = e^\beta - \tfrac{1}{2} > \tfrac{1}{2} \quad \text{für alle } \ \beta > 0.$$

Für $\beta \to 0$ sieht man $E_M(v) = \frac{1}{2}$. Somit existiert kein M-Proximum von v. Eine Minimalfolge ist durch jede Funktionenfolge in M mit $\beta \to 0$ gegeben.

Folgendes Lemma liefert ein hinreichendes Kriterium für die Existenz einer Bestapproximation.

Lemma 2.1 *Sei $M \subset V$ mit $M \neq \emptyset$ für einen normierten Raum $(V, \|\cdot\|)$ und $(u_i)_{i\in\mathbb{N}}$ eine M-Minimalfolge an $v \in V$. Besitzt die Folge einen Häufungspunkt $\hat{u} \in M$, dann ist $\hat{u}$ ein M-Proximum an v.*

Beweis:

Ist $\hat{u} \in M$ ein Häufungspunkt der Folge, dann gibt es eine Teilfolge $(u_{i_k})_{k\in\mathbb{N}}$, die gegen $\hat{u}$ konvergiert, d.h.

$$\lim_{k\to\infty} \|u_{i_k} - \hat{u}\| = 0.$$

Wir setzen an

$$\|\hat{u} - v\| \leq \|\hat{u} - u_{i_k}\| + \|u_{i_k} - v\| \quad \text{für alle } \ k$$

und somit

$$\|\hat{u} - v\| \leq \lim_{k\to\infty} \|\hat{u} - u_{i_k}\| + \|u_{i_k} - v\| = 0 + E_M(v) = E_M(v).$$

Wegen der Abschätzung $E_M(v) \leq \|u - v\|$ für alle u folgt die Behauptung. □

Satz 2.5 *Ist $K \subset V$ mit $K \neq \emptyset$ eine kompakte Teilmenge in einem normierten Raum $(V, \|\cdot\|)$, dann existiert für jedes $v \in V$ ein K-Proximum an v.*

Beweis:

Es existiert zu jeder nichtleeren Teilmenge $K \subset V$ eine K-Minimalfolge. Als Folge in einer kompakten Teilmenge existiert nach Satz 2.1 eine konvergente Teilfolge mit Grenzwert in K. Da der Grenzwert ein Häufungspunkt der K-Minimalfolge ist, stellt dieser Grenzwert laut Lemma 2.1 ein K-Proximum dar. □

Lemma 2.2 *Ist $M \subset V$ für einen normierten Vektorraum $(V, \|\cdot\|)$, dann ist jede M-Minimalfolge (an beliebiges $v \in V$) beschränkt.*

Beweis:

Sei $(u_i)_{i\in\mathbb{N}}$ eine M-Minimalfolge für $v \in V$. Dann gibt es ein $n^* \in \mathbb{N}$ mit

$$E_M(v) \leq \|u_i - v\| \leq E_M(v) + 1$$

für alle $i \geq n^*$. Es folgt

$$\|u_i\| \leq \|v\| + \|u_i - v\| \leq E_M(v) + 1 + \|v\| =: K_1$$

für $i \geq n^*$. Mit $K_2 = \max\{\|u_i\| \ : \ i < n^*\}$ erhalten wir

$$\|u_i\| \leq \max\{K_1, K_2\} \quad \text{für alle } \ i \in \mathbb{N}.$$

Damit ist die Folge beschränkt. □

Das nächste Resultat nennt man auch den Fundamentalsatz der Approximationstheorie in normierten Vektorräumen.

Satz 2.6 *Ist $U \subseteq V$ ein endlichdimensionaler Untervektorraum eines normierten Vektorraums $(V, \|\cdot\|)$, dann existiert für jedes gegebene $v \in V$ ein U-Proximum an v.*

Beweis:

Sei $(u_i)_{i\in\mathbb{N}}$ eine U-Minimalfolge für $v \in V$. Nach Lemma 2.2 ist diese Folge beschränkt in V. Somit ist die Folge trivialerweise auch beschränkt in U, d.h. $u_i \in K$ für alle i mit $K = \{u \in U \ : \ \|u\| \leq c\}$ bei hinreichend hohem c. Die Menge K ist als abgeschlossene und beschränkte Teilmenge eines endlichdimensionalen Vektorraums kompakt in U. Satz 2.1 zeigt die Existenz einer konvergenten Teilfolge mit Grenzwert in K. Da der Grenzwert ein Häufungspunkt der U-Minimalfolge ist, stellt dieser Grenzwert laut Lemma 2.1 wieder ein U-Proximum dar. □

Man beachte, dass in diesem Beweis der Satz 2.5 nicht direkt angewendet wurde, denn dafür müsste man zuerst nachweisen, dass K auch kompakt in V und nicht nur in U ist.

Beispiel: Wir betrachten den normierten Raum $(C[a,b], \|\cdot\|_\infty)$. Es bezeichnet $\mathcal{P}_n$ die Menge der Polynome mit Grad höchstens n. Somit ist $\mathcal{P}_n \subset C[a,b]$ ein Unterraum der Dimension $n+1$. Satz 2.6 garantiert die Existenz einer Bestapproximation in $\mathcal{P}_n$ zu beliebigem $f \in C[a,b]$.

Nachdem wir hinreichende Bedingungen für die Existenz einer Bestapproximation gefunden haben wenden wir uns nun der Eindeutigkeitsfrage zu.

Eine strenge Konvexität impliziert die Eindeutigkeit, jedoch nicht die Existenz.

Satz 2.7 *Ist $M \subseteq V$ mit $M \neq \emptyset$ eine streng konvexe Teilmenge eines normierten Vektorraums $(V, \|\cdot\|)$, dann existiert höchstens ein M-Proximum an $v \in V$.*

Beweis:

Angenommen es existiert mindestens ein Proximum. Es seien $\hat{u}_1$ und $\hat{u}_2$ beide M-Proxima an v sowie $\hat{u}_1 \neq \hat{u}_2$. Dann folgt

$$E_M(v) \leq \|v - \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2)\| \leq \tfrac{1}{2}\|v - \hat{u}_1\| + \tfrac{1}{2}\|v - \hat{u}_2\| = \tfrac{1}{2}E_M(v) + \tfrac{1}{2}E_M(v) = E_M(v).$$

Dadurch haben wir $\|v - \frac{1}{2}(\hat{u}_1 + \hat{u}_2)\| = E_M(v)$.

Weil M streng konvex ist gibt es für $\lambda = \frac{1}{2}$ Werte μ mit

$$\tilde{u} = \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2) + \mu(v - \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2)) \in M$$

für alle $|\mu| < \mu_0$ mit $\mu_0 > 0$ hinreichend klein. Sei $0 < \hat{\mu} < 1$ einer dieser Werte. Wir erhalten

$$\|\tilde{u} - v\| = \|\tfrac{1}{2}(1-\hat{\mu})(\hat{u}_1 + \hat{u}_2) - (1-\hat{\mu})v\| = |1-\hat{\mu}| \cdot \|\tfrac{1}{2}(\hat{u}_1 + \hat{u}_2) - v\| = (1-\hat{\mu})E_M(v).$$

Also folgt $\|\tilde{u} - v\| < E_M(v)$ im Widerspruch zu $E_M(v) \leq \|u - v\|$ für beliebiges $u \in M$. □

Die nächste Schlußfolgerung ergibt sich direkt aus Satz 2.5 und Satz 2.7.

Korollar 2.1 *Ist $K \subset V$ mit $K \neq \emptyset$ eine kompakte und steng konvexe Teilmenge eines normierten Vektorraums $(V, \|\cdot\|)$, dann existiert zu jedem $v \in V$ ein eindeutiges K-Proximum an v.*

Nach Satz 2.6 existiert für einen endlichdimensionalen Untervektorraum die Bestapproximation. In strikt normierten Räumen erhalten wir auch die Eindeutigkeit.

Lemma 2.3 *Ein normierter Vektorraum $(V, \|\cdot\|)$ ist genau dann strikt normiert, wenn für beliebige $v, w \in V\backslash\{0\}$ gilt*

$$\|v\| + \|w\| = \|v + w\| \quad \Rightarrow \quad \exists \alpha \in \mathbb{R} \ : \ w = \alpha v.$$

Beweis:

Wir zeigen nur die wichtigere Implikation. Für die Umkehrung siehe [16], Theorem 15.18.

Sei $(V, \|\cdot\|)$ strikt normiert. Wir verwenden $v, w \in V\backslash\{0\}$ mit der Eigenschaft

$$\|v\| + \|w\| = \|v + w\|.$$

O.E.d.A. sei $\|v\| \leq \|w\|$. Es folgt mit der Dreiecksungleichung

$$\left\| \frac{v}{\|v\|} + \frac{w}{\|v\|} \right\| \leq \left\| \frac{v}{\|v\|} + \frac{w}{\|w\|} \right\| + \left\| \frac{w}{\|v\|} - \frac{w}{\|w\|} \right\|$$

und damit

$$\begin{aligned} \left\| \frac{v}{\|v\|} + \frac{w}{\|w\|} \right\| &\geq \left\| \frac{v}{\|v\|} + \frac{w}{\|v\|} \right\| - \left\| \frac{w}{\|v\|} - \frac{w}{\|w\|} \right\| \\ &= \frac{\|v + w\|}{\|v\|} - \left(\frac{1}{\|v\|} - \frac{1}{\|w\|} \right) \|w\| \\ &= \frac{\|v\| + \|w\|}{\|v\|} - \left(\frac{\|w\|}{\|v\|} - \frac{\|w\|}{\|w\|} \right) \ = \ 2. \end{aligned}$$

Also gilt

$$\left\| \tfrac{1}{2} \cdot \frac{v}{\|v\|} + \tfrac{1}{2} \cdot \frac{w}{\|w\|} \right\| \geq 1.$$

Wegen der strikten Normiertheit ist diese Ungleichung nur möglich falls $\frac{v}{\|v\|} = \frac{w}{\|w\|}$. Somit erhalten wir $w = \alpha v$ mit $\alpha = \frac{\|w\|}{\|v\|}$. □

Satz 2.8 *Sei $(V, \|\cdot\|)$ ein strikt normierter Vektorraum und $U \subseteq V$ ein Untervektorraum. Dann existiert höchstens ein U-Proximum zu einem beliebigen $v \in V$.*

Beweis:

Gilt $v \in U$, dann ist das U-Proximum eindeutig v selbst. Sei daher $v \notin U$. Sind $\hat{u}_1$ und $\hat{u}_2$ beide U-Proxima, dann folgt

$$E_U(v) \leq \|v - \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2)\| \leq \tfrac{1}{2}\|v - \hat{u}_1\| + \tfrac{1}{2}\|v - \hat{u}_2\| = \tfrac{1}{2}E_U(v) + \tfrac{1}{2}E_U(v) = E_U(v).$$

Somit gilt

$$\|\tfrac{1}{2}(v-\hat{u}_1)+\tfrac{1}{2}(v-\hat{u}_2)\| = \|v-\tfrac{1}{2}(\hat{u}_1+\hat{u}_2)\| = \|\tfrac{1}{2}(v-\hat{u}_1)\|+\|\tfrac{1}{2}(v-\hat{u}_2)\|.$$

Da der Raum strikt normiert ist können wir Lemma 2.3 anwenden und erhalten ein $\alpha \in \mathbb{R}$, so dass

$$\tfrac{1}{2}(v-\hat{u}_1) = \alpha\tfrac{1}{2}(v-\hat{u}_2) \qquad \text{und damit} \qquad (1-\alpha)v = \hat{u}_1 - \alpha\hat{u}_2.$$

Weil $v \notin U$ angenommen ist, erhalten wir $1-\alpha = 0$, d.h. $\alpha = 1$. Also folgt $\hat{u}_1 = \hat{u}_2$. □

Als Folgerung aus Satz 2.6 und Satz 2.8 erhalten wir die nächste Aussage.

Korollar 2.2 *Ist $U \subseteq V$ ein endlichdimensionaler Untervektorraum eines strikt normierten Vektorraums $(V, \|\cdot\|)$, dann existiert für jedes gegebene $v \in V$ ein eindeutiges U-Proximum an v.*

Falls die Eindeutigkeit nicht gegeben ist, dann liefert folgender Satz eine Information zur Struktur der Bestapproximationen.

Satz 2.9 *Ist $M \subset V$ mit $M \neq \emptyset$ eine konvexe Teilmenge eines normierten Vektorraums $(V, \|\cdot\|)$, dann ist für jedes $v \in V$ die Menge der M-Proxima an v konvex.*

Beweis:

Seien u_1 und u_2 beide M-Proxima an v und $\lambda \in [0,1]$. Die Konvexität garantiert hier $(1-\lambda)u_1 + \lambda u_2 \in M$. Es folgt

$$\begin{aligned} \|(1-\lambda)u_1+\lambda u_2 - v\| &= \|(1-\lambda)(u_1-v)+\lambda(u_2-v)\| \\ &\leq |1-\lambda|\cdot\|u_1-v\| + |\lambda|\cdot\|u_2-v\| \\ &= (1-\lambda)E_M(v) + \lambda E_M(v) \;=\; E_M(v). \end{aligned}$$

Wegen $E_M(v) \leq \|u-v\|$ für alle $u \in M$ folgt $E_M(v) = \|(1-\lambda)u_1+\lambda u_2 - v\|$, d.h. $(1-\lambda)u_1+\lambda u_2$ ist ebenfalls ein M-Proximum an v. □

Gibt es also zwei verschiedene Bestapproximationen an v, so existiert bereits ein Kontinuum aus Bestapproximationen zu v. Satz 2.9 gilt insbesondere für (unendlichdimensionale) Untervektorräume.

Bestapproximation in Hilbert-Räumen

In Vektorräumen mit Skalarprodukt lassen sich die Bestapproximationen zu Untervektorräumen leicht charakterisieren und bestimmen. Laut Satz 2.4 ist ein (Prä-)-Hilbert-Raum strikt normiert und das Proximum existiert jeweils und ist eindeutig.

Def. 2.6 *Ist V ein Vektorraum mit Skalarprodukt $\langle\cdot,\cdot\rangle$ und $M \subseteq V$, dann ist das orthogonale Komplement von M gegeben durch*

$$M^\perp = \{v \in V \ : \ \langle v, u\rangle = 0 \ \textit{für alle} \ u \in M\}.$$

Diese Definition wird insbesondere für Untervektorräume eingesetzt.

Satz 2.10 *Sei $(V, \langle\cdot,\cdot\rangle)$ ein (Prä-)Hilbert-Raum und $U \subseteq V$ ein Untervektorraum. Ein Element $\hat{u} \in U$ ist genau dann ein U-Proximum an v, wenn $\hat{u} - v \in U^\perp$ gilt.*

Beweis:

Sei $\hat{u} \in U$ und $\hat{u} - v \in U^\perp$. Für beliebiges $u \in U$ folgern wir

$$\begin{aligned} \|u - v\|^2 &= \|\hat{u} - v + u - \hat{u}\|^2 \\ &= \|\hat{u} - v\|^2 + 2\underbrace{\langle \hat{u} - v, u - \hat{u}\rangle}_{=0} + \|u - \hat{u}\|^2 \ \geq \ \|\hat{u} - v\|^2. \end{aligned}$$

Somit ist $\hat{u}$ ein U-Proximum an v.

Sei nun $\hat{u} \in U$ und $\hat{u} - v \notin U^\perp$. Dadurch existiert ein $w \in U \backslash \{0\}$ mit $\langle \hat{u} - v, w\rangle \neq 0$. Mit $t \in \mathbb{R}$ erhalten wir

$$\| \underbrace{\hat{u} + tw}_{\in U} - v\|^2 = \|\hat{u} - v\|^2 + 2t\langle \hat{u} - v, w\rangle + t^2\|w\|^2 =: f(t),$$

d.h. ein Polynom zweiten Grades in t. Differentiation zeigt

$$f'(t) = 2\langle \hat{u} - v, w\rangle + 2t\|w\|^2.$$

Ein Minimum von f liegt daher bei $t^* = \frac{\langle \hat{u} - v, w\rangle}{\|w\|^2} \neq 0$ vor. Also ist $\hat{u} + t^* w$ eine bessere Approximation und somit $\hat{u}$ kein U-Proximum an v. $\square$

Der nächste Satz zeigt die Konstruktion der Bestapproximation.

Satz 2.11 *Sei $(V, \langle\cdot,\cdot\rangle)$ ein (Prä-)Hilbert-Raum und $U \subseteq V$ ein endlichdimensionaler Untervektorraum mit gegebener Basis $\{u_1, \ldots, u_n\}$. Es ist $\hat{u} \in U$ mit*

$$\hat{u} = \sum_{k=1}^{n} \alpha_k u_k$$

genau dann das U-Proximum an $v \in V$, wenn die Normalgleichungen

$$\sum_{k=1}^{n} \langle u_k, u_j \rangle \alpha_k = \langle v, u_j \rangle \qquad \textit{für } j = 1, 2, \ldots, n \tag{2.2}$$

erfüllt sind.

Beweis:

Die Normalgleichungen lassen sich äquivalent umschreiben in

$$\left\langle \left(\sum_{k=1}^{n} \alpha_k u_k \right) - v, u_j \right\rangle = 0 \qquad \text{für } j = 1, 2, \ldots, n.$$

Nach Satz 2.10 ist $\hat{u}$ genau dann ein U-Proximum an v, wenn $\hat{u} - v \in U^\perp$. Ist demnach $\hat{u} - v \in U^\perp$, dann gilt $\langle \hat{u} - v, u_j \rangle = 0$ für alle j und die Normalgleichungen sind erfüllt. Seien umgekehrt die Normalgleichungen erfüllt und $\tilde{u} \in U$ beliebig mit

$$\tilde{u} = \sum_{j=1}^{n} \beta_j u_j.$$

Daraus folgt

$$\langle \hat{u} - v, \tilde{u} \rangle = \left\langle \hat{u} - v, \sum_{j=1}^{n} \beta_j u_j \right\rangle = \sum_{j=1}^{n} \beta_j \underbrace{\langle \hat{u} - v, u_j \rangle}_{=0} = 0,$$

wodurch $\hat{u} - v \in U^\perp$ nachgewiesen ist. □

Die Normalgleichungen (2.2) bilden ein lineares Gleichungssystem $Ax = b$ für die unbekannten Koeffizienten $\alpha_1, \ldots, \alpha_n$ der Bestapproximation. Die beteiligte Matrix

$$A = \begin{pmatrix} \langle u_1, u_1 \rangle & \cdots & \langle u_1, u_n \rangle \\ \vdots & & \vdots \\ \langle u_n, u_1 \rangle & \cdots & \langle u_n, u_n \rangle \end{pmatrix}$$

nennt man Gramsche Matrix. Diese Matrix ist offensichtlich symmetrisch. Sie ist auch positiv definit wegen

$$x^\top A x = \sum_{i,j=1}^{n} x_i x_j \langle u_i, u_j \rangle = \left\langle \sum_{i=1}^{n} x_i u_i, \sum_{j=1}^{n} x_j u_j \right\rangle = \left\| \sum_{j=1}^{n} x_j u_j \right\|^2 > 0$$

falls $x \neq 0$. Das lineare Gleichungssystem kann somit durch die Cholesky-Zerlegung gelöst werden.

Besonders günstig sind hier Orthonormalbasen des endlichdimensionalen Teilraums, d.h.

$$\langle u_i, u_j \rangle = \begin{cases} 0 & \text{für } i \neq j, \\ 1 & \text{für } i = j. \end{cases}$$

In diesem Fall reduziert sich die Gramsche Matrix zur Einheitsmatrix und die Koeffizientendarstellung der Bestapproximation wird direkt

$$\alpha_k = \langle v, u_k \rangle \qquad \text{für } k = 1, 2, \ldots, n.$$

Ist eine beliebige Basis von U gegeben, so lässt sich daraus eine Orthonormalbasis mit dem Verfahren von Gram-Schmidt konstruieren, siehe [6] Abschnitt 5.4.9.

Approximationsprinzip von Korovkin

Wir betrachten in diesem Abschnitt den normierten Raum $(C[a,b], \|\cdot\|_\infty)$ und setzen als Abkürzung $I = [a,b]$ für $a < b$. Diskutiert werden lineare Operatoren auf diesem Vektorraum, wobei nur stetig Operatoren zugelassen sind. Die Stetigkeit ist äquivalent zur Beschränktheit des Operators.

Def. 2.7 *Eine Abbildung $L : C(I) \to C(I)$ heißt*

(i) *linear, falls*

$$L(\alpha f + \beta g) = \alpha Lf + \beta Lg$$

für alle $f, g \in C(I)$ und alle $\alpha, \beta \in \mathbb{R}$ gilt.

(ii) *monoton, wenn die Folgerung*

$$f \leq g \quad \Rightarrow \quad Lf \leq Lg \qquad \text{für } f, g \in C(I)$$

gilt, wobei die Ungleichungen punktweise für alle $x \in I$ gelten.

(iii) *positiv, falls*

$$0 \leq f \quad \Rightarrow \quad 0 \leq Lf \qquad \text{für } f \in C(I).$$

(iv) *beschränkt, falls*

$$\sup \{ \|Lf\|_\infty \ : \ f \in C(I),\ \|f\|_\infty \leq 1 \} < \infty.$$

Man kann leicht zeigen, dass für eine lineare Abbildung dann Monotonie und Positivität äquivalent sind.

Beispiele:

1. *Polynominterpolationsoperator*
Zu einer Menge von Stützstellen $a \leq x_0 < x_1 < \cdots < x_n \leq b$ lautet der Operator der Polynominterpolation

$$(P_n f)(x) = \sum_{j=0}^{n} f(x_j) \cdot \prod_{k=0, k\neq j}^{n} \frac{x - x_k}{x_j - x_k}.$$

Dieser Operator ist linear und beschränkt. Jedoch ist der Operator nicht positiv für z.B. alle $n \geq 3$. Man definiere die stetige Funktion $f \geq 0$ durch

$$f(x) = \left| \prod_{j=1}^{n} x - x_j \right|.$$

Das Interpolationspolynom $P_n f$ besitzt dann die Nullstellen $x_1, \ldots, x_n$ und ist von null verschieden wegen $(P_n f)(x_0) > 0$. Es folgt, dass alle Nullstellen einfach sind und daher ein Vorzeichenwechsel stattfinden muss. Also nimmt $P_n f$ auch negative Werte in $[a, b]$ an.

2. *Bernsteinoperator*
Auf $I = [0, 1]$ definiert sich dieser Operator durch

$$(B_n f)(x) = \sum_{i=0}^{n} f(\tfrac{i}{n}) \cdot \binom{n}{i} x^i (1 - x)^{n-i},$$

vergleiche (1.3). Der Operator basiert auf den Bernstein-Polynomen $B_{i,n}$. Für diese gilt $0 \leq B_{i,n}(x) \leq 1$ für alle $x \in I$. Somit ist der Operator linear, beschränkt und positiv.

Um das Prinzip von Korovkin anzuwenden brauchen wir noch folgenden Begriff. Dabei sei $e_1 \in C(I)$ die Funktion, welche konstant eins ist.

Def. 2.8 *Eine Menge $Q \subset C(I)$ mit $Q = \{f_1, \ldots, f_K\}$ und $e_1 \in Q$ heißt Testmenge, wenn es eine Funktion $p \in C(I \times I)$ gibt mit den Eigenschaften:*

(i) Es existieren Funktionen $a_1, \ldots, a_K \in C(I)$ mit

$$p(t, x) = \sum_{k=1}^{K} a_k(t) f_k(x).$$

(ii) $p(t,x) \geq 0$ *für alle* $(t,x) \in I \times I$.

(iii) $p(t,t) = 0$ *für alle* $t \in I$.

Hilfreich sind noch die nächsten beiden Begriffe.

Def. 2.9 *Zu* $g \in C(I \times I)$ *lautet die* Nullstellenmenge

$$Z(g) = \{(t,x) \in I \times I \ : \ g(t,x) = 0\}.$$

Zu $f \in C(I)$ *ergibt sich die zugehörige* Differenzfunktion

$$d_f(t,x) = f(x) - f(t).$$

Offensichtlich gilt $d_f(t,t) = f(t) - f(t) = 0$ und somit

$$(t,t) \in Z(d_f) \qquad \text{für alle } \ t \in I.$$

Der folgende Satz stellt eine Verallgemeinerung eines Approximationssatzes von P.P. Korovkin aus dem Jahre 1953 dar.

Satz 2.12 *Sei* $(L_n)_{n\in\mathbb{N}}$ *mit* $L_n : C(I) \to C(I)$ *eine Folge monotoner linearer Operatoren und sei* Q *eine Testmenge mit zugehöriger Funktion* p. *Gilt*

$$\lim_{n\to\infty} \|L_n f - f\|_\infty = 0 \quad \textit{für alle } \ f \in Q,$$

dann folgt sogar

$$\lim_{n\to\infty} \|L_n f - f\|_\infty = 0 \quad \textit{für alle } \ f \in C(I),$$

die die Bedingung $Z(p) \subseteq Z(d_f)$ *erfüllen.*

Beweis:

(i) Wir zeigen zuerst, dass die Bedingung

$$\lim_{n\to\infty} \max_{t\in I} |(L_n d_f(t,\cdot))(t)| = 0 \tag{2.3}$$

hinreichend ist für

$$\lim_{n\to\infty} \|L_n f - f\|_\infty = 0.$$

Wegen $d_f(t,\cdot) = f - f(t)e_1$ folgt $f = f(t)e_1 + d_f(t,\cdot)$ und weiter

$$f - L_n f = f - f(t)L_n e_1 - L_n d_f(t,\cdot).$$

Für $t \in I$ erhalten wir die gleichmäßige Abschätzung

$$\begin{aligned} |f(t) - (L_n f)(t)| &\leq |f(t)e_1 - f(t)(L_n e_1)(t)| + |(L_n d_f(t,\cdot))(t)| \\ &\leq \|f\|_\infty \|e_1 - L_n e_1\|_\infty + \max_{t\in I} |(L_n d_f(t,\cdot))(t)|. \end{aligned}$$

Da $e_1 \in Q$ gilt $\|e_1 - L_n e_1\|_\infty \to 0$ und die Bedingung (2.3) liefert $\|f - L_n f\|_\infty \to 0$.

(ii) Als zweites weisen wir nach, dass die Bedingung (2.3) für alle $f \in C(I)$ mit der Eigenschaft $Z(p) \subseteq Z(d_f)$ erfüllt ist.

Die Differenzfunktion ist stetig in x und t. Zu jedem $\varepsilon > 0$ gibt es eine offene Umgebung U von $Z(d_f)$ mit

$$|d_f(t,x)| < \varepsilon \qquad \text{für alle } (t,x) \in U.$$

Für die Diagonale gilt

$$\{(t,x) \in I \times I \ : \ t = x\} \subseteq Z(d_f)$$

bei beliebigem f. Die Bedingung $Z(p) \subseteq Z(d_f)$ impliziert

$$p(t,x) > 0 \quad \text{für alle } (t,x) \in U^{\mathrm{C}} = (I \times I) \backslash U.$$

Ist $U^{\mathrm{C}} = \emptyset$, dann folgt $|d_f(t,x)| < \varepsilon$ für alle $(t,x) \in I \times I$. Anderenfalls ist $U^{\mathrm{C}} \neq \emptyset$ abgeschlossen und somit kompakt in $I \times I$. Dadurch existiert das Minimum

$$M = \min_{(t,x)\in U^{\mathrm{C}}} p(t,x) > 0.$$

Wir folgern

$$|d_f(t,x)| \leq \|d_f\|_\infty \leq \|d_f\|_\infty \frac{p(t,x)}{M} \qquad \text{für alle } (t,x) \in U^{\mathrm{C}}$$

und somit

$$|d_f(t,x)| \leq \frac{\|d_f\|_\infty}{M} p(t,x) + \varepsilon \qquad \text{für alle } (t,x) \in I \times I.$$

Anwendung des monotonen Operators L_n bezüglich x bei festem t zeigt

$$|(L_n d_f(t,\cdot))(x)| \leq \frac{\|d_f\|_\infty}{M} (L_n p(t,\cdot))(x) + \varepsilon (L_n e_1)(x)$$

($L_n p(t,\cdot) \geq 0$ wegen $p(t,\cdot) \geq 0$) und mit $x = t$

$$|(L_n d_f(t,\cdot))(t)| \leq \frac{\|d_f\|_\infty}{M} \max_{t\in I} (L_n p(t,\cdot))(t) + \varepsilon \|L_n e_1\|_\infty \tag{2.4}$$

gleichmäßig für alle $t \in I$. Wegen $p(t,t) = 0$ für alle t gilt

$$\sum_{k=1}^{K} a_k(t) f_k(t) = 0 \qquad \text{für alle } \ t \in I.$$

Andererseits ist

$$L_n p(t,\cdot) = L_n \left(\sum_{k=1}^{K} a_k(t) f_k(\cdot) \right) = \sum_{k=1}^{K} a_k(t) (L_n f_k)$$

und damit

$$(L_n p(t,\cdot))(t) = \sum_{k=1}^{K} a_k(t)(L_n f_k)(t) - \sum_{k=1}^{K} a_k(t) f_k(t) = \sum_{k=1}^{K} a_k(t) \left[(L_n f_k)(t) - f_k(t) \right].$$

Die Konvergenz der Folge $(L_n)_{n\in\mathbb{N}}$ auf $\operatorname{span}(Q)$ impliziert

$$\lim_{n\to\infty} \max_{t\in I} (L_n p(t,\cdot))(t) = 0. \tag{2.5}$$

Desweiteren gilt

$$\|L_n e_1\|_\infty \le \|L_n e_1 - e_1\|_\infty + \|e_1\|_\infty = \|L_n e_1 - e_1\|_\infty + 1$$

und mit $e_1 \in Q$ folgt die Konvergenz der Operatoren. Also ist die betrachtete Folge $(\|L_n e_1\|_\infty)_{n\in\mathbb{N}}$ beschränkt. Da $\varepsilon > 0$ beliebig ist, folgt mit (2.4)

$$|(L_n d_f(t,\cdot))(t)| \le \frac{\|d_f\|_\infty}{M} \max_{t\in I} (L_n p(t,\cdot))(t)$$

gleichmäßig für $t \in I$. Mit (2.5) folgt somit (2.3). □

Beispiel: *Bernstein-Operatoren*
Wir folgern mit dem Prinzip von Korovkin nun die gleichmäßige Konvergenz der Bernstein-Polynome B_n aus (1.3) gegen die zu approximierende Funktion f in $I = [0,1]$. Als Testmenge verwenden wir $Q = \{e_1, e_2, e_3\}$ mit

$$e_1 = 1, \qquad e_2 = x, \qquad e_3 = x^2.$$

Wir definieren

$$p(t,x) = (t-x)^2 = t^2 - 2tx + x^2$$

und verifizieren

(i) $p(t,x) = a_1(t)e_1(x) + a_2(t)e_2(x) + a_3(t)e_3(x) = a_1(t) + a_2(t)x + a_3(t)x^2$ mit $a_1(t) = t^2$, $a_2(t) = -2t$, $a_3(t) = 1$,

(ii) $p(t,x) = (t-x)^2 \geq 0$,

(iii) $p(t,t) = (t-t)^2 = 0$.

Desweiteren gilt $p(t,x) > 0$ für $x \neq t$, wodurch $Z(p) \subseteq Z(d_f)$ für beliegiges f folgt.

Man kann zeigen, dass gilt

$$\lim_{n\to\infty} \|L_n e_k - e_k\|_\infty = 0 \quad \text{für } k = 1,2,3.$$

Für e_1 folgt dies direkt aus der Zerlegung der Eins in (1.2) und für e_2, e_3 aus den ersten beiden Formeln im Beweis von Lemma 1.1. Somit sind die Voraussetzungen für Satz 2.12 erfüllt und es folgt die Konvergenz der Approximationen für beliebiges $f \in C(I)$. Dies stellt einen alternativen Beweis von Satz 1.8 dar.

2.2 Fourier-Reihen

In diesem Abschnitt betrachten wir periodische Funktionen mit fester Periode $L > 0$. O.E.d.A. sei $L = 2\pi$, wodurch wir uns auf das kompakte Intervall $[-\pi, +\pi]$ zurückziehen können.

Trigonometrische Polynome

Def. 2.10 *Wir bezeichnen die Menge der stetigen bzw. quadratintegrablen periodischen Funktionen mit Periode 2π durch*

$$\begin{aligned} C_{\mathrm{p}} &= \{f \in C(\mathbb{R}) \,:\, f(x+k2\pi) = f(x) \text{ für alle } k \in \mathbb{Z} \text{ und } x \in \mathbb{R}\}, \\ L^2_{\mathrm{p}} &= \{f : \mathbb{R} \to \mathbb{R} \,:\, f|_{[-\pi,+\pi]} \in L^2([-\pi,+\pi]), \\ &\qquad f(\cdot + k2\pi) = f(\cdot) \text{ für alle } k \in \mathbb{Z}\}. \end{aligned}$$

Ist $f \in C[-\pi,+\pi]$ mit $f(-\pi) = f(\pi)$, dann kann f erweitert werden zu einer Funktion in C_{p}. Jedes $f \in L^2[-\pi,+\pi]$ kann zu einer Funktion in L^2_p fortgesetzt werden. Umgekehrt kann jedes $f \in L^2_p$ auf den Raum $L^2[-\pi,+\pi]$ eingeschränkt werden. Es gilt $C_{\mathrm{p}} \subset L^2_{\mathrm{p}}$.

Def. 2.11 *Ein reelles trigonometrisches Polynom ist eine Funktion der Gestalt*

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \tag{2.6}$$

mit Koeffizienten $a_0, \ldots, a_n \in \mathbb{R}$ und $b_1, \ldots, b_n \in \mathbb{R}$. Es bezeichnet n den Grad des Polynoms ($a_n \neq 0$ oder $b_n \neq 0$). Sei $\mathcal{T}_n$ die Menge aller trigonometrischen Polynome mit Grad höchstens n.

Offensichtlich ist jedes trigonometrische Polynom in C_{p}. Die trigonometrischen Polynome $\mathcal{T}_n$ bilden einen Untervektorraum, d.h. $\mathcal{T}_n \subset C_{\mathrm{p}} \subset L^2_{\mathrm{p}}$.

Auf $L^2[-\pi, +\pi]$ verwenden wir das modifizierte Skalarprodukt

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) g(x) \, \mathrm{d}x \qquad \text{für } \; f, g \in L^2[-\pi, +\pi]. \tag{2.7}$$

Die von diesem Skalarprodukt erzeugte Norm ist äquivalent zur üblichen L^2-Norm und wir bezeichnen sie mit $\| \cdot \|_{L^2}$. Das System

$$\left\{ \tfrac{1}{\sqrt{2}}, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(nx), \cos(nx) \right\}.$$

ist eine Orthonormalbasis von $\mathcal{T}_n$ bezüglich des Skalarprodukts (2.7).

Def. 2.12 *Zu $f \in L^2[-\pi, +\pi]$ lauten die zugehörigen Fourier-Koeffizienten*

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx) \, \mathrm{d}x \qquad \textit{für } \; k = 0, 1, 2, \ldots \, , \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx) \, \mathrm{d}x \qquad \textit{für } \; k = 1, 2, \ldots \, . \end{aligned}$$

Die Fourier-Koeffizienten implizieren die Operatoren S_n mit

$$(S_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx).$$

Die Eigenschaft $f \in L^2[-\pi, +\pi]$ garantiert die Existenz der Fourierkoeffizienten als reelle Zahlen. Es gilt

$$a_k = \langle f(x), \cos(kx) \rangle \quad \text{für } \; k \geq 0, \qquad b_k = \langle f(x), \sin(kx) \rangle \quad \text{für } \; k \geq 1.$$

Speziell für $k = 0$ haben wir $a_0 = \sqrt{2} \langle f(x), \frac{1}{\sqrt{2}} \rangle$ und es ist $\frac{a_0}{2} = \langle f(x), \frac{1}{\sqrt{2}} \rangle \frac{1}{\sqrt{2}}$.

Konvergenz der Fourier-Reihe

Mit Satz 2.11 folgt, dass $S_n f$ die Bestapproximation von f in $\mathcal{T}_n$ bezüglich der Norm $\| \cdot \|_{L^2}$ ist. Daraus ergibt sich die Frage, ob diese Bestapproximationen auch gegen die Funktion konvergieren.

Satz 2.13 *Für $f \in L^2[-\pi, +\pi]$ konvergiert die Fourier-Reihe im quadratischen Mittel gegen f, d.h. es gilt*

$$\lim_{n\to\infty} \|f - S_n f\|_{L^2} = 0.$$

Beweis: siehe [24], Satz V4.9.

Somit konvergiert die Fourier-Reihe im quadratischen Mittel gegen die zu approximierende Funktion. Eine zentrale Frage der Approximationstheorie war, ob für eine stetige Funktion die Fourier-Reihe auch gleichmäßig konvergiert.

Satz 2.14 *Sei $f \in C_{\mathrm{p}}$ eine stückweise stetig differenzierbare Funktion, d.h. es gibt eine Unterteilung $-\pi = x_0 < x_1 < \cdots < x_m = \pi$, so dass $f|_{(x_{j-1},x_j)}$ für $j = 1, \ldots, m$ stetig differenzierbar in $[x_{j-1}, x_j]$ ist. Dann konvergiert die Fourier-Reihe gleichmäßig gegen f.*

Beweis:

Es bezeichne $g_j : [x_{j-1}, x_j] \to \mathbb{R}$ die erste Ableitung von $f|_{(x_{j-1},x_j)}$. Dann sei $g : \mathbb{R} \to \mathbb{R}$ die periodische Funktion, die auf (x_{j-1}, x_j) mit g_j für alle j übereinstimmt. Die Fourier-Koeffizienten von g erfüllen die Besselsche Ungleichung

$$\frac{|a_0(g)|^2}{4} + \sum_{k=1}^{\infty} |a_k(g)|^2 + |b_k(g)|^2 \le \|g\|_{L^2}^2.$$

Für $k \neq 0$ zeigt partielle Integration

$$\int_{x_{j-1}}^{x_j} f(x)\cos(kx)\,\mathrm{d}x = \left[\frac{1}{k} f(x)\sin(kx)\right]_{x=x_{j-1}}^{x=x_j} - \frac{1}{k}\int_{x_{j-1}}^{x_j} g(x)\sin(kx)\,\mathrm{d}x$$

und damit weil sich die Randterme gegenseitig aufheben

$$a_k(f) = \frac{1}{\pi}\sum_{j=1}^{m}\int_{x_{j-1}}^{x_j} f(x)\cos(kx)\,\mathrm{d}x = -\frac{1}{k\pi}\sum_{j=1}^{m}\int_{x_{j-1}}^{x_j} g(x)\sin(kx)\,\mathrm{d}x = -\frac{1}{k} b_k(g).$$

Analog folgt $b_k(f) = \frac{1}{k} a_k(g)$. Man beachte, dass $g \in L^2[-\pi, +\pi]$ ist.

Für alle $v, w \in \mathbb{R}$ gilt $vw \le \frac{1}{2}(v^2 + w^2)$. Wir erhalten

$$|a_k(f)| = \frac{|b_k(g)|}{k} \le \frac{1}{2}\left(\frac{1}{k^2} + |b_k(g)|^2\right),\ |b_k(f)| = \frac{|a_k(g)|}{k} \le \frac{1}{2}\left(\frac{1}{k^2} + |a_k(g)|^2\right).$$

Da die Reihen

$$\sum_{k=1}^{\infty}\frac{1}{k^2}, \qquad \sum_{k=1}^{\infty}|a_k(g)|^2, \qquad \sum_{k=1}^{\infty}|b_k(g)|^2$$

wegen der Besselschen Ungleichung alle konvergent sind, folgt

$$\frac{|a_0(f)|}{2} + \sum_{k=1}^{\infty} |a_k(f)| + |b_k(f)| < \infty.$$

Die Fourier-Reihe von f konvergiert wegen $|\sin(kx)| \leq 1$, $|\cos(kx)| \leq 1$ für alle k damit absolut und gleichmäßig. Es existiert also eine stetige Funktion $\tilde{f}$ als Grenzwert. Aus gleichmäßiger Konvergenz folgt auch die Konvergenz im quadratischen Mittel. Die Fourier-Reihe von f konvergiert jedoch auch gegen f selbst nach Satz 2.13. Da der Grenzwert eindeutig ist folgt $f = \tilde{f}$ in $L^2[-\pi, +\pi]$. Da f und $\tilde{f}$ stetig sind folgt daraus $f \equiv \tilde{f}$. □

In Satz 2.14 wird vorausgesetzt, dass bei Unstetigkeitsstellen der ersten Ableitung die links- und rechtsseitigen Grenzwerte jeweils existieren.

Desweiteren hängt die Konvergenzgeschwindigkeit von der Glattheit der Funktion ab. Für das Abklingverhalten der Fourier-Koeffizienten gibt der folgende Satz eine Auskunft.

Satz 2.15 *Ist $f \in C_{\mathrm{p}}$ eine m-mal stetig differenzierbare Funktion ($m \geq 1$), dann gibt es eine Konstante $C_f > 0$ und ein $n^* \in \mathbb{N}$ mit*

$$\max\{|a_n|, |b_n|\} \leq C_f \cdot \frac{1}{n^m} \qquad \textit{für } n \geq n^*$$

und desweiteren gilt

$$\|f - S_n f\|_\infty \leq K \, \|f^{(m)}\|_\infty \frac{\ln n}{n^m} \qquad \textit{für } n \geq 2$$

mit einer Konstanten $K > 0$ und der Maximumnorm $\|\cdot\|_\infty$.

Der Beweis der ersten Aussage erfolgt analog zum Beweis von Satz 2.14 mit partieller Integration. Zum Beweis der zweiten Aussage siehe [21].

Jedoch ist die Stetigkeit allein nicht hinreichend für die gleichmäßige Konvergenz. Dazu betrachten wir ein Gegenbeispiel. Die Reihe

$$f(x) = \sum_{k=1}^{\infty} \frac{\sin(2^{k^3} x)}{k^2} \qquad \text{für } 0 \leq x \leq \pi$$

konvergiert absolut und gleichmäßig. Der Grenzwert f existiert damit und ist stetig. Es gilt $f(0) = f(\pi) = 0$. Wir erweitern diese Funktion zu $\tilde{f} : [-\pi, \pi] \to \mathbb{R}$ durch $\tilde{f}(x) = f(|x|)$, d.h. $\tilde{f}$ ist eine gerade Funktion. Die Fourier-Reihe von $\tilde{f}$

besteht dann nur aus Cosinus-Termen. Man kann zeigen, dass diese Fourier-Reihe bei $x = 0$ divergiert, siehe [5].

Approximation mit Fejér-Operatoren

Erfreulicherweise ergibt sich die gleichmäßige Konvergenz mit trigonometrischen Polynomen für eine leichte Modifikation der Fourier-Summen. Dabei wird aus den Fourier-Summen ein arithmetisches Mittel konstruiert mittels der Cesàro-Summation.

Def. 2.13 *Zu $f \in L^2[-\pi, +\pi]$ wird mit den Operatoren S_n der Fourier-Summen der Fejér-Operator*

$$F_n f = \frac{1}{n} \sum_{j=0}^{n-1} S_j f$$

für $n \in \mathbb{N}$ gebildet.

Desweiteren existieren für beide Operatoren Kerne.

Def. 2.14 *Der n-te Dirichlet-Kern ist die Funktion*

$$K_n^{\mathrm{D}}(x) = \frac{\sin((2n+1)\frac{x}{2})}{\sin(\frac{x}{2})}$$

und der n-te Fejér-Kern ist die Funktion

$$K_n^{\mathrm{F}}(x) = \frac{1}{n} \left(\frac{\sin(n\frac{x}{2})}{\sin(\frac{x}{2})} \right)^2 .$$

Abbildung 8 zeigt ein Beispiel für diese Funktionen. Jetzt können wir die Operatoren durch Integrale mit den Kernen darstellen. Man beachte, dass diese Repräsentation nur in der Theorie von Interesse ist. Für die numerische Auswertung der Operatoren werden diese Formeln nicht eingesetzt.

Satz 2.16 *Für $f \in C[-\pi, \pi]$ gilt*

$$(S_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) K_n^{\mathrm{D}}(t - x) \, \mathrm{d}t$$

und

$$(F_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) K_n^{\mathrm{F}}(t - x) \, \mathrm{d}t$$

für alle $x \in \mathbb{R}$ und alle $n \geq 0$.

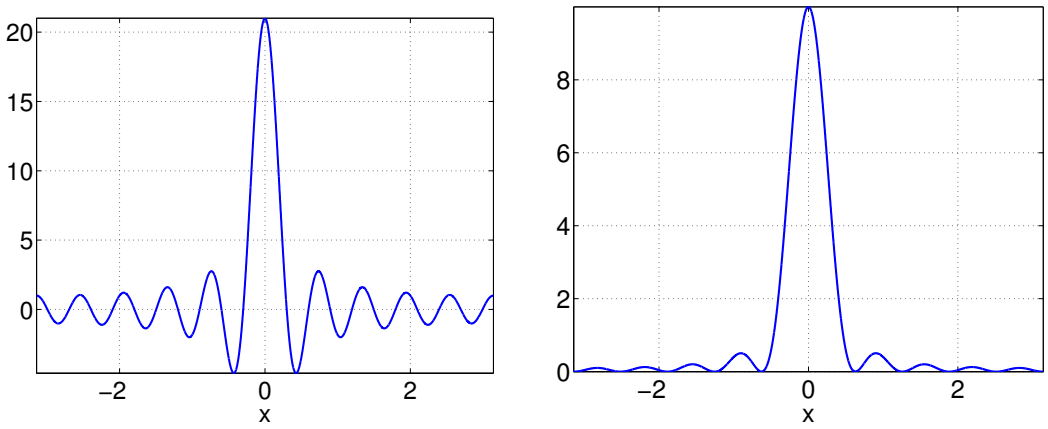


Abbildung 8: Dirichlet-Kern (links) und Fejér-Kern (rechts) für $n = 10$ in $[-\pi, +\pi]$.

Beweis:

Wir erhalten die Hilfsformel

$$\frac{1}{2} + \sum_{k=1}^{n} \cos(kx) = \frac{1}{2} \cdot \frac{\sin((n+\frac{1}{2})x)}{\sin(\frac{x}{2})} \qquad \text{für alle } n$$

aus

$$\begin{aligned}
\frac{1}{2} + \sum_{k=1}^{n} \cos(kx) &= \frac{1}{2} \sum_{k=-n}^{n} \cos(kx) &&= \frac{1}{2} \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}kx} \\
&= \frac{1}{2} \mathrm{e}^{-\mathrm{i}nx} \sum_{k=0}^{2n} \mathrm{e}^{\mathrm{i}kx} &&= \frac{1}{2} \mathrm{e}^{-\mathrm{i}nx} \frac{\mathrm{e}^{\mathrm{i}(2n+1)x} - 1}{\mathrm{e}^{\mathrm{i}x} - 1} \\
&= \frac{1}{2} \frac{\mathrm{e}^{\mathrm{i}(n+\frac{1}{2})x} - \mathrm{e}^{-\mathrm{i}(n+\frac{1}{2})x}}{\mathrm{e}^{\mathrm{i}\frac{x}{2}} - \mathrm{e}^{-\mathrm{i}\frac{x}{2}}} &&= \frac{1}{2} \frac{\sin((n+\frac{1}{2})x)}{\sin(\frac{x}{2})},
\end{aligned}$$

wobei die Formeln für Sinus und Cosinus im Komplexen sowie der Wert der endlichen geometrischen Reihe verwendet wurden.

Mit den Additionstheoremen für trigonometrische Funktionen, der Periodizität

und der obigen Formel folgt

$$\begin{aligned}
(S_n f)(x) &= \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \\
&= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \left[\frac{1}{2} + \sum_{k=1}^{n} \cos(kt)\cos(kx) + \sin(kt)\sin(kx) \right] \mathrm{d}t \\
&= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \left[\frac{1}{2} + \sum_{k=1}^{n} \cos(k(t-x)) \right] \mathrm{d}t \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \, \frac{\sin((n+\frac{1}{2})(t-x))}{\sin(\frac{t-x}{2})} \, \mathrm{d}t.
\end{aligned}$$

Das Additionstheorem $\sin(\alpha)\sin(\beta) = \frac{1}{2}(\cos(\alpha-\beta) - \cos(\alpha+\beta))$ zeigt uns

$$\begin{aligned}
\sum_{j=0}^{n-1} \sin((j+\tfrac{1}{2})x)\sin(\tfrac{x}{2}) &= \frac{1}{2} \sum_{j=0}^{n-1} \cos(jx) - \cos((j+1)x) \\
&= \tfrac{1}{2}\left[1 - \cos(nx)\right] = \sin^2(\tfrac{nx}{2}).
\end{aligned}$$

Aus Def. 2.13 und der obigen Formel für S_n folgt

$$\begin{aligned}
(F_n f)(x) &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \frac{1}{n} \left[\sum_{j=0}^{n-1} \frac{\sin((j+\frac{1}{2})(t-x))}{2\sin(\frac{t-x}{2})} \right] \mathrm{d}t, \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \, \frac{1}{n} \, \frac{\sin^2(\frac{n(t-x)}{2})}{\sin^2(\frac{t-x}{2})} \, \mathrm{d}t,
\end{aligned}$$

wodurch die Aussage gezeigt ist. □

Der Fejér-Operator ist offensichtlich linear. Mit der Darstellung des Operators aus Satz 2.16 folgt, dass der Operator auch positiv ist, denn der Kern ist eine nichtnegative Funktion. Dadurch können wir Satz 2.12 anwenden, wobei wir uns auf den Raum $C[-\pi, +\pi]$ zurückziehen. Wir definieren eine Testmenge Q : $\{e_1, e_2, e_3\} \in C_\mathrm{p}$ mit

$$e_1 = 1, \quad e_2 = \cos(x), \quad e_3 = \sin(x)$$

sowie die Funktion

$$p(t,x) = 1 - \cos(t-x) = 1 - \cos(t)\cos(x) - \sin(t)\sin(x).$$

Wir verifizieren

(i) $p(t,x) = a_1(t)e_1(x) + a_2(t)e_2(x) + a_3(t)e_3(x)$
mit $a_1(t) = 1$, $a_2(t) = -\cos(t)$, $a_3(t) = -\sin(t)$,

(ii) $p(t,x) \geq 0$ wegen $-1 \leq \cos(t-x) \leq 1$,

(iii) $p(t,t) = 1 - \cos(0) = 0$.

Die Nullstellenmenge von p in $[-\pi,+\pi]^2$ ist

$$Z(p) = D \cup \{(-\pi,+\pi),(+\pi,-\pi)\}.$$

mit der Diagonalen $D = \{(t,t) : t \in [-\pi,+\pi]\}$. Für die Differenzfunktion gilt immer $D \subseteq Z(d_f)$. Da f periodisch ist, d.h. $f(-\pi) = f(\pi)$, folgt hier $\{(-\pi,+\pi),(+\pi,-\pi)\} \subset Z(d_f)$. Die Eigenschaft $Z(p) \subseteq Z(d_f)$ ist somit für alle relevanten Funktionen f erfüllt. Es verbleibt zu zeigen, dass die Fejér-Operatoren für jede Funktion aus der Testmenge Q gleichmäßig konvergieren. Dies folgt aus

$$S_n e_1 = 1 \text{ für } n \geq 0, \quad S_n e_2 = \cos(x) \text{ für } n \geq 1, \quad S_n e_3 = \sin(x) \text{ für } n \geq 1,$$

denn wir erhalten

$$F_n e_1 = 1, \quad F_n e_2 = \tfrac{n-1}{n}\cos(x), \quad F_n e_3 = \tfrac{n-1}{n}\sin(x) \quad \text{für } n \geq 1$$

und somit

$$\lim_{n\to\infty} \|e_i - F_n e_i\|_\infty = 0 \qquad \text{für } \ i = 1,2,3.$$

Wir notieren als Schlussfolgerung:

Satz 2.17 *Ist $f \in C_{\mathrm{p}}$, dann gilt mit den Fejér-Operatoren*

$$\lim_{n\to\infty} \|f - F_n f\|_\infty = 0,$$

wobei $\|\cdot\|_\infty$ die Maximumnorm auf $[-\pi,+\pi]$ bezeichnet.

Die nachfolgende direkte Konsequenz daraus nennt man auch den Weierstraßschen Approximationssatz für periodische Funktionen.

Korollar 2.3 *Zu jeder Funktion $f \in C_{\mathrm{p}}$ existiert eine Folge von trigonometrischen Polynomen, welche gleichmäßig gegen f konvergiert.*

2.3 Wavelets

In diesem Abschnitt behandeln wir eine Approximationsmethode für Funktionen in $L^2(\mathbb{R})$. Somit liegen keine Periodizitäten vor.

Motivation

Wir definieren die Funktionenräume

$$L^p(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{C} \ : \ f \text{ messbar}, \ \int_{-\infty}^{\infty} |f(x)|^p \, \mathrm{d}x < \infty \right\}$$

für ganzzahliges $p \geq 1$. Die zugehörigen Normen lauten

$$\|f\|_{L^p} = \sqrt[p]{\int_{-\infty}^{\infty} |f(x)|^p \, \mathrm{d}x}.$$

Nur im Fall $p = 2$ wird die Norm von einem Skalarprodukt (d.h. hermitesche, positiv definite Sesquilinearform) erzeugt, nämlich

$$\langle f, g \rangle_{L^2} = \int_{-\infty}^{\infty} f(x) \, \overline{g(x)} \, \mathrm{d}x \qquad \text{für } \ f, g \in L^2(\mathbb{R}).$$

Eine Teilmenge bilden die reellwertigen Funktionen der Gestalt $f : \mathbb{R} \to \mathbb{R}$.

Eine reellwertige Funktion $f \in L^2(\mathbb{R})$ soll nun approximiert werden. Polynome eignen sich in diesem Raum nicht zur Approximation, denn es gilt $q \notin L^2(\mathbb{R})$ für alle Polynome $q \not\equiv 0$. Ebenso gilt $q \notin L^2(\mathbb{R})$ für jedes trigonometrische Polynom $q \not\equiv 0$. Eine Approximation mit trigonometrischen Funktionen wäre hier zudem nicht sinnvoll, da die Funktion f nicht periodisch vorausgesetzt ist.

Eine naheliegende Idee wäre, für die aperiodische Funktion $f \in L^2(\mathbb{R})$ die kontinuierliche Fourier-Transformation anzuwenden, d.h. eine Integraltransformation. Falls $f \in L^1(\mathbb{R})$, dann lautet die Fourier-Transformierte

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) \, \mathrm{e}^{-\mathrm{i}\omega x} \, \mathrm{d}x \qquad \text{für } \ \omega \in \mathbb{R}.$$

Man beachte, dass die Funktionen $\mathrm{e}^{\mathrm{i}\omega x}$ zwar beschränkt sind, jedoch nicht in $L^2(\mathbb{R})$ liegen. Für $f \in L^2(\mathbb{R})$ kann die Fourier-Transformierte definiert werden durch

$$\hat{f}(\omega) = \lim_{R \to \infty} \frac{1}{\sqrt{2\pi}} \int_{-R}^{+R} f(x) \, \mathrm{e}^{-\mathrm{i}\omega x} \, \mathrm{d}x,$$

wobei der Grenzwert in $L^2(\mathbb{R})$ zu bilden ist. Es gilt die Eigenschaft

$$\langle \hat{f}, \hat{g} \rangle_{L^2} = \langle f, g \rangle_{L^2} \qquad \text{für } \ f, g \in L^2(\mathbb{R}).$$

Insbesondere ist die Fourier-Transformation damit eine Isometrie, d.h.

$$\|\hat{f}\|_{L^2} = \|f\|_{L^2} \qquad \text{für } f \in L^2(\mathbb{R}).$$

Die Transformierte $\hat{f} \in L^2(\mathbb{R})$ ist selbst wieder eine Funktion, welche keine Approximation von f liefert.

Unser Wunsch ist es, ein Orthonormalsystem $(\psi_n)_{n\in\mathbb{N}}$ zu finden, welches vollständig ist, d.h. $\overline{\text{span}\{\psi_n : n \in \mathbb{N}\}} = L^2(\mathbb{R})$. Insbesondere gilt dann

$$f(x) = \sum_{n=1}^{\infty} \langle f, \psi_n \rangle_{L^2} \, \psi_n(x) \qquad \text{für alle } f \in L^2(\mathbb{R}),$$

d.h. f kann in der Norm $\|\cdot\|_{L^2}$ beliebig genau approximiert werden. Zudem soll zu jedem bzw. möglichst vielen f jeweils eine endliche Indexmenge $J(f) \subset \mathbb{N}$ mit wenigen Elementen existieren, wobei die Fehlernorm

$$\Big\| f - \sum_{n\in J(f)} \langle f, \psi_n \rangle_{L^2} \, \psi_n \Big\|_{L^2}$$

relativ klein ist, d.h. f kann mit nur wenigen Basisfunktionen hinreichend genau approximiert werden.

Kontinuierliche Wavelet-Transformation

Die kontinuierlierliche Fourier-Transformation in $L^2(\mathbb{R})$ basiert auf den trigonometrischen Funktionen $v(x,\omega) = \mathrm{e}^{\mathrm{i}\omega x}$, welche keine Basisfunktionen wegen $v(\cdot,\omega) \notin L^2(\mathbb{R})$ für jedes ω. Eine Verschiebung einer Funktion v im Argument x ergibt kein qualitativ unterschiedliches Verhalten, denn es gilt

$$v(x-b,\omega) = \mathrm{e}^{\mathrm{i}\omega(x-b)} = \mathrm{e}^{-\mathrm{i}\omega b}\mathrm{e}^{\mathrm{i}\omega x} =: C(\omega,b)v(x,\omega)$$

und die neue Konstante verschwindet wieder in einer Normierung.

Wir definieren die Wavelet-Transformation wie folgt.

Def. 2.15 *Eine Funktion $\psi : \mathbb{R} \to \mathbb{R}$ mit $\psi \in L^2(\mathbb{R})$ nennt man Wavelet, wenn ihre Fourier-Transformierte $\hat{\psi}(\omega)$ die Bedingung*

$$c_\psi = 2\pi \int_{\mathbb{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} \, \mathrm{d}\omega < \infty. \tag{2.8}$$

erfüllt. Die zugehörige Wavelet-Transformation einer Funktion $f \in L^2(\mathbb{R})$ ist gegeben durch

$$(L_\psi f)(a,b) = \frac{1}{\sqrt{c_\psi}} \, |a|^{-\frac{1}{2}} \int_{\mathbb{R}} f(x) \, \psi\left(\frac{x-b}{a}\right) \, \mathrm{d}x \tag{2.9}$$

für $a \in \mathbb{R}\backslash\{0\}$ und $b \in \mathbb{R}$ mit dem linearen Operator L_ψ.

Erstaunlicherweise liegt die Menge der Wavelets dicht in $L^2(\mathbb{R})$, d.h. es existiert eine Vielzahl von Wavelets. Die Fourier-Transformierte $\hat{\psi}$ ist für eine Funktion $\psi \in L^1(\mathbb{R})$ stetig. Erfüllt somit $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ die Bedingung (2.8), dann folgt

$$\int_{\mathbb{R}} \psi(x)\, \mathrm{d}x = \int_{\mathbb{R}} \psi(x)\, \mathrm{e}^{\mathrm{i}\cdot 0\cdot x}\, \mathrm{d}x = \sqrt{2\pi}\, \hat{\psi}(0) = 0. \tag{2.10}$$

Dadurch verschwindet der Mittelwert eines Wavelets. Besitzt ψ einen kompakten Träger und ist stückweise stetig, dann folgt $\psi \in L^p(\mathbb{R})$ für alle $p \geq 1$.

Beispiele:

Die folgenden Beispiele sind in Abbildung 9 dargestellt.

- Das Haar-Wavelet ist definiert durch

$$\psi(x) = \begin{cases} 1 & \text{für } 0 \leq x < \frac{1}{2} \\ -1 & \text{für } \frac{1}{2} \leq x \leq 1 \\ 0 & \text{sonst.} \end{cases}$$

  Somit ist ψ stückweise konstant und unstetig. Dieses Wavelet besitzt einen kompakten Träger. Es gilt $c_\psi = 2 \ln 2$.

- Der Mexikanische Hut ergibt sich aus

$$\psi(x) = (1 - x^2)\mathrm{e}^{-\frac{1}{2}x^2},$$

  wobei $\psi \in C^\infty$ und $c_\psi = 1$ gilt.

- Das komplex-wertige Morlet-Wavelet besitzt den Realteil

$$\mathrm{Re}\ \psi(x) = \mathrm{e}^{-\frac{1}{2}x^2} \cos(2\pi\nu x)$$

  mit einem reellwertigen Parameter ν. Das Morlet-Wavelet erfüllt nicht die Bedingung (2.8). Jedoch können Wavelets ähnlicher Gestalt konstruiert werden, die (2.8) aufweisen.

Wir definieren eine Modifikation des Funktionenraums $L^2(\mathbb{R}^2)$ durch

$$L^2\Big(\mathbb{R}^2, \frac{\mathrm{d}a\mathrm{d}b}{a^2}\Big) = \Big\{ g : \mathbb{R}^2 \to \mathbb{R} : \int_{\mathbb{R}} \int_{\mathbb{R}\setminus\{0\}} |g(a,b)|^2\, \frac{1}{a^2}\, \mathrm{d}a\mathrm{d}b < \infty \Big\}.$$

Falls $g \in L^2(\mathbb{R}^2, a^{-2}\mathrm{d}a\mathrm{d}b)$ stetig in $\mathbb{R}\setminus\{0\} \times \mathbb{R}$ ist, dann gilt notwendigerweise

$$\lim_{a\to 0} g(a,b) = 0 \quad \text{für jedes } \ b \in \mathbb{R}$$

vorausgesetzt der Grenzwert existiert.

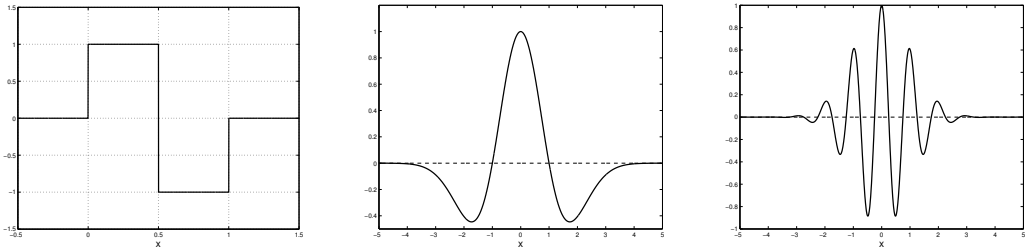


Abbildung 9: Haar-Wavelet, Mexikanischer Hut und Morlet-Wavelet.

Satz 2.18 *Die Bedingung (2.8) garantiert, dass der Operator*

$$L_\psi : L^2(\mathbb{R}) \to L^2\Big(\mathbb{R}^2, \frac{\mathrm{d}a\mathrm{d}b}{a^2}\Big)$$

aus (2.9) eine Isometrie zwischen diesen Hilbert-Räumen darstellt.

Beweis: siehe [14], Satz 1.1.8.

Dadurch ist der Operator L_ψ injektiv. Jedoch ist der Operator nicht surjektiv. Für eine Funktion $g \in \mathrm{Im}(L_\psi)$ im Bild des Operators lautet die inverse Transformation bezüglich (2.9) dann

$$f(x) = \frac{1}{\sqrt{c_\psi}} \int_{\mathbb{R}} \int_{\mathbb{R}\setminus\{0\}} |a|^{-\frac{1}{2}}\ \psi\left(\frac{x-b}{a}\right) g(a,b)\ \frac{1}{a^2}\ \mathrm{d}a\mathrm{d}b. \tag{2.11}$$

Als Beispiel betrachten wir die Funktion

$$f(x) = \begin{cases} 1 & \text{für } -1 \le x < 0 \text{ oder } 1 \le x \le \frac{3}{2} \\ 2+x & \text{für } -2 \le x \le -1 \\ 0 & \text{sonst} \end{cases} \tag{2.12}$$

aus Abbildung 10. Unstetigkeiten dieser Funktion und ihrer ersten Ableitung liegen vor. Abbildung 11 zeigt die kontinuierliche Wavelet-Transformation basierend auf dem Haar-Wavelet und dem Mexikanischen Hut. Wir erkennen eine Glättung der Funktion für ansteigenden Parameter a.

Zum besseren Verständnis dieser Transformation führen wir noch zwei Operatoren ein.

Def. 2.16 *Für gegebene reelle Zahl $a \neq 0$ lautet der Dilatations-Operator*

$$D^a : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad f \mapsto D^a f = |a|^{-\frac{1}{2}} f(\cdot/a).$$

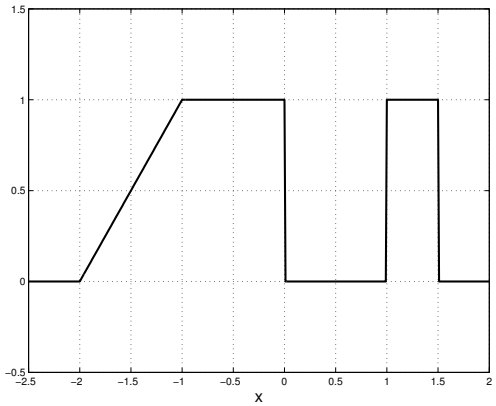


Abbildung 10: Funktion (2.12) in Anwendung der Wavelet-Transformation.

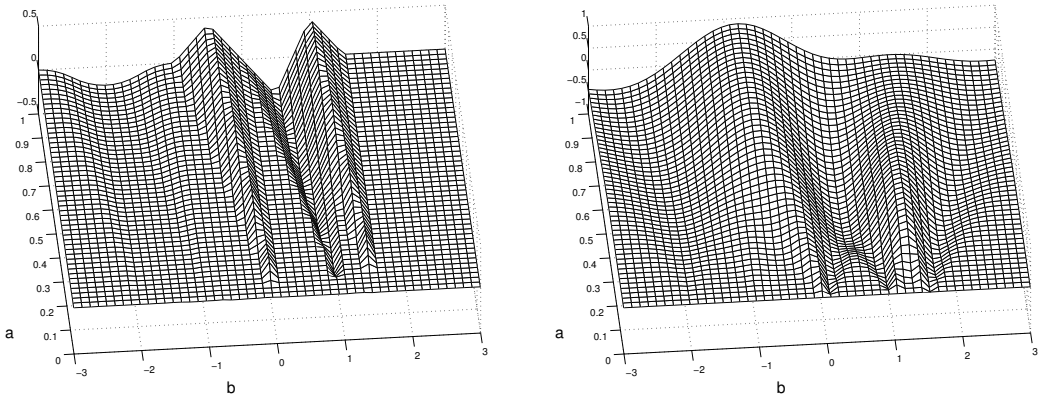


Abbildung 11: Wavelet-Transformation der Funktion (2.12), siehe Abbildung 10, mit dem Haar-Wavelet (links) und dem Mexikanischen Hut (rechts).

Für gegebene reelle Zahl b ist der Translations-Operator

$$T^b : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad f \mapsto T^b f = f(\cdot - b).$$

Beide Operatoren sind Isometrien. Es folgt

$$(L_\psi f)(a,b) = c_\psi^{-\frac{1}{2}} \langle f, T^b D^a \psi \rangle_{L^2}.$$

Die inversen Operatoren sind $(D^a)^{-1} = D^{1/a}$ und $(T^b)^{-1} = T^{-b}$.

Man kann nun die Wavelet-Transformation als Filter interpretieren. Die Filtration von Frequenzen aus einem Signal $f \in L^2(\mathbb{R})$ wird oft über die Faltung mit einer speziellen Funktion $\varphi \in L^2(\mathbb{R})$ durchgeführt, d.h.

$$f_\varphi(x) = (f * \varphi)(x) = \int_{\mathbb{R}} f(s)\varphi(x-s)\,\mathrm{d}s.$$

Die zugehörige Fourier-Transformierte erfüllt die Gleichung

$$\widehat{f_\varphi} = \sqrt{2\pi}\ \hat{f} \cdot \hat{\varphi}.$$

Abhängig vom Verhalten von $\hat{\varphi}$ ergeben sich drei Fälle:

- Tiefpass-Filter: $\hat{\varphi} \approx \chi([-C, C])$,
- Bandpass-Filter: $\hat{\varphi} \approx \chi([-C_2, -C_1] \cup [C_1, C_2])$ $(C_1, C_2 > 0)$,
- Hochpass-Filter: $\hat{\varphi} \approx \chi((-\infty, -C] \cup [C, +\infty))$.

Darin wird die charakteristische Funktion definiert durch $(\chi(U))(x) = 1$ für $x \in U$ und $(\chi(U))(x) = 0$ für $x \notin U$ mit $U \subset \mathbb{R}$ verwendet. Für festes $a \neq 0$ können wir die Wavelet-Transformation (2.9) als Faltung schreiben

$$\begin{aligned}(L_\psi f)(a,b) &= \frac{1}{\sqrt{c_\psi}}\,|a|^{-\frac{1}{2}} \int_{\mathbb{R}} f(x)\ \psi\left(\frac{b-x}{-a}\right)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{c_\psi}}\,|a|^{-\frac{1}{2}} \left(f * \psi\left(\frac{\cdot}{-a}\right)\right)(b).\end{aligned}$$

Wir bestimmen die Fourier-Transformierte von $\psi(x/a)$

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \psi\left(\frac{x}{a}\right) \mathrm{e}^{-\mathrm{i}\omega x}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \psi(s)\mathrm{e}^{-\mathrm{i}\omega a s} a\,\mathrm{d}s = a\hat{\psi}(a\omega).$$

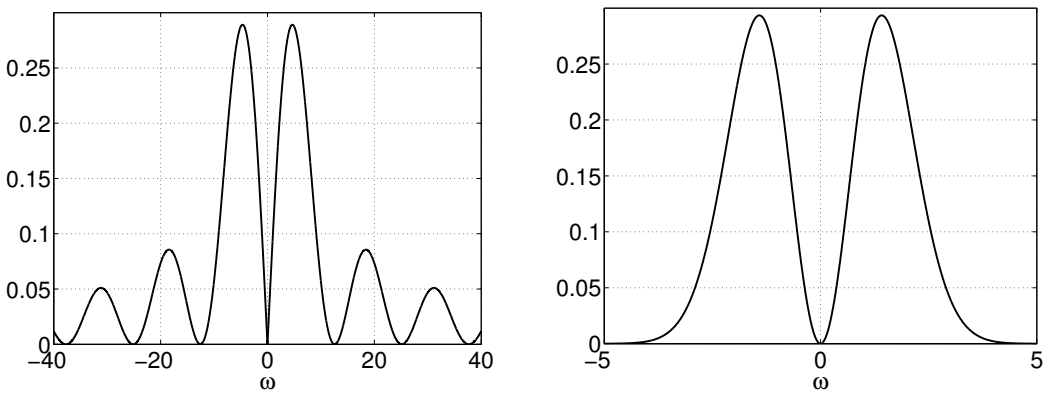


Abbildung 12: Fourier-Transformierte $|\hat{\psi}(\omega)|$ für das Haar-Wavelet (links) und den Mexikanischen Hut (rechts).

Für beliebiges $a \neq 0$ erhalten wir

$$\widehat{\psi(\cdot/a)} = a\hat{\psi}(a\omega).$$

Mit einem Wavelet $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ ist die Fourier-Transformierte $\hat{\psi}$ stetig und wir nehmen an, dass der Grenzwert im Unendlichen existiert, wodurch

$$\lim_{\omega \to \pm\infty} \hat{\psi}(\omega) = 0$$

folgt. Die Aussage (2.10) liefert $\hat{\psi}(0) = 0$. Die Funktion $\hat{\psi}(a\cdot)$ erbt diese Eigenschaften. Daher agiert das Wavelet wie ein Bandpass-Filter. Man beachte, dass $|\hat{\psi}(\omega)| = |\hat{\psi}(-\omega)|$ gilt. Sei die Fourier-Transformierte $\hat{\psi}$ um die Frequenz ω_0 herum konzentriert. Folglich ist $\hat{\psi}(a\cdot)$ um die Frequenz ω_0/a konzentriert. Somit enthält die Funktion $(L_\psi f)(a, \cdot)$ nur Informationen bezüglich der Frequenzen um ω_0/a in f. Deshalb nennt man a den Frequenzparameter. Abbildung 12 zeigt als Beispiel die Fourier-Transformierten zum Haar-Wavelet und zum Mexikanischen Hut.

Besitzt das Wavelet ψ einen kompakten Träger, dann gilt

$$\operatorname{supp}(\psi) \subseteq [-C, C] \quad \Rightarrow \quad \operatorname{supp}(\psi((\cdot - b)/a)) \subseteq [-aC + b, aC + b].$$

Dementsprechend hängt der Wert $(L_\psi f)(a, b)$ nur von der Funktion f im Intervall $[-aC + b, aC + b]$ ab. Eine ähliche Interpretation ist möglich für ein allgemeineres Wavelet, das nur in einem Intervall $[-C, C]$ konzentriert ist.

Einerseits hängt der Parameter a mit den inherenten Frequenzen im zu transformierenden Signal zusammen. Andererseits erlaubt der Parameter b auf einen

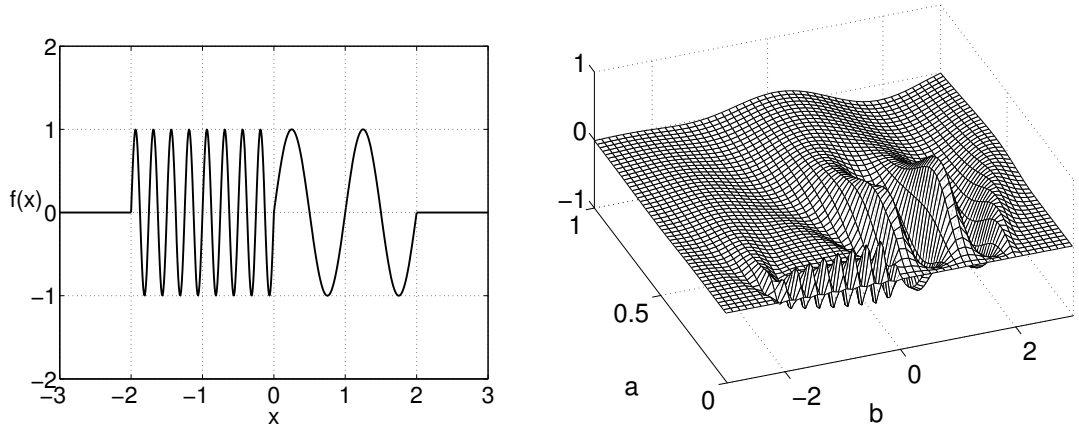


Abbildung 13: Funktion f (links) und Wavelet-Transformierte $L_\psi f$ bezüglich des Mexikanischen Huts (rechts).

bestimmten Zeitpunkt zu fokussieren. Dadurch ist eine Lokalisation sowohl in der Zeit als auch in der Frequenz möglich.

Als Beispiel zur kontinuierlichen Wavelet-Transformation betrachten wir noch die Funktion f in Abbildung 13 (links), welche einen kompakten Träger und im wesentlichen zwei Frequenzanteile $\omega = 2\pi$ ($x \in [0, 2]$) sowie $\omega = 8\pi$ ($x \in [-2, 0]$) besitzt. Die Wavelet-Transfomierte bezüglich des Mexikanischen Huts ist in Abbildung 13 (rechts) dargestellt. Wir erkennen die Frequenzfilterungen für die unterschiedlichen festen Parameter a unter der Veränderlichen b.

Diskrete Wavelet-Transformation

Die Wavelet-Transformation (2.9) erweist sich als redundant, denn wir erhalten eine Funktion f mit der inversen Transformation (2.11) zurück ohne $L_\psi f$ an allen Stellen (a, b) kennen zu müssen. Insbesondere werden nur positive Werte $a > 0$ benötigt. Wir betrachten das zweidimensionale Gitter

$$\{(a, b) = (a_0^m, nb_0a_0^m) : m, n \in \mathbb{Z}\} \subset \mathbb{R}^+ \times \mathbb{R} \tag{2.13}$$

mit Konstanten $a_0 > 1$, $b_0 > 0$. Das zugehörige Gitter im Frequenz-Zeit-Bereich lautet

$$\big\{(\omega, t) = (a_0^{-m}\omega_0, nb_0a_0^m) : m, n \in \mathbb{Z}\big\} \subset \mathbb{R}^+ \times \mathbb{R}$$

mit einer Konzentration um den Frequenzpunkt

$$\omega_0 = \int_0^\infty \omega |\hat{\psi}(\omega)|^2 \, \mathrm{d}\omega.$$

Abbildung 14 verdeutlicht dieses Gitter. Für Wavelets ψ mit gewissen Eigenschaften existieren spezielle Paare a_0, b_0, so dass die inverse Transformation (2.11)

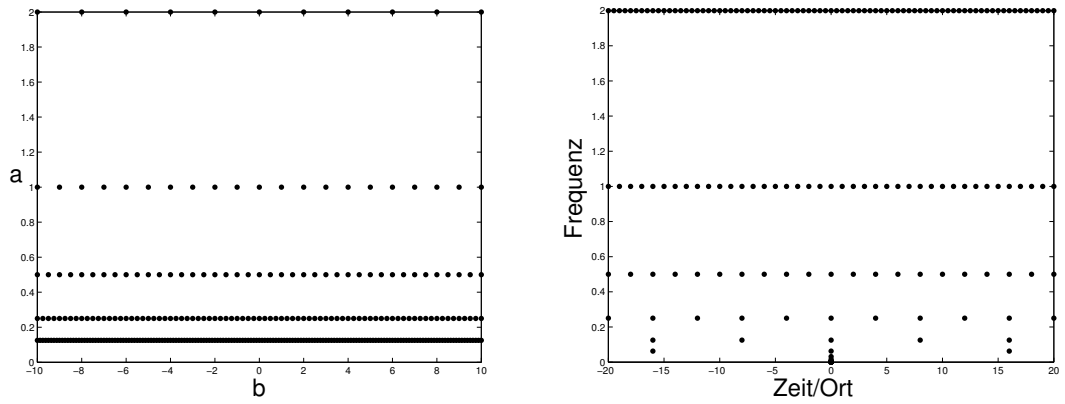


Abbildung 14: Gitter $(a_0^m, nb_0a_0^m)$ im Parameterraum (links) und zugehöriges Gitter $(a_0^{-m}\omega_0, nb_0a_0^m)$ im Frequenz-Zeit-Bereich (rechts) für $a_0 = 2, b_0 = 1, \omega_0 = 1$.

durchgeführt werden kann mittels nur dieser Gitterpunkte. Die Wahl $a_0 = 2$ und $b_0 = 1$ ist oft erwünscht.

Wavelets können für eine Multiskalenanalyse genutzt werden.

Def. 2.17 *Eine Multiskalenanalyse (MSA) von $L^2(\mathbb{R})$ besteht aus einer Folge von abgeschlossenen Unterräumen V_m mit*

$$\{0\} \subset \cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots \subset L^2(\mathbb{R}),$$

welche die Eigenschaften

$$\overline{\bigcup_{m\in\mathbb{Z}} V_m} = L^2(\mathbb{R}), \qquad \bigcap_{m\in\mathbb{Z}} V_m = \{0\},$$

$$f(\cdot) \in V_m \quad \Leftrightarrow \quad f(2^m\cdot) \in V_0$$

erfüllen. Desweiteren soll eine Funktion $\varphi \in L^2(\mathbb{R})$ existieren, deren Translationen $\varphi(\cdot - k)$ eine Riesz-Basis von V_0 bilden, d.h.

$$V_0 = \overline{\operatorname{span}\{\varphi(\cdot - k) : k \in \mathbb{Z}\}}$$

und es gibt Konstanten $A, B > 0$ mit

$$A\sum_{k\in\mathbb{Z}} c_k^2 \le \Big\| \sum_{k\in\mathbb{Z}} c_k\varphi(\cdot - k)\Big\|_{L^2}^2 \le B\sum_{k\in\mathbb{Z}} c_k^2 \qquad \text{für jedes } (c_k)_{k\in\mathbb{Z}} \in \ell^2.$$

Die Abbildung φ heißt Skalierungsfunktion.

Bemerkungen:

- Für jedes $f \in V_0$ gilt
$$f(x) = \sum_{k\in\mathbb{Z}} c_k(f)\varphi(x-k)$$
mit eindeutig bestimmten Koeffizienten $c_k(f) \in \mathbb{R}$ und $(c_k(f))_{k\in\mathbb{Z}} \in \ell^2$.

- Der Raum V_0 ist invariant bezüglich ganzzahliger Translationen, d.h.
$$f \in V_0 \quad \Leftrightarrow \quad f(\cdot - k) \in V_0 \qquad \text{für jedes } k \in \mathbb{Z}.$$
Desweiteren gilt
$$f \in V_m \quad \Leftrightarrow \quad f(\cdot - 2^m k) \in V_m \qquad \text{für jedes } k \in \mathbb{Z}.$$

- Die Funktionen
$$\varphi_{m,k}(x) = 2^{-m/2}\varphi(2^{-m}x - k)$$
erzeugen die Räume V_m, d.h. es ist $V_m = \overline{\text{span}\{\varphi_{m,k} : k \in \mathbb{Z}\}}$. Zudem gilt $\|\varphi_{m,k}\|_{L^2} = \|\varphi\|_{L^2}$. Die Vorfaktoren $2^{-m/2}$ dienen nur zur Normierung.
Denn ist $f \in V_m$, dann folgt nach Definition $f(2^m\cdot) \in V_0$. Somit gilt
$$f(2^m x) = \sum_{k\in\mathbb{Z}} c_k(f(2^m\cdot))\varphi(x-k).$$
Die Substitution $\tilde{x} = 2^m x$ liefert
$$f(\tilde{x}) = \sum_{k\in\mathbb{Z}} \left(2^{m/2}c_k(f(2^m\cdot))\right) 2^{-m/2}\varphi(2^{-m}\tilde{x} - k).$$

Das folgende Lemma wird später bei einem Beispiel benötigt.

Lemma 2.4 *Zu jeder Skalierungsfunktion φ gibt es eine Folge $(h_k)_{k\in\mathbb{Z}} \subset \mathbb{R}$ mit*
$$\varphi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k\varphi(2x-k). \tag{2.14}$$

Beweis:

Es gilt
$$\varphi \in V_0 \subset V_{-1} = \overline{\text{span}\{\sqrt{2}\varphi(2x-k) : k \in \mathbb{Z}\}}.$$
Weil die Funktionen $\varphi(x-k)$ eine Riesz-Basis von V_0 bilden, stellen die Funktionen $\varphi(2x-k)$ eine Riesz-Basis von V_{-1} dar. Es folgt die Existenz der Entwicklung (2.14). □

Die nächste Definition verwendet die direkte Summe von Unterräumen.

Def. 2.18 *Der Unterraum W_m sei das orthogonale Komplement von V_m im Unterraum V_{m-1}, d.h.*

$$V_{m-1} = V_m \oplus W_m \quad mit \quad V_m \perp W_m.$$

Wir können diese Zerlegung sukzessive anwenden, z.B.

$$V_0 = V_1 \oplus W_1 = V_2 \oplus W_2 \oplus W_1 = V_3 \oplus W_3 \oplus W_2 \oplus W_1 = V_m \oplus W_m \oplus \cdots \oplus W_1.$$

Dadurch erhalten wir durch Grenzübergang

$$V_m = \bigoplus_{j \geq m+1} W_j, \qquad L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j.$$

Die Unterräume W_m besitzen die Skalierungseigenschaft

$$f \in W_m \quad \Leftrightarrow \quad f(2^m \cdot) \in W_0.$$

Wir beweisen eine Richtung: Sei $f \in W_m$. Es folgt $f \in V_{m-1}$ wegen $V_{m-1} = V_m \oplus W_m$. Wir haben $f(2^m \cdot) \in V_{-1} = V_0 \oplus W_0$. Ist $g \in V_0$, so ist $g(2^{-m} \cdot) \in V_m$. Es folgt $\langle f(2^m \cdot), g \rangle_{L^2} = 2^{-m} \langle f, g(2^{-m} \cdot) \rangle_{L^2} = 0$. Daher gilt $f(2^m \cdot) \perp V_0$ und somit $f(2^m \cdot) \in W_0$.

Die Räume V_m sind alle als abgeschlossen vorausgesetzt. Es ist leicht zu zeigen, dass dann auch alle Räume W_m abgeschlossen sind (bilde Cauchy-Folge $(f_n)_{n \in \mathbb{N}} \subset W_m$, dann $f_n \to \hat{f} \in V_{m-1}$ und $\hat{f} \perp V_m$). Daher exisitieren die Orthogonalprojektionen $P_m : L^2(\mathbb{R}) \to V_m$ und $Q_m : L^2(\mathbb{R}) \to W_m$ auf die Unterräume V_m bzw. W_m, siehe Theorem V.3.4 in [24]. Insbesondere gilt $\text{kern}(P_m) = V_m^\perp = W_m$ und $\text{bild}(P_m) = V_m$. Somit folgt

$$P_{m-1} = Q_m + P_m.$$

Die wiederholte Anwendung dieser Projektion ist in folgendem Diagramm dargestellt:

$$\begin{array}{ccccccccccccc} L^2(\mathbb{R}) & \cdots & \longrightarrow & V_{-1} & \longrightarrow & V_0 & \longrightarrow & V_1 & \longrightarrow & V_2 & \longrightarrow & \cdots & \{0\} \\ & & & & \searrow & & \searrow & & \searrow & & & & \\ & & & & & W_0 & & W_1 & & W_2 & & & \end{array}$$

In einigen Anwendungen sind Zerlegungen

$$V_J = V_0 \oplus \bigoplus_{j=J+1}^{0} W_j$$

mit $J < 0$ nützlich, wobei die Räume V_J dann numerische Approximationen eines bestimmten Problems enthalten.

Das Hauptergebnis zur MSA ist in folgendem Satz enthalten.

Satz 2.19 *Zu jeder MSA existiert ein (nicht eindeutiges) Wavelet ψ, so dass die Funktionen*

$$\psi_{m,k}(x) = 2^{-m/2}\psi(2^{-m}x - k) \tag{2.15}$$

eine Orthonormalbasis von W_m für festes $m \in \mathbb{Z}$ bilden. Ein solches Wavelet kann explizit aus der Skalierungsfunktion konstruiert werden.

Der Beweis ergibt sich aus mehreren anderen Sätzen, siehe [14]. Da alle Funktionen $\psi_{m,k}$ durch ψ definiert sind, nennt man die Funktion ψ das Mutter-Wavelet.

Beispiel:

Eine einfache MSA ergibt sich aus der Skalierungsfunktion

$$\varphi(x) = \begin{cases} 1 & \text{für } 0 \le x < 1, \\ 0 & \text{sonst.} \end{cases}$$

Für jedes feste $m \in \mathbb{Z}$ gilt

$$\begin{aligned} V_m &= \overline{\operatorname{span}\{\varphi_{m,k} : k \in \mathbb{Z}\}} \\ &= \{f \in L^2(\mathbb{R}) : f \text{ konstant auf } [2^m k, 2^m(k+1)[\text{ für alle } k \in \mathbb{Z}\}\,. \end{aligned}$$

Die Projektion P_m resultiert aus

$$(P_m f)(x) = 2^{-m} \int_{2^m k}^{2^m(k+1)} f(t)\,\mathrm{d}t \quad \text{für } x \in [2^m k, 2^m(k+1)[.$$

In diesem Beispiel stellen die Funktionen $\{\varphi_{m,k} : k \in \mathbb{Z}\}$ eine Orthonormalbasis von V_m dar. Dadurch erhalten wir eine Darstellung von $P_m f$ als

$$P_m f = \sum_{k\in\mathbb{Z}} c_k^m(f)\varphi_{m,k}$$

mit

$$c_k^m(f) = \langle P_m f, \varphi_{m,k}\rangle_{L^2} = 2^{-m/2} \int_{2^m k}^{2^m(k+1)} f(t)\,\mathrm{d}t.$$

Die Skalierungseigenschaft (2.14) erhalten wir durch

$$\varphi(x) = \sqrt{2}\left(\tfrac{1}{\sqrt{2}}\varphi(2x) + \tfrac{1}{\sqrt{2}}\varphi(2x-1)\right).$$

Daraus folgt die Entwicklung von $\varphi_{m+1,k}$ in der Basis $\{\varphi_{m,k} : k \in \mathbb{Z}\}$

$$\varphi_{m+1,k} = \tfrac{1}{\sqrt{2}}(\varphi_{m,2k} + \varphi_{m,2k+1}).$$

Für die Koeffizienten von $P_{m+1}f \in V_{m+1} \subset V_m$ folgt

$$c_k^{m+1}(f) = \tfrac{1}{\sqrt{2}}(c_{2k}^m(f) + c_{2k+1}^m(f)).$$

Wir erhalten für die Projektion auf W_{m+1}

$$\begin{aligned} Q_{m+1}f &= P_m f - P_{m+1}f \\ &= \sum_{k\in\mathbb{Z}} c_k^m \varphi_{m,k} - \sum_{k\in\mathbb{Z}} c_k^{m+1}\varphi_{m+1,k} \\ &= \sum_{k\in\mathbb{Z}} (c_{2k}^m \varphi_{m,2k} + c_{2k+1}^m \varphi_{m,2k+1}) - \tfrac{1}{2}\sum_{k\in\mathbb{Z}} (c_{2k}^m + c_{2k+1}^m)(\varphi_{m,2k} + \varphi_{m,2k+1}) \\ &= \tfrac{1}{2}\sum_{k\in\mathbb{Z}} (c_{2k}^m - c_{2k+1}^m)(\varphi_{m,2k} - \varphi_{m,2k+1}). \end{aligned}$$

Dadurch kann die Funktion $Q_{m+1}f \in W_{m+1}$ als eine Reihe geschrieben werden mit den Basisfunktionen

$$\psi_{m+1,k} = \tfrac{1}{\sqrt{2}}(\varphi_{m,2k} - \varphi_{m,2k+1}).$$

Man kann zeigen, dass die Funktionen $\psi_{m+1,k}$ eine Orthonormalbasis von W_{m+1} bilden. Das Mutter-Wavelet $\psi = \psi_{0,0}$ ergibt sich zu

$$\psi(x) = \varphi(2x) - \varphi(2x-1) = \begin{cases} 1 & \text{für } 0 \le x < \frac{1}{2} \\ -1 & \text{für } \frac{1}{2} \le x < 1 \\ 0 & \text{sonst,} \end{cases}$$

welches gerade das Haar-Wavelet ist.

Def. 2.19 *In einer MSA heißt eine Skalierungsfunktion φ orthogonal, wenn ihre ganzzahligen Translationen $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$ eine Orthogonalbasis des Unterraums V_0 bilden.*

In obigem Beispiel ist die Konstruktion des Wavelets einfach, weil die Skalierungsfunktion orthogonal ist. Falls eine Skalierungsfunktion keine Orthogonalbasis erzeugt, dann kann man daraus eine orthogonale Skalierungsfunktion konstruieren, welche die erforderlichen weiteren Schritte ermöglicht.

Satz 2.20 *Sei $(V_m)_{m\in\mathbb{Z}}$ eine MSA, welche von einer orthogonalen Skalierungsfunktion $\varphi \in V_0$ erzeugt wird. Die Funktion $\psi \in V_{-1}$ definiert durch*

$$\psi(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} g_k \varphi(2x-k) = \sum_{k\in\mathbb{Z}} g_k \varphi_{-1,k}(x) \tag{2.16}$$

mit $g_k = (-1)^k h_{1-k}$, wobei $(h_k)_{k\in\mathbb{Z}}$ die Koeffizienten aus (2.14) sind, besitzt die folgenden Eigenschaften mit $\psi_{m,k}(x) = 2^{-m/2}\psi(2^{-m}x-k)$:

(i) $\{\psi_{m,k} : k \in \mathbb{Z}\}$ *bildet eine Orthonormalbasis von* W_m,

(ii) $\{\psi_{m,k} : m,k \in \mathbb{Z}\}$ *stellt eine Orthonormalbasis von* $L^2(\mathbb{R})$ *dar,*

(iii) ψ *ist ein Wavelet mit* $c_\psi = 2\ln 2$.

Die Eigenschaft (ii) aus Satz 2.20 impliziert

$$f(x) = \sum_{m,k\in\mathbb{Z}} \langle f, \psi_{m,k}\rangle_{L^2}\ \psi_{m,k}(x) \qquad \text{für jedes } f \in L^2(\mathbb{R}). \tag{2.17}$$

Genauer bedeutet dies

$$\lim_{n\to\infty} \left\| f(x) - \sum_{m=-n}^{n}\sum_{k=-n}^{n} \langle f, \psi_{m,k}\rangle_{L^2}\ \psi_{m,k}(x) \right\|_{L^2} = 0.$$

Desweiteren gilt wegen $\psi_{m,k} = 2^{-m/2}\psi(\frac{x-2^m k}{2^m})$

$$(L_\psi f)(2^m, 2^m k) = \tfrac{1}{\sqrt{c_\psi}}\langle f, \psi_{m,k}\rangle_{L^2} \qquad \text{für alle } m,k \in \mathbb{Z}.$$

Dementsprechend kann die ursprüngliche Funktion f vollständig aus den Werten ihrer Wavelet-Transformierten nur in den Gitterpunkten (2.13) mit $a_0 = 2$ und $b_0 = 1$ rekonstruiert werden. Die Koeffizientenfolgen $(g_k)_{k\in\mathbb{Z}}$ und $(h_k)_{k\in\mathbb{Z}}$ ermöglichen eine schnelle Wavelet-Transformation.

Eine Umkehrung von Satz 2.20 gilt jedoch nicht. Es gibt Wavelets ψ, die durch $\psi_{m,k} = 2^{-m/2}\psi(2^{-m}x-k)$ sogar eine Orthonormalbasis von $L^2(\mathbb{R})$ erzeugen, und gleichzeitig keine MSA existiert, welche ψ als Mutter-Wavelet zulässt.

Die kontinuierliche Wavelet-Transformation (2.9) stellt eine Integraltransformation für Funktionen in $L^2(\mathbb{R})$ dar. Demgegenüber liefert die diskrete Wavelet-Transformation eine Entwicklung (2.17) für Funktionen in $L^2(\mathbb{R})$ mit einem abzählbar unendlichen System aus Basisfunktionen. Desweiteren existiert noch die sogenannte schnelle Wavelet-Transformation, welche aus einer diskreten endlichen Datenmenge eine äquivalente transformierte endliche Darstellung erzeugt.

Als Ausblick seien noch die folgenden Anwendungen der schnellen Wavelet-Transformation erwähnt:

- *Signalanalyse*:
  Mit dem Prinzip der Multiskalen-Analyse kann hier nach bestimmten Anteilen bzw. Mustern in einem Signal gesucht werden. Ein Beispiel ist die Untersuchung von Signalen aus einem Elektrokardiogramm (EKG), siehe Abschnitt 3.1.2 in [14].

- *Datenkompression*:
  Hier wird in die endliche Wavelet-Darstellung transformiert und relativ kleine Koeffizienten werden auf null gesetzt bzw. nicht mehr mit abgespeichert. Hauptanwendungsbereich ist die Speicherplatzersparnis bei Bildern oder Fotos, d.h. zweidimensionalen Daten, siehe Abschnitt 3.3 in [14]. Der Kompressionsstandard JPEG2000 beruht auf dieser Wavelet-Transformation.

- *Rauschreduktion/Entrauschen*:
  Es wird in die endliche Wavelet-Darstellung transformiert und bestimmte Koeffizienten zu hohen Frequenzen werden auf null gesetzt (auch wenn diese Koeffizienten möglicherweise groß sind). Beispiele sind hier die Entrauschung von eindimensionalen Daten wie Ton bzw. Musik und zweidimensionalen Daten wie Bildern bzw. Fotos.

Anwendungsbeispiele zur (diskreten) Wavelet-Transformation können auch in Kapitel 4 aus [2] gefunden werden.

3 Parameterbestimmung

In diesem Kapitel behandeln wir die geeignete Bestimmung der Zahlwerte von (physikalischen) Parametern in mathematischen Modellen.

3.1 Problemstellung und Beispiele

Es bezeichne $\mathcal{S}$ ein physikalisches System. Beispiele sind die Planeten mit der Sonne im Sonnensystem, elektrische Schaltungen oder Substanzen unter chemischen Reaktionen in einem Behälter mit Wasser. Das System $\mathcal{S}$ ist oft zeitabhängig. Ein zeitabhängiges System enthält häufig Eingangssignale $u(t)$. Existieren keine zeitabhängigen Eingangssignale, so spricht man von autonomen Systemen. Oft sind nicht alle zeitabhängigen Anteile des Systems $\mathcal{S}$ messbar, sondern nur bestimmte beobachtbare Größen $y(t)$, die man auch als Ausgangssignale interpretieren kann.

Die typischen Schritte in der Modellbildung sind:

1. *Parametrisierung des Systems* :
   Identifiziere eine minimale Menge von Modellparametern (physikalische Parameter und/oder künstliche Parameter), die das System $\mathcal{S}$ vollständig charakterisieren. (Zahlwerte der Parameter sind hier noch nicht festgelegt.)

2. *Vorwärts-Modellierung* :
   Bestimme die physikalischen Gesetze, die es erlauben aus den Modellparametern dann Vorhersagen zu beobachtbaren Größen zu machen. Hier wird im wesentlichen ein mathematisches Modell konstruiert.

3. *Rückwärts-Modellierung* :
   Verwende die Ergebnisse aus Messungen von beobachtbaren Größen um die Zahlwerte der Modellparameter festzulegen.

Es besteht eine starke Interaktion bzw. Rückkopplung zwischen diesen drei Schritten. Auch kann Schritt 2 zuerst erfolgen, wobei man sich dabei über die benötigten Parameter aus Schritt 1 im klaren wird. Wir wollen in diesem Kapitel hauptsächlich Verfahren zur Durchführung von Schritt 3 besprechen.

Zudem unterscheiden wir zwischen zwei Modelltypen:

1. *statische Modelle*,

2. *dynamische Modelle*.

Mathematische Modelle vom Typ 2 sind häufig durch Differentialgleichungen gegeben.

Beispiele von jeweils zwei statischen Modelle und zwei dynamischen Modellen sind:

(i) *Bethe-Weizsäcker-Formel*
Ein Atomkern besteht aus N Neutronen und Z Protonen. Die Nukleonenzahl ist entsprechend $A = N+Z$. Für die Gesamtbindungsenergie des Kerns wird folgendes Modell aufgestellt

$$E_{\text{Bindung}} = E_{\text{Volumen}} - E_{\text{Oberfläche}} - E_{\text{Coulomb}} - E_{\text{Symmetrie}} \pm E_{\text{Paarung}}.$$

Es entsteht die Formel

$$\begin{aligned} E_{\text{Bindung}}(N,Z) &= a_{\text{V}}A - a_{\text{O}}A^{\frac{2}{3}} - a_{\text{C}}Z(Z-1)A^{-\frac{1}{2}} - a_{\text{S}}\frac{(N-Z)^2}{4A} \\ &\quad + \begin{cases} +a_{\text{P}}A^{-\frac{1}{2}} & \text{für } N,Z \text{ gerade,} \\ -a_{\text{P}}A^{-\frac{1}{2}} & \text{für } N,Z \text{ ungerade,} \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

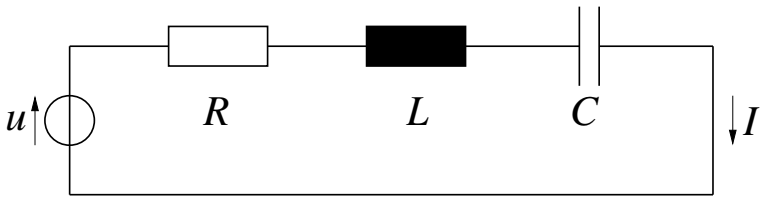
Die dabei auftretenden Modellparameter sind die Proportionalitätskonstanten $a_{\text{V}}, a_{\text{O}}, a_{\text{C}}, a_{\text{S}}, a_{\text{P}}$. Aus Messungen der Bindungsenergien von mindestens fünf verschiedenen Atomkernen lassen sich diese Parameter identifizieren. Je nach Auswahl dieser Referenzkerne entstehen in der Literatur leicht verschiedene Parameterwerte. Für weitere Informationen hierzu siehe Abschnitt 18.1.3 in [8].

(ii) *Kettenlinie*
Eine hängende Kette in einer zweidimensionalen Ebene nimmt die Form des Graphen der Funktion

$$y(x) = \tfrac{1}{a}\cosh(a(x+C_1)) + C_2 \tag{3.1}$$

ein. Daher sind die Modellparameter $a > 0$ und $C_1, C_2 \in \mathbb{R}$. Die Parameter C_1, C_2 lassen sich aus einer Anfangswertvorgabe $y(x_0) = y_0$, $y'(x_0) = y_0'$ oder Randwertvorgabe $y(x_0) = y_0$, $y(x_{\text{end}}) = y_{\text{end}}$ bei festem a bestimmen. Der Parameter a steht im Zusammenhang mit der Gesamtlänge der Kette.

(iii) *elektromagnetischer Schwingkreis*
Wir betrachten eine elektrische Schaltung bestehend aus einem Widerstand mit Wert R, einer Spule mit Induktivität L, einem Kondensator mit Kapazität C und einer Spannungsquelle in Reihe geschaltet.



Die Spannungsquelle liefert ein zeitabhängiges Eingangssignal $u(t)$. Als Ausgangssignal wird der Strom $I(t)$ durch die Elemente definiert. Als mathematisches Modell folgt ein System aus zwei Differentialgleichungen erster Ordnung

$$\begin{aligned} \frac{\mathrm{d}I}{\mathrm{d}t} &= -\frac{1}{LC}Q(t) - \frac{R}{L}I(t) + \frac{1}{L}u(t) \\ \frac{\mathrm{d}Q}{\mathrm{d}t} &= I(t) \end{aligned} \tag{3.2}$$

mit der Lösung $(I(t), Q(t))^\top$. Die enthaltenen Modellparameter sind somit $R, L, C > 0$. Gilt $u \equiv 0$, dann ist die Lösung des Systems (3.2) gegeben durch

$$I(t) = \mathrm{e}^{-\delta t}\left[A_0 \sin(\omega t) + B_0 \cos(\omega t)\right]$$

mit $\delta = \frac{R}{2L}$ und $\omega^2 = \frac{1}{LC} - \frac{R^2}{4L^2}$. Die Konstanten A_0, B_0 bestimmen sich aus Anfangswertvorgaben.

(iv) *Hydrolyse von Harnstoff*
Als ein Beispiel mit chemischen Reaktionen sei der Abbau von Harnstoff betrachtet. Die beteiligten Reaktionsgleichungen sind

$$\begin{aligned} \mathrm{U} + \mathrm{E} &\xrightarrow{k_1} \mathrm{UE} \\ \mathrm{UE} &\xrightarrow{k_2} \mathrm{U} + \mathrm{E} \\ \mathrm{UE} + 2\,\mathrm{H_2O} &\xrightarrow{k_3} \mathrm{A} + \mathrm{E} \end{aligned}$$

mit den Stoffen U: Harnstoff (Urea), E: Enzym (Urease), UE: Komplex aus U und E, A: Ammoniumkarbonat, H_2O: Wasser. Die Modellparameter sind die Reaktionsgeschwindigkeiten $k_1, k_2, k_3 > 0$. Als mathematische Beschreibung entsteht ein zeitabhängiges System aus Differentialgleichungen erster Ordnung

$$\begin{aligned} c'_{\mathrm{U}} &= -k_1 c_{\mathrm{U}} c_{\mathrm{E}} + k_2 c_{\mathrm{UE}} \\ c'_{\mathrm{E}} &= -k_1 c_{\mathrm{U}} c_{\mathrm{E}} + k_2 c_{\mathrm{UE}} + k_3 c_{\mathrm{UE}} \\ c'_{\mathrm{UE}} &= k_1 c_{\mathrm{U}} c_{\mathrm{E}} - k_2 c_{\mathrm{UE}} - k_3 c_{\mathrm{UE}} \\ c'_{\mathrm{A}} &= k_3 c_{\mathrm{UE}} \end{aligned}$$

für die Konzentrationen $c_{\mathrm{U}}(t)$, $c_{\mathrm{E}}(t)$, $c_{\mathrm{UE}}(t)$, $c_{\mathrm{A}}(t)$ der vier Stoffe. Da keine Eingangssignale vorliegen ist dieses System autonom.

3.2 Ausgleichsrechnung

Wir bezeichnen die Modellparameter im folgenden als $\theta \in \Theta$ mit einer Menge $\Theta \subseteq \mathbb{R}^n$ aus zulässigen Parameterwerten bzw. physikalisch sinnvollen Parameterwerten. Eine einfache allgemeine Notation eines Modells ist

$$y = f(\theta)$$

mit den beobachtbaren Ergebnissen $y \in \mathbb{R}^k$ und einer Funktion $f : \Theta \to \mathbb{R}^k$. Bei zeitabhängigen physikalischen Systemen modifiziert sich die Darstellung zu

$$y(t) = f(t, \theta)$$

mit einer Funktion $f : [t_0, t_{\text{end}}] \times \Theta \to \mathbb{R}^k$.

Sei nun zur Vereinfachung $k = 1$. Die Parameter sollen aus Messungen der beobachtbaren Ergebnisse geeignet bestimmt werden. Es bezeichne $\hat{y}(t)$ die in der „Realität“ exakt vorliegende Größe. Zu den Zeitpunkten $t_1 < t_2 < \cdots < t_m$ werden Messungen durchgeführt, wodurch die Ergebnisse

$$\hat{y}(t_i) = y_i + \varepsilon_i \qquad \text{für } i = 1, \ldots, m$$

folgen. Es sind $\varepsilon_i \in \mathbb{R}$ die stets auftretenden Messfehler. Die Messfehler lassen sich gegebenenfalls nur durch genauere Messverfahren reduzieren, d.h. hier haben wir keinen direkten Einfluss. Demgegenüber weist unser Modell auch immer einen Modellfehler gegenüber der Realität auf, d.h.

$$y(t) = f(t, \theta) = \hat{y}(t) + \delta(t, \theta)$$

mit dem Modellfehler $\delta : [t_0, t_{\text{end}}] \times \Theta \to \mathbb{R}$. Im allgemeinen ist es nicht möglich den Modellierungsfehler zu beseitigen. Jedoch hängt die Größe von δ nun von der Wahl der Parameter θ ab. Ziel ist daher, diesen Anteil möglichst klein zu halten.

Es sei $\delta_i(\theta) = \delta(t_i, \theta)$ für $i = 1, \ldots, m$. Unter Einbezug der Messungen entsteht die Formel

$$y(t_i) = f(t_i, \theta) = y_i + \delta_i(\theta) + \varepsilon_i \qquad \text{für } i = 1, \ldots, m.$$

Die Wahl der Parameter θ beeinflusst die Messfehler nicht. Wir nehmen daher zunächst $\varepsilon_i = 0$ für alle i an. Die Modellfehler sollen minimiert werden. Wir setzen voraus, dass die Anzahl der Messungen immer mindestens so hoch ist wie die Anzahl der zu bestimmenden Parameter, d.h. $m \geq n$. Mit der Methode der kleinsten Quadrate entsteht das Minimierungsproblem

$$\min_{\theta \in \Theta} \sum_{i=1}^{m} \delta_i(\theta)^2 = \sum_{i=1}^{m} \left(y_i - f(t_i, \theta)\right)^2. \tag{3.3}$$

Je nach dem funktionalen Zusammenhang entsteht ein lineares oder nichtlineares Ausgleichsproblem. Ein lineares Problem liegt genau dann vor, wenn gilt

$$f(t_i, \theta) = \varphi_{i1}\theta_1 + \varphi_{i2}\theta_2 + \cdots + \varphi_{in}\theta_n \qquad \text{für } i = 1, \ldots, m$$

mit Koeffizienten $\varphi_{ij} \in \mathbb{R}$, die unabhängig von θ sind. Insbesondere ist dies der Fall, wenn als Modell ein funktionaler Zusammenhang

$$f(t, \theta) = f_1(t)\theta_1 + f_2(t)\theta_2 + \cdots + f_n(t)\theta_n$$

angesetzt wird mit vorgegebenen stetigen Funktionen $f_j : [t_0, t_{\text{end}}] \to \mathbb{R}$ für $j = 1, \ldots, n$. Dabei wird die lineare Unabhängigkeit dieser Funktionenmenge gefordert. Es folgt dann $\varphi_{ij} = f_j(t_i)$.

Ordnen wir die Koeffizienten in der Matrix $\Phi \in \mathbb{R}^{m \times n}$ und die Messungen im Vektor $y = (y_1, \ldots, y_m)^\top$ an, dann kann das lineare Ausgleichsproblem geschrieben werden als

$$\min_{\theta \in \Theta} \|y - \Phi\theta\|_2 \tag{3.4}$$

mit der Euklidischen Norm $\|\cdot\|_2$. Das lineare Ausgleichsproblem ist genau dann eindeutig lösbar in $\mathbb{R}^n$, wenn der Rang der Matrix Φ gleich n ist. In diesem Fall ist das Minimum gegeben als Lösung der Normalgleichungen

$$\left(\Phi^\top \Phi\right) \theta = \Phi^\top y,$$

welches ein lineares Gleichungssystem mit quadratischer Matrix darstellt.

Die Verwendung des Ansatzes der kleinsten Quadrate (3.3) bleibt auch noch sinnvoll, wenn die Messfehler ungleich null sind, d.h. $\varepsilon_i \neq 0$, sofern keine systematischen Messfehler vorliegen.

Beispiele:

(i) *Polynome* :
Als funktionalen Zusammenhang sei ein Polynom vom Grad $n-1$ gegeben

$$f(t, \theta) = \theta_0 + \theta_1 t + \theta_2 t^2 + \cdots + \theta_{n-2} t^{n-2} + \theta_{n-1} t^{n-1}$$

mit den n Koeffizienten $\theta_0, \ldots, \theta_{n-1}$. Man beachte, dass ein Polynom mit Grad höher als eins zwar eine nichtlineare Funktion ist, jedoch die Abhängigkeit der Funktion von ihren Koeffizienten linear ist. Es entsteht ein lineares Ausgleichsproblem (3.4) mit der Koeffizientenmatrix

$$\Phi = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^{n-1} \end{pmatrix},$$

die in dieser Anwendung auch Van-der-Monde-Matrix genannt wird. Man kann zeigen, dass bei paarweise verschiedenen Messpunkten t_i der Rang dieser Matrix gleich n ist. Dieses Beispiel trat auch in Abschnitt 1.2 bei (1.5) auf.

(ii) *Kettenlinie* :
Bei der Kettenlinie haben wir statt der Zeitpunkte nun Ortspunkte gegeben. Es seien (x_i, y_i) für $i = 1, \ldots, m$ Messungen von Positionen einer hängenden Kette. Mit dem funktionalen Zusammenhang (3.1) entsteht das Ausgleichsproblem

$$\min_{\theta \in \Theta} \sum_{i=1}^{m} \left(y_i - \frac{1}{\theta_3} \cosh\left(\theta_3(x_i + \theta_1)\right) + \theta_2 \right)^2$$

mit der Parametermenge $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$. Dieses Ausgleichsproblem ist nichtlinear, da Parameterwerte innerhalb der nichtlinearen Cosinushyperbolicusfunktion auftreten.

Die Lösung von linearen Ausgleichsproblemen (3.4) kann mit numerischen Verfahren direkt erfolgen. Typischerweise wird eine QR-Zerlegung der Koeffizientenmatrix Φ bestimmt, wobei entweder Householder-Transformationen oder Givens-Rotationen zum Einsatz kommen. Demgegenüber kann eine numerische Lösung von nichtlinearen Ausgleichsproblemen nur mit iterativen Verfahren erreicht werden, beispielsweise mit dem Gauß-Newton-Verfahren. Näheres ist in Abschnitt 4.8 bei [20] beschrieben.

Konvergenz bei linearem Ausgleichsproblem

Im folgenden behandeln wir Resultate aus dem Bereich der linearen Regression. Wir nehmen als Idealisierung an, dass die Modellfehler identisch null sind für eine eindeutige Wahl $\hat{\theta} \in \Theta$ der Modellparameter. Dementsprechend liefern die Messungen bis auf die Messfehler das beobachtbare Ergebnis unseres Modells. Wir fragen danach, ob bei einer in gewissem Sinne gleichmäßen Verteilung der Messfehler eine Konvergenz der Lösung unseres linearen Ausgleichsproblems gegen die optimalen Parameter $\hat{\theta}$ im System vorliegt. Die Messfehler werden dazu als zufallsabhängig angenommen.

Def. 3.1 *Eine k-dimensionale Zufallsvariable X ist normalverteilt mit Erwartungswert $\mu \in \mathbb{R}^k$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^{k \times k}$, wenn sie eine Dichtefunktion besitzt der Gestalt*

$$f(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \quad \textit{mit } x \in \mathbb{R}^k.$$

Die Kovarianzmatrix wird dabei als symmetrisch und positiv definit vorausgesetzt. Über lineare Transformationen von Normalverteilungen gibt das folgende Lemma eine Auskunft.

Lemma 3.1 *Sei X eine k-dimensionale normalverteilte Zufallsvariable mit Erwartungswert μ und Kovarianzmatrix Σ. Für $T \in \mathbb{R}^{\ell \times k}$ mit $\ell \leq k$ und* $\operatorname{rang}(T) = \ell$ *ist $Y = TX + \eta$ eine ℓ-dimensionale normalverteilte Zufallsvariable mit Erwartungswert $T\mu + \eta$ und Kovarianzmatrix $T\Sigma T^\top$.*

Beweis: siehe Abschnitt 6.7 in [13].

Im Spezialfall $\ell = k$ ist somit eine reguläre Transformationsmatrix gefordert. Sei X eine k-dimensionale standard-normalverteilte Zufallsvariable, d.h. $\mu = 0$ und $\Sigma = I$. X besteht damit aus k unabhängigen standard-normalverteilten Zufallsvariablen. Sei Y eine k-dimensionale normalverteilte Zufallsvariable mit Erwartungswert η und Kovarianzmatrix $\tilde{\Sigma}$. Bilden wir die Cholesky-Zerlegung $\tilde{\Sigma} = LL^\top$, so folgt, dass die Zufallsvariable $\tilde{Y} = LX + \eta$ die gleiche Verteilung wie Y besitzt wegen $LIL^\top = \tilde{\Sigma}$. Aus dieser Überlegung notieren wir die folgende Interpretation.

Korollar 3.1 *Y ist genau dann eine k-dimensionale normalverteilte Zufallsvariable, wenn es eine k-dimensionale Zufallsvariable X mit unabhängigen standard-normalverteilten Komponenten und eine reguläre Matrix $T \in \mathbb{R}^{k \times k}$ sowie $\eta \in \mathbb{R}^k$ gibt, so dass $\tilde{Y} = TX + \eta$ die gleiche Verteilung wie Y besitzt.*

Sei nun ein Modell mit einem linearen Struktur gegeben, d.h.

$$y(t) = f_1(t)\theta_1 + f_2(t)\theta_2 + \cdots + f_n(t)\theta_n$$

mit linear unabhängigen stetigen Ansatzfunktionen. Die Dimension n ist damit a priori fest gewählt. Es seien Messungen $y_1, \ldots, y_m$ gegeben

$$\hat{y}(t_i) = y_i + \varepsilon_i \qquad \text{für } \; i = 1, \ldots, m$$

mit den Messfehlern ε_i, wobei $m \gg n$ gelte. Die Messfehler werden als zufallsabhängig ohne einen systematischen Fehler angenommen. Dies geschieht durch die Annahme unabhängiger Zufallsvariablen mit jeweils einem Erwartungswert null.

Es folgt ein lineares Ausgleichsproblem (3.4) für die zu bestimmenden Parameter

$\theta_m = (\theta_1, \ldots, \theta_n)^\top$ mit Matrix $\Phi_m \in \mathbb{R}^{m\times n}$ und Vektor $z_m \in \mathbb{R}^m$

$$\Phi_m = \begin{pmatrix} f_1(t_1) & f_2(t_1) & \cdots & f_n(t_1) \\ f_1(t_2) & f_2(t_2) & \cdots & f_n(t_2) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ f_1(t_m) & f_2(t_m) & \cdots & f_n(t_m) \end{pmatrix}, \quad z_m = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{pmatrix}.$$

Der nächste Satz liefert ein Resultat über die Konvergenz in Wahrscheinlichkeit. O.E.d.A. betrachten wir das Intervall $t \in [0, 1]$.

Satz 3.1 *Es sei $\hat{\theta} \in \mathbb{R}^n$ das eindeutige Minimum der Modellfehler in der L^2-Norm, d.h.*

$$\int_0^1 \delta(t, \hat{\theta})^2 \, \mathrm{d}t \le \int_0^1 \delta(t, \theta)^2 \, \mathrm{d}t \qquad \textit{für alle } \theta \in \mathbb{R}^n.$$

Die Zeitpunkte seien $t_i = ih$ für $i = 1, \ldots, m$ mit Schrittweite $h = \frac{1}{m}$. Sind die Messfehler ε_i für $i = 1, \ldots, m$ unabhängige Zufallsvariablen mit identischen Normalverteilungen zu Erwartungswert null und Varianz σ^2, dann folgt

$$\forall \eta > 0 \; : \; \lim_{m\to\infty} P\left(\left\| \theta_m - \hat{\theta} \right\| > \eta \right) = 0$$

mit einer beliebigen Vektornorm $\|\cdot\|$ auf dem $\mathbb{R}^n$.

Beweis:

i) Es sei $f(\cdot, \hat{\theta})$ die Bestapproximation an $\hat{y}$ im endlichdimensionalen Untervektorraum $U = \mathrm{span}\{f_1, \ldots, f_n\}$ bezüglich der L^2-Norm. Diese Bestapproximation existiert und ist eindeutig, weil ein Hilbert-Raum vorliegt. Wegen Satz 2.11 erfüllt $\hat{\theta}$ damit die Normalgleichungen $B\hat{\theta} = c$ mit den Koeffizienten

$$b_{ij} = \int_0^1 f_i(t) f_j(t) \, \mathrm{d}t, \qquad c_i = \int_0^1 f_i(t) \hat{y}(t) \, \mathrm{d}t \qquad \text{für } \; i, j = 1, \ldots, n.$$

ii) Die Lösung θ_m des linearen Ausgleichproblems ist gegeben durch

$$\theta_m = (\Phi_m^\top \Phi_m)^{-1} \Phi_m^\top z_m = (\Phi_m^\top \Phi_m)^{-1} \Phi_m^\top \hat{z}_m - (\Phi_m^\top \Phi_m)^{-1} \Phi_m^\top e_m.$$

mit $\hat{z}_m = (\hat{y}(t_1), \ldots, \hat{y}(t_m))^\top$ und $e_m = (\varepsilon_1, \ldots, \varepsilon_m)^\top$. Für die Einträge in der Matrix $\Phi_m^\top \Phi_m$ erhalten wir

$$(\Phi_m^\top \Phi_m)_{i,j} = \sum_{k=1}^m f_i(t_k) f_j(t_k) = m \left(\sum_{k=1}^m h f_i(t_k) f_j(t_k) \right)$$

für $i, j = 1, \ldots, n$. Da die Ansatzfunktionen als stetig vorausgesetzt sind, konvergieren die Audrücke gegen die Riemann-Integrale

$$\lim_{m\to\infty} \sum_{k=1}^{m} h f_i(t_k) f_j(t_k) = \int_0^1 f_i(t) f_j(t)\ \mathrm{d}t = b_{ij}$$

für $i, j = 1, \ldots, n$. Wir erhalten somit die Matrix $B \in \mathbb{R}^{n\times n}$ aus (i), welche symmetrisch und positiv definit ist wegen der linearen Unabhängigkeit der Funktionen $f_1, \ldots, f_n$. Es folgt

$$\lim_{m\to\infty} \tfrac{1}{m}(\Phi_m^\top \Phi_m) = B \qquad \text{und} \qquad \lim_{m\to\infty} m(\Phi_m^\top \Phi_m)^{-1} = B^{-1}.$$

Somit ist notwendigerweise $(\Phi_m^\top \Phi_m)^{-1} = \mathcal{O}(\frac{1}{m})$, d.h. jede Komponente der Matrix $(\Phi_m^\top \Phi_m)^{-1}$ konvergiert gegen null. Analog gilt

$$\lim_{m\to\infty} \tfrac{1}{m}(\Phi_m^\top \hat{z}_m)_i = \lim_{m\to\infty} \sum_{k=1}^{m} h f_i(t_k)\hat{y}(t_k) = \int_0^1 f_i(t)\hat{y}(t)\ \mathrm{d}t = c_i$$

für $i = 1, \ldots, n$. Wir erhalten dadurch

$$\lim_{m\to\infty} \hat{\theta}_m = \hat{\theta}$$

für die Lösungen aus den Normalgleichungen $(\Phi_m^\top \Phi_m)\hat{\theta}_m = \Phi_m^\top \hat{z}_m$.

iii) Der Vektor e_m ist eine m-dimensionale normalverteilte Zufallsvariable mit Erwartungswert null und Kovarianzmatrix $\sigma^2 I$ mit der Einheitsmatrix $I \in \mathbb{R}^{m\times m}$. Die involvierte Matrix $T = (\Phi_m^\top \Phi_m)^{-1}\Phi_m^\top \in \mathbb{R}^{n\times m}$ besitzt den Rang n nach Voraussetzung. Laut Lemma 3.1 ist θ_m jeweils normalverteilt mit Erwartungswert $\mu_m \in \mathbb{R}^n$

$$\mu_m = (\Phi_m^\top \Phi_m)^{-1}\Phi_m^\top \hat{z}_m = \hat{\theta}_m$$

und Kovarianzmatrix $\Sigma_m \in \mathbb{R}^{n\times n}$

$$\begin{aligned}
\Sigma_m &= \left((\Phi_m^\top \Phi_m)^{-1}\Phi_m^\top\right)(\sigma^2 I)\left((\Phi_m^\top \Phi_m)^{-1}\Phi_m^\top\right)^\top \\
&= \sigma^2(\Phi_m^\top \Phi_m)^{-1}\Phi_m^\top \Phi_m((\Phi_m^\top \Phi_m)^{-1})^\top \\
&= \sigma^2(\Phi_m^\top \Phi_m)^{-1}
\end{aligned}$$

wegen der Symmetrie von $\Phi_m^\top \Phi_m$. Somit folgt komponentenweise wegen (ii)

$$\lim_{m\to\infty} \Sigma_m = \sigma^2 \lim_{m\to\infty} (\Phi_m^\top \Phi_m)^{-1} = 0.$$

Die Kovarianzmatrix enthält auf ihrer Diagonale die Varianzen der Zufallsvariablen $\theta_m = (\theta_{m,1}, \ldots, \theta_{m,n})^\top$.

iv) Die Tschebyscheff-Ungleichung liefert komponentenweise

$$P\left[\left|\theta_{m,j} - \hat{\theta}_{m,j}\right| \geq \eta\right] \leq \frac{\mathrm{Var}(\theta_{m,j})}{\eta^2} \qquad \text{für } j = 1, \ldots, n.$$

Wir verwenden o.E.d.A. die Vektornorm $\|\cdot\|_\infty$. Die Monotonie des Wahrscheinlichkeitsmaßes erlaubt die Abschätzung

$$\begin{aligned} P\left[\left\|\theta_m - \hat{\theta}_m\right\|_\infty \geq \eta\right] &= P\left[\max_{j=1,\ldots,n}\left|\theta_{m,j} - \hat{\theta}_{m,j}\right| \geq \eta\right] \\ &= P\left(\bigcup_{j=1}^{n}\left\{\omega \in \Omega \,:\, \left|\theta_{m,j}(\omega) - \hat{\theta}_{m,j}\right| \geq \eta\right\}\right) \\ &\leq \sum_{j=1}^{n} P\left[\left|\theta_{m,j} - \hat{\theta}_{m,j}\right| \geq \eta\right] \leq \frac{1}{\eta^2}\sum_{j=1}^{n}\mathrm{Var}(\theta_{m,j}). \end{aligned}$$

Weil die Dimension n fest ist und alle Varianzen jeweils gegen null konvergieren (siehe (iii)), folgt die Konvergenz

$$\lim_{m\to\infty} P\left[\left\|\theta_m - \hat{\theta}_m\right\|_\infty \geq \eta\right] = 0.$$

v) Für $\eta > 0$ gegeben existiert wegen (ii) ein m' mit $\|\hat{\theta}_m - \hat{\theta}\| < \frac{\eta}{2}$ für alle $m \geq m'$. Sowohl $\hat{\theta}_m$ als auch $\hat{\theta}$ sind nicht zufallsabhängig. Wegen der Implikationen

$$\|\theta_m - \hat{\theta}\| > \eta \quad \Rightarrow \quad \|\theta_m - \hat{\theta}_m\| + \|\hat{\theta}_m - \hat{\theta}\| > \eta \quad \Rightarrow \quad \|\theta_m - \hat{\theta}_m\| > \tfrac{\eta}{2}$$

für alle $m \geq m'$ folgt mit der Monotonie des Wahrscheinlichkeitsmaßes

$$0 \leq P\left[\left\|\theta_m - \hat{\theta}\right\| > \eta\right] \leq P\left[\left\|\theta_m - \hat{\theta}_m\right\| > \frac{\eta}{2}\right] \qquad \text{für } m \geq m'.$$

Wir haben in (iv) gezeigt, dass die Wahrscheinlichkeiten auf der rechten Seite gegen null konvergieren. Damit erhalten wir die behauptete Konvergenz in Wahrscheinlichkeit. □

Die Konvergenzaussage des Satzes 3.1 kann statt für äquidistante Zeitpunkte auch für eine größere Menge von Gitterfolgen erhalten werden.

Beispiel:

Wir geben nun die drei Ansatzfunktionen $f_1(t) = 1$, $f_2(t) = \sin(2\pi t)$, $f_3(t) = \cos(2\pi t)$ vor, d.h. das Modell

$$y(t) = \theta_1 + \theta_2 \sin(2\pi t) + \theta_3 \cos(2\pi t)$$

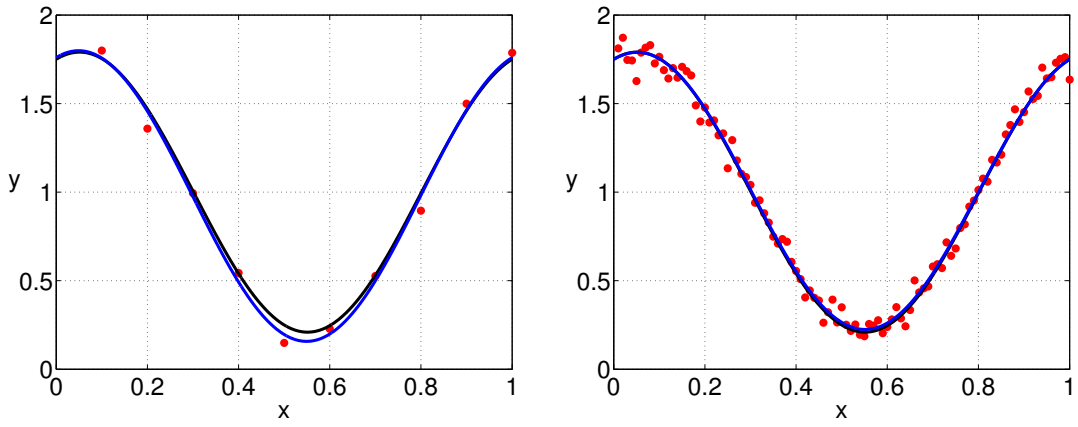


Abbildung 15: Exakte Funktion $\hat{y}(t)$ (schwarz) zusammen mit Messwerten (rot) und daraus konstruierter Approximation $y(t)$ (blau) für $m = 10$ (links) und $m = 100$ (rechts).

für $t \in [0, 1]$. Die spezielle Wahl $\hat{\theta}_1 = 1$, $\hat{\theta}_2 = \frac{1}{4}$, $\hat{\theta}_3 = \frac{3}{4}$ wird festgelegt und die Ergebnisse bezeichne $\hat{y}(t)$. In diesem künstlichen Fall tritt somit bei $\hat{\theta}$ kein Modellfehler auf, während $\theta \neq \hat{\theta}$ einen Modellfehler verursacht.

Wir simulieren Messungen mit Messfehlern durch $y_i = \hat{y}(t_i) + \varepsilon_i$ mit Pseudo-Zufallszahlen ε_i zu unabhängigen Normalverteilungen mit Erwartungswert 0 und Standardabweichung $\sigma = 0.05$. Es sei $t_i = \frac{i}{m}$ für $i = 1, \ldots, m$. Abbildung 15 zeigt die für $m = 10$ und $m = 100$ berechneten Modelle. Wir erkennen eine bessere Approximation für die höhere Anzahl von Messungen. Tabelle 1 zeigt die Lösungen aus den linearen Ausgleichsproblemen für verschiedene Anzahlen von Messungen. Wir stellen die Konvergenz gegen die gesuchten Parameterwerte fest.

Tabelle 1: Lösungen des Ausgleichproblems für verschiedene Anzahlen von Messungen.

Anzahl Messungen	10	100	1000	10000	100000
θ_1	0.9780	1.0071	1.0009	0.9994	1.0000
θ_2	0.2495	0.2484	0.2521	0.2493	0.2503
θ_3	0.7817	0.7420	0.7476	0.7498	0.7504

Nichtlineare Ausgleichsrechnung

Wir betrachten nun ein Modell $y(t) = f(t, \theta)$ mit $f : [t_0, t_{\text{end}}] \times \Theta \to \mathbb{R}$ und Parametermenge $\Theta \subseteq \mathbb{R}^n$, wobei f nichtlinear von den Parametern θ abhängt. Es sei f stetig in t und stetig differenzierbar in θ. Zu Zeitpunkten $t_0 \leq t_1 < t_2 < \cdots < t_m \leq t_{\text{end}}$ werden wieder Messungen $y_i = \hat{y}(t_i) + \varepsilon_i$ für $i = 1, \ldots, m$ durchgeführt. Wieder bezeichnet $\hat{y}(t)$ die „Realität“ und ε_i die Messfehler. Die Methode der kleinsten Quadrate zur Bestimmung der Parameter lautet

$$\min_{\theta \in \Theta} \sum_{i=1}^{m} \left(y_i - f(t_i, \theta)\right)^2 .$$

Um das Problem allgemeiner zu formulieren erfolgt ein Notationswechsel: $x = \theta$, $D = \Theta$, $F : D \to \mathbb{R}^m$ mit $F = (F_1, \ldots, F_m)^\top$, $F_i = y_i - f(t_i, \theta)$.

Def. 3.2 *Sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^m$ eine stetige nichtlineare Funktion. Das zugehörige (nichtlineare) Ausgleichsproblem lautet*

$$\min_{x \in D} \tfrac{1}{2} \|F(x)\|^2 \tag{3.5}$$

mit der Euklidischen Norm $\|\cdot\|$, wobei Existenz und Eindeutigkeit des Minimums vorausgesetzt wird.

Das Minimierungsproblem (3.5) bedeutet gerade

$$\min_{x \in D} \tfrac{1}{2} \sum_{i=1}^{m} F_i(x)^2.$$

Im allgemeinen wird F auch als stetig differenzierbar vorausgesetzt. Dadurch existiert die Jacobi-Matrix

$$\mathrm{D}F : D \to \mathbb{R}^{m \times n}, \quad (\mathrm{D}F)_{i,j} = \frac{\partial F_i}{\partial x_j} \quad \text{für } \; i = 1, \ldots, m, \; j = 1, \ldots, n.$$

Minimiert werden soll also die Funktion

$$r(x) = \tfrac{1}{2} \|F(x)\|^2.$$

Der Gradient dieser Funktion ist

$$\operatorname{grad} r(x) = \mathrm{D}F(x)^\top F(x) \in \mathbb{R}^n$$

für $x \in D$. Prinzipiell ist jedes Verfahren zur Minimierung von r anwendbar um das Ausgleichsproblem (3.5) numerisch zu lösen. Dies sind typischerweise iterative Verfahren, welche eine Folge von Näherungen $(x^{(k)})_{k \in \mathbb{N}}$ erzeugen. Sie benötigen jeweils einen Startwert $x^{(0)} \in D$.

Wir betrachten drei wichtige Verfahren.

1. *Gradientenverfahren*

Sei $x^{(k)} \in D$ gegeben. Wir definieren als Suchrichtung den negativen Gradienten

$$\Delta x^{(k)} = -\mathrm{D}F(x^{(k)})^\top F(x^{(k)}). \tag{3.6}$$

Daher nennt man dieses Verfahren auch die Methode des steilsten Abstiegs.

Man setze $\lambda = 1$. Der Ansatz für die neue Näherung ist

$$x^{(k+1)} = x^{(k)} + \lambda \Delta x^{(k)}. \tag{3.7}$$

Falls $r(x^{(k+1)}) < r(x^{(k)})$ gilt, dann wird die Näherung $x^{(k+1)}$ akzeptiert und der nächste Iterationsschritt begonnen. Anderenfalls wird $\lambda = \frac{\lambda}{2}$ gesetzt und erneut das Abstiegskriterium überprüft.

Man kann zeigen, dass ein $\lambda_0 > 0$ existiert, so dass

$$r(x^{(k)} + \lambda \Delta x^{(k)}) < r(x^{(k)}) \qquad \text{für alle } \lambda \in (0, \lambda_0).$$

Das Abstiegskriterium ist also irgendwann für ein hinreichend kleines λ erfüllt.

Das Gradientenverfahren läßt sich zur Minimierung jeder reellwertigen differenzierbaren Funktion $r : D \to \mathbb{R}$ einsetzen. Leider ist die Konvergenz häufig langsam. Zudem besteht die Gefahr in einem lokalen Minimum stehen zu bleiben.

2. *Gauß-Newton-Verfahren*

Diese Methode stellt eine Übertragung des Newton-Verfahrens für Systeme aus nichtlinearen Gleichungen auf nichtlineare Ausgleichsprobleme dar.

Die Funktion F wird hier um einen gegebenen Wert $x^{(k)} \in D$ linearisiert, d.h. man approximiert

$$F(x) \approx F(x^{(k)}) + \mathrm{D}F(x^{(k)})(x - x^{(k)}).$$

Statt der exakten Funktion F wird nun diese Approximation minimiert

$$\min_{\Delta x^{(k)} \in \mathbb{R}^n} \tfrac{1}{2} \left\| F(x^{(k)}) + \mathrm{D}F(x^{(k)}) \Delta x^{(k)} \right\|^2 \tag{3.8}$$

mit $\Delta x^{(k)} = x - x^{(k)}$. Dies ist gerade ein lineares Ausgleichsproblem für die Unbekannte $\Delta x^{(k)}$. Die Lösung ist somit gegeben durch die Normalgleichungen

$$\left(\mathrm{D}F(x^{(k)})^\top \mathrm{D}F(x^{(k)})\right) \Delta x^{(k)} = -\mathrm{D}F(x^{(k)})^\top F(x^{(k)}). \tag{3.9}$$

Der Algorithmus für einen Iterationsschritt des Gauß-Newton-Verfahrens ergibt sich zu:

1. Werte $F(x^{(k)})$ und $\mathrm{D}F(x^{(k)})$ aus.
2. Löse lineares Ausgleichsproblem (3.8) nach Unbekannten $\Delta x^{(k)}$.
3. Setze $x^{(k+1)} = x^{(k)} + \Delta x^{(k)}$.

Da sogar für Werte nahe des exakten Mininums diese Iteration divergieren kann, wird ein gedämpftes Gauß-Newton-Verfahren eingesetzt. Der Dämpfungsfaktor sei $\lambda_k \in (0, 1]$. Dazu werden Konstanten $0 < \delta < \frac{1}{2}$, z.B. $\delta = 0.01$, und $0 < \alpha < 1$, z.B. $\alpha = \frac{1}{2}$ gewählt. Der obige 3. Schritt wird ersetzt durch (3.7) und λ_k wird maximal in der Folge $1, \alpha, \alpha^2, \alpha^3, \ldots$ gewählt, so dass das Kriterium

$$r(x^{(k)}) - r(x^{(k+1)}) \geq \delta \, \lambda_k \, \|\mathrm{D}F(x^{(k)})\Delta x^{(k)}\|^2$$

erfüllt ist, vgl. auch [1]. Insbesondere gilt dann $r(x^{(k+1)}) < r(x^{(k)})$.

Für die Konvergenz dieser Methode gilt das folgende Resultat, siehe Satz 7.3 in [22].

Satz 3.2 *Sei $D \subset \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^m$ mit $m \geq n$ und $F \in C^2(D)$. Desweiteren existiere ein $\bar{y} \in D$, so dass die Menge*

$$L = \{x \in D \ : \ r(x) \leq r(\bar{y})\}$$

kompakt ist und $\mathrm{rang}(\mathrm{D}F(x)) = n$ *für alle $x \in L$ gilt. Dann konvergiert das gedämpfte Gauß-Newton-Verfahren für jedes $x^{(0)} \in L$ gegen ein lokales Minimum von r in L.*

Dieser Satz garantiert die Konvergenz gegen eine Minimalstelle, jedoch nicht notwendigerweise gegen ein globales Minimum falls mehrere lokale Minima existieren. Die Konvergenzgeschwindigkeit des gedämpften Gauß-Newton-Verfahrens wird aber nicht immer lokal quadratisch, d.h. es kann auch nur lineare Konvergenz über die gesamte Iteration hinweg vorliegen.

3. *Levenberg-Marquardt-Verfahren*

In dieser Methode wird zu gegebenem $x^{(k)} \in D$ die Korrektur $\Delta x^{(k)}$ mittels einer Modifikation von (3.8) aus dem Gauß-Newton-Verfahren bestimmt, nämlich

$$\min_{\Delta x^{(k)} \in \mathbb{R}^n} \left\| F(x^{(k)}) + \mathrm{D}F(x^{(k)}) \Delta x^{(k)} \right\|^2 + \mu^2 \left\| \Delta x^{(k)} \right\|^2$$

mit einem Regularisierungsparameter $\mu \geq 0$. Dieses Ausgleichsproblem kann auch geschrieben werden als

$$\min_{\Delta x^{(k)} \in \mathbb{R}^n} \left\| \begin{pmatrix} F(x^{(k)}) \\ 0 \end{pmatrix} + \begin{pmatrix} \mathrm{D}F(x^{(k)}) \\ \mu I \end{pmatrix} \Delta x^{(k)} \right\|^2 \tag{3.10}$$

mit der Einheitsmatrix $I \in \mathbb{R}^{n \times n}$. Aus (3.10) erhält man die Normalgleichungen

$$\left(\mathrm{D}F(x^{(k)})^\top \mathrm{D}F(x^{(k)}) + \mu^2 I \right) \Delta x^{(k)} = -\mathrm{D}F(x^{(k)})^\top F(x^{(k)})$$

für die unbekannte Lösung. Interessant sind nun die Grenzfälle für kleine und große Regularisierungsparameter.

1. Fall: $\mu \to 0$
Für $\mu = 0$ entsteht gerade die Vorschrift (3.9), d.h. die Iteration wird zum (ungedämpften) Gauß-Newton-Verfahren.

2. Fall: $\mu \to \infty$
Für $\mu \gg 1$ wird die Matrix $\mathrm{D}F(x^{(k)})^\top \mathrm{D}F(x^{(k)})$ vernachlässigbar gegenüber $\mu^2 I$. Die Korrektur wird dadurch näherungsweise

$$\Delta x^{(k)} \approx -\tfrac{1}{\mu^2} \mathrm{D}F(x^{(k)})^\top F(x^{(k)}),$$

d.h. die Suchrichtung (3.6) liegt vor. Es entsteht hier das Gradientenverfahren, wobei wegen $\frac{1}{\mu^2} \ll 1$ jedoch nur ein kleines Stück in Richtung des steilsten Abstiegs gegangen wird.

In der Praxis kann eine geeignete Folge μ_k von Regularisierungsparametern während der Iteration bestimmt werden. Diese sollen nicht zu klein sein, da oft Divergenz resultiert. Sie sollen auch nicht zu groß sein, weil sonst nur kleine Korrekturen geliefert werden.

Ein Vorteil des Levenberg-Marquardt-Verfahrens ist, dass der Konvergenzbereich häufig größer als beim Gauß-Newton-Verfahren ist und wenn die Iteration später nahe der gesuchten Lösung ist wird automatisch $\mu \approx 0$ verwendet, wodurch die lokale quadratische Konvergenz der Gauß-Newton-Methode erhalten wird.

3.3 Parameterbestimmung bei dynamischen Systemen

In diesem Abschnitt behandeln wir die Identifikation von Parametern in Modellen für zeitabhängige Prozesse. Das zugrunde liegende physikalische System besitzt eine Eingabe von Signalen $u(t) \in \mathbb{R}^{n_{\text{in}}}$ und eine Ausgabe von beobachtbaren Ergebnissen $y(t) \in \mathbb{R}^{n_{\text{out}}}$, siehe das Schema in Abbildung 16. Zur Vereinfachung betrachten wir in diesem Abschnitt nur Systeme mit einem einzelnen Eingabesignal, d.h. $n_{\text{in}} = 1$.

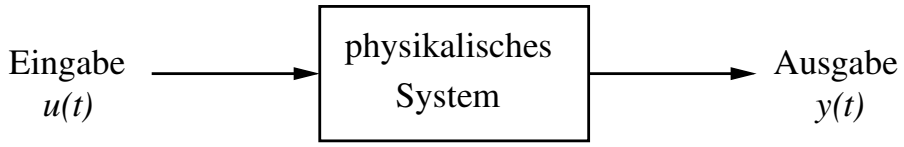


Abbildung 16: Schema eines dynamischen Systems.

Validierung

Es bezeichne $\hat{y}$ die „Realität“ aus dem physikalischen System, y die exakte Lösung eines mathematischen Modells (Gleichungen) und $\tilde{y}$ die Näherungslösung aus einem numerischen Verfahren. Der Gesamtfehler kann zerlegt werden in

$$\underbrace{\hat{y} - \tilde{y}}_{\text{Gesamtfehler}} = \underbrace{\hat{y} - y}_{\text{Modellfehler}} + \underbrace{y - \tilde{y}}_{\text{Verfahrensfehler}} .$$

Bei der Analyse der Fehleranteile unterscheidet man zwei Begriffe:

- *Validierung* bedeutet die Überprüfung, ob der Modellfehler hinreichend klein ist, d.h. die Qualität eines Modells wird analysiert. (Fragestellung: „Werden die richtigen Gleichungen gelöst?“)
- *Verifizierung* bedeutet die Überprüfung, ob der Verfahrensfehler hinreichend klein ist, d.h. die Qualität einer numerischen Methode wird analysiert. (Fragestellung: „Werden die Gleichungen richtig gelöst?“)

In diesem Abschnitt ist nur die Validierung relevant. Wir nehmen gegebenenfalls an, dass die Verfahrensfehler vernachläßigbar klein gegenüber den Modellfehlern sind.

Für die Parameterbestimmung mit nachfolgender Validierung werden bei dynamischen Modellen typischerweise vier Schritte durchgeführt:

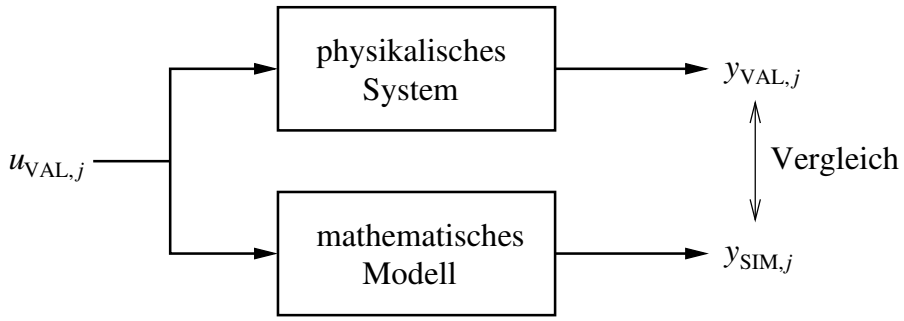


Abbildung 17: Schema der Validierung eines Modells.

1. *Messphase* :
   An das physikalische System werden nacheinander die verschiedenen Eingaben $u_1, \ldots, u_\ell$ angelegt und die zugehörigen Ausgaben $y_1, \ldots, y_\ell$ gemessen. Die Eingaben liegen dabei meist als zeitkontinuierliche Funktionen vor, während die Messungen nur an diskreten Zeitpunkten erfolgen.

2. *Aufteilung der Daten* :
   Es erfolgt eine Aufspaltung der Eingaben und Ausgaben in zwei Gruppen: PB für Parameterbestimmung und VAL für Validierung, d.h.

$$\begin{array}{ll} u_{\mathrm{PB}} : \{u_1, \ldots, u_k\}, & u_{\mathrm{VAL}} : \{u_{k+1}, \ldots, u_\ell\}, \\ y_{\mathrm{PB}} : \{y_1, \ldots, y_k\}, & y_{\mathrm{VAL}} : \{y_{k+1}, \ldots, y_\ell\}. \end{array}$$

   Als Faustregel sollen beide Gruppen etwa gleich groß sein.

3. *Parameterschätzung* :
   Verwende u_{PB} und y_{PB} in einer Ausgleichsrechnung, um geeignete Parameter θ zu bestimmen. Hier kommen die Methoden aus Abschnitt 3.2 zum Einsatz.

4. *Validierung* :
   Das mathematische Modell mit der Parameterwahl aus dem 3. Schritt wird für die Eingaben u_{VAL} numerisch simuliert. Die Ergebnisse y_{SIM} vergleicht man mit den Messungen y_{VAL}. Liegt eine gute Übereinstimmung vor, so deutet dies auf eine geeignete Wahl der Parameter hin.

Ein Anteil des 1. Schritts und der 4. Schritt sind in Abbildung 17 schematisch dargestellt.

Wahl des Eingangsignals

Wir stellen zwei Beispiele für typische Wahlen des Eingangsignals u dar:

(i) *Pseudo-Zufall-Binär-Signal*
Dieses Signal nimmt nur zwei verschiedene Werte an, d.h.

$$u(t) = \begin{cases} +U_0 & \text{für } t \in S \\ -U_0 & \text{für } t \notin S \end{cases}$$

mit einer Konstante $U_0 > 0$ und einer Teilmenge $S \subseteq [t_0, t_{\text{end}}]$. Der Zeitbereich $[t_0, t_{\text{end}}]$ wird in Teilintervalle identischer Länge Δt aufgespaltet und u ist auf jedem Teilintervall konstant. Vorgegeben wird auch eine Übergangswahrscheinlichkeit p mit $0 < p < 1$. Nach jedem Teilintervall erfolgt mit Wahrscheinlichkeit p ein Sprung zum anderen möglichen Wert und mit Wahrscheinlichkeit $1-p$ wird der aktuelle Wert beibehalten. Dieses Zufallsexperiment wird auf dem Rechner mit Pseudo-Zufallszahlen simuliert, siehe z.B. Abschnitt 5.2 in [9]. Man beachte, dass dieses Eingangssignal unstetig in t ist, d.h. in einem entsprechenden mathematischen Modell müssen unstetige Funktionen u zugelassen sein. Abbildung 18 (links) zeigt ein Beispiel.

(ii) *Chirp-Signal*
Das Signal stellt hier eine frequenzmodulierte Schwingung mit in der Zeit zunehmender Frequenz dar. Man definiert

$$u(t) = U_0 \sin(2\pi f(t)t) \tag{3.11}$$

mit einer vorzugebenden Konstante $U_0 > 0$. Für die Frequenz kann beispielsweise $f(t) = K_f t$ gesetzt werden, wobei dann die Festsetzung der Konstante $K_f > 0$ ein Freiheitsgrad darstellt. Ein Vorteil ist, dass dieses Signal beliebig oft stetig differenzierbar ist. Abbildung 18 (rechts) verdeutlicht diesen Signaltyp.

Diese Signale wurden z.B in [12] verwendet.

Anwendung bei Differentialgleichungen

Wir betrachten ein System aus gewöhnlichen Differentialgleichungen

$$x'(t) = g(t, x(t), \theta) \tag{3.12}$$

mit $g : [t_0, t_{\text{end}}] \times \mathbb{R}^k \times \Theta \to \mathbb{R}^k$. Eingangssignale $u(t)$ können somit in g enthalten sein. Die Lösung $x : [t_0, t_{\text{end}}] \to \mathbb{R}^k$ hängt damit auch von den Parametern $\theta \in \Theta \subseteq \mathbb{R}^n$ ab. Die Parameter sollen nun geschätzt werden.

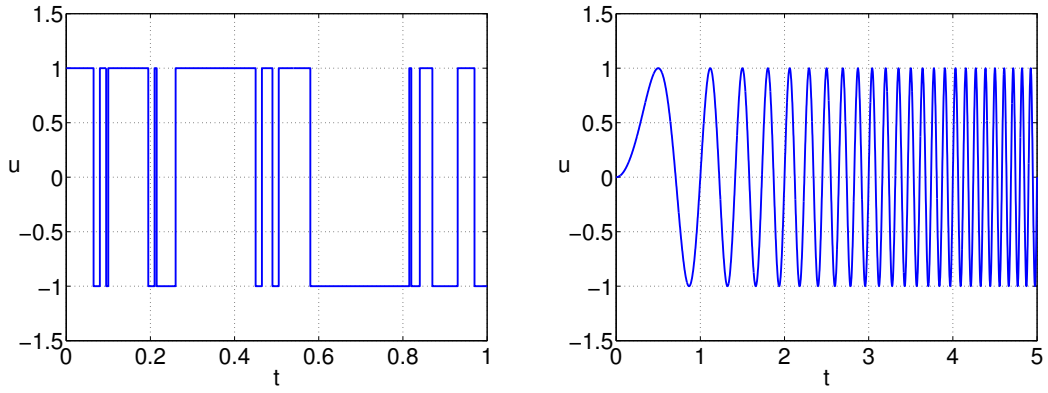


Abbildung 18: Beispiel eines Pseudo-Zufall-Binär-Signals mit $\Delta t = 0.005$ und $p = 0.1$ (links) und Chirp-Signal mit $K_f = 1$ (rechts).

1. Strategie: „Setze Messungen in Differenzengleichung ein."

Wir nehmen an, dass Messungen $x_1, x_2, \ldots, x_m$ aller Lösungskomponenten zu den Zeitpunkten $t_0 \leq t_1 < t_2 < \cdots < t_m$ vorliegen. Wir betrachten eine Teilmenge des Gleichungssystems spezifiziert durch die Indexmenge $\mathcal{I} \subseteq \{1, \ldots, k\}$. Voraussetzung ist eine rechte Seite in (3.12) der Form

$$g_\ell(t, x, \theta) = \theta_1 g_{\ell 1}(t, x) + \theta_2 g_{\ell 2}(t, x) + \cdots + \theta_n g_{\ell n}(t, x) + s_\ell(t, x) \tag{3.13}$$

für alle $\ell \in \mathcal{I}$. Ableitungen auf der linken Seite in (3.12) werden durch Differenzenquotienten ersetzt. Es bestehen folgende Möglichkeiten der Approximation der Ableitungen

$$\begin{aligned}
x'(t_i) &\approx \tfrac{1}{t_{i+1}-t_i}\,[x(t_{i+1}) - x(t_i)] && \text{(Vorwärts-Diff.)},\\
x'(t_i) &\approx \tfrac{1}{t_i-t_{i-1}}\,[x(t_i) - x(t_{i-1})] && \text{(Rückwärts-Diff.)},\\
x'(t_i) &\approx \tfrac{1}{t_{i+1}-t_{i-1}}\,[x(t_{i+1}) - x(t_{i-1})] && \text{(symmetrische Diff.)}.
\end{aligned}$$

Sei $z : [t_0, t_{\text{end}}] \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Für Messungen $z_i = z(t_i) + \varepsilon_i$ gelte $|\varepsilon_i| \leq \bar{\varepsilon}$ für alle i. Für den Approximationsfehler der Vorwärts-Differenzen folgt

$$\left| z'(t_i) - \frac{z_{i+1} - z_i}{\Delta t} \right| \leq C_1 \Delta t + C_2 \frac{\bar{\varepsilon}}{\Delta t}$$

mit Konstanten $C_1, C_2 > 0$ und $\Delta t = t_{i+1} - t_i$. Analoge Formeln gelten für die beiden anderen Differenzenquotienten. Eine zu kleine Zeitschrittweite bewirkt daher eine Verstärkung der Messfehler. Dagegen führt eine zu große Zeitschrittweite auch auf eine schlechte Approximation der Ableitung. Nähere Erläuterungen kann man in [23] finden.

Die ℓ-te Differentialgleichung des Systems (3.12) wird nun durch eine Differenzengleichung ersetzt, d.h.

$$\frac{1}{t_{i+1}-t_i}\left[x_{\ell,i+1} - x_{\ell,i}\right] = \theta_1 g_{\ell 1}(t_i, x_i) + \cdots + \theta_n g_{\ell n}(t_i, x_i) + s_\ell(t_i, x_i)$$

für $i = 1, \ldots, m-1$ und alle $\ell \in \mathcal{I}$. Dadurch entsteht ein lineares Ausgleichsproblem

$$\min_{\theta \in \Theta} \|\Phi\theta - y\|^2$$

für die Parameterwerte θ mit $(m-1)|\mathcal{I}|$ Gleichungen. Im Spezialfall $\mathcal{I} = \{\ell\}$ folgt im Ausgleichsproblem eine Matrix Φ mit Komponenten

$$\Phi_{ij} = g_{\ell j}(t_i, x_i) \qquad \text{für } \; i = 1, \ldots, m-1 \; \text{ und } \; j = 1, \ldots, n$$

und ein Vektor y mit Komponenten

$$y_i = \frac{1}{t_{i+1}-t_i}\left[x_{\ell,i+1} - x_{\ell,i}\right] - s_\ell(t_i, x_i) \qquad \text{für } \; i = 1, \ldots, m-1.$$

Vorteile:

- Lineares Ausgleichsproblem ist direkt lösbar.
- Wenig Rechenaufwand. Differentialgleichungen brauchen nicht gelöst zu werden.

Nachteile:

- Differentialgleichungen müssen als Formeln gegeben sein und betrachtet bzw. umgeformt werden.
- Messungen müssen für alle Komponenten x_j vorliegen, die in den Gleichungen $\ell \in \mathcal{I}$ auftreten.
- Messfehler wirken sich bei Näherungen der Ableitungen stark aus.

Die Strategie 1 kann auch bei allgemeiner rechter Seite in (3.12) angewendet werden. Jedoch entsteht dann ein nichtlineares Ausgleichsproblem.

2. Strategie: „Vergleiche Messungen mit Lösung der Differentialgleichung."

Zu dem Differentialgleichungssystem (3.12) muss jetzt eine Anfangswertvorgabe

$$x(t_0) = x_0 \tag{3.14}$$

gemacht werden, wobei x_0 geeignet festzulegen ist. Desweiteren wird aus der Lösung des Systems (3.12) eine Ausgabe

$$y(t) = f(x(t)) \qquad \text{mit } f : \mathbb{R}^k \to \mathbb{R}$$

definiert. Beispielsweise kann dies einfach eine Lösungskomponente von x sein, d.h. $y(t) = x_j(t)$ für ein $j \in \{1, \ldots, k\}$. Zum einen seien in Zeitpunkten $t_0 \leq t_1 < t_2 < \cdots < t_m$ Messungen $y_1, y_2 \ldots, y_m$ der Ausgabe gegeben. Zum anderen erhalten wir aus der Lösung des Anfangswertproblems (3.12),(3.14) mit fester Parameterwahl aus einem numerischen Verfahren Näherungen

$$x(t_i) \approx \tilde{x}_i, \quad y(t_i) \approx \tilde{y}_i = f(\tilde{x}_i) \quad \text{für } i = 1, \ldots, m.$$

Dadurch hängt $\tilde{y}_1, \ldots, \tilde{y}_m$ von θ ab. Ein Vergleich der Messungen mit den Näherungen der Lösung der Differentialgleichungen führt auf das nichtlineare Ausgleichsproblem

$$\min_{\theta \in \Theta} \sum_{i=1}^{m} \left(y_i - \tilde{y}_i(\theta)\right)^2. \tag{3.15}$$

Die Auswertung von $\tilde{y}_1, \ldots, \tilde{y}_m$ für festes θ erfordert jeweils die numerische Lösung des Anfangswertproblems (3.12),(3.14) über das gesamte Zeitintervall $[t_0, t_m]$.

Vorteile:

- Differentialgleichungen brauchen nicht als Formeln gegeben zu sein und müssen nicht betrachtet werden.
- Messungen sind nur für die Ausgabe $y(t)$ erforderlich.
- Messfehler wirken sich oft nicht stark aus.

Nachteile:

- Hoher Rechenaufwand. Anfangswertprobleme von Differentialgleichungen müssen numerisch gelöst werden.
- Ein Iterationsverfahren zur Lösung des nichtlinearen Ausgleichsproblems kann divergieren, d.h. keine Näherungslösung wird geliefert.

Die Strategie 2 läßt sich auch bei dymanischen Modellen einsetzen, die nicht auf Differentialgleichungen basieren.

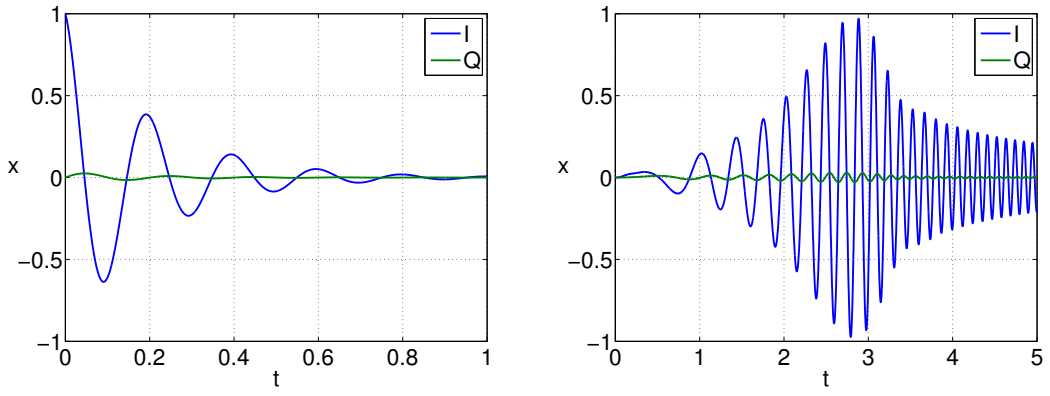


Abbildung 19: Lösung des Differentialgleichungssystems (3.2) ohne Eingangssignal (links) und mit frequenzmoduliertem Eingangssignal (rechts).

Beispiel:

Wir greifen als Beispiel das Differentialgleichungsmodell (3.2) einer elektrischen Schaltung aus Abschnitt 3.1 auf. Es sei $R = 1$, $L = 0.1$, $C = 0.01$. Abbildung 19 zeigt die Lösung von zwei Anfangswertproblemen: für Eingabesignal $u \equiv 0$ und für das Chirp-Signal (3.11) mit $U_0 = 1$ und $f(t) \equiv t$ als Eingabe.

Wir wenden Strategie 1 an. Die erste Gleichung im System (3.2) können wir schreiben als

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\theta_1 Q(t) - \theta_2 I(t) + \theta_3 u(t)$$

mit $\theta_1 = \frac{1}{LC}$, $\theta_2 = \frac{R}{L}$, $\theta_3 = \frac{1}{L}$, wodurch wir die Form (3.13) erzeugt haben. Die gesuchten Parameter folgen dann aus $R = \frac{\theta_2}{\theta_3}$, $L = \frac{1}{\theta_3}$, $C = \frac{\theta_3}{\theta_1}$. Als Eingabe $u(t)$ verwenden wir das Chirp-Signal (3.11) mit $f(t) \equiv t$ und $U_0 = 1$. Im Zeitintervall $[0, 5]$ benötigen wir Messungen für I und Q in Zeitpunkten mit Abstand Δt. Wir simulieren diese Messungen, indem zu einer Näherungslösung eines Anfangswertproblems in den Zeitpunkten Zufallszahlen für unabhängige Normalverteilungen mit Erwartungswert null und Varianz σ^2 hinzuaddiert werden. Es entsteht ein lineares Ausgleichsproblem mit $m - 1 = \frac{5}{\Delta t} - 1$ Gleichungen.

Tabelle 2 zeigt die Ergebnisse für verschiedene Wahlen der Zeitschrittweite und der Standardabweichungen. Wir erkennen, dass die Güte der Schätzung kritisch von den Messfehlern abhängt. Für hinreichend kleine Messfehler entsteht jedoch eine sehr gute Approximation der gesuchten Parameter.

Wir setzen nun Strategie 2 für dieses Problem ein. Hier kann direkt $\theta_1 = R$, $\theta_2 = L$, $\theta_3 = C$ definiert werden. Das gleiche Chirp-Signal wie in Strategie 1 wird verwendet. Als Zeitintervall wird $[0, 2]$ und als Zeitschrittweite $\Delta t = 0.01$

Tabelle 2: Parameterschätzungen aus Strategie 1 für Beispiel des elektromagnetischen Schwingkreises.

	$\Delta t = 0.01$			$\Delta t = 0.001$		
	$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$	$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$
R	0.9549	0.9104	0.9978	0.9118	0.9752	1.0002
L	0.1515	0.1188	0.1059	0.1511	0.1179	0.1007
C	0.4445	0.0147	0.0098	0.6356	0.0156	0.0100

Tabelle 3: Parameterschätzungen aus Strategie 2 für Beispiel des elektromagnetischen Schwingkreises.

	$\Delta t = 0.01$		
	$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$
R	1.1259	1.0045	1.0036
L	0.1098	0.1045	0.0996
C	0.0099	0.0098	0.0100
Anz. It.	16	13	13

angesetzt, d.h. $m = 200$ diskrete Zeitpunkte entstehen. Die Anfangswerte sind stets $I(0) = Q(0) = 0$. Die Anfangswertprobleme der Differentialgleichungen werden näherungsweise mit der Trapez-Regel gelöst. Als Ausgabe sei nur die Variable I definiert, d.h. Messungen benötigen wir jetzt nur noch für diesen Strom. Das nichtlineare Ausgleichsproblem (3.15) wird dann mit dem gedämpften Gauß-Newton-Verfahren (3.9) aus Abschnitt 3.2 iterativ gelöst. Als Startwerte verwenden wir $R_0 = 2$, $L_0 = 0.5$, $C_0 = 0.1$.

Tabelle 3 enthält die Ergebnisse zu diesem Ansatz sowie die benötigte Anzahl an Iterationsschritten. Wir erkennen, dass sich die Näherungslösung viel robuster bezüglich der Messfehler verhält als in Strategie 1. Jedoch sei erwähnt, dass auch das gedämpfte Gauß-Newton-Verfahren für schlechtere Startwerte (z.B. $R_0 = L_0 = C_0 = 1$) in diesem Beispiel divergiert, d.h. wir erhalten dann keine sinnvolle Parameterschätzung.

Literatur

[1] A. Björck: Numerical Methods for Least Squares Problems. SIAM, Philadelphia, 1996.

[2] O. Christensen, K.L. Christensen: Approximation Theory: From Taylor Polynomials to Wavelets. (2. Aufl.) Birkhäuser, Boston, 2005.

[3] R.A. DeVore, G.G. Lorentz: Constructive Approximation. Springer, Berlin, 1993.

[4] G. Faber: Über die interpolatorische Darstellung stetiger Funktionen. Jber. Deutsch. Math.-Verein. 23 (1914), 192–210.

[5] L. Fejér: Beispiele stetiger Funktionen mit divergenter Fourierreihe. Journal für Mathematik. Bd. 137, Heft 1 (1909).

[6] G. Fischer: Lineare Algebra. (18. Aufl.) Springer, Wiesbaden, 2014.

[7] O. Forster: Analysis 2. (10. Aufl.) Springer, Wiesbaden, 2013.

[8] D. Meschede (Hrsg.): Gerthsen Physik. (23. Aufl.) Springer, Berlin 2006.

[9] M. Günther, A. Jüngel: Finanzderivate mit MATLAB. (2. Aufl.) Vieweg + Teubner, Wiesbaden, 2010.

[10] C.A. Hall: On error bounds for spline interpolation. J. Approximation Theory 1 (1968), 209–218.

[11] G. Hämmerlin, K.-H. Hoffmann: Numerische Mathematik. (4. Aufl.) Springer, Berlin, 1994.

[12] H.-P. Halvorsen: System Identification and Estimation in LabVIEW. Tutorial, Telemark University College, Norwegen, 2011.

[13] G. Hübner: Stochastik. (5. Aufl.) Vieweg + Teubner, Wiesbaden, 2009.

[14] A.K. Louis, P. Maaß, A. Rieder: Wavelets. Teubner, Stuttgart, 1994.

[15] J. Marcinkiewicz: Sur l'interpolation d'operations. C. R. Acad. des Sciences 208 (1939), 1272–1273.

[16] R. Plato: Numerische Mathematik kompakt. (4. Aufl.) Vieweg + Teubner, Wiesbaden, 2010.

[17] C. Reinsch: Smoothing by spline functions. Numer. Math. 10 (1967), 177–183.

[18] H.R. Schwarz, N. Köckler: Numerische Mathematik. (8. Aufl.) Vieweg + Teubner, Wiesbaden, 2011.

[19] A. Sharma, A. Meir: Degree of approximation of spline interpolation. J. Math. Mech. 15 (1966), 749–768.

[20] J. Stoer: Numerische Mathematik 1. (9. Aufl.) Springer, Berlin, 2005.

[21] G. Steidl, M. Tasche: Schnelle Fouriertransformation – Theorie und Anwendungen. Lehrbriefe der Fern Universität Hagen, 1996.

[22] W. Törning, P. Spellucci: Numerische Mathematik für Ingenieure und Physiker, Band 1, Numerische Methoden der Algebra. (2. Aufl.) Springer, Berlin, 1988.

[23] W. Törning, P. Spellucci: Numerische Mathematik für Ingenieure und Physiker, Band 2, Numerische Methoden der Analysis. (2. Aufl.) Springer, Berlin, 1990.

[24] D. Werner: Funktionalanalysis. (7. Auf.) Springer, Berlin, 2011.